

Contents **i**

I Data in the World **1**

1 Studies **3**

1.1	Designed experiments	3
	Polio	3
	Gold star designs	10
	The placebo effect	14
	Summary of controls	15
	Moral	19
1.2	Observational studies	20
	Pellagra	20
	Association does not imply causation	22
	Berkeley Graduate School Admissions — Sex bias?	23
	Simpson's paradox	26
	Smoking and death	27
	Does TV harm children's psyche?	29
	Moral	31
	Smoking and lung cancer	32
	Moral	41

2 Looking at Data **43**

2.1	Histograms	43
	Reading a histogram	45
	Creating histograms	49
	Drawing the histogram	53
2.2	The average, median, and standard deviation	58

Average	59
Median	61
Balancing a histogram	65
Average vs. median	70
The standard deviation	73
The root mean square deviation	77
Rule of thumb	79
3 The Normal Curve	83
3.1 The normal curve table	87
3.2 The normal curve for data	94
4 Correlation	101
4.1 Scatter plots	101
4.2 Correlation	106
4.3 The correlation coefficient	110
Calculating the correlation coefficient	113
The point of averages and SD line	118
4.4 Ecological correlations	123
Morals	128
5 Regression	129
5.1 The regression effect	129
National League batting leaders	129
Galton's height data	131
The regression effect in a scatter plot	134
Moral	135
5.2 The Regression Line	136
The equation of a line	140
Interpreting the slope	146
5.3 Errors from the regression Line	149
SD of errors	152
Calculating the SD_{errors}	153
The rule of thumb for scatter plots	155
5.4 Using the normal curve	157

II	Box Models	161
6	Chance	163
6.1	Drawing from a box	164
6.2	Calculating chances	166
	Drawing more than one	169
	Multiplying the chances	171
6.3	Conditional chance	181
6.4	Independence	185
	The rule for independence	187
	Multiplying chances	194
	Robbery	196
	Moral	197
6.5	Monty Hall	198
7	The Sum of the Draws	201
7.1	Drawing from a box	201
	Roulette	202
	Sum of the draws	205
7.2	Expected values	209
7.3	Standard errors	213
	The square root law	214
	SD in the box	215
	Rule of Thumb	220
8	Probability Histograms	223
8.1	Probability histograms of the sums of the draws	227
	When can you add up chances?	228
	Moral	236
8.2	Using the normal curve	237
	Summary	240
III	What Can You Say about the Population?	241
9	Survey Sampling	243
9.1	Literary Digest poll	243
	The Gallup poll	245

Quota sampling	245
9.2 Probability sampling	248
Simple random samples	248
Multistage cluster samples	249
9.3 Summary	251
10 Estimating Population Percentages and Averages	253
10.1 Percentages	255
Errors	256
The law of averages for percentages	257
Warning	257
The standard error of the percentage	258
10.2 The law of average for any box	262
The standard error of the average of the draws	265
Summary of standard errors	267
Standard errors when drawing without replacement	268
The correction factor: What you need to know	272
10.3 Confidence intervals for percentages	273
Confidence interval for the Gallup poll	275
What is a confidence interval?	277
10.4 Confidence intervals for averages	280
Summary	284
11 Hypothesis Testing	285
11.1 The null hypothesis	286
11.2 The null box	290
11.3 The logic of hypothesis testing	291
The logical steps in the caffeine example	293
11.4 The p-value	294
11.5 Summary	300
Illinois State Lottery	301
Interpretation of “Fail to reject”	305
11.6 Student’s t	307
When to use	308
Example	309
Degrees of freedom	311
Student’s t table	312

Summary	315
11.7 Two samples	316
A box model for two samples	317
The SE for two independent samples	318
Comparing two percentages	324
Randomized experiments	328
11.8 Independence	331
Box model	333
Expected values	335
The chi-square statistic	341
The chi-square table	343
A larger table	344
A Appendix	353
A.1 Practice exam 1	353
A.2 Practice exam 2	373
A.3 Normal table	389
A.4 Student's t table	390
A.5 Chi-square table	391

Part I

Data in the World

1.1 Designed experiments

In a designed experiment, the experimenters decide on various treatments, e.g., in agriculture, they must decide which crops to plant where, or which fertilizers to use on which plots.

Polio

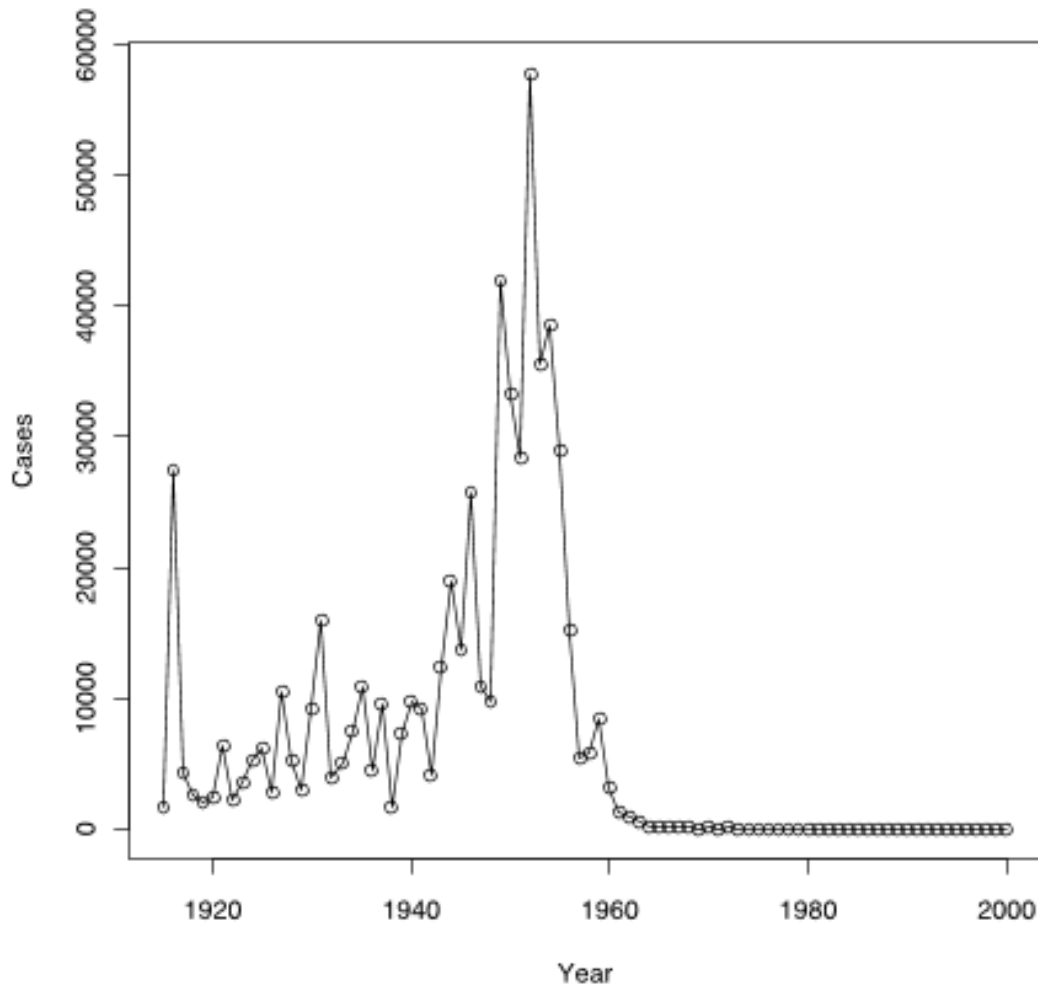
One of the more famous designed experiments tested the effectiveness of a polio vaccine. Polio was a scary disease that attacked mostly children.



- polio can cause paralysis
- It is caused by virus
- There were over 30,000 cases per year in the years 1949–1955

The little boy in the picture is a victim of polio, needing to live in an iron lung in order to breathe.

The graph below shows the number of cases in the US by year. (The horizontal axis has the years, and the vertical axis has the numbers of cases.)



Notice

- The number of cases varied substantially from year to year.
- Up until the early 1950's, the number of cases averaged about 5,000 to 10,000 per year.
- In the early 1950's, the number of cases increased dramatically, peaking at about 57,000.

President Roosevelt

President Franklin Roosevelt was diagnosed with polio in 1921. His legs were paralyzed for the rest of his life. As president, he helped found the March of Dimes Foundation, originally called the **National Foundation for Infantile Paralysis**.

(Now, there are some questions about whether it was really Guillain-Barre syndrome. See Wikipedia.)



Basic facts

- Attacks children
- Prone to epidemics
- Higher incidence among more affluent families: Children in less affluent areas catch polio when they are young, still protected by mother's immunity. Then they are immune.
- Not easily diagnosed always.

Dr. Jonas Salk developed killed-virus vaccines. The idea is to inject a person with the dead virus, so that the body will develop anti-bodies to the virus. Then the hope is that if the person later comes in contact with a live virus, the body will be able to fight it off.

Experiment to see if Salk vaccine works

The question was whether the Salk vaccine will actually prevent polio. The March of Dimes Foundation wanted to conduct an experiment, giving some children the vaccine, and comparing them to children who did not get the vaccine.

Controls: The people that do not get the vaccine are called *controls*.

How do you decide who gets the vaccine and who are the controls? Some possibilities:

- Year by year? One year disseminate the vaccine as widely as possible, and compare the polio rate to the previous year (controls). (Note: Without the vaccine, from 1948 to 1949, the number of cases went from 10000 to 40000; from 1952 to 1953, the number went from 58000 to 36000)
- By area? Chicago gets the vaccine, New York doesn't?

? What problems would either of these plans have. (That is, could there be a reason two years, or two cities, could have different rates besides whether they got the vaccine or not?)

1954: March of Dimes Foundation plan

The March of Dimes instituted an experiment. In (many) selected school districts, they decided to give the vaccine to second graders, and have the first and third graders as controls (no vaccine). For ethical reasons, they couldn't (nor shouldn't) force anyone to take the experimental vaccine, so only children whose parents volunteered them received the vaccine:

- 2nd graders: Vaccine (Only volunteers)
- 1st and 3rd graders: Control (nothing)

? What problems would either of these plans have? (That is, could there be a reason the second grade volunteered students could have different rates than the first and third graders, besides whether they got the vaccine or not?)

Here are the results from the March of Dimes study:

	# Studied	# cases	Rate per 100,000
Vaccine: 2nd graders (Volunteers)	221,998	56	25
Control: 1st and 3rd (Everyone)	725,173	391	54
Others: 2nd graders (non-Volunteers)	123,605	54	44

So of the 221,998 second grade volunteers studied, there were 56 cases of polio. The “Rate per 100,000” is calculated by

$$\text{Rate per 100,000} = \frac{\# \text{ of cases}}{\# \text{ studied}} \times 100,000.$$

So for the vaccinated:

$$\text{Rate per 100,000} = \frac{56}{221,998} \times 100,000 = 25.225.$$

We round that to 25.

The vaccine looks like it helped: The rate for the vaccinated is about half for the controls.

? Why are the non-vaccinated 2nd graders better off than non-vaccinated 1st and 3rd graders?

The volunteer effect

Why are the non-vaccinated 2nd graders better off than non-vaccinated 1st and 3rd graders? It is due to the difference between volunteers and non-volunteers:

Volunteers different than non-volunteers → more likely to be
to be affluent
→ more likely
to get polio

The 1st + 3rd graders had everyone; the 2nd graders nonvolunteers were less likely to get polio.

Other problems:

1. Diagnosis: Doctors diagnosis may be affected by knowing the treatment.
2. Behavior: Knowing whether you got the vaccine may effect your behavior.

? In what ways could the doctors' diagnoses be affected by knowing who got the vaccine and who didn't? In what ways could the kids' behaviors be affected by knowing whether they (or their parents) got the vaccine or not.

Gold star designs

Some school districts were wary of the March of Dimes design.

They instead used a gold star design:¹



Randomized control, double blind experiment with **placebo**.

Scheme involved only volunteers, from all three grades.

- Half were randomly assigned to Vaccine, half to Control. (So the comparison is fair: **Everyone is a volunteer.**)
- Done randomly to avoid bias. (Flip a coin to decide who gets the vaccine, who doesn't.)
- Double blind means that
 - Neither the children and their families,
 - Nor the doctors doing the diagnoses,knew who got what.

But wouldn't people notice getting injected?

Placebo

The controls had a **placebo**: An injection of something inert, in this case just salt water. But no one could tell.

In general, a **placebo** is a fake treatment that seems like the real treatment in every way, except for the active ingredient.

¹"It is a pity that explicit credit is not given to whomever was responsible for this change. However, only 41 percent of the trial was rescued and the remaining 59 percent blundered along its stupid and futile path." (K.A. Brownlee)

Summary of gold star designs

Randomized control, double blind experiment with **placebo**.

Main points of a gold star study:

1. Randomization: Prevents two groups from being too different.
2. Double blind: Takes care of possible bias in diagnosis and difference in behavior.
3. Placebo: Allows there to be double-blindedness.

Results of the gold star study

Here are the results of the gold star study.

	# Studied	# cases	Rate per 100,000
Vaccine: (Volunteers)	200,745	57	28
Control: (Volunteers)	201,229	142	71
Others: (Non-Volunteers)	338,778	157	46

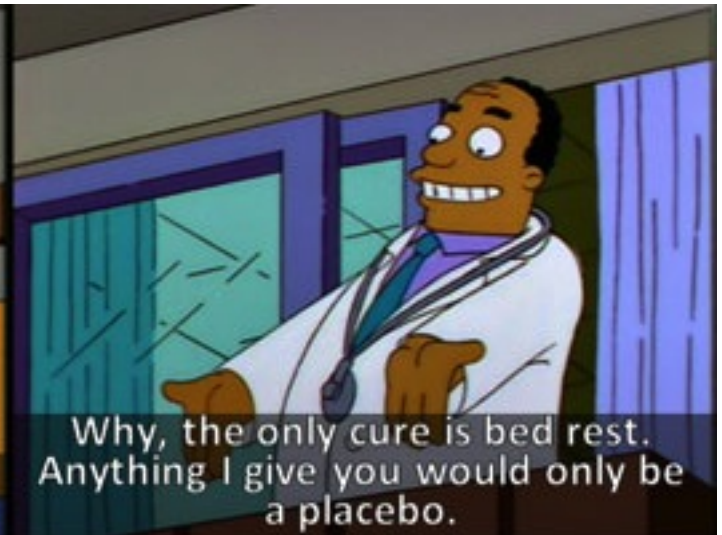
The third row is not really part of the official comparison of the vaccine group vs. the control group, because the comparison should be made only between groups made up of volunteers.

What does this tell us?

- The vaccine was even more effective than we thought. (March of Dimes study seems to have been somewhat biased against vaccine.)
- Volunteer controls did worse than non-volunteers.
- Volunteer effect: It is good they went with randomized controls.

? What were the rates of polio in the vaccine and control groups for the March of Dimes study?

? Why did the controls have a higher rate than the non-volunteers? Neither of these groups got the vaccine.



The placebo effect

When a new drug or medical procedure is introduced, people often find that it works. For example, one treatment for angina (suffocating chest pains) was to tie off the mammary artery. Two studies reported 68% and 91% of the patients improving from the surgery. This procedure subsequently became quite popular in the years 1955-1960.

Popularity plummeted after two more experiments **with controls**:

	% Improved
Treatment	67%
Control	71%

The controls received a placebo in the form of a skin incision that did not effect the artery. It appears as though the improvements people felt were based on the **placebo effect**. The actual surgery did not provide the relief. Without carefully controlled experiments comparing the treatment to the placebo, surgeons would likely have continued performing a dangerous but useless procedure.

The placebo effect occurs when people getting the fake treatment get better just because they *think* they are getting the real treatment. It can occur either when the people get an actual placebo, or when they get a real treatment that has no effect.

? In one study, without controls, the surgery had a 91% success rate. In the studies with controls, the surgery had a 67% success rate. Why is there a difference? Both groups got the surgery.

Summary of controls: Randomized, non-randomized, or nonexistent?

A study should have controls -- You want to know whether the treatment worked better than no treatment. The control group should be as much like the treatment group as possible — The only difference should be the actual treatment.

- **Randomized controls** are usually best: Start with everyone, then flip a coin to randomly assign people to control or treatment. There is no discretion on the part of the experimenter.
- **Non-randomized controls:** Either the person designing the experiment, or the people in the experiment, or something else non-objective, decides who gets the treatment & who the control. There's always the possibility of bias, maybe unintentional, creeping in.
 - **Historical controls** (a particular type of non-randomized controls): The treatment is given to people who are currently around. The controls are people from the past. (Maybe not the far past.) Especially in hospital settings, where they have the records of previous people who got the “old” treatment.
- **No controls:** Nothing to compare the treatment to.

Liver disease

The text has a study of studies on the portacaval shunt, which is a surgical treatment for cirrhosis of the liver hemorrhaging. The results of the studies:

Conclusion →	It worked!	Maybe	No good	Total
No Controls (Bad)	24 (75%)	7 (22%)	1 (3%)	32
Controls, not randomized (Ok, maybe)	10 (67%)	3 (20%)	2 (13%)	15
Randomized controls (Good)	0 (0%)	1 (25%)	3 (75%)	4
Total	34	11	6	51

? In which type of studies (no controls, non-randomized controls, or randomized controls) did the surgery look best? In which type of studies did the surgery look worst?

? Based on these results, do you think the surgery is an effective treatment?

? What are possible problems with giving a treatment that may not be effective?

Why did the studies with no or non-randomized controls have better results for the surgery? Surgery is most likely to be conducted on relatively healthy patients. Controls in non-randomized studies are likely to be less healthy than the treated people.

In randomized controls, generally everyone is fairly healthy, and half get the knife, half don't. It's fair.

For all those studies, what are the three-year survival rates?

	% Survived
Treatment	60%
Control in Randomized control study	60%
Control in non-randomized control study	45%


? Why did the controls from the non-randomized study have a lower survival rate than the controls for the randomized control study? (None of them had the surgery.)

Another example: Coronary bypass surgery.

- Of 8 randomized control studies, 1 had positive results (that is, the researchers concluded that the treatment worked), 7 negative (the treatment didn't work). (17% were positive)
- Of 21 Historical (non-randomized) control studies, 16 had positive results and 5 had negative results (76% were positive)

? Based on these results, should people get coronary bypass surgery? Why or why not?

Moral

- No controls — Bad
- Non-randomized controls — Better, but likely to have bias
- Randomized controls — Good
 - Even better if it is double blind. 
 - * With placebos

The key idea: Treatment & control groups should be as similar as possible,

except for who gets the treatment.

1.2 Observational studies

An observational study is different from a designed experiment in that the researcher has no control over who gets the treatment and who doesn't. People assign themselves to groups, e.g., whether they smoke; or they are assigned by fate, e.g., whether they are male or female.

Pellagra

Wikipedia says that

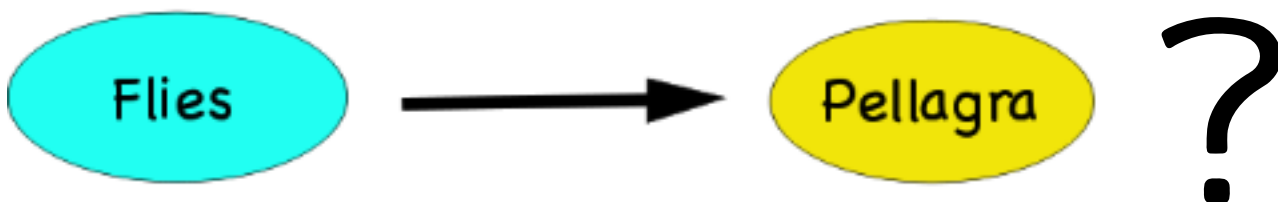
Pellagra is a vitamin deficiency disease most commonly caused by a chronic lack of niacin (vitamin B3) in the diet.

...
Pellagra is classically described by "the four D's": diarrhea, dermatitis, dementia and death.



In nineteenth century Europe, it was observed that pellagra was most prevalent in villages and households that had more flies. Consider this conclusion:

Flies cause pellagra.



? Is that conclusion justified? Would eliminating flies prevent pellagra?

This study is an **observational** study.

In a designed experiment, the researcher decides on the mechanism for who gets the treatment and who doesn't. In the polio example, the researchers decided who got the vaccine (whether by grade, or randomly).

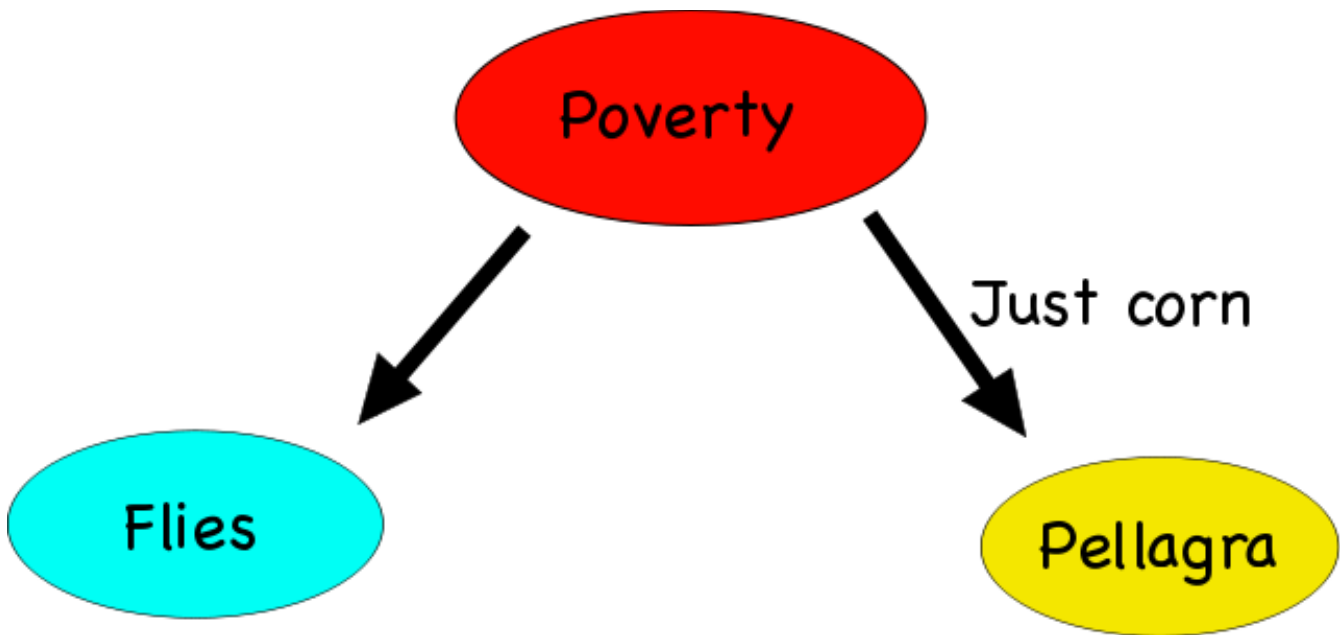
In the pellagra study, did the researchers decide which households get the flies? Did they go around randomly putting flies in some houses and not others? Did they have placebo flies? No, they just observed which households had flies.

- **Designed experiments.** The researcher decides how people are assigned the treatment group or control group.
- **Observational studies.** The researchers do not decide how people are assigned to the treatment group or control group. The people themselves may choose the group they are in, or it may be fate, or it may be some other reason.

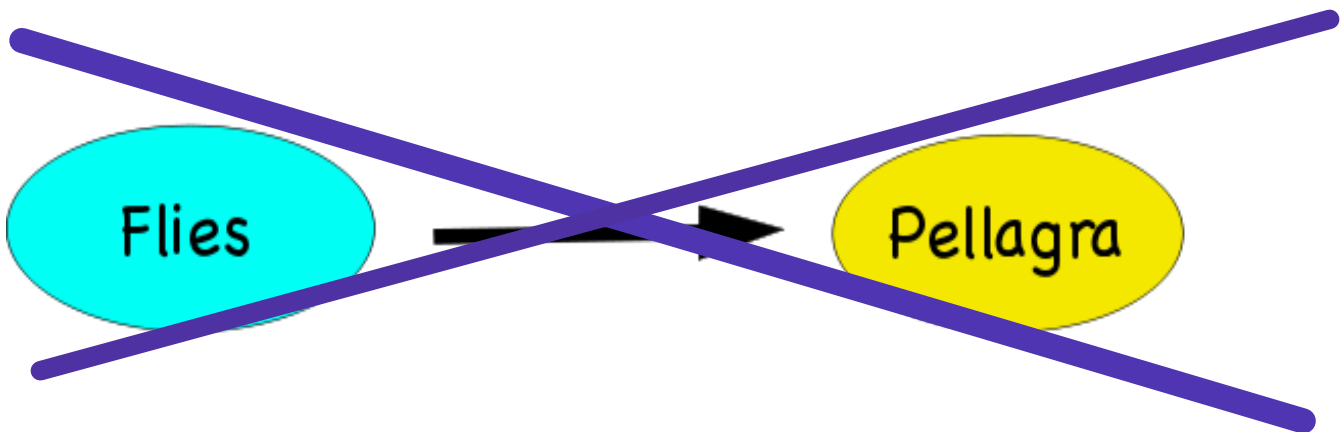
? Could there be a third factor that explains the association between flies and pellagra?

Association does not imply causation

Actually, lack of niacin causes pellagra. Eating just corn causes a lack of niacin. Poverty causes eating just corn. Poverty causes flies. There was a third factor: Poverty. Poverty caused both.



Pellagra and flies were **associated**. But flies **did not cause** pellagra: both were caused by poverty. Thus wiping out the flies would do nothing for pellagra. To eliminate pellagra, the people needed food with niacin.



Poverty was a **third factor** that caused both.

Berkeley Graduate School Admissions — Sex bias?

1973. 8,442 men and 4,321 women applied for admission to the graduate school at Berkeley; 44% of the men got in, and 35% of the women got in.

	# Applied	# Admitted	% Admitted
Men	8442	3738	44%
Women	4321	1494	35%

Does this table show evidence of bias?



? Does being male *cause* one to be more likely to be admitted?

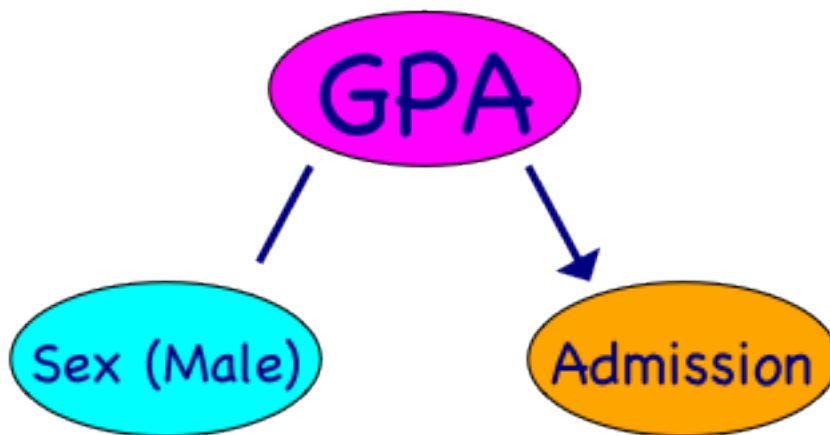
? Is this

- a designed experiment, or
- an observational study?

A third factor?

Could there be a third factor explaining the association?

- College GPA? Do men have better GPA's?
- GRE's?
- Undergrad college?



? Any other possible third factors?

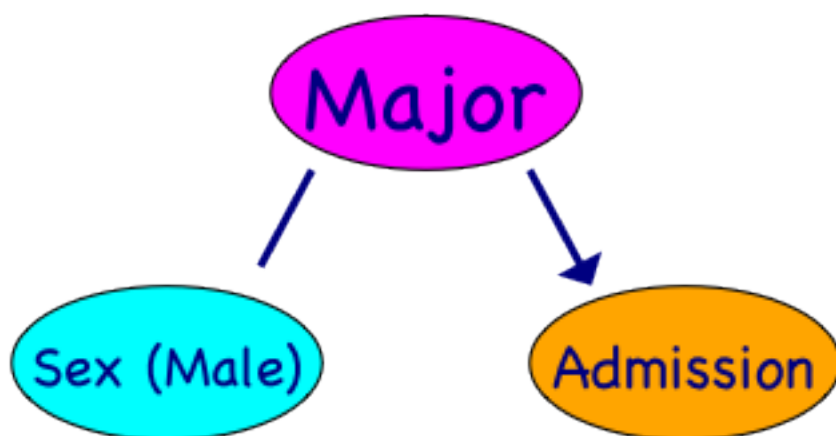
Third factor: Major

Break it down by major (just showing six of the 101 majors):

	Men		Women	
Major	#	% Admitted	#	% Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
All	8442	44%	4321	35%

Women did better in major A. In the others, it was fairly close. So it actually looks like women generally have a better admission rate than men. Men tended to apply to the majors with higher admission rates.

So again we see that association does not imply causation. The association between gender and admission could be explained by major:



(There may be other factors as well.)

Simpson's paradox

The Berkeley admissions data provides an example of **Simpson's Paradox**:

Overall, men do better than women, but major by major, women do better (or about the same).

Simpson's paradox occurs when

- Overall, group A does better than group B, but
- When broken down by a third factor, group B does better than group A within each level of the factor.

For the Berkeley data,

Group A = Men,
Group B = Women,
Third factor = Major,
Levels of third factor = The different majors.

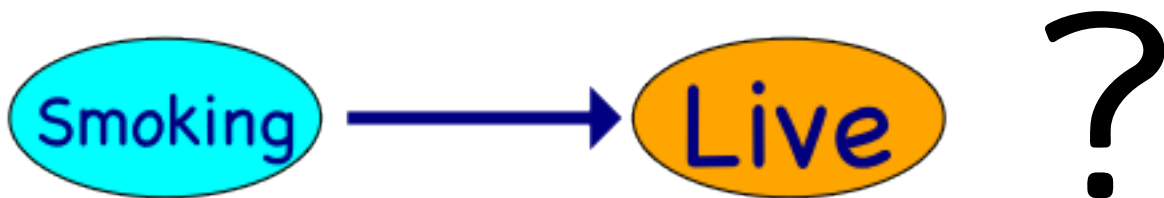
(D'oh!)

Smoking and death

A study was conducted on people near Newcastle on Tyne (in England) in 1972-74, and followed up twenty years later. Focus on the 1314 women that were in the study. We want to compare the death rates (as of 1994) depending on whether they were smokers in 1974.

Overall: $139/582 = 23.9\%$ of the smokers had died, while $230/732 = 31.4\%$ of the nonsmokers had died.

	# in 1974	# Died by 1994	% Died by 1994
Smokers	582	139	23.9%
Non-smokers	732	230	31.4%



Is this conclusion valid?:

Smoking helps you live longer.

? Is this a designed experiment or observational study? Could there be a third factor explaining why smokers were more likely to live?

Third factor: Age

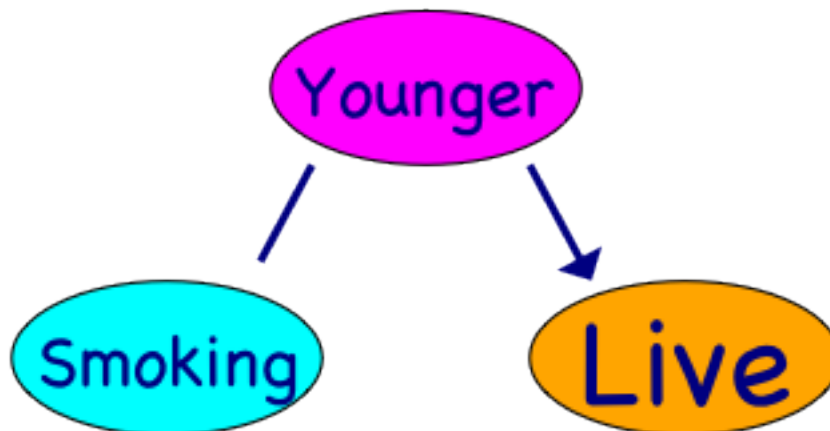
Consider the ages of the women in 1974, and see whether smokers were more likely to live in each age group:

Age in 1974	Smokers		Non-smokers	
	#	% Died by 1994	#	% Died by 1994
Young (18-34)	179	2.8%	219	2.7%
Middle (34-64)	354	26.0%	320	18.4%
Old (65+)	49	85.7%	193	85.5%
All	582	23.9%	732	31.4%

For each group, non-smokers were more likely to live. (Mostly it's close, except for the middle-aged women.)

45% of young people smoked, 52.5% of middle aged people smoked, and 20% of older people smoked.

Old people didn't smoke much in 1974, but since they were old, more likely to die:



? Is this another example of Simpson's paradox?

Does TV harm children's psyche?

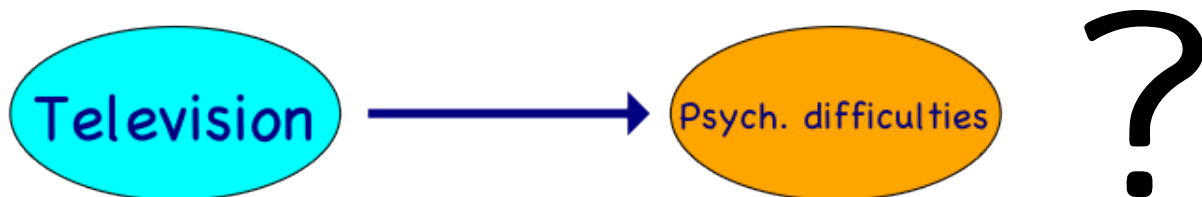
A headline Discovery News:

TOO MUCH TV HARMS KIDS PSYCHOLOGICALLY

Here's a quote from the article²:

"The researchers got 1,013 children between the ages of 10 and 11 to self-report average daily hours spent watching television or playing — not doing homework — on a computer. Responses ranged from zero to around five hours per day. The children also completed a 25-point questionnaire to assess their psychological state...

"The researchers found that children who spent two hours or more a day watching television or playing on a computer were more likely to get high scores on the questionnaire, indicating they had more psychological difficulties than kids who did not spend a lot of time in front of a screen."



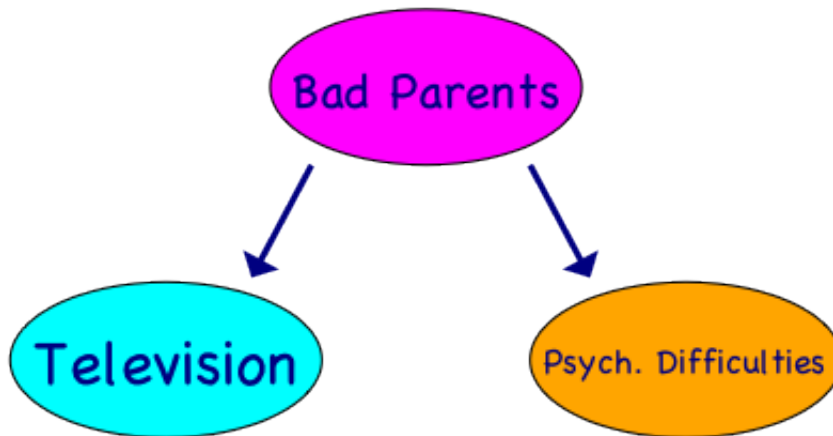
? Which is it? (Did the researcher assign how much TV the kids were to watch?)

- Designed experiment
- Observational study

²The researchers did not claim causation, by the way.

Third factor? Reverse causation?

Could there be a third factor, such as bad parenting?



Or could the causation be reversed?




? Can you think of other possible third factors that explain the association? Is it plausible that psychological difficulties could cause more TV watching?

Moral

- In observational studies, you cannot prove causation just by proving an association. There may be a third factor that explains the association. Or the causation may be reversed!
- In a gold star designed experiment, those third factors aren't likely to creep in. You randomly assign people to the treatment or control.

Smoking and lung cancer

Back in the 40's and 50's, tobacco companies advertised the health benefits of smoking.



"I've been a two-pack-a-day man for fifteen years and I've found much milder Chesterfield is best for me."
Perry Como

NOW...10 Months Scientific Evidence For Chesterfield

A MEDICAL SPECIALIST is making regular bi-monthly examinations of a group of people from various walks of life. 45 percent of this group have smoked Chesterfield for an average of over ten years.

After ten months, the medical specialist reports that he observed...

no adverse effects on the nose, throat and sinuses of the group from smoking Chesterfield.

MUCH Milder
CHESTERFIELD
IS BEST FOR YOU

First and Only Premium Quality Cigarette in Both Regular and King-Size

CHESTERFIELD CIGARETTES

CONTAINS TOBACCOS OF BETTER QUALITY AND HIGHER PRICE THAN ANY OTHER KING-SIZE CIGARETTE

Copyright 1955, LOUW & MILES Tobacco Co.

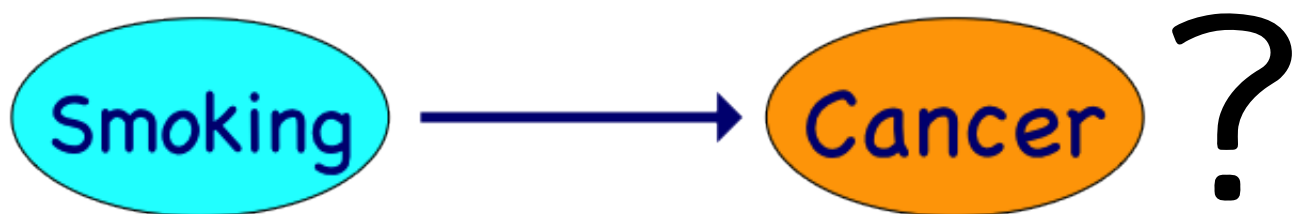
But it was clear by the 1950's that smoking and lung cancer were associated.



A study by Doll and Hill (observational, with historical controls from hospital records), used the following data from 1948-49:

	# Studied	# Smokers	% Smokers
Lung cancer	709	688	97.0%
No lung cancer	709	640	90.3%

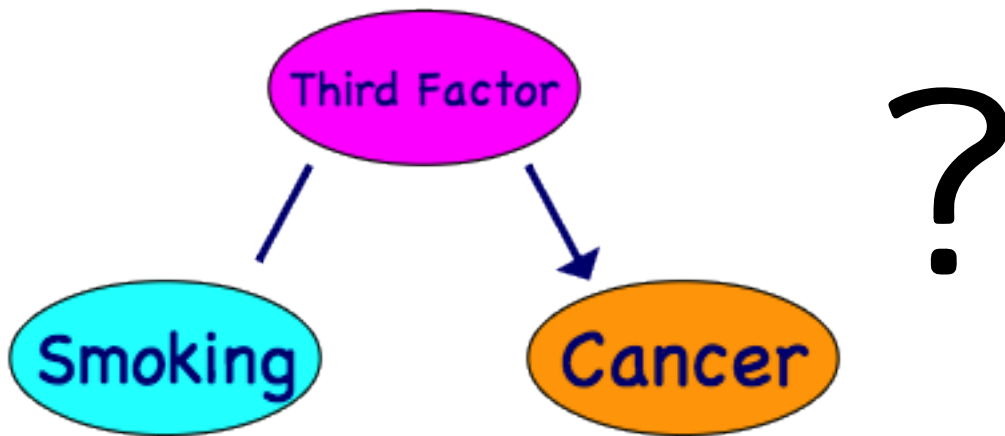
Notice that lots of people smoked. But the group with lung cancer had a higher percentage of smokers than the group without lung cancer.



? Does this table prove that smoking causes lung cancer?

Third factors?

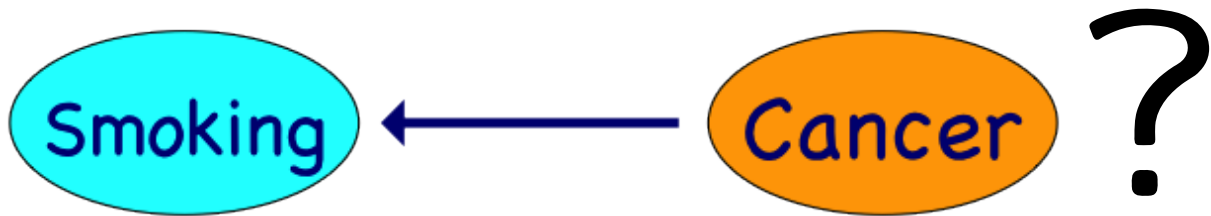
Sir Ronald Fisher (1890-1962), perhaps the most famous and influential statistician of all time, objected. Could there be a third factor explaining the association? Gender? (Age? Genetics? Type of employment? Socio-economic status?)



? Can you think of other possibilities?

Reverse causation?

Fisher even suggested the causation might go the other way:

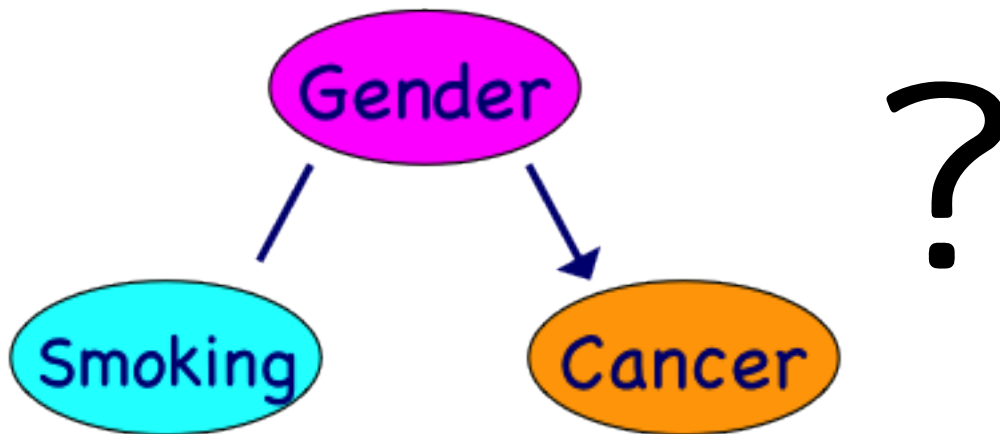


For example, people with the early stages of a long-developing disease like lung cancer may feel slightly uncomfortable, and smoking can help alleviate the discomfort. As Fisher said:

And to take the poor chap's cigarettes away from him would be rather like taking away his stick from a blind man. It would make an already unhappy person a little more unhappy than he need be.

? Does this seem like a reasonable possibility?

A possible third factor is gender. It was noticed that men smoked more than women, and men had more lung cancer than women, so perhaps gender explains the association between smoking and lung cancer:



Here's the data, split into the men's and women's data:

	Women			Men		
	# Stud-ied	# Smokers	% Smokers	# Stud-ied	# Smokers	% Smokers
Lung cancer	60	41	68.3%	649	647	99.7%
No lung cancer	60	28	46.7%	649	622	95.8%

? Among the women, are the people with lung cancer more likely to be smokers?

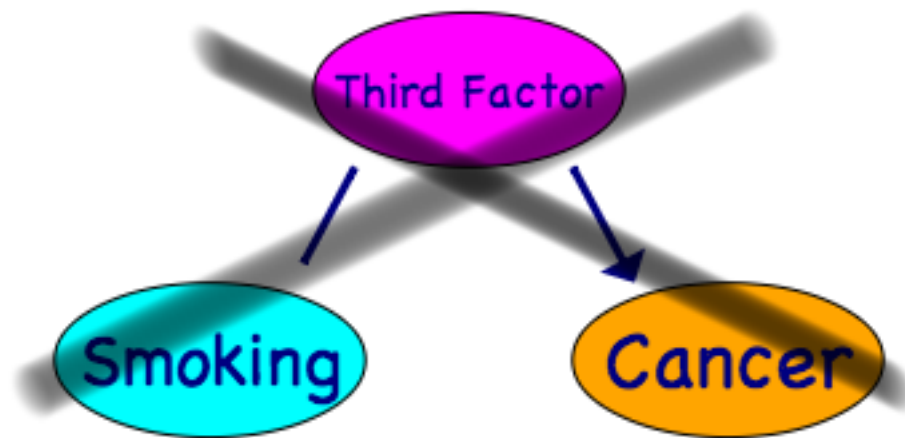
Among the men, are the people with lung cancer more likely to be smokers?

Is there an association between smoking and lung cancer within the men? Within the women?

Does gender explain the overall association between lung cancer and smoking?

Third factors?

The previous page suggests that gender *does not* explain the association between lung cancer and smoking, because the association still remains within the men, and within the women. Other studies looked at other potential third factors. In no case do they find the third factor explains the association:



In 1957, the British Medical Journal editorialized that smoking does indeed cause lung cancer, citing "... the painstaking investigations of statisticians that seem to have closed every loophole of escape for tobacco as the villain in the piece."



Designed experiment?

It is always better to have a gold star study, rather than an observational one. How could you implement a designed experiment for seeing whether smoking causes lung cancer?

- Randomized controls? Flip a coin to decide who's going to smoke for the rest of their lives, who's not going to smoke ever.
- Placebo? Placebo cigarettes that look just like regular cigarettes, and have the same effect on people. (If they did not, then it would not be long before the subjects would figure out which group they are in.) But these placebo cigarettes cannot have the cancer-causing effects of regular cigarettes.
- These people would have to be followed, and supplied with the correct type of cigarettes, until they die, or say, for 40 years. Then the researchers can analyze who died of lung cancer and who did not.

? This would be a convincing experiment, but would it be possible to carry out?

Collective force

Though designed experiments are best, in many important situations, they are impossible to carry out.

But the collective force of many observational studies can be convincing. Rule out all third factors, one-by-one. (Gender? Age? Genetics? Type of employment? Socio-economic status?) The Surgeon General surveyed a huge number of such studies, and in his 1964 report said,

“Cigarette smoking is causally related to lung cancer in men; the magnitude of the effect of smoking far outweighs all other factors.”



Not everyone bought into this conclusion, especially the tobacco companies. In 1979, the then Surgeon General produced another (heftier) report, even more comprehensive and even more damning for smoking.

Tobacco companies still didn't buy it.



John Duricka/Associated Press

“Tobacco executives told Congress in 1994 that they did not believe there was a proven link between smoking and cancer. Later, the Justice Department sought to prove that they had lied about the dangers of smoking.”(NY Times, 3/27/2009)

Finally, in January of 1998, even the tobacco companies had to admit, “We recognize that there is a substantial body of evidence which supports the judgment that cigarette smoking plays a causal role in the development of lung cancer and other diseases in smokers.” (Geoffrey C. Bible, Chairman and Chief Executive Officer of Philip Morris. Cos., Inc.)

As Fletcher Knebel said, “It is now proved beyond doubt that smoking is one of leading causes of statistics.”

Moral

Observational studies are valuable, but it takes many more of them, and a wide variety of types, to collect evidence equal to good randomized control studies. To compare:

- The gold star polio vaccine designed experiment was conducted in 1954. People were convinced of the vaccine's efficacy immediately, and polio was virtually wiped out in the US in ten years.
- Observational studies on smoking & cancer started in the 40's, and it wasn't until 50 years later that finally, the tobacco companies admitted that smoking causes cancer.

2.1 Histograms

How many pairs of shoes do you own? Imelda Marcos reportedly owned 2700 pairs. So many, they put them in a shoe museum.

Here are the answers for STAT100 students in the spring:



30 40 9 4 20 6 12 2 10 13 5 10 25 6 5
 5 40 29 22 16 13 25 7 12 2 10 30 15
 25 4 21 15 40 15 16 10 7 10 6 5 16 4
 7 10 3 8 100 4 12 12 12 25 20 4 4 3
 40 8 19 14 13 60 9 15 2 10 5 20 30 15 12 53 4 9 15 20 5 5 6 40 10
 8 5 20 20 60 11 3 20 16 4 30 20 25 22 5 6 20 35 4 10 31 13 18 6 10
 40 10 30 10 4 6 7 12 20 15 10 5 34 11 10 4 20 14 20 12 15 20 8 13
 30 30 25 15 5 15 18 25 5 10 13 20 40 9 25 13 15 5 20 22 6 20 8 40
 35 30 20 20 10 6 9 16 5 12 17 5 20 25 6 13 10 20 30 7 31 12 10 20
 6 15 18 3 10 4 38 50 15 10 26 20 15 4 9 25 11 20 25 20 14 15 11 12
 5 30 5 23 10 25 4 25 3 5 3 5 4 8 4 24 10 6 12 9 25 6 7 12 26 4 5 15
 30 10 8 35 20 7 10 20 34 20 30 25 16 8 8 6 10 23 4 9 50 5 31 8 2 3
 5 35 10 2 15 30 25 8 6 30 9 8 7 20 43 8 15 2 6 15 30 7 12 12 7 2 10
 51 8 22 10 18 20 30 25 6 14 3 90 3 12 15 5 17 30 20 10 10 50 14 8
 11 3 10 3 40 15 25 10 14 6 10 30 30 10 57 8 40 30 3 11 5 50 10 15
 10 20 10 26 15 10 9 15 4 14 2 15 3 10 20 19 10 10 11 3 7 6 7 30 23
 25 20 9 6 5 8 5 20 4 15 70 10 10 4 12 23 17 24 5 15 13 14 25 9 6 50
 30 5 8 12 7 13 20 15 26 2 12 6 10 8 5 10 10 3 20 20 8 5 3 2 4 10 20
 20 5 15 100 10 2 10 10 20 12 8 3 12 10 25 24 30 5 23 15 14 14 35
 15 10 8 18 15 20 4 9 6 6 4 3 20 15 12 16 15 10 10 7 12 16 3 3 2 8 10

15 9 11 9 8 5 4 14 12 10 5 4 10 20 14 9 7 10 4 20 6 9 12 14 10 9 5
18 5 4 6 15 7 16 10 3 25 20 9 20 15 40 8 7 13 20 15 4 15 14 7 22 30
21 5 7 20 5 15 13 20 9 30 15 10 13 15 3 26 20 13 4 20 10 20 40 100
3 16 10 24 25 30 25 30 21 7 3 60 20 15 40 17 31 11 12 14 7 30 3 12
6 4 15 3 6 2 10 5 8 8 30 30 6 14 17 15 4 12 40 12 30 4 34 20 12 15
3 8 10 10 12 6 7 5 10 24 12 2 6 35 15 15 3 20 20 20 20 25 6 15 15 6
7 50 5 3 50 10 5 9 30 12 18 3 10 7 14 20 8 60 25 20 12 30 30 10 9
10 4 16 7 21 15 22 15 7 4 15 12 30 7 15 12 14 14 25 5 6 19 27 30 16
20 7 2 13 5 11 15 11 10 14 2 12 7 11 15 1 10 1 18 6 12 40 3 4 20 4
6 27 30 10 8 6 5 17 5 20 27 70 13 25 8 25 23 15 10 10 14 10 17 18

There are 712 observations.

? What do you see? What is the largest number of shoes in the list? Smallest? What numbers seem to be most plentiful?

Reading a histogram

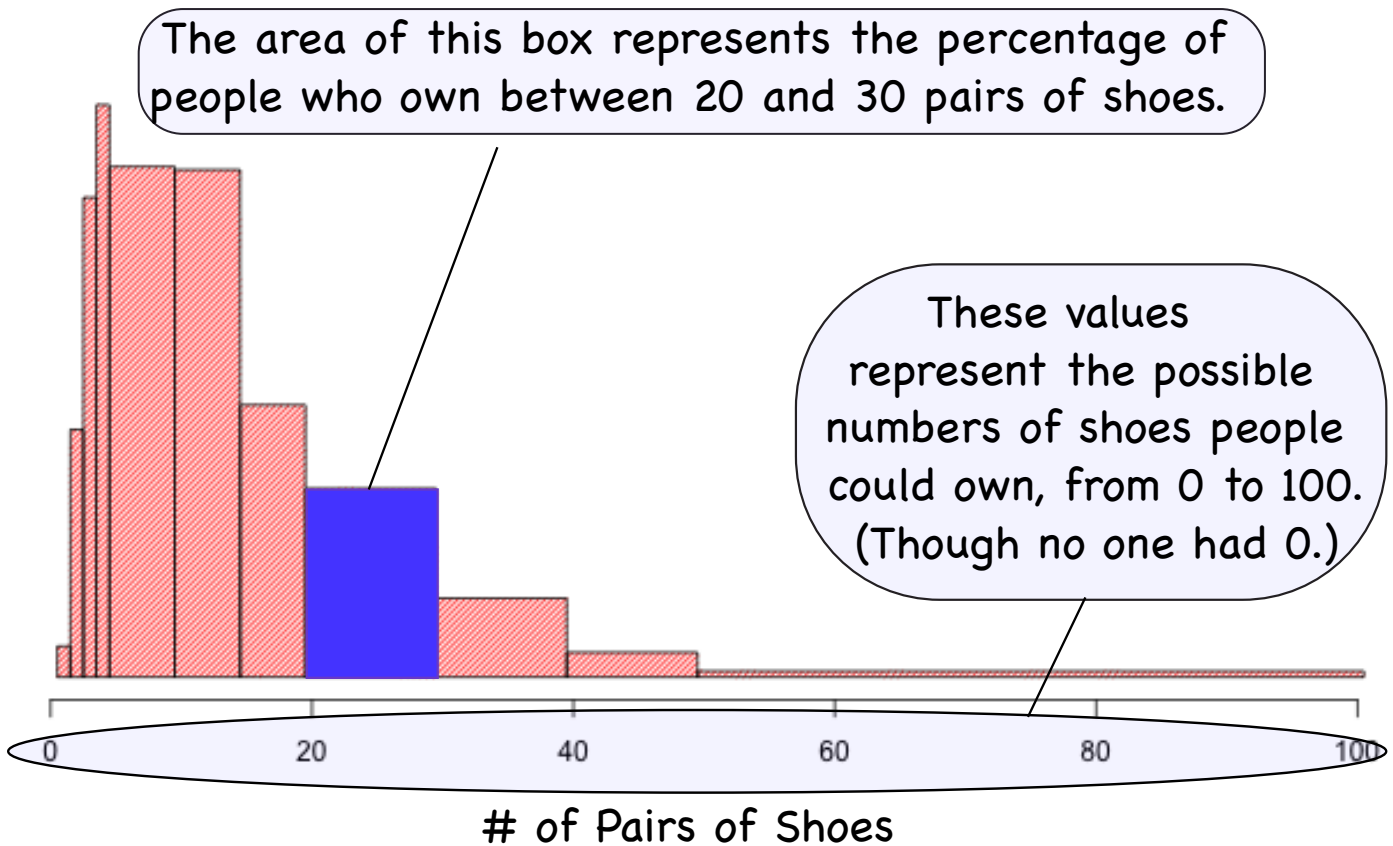
You can look through the numbers and get some idea of how many shoes people have in general, but overall, just a list of the numbers is hard to take in. A **histogram** is a type of graph that gives a good overall picture of the data. Here is one for the shoes:



How do you read such a graph?

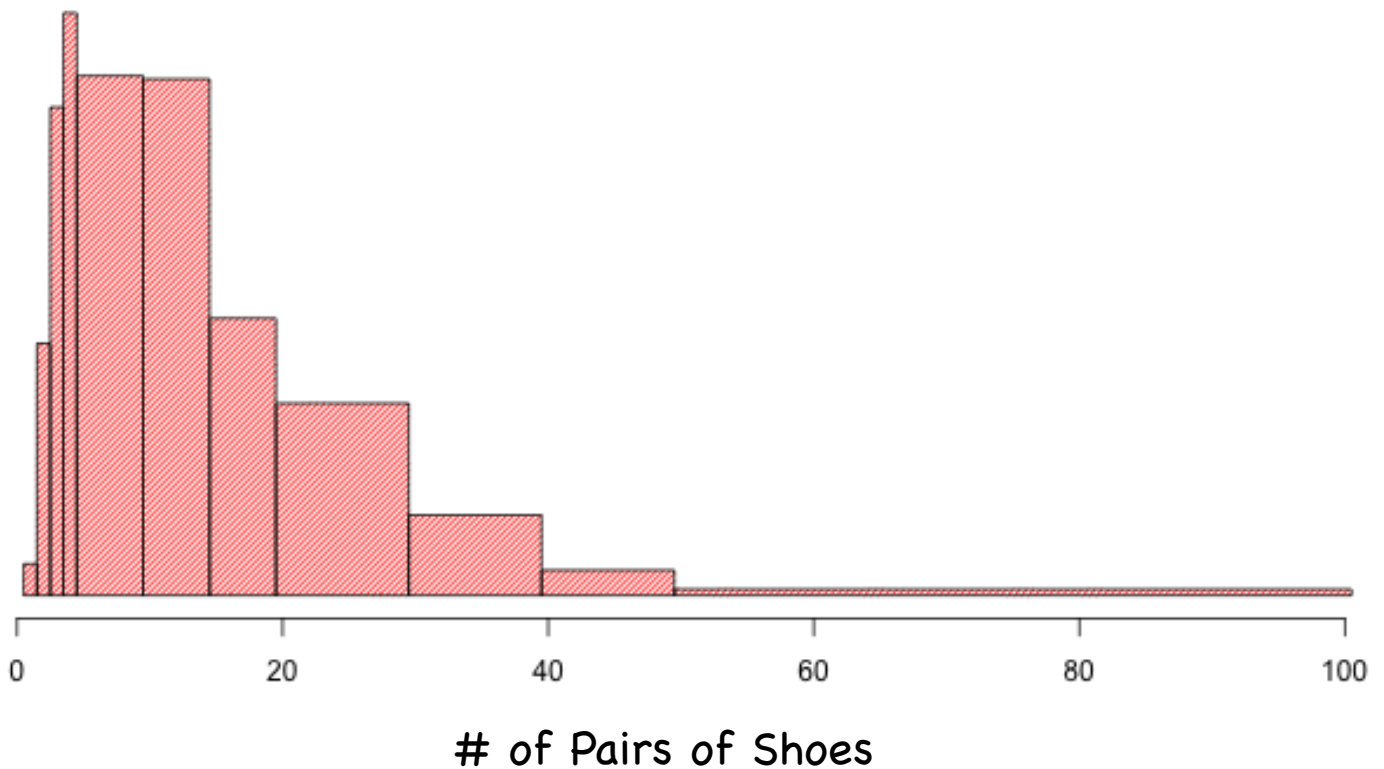
- The horizontal axis represents the numbers of pairs of shoes.
- The area of a box above an interval on the horizontal axis represents the percentage of people who have the numbers of pairs of shoes within that interval.
- The total area of the boxes is 100%.

Look at the solid box in the picture below. Its base rests on the horizontal axis, from about 20 to about 30.



- At the far left, there is a small narrow box. Its area represents the percentage of people who own just one pair of shoes. It's a very small percentage (less than 1%).
- A higher percentage has 2, 3 or 4 pairs. But there is way larger percent, maybe 95%, that has more than 4.
- More than half have between 1 and 20 pairs, but there are substantial percentages between 20 and 30, not as many, but some between 30 and 40, then maybe 5-10% have more than 40.
- About half the people have over 15 pairs, half under 15.

Here is the histogram again:



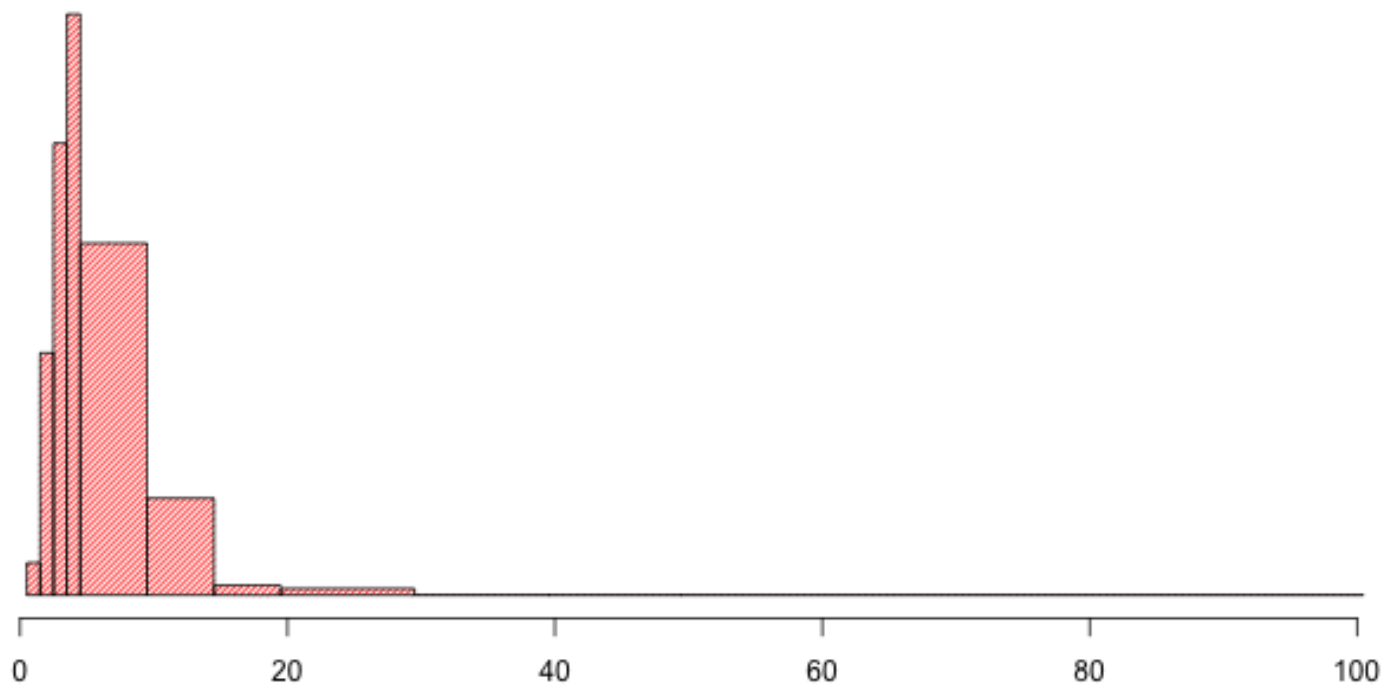
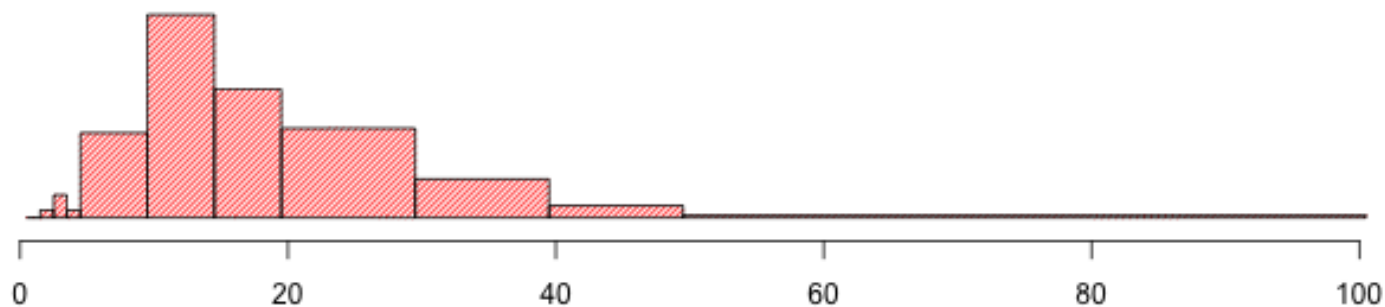
? Which boxes represent the people who have between 0 and 10 pairs of shoes? Between 10 and 20?

Which has the highest percentage: The people who have between 0 and 10 pairs, between 10 and 20 pairs, between 20 and 30 pairs, or between 30 and 40 pairs?

The percentage of people who have between 0 and 50 pairs is _____ the percentage who have between 50 and 100 pairs. [Fill in the blank with one of these choices: (a) much higher than; (b) a little higher than; (c) about the same as; (d) a little less than; (e) much less than]

Comparing two histograms

Men vs. Women



Split the data into two groups: men (227 of them) and women (485 of them). For each group, we can draw a histogram.

? For the top one, what do you guess is the percentage with fewer than 20 pairs? For the bottom one? Can you guess which is for men and which for women?

Creating histograms

The next table contains the ages at inauguration and death of US Presidents (not including the ones that are still alive).

	Inauguration	Death
Washington	57	67
J_Adams	61	90
Jefferson	57	83
Madison	57	85
Monroe	58	73
J_Q_Adams	57	80
Jackson	57	78
Van_Buren	54	79
W_H_Harrison	68	68
Tyler	51	71
Polk	49	53
Taylor	64	65
Fillmore	50	74
Pierce	48	64
Buchanan	65	77
Lincoln	52	56
A_Johnson	56	66
Grant	46	63
Hayes	54	70

	Inauguration	Death
Garfield	49	49
Arthur	50	57
Cleveland	47	71
B_Harrison	55	67
McKinley	54	58
T_Roosevelt	42	60
Taft	51	72
Wilson	56	67
Harding	55	57
Coolidge	51	60
Hoover	54	90
F_D_Roosevelt	51	63
Truman	60	88
Eisenhower	62	78
Kennedy	43	46
L_Johnson	55	64
Nixon	56	81
Ford	61	93
Reagan	69	93

They had to live long enough to become president, so they lived reasonably long as a rule.

? Which presidents died before age 50? (What did they die from?)

Which presidents lived the longest?

Grouping the values

A histogram gives an understandable picture of all the data. Here's the ages at death again, put in order:

46 49 53 56 57 57 58 60 60 63 63 64 64 65 66 67 67 67 68 70 71 71
72 73 74 77 78 78 79 80 81 83 85 88 90 90 93 93

To create a histogram, we break the data into convenient groups, called class **intervals**. We will use the intervals for ages

$$\begin{array}{llll} 44\frac{1}{2} \text{ to } 49\frac{1}{2} & 49\frac{1}{2} \text{ to } 59\frac{1}{2} & 59\frac{1}{2} \text{ to } 64\frac{1}{2} & 64\frac{1}{2} \text{ to } 69\frac{1}{2} \\ 69\frac{1}{2} \text{ to } 79\frac{1}{2} & 79\frac{1}{2} \text{ to } 89\frac{1}{2} & 89\frac{1}{2} \text{ to } 94\frac{1}{2} & \end{array}$$

Why the $\frac{1}{2}$'s? So we do not have to worry about ages on the edge of two intervals.

Now count the number of presidents in each class interval:

- Between $44\frac{1}{2}$ and $49\frac{1}{2}$: There are 2 such ages at death, 46 (Kennedy) and 49 (Garfield)
- Between $49\frac{1}{2}$ and $59\frac{1}{2}$: There are 5 (53, 56, 57, 57, 58).
- Between $59\frac{1}{2}$ and $64\frac{1}{2}$: There are 6 (60, 60, 63, 63, 64, 64).

? Count the numbers of values in the other intervals:

- Between $64\frac{1}{2}$ and $69\frac{1}{2}$:
- Between $69\frac{1}{2}$ and $79\frac{1}{2}$:
- Between $79\frac{1}{2}$ and $89\frac{1}{2}$:
- Between $89\frac{1}{2}$ and $94\frac{1}{2}$:

We have seven intervals, and there are 38 ages (presidents).

Class interval	# of presidents
$44\frac{1}{2}-49\frac{1}{2}$	2
$49\frac{1}{2}-59\frac{1}{2}$	5
$59\frac{1}{2}-64\frac{1}{2}$	6
$64\frac{1}{2}-69\frac{1}{2}$	6
$69\frac{1}{2}-79\frac{1}{2}$	10
$79\frac{1}{2}-89\frac{1}{2}$	5
$89\frac{1}{2}-94\frac{1}{2}$	4
Total:	38

Next, we find the percentage of observations in each interval:

$$\text{Percentage in interval} = 100 \times \frac{\# \text{ in interval}}{\text{Total } \#}$$

So for the first interval, the $44\frac{1}{2}-49\frac{1}{2}$ one, we have 2 presidents (and total $\# = 38$):

$$\text{Percentage in interval} = 100 \times \frac{2}{38} = 5.26\%$$

For the second interval, the $49\frac{1}{2}-59\frac{1}{2}$ one, we have 5 presidents:

$$\text{Percentage in interval} = 100 \times \frac{5}{38} = 13.16\%$$

We do the same for all the intervals.

$$\text{Percentage in interval} = 100 \times \frac{\# \text{ in interval}}{n}$$

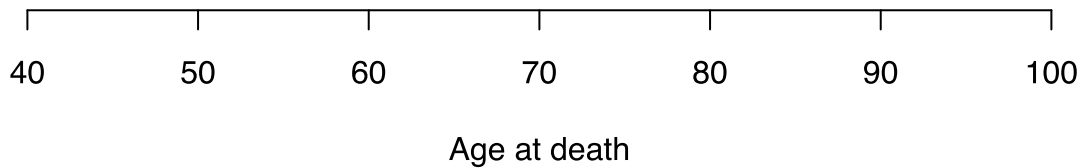
Class interval	# of presidents	Percentage
$44\frac{1}{2}-49\frac{1}{2}$	2	5.26%
$49\frac{1}{2}-59\frac{1}{2}$	5	13.16%
$59\frac{1}{2}-64\frac{1}{2}$	6	15.79%
$64\frac{1}{2}-69\frac{1}{2}$	6	15.79%
$69\frac{1}{2}-79\frac{1}{2}$	10	26.32%
$79\frac{1}{2}-89\frac{1}{2}$	5	
$89\frac{1}{2}-94\frac{1}{2}$	4	
Total:	38	100.01%

(The total percentage should be exactly 100%, but there is some round-off error in the individual percentages, which need not be worrisome.)

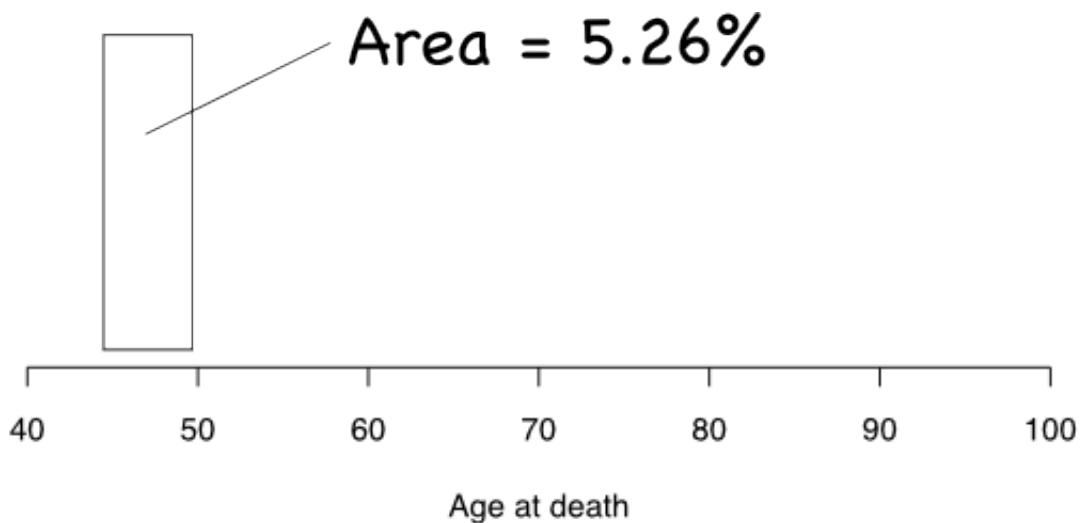
? Fill in the two missing percentages.

Drawing the histogram

First, we need to draw the horizontal axis: the line which will represent the ages. The ages run from 46 to 93, so we make this line a little longer, going from 40 to 100:



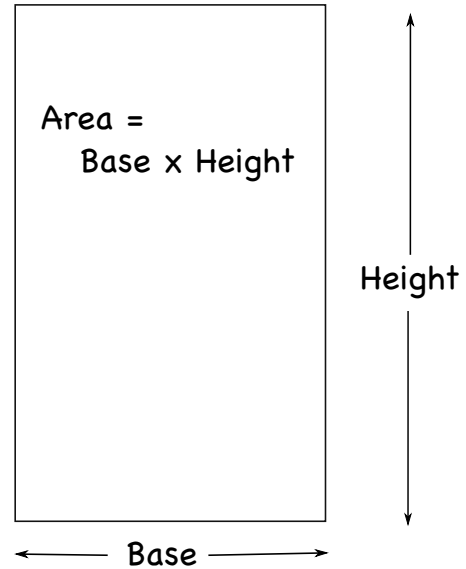
For each interval, we draw a box above the interval on top of the horizontal axis. The key is that the box's area should equal the percentage of presidents in that interval. So for the first interval, the area should be 5.26%.



The base of the box goes from 44.5 to 49.5. We also need to know the height of the box in order to draw it.

To find the height, we need to know that the area of a general rectangle is $\text{Base} \times \text{Height}$. So the Height is Base/Area :

$$\text{Area} = \text{Base} \times \text{Height} \implies \text{Height} = \frac{\text{Area}}{\text{Base}}$$

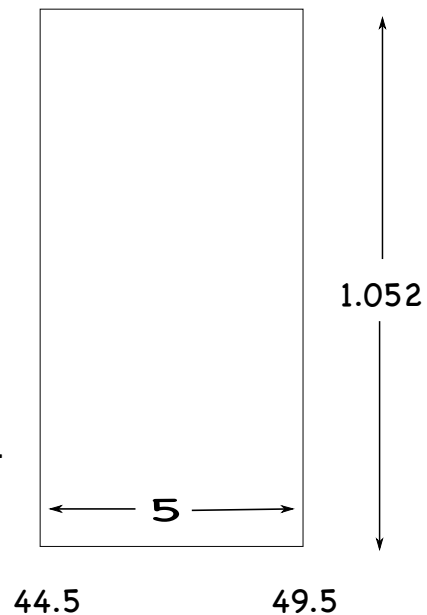


The first interval is $44\frac{1}{2}$ to $49\frac{1}{2}$, so the base is the difference between those two endpoints:

$$\text{Base} = 49\frac{1}{2} - 44\frac{1}{2} = 5.$$

The percentage = area of our first rectangle is 5.26%, so the height is

$$\text{Height} = \frac{\text{Area}}{\text{Base}} \implies \text{Height} = \frac{5.26}{5} = 1.052.$$



? The second interval is $49\frac{1}{2}$ to $59\frac{1}{2}$. Find the base.

The percentage of the second interval is 13.16%. Find the area and the height of the rectangle.

And so to all the intervals:

Class interval	# of presidents	Percentage	Base	Height
$44\frac{1}{2}-49\frac{1}{2}$	2	5.26%	5	1.052
$49\frac{1}{2}-59\frac{1}{2}$	5	13.16%	10	1.316
$59\frac{1}{2}-64\frac{1}{2}$	6	15.79%	5	3.158
$64\frac{1}{2}-69\frac{1}{2}$	6	15.79%	5	3.158
$69\frac{1}{2}-79\frac{1}{2}$	10	26.32%	10	
$79\frac{1}{2}-89\frac{1}{2}$	5	13.16%		
$89\frac{1}{2}-94\frac{1}{2}$	4	10.53%		
Total:	38	100.01%		

Again, the area of a rectangle is $\text{Base} \times \text{Height}$. So the Height is Base/Area :

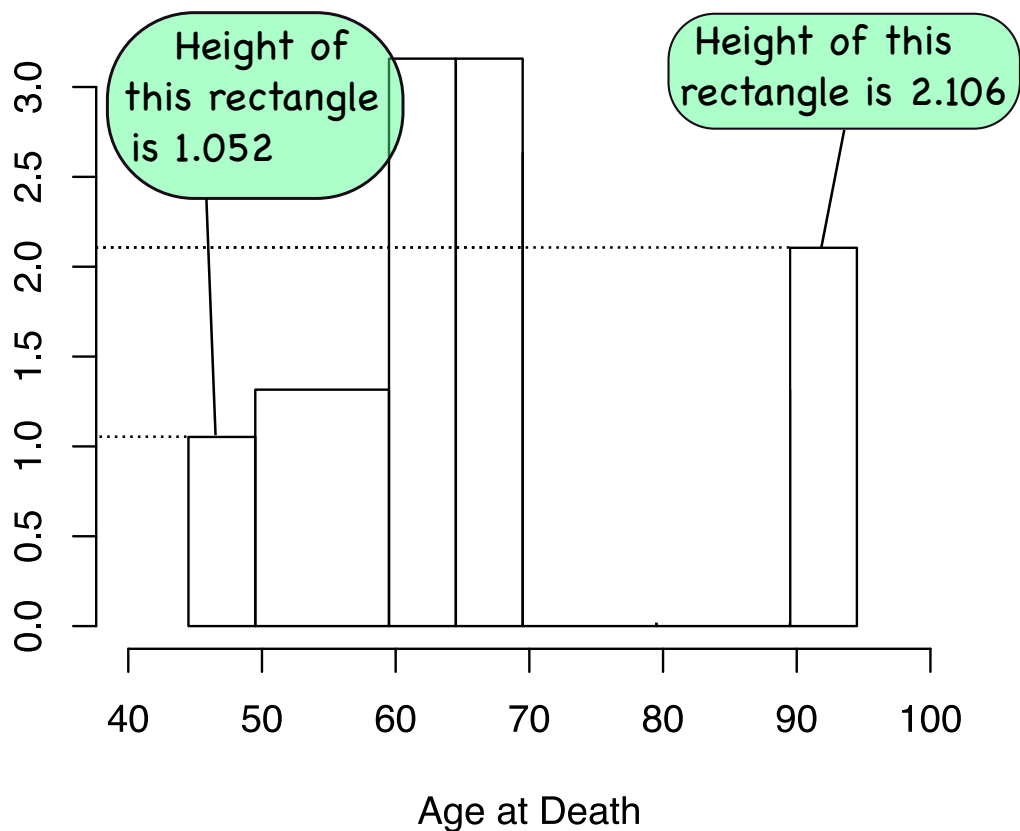
$$\text{Area} = \text{Base} \times \text{Height} \implies \text{Height} = \frac{\text{Area}}{\text{Base}}$$

? Fill in the five blank spaces in the table. (Area = what?)

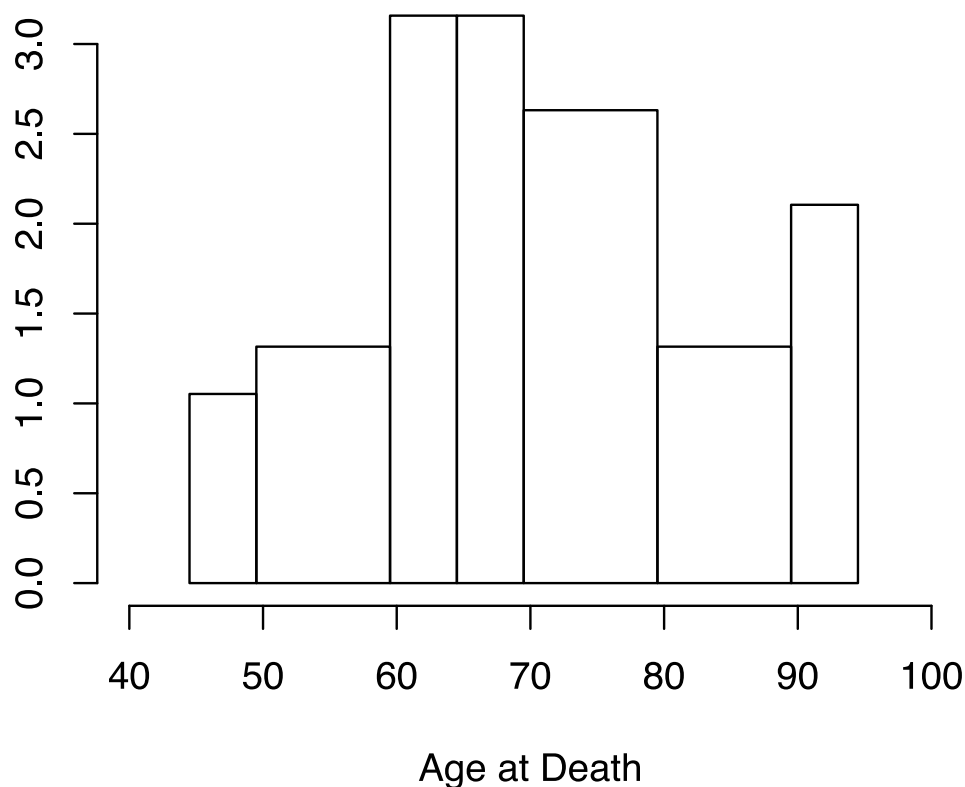
Here is the full table:

Class interval	# of presidents	Percentage	Base	Height
$44\frac{1}{2}-49\frac{1}{2}$	2	5.26%	5	1.052
$49\frac{1}{2}-59\frac{1}{2}$	5	13.16%	10	1.316
$59\frac{1}{2}-64\frac{1}{2}$	6	15.79%	5	3.158
$64\frac{1}{2}-69\frac{1}{2}$	6	15.79%	5	3.158
$69\frac{1}{2}-79\frac{1}{2}$	10	26.32%	10	2.632
$79\frac{1}{2}-89\frac{1}{2}$	5	13.16%	10	1.316
$89\frac{1}{2}-94\frac{1}{2}$	4	10.53%	5	2.106
Total:	38	100.01%		

? The graph has five of the seven boxes drawn in. You are to draw in the other two (the fifth and sixth ones).



Here is the histogram again:

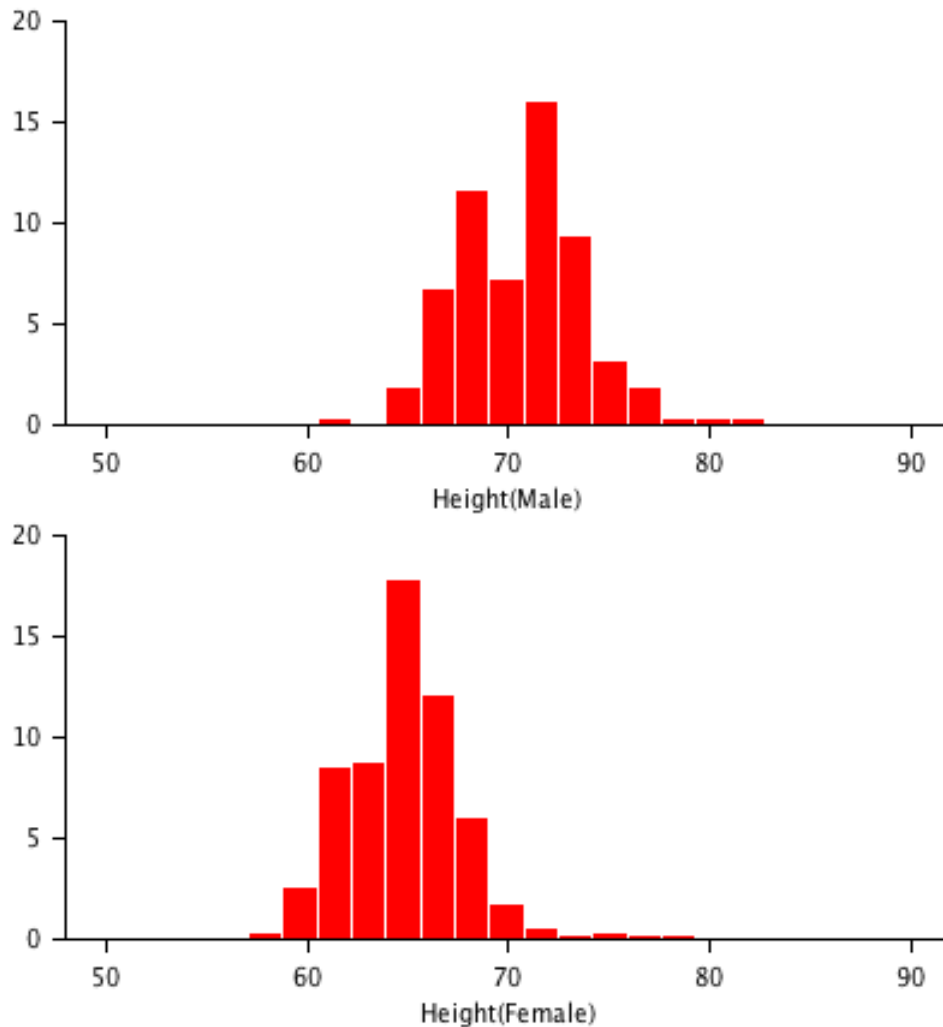


? In the fourth interval, $64\frac{1}{2}$ - $69\frac{1}{2}$, there are 6 presidents, while in the fifth interval, $69\frac{1}{2}$ - $79\frac{1}{2}$, there are ten. So the fourth interval has fewer people, but a higher bar. Why?

What would you estimate the percentage who died between 70 and 75 years of age?

Which ten-year age group (i.e., 40's, 50's, ..., 90's) has the highest percentage? Second highest?

2.2 The average, median, and standard deviation



Here are the histograms of the heights of men and women in a STAT100 class. What is a “typical” height of the men and women represented in the histograms? There is clearly a wide range of height: Men go from about 60 inches (5 feet) to over 80 inches. Women go from about 58 inches to close to 80 inches. (Must be some basketball players in there.)

What are middle-ish values? Maybe for men around 70-71 inches (just under 6 feet); For women, about 65 inches (5 foot 5). Men are generally taller, but there is a lot of overlap.

Average

The average¹ is one measure of the typical value in a batch of data.

If everyone in the class put their money in a big pot, then split it equally, each person would end up with the average amount of money. To calculate the average, add up all the numbers, and divide by how many there are.

A simple example: Data: 1, 6, 10, 10, 10, 0, 5 (There are 7 of them.)

$$\text{Average} = \frac{\text{Sum}}{\# \text{ of values}} = \frac{1 + 6 + 10 + 10 + 10 + 0 + 5}{7} = \frac{42}{7} = 6.$$

(They add up to 42.)

A couple of things to be aware of:

- There are three 10's in the data, so we add 10 in three times.
- There is a 0 in the data. It doesn't affect the sum, but we still have to count it as one of the seven values.

? Here are the heights of five of the men: 74 69 72 75 69.

What is their sum?

What is their average?

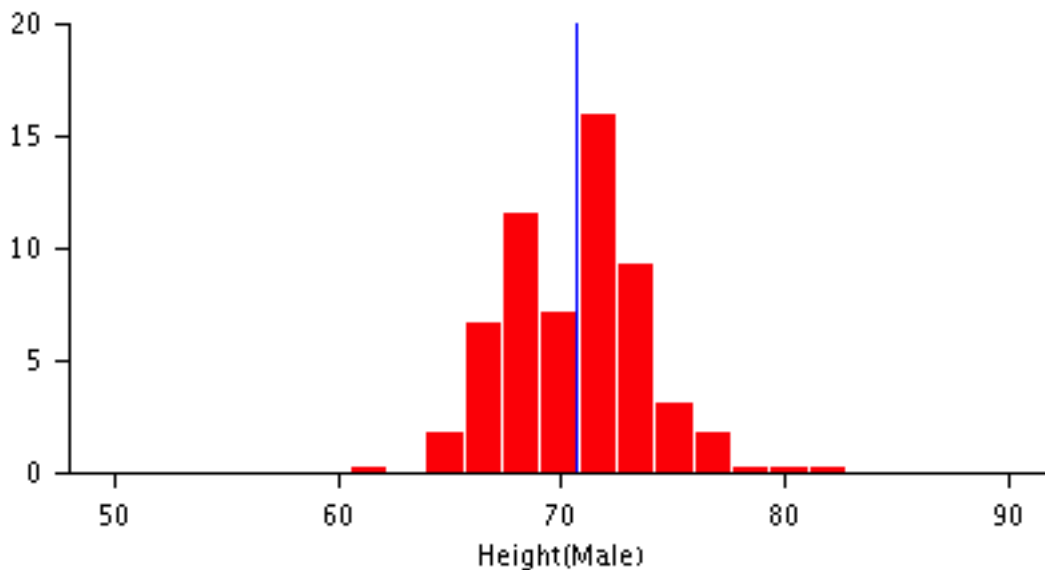
¹The average is also often called the *mean*.

Here are all of the men's heights:

74 69 72 75 69 68 74 71 67 72 72 73 72 68 74 66 70 70 75 72 70 65
 72 66 73 69 72 82 67 72 65 67 71 77 69 74 69 71 74 70 70 68 74 69
 74 70 70 71 67 73 70 64 73 69 73 75 74 69 67 68 76 62 72 70 70 72
 75 67 68 75 69 70 72 70 71 70 71 74 72 65 72 68 72 67 69 73 70 74
 69 70 70 72 67 73 69 76 73 70 69 77 70 70 69 71 75 68 68 69 72 71
 72 67 75 73 70 68 72 73 73 66 75 73 70 66 71 69 67 72 71 74 71 75
 68 68 67 72 68 73 69 67 74 72 70 72 71 67 70 68 71 72 66 71 69 66
 78 69 71 68 71 74 67 72 74 76 72 73 68 71 73 72 71 69 71 68 72 72
 67 73 68 72 68 72 67 69 72 72 71 70 70 65 68 64 75 70 69 70 72 69
 74 67 71 81 68 73 75 71 69 64 71 70 73 72 74 72 72 67 67 75 76 68
 76 72 67 72 73 73 71

There are 227 of them. You'd not want to do this by hand, but the average again sums the numbers, and divides by 227:

$$\begin{aligned}\text{Average} &= \frac{\text{Sum}}{\# \text{ of values}} = \frac{74 + 69 + 72 + \cdots + 73 + 71}{227} \\ &= \frac{16053}{227} = 70.718.\end{aligned}$$



The vertical blue line represents the average, 70.718. It is in the midst of all the data.

Median

Here is the small set of data again:

Data: 1, 6, 10, 10, 10, 0, 5

(There are 7 of them.) The median is the middle number.

First, write the data in order, from smallest to largest:

Data in order: 0, 1, 5, 6, 10, 10, 10

The middle number is the fourth one, since there are three below it, and three above it. In this case, the fourth one is 6, so the median is 6.

? Put these numbers in order, from smallest to largest: 74 69 72 75
69

What is the median?

The median with an even number of data points

If there is an even number of data points, you take half-way between the two middle ones. For example:

Data: 25, 33, 32, 50, 12, 15

Data in order: 12, 15, 25, 32, 33, 50

The median is half-way between 25 and 32:

$$\text{Median} = \frac{25 + 32}{2} = \frac{57}{2} = 28.5.$$

? Put these numbers in order, from smallest to largest:

74 69 72 75 69 68

What are the middle two numbers?

What is the median?

The median when there are ties

If there are ties, you still find the middle one after putting the data in order:

Data: 4, 6, 3, 2, 6, 3, 3, 4, 4, 4, 7 (11 of them)

Data in order: 2, 3, 3, 3, 4, 4, 4, 4, 6, 6, 7

There are several 4's, but one of the 4's is in the middle, so it is the median.

? Suppose the data consist of six 10's and nine 20's. Write the fifteen values in order:

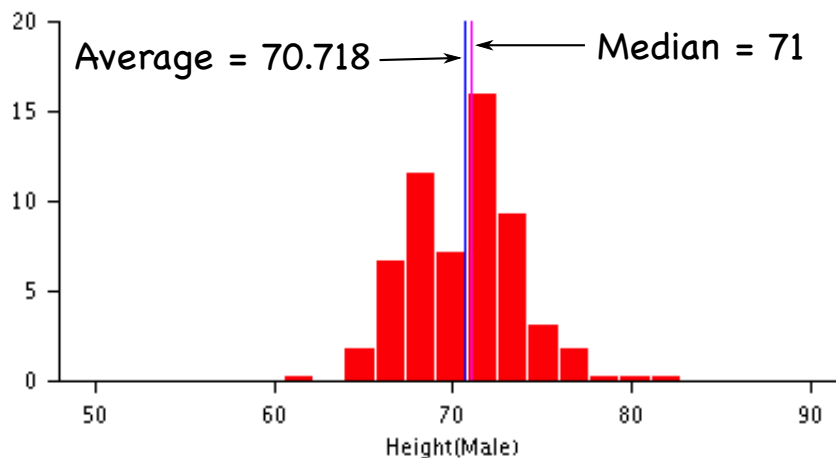
What is the median?

The median of the men's heights

Here are the men's heights, but now in order:

62 64 64 64 65 65 65 65 66 66 66 66 66 66 67 67 67 67 67 67 67 67
 67 67 67 67 67 67 67 67 67 67 67 67 68 68 68 68 68 68 68 68 68 68
 68 68 68 68 68 68 68 68 68 68 68 68 69 69 69 69 69 69 69 69 69 69
 69 69 69 69 69 69 69 69 69 69 69 69 69 69 70 70 70 70 70 70 70 70
 70 71 71 71
 71 71 71 **71** 71 71 71 71 71 71 71 71 71 71 71 71 71 71 71 71 71 72
 72
 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 73 73 73 73 73 73
 73 73 73 73 73 73 73 73 73 73 73 73 73 73 74 74 74 74 74 74 74 74
 74 74 74 74 74 74 74 74 75 75 75 75 75 75 75 75 75 75 75 75 76 76 76
 76 76 77 77 78 81 82

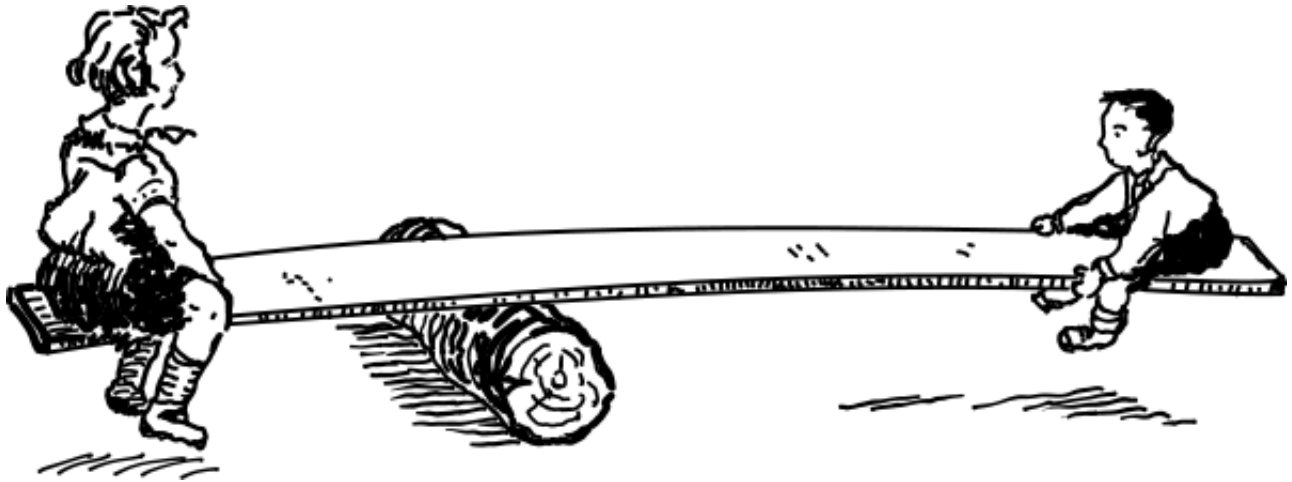
There are 227 men, so the middle one turns out to be the 114th (There are 113 below that one, and 113 above it.) The median is 71. Notice that there are quite a few 71's.



The median is also in the midst of the data. For these heights, the mean and the median are very close to each other. It does not always happen that way, as we'll see.

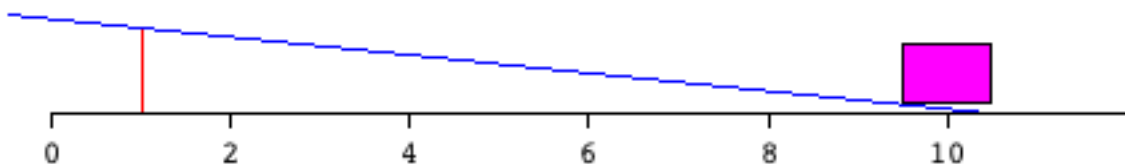
Balancing a histogram

How do you balance two people on ends of a teeter-totter, when one is big and the other small?



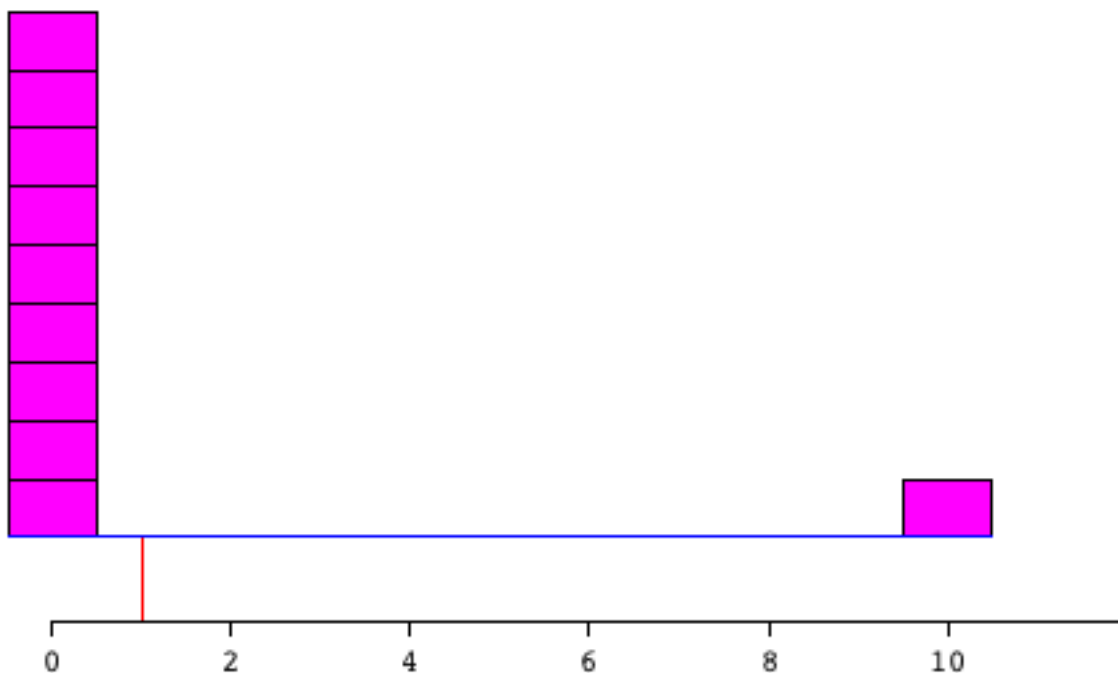
You don't balance the board right in the middle between the two people. The smaller person should be farther away from the balancing point. Being farther away has the effect of having more weight.

On the teeter-totter below, how many blocks should you put at 0 to balance the one at 10? The fulcrum (the vertical stick that the teeter-totter should balance upon) is at the value 1.



? How many blocks should you put at 0 to balance the one at 10?

It turns out you need nine 0's.



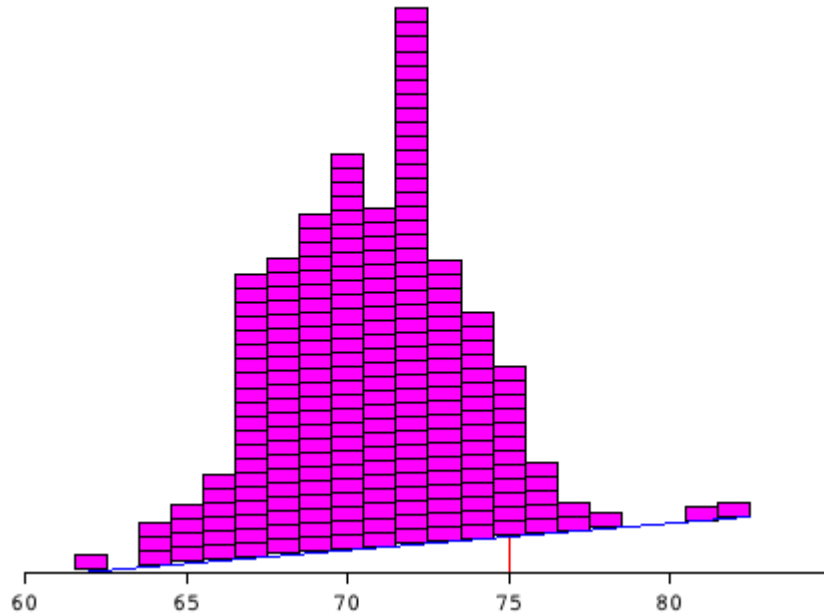
The average of nine 0's and one 10 is 1. So if you think of boxes as making up a histogram, the histogram balances when the fulcrum is at the average:

That is,

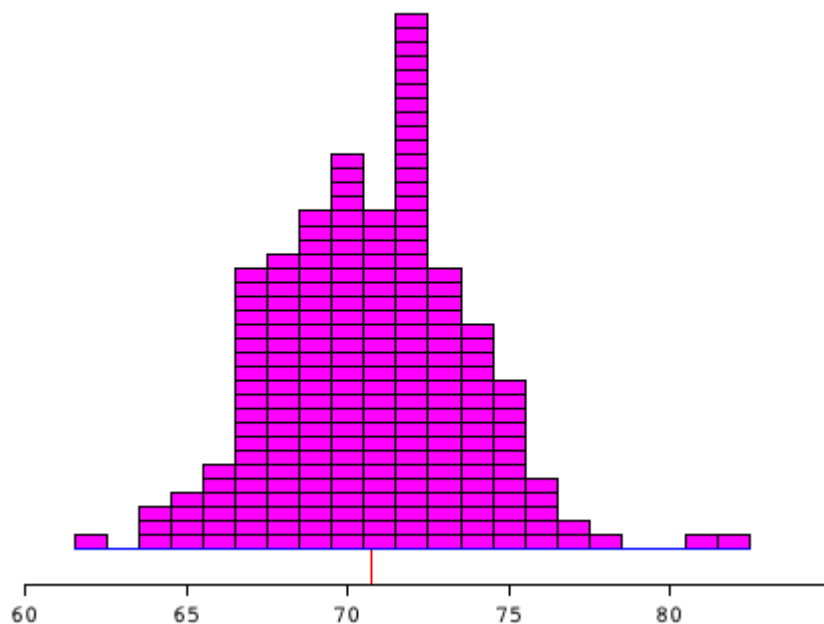
The histogram balances at the average

Balance a histogram at the average

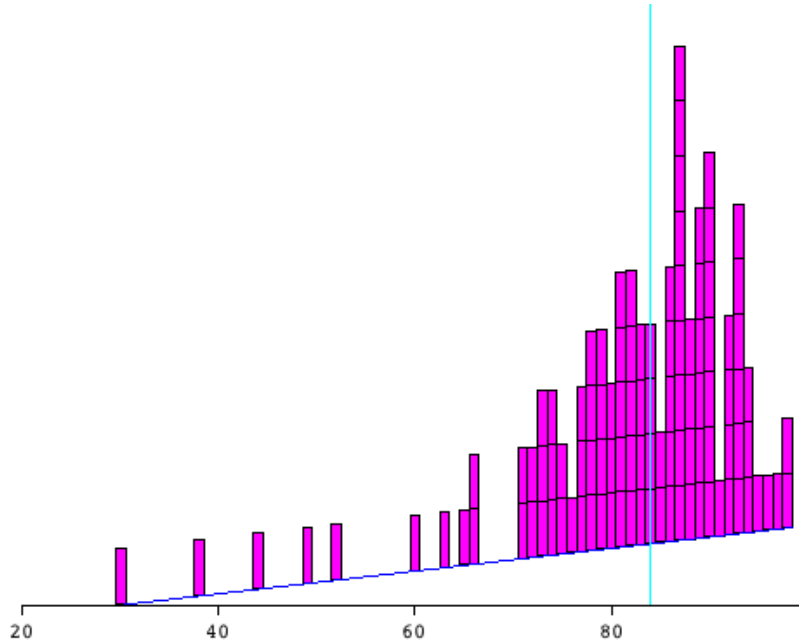
Here are the heights of the men again:



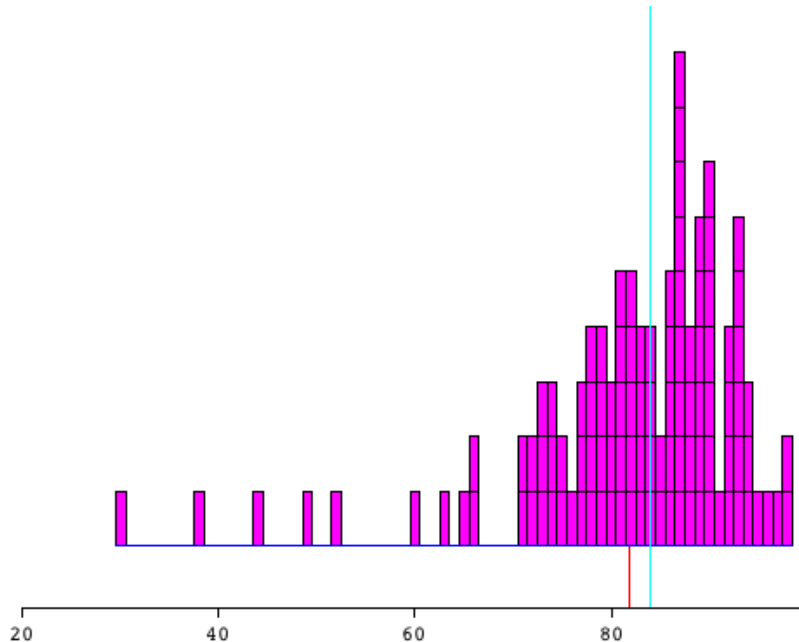
Putting the fulcrum at 75 does not work. There is too much weight to the left. But if we put the fulcrum at the average, 70.718, then the histogram does balance:



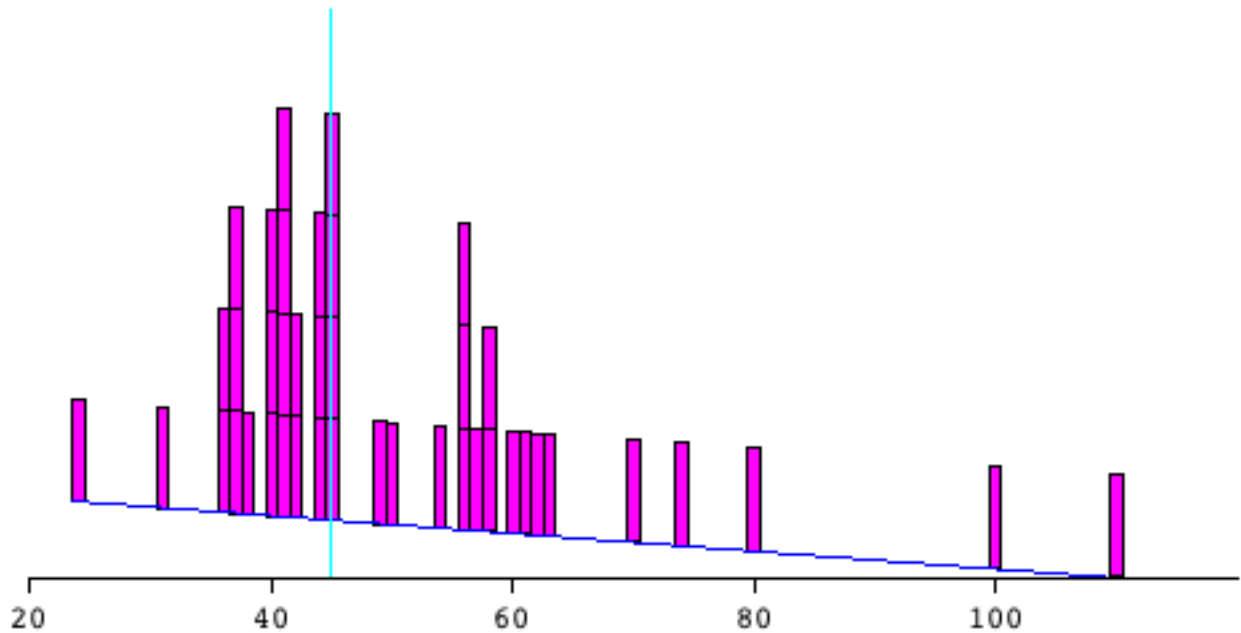
The median cuts the area of a histogram into two equal halves (each 50%). Here is a histogram for the scores on the homework in a previous class, where the fulcrum is at the median = 84.



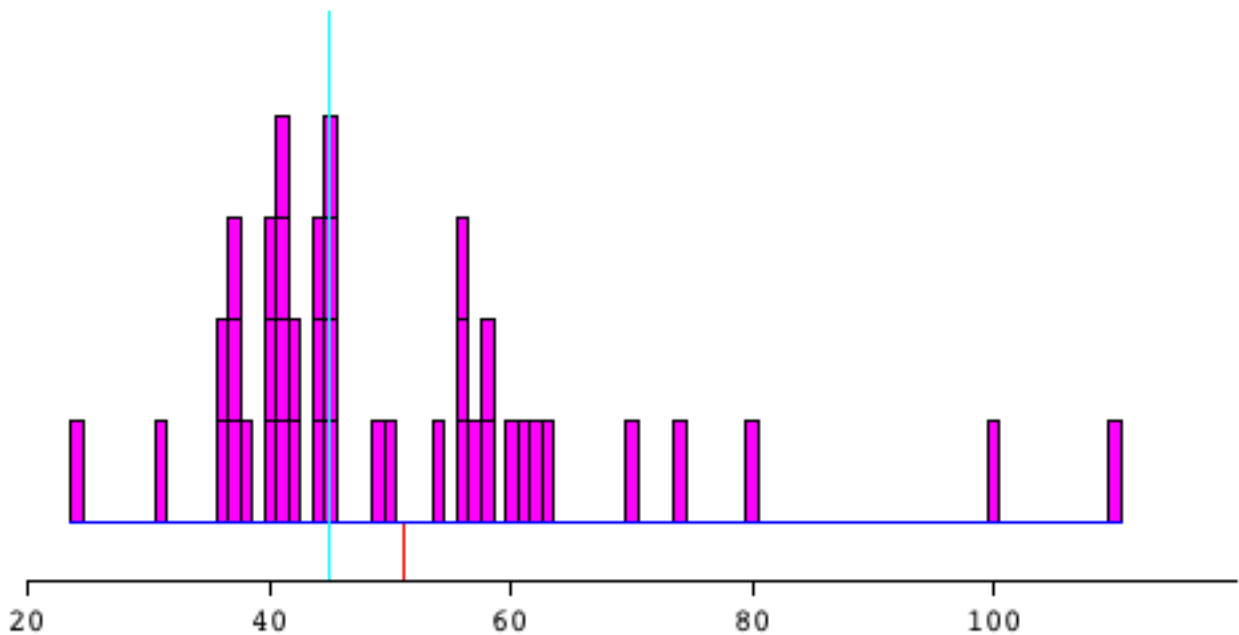
It doesn't balance at the median. To balance those few bars at the far left, we have to move the fulcrum to the left. Move it to the average = 81.78. So the average is **less than** the median.



Here are the numbers of stories of 42 Chicago buildings. The median is 45 stories.



But now, with a couple of super-tall buildings, we have to move to the right to have the histogram balance at the average = 51.



In this case, the average is **greater than** the median.

Average vs. median

The average and median are both ways to describe the typical value of a set of data. Although often they are very similar, they do measure different aspects of a histogram.

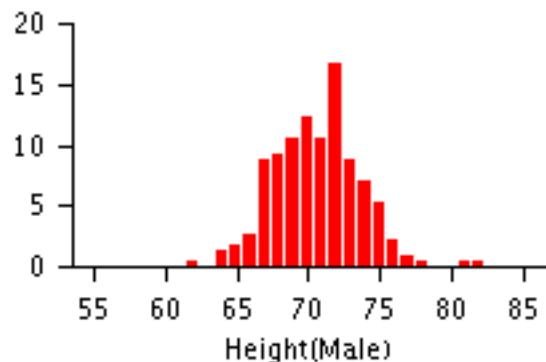
- The average is the point at which the histogram balances.
- The median is the value that splits the area of the histogram into two equal halves.

You can often decide which is larger by looking at the histogram. The basic idea is to find where the bulk of the data are, then see whether the data far away from the bulk tends to be to the right (larger), or to the left (smaller).

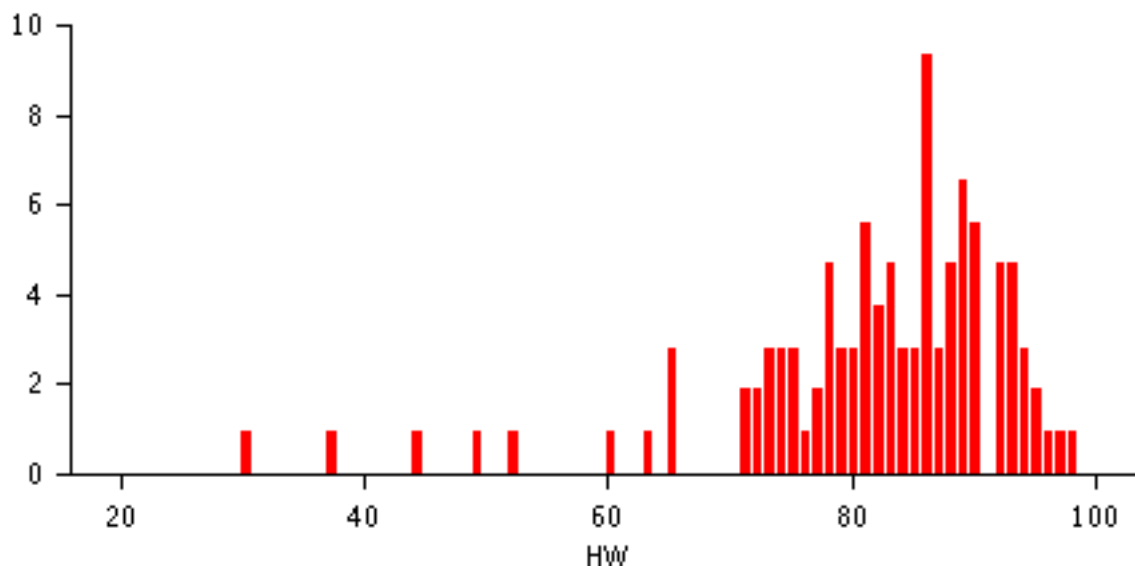
- If the far away values are to the **right**, then the average is likely to be **larger than** the median.
- If the far away values are to the **left**, then the average is likely to be **smaller than** the median.

The reason is due to the teeter-totter effect: If there are very large values, we have to move the average towards the right. If there are very small values, we have to move the average to the left.

If the histogram is fairly symmetric, that is, histogram to the left of the median is approximately a mirror image of the histogram to the right of the median, then the average and median should be about the same. It would be very difficult to decide which is larger. For example, the men's heights is fairly symmetric. The average is 70.72 and the median is 71, very close.



Here is a histogram of homework scores:



? Where are the bulk of the data?

Are the values far away from the bulk to the right or to the left?

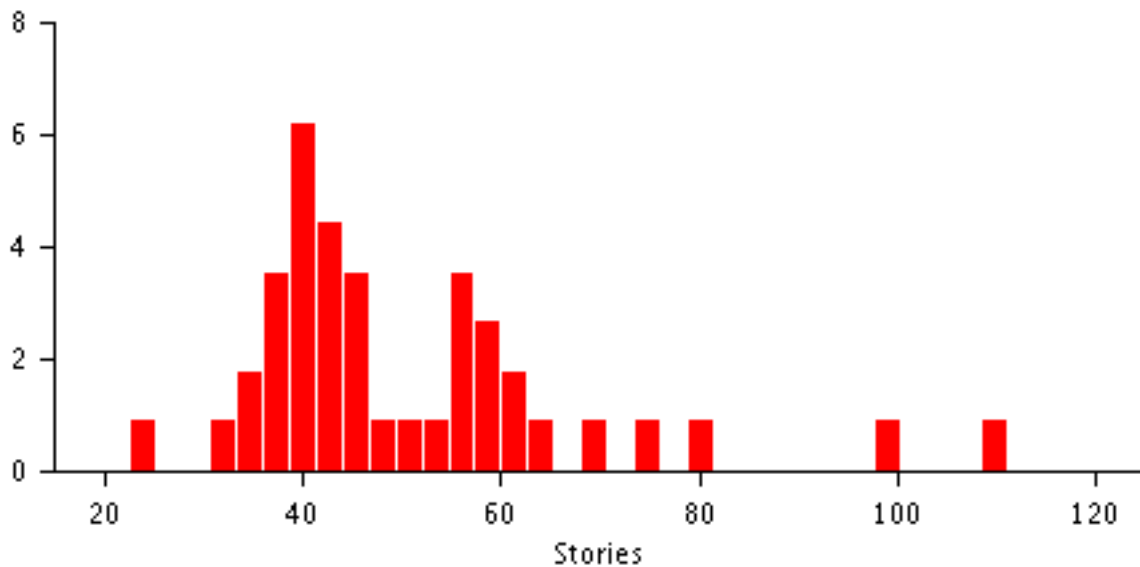
Recall:

- If the far away values are to the **right**, then the average is likely to be **larger than** the median.
- If the far away values are to the **left**, then the average is likely to be **smaller than** the median.

Is the average larger than or smaller than the median?²

²See page 68.

Next is a histogram of the numbers of stories of tall Chicago buildings:



? Where are the bulk of the data?

Are the values far away from the bulk to the right or to the left?

Recall:

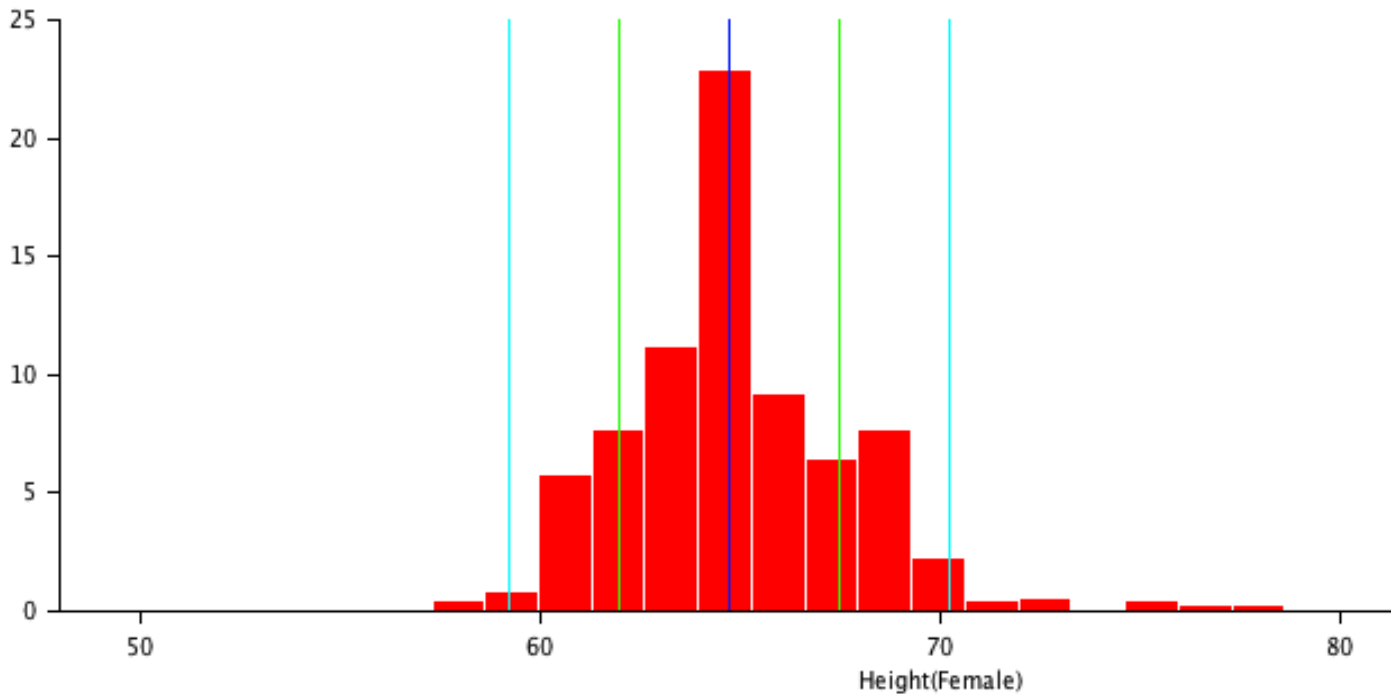
- If the far away values are to the **right**, then the average is likely to be **larger than** the median.
- If the far away values are to the **left**, then the average is likely to be **smaller than** the median.

Is the average larger than or smaller than the median?³

³See page 69.

The standard deviation

Here are the heights of the women again. The middle (blue) line is the average, 64.72 inches.



But are all women 64.72 inches? Or about 64.72 inches? Or are most about 64.72 inches? The average (& the median) do not give an idea of how variable heights can be.

The graph has other vertical lines:

- About 68% of the area in the histogram is between the second and fourth (light green) lines. (Seem reasonable?)
- About 95% of the area is between the two outer (light blue) lines.

The distance between two adjacent lines is called the **standard deviation**. Instead of just having the average, one also has a “plus-or-minus” number to give an idea how close people are to the average. That \pm number is the standard deviation.

Deviations

The average height of the women is 64.72 inches. The deviation of a particular woman is how much taller or shorter she is than the average. So if a woman has height = 70 inches, the deviation is

$$\text{Deviation} = \text{Height} - \text{Average height} = 70 - 64.72 = 5.28 \text{ inches.}$$

She's 5.28 inches taller than average. Or, for someone 62 inches, say:

$$\text{Deviation} = \text{Height} - \text{Average height} = 62 - 64.72 = -2.72 \text{ inches.}$$

This is a negative deviation, meaning she is 2.72 inches shorter than average.

? Find the deviation for a woman with height = 65.

For height = 63.

Say Woman A has height 70, woman B has height 62, and woman C has height 65.

- Which of the three has the smallest deviation, ignoring sign. (So is closest to the average.)
- Which of the three has the largest deviation, ignoring sign. (So is farthest from the average.)
- Which have positive deviations? Which are taller than average?

The “standard” deviation

Do you want to see the deviations for all 485 women?

3.28 -0.72 0.28 -0.72 -0.72 -3.72 1.28 1.28 -0.72 -0.72 -2.72 5.28 -0.72 -2.72 0.28 -0.72 0.28 4.28 -0.72 1.28
 0.28 3.28 3.28 -2.72 0.28 3.28 -1.72 1.28 -0.72 0.28 -2.72 -4.72 1.28 -2.72 -2.72 1.28 -3.72 -1.72 -0.72 3.28
 4.28 2.28 -0.72 -1.72 -0.72 -0.72 -0.72 1.28 -1.72 -0.72 -0.72 -0.72 -1.72 -4.72 1.28 0.28 0.28 0.28 -2.72 7.28
 2.28 -0.72 -1.72 3.28 -2.72 5.28 -5.72 -1.72 3.28 -2.72 -1.72 -0.72 0.28 2.28 -1.72 -0.72 -2.72 -1.72 -1.72 -5.72
 0.28 1.28 1.28 -4.72 -0.72 -0.72 0.28 -3.72 -1.72 2.28 -1.72 -4.72 -1.72 -2.72 -2.72 0.28 -0.72 -0.72 -0.72 5.28
 1.28 -0.72 -4.72 0.28 -2.72 -2.72 0.28 0.28 -1.72 0.28 -3.72 -0.72 -2.72 2.28 -1.72 -2.72 2.28 -0.72 -5.72 1.28
 -0.72 -2.72 -1.72 2.28 0.28 0.28 -0.72 0.28 -0.72 0.28 1.28 0.28 6.28 -4.72 3.28 -2.72 1.28 -1.72 -2.72 1.28
 -2.72 -2.72 -1.72 5.28 -1.72 3.28 -0.72 -1.72 -2.72 -1.72 3.28 -0.72 -2.72 -1.72 2.28 -4.72 2.28 -0.72 -2.72 2.28
 -1.72 -0.72 4.28 -1.72 5.28 -1.72 3.28 -1.72 5.28 0.28 -2.72 1.28 -0.72 5.28 -1.72 -1.72 -0.72 0.28 -1.72 -1.72
 0.28 1.28 -4.72 -3.72 1.28 2.28 1.28 -3.72 2.28 -1.72 3.28 0.28 -3.72 5.28 2.28 1.28 -6.72 -1.72 -4.72 0.28 1.28
 -2.72 1.28 -0.72 4.28 -3.72 2.28 0.28 2.28 4.28 -0.72 3.28 -2.72 1.28 -0.72 -2.72 1.28 1.28 1.28 -0.72 1.28 -0.72
 -2.72 -1.72 -0.72 -0.72 0.28 1.28 -0.72 3.28 -2.72 1.28 -0.72 -4.72 -0.72 0.28 -0.72 0.28 -2.72 -0.72 2.28 2.28
 0.28 0.28 -1.72 1.28 0.28 3.28 2.28 -2.72 2.28 -1.72 -0.72 -1.72 3.28 0.28 -0.72 5.28 -4.72 -1.72 -1.72 -1.72
 3.28 1.28 3.28 5.28 -0.72 -0.72 1.28 -1.72 2.28 1.28 2.28 0.28 -1.72 -3.72 -0.72 4.28 -4.72 2.28 -2.72 1.28 1.28
 -0.72 -0.72 3.28 0.28 -1.72 1.28 3.28 2.28 -0.72 -0.72 2.28 -1.72 -2.72 0.28 -0.72 -0.72 -3.72 0.28 12.28 0.28
 -1.72 1.28 1.28 3.28 -2.72 3.28 -4.72 2.28 -0.72 -1.72 -0.72 -1.72 5.28 -0.72 -1.72 2.28 -1.72 -4.72 -3.72 -0.72
 1.28 -0.72 -4.72 3.28 -2.72 1.28 3.28 -2.72 -0.72 -0.72 -3.72 -1.72 3.28 0.28 -2.72 -1.72 1.28 3.28 0.28 -2.72
 0.28 -1.72 -0.72 -1.72 1.28 -3.72 -1.72 1.28 -0.72 1.28 2.28 -0.72 -0.72 0.28 2.28 0.28 -5.72 6.28 1.28 3.28 1.28
 2.28 -0.72 -0.72 1.28 -3.72 -2.72 1.28 0.28 -0.72 0.28 7.28 -0.72 0.28 2.28 3.28 -2.72 0.28 -0.72 -6.72 -2.72
 4.28 -0.72 -2.72 -2.72 10.28 0.28 1.28 2.28 -1.72 1.28 2.28 -0.72 2.28 2.28 1.28 2.28 1.28 -0.72 0.28 -3.72 3.28
 4.28 -0.72 -2.72 3.28 -0.72 5.28 4.28 -2.72 0.28 -0.72 -2.72 -1.72 0.28 -3.72 0.28 -1.72 1.28 -3.72 -3.72 0.28
 -1.72 3.28 1.28 1.28 -4.72 0.28 2.28 1.28 -0.72 -1.72 -5.72 -1.72 -3.72 1.28 -1.72 -1.72 -1.72 2.28 1.28 -2.72
 4.28 2.28 10.28 4.28 0.28 4.28 2.28 -1.72 -0.72 -0.72 -1.72 -0.72 -1.72 4.28 -1.72 3.28 -0.72 8.28 4.28 -3.72
 5.28 -3.72 -1.72 -1.72 -0.72 -1.72 4.28 -1.72 0.28 5.28 2.28 2.28 13.28 -2.72 4.28 0.28 -1.72 1.28 3.28 0.28

The **standard** deviation is the typical size of the deviations. We cannot just take the average, however. The plusses and minusses would cancel, leaving us nothing (0):

The average of the deviations is always 0.

? What would you guess is a typical size of these deviations, ignoring signs?

Calculating the Standard Deviation

First, let's call it the SD, for short. There are four steps for calculating the SD:

1. Deviation: Calculate the deviations
2. Square: Square the deviations
3. Mean: Take the average of the squared deviations
4. Root: Take the square root

Start with the data: 18, 20, 20, 23, 19. (So there are 5 values.)

The average is

$$\text{Average} = \frac{18 + 20 + 20 + 23 + 19}{5} = \frac{100}{5} = 20.$$

Step 1: Deviations — subtract 20 from each of the data values.

Step 2: Square them. The results are all positive.

Value - Average	Deviation	Squared
18 - 20	-2	$(-2)^2 = 4$
20 - 20	0	$(0)^2 = 0$
20 - 20	0	$(0)^2 = 0$
23 - 20	3	$(3)^2 = 9$
19 - 20	-1	$(-1)^2 = 1$

Step 3: Find the average of those squared deviations:

$$\frac{4 + 0 + 0 + 9 + 1}{5} = \frac{14}{5} = 2.8.$$

Step 4: Take the square root of that average:

$$\text{SD} = \sqrt{2.8} = 1.67.$$

The root mean square deviation

To recap: The data 18, 20, 20, 23, 19,

have deviations -2, 0, 0, 3, -1. (Note that the deviations do have an average of 0.)

The $SD = 1.67$. Is that a reasonable value for the typical size of the deviations? (Should be at least ok.)

We calculated with the steps

1. Deviation: Calculate the deviations
2. Square: Square the deviations
3. Mean: Take the average of the squared deviations
4. Root: Take the square root

The SD is often called the **root mean square deviation** of the data, which is the four steps, listed backwards.

The SD of the women's heights

Here again are some of the deviations for the women's heights ($n=485$):

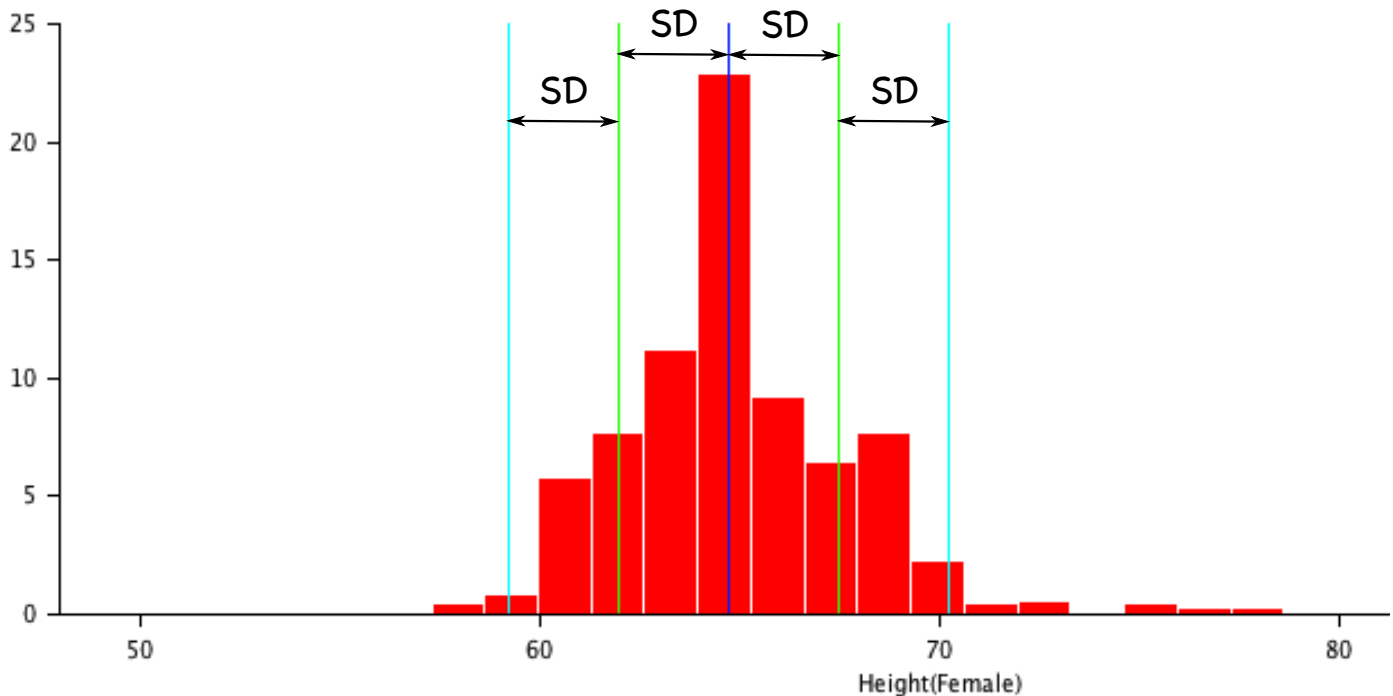
3.28 -0.72 0.28 -0.72 -0.72 -3.72 1.28 1.28 -0.72 -0.72 -2.72 5.28 -0.72 -2.72 0.28
-0.72 0.28 4.28 -0.72 1.28 0.28 3.28 3.28 -2.72 0.28 3.28 -1.72 1.28 -0.72 0.28 -2.72
-4.72 1.28 ... 0.28 -1.72 1.28 3.28 0.28

To find the SD, square those, then find the average of the squared deviations, which turns out to be 7.56.

? Find the SD.

For the women's heights, then,

Average = 64.72 inches, SD = 2.75 inches



The middle blue line is the average, 64.72 inches. The second and fourth light green lines are one SD ($SD = 2.75$) away from the average:

$$\begin{aligned} \text{Average} \pm SD &= 64.72 \pm 2.75 \\ &= (64.72 - 2.75, 64.72 + 2.75) \\ &= (61.97, 67.47). \end{aligned}$$

The outer two light blue lines are two SD's away from the average:

$$\begin{aligned} \text{Average} \pm 2 \times SD &= 64.72 \pm 2 \times 2.75 \\ &= (64.72 - 2 \times 2.75, 64.72 + 2 \times 2.75) \\ &= (59.22, 70.22). \end{aligned}$$

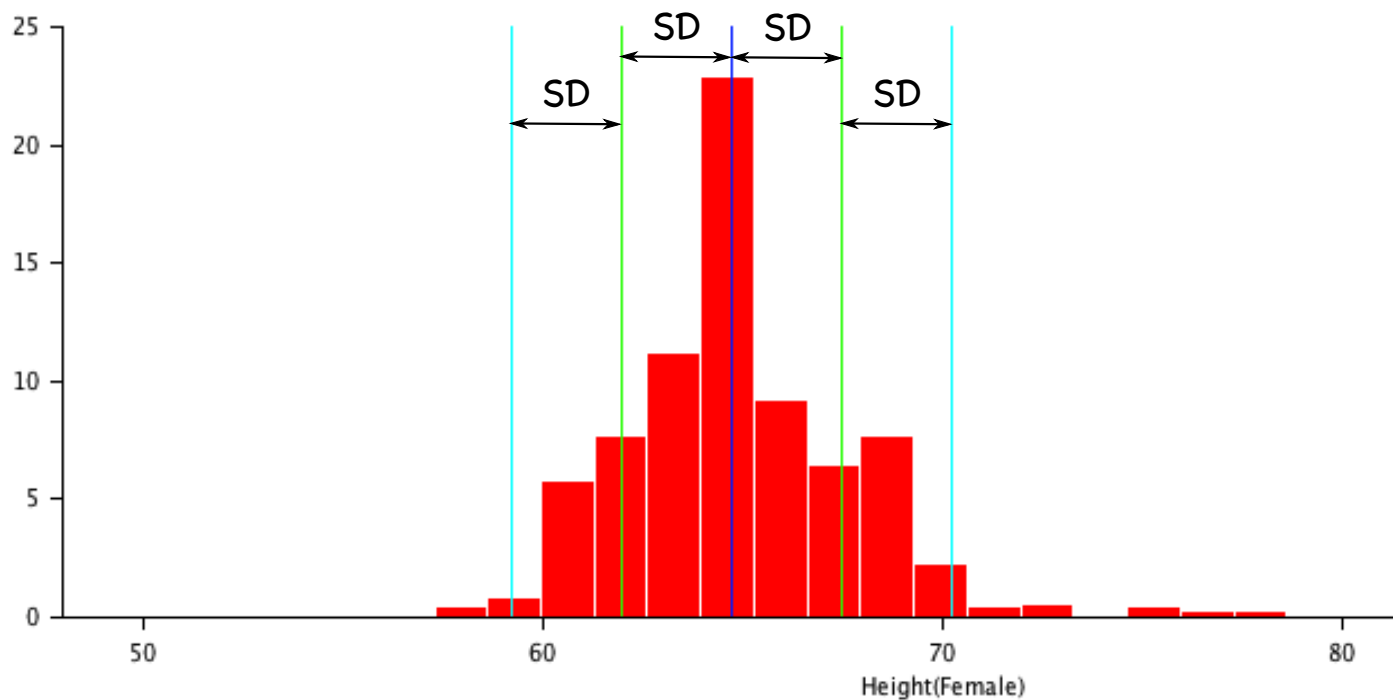
Rule of thumb

- About 68% of the data is between

$$\text{Average} \pm \text{SD} = (61.97, 67.47).$$

- About 95% of the data is between

$$\text{Average} \pm 2 \times \text{SD} = (59.22, 70.22).$$



? Here are the scores on a final exam for six students: 70, 72, 74, 78, 74, 58 The average of these scores is 71. Find the deviations:

Value	Deviation
70	
72	
74	
78	
74	
58	

What is the sum of those deviations?

What is the average of those deviations?

Match the following descriptions of deviations with the corresponding values:

- Has the smallest positive deviation.
- Has the negative deviation farthest from 0.
- Has the largest positive deviation.
- Has the deviation farthest from 0.

Without doing any more calculations, just looking at the deviations, what would you guess is the SD?

? Continuing, find the squared deviations:

Value	Deviation	Squared deviation
70	-1	
72	1	
74	3	
78	7	
74	3	
58	-13	

What is the sum of those squared deviations?

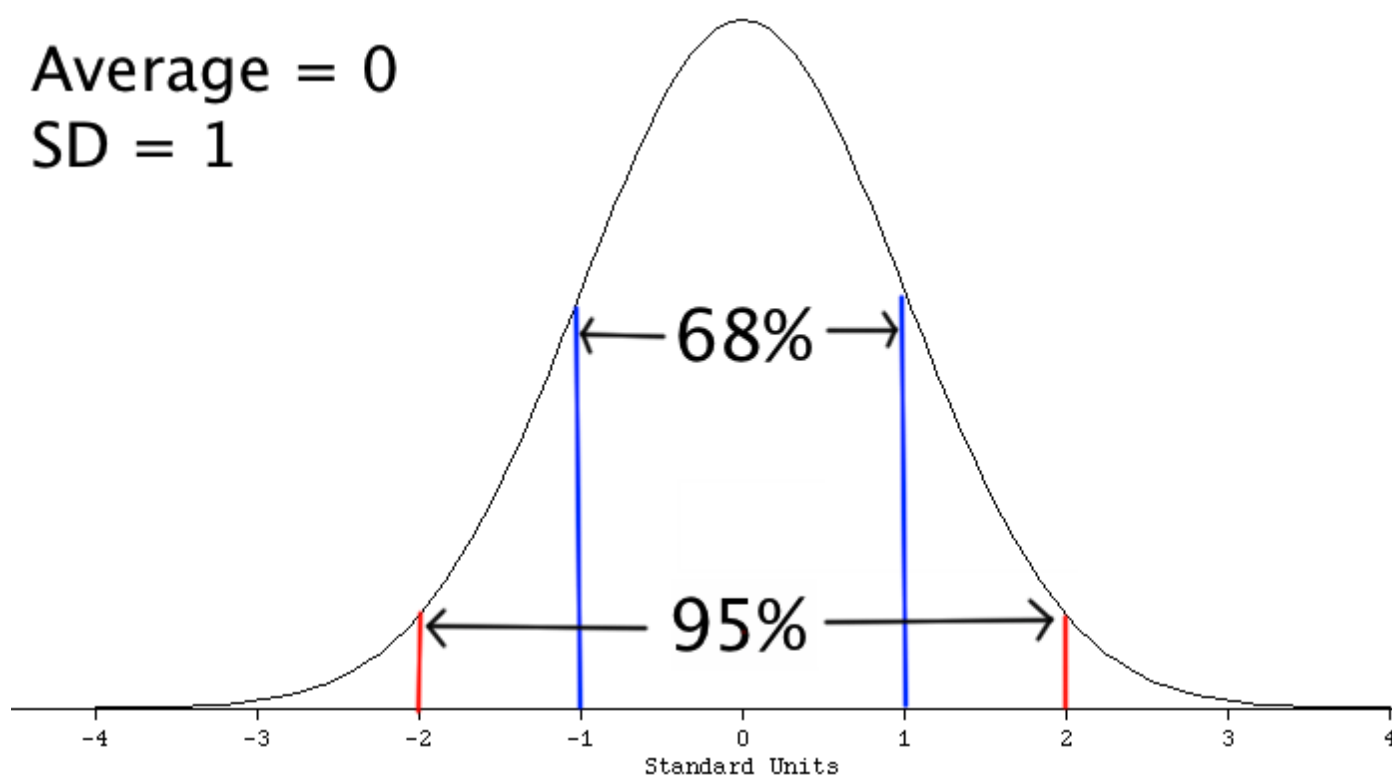
What is the average of those squared deviations?

What is the SD?

How good was your guess on the previous page?

The Normal Curve

The normal curve is an idealized histogram. The horizontal axis is not in inches or pounds or pairs of shoes, but in *standard units*:

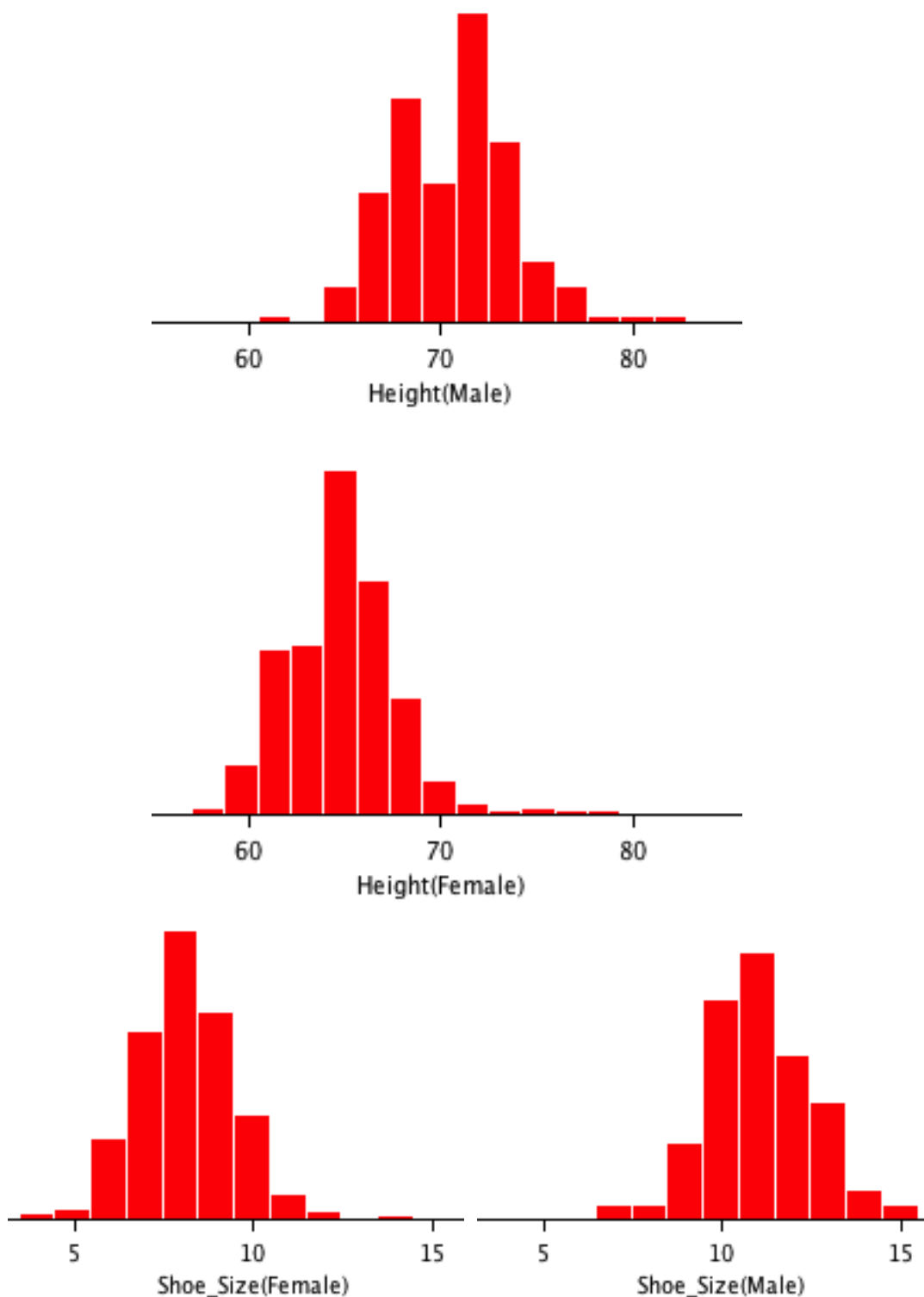


This histogram has average = 0 and SD = 1.

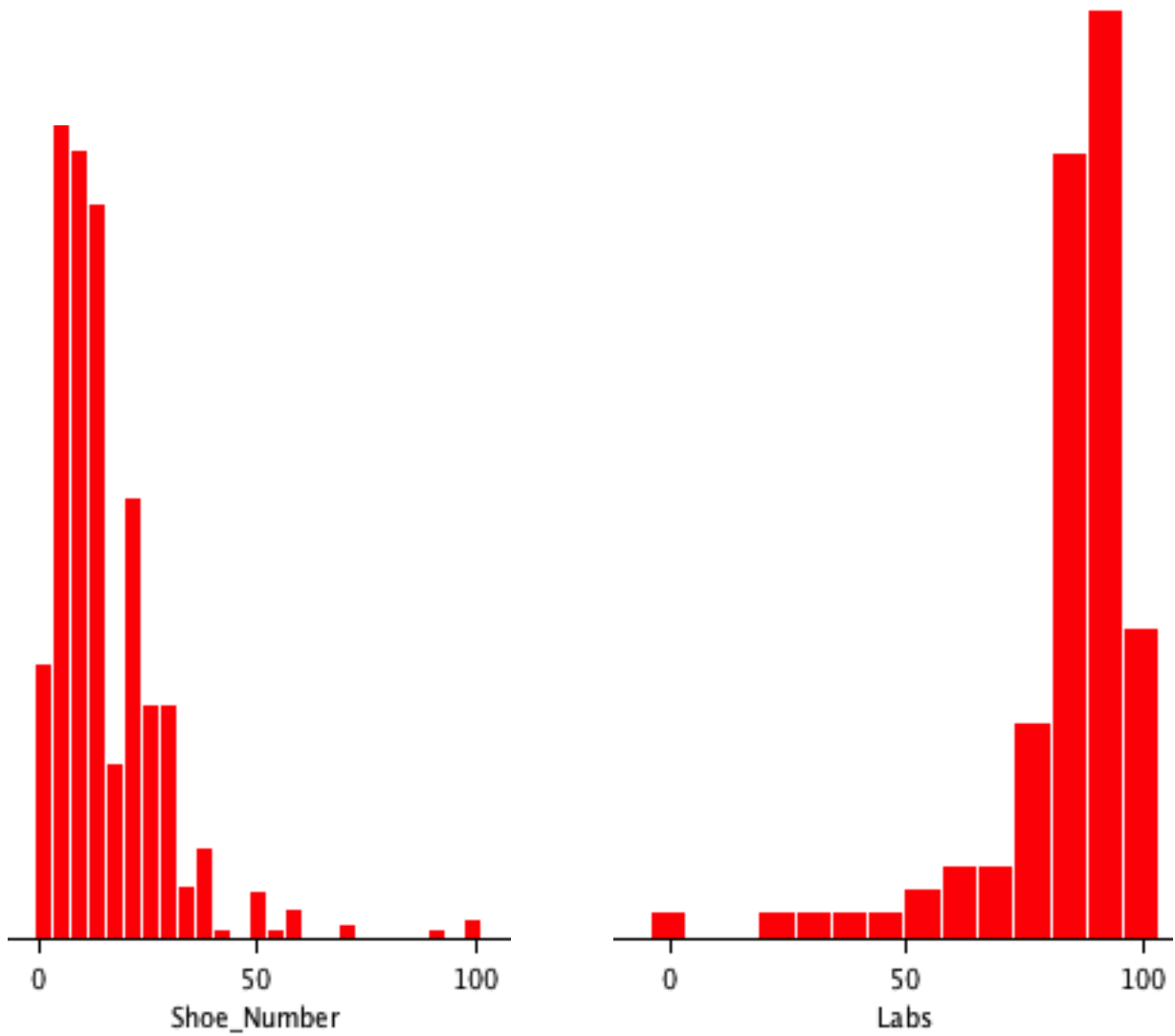
The rule of thumb works out so that

- About 68% of the area is between ± 1
- About 95% of the area is between ± 2

Some histograms look reasonably like a normal curve, such as those below. The horizontal axes are different (they are in inches, or shoe sizes, not standard units), but the shape is roughly similar to a normal curve.

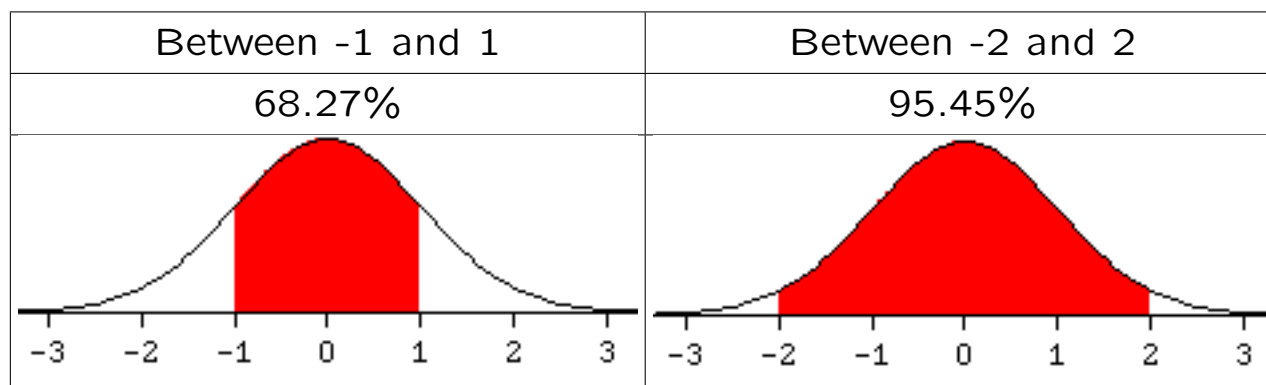


These two histograms do not look like normal curves:



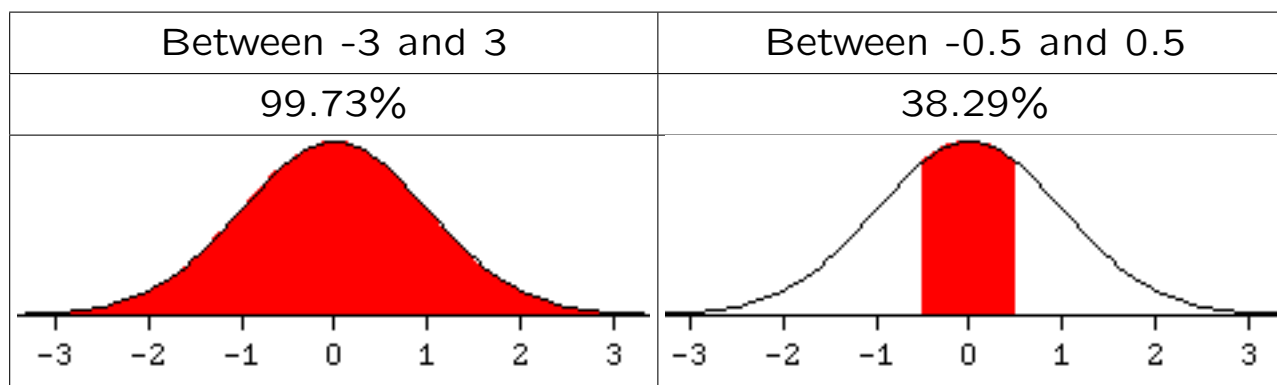
These two have values far away from the bulk of the data, either to the left or to the right.

We can use the normal curve for approximating percentages in histograms, not just 68% and 95%, but any percentage.

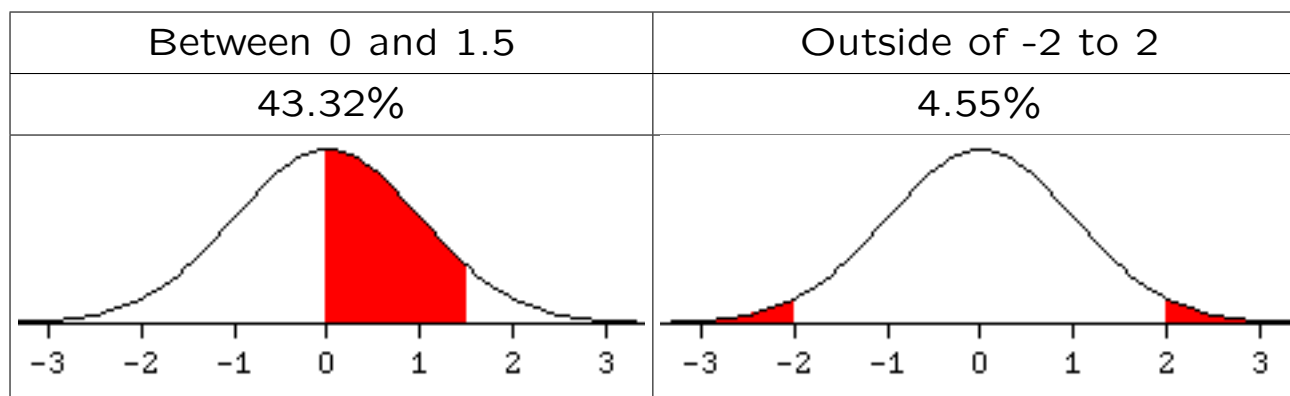


These are more exact than 68% and 95%.

Between -3 and +3 is almost everything (99.73%).



And areas not just \pm something:



So about 5% (4.55%) are outside of the ± 2 area.

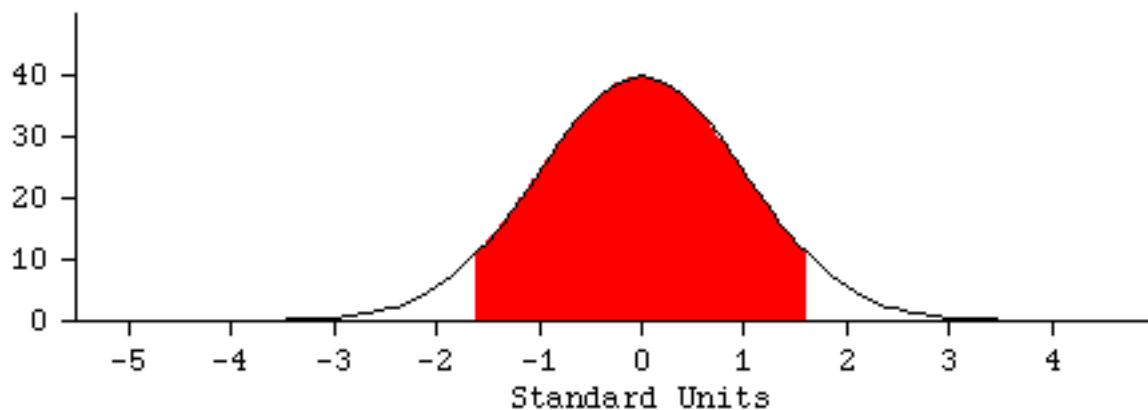
3.1 The normal curve table

How do we find areas under the normal curve? Use the Normal Table at the back of the text (or on page 389 in these notes). For numbers between 0 and 4.45, the table gives the area between plus and minus that number. Here is a small part of the table:

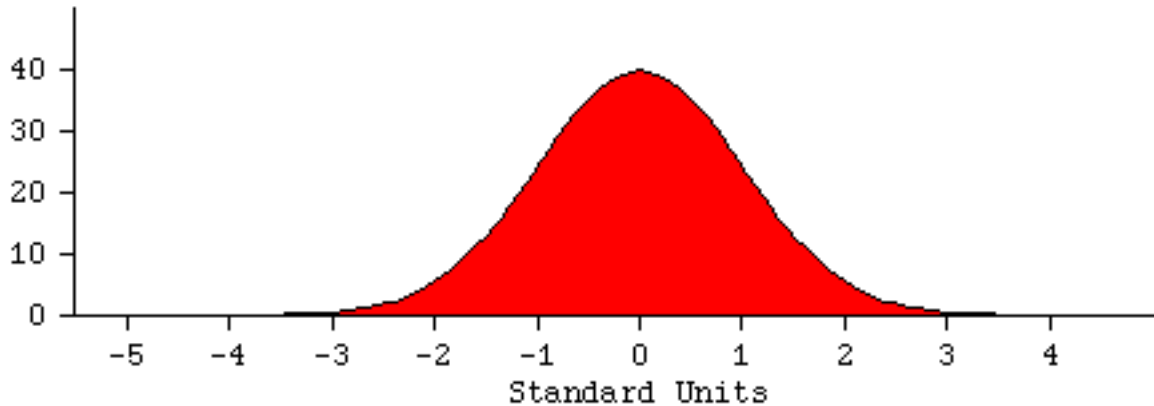
z	Height	Area
1.50	12.95	86.64
1.55	12.00	87.89
1.60	11.09	89.04
1.65	10.23	90.11
1.70	9.40	91.09

The *Area* is the area between $-z$ and $+z$, where z is your number of interest. (The heights are crossed out since we never have to use them.)

So, if you want the area between ± 1.60 , you look for 1.60 in the first column, and find the area in the third column: 89.04%.



The smallest z in the table is 0. The area between ± 0 is 0. The largest z in the table is 4.45. The area between -4.45 and $+4.45$ is 99.9991%, basically 100%.



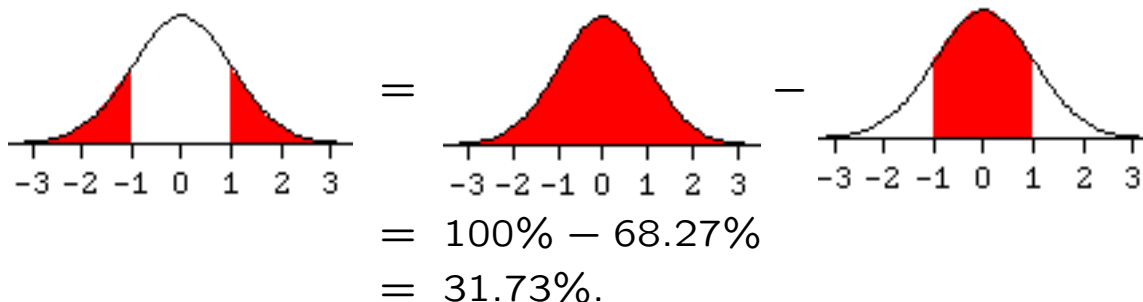
You cannot even see the area outside of ± 4.45 .

What about other areas, not just between $\pm z$?

You can work anything out, using some characteristics of the normal curve.

Example 1. The area outside of $\pm z$

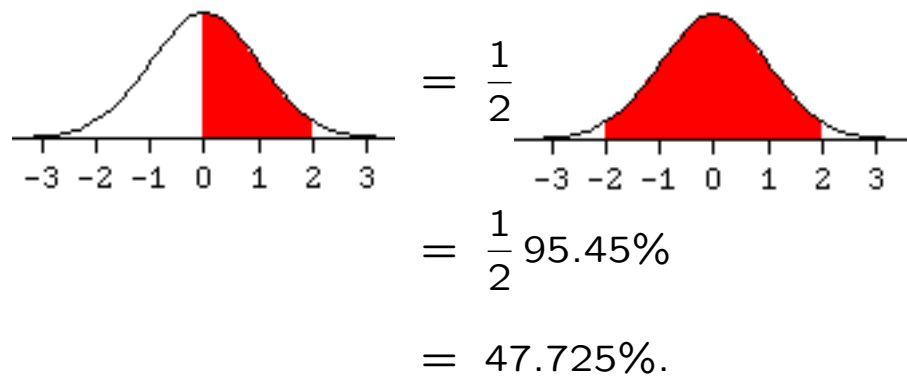
The total area is 100%, so if you want the area *outside* of -1 to 1 , you take 100% minus the area *between* -1 and 1 . The area between -1 and 1 is 68.27%, so



Area outside of -1 to $1 = 100\% - 68.27\% = 31.73\%$.

Example 2. The area from 0 to z

The normal curve is symmetric, which means the left side is a mirror image of the right side. So to find the area between 0 and 2, you take half the area between -2 and 2. The area between -2 and 2 is 95.45%, so



So the area between 0 and 2 = $\frac{1}{2} 95.45\% = 47.725\%$.

? Find the area between $-.5$ and $+.5$.

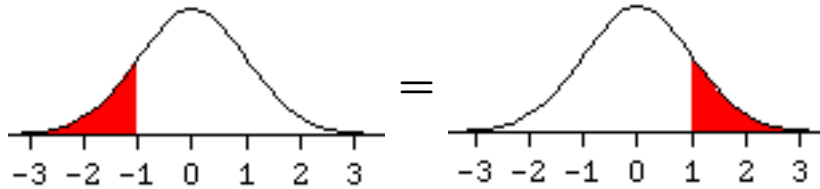
Find the area outside of $-.5$ to $+.5$.

Find the area between 0 and $+.5$.

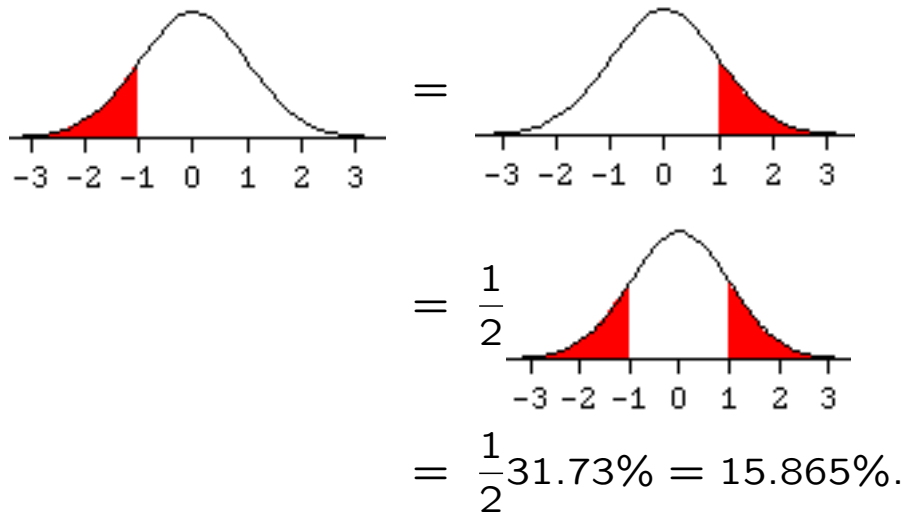
Find the area between $-.5$ and 0.

Example 3. The area above 1?

Because the normal curve is symmetric, the area above 1 is the same as the area below -1:



So they are both half of the area outside of -1 to 1. The area outside of -1 to 1 is 31.73%, from Example 1.



Area above 1 equals $\frac{1}{2} 31.73\% = 15.865\%$. The area below -1 is also 15.865%.

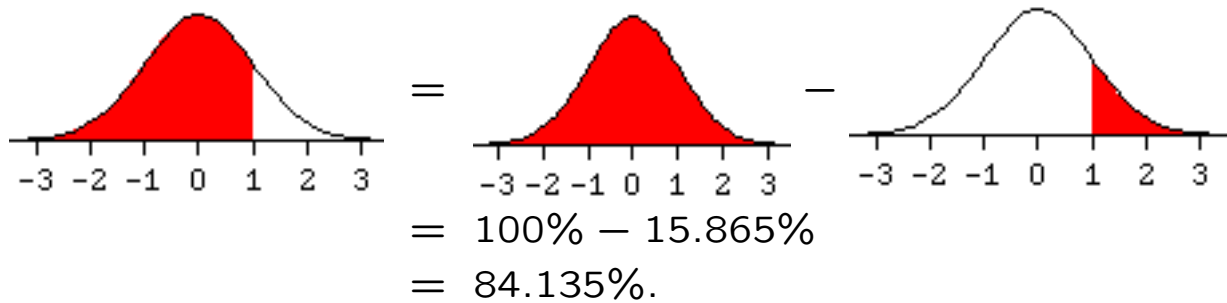
? Find the area above 0.

Find the area above $+.5$.

Find the area below $-.5$.

Example 4. The area below 1?

The area below 1 is the 100% minus the area above 1. From the previous example, the area above 1 is 15.865%.



The area below 1 = $100\% - 15.865\% = 84.135\%$.

The area above -1 is also 84.135%.

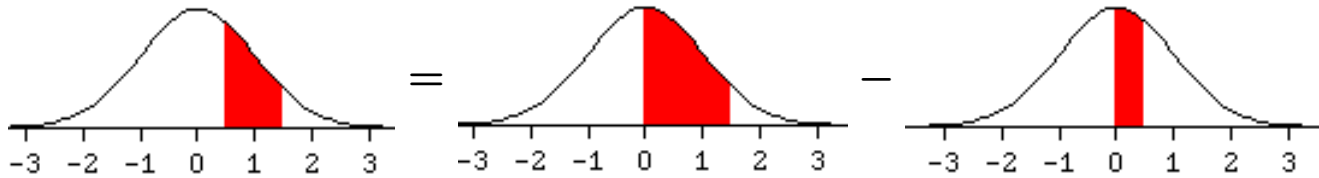
? What is the area below +.5?

What is the area above $-.5$?

What is the area below $+5$?

Example 5. The area between 0.5 and 1.5?

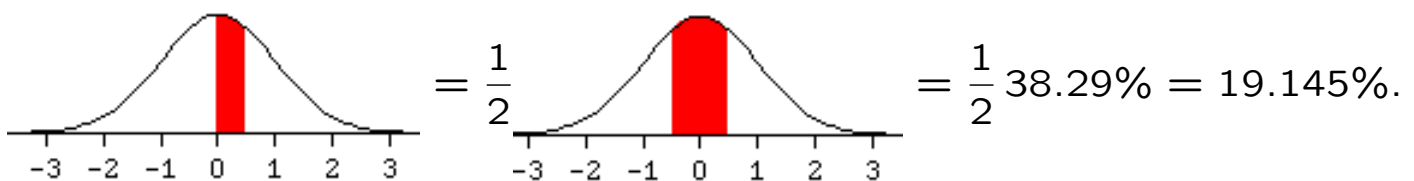
First, we know that the area between 0.5 and 1.5 is the area between 0 and 1.5 minus the area between 0 and 0.5:



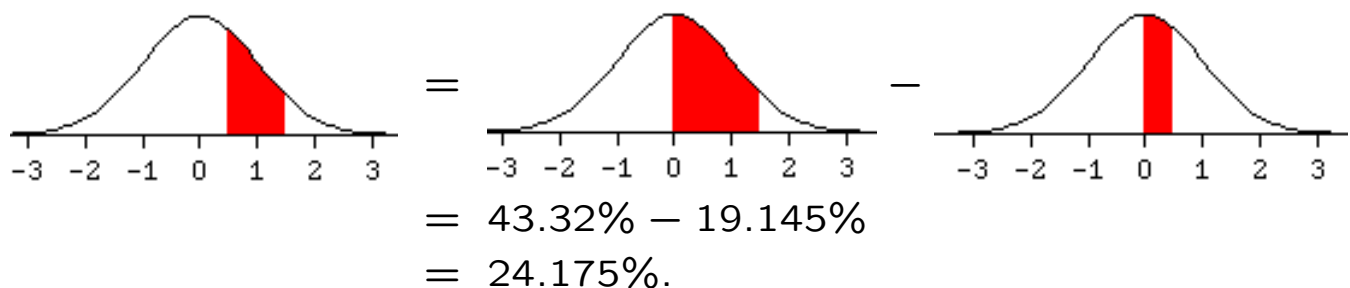
The area from 0 to z is half the area from $-z$ to z , so we do it twice, once for $z = 1.5$, and once for $z = 0.5$. The area between -1.5 and 1.5 is 86.64% (from the Normal Table), so the area between 0 and 1.5 is



The area between -0.5 and 0.5 is 38.29% (from the Normal Table), so the area between 0 and 0.5 is



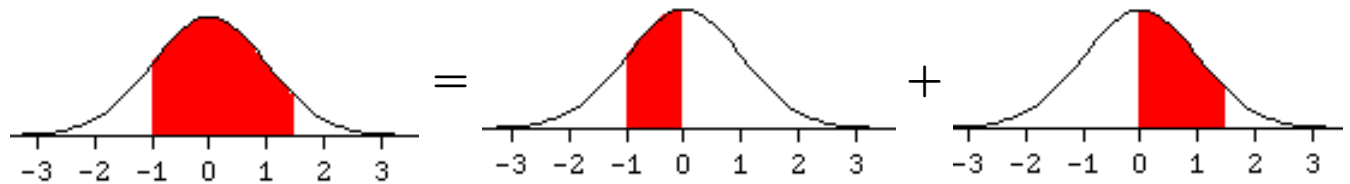
Finally,



The area between 0.5 and 1.5 is then 24.175%.

Example 6. The area between -1 and 1.5?

We can split it up into the area between -1 and 0, and the area between 0 and 1.5:



Those two areas can be found as before.

? Show that the area between -1 and 0 is 34.135%.

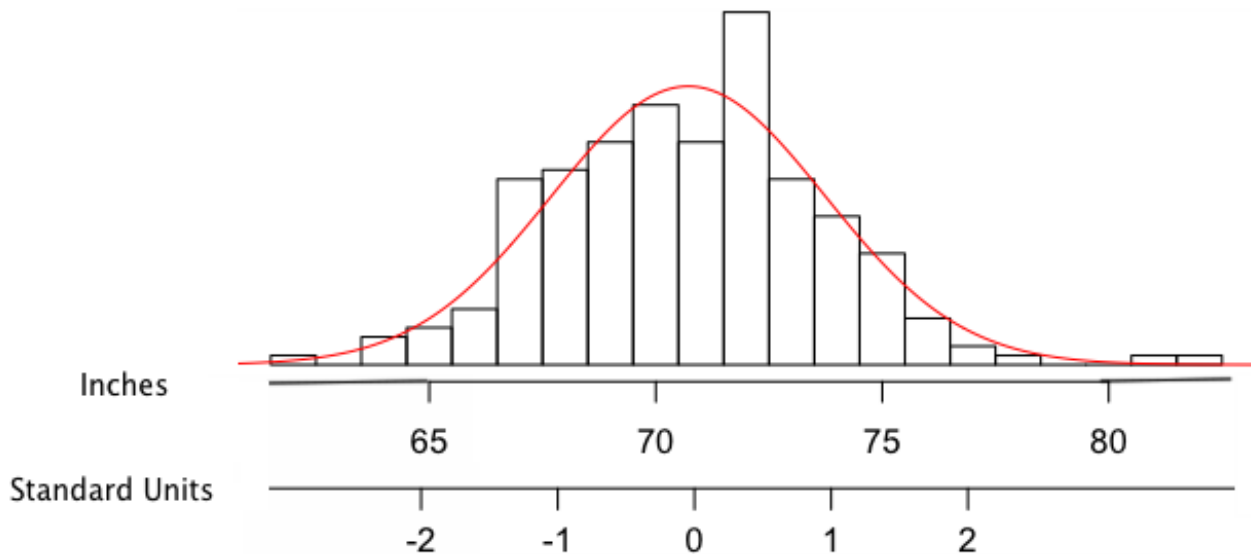
Show that the area between 0 and 1.5 is 43.32%.

Now add those two areas, so we have that

$$\text{Area between } -1 \text{ and } 1.5 = 77.455\%.$$

3.2 The normal curve for data

Here is the histogram of men's heights and the normal curve on the same graph:



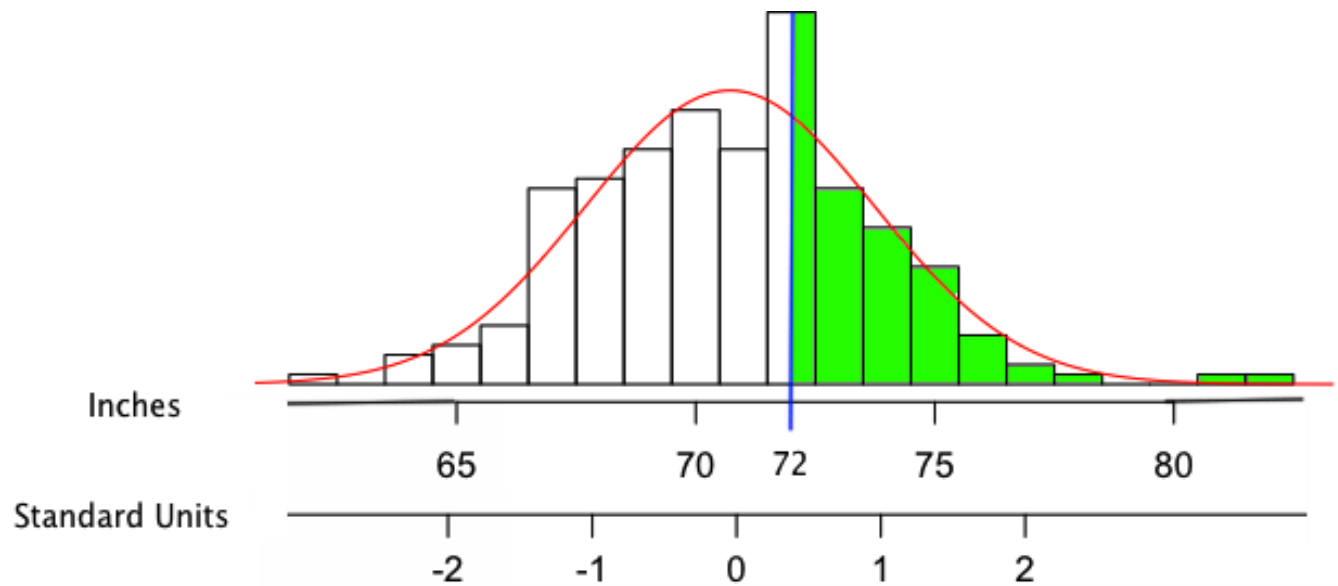
The histogram is in inches, going from about 60 to over 80. The normal curve is in standard units, going from -3 to 3. How do the inches match up with the standard units? The key is that the average connects with 0 standard units, and the average \pm SD connects with ± 1 , and the average ± 2 SD's connects with ± 2 .

Here the average = 70.72 inches and the SD = 3.01 inches.

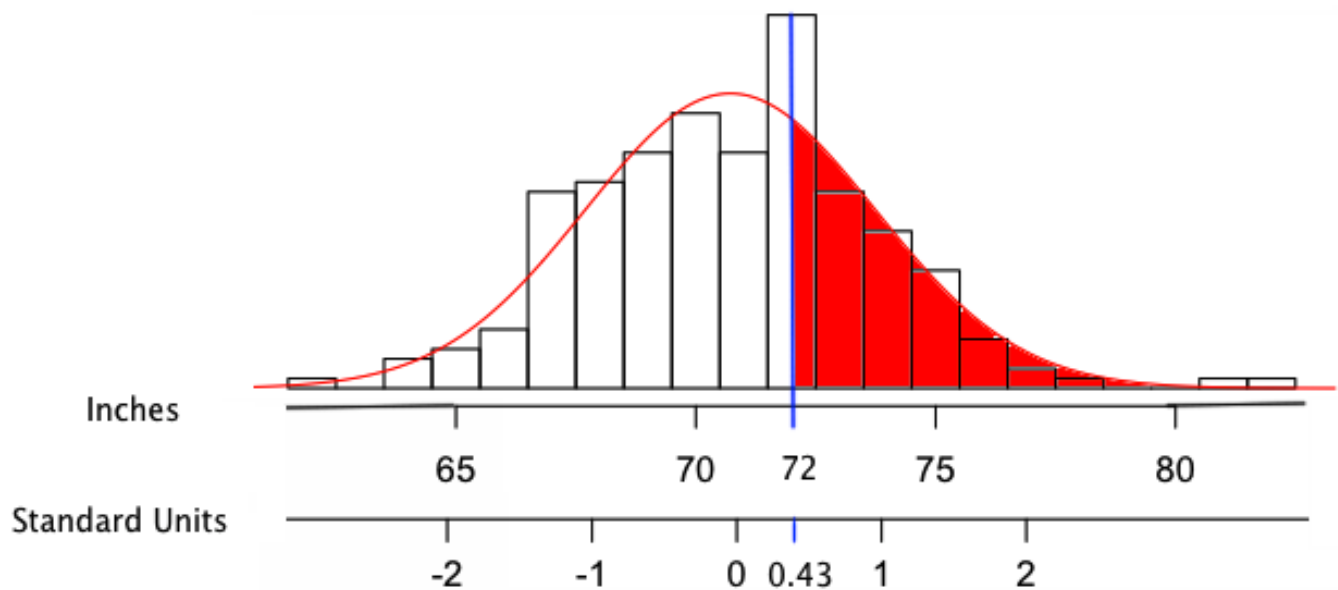
Standard Units	Data	Inches
0	Average	70.72
1	Average + SD	$70.72 + 3.01 = 73.73$
-1	Average - SD	$70.72 - 3.01 = 67.71$
2	Average + 2×SD	$70.72 + 2 \times 3.01 = 76.74$
-2	Average - 2×SD	$70.72 - 2 \times 3.01 = 64.70$

Estimating an area in the histogram

We want to use the normal curve to estimate percentages in the histogram. For example, estimate the percentage of men whose height is over 72 inches.



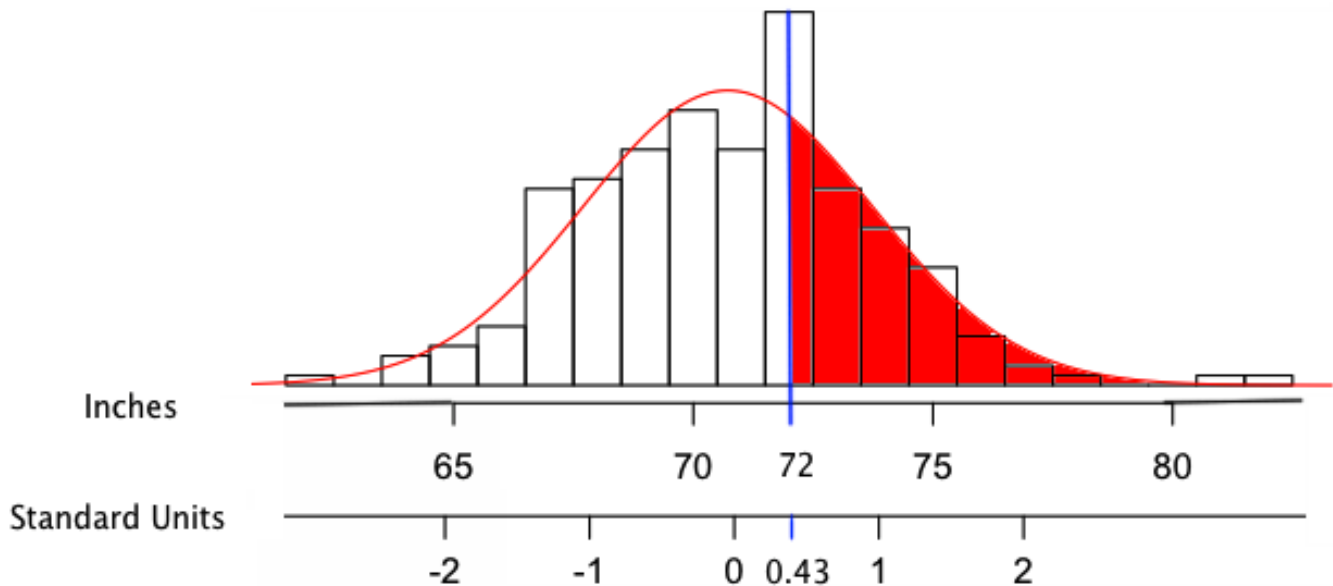
What is the corresponding area for the normal curve?



It is the area under the normal curve above 0.43.

Why does 72 inches correspond to 0.43 standard units?

Translating data into standard units



If we want to find the area in the normal curve that corresponds to heights above 72 inches, we have to translate inches to standard units. Here is the process:

$$\text{Standard units} = \frac{\text{Value} - \text{Average}}{\text{SD}}.$$

The heights have Average = 70.72 and SD = 3.01. So standard units for 72 inches is

$$\text{Standard units} = \frac{\text{Value} - \text{Average}}{\text{SD}} = \frac{72 - 70.72}{3.01} = 0.43.$$

We go to the normal curve table, to find the area above 0.43. (See Example 3 on page 90.)

$$\text{Area above } 0.43 = \frac{1}{2} (100\% - \text{Area between } -0.43 \text{ and } 0.43).$$

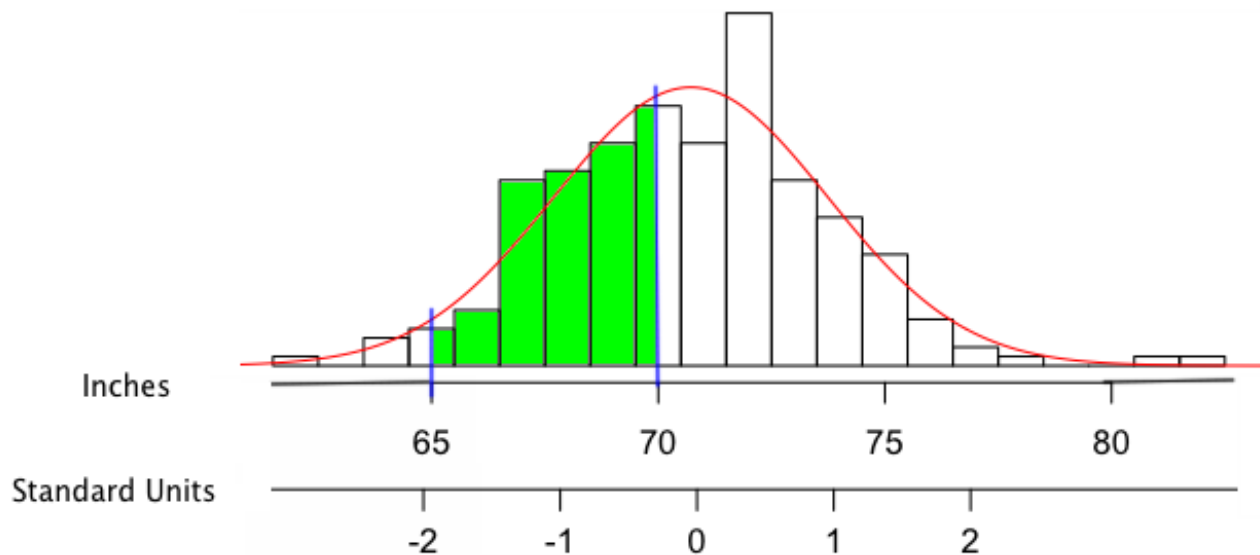
The table does not have 0.43, but the closest is 0.45. Then the area between -0.45 and 0.45 is 34.73%, so that

$$\text{Area above } 0.43 = \frac{1}{2} (100\% - 34.73\%) = 32.64\%.$$

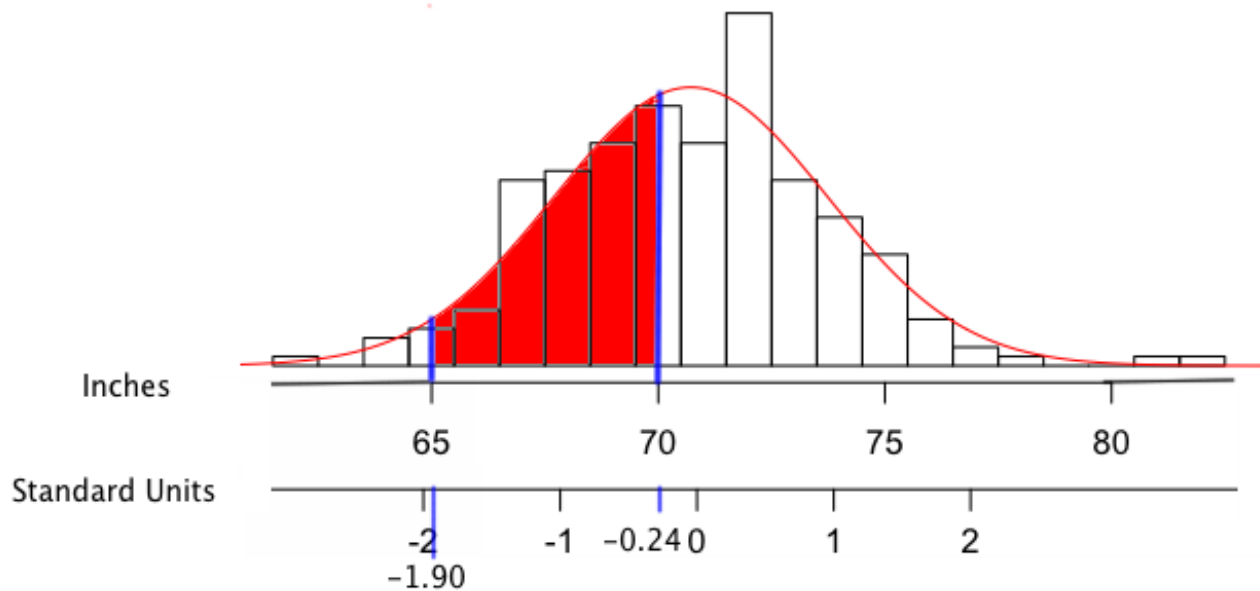
So about 33% of the men are over 72 inches.

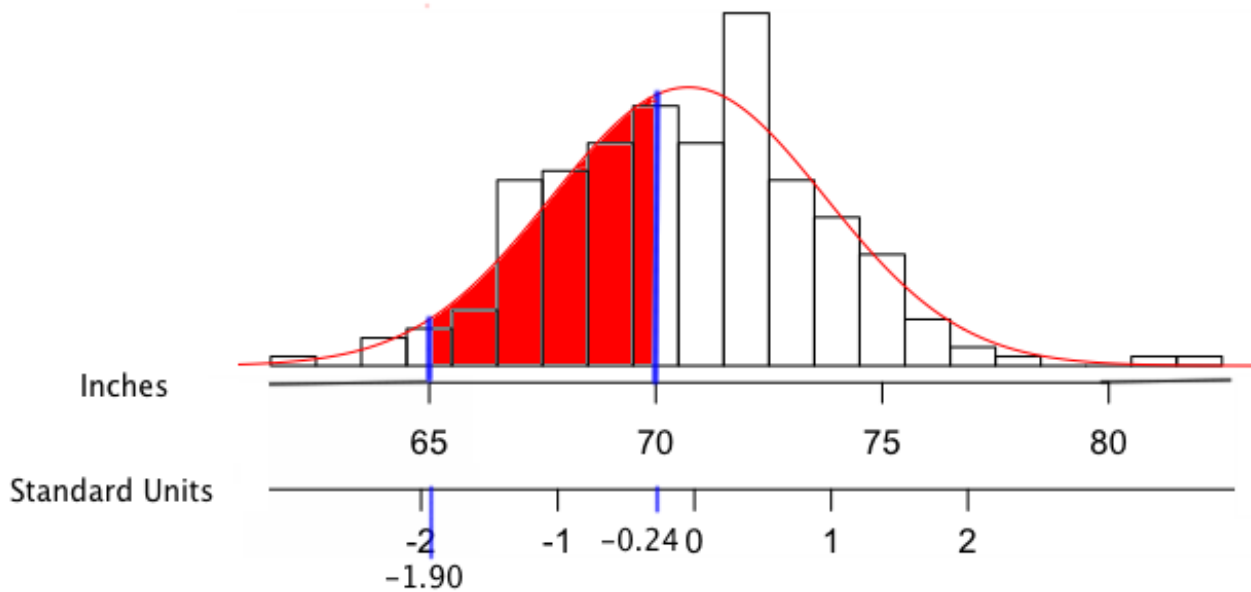
The area between 65 and 70 inches

Consider using the normal curve to approximate the percentage of men with heights between 65 and 70 inches (so between 5 foot 5 and 5 foot 10). The area for the histogram:



The area under the normal curve:





There are two standard units we need, corresponding to 65 inches and 70 inches. The average = 70.72 and the SD = 3.01, as before, so

$$65 \text{ inches} \rightarrow \frac{65 - \text{Average}}{\text{SD}} = \frac{65 - 70.72}{3.01} = -1.90 \text{ (standard units),}$$

$$70 \text{ inches} \rightarrow \frac{70 - \text{Average}}{\text{SD}} = \frac{70 - 70.72}{3.01} = -0.24 \text{ (standard units)}$$

We need the area between -1.90 and -0.24 under the normal curve. (See Example 5 on page 92, where there the numbers were both positive.) From the table:

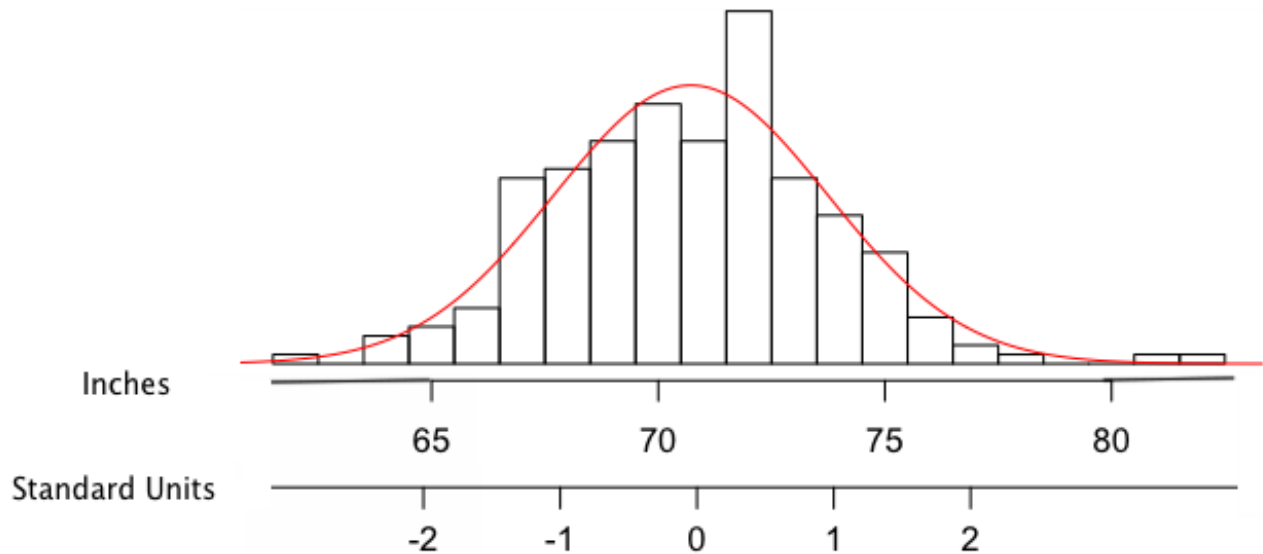
- Area between -1.90 and 1.90 is 94.26%.
- Area between -0.24 and 0.24 is about the area between -0.25 and 0.25, which is 19.74%.
- Then

$$\text{Area between -1.90 and -0.24} \approx \frac{1}{2} (94.26\% - 19.74\%) = 37.26\%.$$

So the answer is 37.26%, or about 37% of the men are between 65 and 70 inches.

The area between 70 and 75 inches

? Estimate the percentage of men between 70 and 75 inches tall. First, indicate on the graph the area under the normal curve you want to find.



Next, find 70 and 75 in standard units. One should be negative, the other positive.

? (Continuing)

Find the area under the normal curve between 0 and the negative value in standard units.

Find the area under the normal curve between 0 and the positive value in standard units.

Finally, to estimate the area between 70 & 75, you either add or subtract those two areas you just figured. Which is it? What is the answer?

4.1 Scatter plots

Here again is the data table on the age at inauguration and age at death of the US presidents.

	Inauguration	Death
Washington	57	67
J_Adams	61	90
Jefferson	57	83
Madison	57	85
Monroe	58	73
J_Q_Adams	57	80
Jackson	57	78
Van_Buren	54	79
W_H_Harrison	68	68
Tyler	51	71
Polk	49	53
Taylor	64	65
Fillmore	50	74
Pierce	48	64
Buchanan	65	77
Lincoln	52	56
A_Johnson	56	66
Grant	46	63
Hayes	54	70

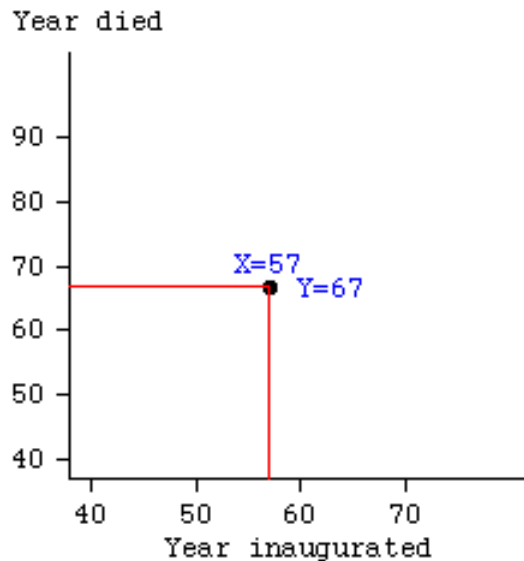
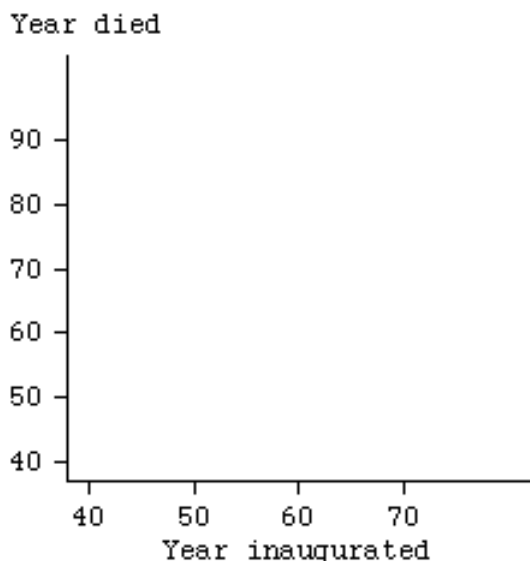
	Inauguration	Death
Garfield	49	49
Arthur	50	57
Cleveland	47	71
B_Harrison	55	67
McKinley	54	58
T_Roosevelt	42	60
Taft	51	72
Wilson	56	67
Harding	55	57
Coolidge	51	60
Hoover	54	90
FDR	51	63
Truman	60	88
Eisenhower	62	78
Kennedy	43	46
L_Johnson	55	64
Nixon	56	81
Ford	61	93
Reagan	69	93

We have seen a histogram of the ages at death. How do we picture the two variables together? **A scatter plot.**

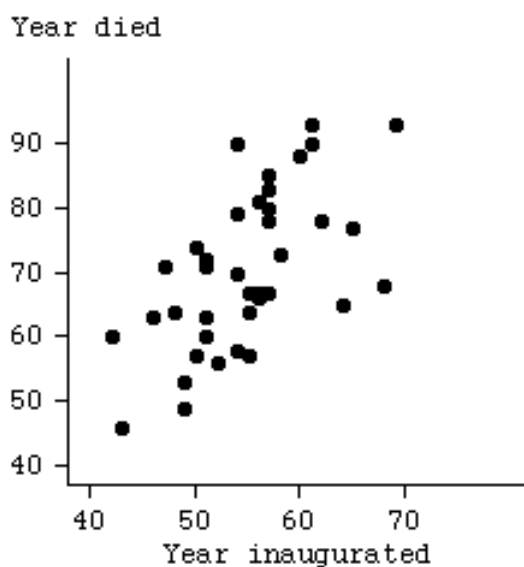
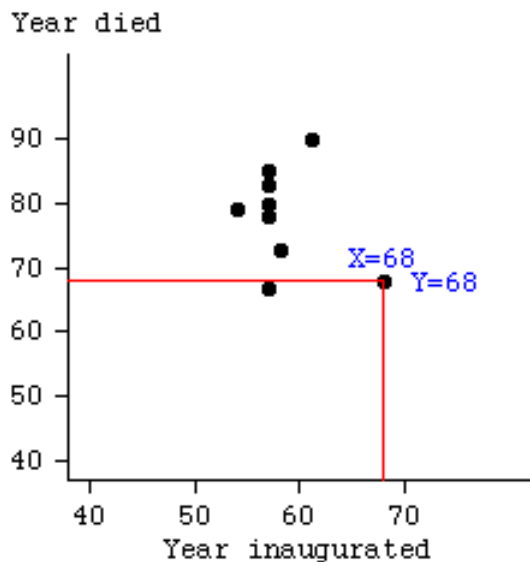
The two variables are X and Y:

X = Age at inauguration, Y = Age at death.

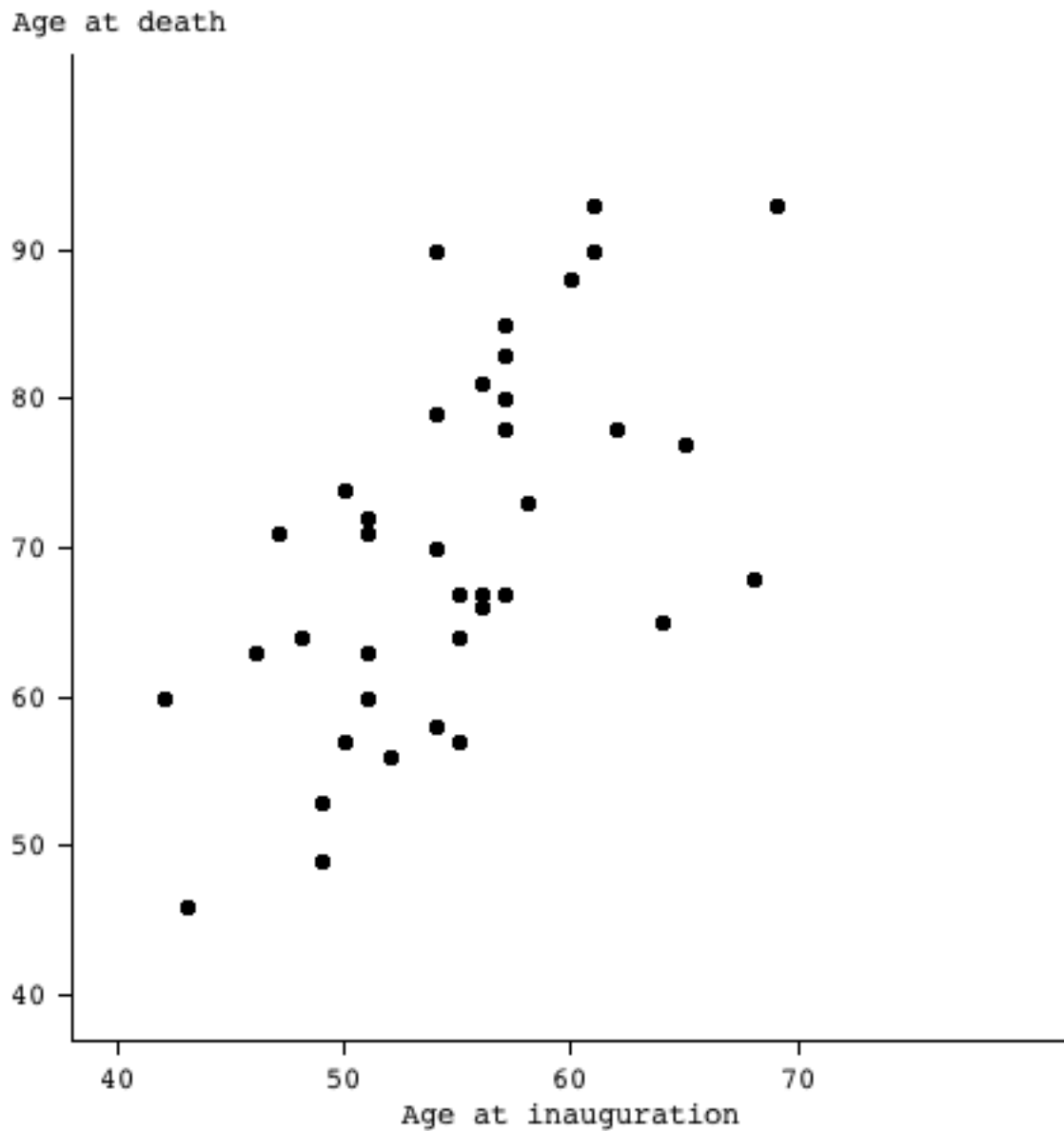
X is represented along the horizontal axis, and Y on the vertical axis, as in the first plot. The first point (George Washington) is $X=57$, $Y=67$. (So he was inaugurated at age 57, and died at 67.) It is shown on the second plot.



Continue, one dot for each president. E.g., the ninth is $X=68$, $Y=68$: William Henry Harrison. He died soon after the inauguration. The last plot has all 38 presidents.



Here is the complete scatter plot again:



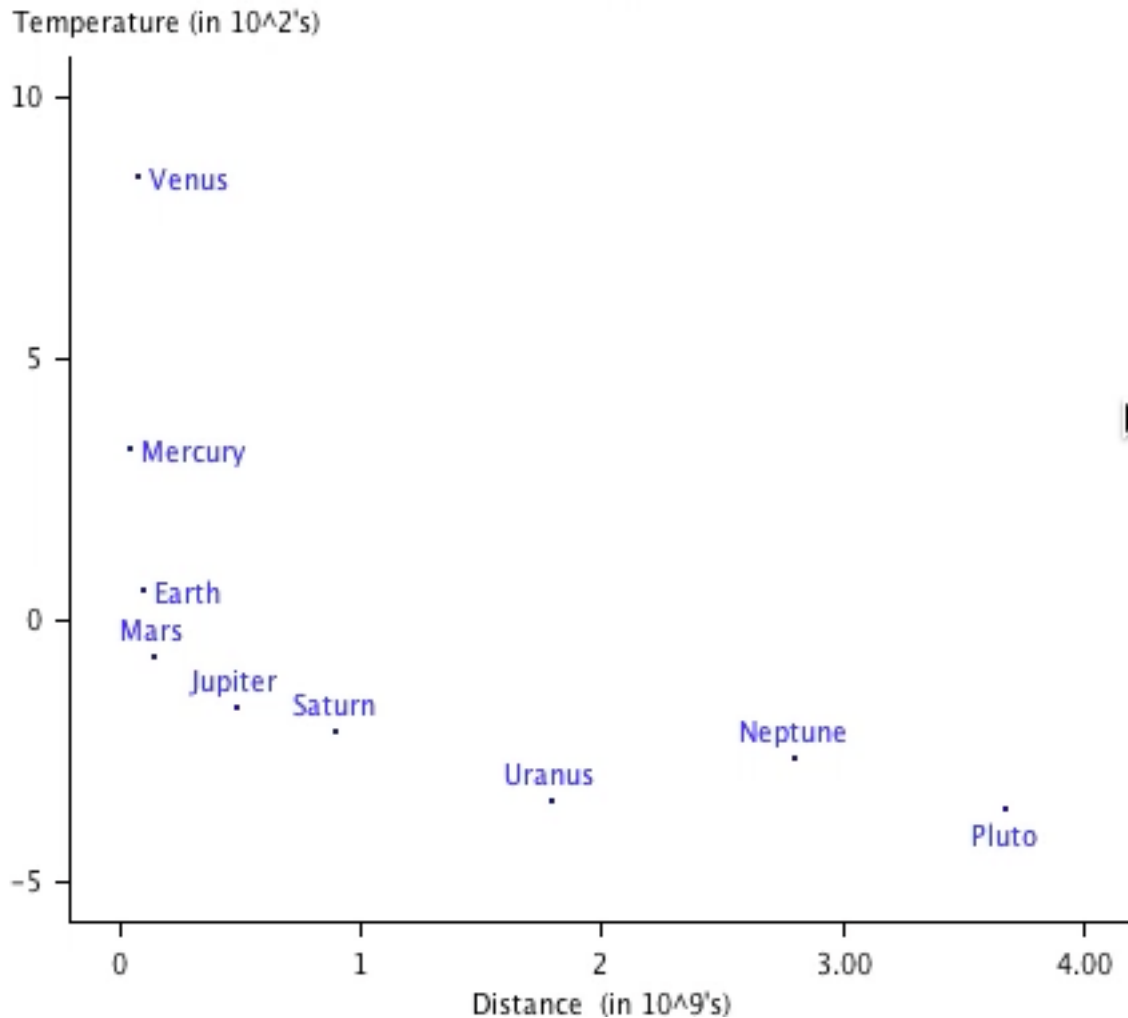
? To check you are reading it correctly, circle the dots corresponding to the following presidents:

	Inauguration	Death
Madison	57	85
Jackson	57	78
Lincoln	52	56

	Inauguration	Death
Garfield	49	49
Hoover	54	90
Reagan	69	93

Next is a scatter plot of the planets (plus Pluto):

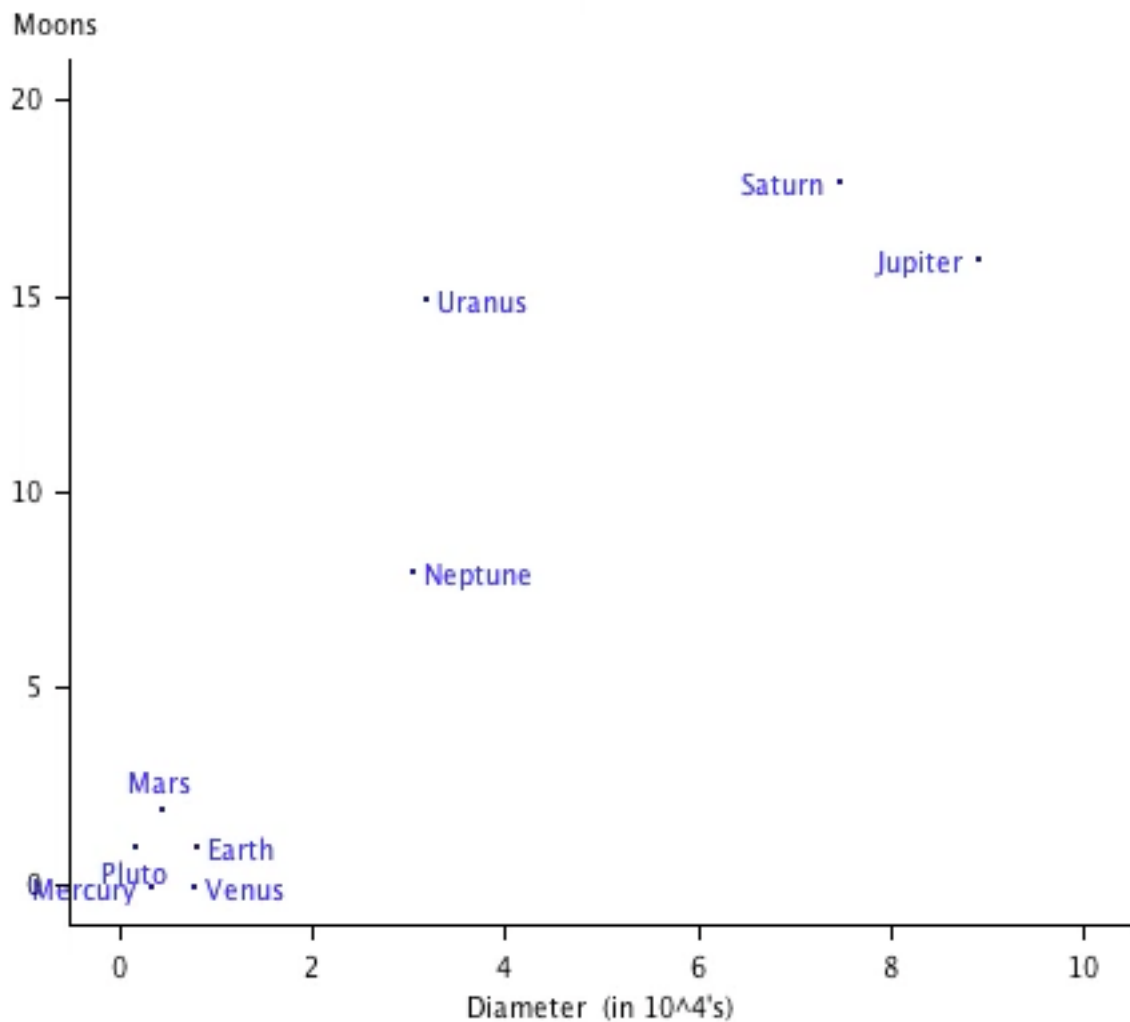
X = Distance from the Sun, Y = Temperature.



? What is the general trend? That is, does temperature tend to go up or go down as the distance from the sun increases?

Are the points in a straight line, or more of a curve?

This plot has $X = \text{Diameter}$, $Y = \text{Number of moons}$.



? Do the larger planets tend to have more or fewer moons than the smaller planets?

Does this plot show clusters of planets? How so?

4.2 Correlation

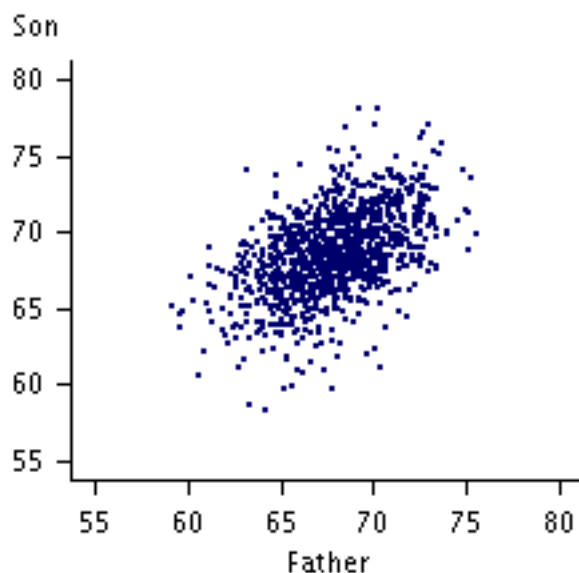
Many scatter plots show an association between the two variables. In the president's data, the older the president was inaugurated, generally the older they died. (Naturally, they couldn't get inaugurated after they were dead.)

The correlation coefficient is a number that measures the correlation, or association, of two variables in a scatter plot.

- Positive correlation: Higher values of the first variable are associated with higher values of the second, and lower values of the first variable are associated with lower values of the second.
- Negative correlation: Higher values of the first variable are associated with lower values of the second, and lower values of the first variable are associated with higher values of the second.

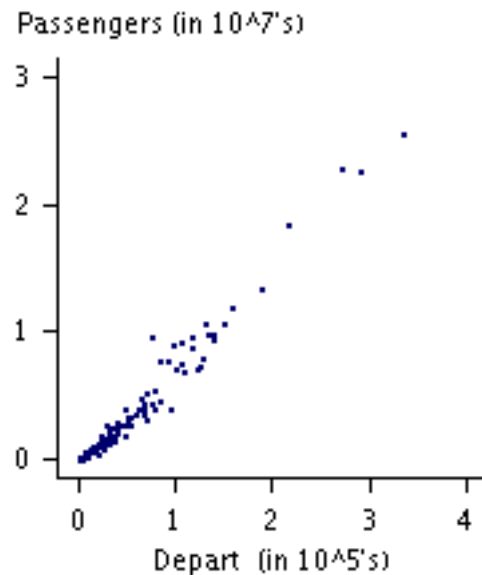
Heights of Fathers and Sons

The next plot has a positive correlation. Taller fathers tend to have taller sons. But there is a lot of variability. The points tend to go from the lower left to the upper right.

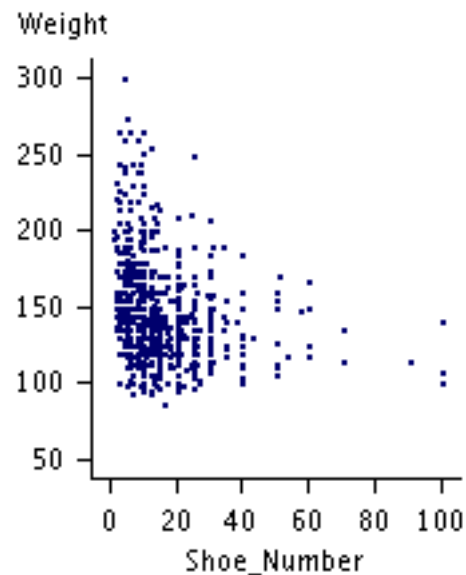


Departures and passengers

For each of 135 airports in the US, the $X = \#$ departing flights, and the $Y =$ total $\#$ of passengers. This plot has a positive correlation. The more flights, the more passengers. The points tend to go from the lower left to the upper right. The pattern is fairly tight, that is, close to a straight line.

**# of pairs of shoes and weight**

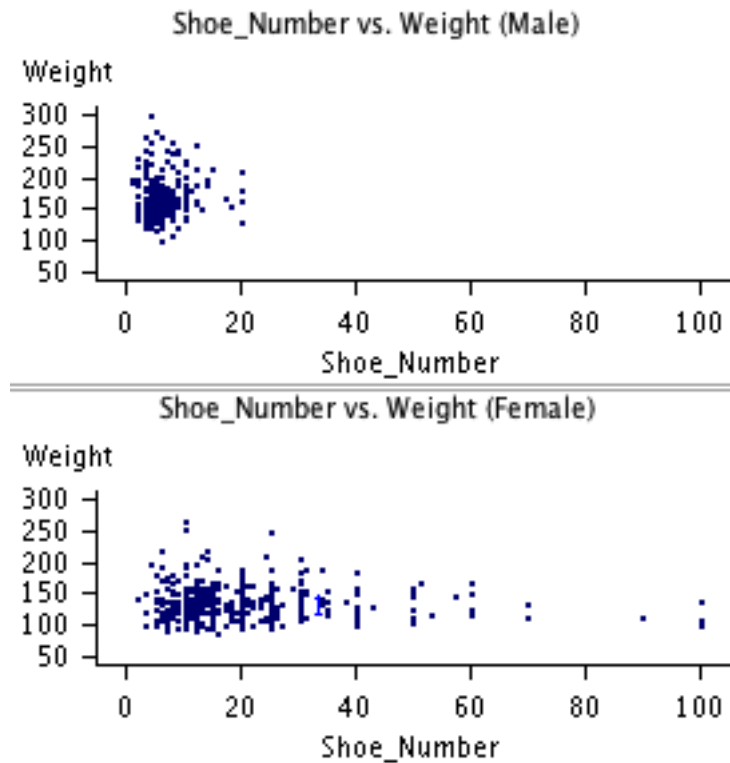
This plot has a negative correlation. The more shoes, the less people weigh. The points tend to go from the upper left to the lower right, but not in a straight line. There's a huge amount of variability.



? Why would people with more pairs of shoes tend to weigh less? Could there be a third factor explaining the association?

of pairs of shoes and weight, split by gender

The top scatter plot below has just the men, and the bottom has just the women.



These plots don't seem to have much correlation one way or the other. So even though there is an association between number of pairs of shoes and weight when looking at the entire class, when splitting up into men and women, neither plot shows much of a trend (although a lot of variation).

That is, gender is a third factor, explaining the negative correlation between weight and # of pairs of shoes.

? In what way does gender explain the correlation?

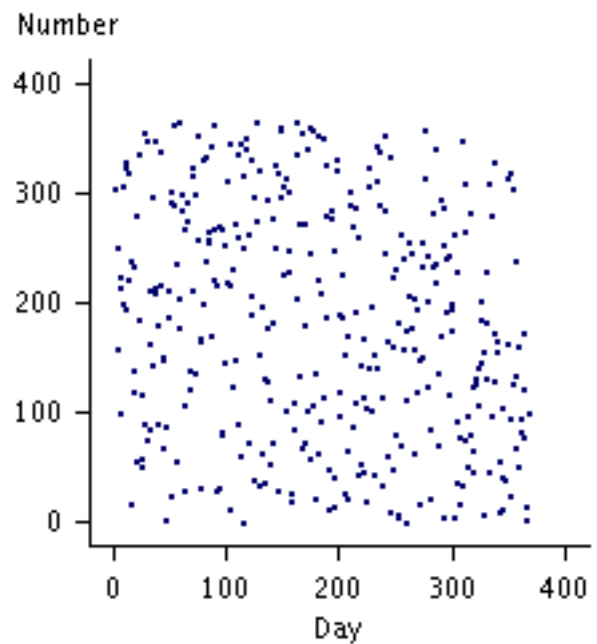
Draft lottery

In the 1969 draft lottery, young men were randomly assigned numbers from 1 to 366, by birthday. People were then drafted in order of their lottery number, #1 first, #2 second, etc. The plot has

X = Birthday (day of the year, from 1 to 366)

Y = Lottery number

Is there a pattern? It just looks like a big random square.



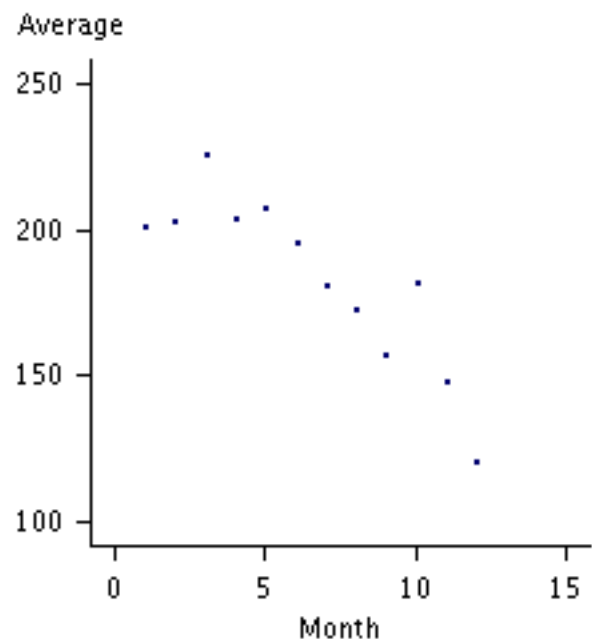
Averaged by month

This plot is based on the same data, but the data points are averaged by month:

X = Month (1 to 12)

Y = Average lottery number for that month

This has a fairly clear negative correlation. The later the month, the lower the lottery number.

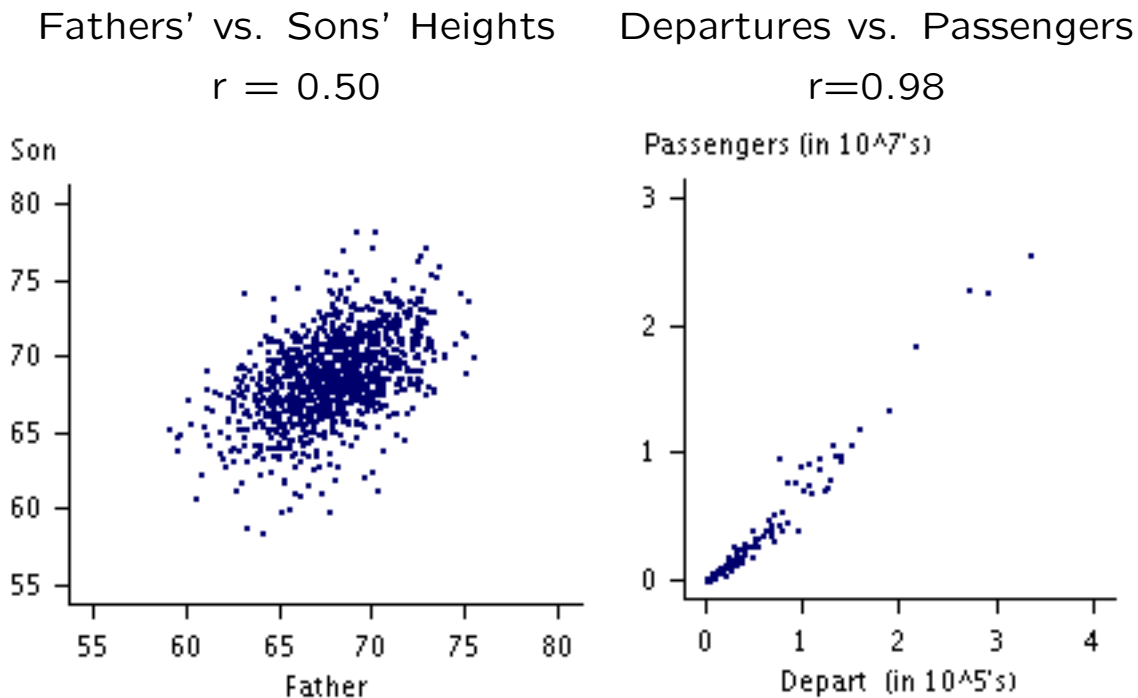


4.3 The correlation coefficient

The correlation coefficient is a number, r , between -1 and 1 , that measures correlation.

- Positive correlation: $r > 0$
- Negative correlation: $r < 0$
- The closer r is to $+1$ or -1 , the closer the points are to a straight line
- $r \approx 0$ means there is not much correlation

What are the correlation coefficients for the scatter plots we just saw?

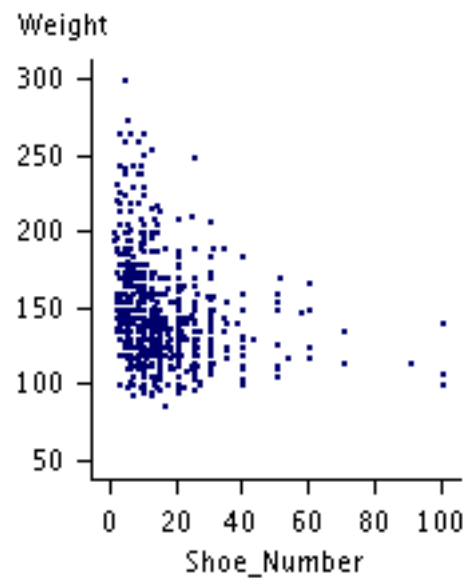


Both these plots have a definite positive correlation. But the departures vs. passengers is much less scattered, very close to a straight line, so its correlation coefficient is nearly 1. The plot with fathers' and sons' heights is quite scattered, and its correlation coefficient turns out to be 0.50, which is not too small but not near 1, either.

By contrast, the plot with number of shoes and weight shows a negative correlation. The points are not very close to a straight line. It's correlation coefficient is

$$r = -0.29,$$

negative, but not particularly close to -1.



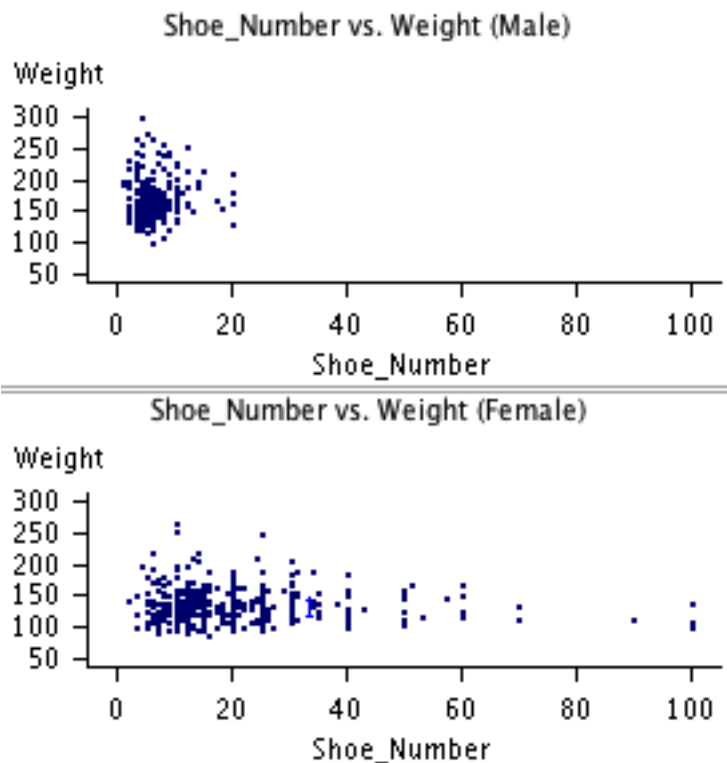
If we split the shoes vs. weight data by gender, we do not see much trend in either plot, and the points are quite scattered. The correlation coefficients are

For men, $r = 0.05$

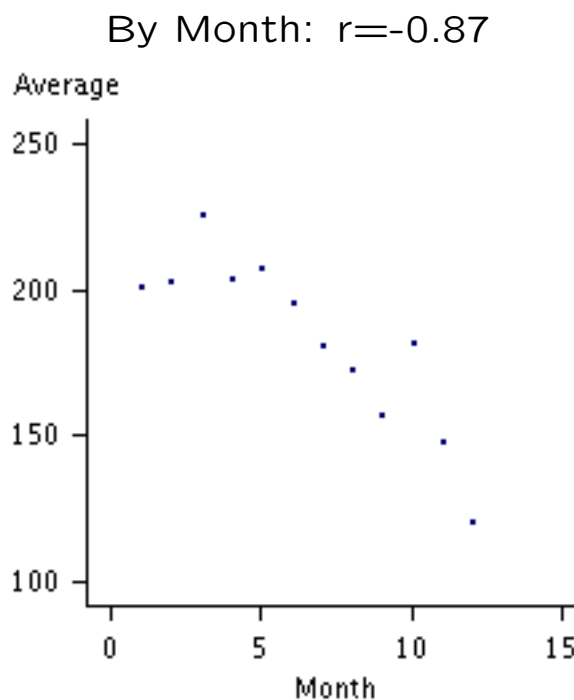
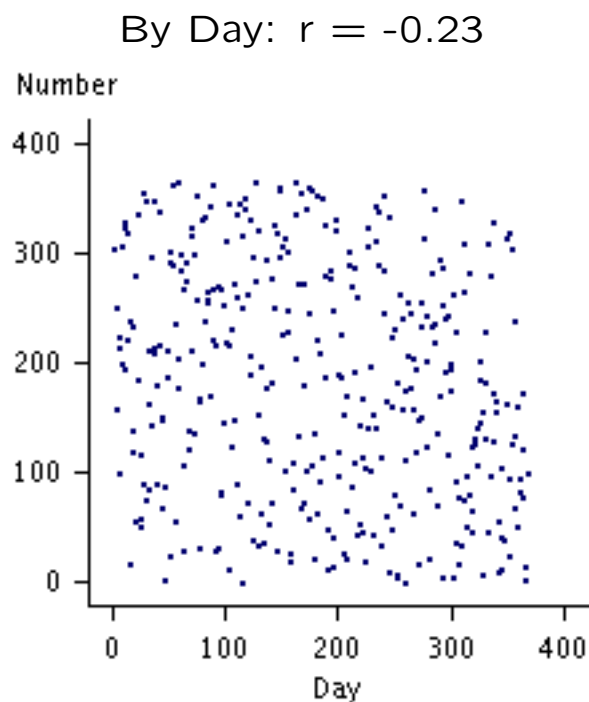
For women, $r = -0.06$

These r 's are practically 0.

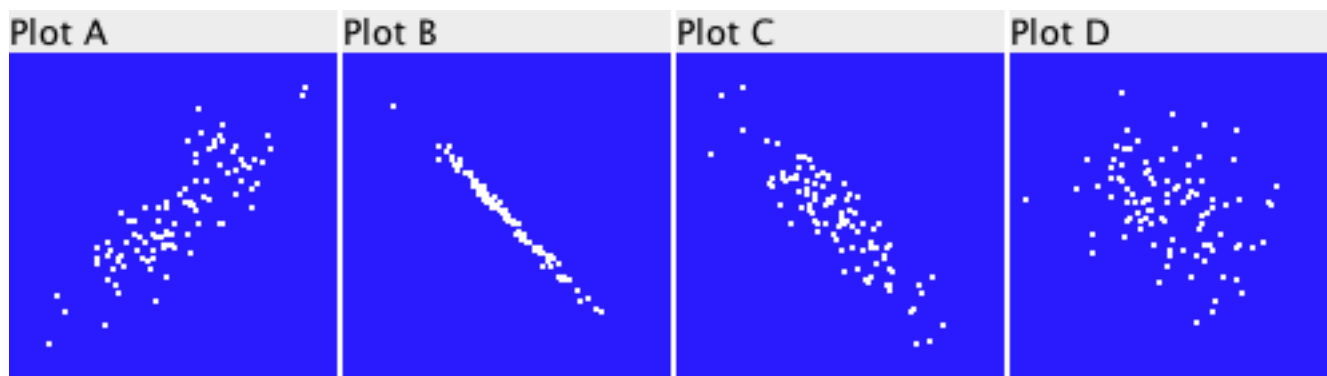
Thus among men, number of shoes and weight are unrelated, and among women, number of shoes and weight are unrelated.



The draft lottery correlations are both negative. It is easy to see in the plot by month, but not obvious for the plot by day. The points in the plot by month are much closer to a straight line, so have a correlation coefficient close to -1.



? Here are four scatter plots:



Guess their correlation coefficients, choosing among the values

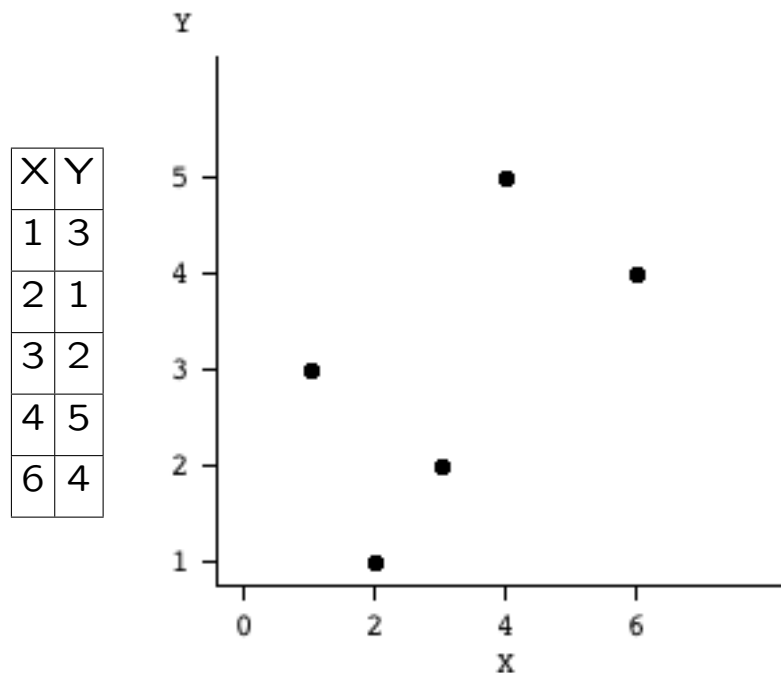
-0.99 , -0.89 , -0.33 , $+0.85$

Calculating the correlation coefficient

There are three main steps to calculating the r :

1. For each data point, find the X value in standard units, and find the Y value in standard units.
2. For each point, find the product of their two standardized values.
3. Find the average of those products.

A small example, with $n = 5$ observations, and variables X and Y :



Step 1. Find the standard units

Before we do anything else, we have to find the average and SD of the X's, and the average and SD of the Y's. Start with the X's.

$$\text{Average}_x = \frac{1 + 2 + 3 + 4 + 6}{5} = 3.2.$$

For the SD, we first find the deviations: -2.2, -1.2, -0.2, 0.8, 2.8

Square the deviations: 4.84, 1.44, 0.04, 0.64, 7.84

$$\begin{aligned} \text{Average of squared deviations} &= \frac{4.84 + 1.44 + 0.04 + 0.64 + 7.84}{5} \\ &= 2.96. \end{aligned}$$

Square root that average: $\text{SD}_x = \sqrt{2.96} = 1.72$.

Then find the standard units:

$$\text{Standard Units} = \frac{\text{Value} - \text{Average}_x}{\text{SD}_x}$$

$$\text{So for } X = 1: \quad \text{Standard Units} = \frac{1 - 3.2}{1.72} = \frac{-2.2}{1.72} = -1.28.$$

$$\text{And for } X = 2: \quad \text{Standard Units} = \frac{2 - 3.2}{1.72} = \frac{-1.2}{1.72} = -0.70.$$

$$\dots \text{ last for } X = 6: \quad \text{Standard Units} = \frac{6 - 3.2}{1.72} = \frac{2.8}{1.72} = 1.63.$$

The table has the standard units for the X's:

X	Standard Units for X	Y
1	-1.28	3
2	-0.70	1
3	-0.12	2
4	0.47	5
6	1.63	4

Next, find the average and SD of the Y's:

$$\text{Average}_Y = \frac{3 + 1 + 2 + 5 + 4}{5} = 3.$$

The deviations are then 0, -2, -1, 2, 1

? Show that the $\text{SD}_Y = 1.41$.

? Find the standard units for the Y's.

Step 1 is to find the standard units for the X's and Y's, which we just did. Here they are:

X	Standard Units for X	Y	Standard Units for Y
1	-1.28	3	0
2	-0.70	1	-1.42
3	-0.12	2	-0.71
4	0.47	5	1.42
6	1.63	4	0.71

Step 2: Find the products of the pairs of standard units.

Now for each observation (row in the table), we multiply the standard units for X times the standard units for Y. So for the first two observations:

$$X = 1, Y = 3 \rightarrow (\text{Standard units for X}) \times (\text{standard Units for Y}) \\ = (-1.28) \times (0) = 0,$$

$$X = 2, Y = 1 \rightarrow (\text{Standard units for X}) \times (\text{standard Units for Y}) \\ = (-0.70) \times (-1.42) = 0.9940.$$

The results for the first four points are in the last column:

X	Standard Units for X	Y	Standard Units for Y	Product of Standard Units
1	-1.28	3	0	0
2	-0.70	1	-1.42	0.9940
3	-0.12	2	-0.71	0.0852
4	0.47	5	1.42	0.6674
6	1.63	4	0.71	

? Find the product for the last observation.

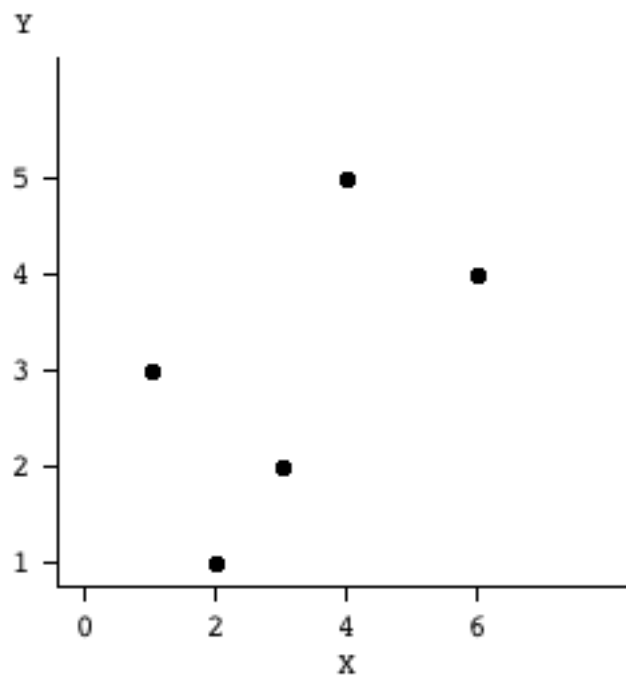
Step 3: Average the products.

X	Standard Units for X	Y	Standard Units for Y	Product of Standard Units
1	-1.28	3	0	0
2	-0.70	1	-1.42	0.9940
3	-0.12	2	-0.71	0.0852
4	0.47	5	1.42	0.6674
6	1.63	4	0.71	1.1573

Now we find the average of those products, to calculate the correlation coefficient r :

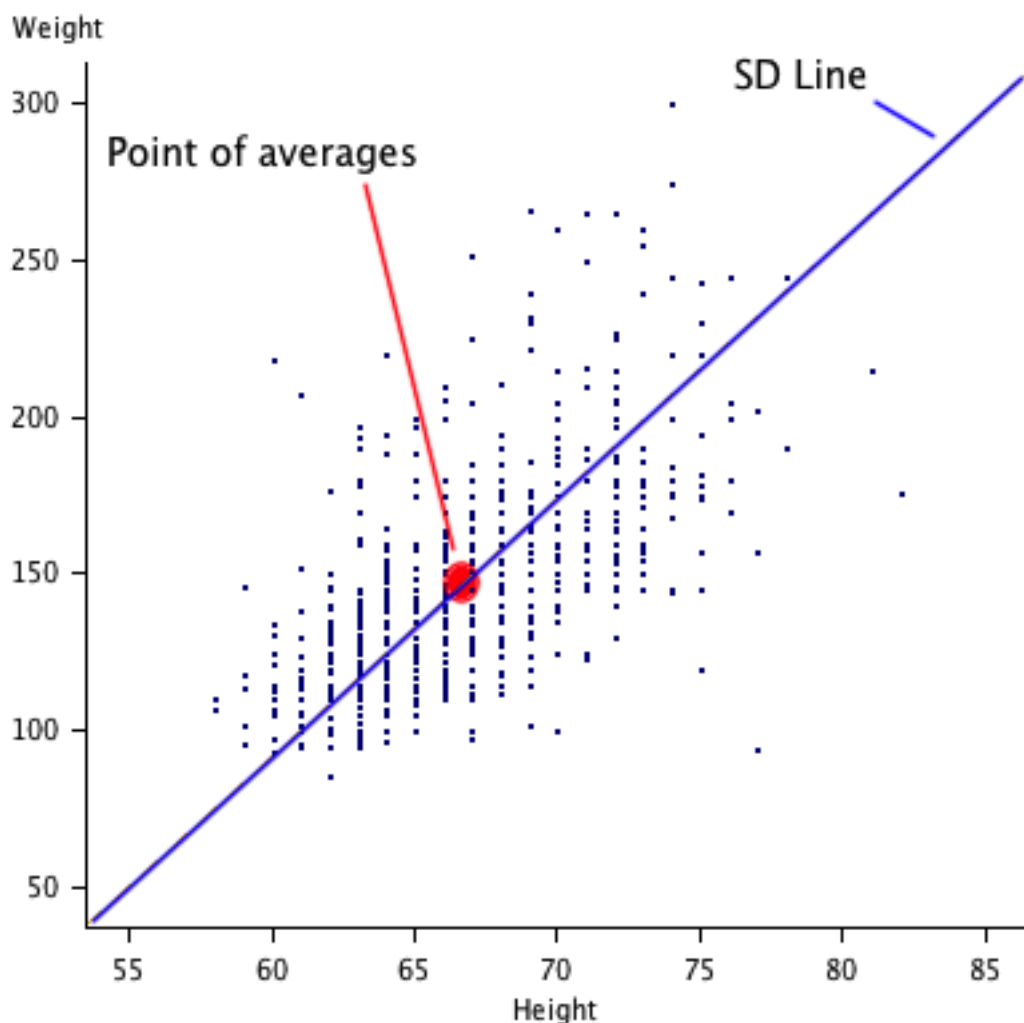
$$\begin{aligned}
 r &= \frac{0 + 0.9940 + 0.0852 + 0.6674 + 1.1573}{5} \\
 &= \frac{2.9039}{5} \\
 &= 0.58.
 \end{aligned}$$

So the correlation coefficient is $r = 0.58$. It is positive, but medium-sized. That should be reasonable, given the scatter plot:



The point of averages and SD line

The correlation coefficient r measures which direction the points go, as well as how close they are to a straight line. Which line? First, the line goes through the **point of averages**, which is where the average of X meets up with the average of Y .

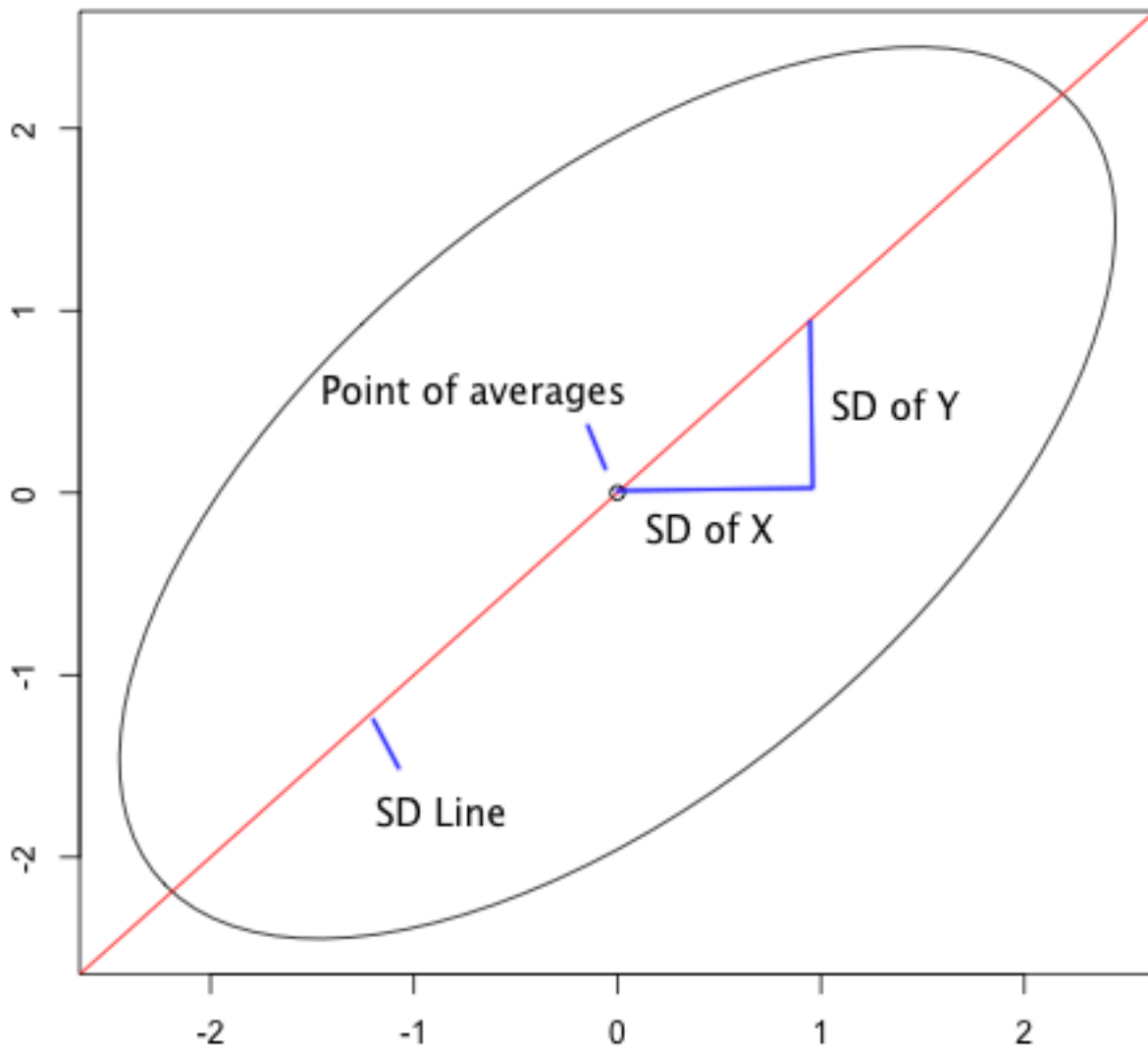


This plot has the heights vs. weights of students in a class. Here, the average of $X = 67$ inches and the average of $Y = 146$ pounds.

If you think of the cloud of points as an ellipse, or *football-shaped*, as the book puts it, the SD line goes through the ends of the football, also passing through the point of averages.

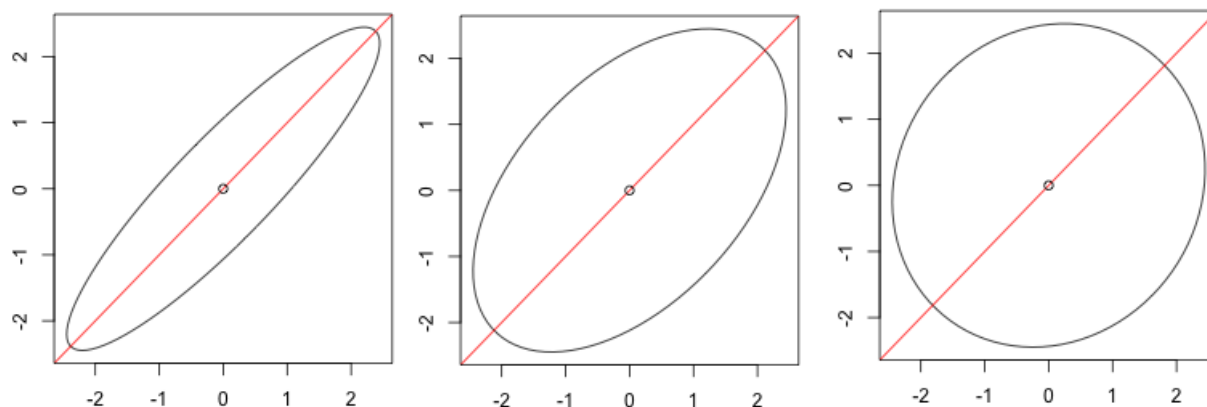
Idealized plots

Sometimes it is useful to sketch the scatter plot as an ellipse — Imagine the points scattered within the ellipse. The point of averages is in the center, and the SD line goes through the ends of the ellipse, or the “ends of the football.”



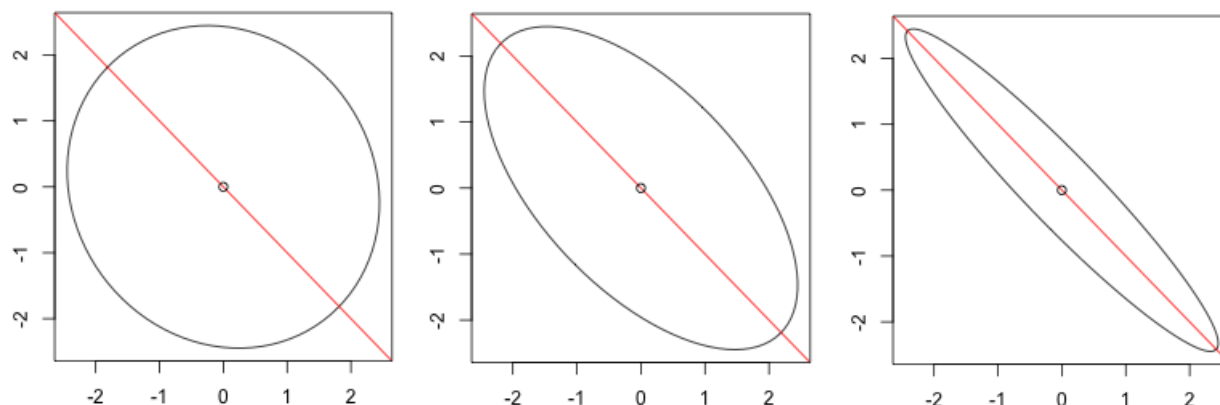
The SD line starts at the point of averages, then goes one SD of X to the right, and one SD of Y up.

If the correlation is positive, the SD line has positive slope:



In the third plot you can hardly tell that the correlation is positive.

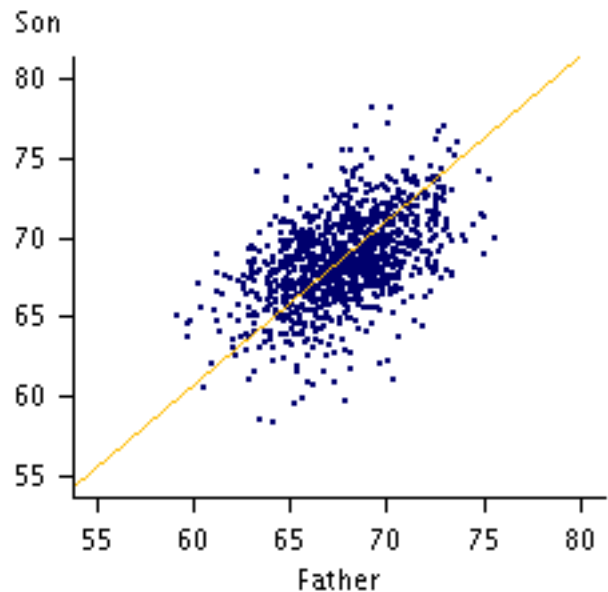
If the correlation is negative, the ellipse is pointing down, so that the SD line has negative slope:



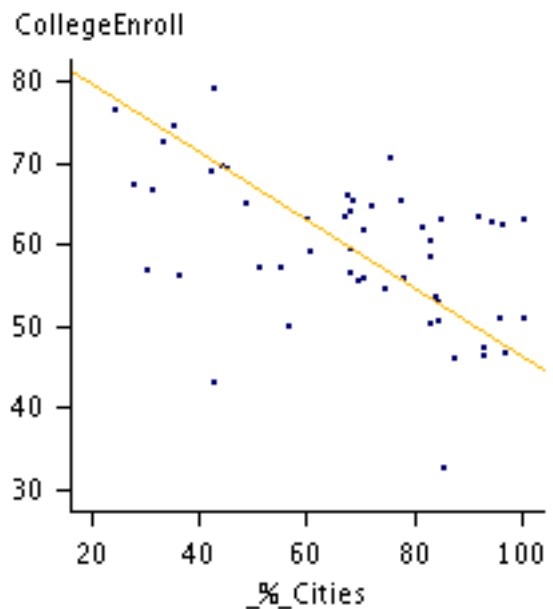
If the ellipse is actually a perfect circle, we cannot find the ends of the football (it is a soccer ball), so we do not know which way the SD line goes.

More examples of the SD line

The heights of fathers and sons has a reasonably football-shaped plot. The points of average here is 68 inches, for the fathers and 69 inches, for the sons.

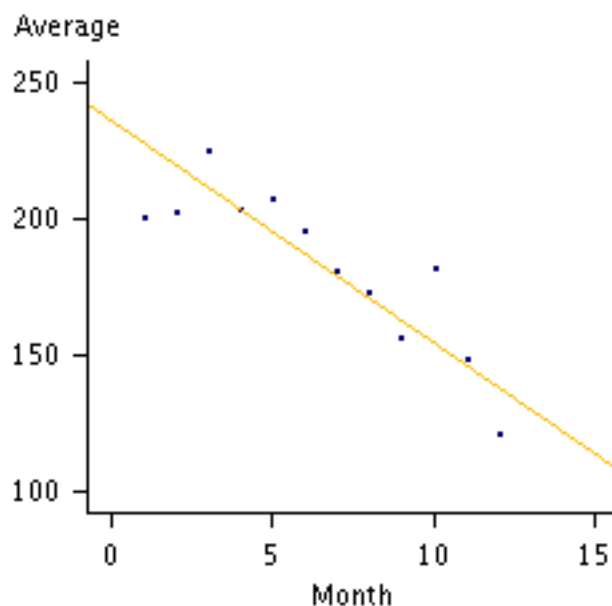
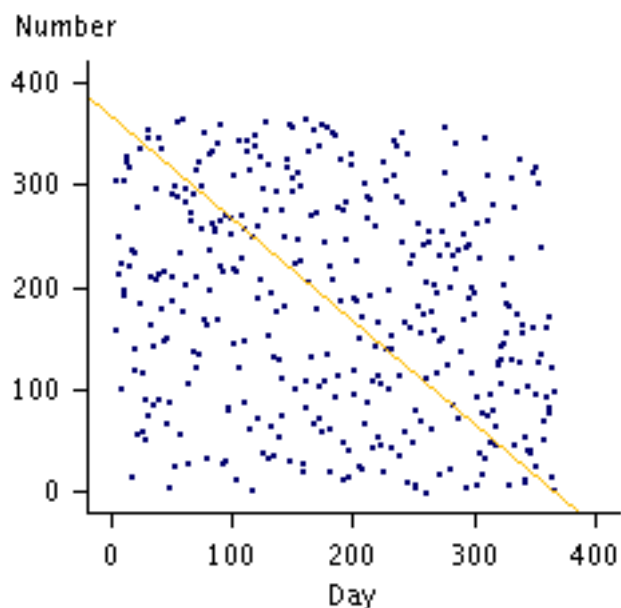


This plot has a negative correlation. The data are on the states, where X = the percentage of the state that is urbanized, and Y is the percentage of college-aged people enrolled full time:

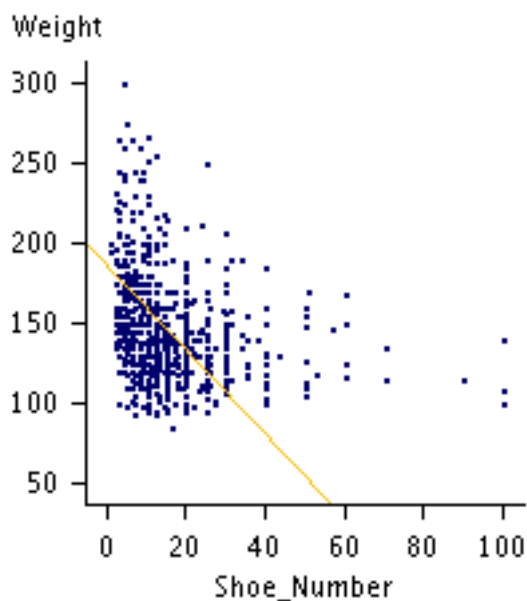


The draft lottery by day data is more a square than a football. It has a negative correlation, so the SD line goes through the upper left and lower right corners.

The draft lottery data averaged by month has a much tighter scatter plot. The SD line follows the data points quite closely.



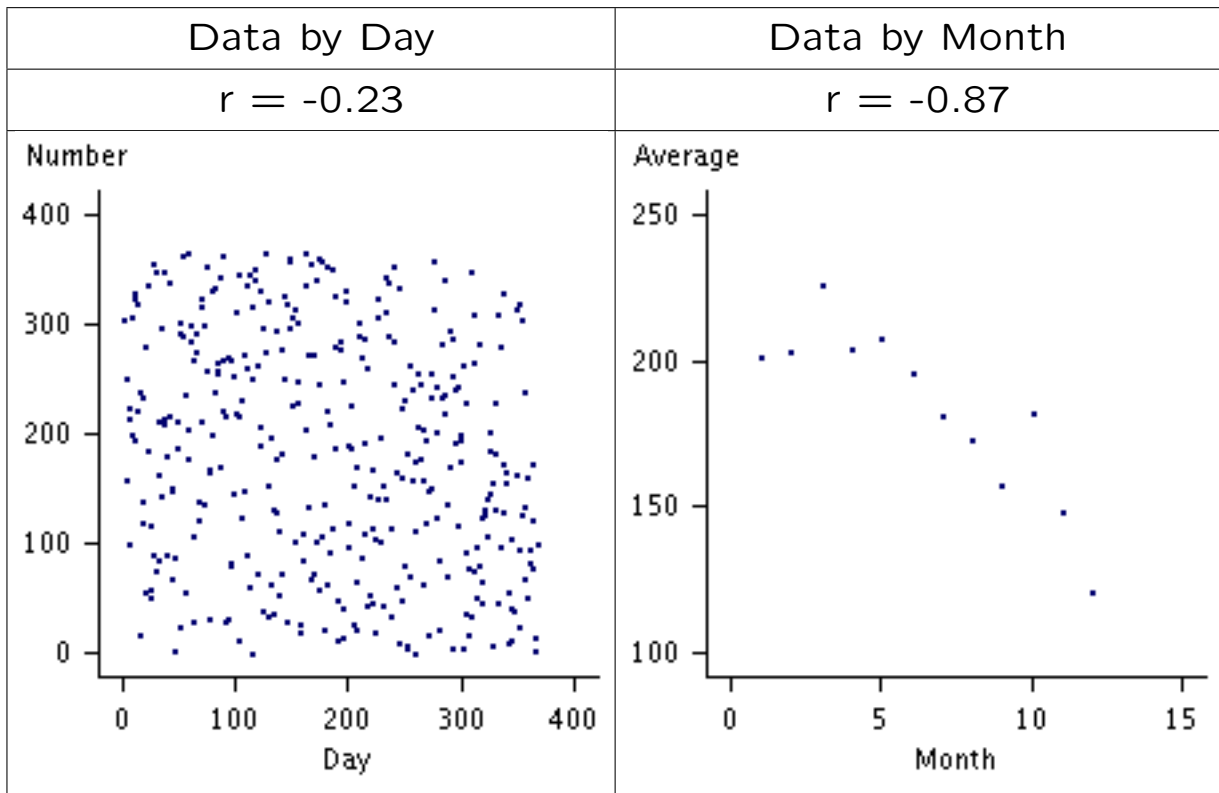
If the scatter plot is far from football-shaped, the SD line might not be very meaningful, as in the data with $X = \#$ of pairs of shoes, and $Y = \text{weight}$.



4.4 Ecological correlations

Correlations based on averaging over groups of data

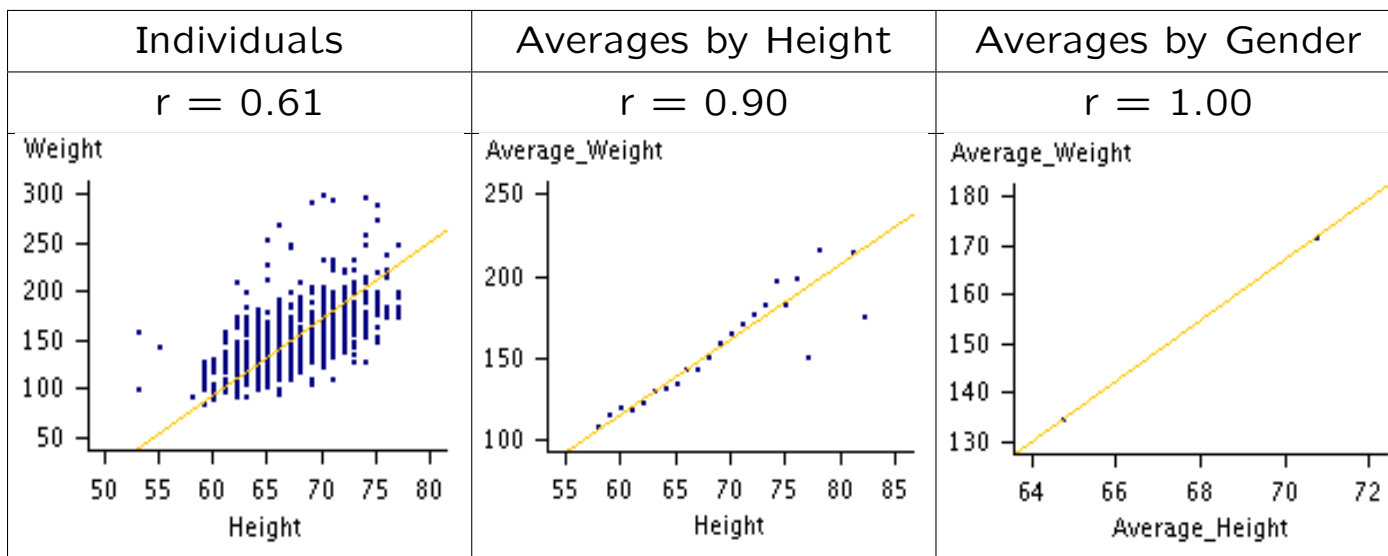
For the draft lottery data, we found two scatter plots. One is based on the data by day, and the other by finding the average lottery number for each month:



Even though these two plots are based on the same data, one shows a fairly weak correlation of -0.23 , while one has a very strong correlation of -0.87 . Why is there a difference? And which one is correct?

Height vs. Weight

Here are three scatter plots for the same data:



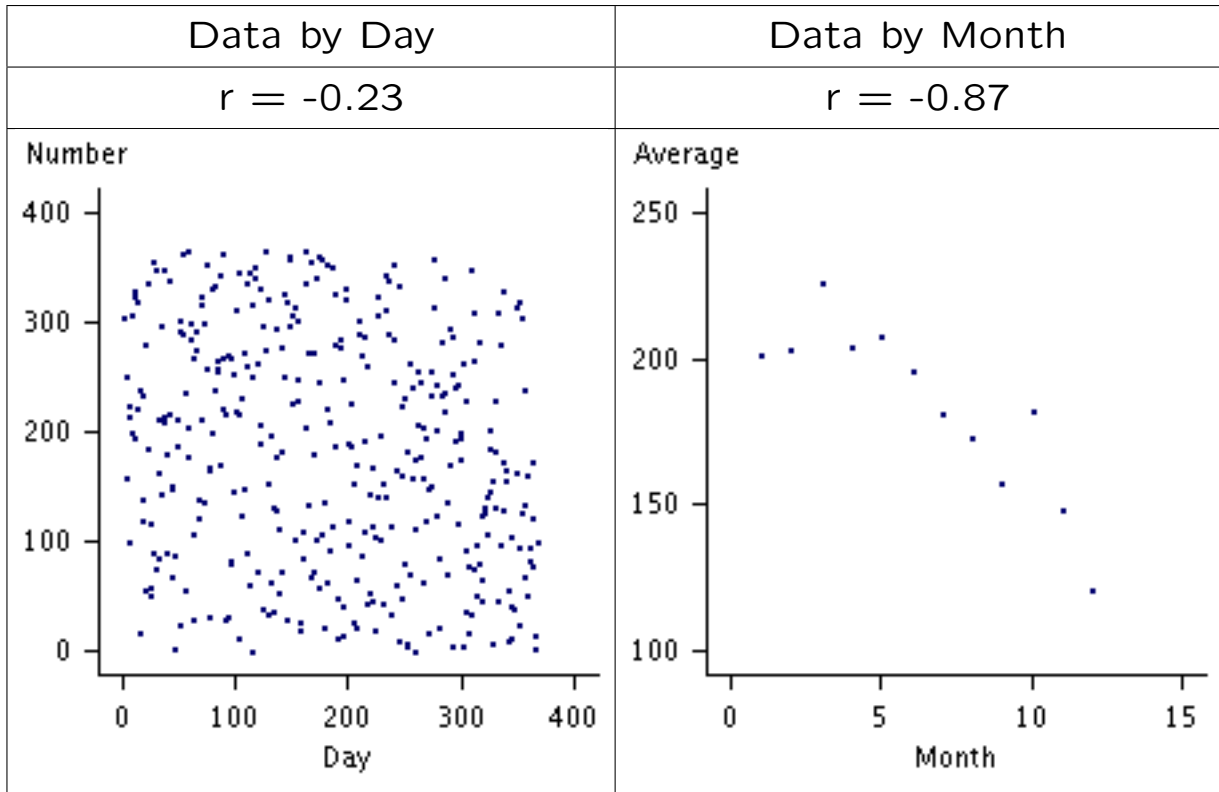
The first plot shows the individual heights and weights. The second shows the average weight for each height. The third shows the two genders, with their average height and weight. Again, even though all three plots are from the same data, they have very different correlation coefficients.

What's happened? In these two examples, the correlation became stronger when we took averages. This often occurs because averaging reduces the scatter, so there are fewer points far from the line. When a correlation is calculated based on grouping of data, it is called an **ecological correlation**¹.

Ecological correlations are not bad. The danger in using ecological correlations is to think that they are actually the correlations for individuals. Using the second plot above, one might think the correlation between height and weight is 0.90. It is not! It is 0.61. The ecological correlation often overstates the correlation for the individuals.

¹The word “ecological” in this context does not necessarily mean having to do with biology or the environment. It comes from the notion of “ecological studies” in sociology, which are studies of groups of humans and how they related to social and other variables.

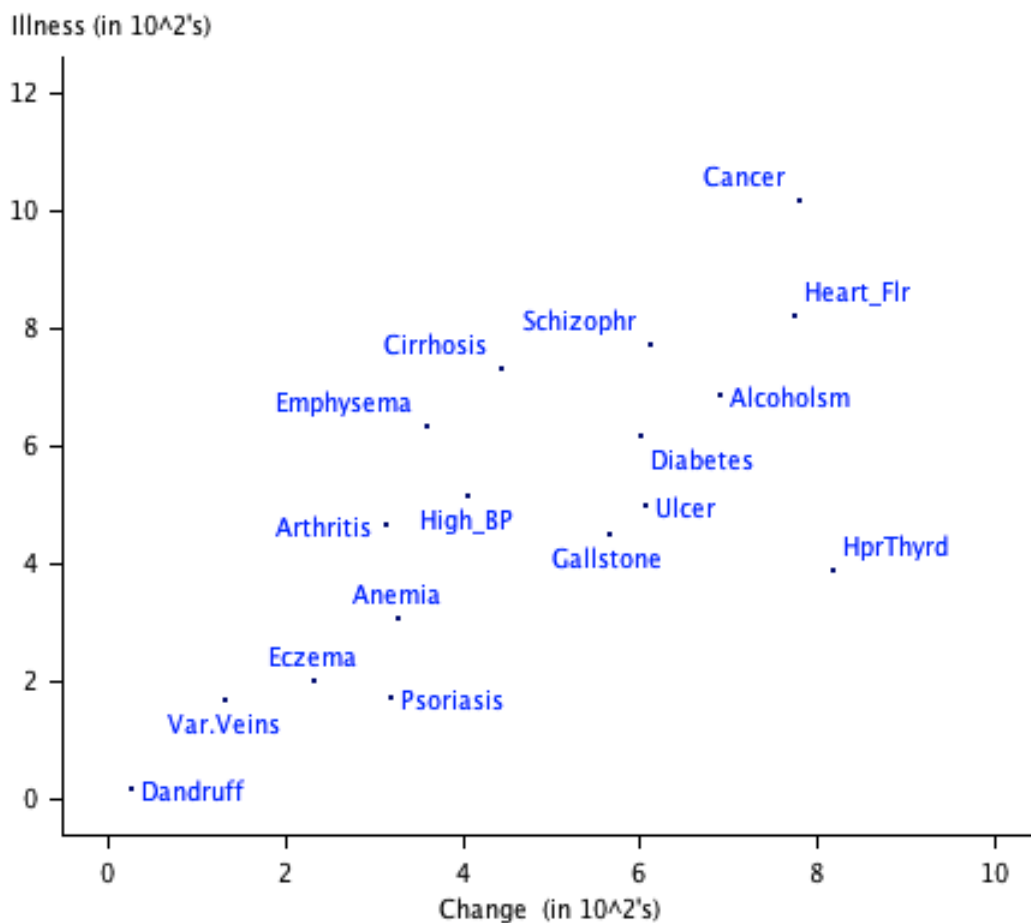
For example, in the **draft lottery**, the data averaged by month shows a very strong downward trend: $r = -0.87$. This r is an ecological correlation. It tells us that there is strong evidence that the lottery was **not** totally random: There is some systematic bias. On the other hand, the correlation for individuals is only $r = -0.23$, which means the impact on individuals is not very strong.



? The lottery was not totally random. Does that mean it was not fair?

Lifestyle Changes and Illness

Are greater life-style changes associated with more serious illnesses? Patients hospitalized for various chronic conditions filled out questionnaires which purported to assess the degree of change in the person's lifestyle in the past two years. For each illness, a score "Illness" representing the seriousness of the illness was obtained, and a score "Change" representing the average life-style change of people with that illness.² The correlation is $r = 0.76$, quite strong.



? Is the correlation an ecological correlation?

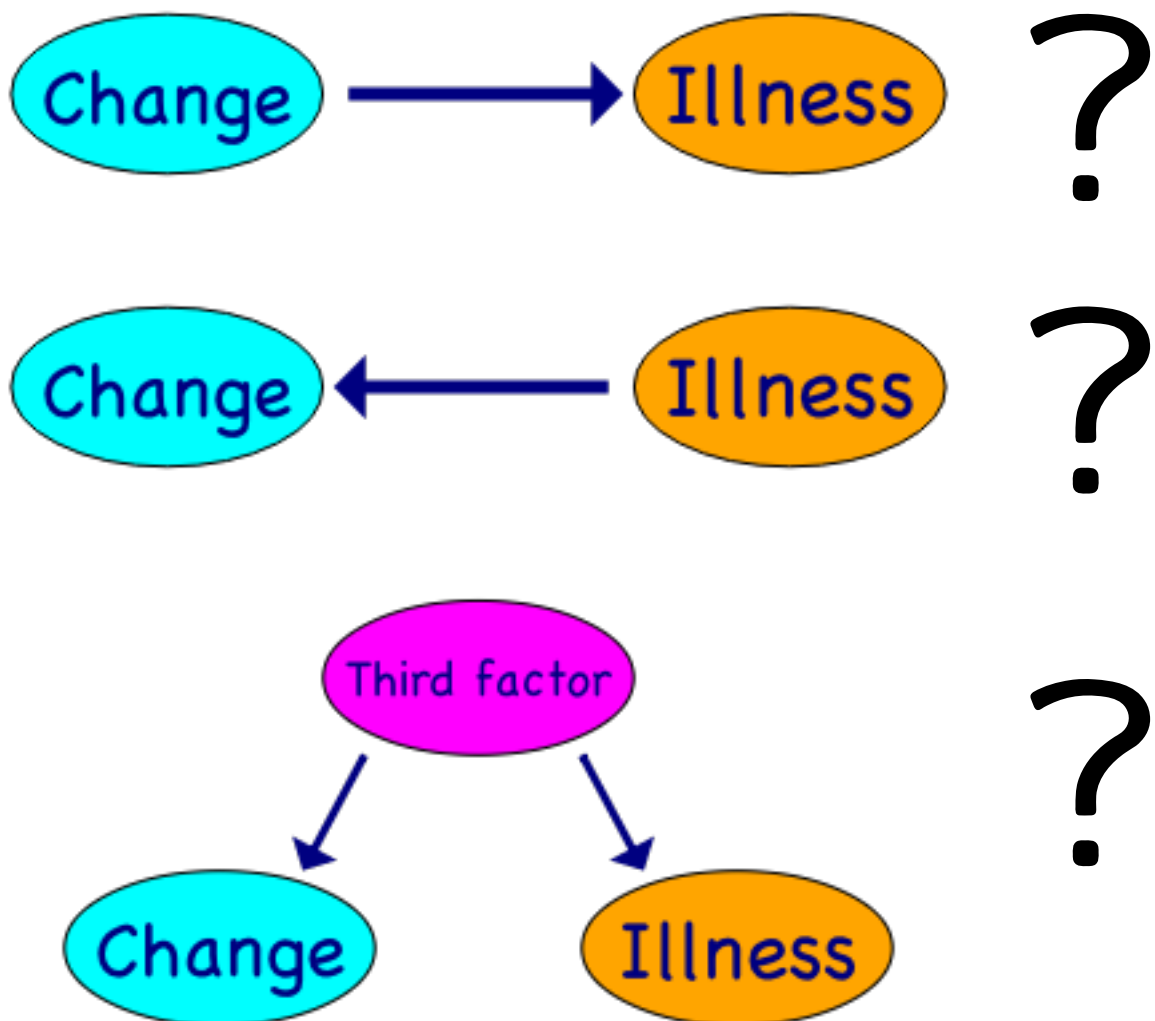
Can we conclude the correlation for individuals is very high?

²From *Statistics in the Real World* by Larsen and Stroup.

? There is a strong correlation between lifestyle changes and severity of illness. Does this correlation prove causation, that is, does it prove that lifestyle changes cause illness?

Are these data from a designed experiment or observational study?

Which of the following, if any, are plausible?



Morals

- Ecological correlations are those based on averages. They tend to be stronger (closer to ± 1) than correlations based on the individuals' data.

Do not assume ecological correlations reflect those for individuals.

- Correlation does not prove causation.

5.1 The regression effect

Imagine a situation that is repeated over time. For example, how an athlete performs from year to year, or how a student's GPA varies from semester to semester, or how much money one makes (or loses) playing poker from month to month. What often happens is that after a particularly good year, the next year is worse, much more like the average. Or after a particularly bad year, the next year is better, much more like the average. This phenomenon is called “**regression to the mean.**” The word “regression” makes it sound bad, but it is actually a neutral term. It just describes a statistical fact that arises from the variability of data.

National League batting leaders

For a baseball player, the “batting average,” or **BA**, is the proportion of times the player got a hit relative to his at bats.

$$BA = \frac{\# \text{ Hits}}{\# \text{ At bats}}.$$

The table on the next page shows the player with the best batting average in the National League for each year starting from 2000, as well as how well they did the following year. (And their career batting average.)

The National League Batting Average Leaders

	Year	BA	BA the next year	Career average
Todd Helton	2000	0.372	0.336	0.320
Larry Walker	2001	0.350	0.338	0.313
Barry Bonds	2002	0.370	0.341	0.298
Albert Pujols	2003	0.359	0.331	0.325
Barry Bonds	2004	0.362	0.286	0.298
Derrek Lee	2005	0.335	0.286	0.281
Freddy Sanchez	2006	0.344	0.304	0.297
Matt Holliday	2007	0.340	0.321	0.313
Chipper Jones	2008	0.364	0.264	0.303
Hanley Ramírez	2009	0.342	0.300	0.298
Carlos González	2010	0.336	0.295	0.299
José Reyes	2011	0.337	0.287	0.291
Buster Posey	2012	0.336	???	0.314
Average		0.350	0.307	0.304

A batting average of 0.300 is quite good, and 0.350 is excellent. Notice in every case, the player who had the best batting average in one year had a worse (though still usually pretty good) average the following year. Why did they do worse?

Became complacent? Opponents next year were more careful with them? They got a big raise and ate too much and got fat? They got older and lost a step?

Maybe. That's what the sports writers like to say, anyway. It could be just a simple statistical phenomenon: **Regression to the mean**

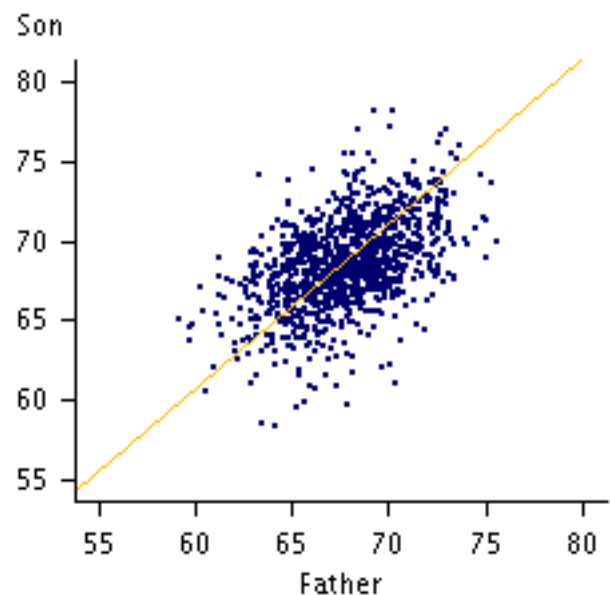
Note that, taken together, we see the average performance the next year was very close to the average career average. That suggests it wasn't because they were lazy or anything. They were still their usual selves. It was just they had one real good year, which can happen by chance. There's no story here.

Galton's height data

Sir Francis Galton (1822-1911) and his student, Karl Pearson (1857-1936), were very interested in inheritance. They looked closely at the heights of parents and their kids.



Here's the data again on 1078 Father/Son pairs, with X = Father's Height and Y = Son's Height. The correlation coefficient is $r = 0.5$. There doesn't seem to be anything unusual.



Average Son's Height for given Father's Height

Consider the fathers who are 65 inches tall. The average height of their sons was about 67 inches tall. So their sons were still short, but not quite as short as their fathers. What about tall fathers, say, those who are 74 inches (6 feet 2 inches) tall? Their sons averaged about 70 inches tall. So their sons were still tall, but not quite as tall as their fathers.

The average height of the sons for fathers of given height

Father's Height	Average Sons' Height
59	64.67
60	64.67
61	65.93
62	65.59
63	66.51
64	66.70
65	67.22
66	67.66
67	68.14

Father's Height	Average Sons' Height
68	69.09
69	69.44
70	69.77
71	70.54
72	70.68
73	72.05
74	70.36
75	71.72

Look at the fathers' heights, then at their sons' average height. Generally, they are not the same. In fact:

- Tall fathers had, on average, sons who were shorter than they (but still fairly tall);
- Short fathers had, on average, sons who were taller than they (but still fairly short).

OMG!

If tall fathers have shorter sons, and short fathers have taller sons, and that continues, eventually everyone will have be the same height! Regression to mediocrity!

? Are people correct to freak out, i.e., worry about everyone becoming average? Or is there another explanation.

Turn it around

Now look at the sons of various heights, and see the average height of their fathers:

- Sons who were 61 inches had fathers, on average, who were 65 inches;
- Sons who were 74 inches had fathers, on average, who were 70 inches.

So short sons had taller fathers, and tall sons had shorter fathers.

OMG! If that continues, will eventually everyone be either very short or very tall?

No. In fact, the SD's for both fathers and sons was about 3 inches. So there is no evidence heights are becoming less spread out nor more spread out. What we are seeing is the

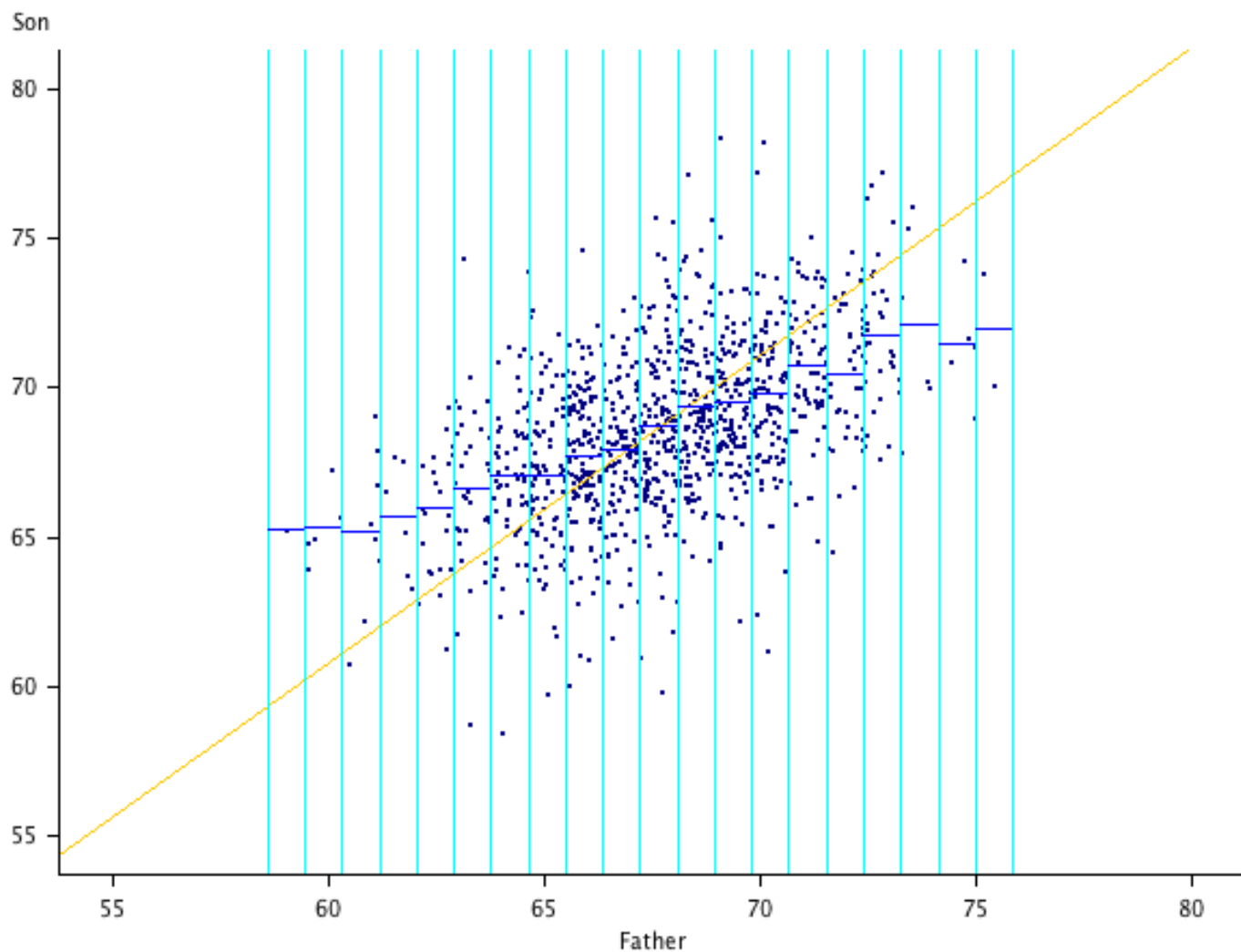
Regression effect: In repeated situations, large values the first time will be associated with smaller values the second time; small values the first time will be associated with larger values the second time.

The **regression fallacy** is to see the regression effect, and think that something is going wrong. For example, the regression fallacy is

- To think baseball players who have the best batting average one year, get lazy and unmotivated the next year;
- To think eventually everyone will be the same height.

The regression effect in a scatter plot

In the scatter plot below of the fathers' and sons' height, the little blue line segments indicate the average of the sons' heights for the points in that strip. Notice that for tall fathers, the lines are below the SD line, and for short fathers, the lines are above the SD line. That is the graphical representation of regression to the mean.



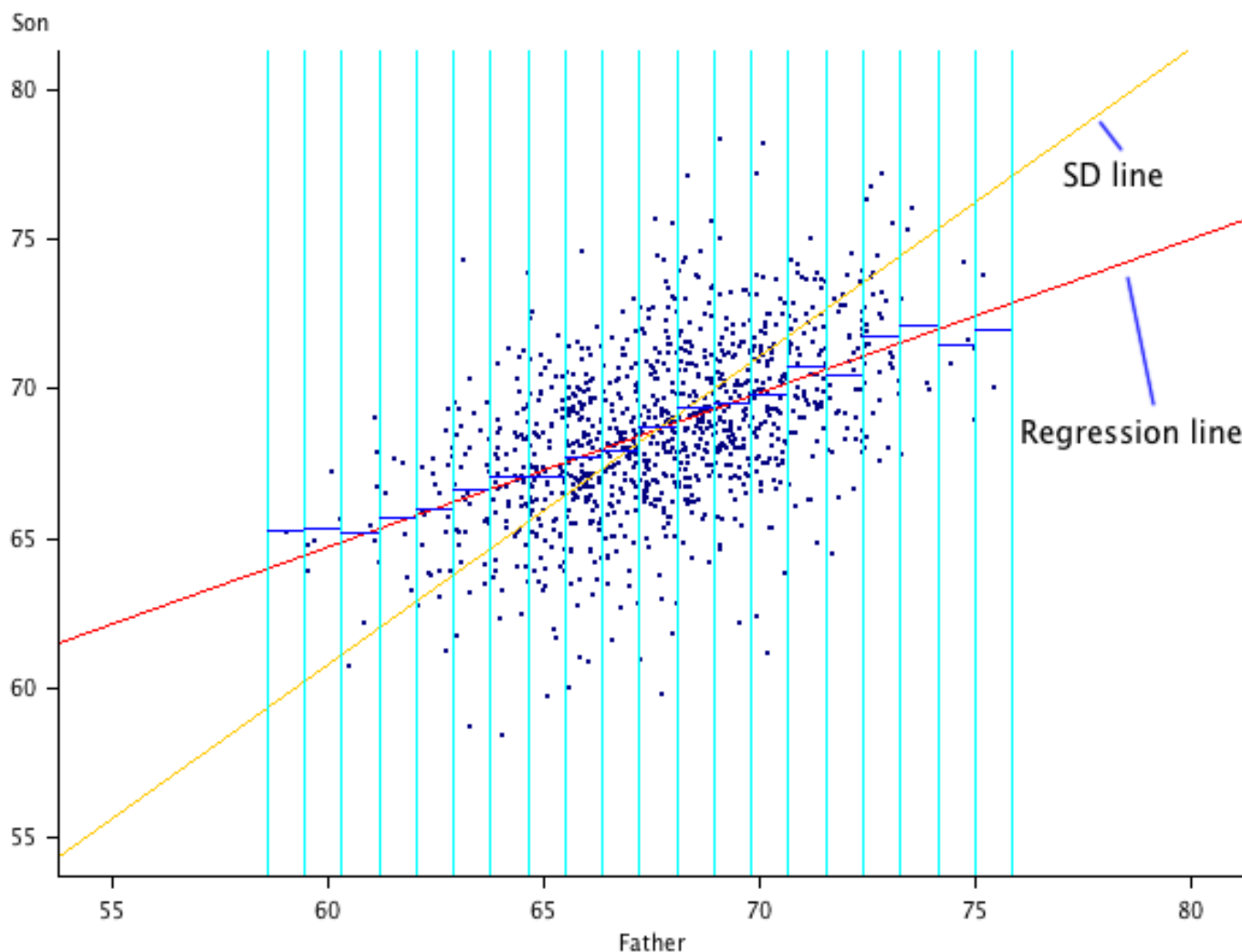
Moral

In situations repeated over the days or years or generations, you will often see very high values one time become smaller the next time, or very low values one time be higher the next time. That phenomenon is called **regression to the mean** or the **regression effect**. It is due to variation in the scatter plot.

The **regression fallacy** is to think something else is going on. That is, it is wrong to assume everyone is eventually going to be the same height, or have the same IQ, or the same batting average. Or that people who excel one year and are not as good the next are getting lazy or complacent.

5.2 The Regression line

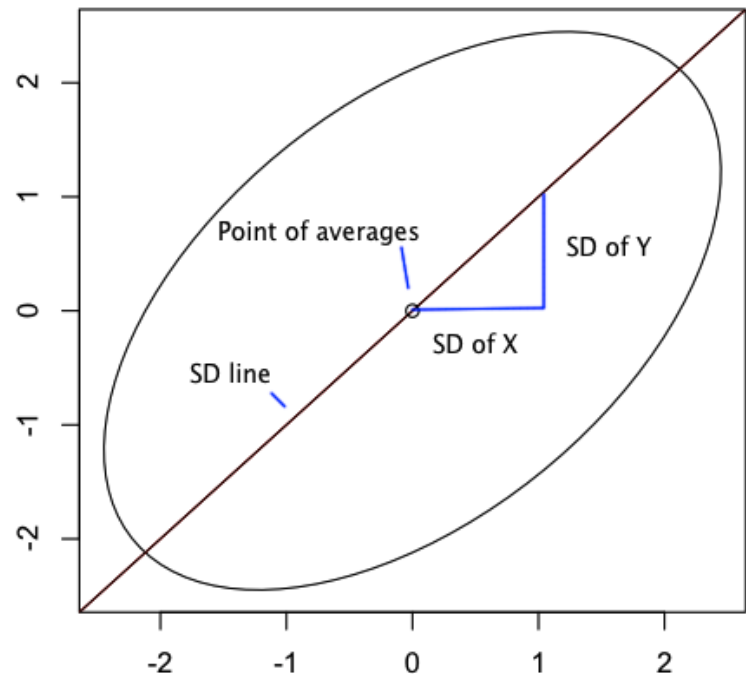
The graph shows the SD line, and the average sons' heights for given fathers' heights. These averages are roughly in a straight line, but they don't follow the SD line. The line through those averages is called the **regression line**, the red line.



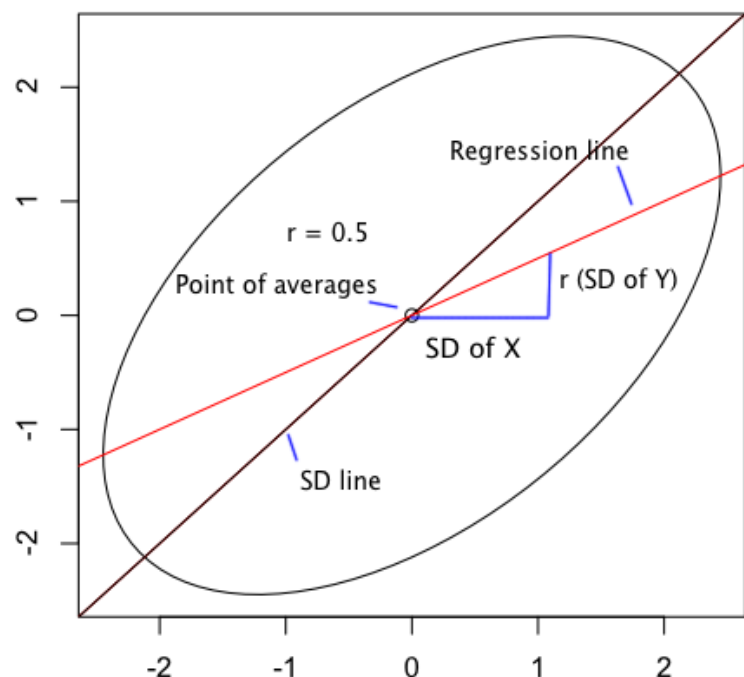
Notice that the regression line is **flatter** than the SD line, which is due to the regression effect.

The SD Line and the Regression Line

Recall that for the SD line, we start at the point of averages, then go over one SD of X , and up one SD of Y . (At least if the correlation $r > 0$.)



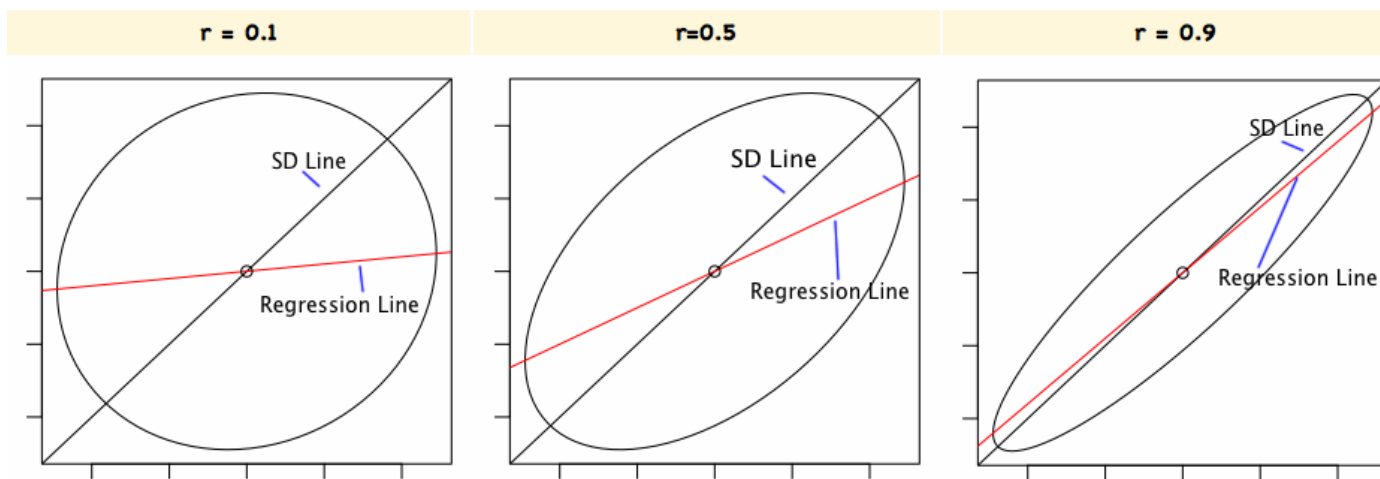
For the regression line, again start at the point of averages. Then go over one SD of X , but don't go up a whole SD of Y . Instead, go up $r \times (\text{SD of } Y)$. So if $r = 0.5$, go up only half the SD of Y .



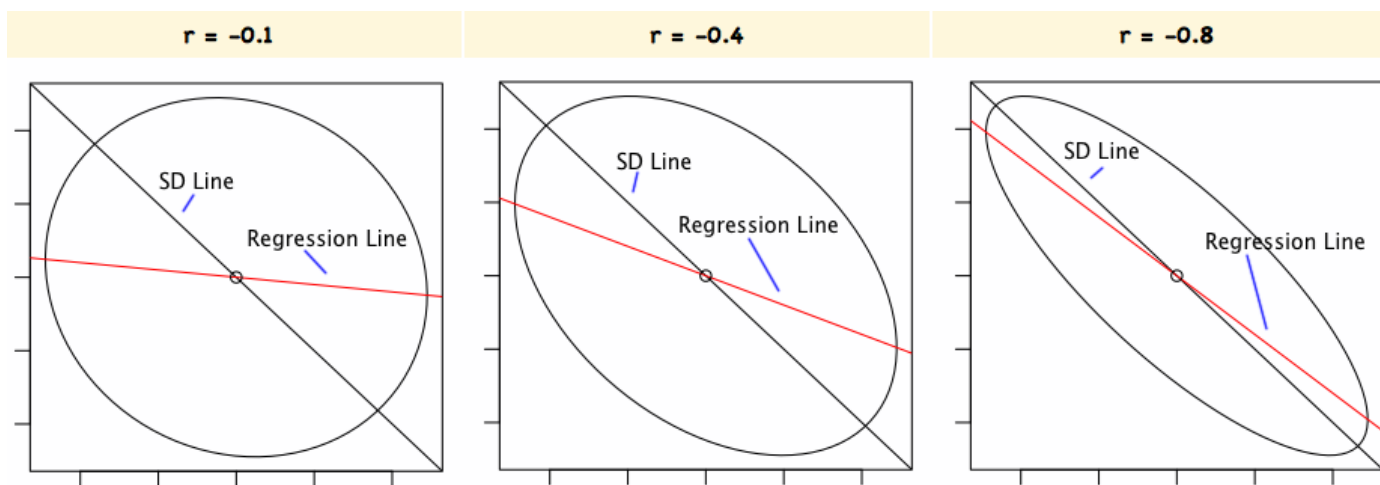
Because we go up only $r \times (\text{SD of } Y)$, the regression line is flatter than the SD line. Notice that that was what we saw in the fathers' vs. sons' height data, where $r = 0.50$.

More examples

The closer the r is to 1, the closer the regression line is to the SD line. The closer r is to 0, the flatter the regression line.

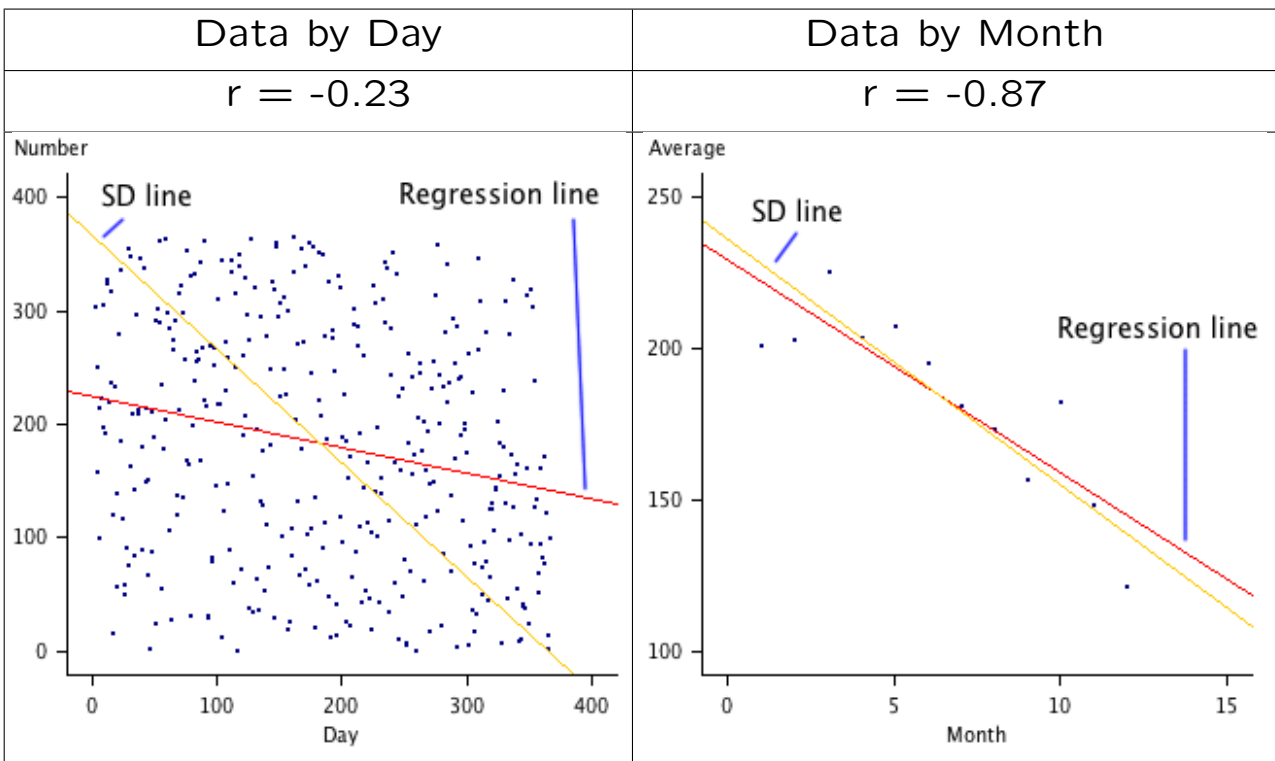


If the correlation r is negative, $r < 0$, the SD line slopes downward. So does the regression line. The regression line is still flatter than the SD line. The closer the r is to -1 , the closer the regression line is to the SD line. The closer r is to 0, the flatter the regression line.



The draft lottery

Here is the draft lottery again, using both the daily data, and the monthly data.



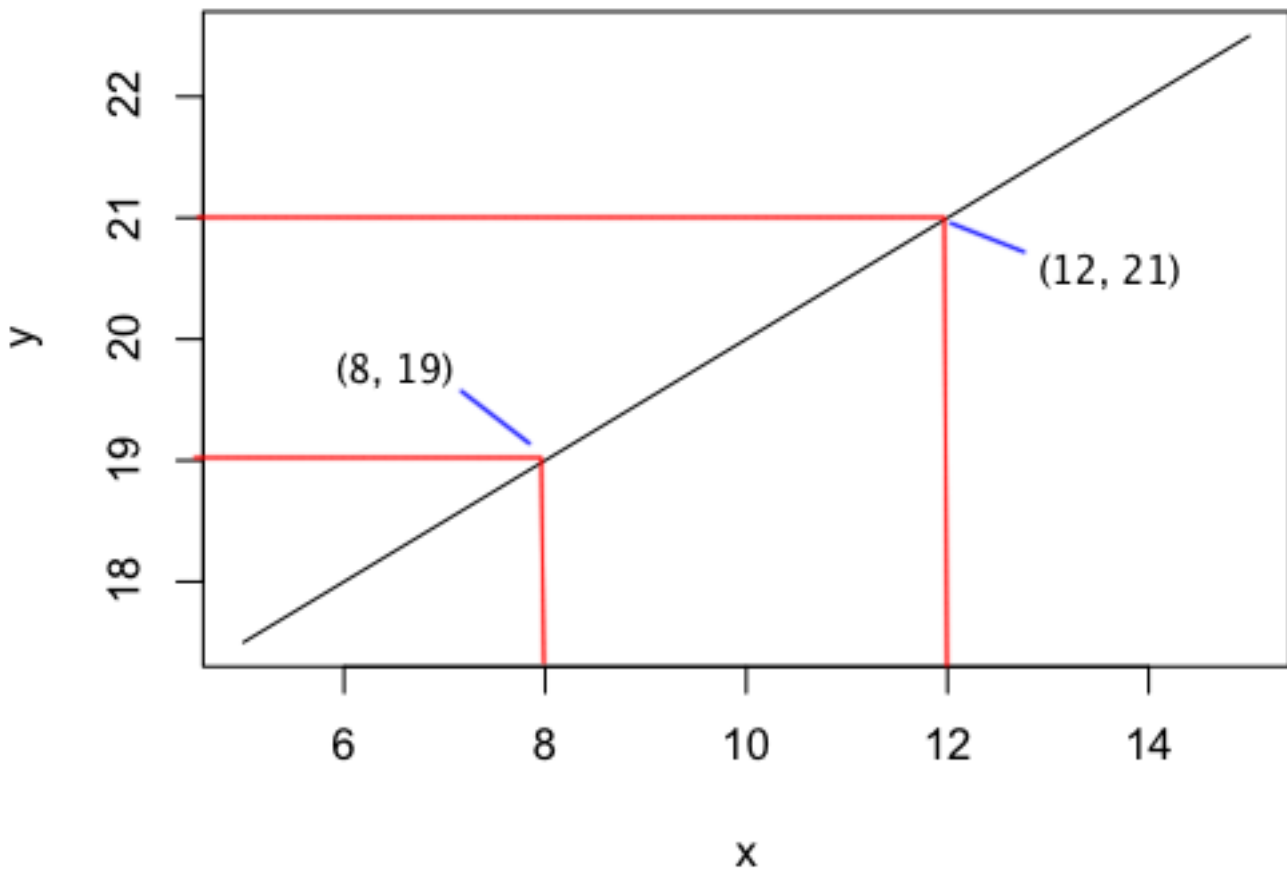
These are negative correlations. The first plot, with the individual days, shows the regression line is much flatter than the SD line, because r is small, $r = -0.23$. The second plot, based on monthly averages has $r = -0.87$ (an ecological correlation), which is fairly close to -1 . Thus the regression line is flatter than the SD line, but not by much.

The equation of a line

The equation of the line in the plot below is

$$Y = \frac{1}{2}X + 15.$$

You can plot the line by finding a couple of points: For example, $X = 8$ goes with $Y = \frac{1}{2} \cdot 8 + 15 = 19$, and $X = 12$ goes with $Y = \frac{1}{2} \cdot 12 + 15 = 21$.



The equation of any straight line is

$$Y = m X + b.$$

The m is the slope and the b is the intercept. That is,

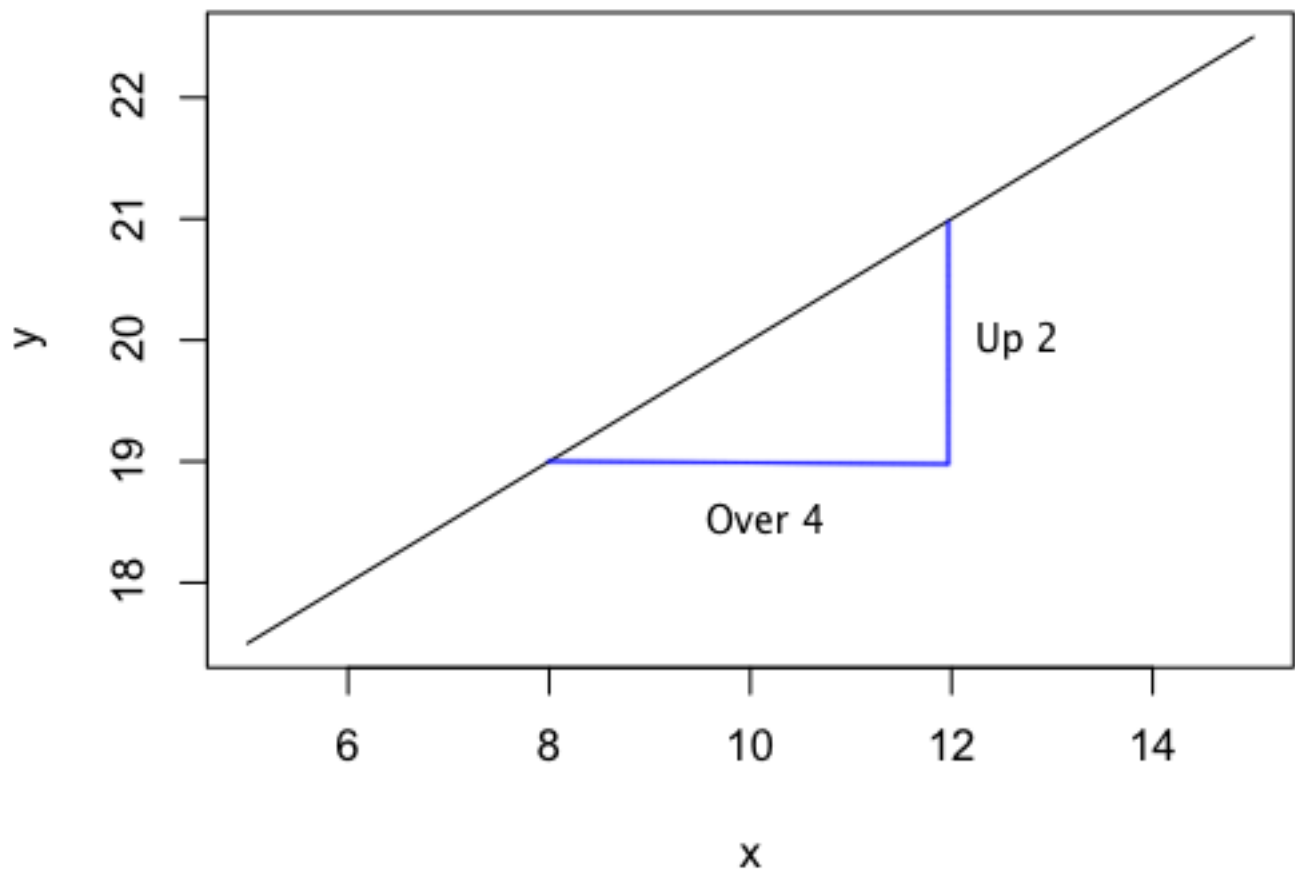
$$Y = (\text{slope}) \times X + (\text{intercept}).$$

So for the line above, slope = $\frac{1}{2}$, and intercept = 15.

The slope of a line

The slope of a line tells how far you go up relative to how far you go over:

$$\text{Slope} = \frac{\text{Up}}{\text{Over}}.$$



In this plot, if you go over from $x = 8$ to $x = 12$, i.e., $12 - 8 = 4$, then you go up from $y = 19$ to $y = 21$, or $21 - 19 = 2$. Thus

$$\text{Over} = 4 \text{ \& Up} = 2 \implies \text{Slope} = \frac{\text{Up}}{\text{Over}} = \frac{2}{4} = \frac{1}{2}.$$

Indeed, the equation is

$$Y = \frac{1}{2}X + 15.$$

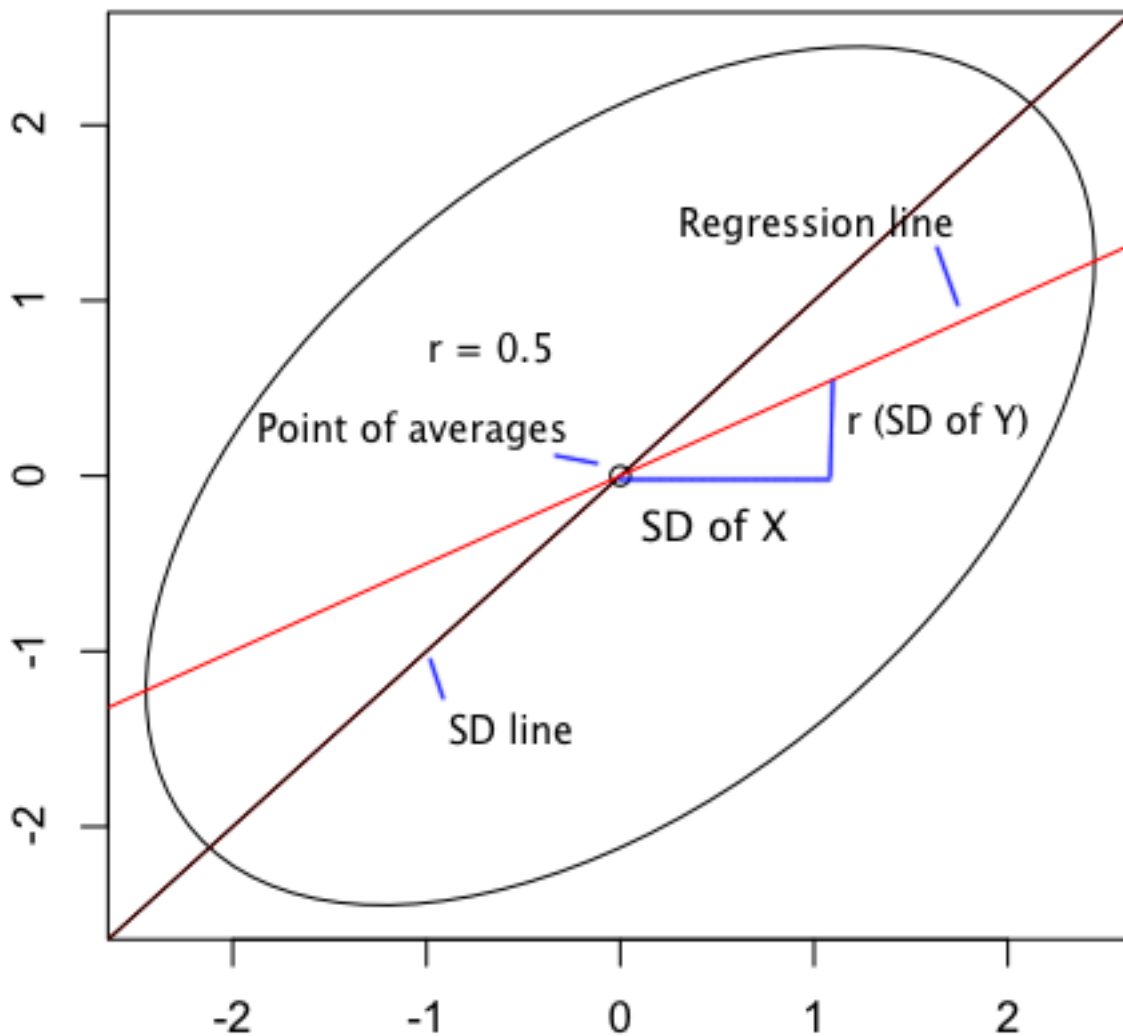
The slope of the regression line

Again, the slope of a line is

$$\text{Slope} = \frac{\text{Up}}{\text{Over}}.$$

For a regression line, you go over one SD of X, and up $r \times (\text{SD of Y})$.
So

$$\text{Slope} = \frac{\text{Up}}{\text{Over}} = r \frac{\text{SD}_Y}{\text{SD}_X}.$$



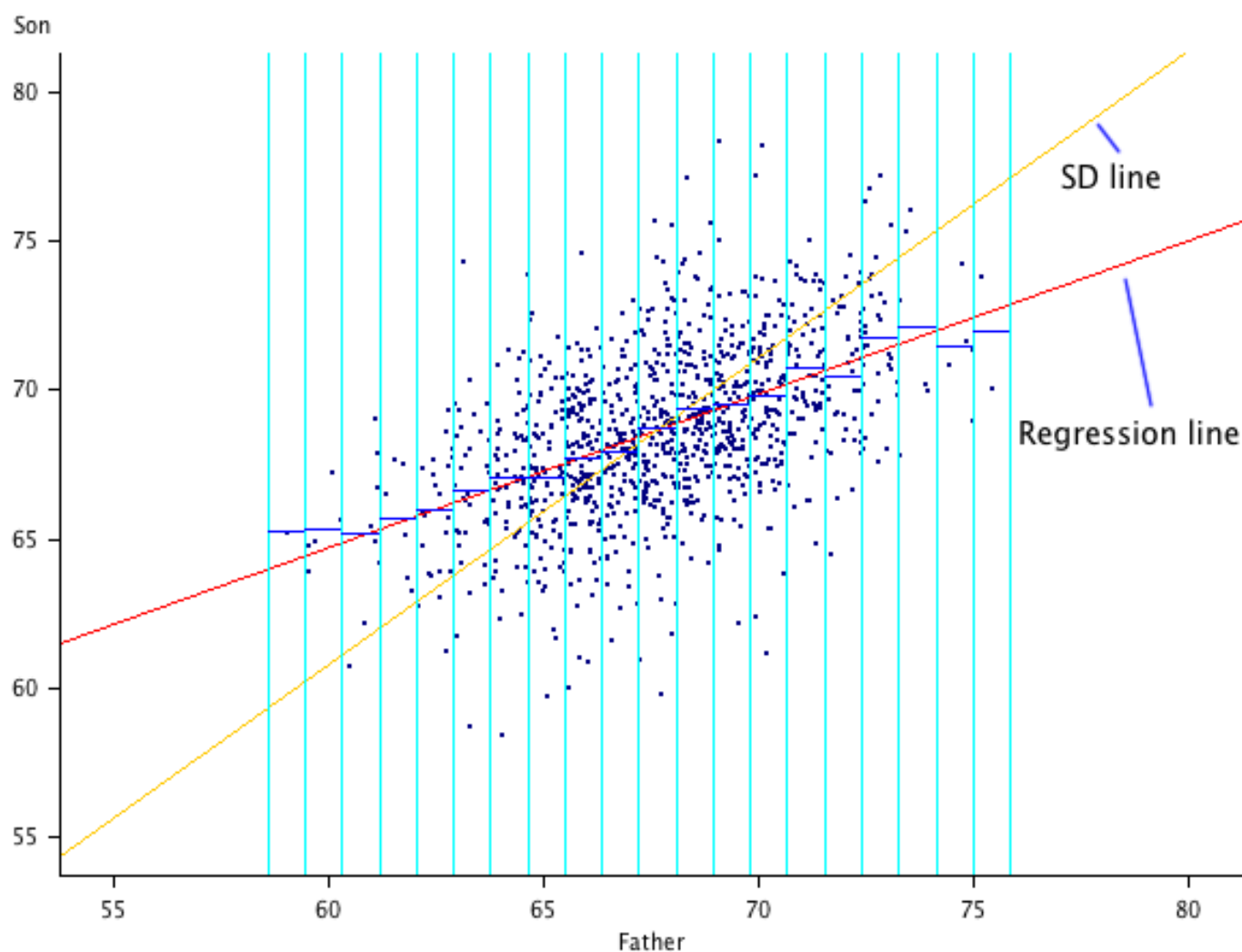
Heights of fathers and sons

For these data,

$$SD_X = 3, \quad SD_Y = 3, \quad r = 0.5$$

so that

$$\text{Slope} = r \frac{SD_Y}{SD_X} = 0.5 \frac{3}{3} = 0.5.$$



The intercept of the regression line

The regression line always goes through the point of averages. Since $Y = (\text{slope}) \times X + (\text{intercept})$,

$$\text{Average}_Y = \text{slope} \times \text{Average}_X + \text{intercept}.$$

For the heights data, the slope = 0.5, the average of the X (fathers) is 68 inches, and the average of the Y (sons) is 69 inches. Plugging in:

$$\begin{aligned} 69 &= 0.5 \times 68 + \text{intercept} \\ &= 34 + \text{intercept} \end{aligned}$$

$$\begin{aligned} \text{so } 69 - 34 &= \text{intercept} \\ 35 &= \text{intercept}. \end{aligned}$$

The intercept is then 35, and the slope is 0.5, so the regression equation is

$$Y = 0.5 \times X + 35,$$

or

$$\text{Son's height} = 0.5 \times (\text{Father's height}) + 35.$$

Using the regression line to estimate average Y for given X

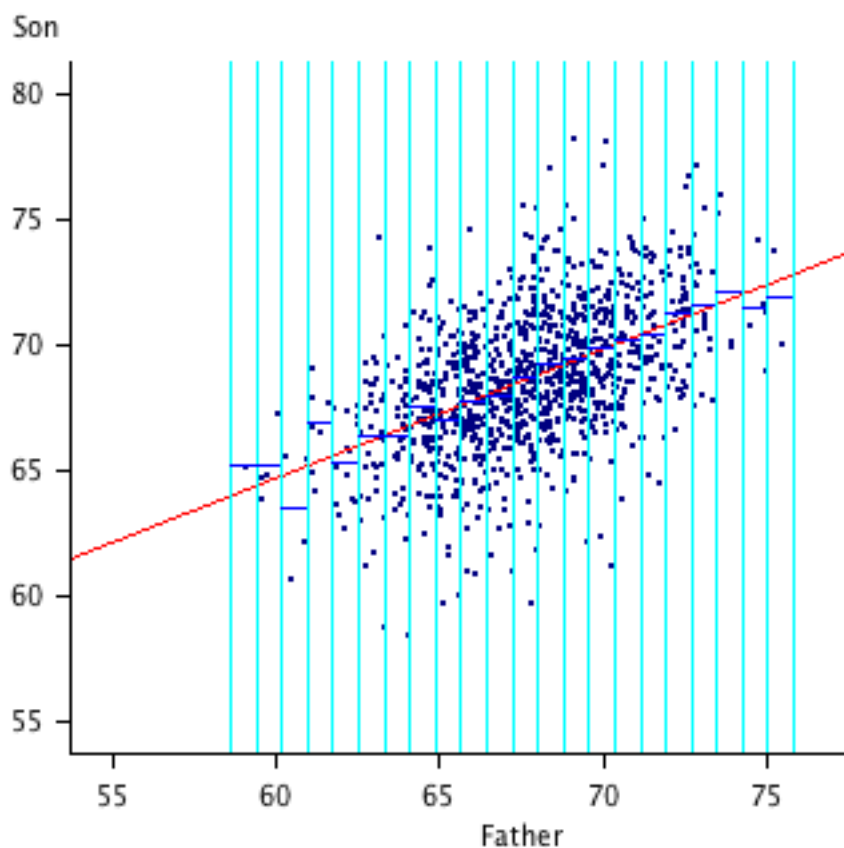
The equation for the regression line for the father/son height data is

$$Y = 0.5 \times X + 35.$$

The regression line roughly goes through the average Y for each X, that is, the average sons' height for given fathers' height. Look at fathers who are 74 inches tall. What is the average height of their sons? Stick 74 in for X:

$$Y = 0.5 \times 74 + 35 = 37 + 35 = 72.$$

Thus we estimate the average height of those sons to be 72 inches.



X	Y
Fathers' height	Average height of sons
60	65
65	
68	69
70	70
74	72
76	

? The table has some other values. Fill in the two blanks (for $X = 65$ and 70). Do you see the regression effect?

Interpreting the slope

The equation again: $Y = 0.5 \times X + 35$.

X	Y
Fathers' height	Average height of sons
60	65
70	70
74	72

The slope is 0.5. What does that mean? It tells you how many additional units of “Y” are associated with one additional unit of “X.”

Look at $X = 70$. The $Y = 70$.

- Now increase X by 4, to 74. How much does Y increase? From 70 to 72: it increases by 2 inches. Half as much.
- Or go from $X = 70$ down by 10, to 60. The Y goes from 70 to 65, or down by 5. Only half as much.

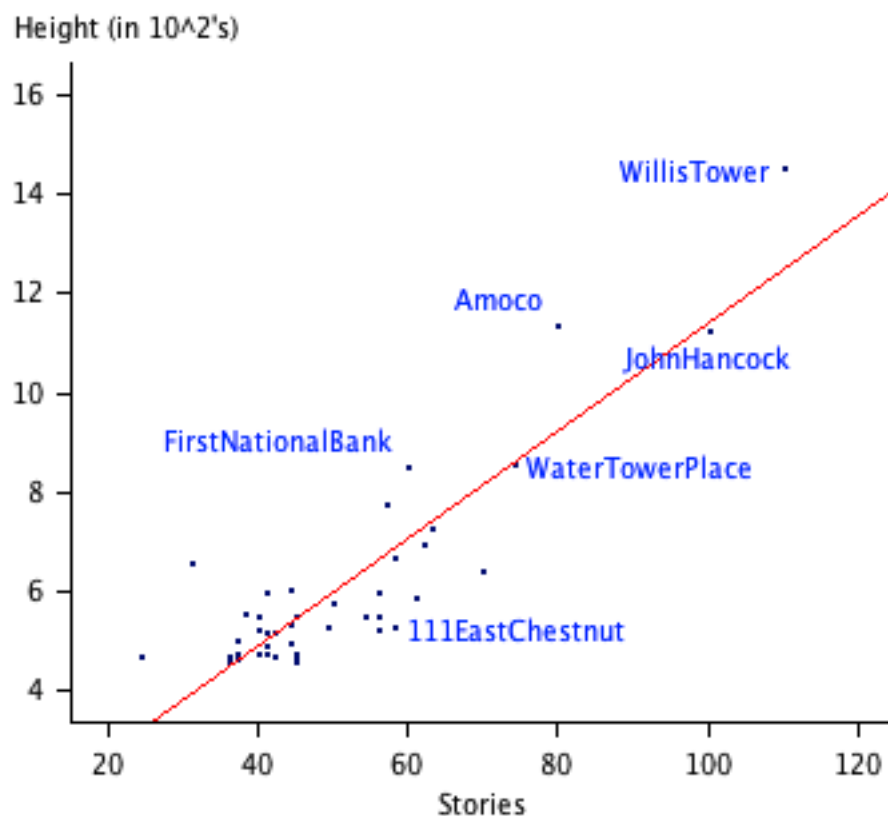
The slope gives the tradeoff: One more inch in X is associated with one-half more inches in Y .

? Start at $X = 74$, so that $Y = 72$. If you go up to $X = 78$, how much does Y go up? What is the new value of Y ?

Start at $X = 70$, so that $Y = 70$. If you go down to $X = 65$, how much does Y go down? What is the new value of Y ?

Buildings

This scatter plot contains 42 tall buildings in Chicago. $X = \#$ of stories, $Y =$ height in feet. Just using your imagination, can you guess (approximately) the slope? Slope = 1? Would each additional story be associated with about one additional foot in height? Those would be pretty low ceilings. Slope = 30? Would each additional story be associated with about 30 additional feet in height? Maybe. Slope = 100? Would each additional story be associated with about 100 additional feet in height? Nah, too much.



? What would you guess the slope would be, approximately? That is, how many additional feet would be associated with one additional story, on average?

Again, for the buildings, $X = \#$ of stories, $Y =$ height in feet. Some statistics:

$$SD_X = 16.8, \quad SD_Y = 204, \quad r = 0.88.$$

That's a high correlation.

? Use the formula from page 142 to calculate the slope of the regression line.

An additional story is associated with how many additional feet, on average?

If building A is two stories shorter than building B, how much shorter would building A be than building B in feet, on average?

5.3 Errors from the regression Line

Is someone who weighs 195 pounds fat? Skinny? These two fellows both weigh approximately 195 pounds. The average weight of men in the class was 172 pounds.



What are these two guys' deviations?

Deviation = value – average = 195 – 172 = 23 pounds.

They are both 23 pounds heavier than average. (Though they were not actually in the class.) Determining whether someone is stocky or lanky is not based on just weight.

There are potentially many factors other than weight:

- Height
- Build
- Ethnicity
- Musculature

? What other factors could there be?

Height vs weight

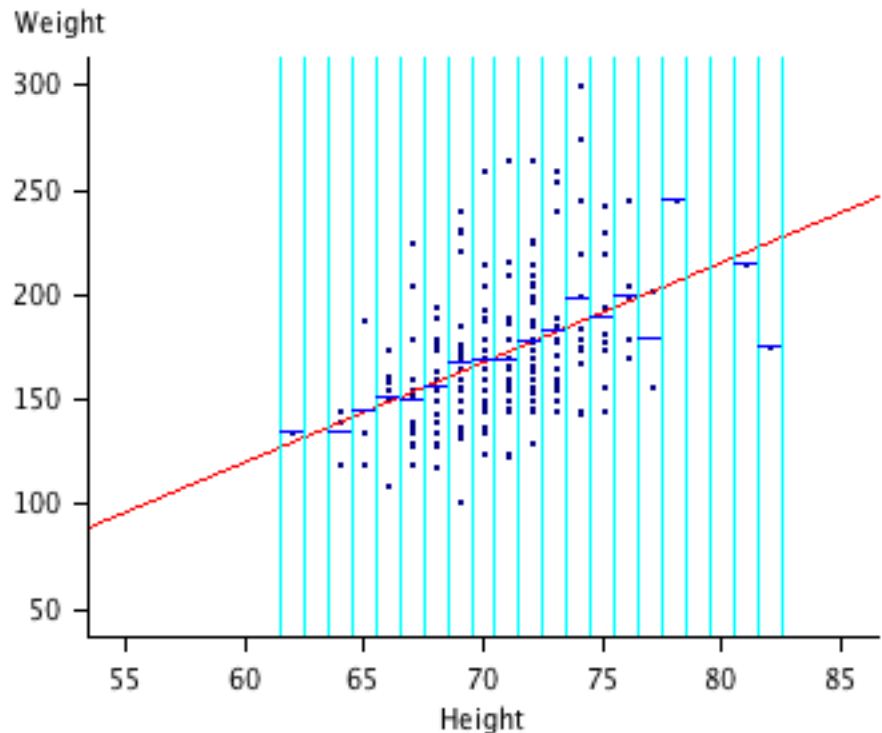
The graph has

$X = \text{Height}$

and

$Y = \text{Weight}$

for males in a class. The little blue lines represent the average weight for each height.



The line is the regression line:

$$\text{Weight} = 4.8 \times (\text{Height}) - 166.$$

What's the average weight for guys who are 65 inches (like George)?

$$\text{Weight} = 4.8 \times 65 - 166 = 146 \text{ pounds.}$$

So if he weighs 195 pounds, he is actually

$$195 - 146 = 49 \text{ pounds heavier}$$

than the average guy his height. What about a guy who is 75 inches (6 foot 3), like Kramer? $\text{Weight} = 4.8 \times 75 - 166 = 194$ pounds. So he's $195 - 194 = 1$ pound heavier than the average for his height.

So, though they are about the same weight, George is 49 pounds heavier than average *for his height*, while Kramer is just about average *for his height*.

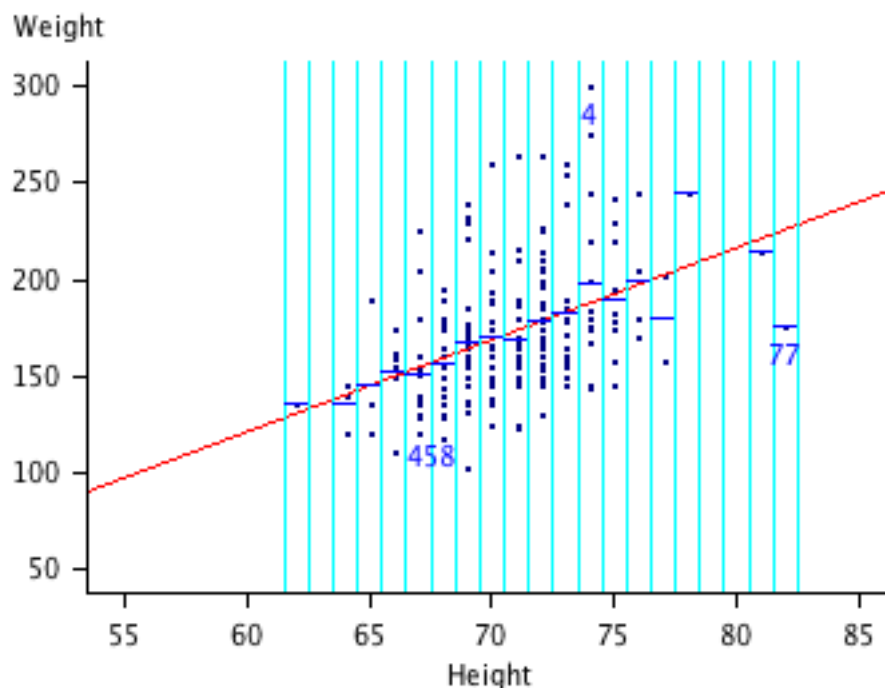
Errors

The difference between someone's actual weight and the average weight for his height is often called the "error." The idea is that knowing the person's height, you'd guess his weight was the average weight for that height. The error is how far off your guess would be. Regression equation:

$$\text{Weight} = 4.8 \times (\text{Height}) - 166.$$

Actual weight (Y)	Height (X)	Average weight for that height = $4.8 \times X - 166$	Error
(George) 195	65	$4.8 \times 65 - 166 = 146$	$195 - 146 = 49$
(Kramer) 195	75	$4.8 \times 75 - 166 = 194$	$195 - 194 = 1$
(person 458) 110	66	$4.8 \times 66 - 166 = 150.8$	
(person 77) 176	82	$4.8 \times 82 - 166 = 227.6$	$176 - 227.6 = -51.6$
(person 4) 300	74	$4.8 \times 74 - 166 = 189.2$	$300 - 189.2 = 110.8$

The table has the height and weight and error for George and Kramer, as well as for three other men in the class. So, for example, person 4 is very tall (82 inches), but weighs only 176 pounds. Average for his height is 227.6 pounds, so his error is $176 - 227.6 = -51.6$.



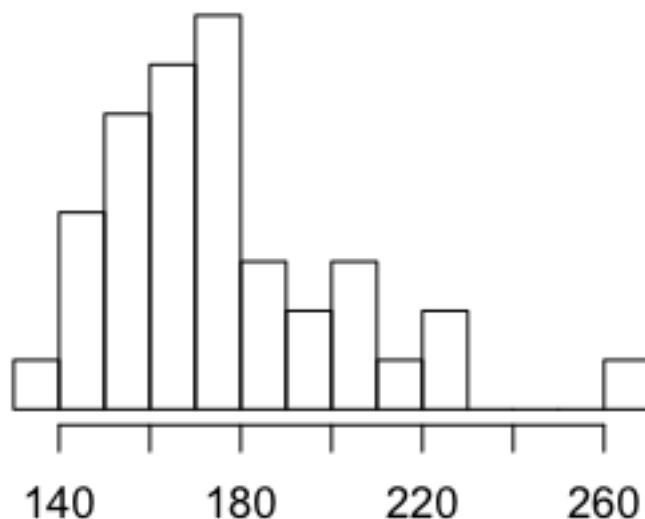
? Fill in the error for person 458.

The typical error from a regression line

X = Height, Y = Weight, for males. The regression line gives an estimate of the average weight for a specific height, but certainly there is quite a bit of variability even for a given height.

There are 38 men who are 72 inches tall. Their weights are

175 175 178 165 150 164 215
 265 180 170 170 146 200 175
 188 170 155 156 160 205 188
 225 170 210 180 227 175 145
 154 206 180 167 155 145 130
 197 186 155



The histogram shows these weights.

Their average weight for men 72 inches tall is about 178 pounds, and the SD is about 27 pounds.

The SD of errors (RMSE)

For each height, there is another average weight, and another SD of weight. If the scatter plot is generally football-shaped, we can estimate these SD's with what is called the **root mean square error** (RMSE), which is the root mean square of all the errors, so

$$\text{RMSE} = \text{SD}_{\text{errors}} = \text{the SD of the errors.}$$

Calculating the SD_{errors}

The SD_{errors} is the root mean square error, so is calculated in much the same way the SD is. Here are the steps:

1. Error: For each point in the scatter plot, find the error
2. Square: Square all the errors
3. Mean: Average the squared errors
4. Root: Take the square root of that average

(Note: Like the deviations, the average of the errors is 0, since the plusses and minuses balance out.)

Fortunately, there is a quick formula for finding the SD_{errors} :

$$RMSE = SD_{\text{errors}} = \sqrt{1 - r^2} \times SD_Y.$$

For the height/weight data, the SD of weights is $SD_Y = 32.85$ pounds, and the correlation between height and weight is $r = 0.438$. Calculating:

$$\begin{aligned} SD_{\text{errors}} &= \sqrt{1 - r^2} \times SD_Y \\ &= \sqrt{1 - 0.438^2} \times 32.85 \\ &= \sqrt{1 - 0.192} \times 32.85 \\ &= \sqrt{0.808} \times 32.85 \\ &= 0.899 \times 32.85 \\ &= 29.53 \end{aligned}$$

So there are two standard deviations:

- The regular SD measures how far the weights are from the overall average weight: $SD_Y = 32.85$
- The SD of errors measures how far the weights are from the average weight for the given height: $SD_{\text{errors}} = 29.53$

More generally,

- The SD_Y measures how far the values are from the overall average of Y .
- The SD_{errors} measures how far the values are from the regression line.

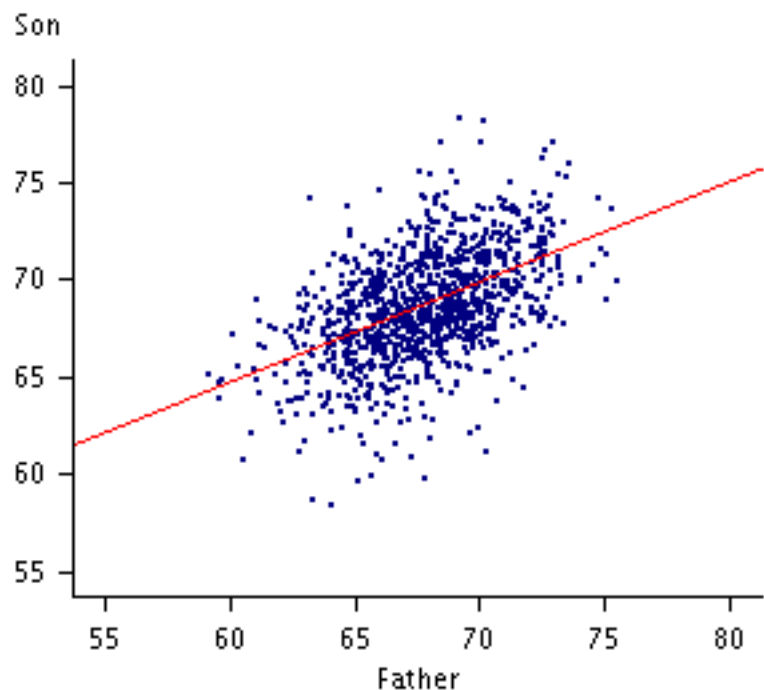
The key formula is

$$SD_{\text{errors}} = \sqrt{1 - r^2} \times SD_Y.$$

? Here is the plot with X = heights of fathers and Y = heights of sons.

There are 1078 father/son pairs in the data. The correlation coefficient is $r = 0.50$, and the SD for the fathers is 3, and the SD for the sons is also 3.

Find the SD_{errors} .

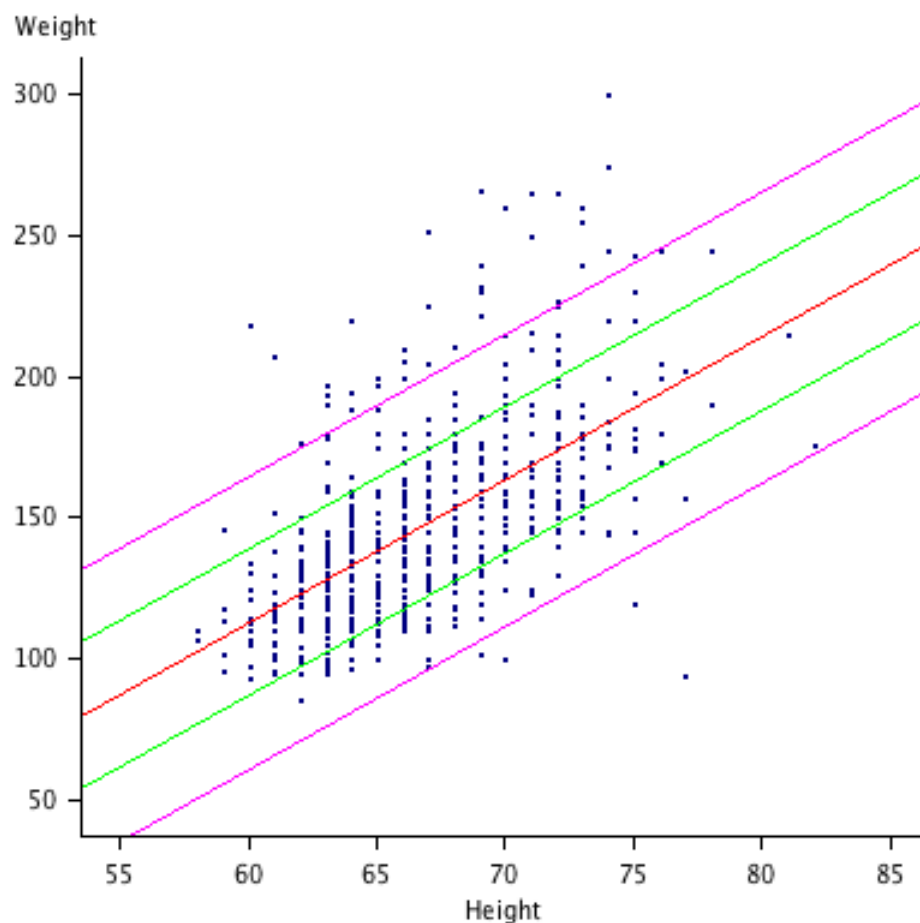


The rule of thumb for scatter plots

If a scatter plot is reasonably football-shaped, then

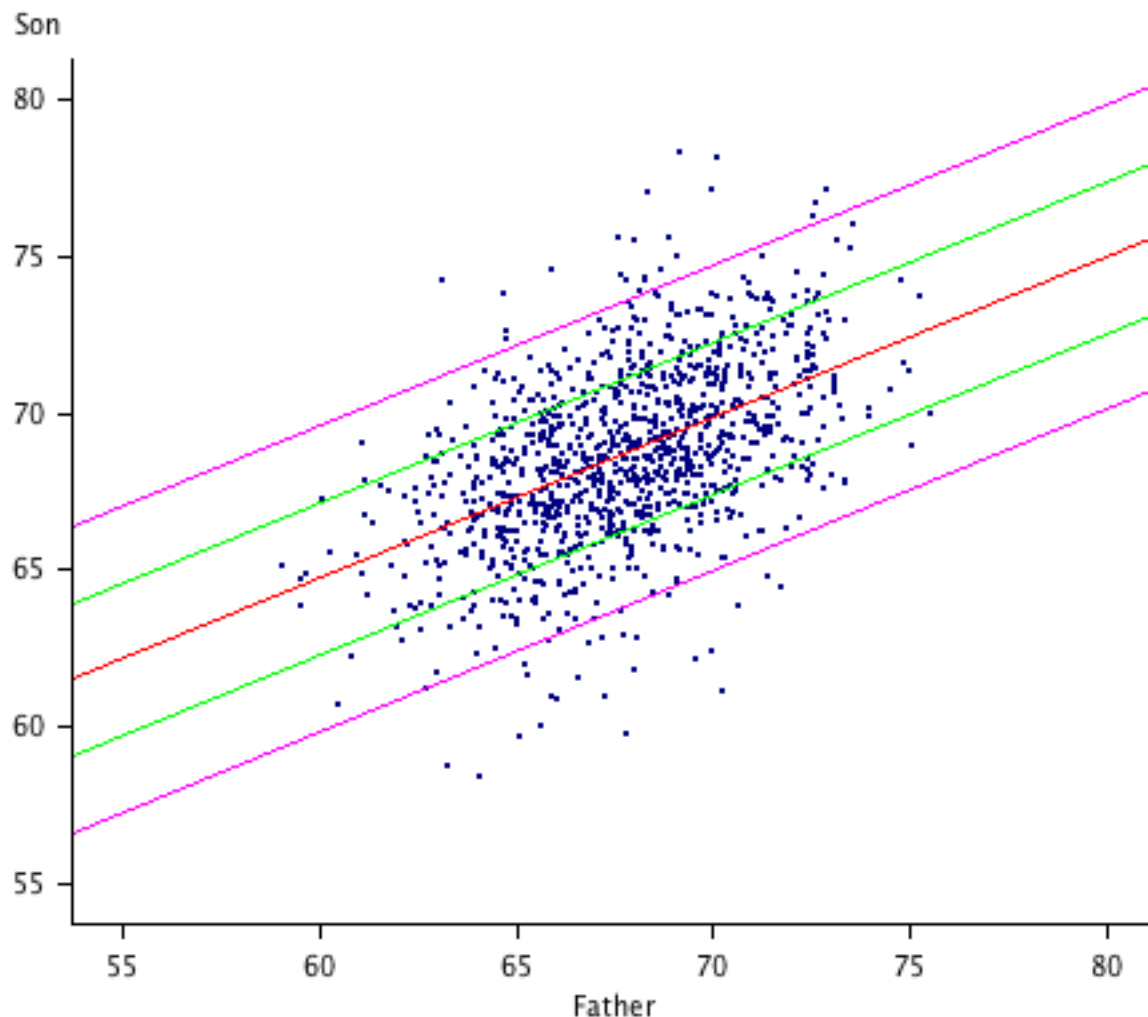
- Approximately 68% of the points are within $\pm SD_{\text{errors}}$ of the regression line.
- Approximately 95% of the points are within $\pm 2 \times SD_{\text{errors}}$ of the regression line.

The plot below has the heights and weights of the men. The red (center) line is the regression line, the green (second and fourth) lines are $\pm SD_{\text{errors}}$ from the regression lines, and the purple (outer) lines are $\pm 2 SD_{\text{errors}}$ from the regression line.



So about 95% of the points are between the two outer lines, and about 68% of the points are between the second and fourth lines.

? Here is the scatter plot for the heights of fathers' and their sons.



The $SD_{\text{errors}} = 2.60$. How far above and below the regression line are the two outer lines?

There are 1078 father/son pairs in the data. According to the rule of thumb, how many points should be outside the two outer lines? About how many points actually are outside the two outer lines?

5.4 Using the normal curve

Consider the heights and weights of men. We can use the normal curve to estimate the percentage of men who weigh more than 170 pounds. But we can also estimate the percentages such as

- The percentage of men who are 65 inches tall who weigh more than 170 pounds
- The percentage of men who are 72 inches tall who weigh more than 170 pounds
- The percentage of men who are 75 inches tall who weigh more than 170 pounds

For each height, we need to get a new average and new SD based on the height.

- The new average weight is found using the regression line.
- The new SD of weight is the SD_{errors} .

Look at the men who are 65 inches tall. The regression line (from page 150) is

$$\text{Weight} = 4.8 \times (\text{Height}) - 166.$$

So if we plug in $\text{Height} = 65$ inches, we get the new average:

$$\text{New average} = 4.8 \times 65 - 166 = 146.$$

The $r = 0.438$ and the $SD_Y = 32.85$, so

$$\begin{aligned} \text{New SD} = SD_{\text{errors}} &= \sqrt{1 - r^2} \times SD_Y \\ &= \sqrt{1 - 0.438^2} \times 32.85 \\ &= 29.53, \end{aligned}$$

as on page 153.

To the normal curve

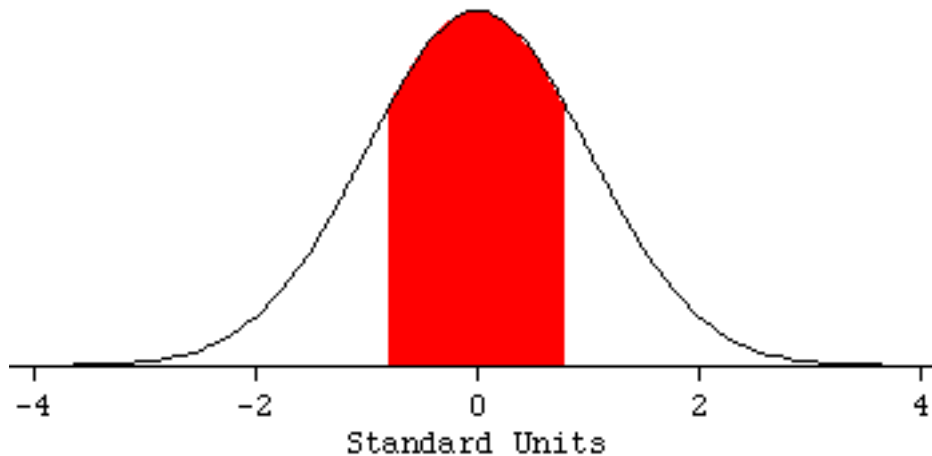
To summarize, we want to estimate the percentage of men who are 65 inches tall who weigh more than 170 pounds. We use the normal curve, with the new average and SD:

$$\text{New average} = 146, \quad \text{New SD} = 29.53.$$

To use the normal curve, we first have to standardize the 170 with the new average and SD:

$$\text{Standard Units} = \frac{\text{Value} - \text{Average}}{\text{SD}} = \frac{170 - 146}{29.53} = 0.81.$$

We want the percentage over 170 pounds, so we need the percentage in the normal curve over 0.81. The area between ± 0.80 (which is as close as we can get) is 57.63%.



Then

$$\text{The percentage over 170 pounds} \approx \frac{1}{2}(100\% - 57.63\%) = 21.18\%.$$

So about 21% of the guys who are 65 inches tall are over 170 pounds.

What about the percentage of men who are 75 inches tall who weigh more than 170 pounds?

We have to again go through the four steps:

1. Find the new average: Use the regression line to find the average Y (weight) for people with X (height) = 75.

$$\text{New average} = 4.8 \times 75 - 166 = 194.$$

2. Find the new SD: Use the SD_{errors} to estimate the SD of the Y for people with X = 75.

$$\text{New SD} = SD_{\text{errors}} = 29.53. \text{ (The same as on the previous page.)}$$

3. Translate the Y value of interest (in this example, 170 pounds) into standard units, using the new average and new SD.

$$\text{Standard Units} = \frac{\text{Value} - \text{New average}}{\text{New SD}} = \frac{170 - 194}{29.53} = -0.81.$$

4. Use the normal curve for data to estimate the percentage of people with Y (weight) over 170 pounds, using the standard units.

We now need the percentage in the normal curve over -0.81.

We just found that area between ± 0.80 to be 57.63%. So the area above -0.80 is the area between ± 0.80 plus the area above 0.80, which we found to be 21.18%. So the percentage over 170 pounds $57.63\% + 21.18\% = 78.81\%$.

We end up saying that about 79% of the guys who are 75 inches tall are over 170 pounds.

And from the previous page, about 21% of the guys who are 65 inches tall are over 170 pounds.

Do those percentages seem reasonable?

We could do the same for other heights.

Height	Average weight for height	Standard Units	Percentage over 170 pounds
62	131.6	1.30	
65	146	0.81	21.19%
70	170.0	0	50%
72	179.6		63.68%
75	194.0	-0.81	78.81%
78			

This table should seem reasonable. The percentage of men over 170 pounds depends on their height. A small percentage of short guys are over 170 pounds, but most tall guys are over 170 pounds. In fact, about 90% of men 78 inches tall are over 170 pounds.

? Fill in the blanks in the table.

Part II

Box Models

The chance something happens is the percentage of time it happens if one repeats the process over & over & over & (to infinity). If you draw five cards from a regular deck of 52, what's the chance you get a pair (two cards of the same number)? We drew 1000 hands. Here are the results (with the last hand showing):

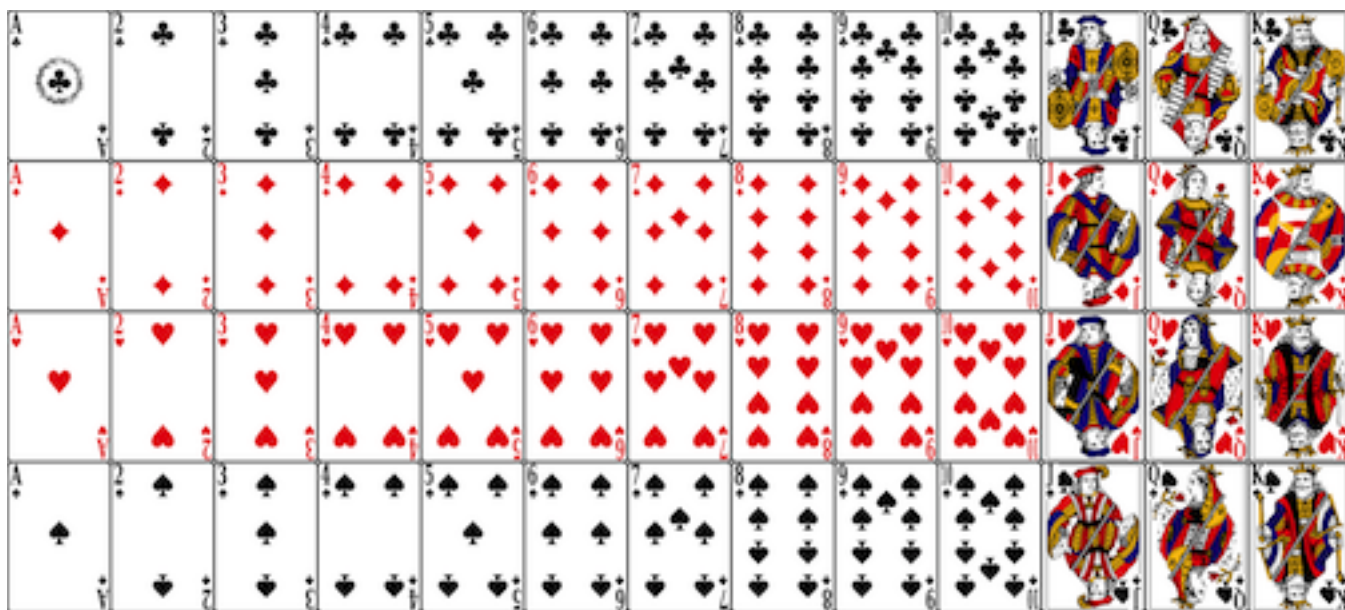


Statistics	Number	Expected #	Percentage	Expected %
Nothing	493	501.5699	49.3000	50.1570
One Pair	431	422.5690	43.1000	42.2569
Two Pair	41	47.5390	4.1000	4.7539
3 of a Kind	19	21.1285	1.9000	2.1128
Straight	8	3.5322	0.8000	0.3532
Flush	1	1.9669	0.1000	0.1967
Full House	4	1.4406	0.4000	0.1441
4 of a Kind	3	0.2401	0.3000	0.0240
Straight Flush		0.0123		0.0012
Royal Flush		0.0015		0.0002
Total	1000	1000	100.0000	100.0000

Of the 1000 hands, 431 had one pair. So we'd guess that the chance of getting one pair is about 43.1%. (The actual theoretical chance is 42.26%, what you'd get if you drew five cards an infinite number of times.)

6.1 Drawing from a box

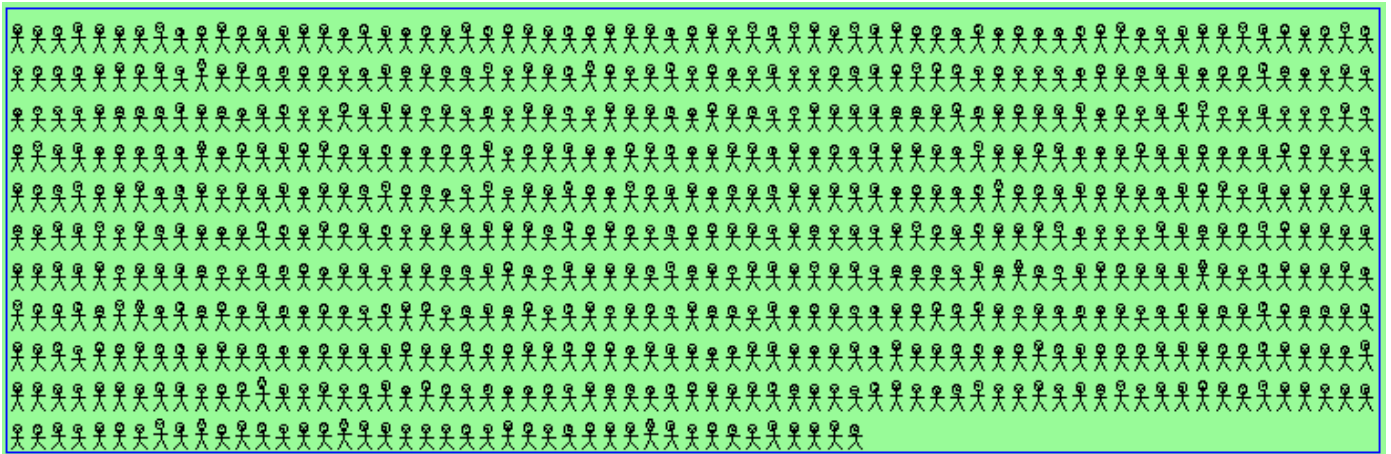
Think of a box with a number of objects in it. Randomly draw one of the objects from the box, in such a way that each object has the same chance of being drawn. What's in the box? It could be a regular deck of cards.



There are 52 cards in the deck. There are four suits, with 13 cards in each suit. The first row in the picture has the clubs, the second has the diamonds, then the hearts, and the last row has the spades.

Within each suit, there is an ace (with just one symbol), then 2, 3, through 10, then the jack, queen and king.

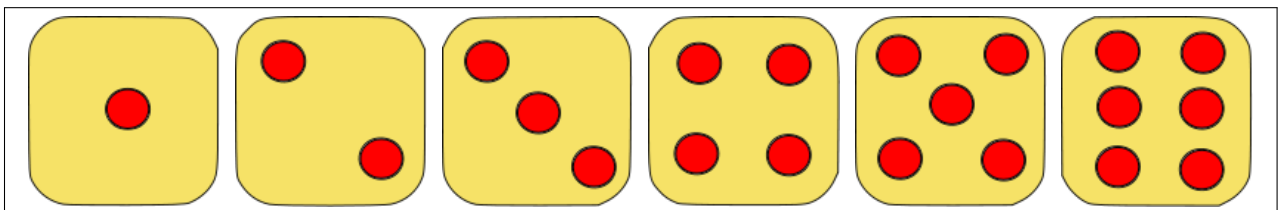
Or the box could contain people:



When considering regular six-sided dice,



the box may include the possible sides of one die:



Usually, the box will have tickets with numbers on them. Here are the numbers of pairs of shoes for 60 people:

30	40	9	4	20	6	12	2	10	13
5	10	25	6	5	5	40	29	22	16
13	25	7	12	2	10	30	15	25	4
21	15	40	15	16	10	7	10	6	5
16	4	7	10	3	8	100	4	12	12
12	25	20	4	4	3	40	8	19	14

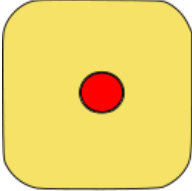
6.2 Calculating chances

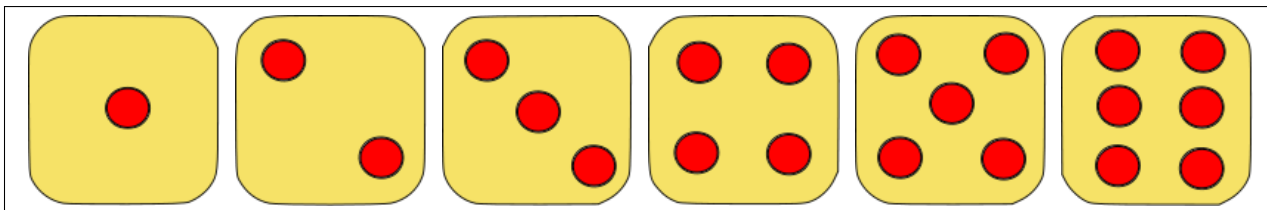
The chance of drawing a particular type of object from the box when drawing so that each object has the same chance of being drawn is the fraction of objects of that type:

$$\text{Chance of drawing particular type} = \frac{\# \text{ of objects of that type}}{\text{Total } \# \text{ of objects in the box}}.$$

A die

The chance of rolling a die and getting a “1” is the same as the chance

of drawing the  from the box





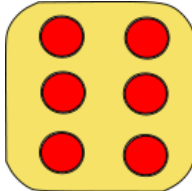
There is one 1, and six objects all together in the box:

$$\text{Chance of drawing a 1} = \frac{\# \text{ of 1's}}{\text{Total \# of objects in the box}} = \frac{1}{6}.$$

We often turn that fraction into a percentage:

$$100 \times \frac{1}{6} = 16.67\%.$$

The chance of rolling an even number is the chance of drawing the

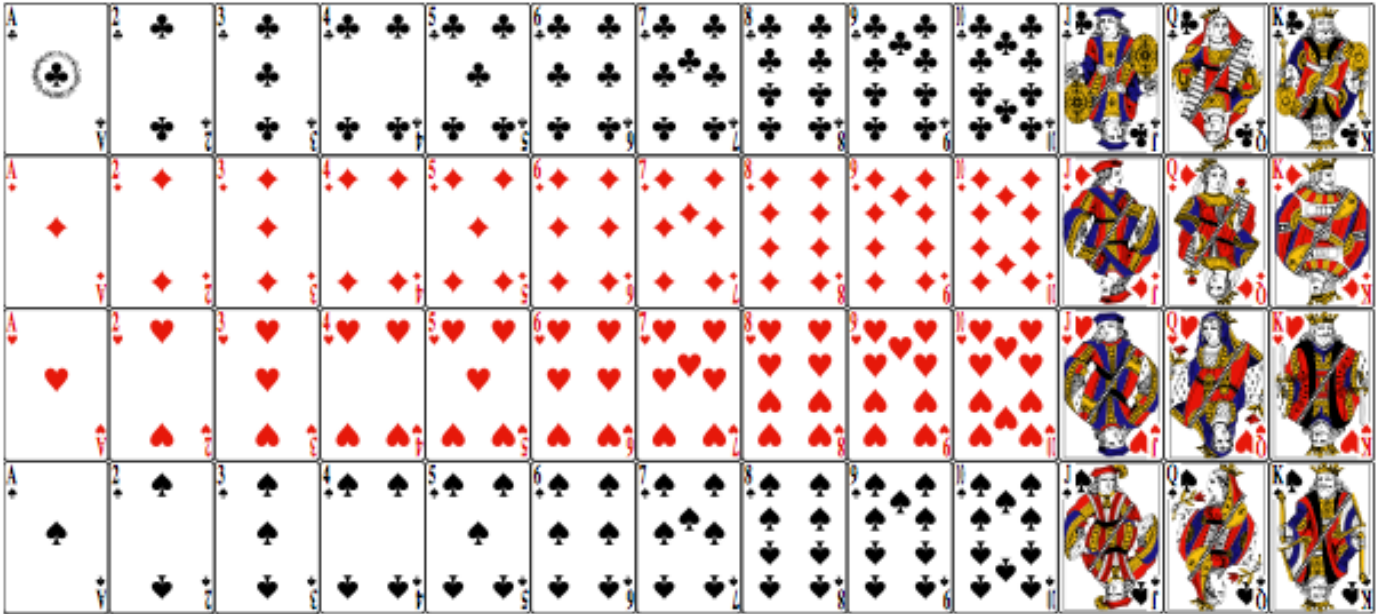
 or  or  from the same box. There are three even numbers (2,4,6), and six objects all together in the box:

$$\begin{aligned} \text{Chance of drawing an even number} &= \frac{\# \text{ of even numbers}}{\text{Total \# of objects in the box}} \\ &= \frac{3}{6} = \frac{1}{2}. \end{aligned}$$

Percentage:

$$100 \times \frac{1}{2} = 50\%.$$

Cards



What is the chance of drawing an ace from a deck of 52 cards? There are four aces (in the first column), and 52 cards total:

$$\begin{aligned}\text{Chance of drawing an ace} &= \frac{\# \text{ of aces}}{\text{Total } \# \text{ of objects in the box}} \\ &= \frac{4}{52} = \frac{1}{13} \rightarrow 7.69\%.\end{aligned}$$

What is the chance of drawing a red card? There are 26 red cards (the diamonds and hearts, in the middle two rows), and 52 cards total:

$$\begin{aligned}\text{Chance of drawing a red} &= \frac{\# \text{ of reds}}{\text{Total } \# \text{ of objects in the box}} \\ &= \frac{26}{52} = \frac{1}{2} \rightarrow 50\%.\end{aligned}$$

? How many face cards are there? (A face card is a card with a face on it. It could be a jack or queen or king.) What is the chance of drawing a face card, as a fraction? As a percentage?

Drawing more than one

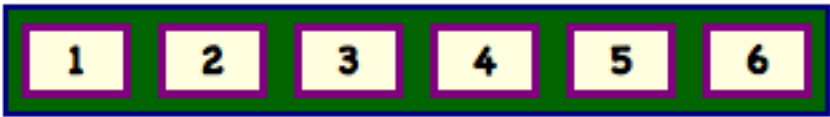
Generally, one makes more than one draw from the box. For example:

- Rolling two dice: It's like drawing twice from the box of the six die sides.
- Being dealt a five-card hand: It's like drawing five cards from the box of cards.
- Taking a random sample for a poll: It's like drawing several people from a box with many people.

There are two ways to draw several objects. Do you put back the object before you draw the next one? Or do you keep it out?

- Drawing **with replacement** means after each draw, you return the object you chose. Then it is possible to draw the same object more than once.
- Drawing **without replacement**. You do not return the drawn objects to the box. Then you have a different set of objects to draw from each time.

Suppose you wish to draw two tickets from the box



If you are drawing **with replacement**, you use the process illustrated by going down the first column below:

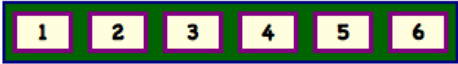
Drawing 2 with replacement	Drawing 2 without replacement
First, draw one ticket (it's a 6). 	First, draw one ticket (it's a 5).
Then replace it (put it back in the box). 	Don't replace it.
Then draw another ticket (it's a 4). 	Draw another ticket from what's left. It's a 1.
Then replace that one 	Don't replace it.
You've drawn a 4 and a 6, drawing with replacement.	You've drawn a 1 and a 5, drawing without replacement.

If you are drawing **without replacement**, you use the process illustrated in the second column.

Multiplying the chances

It is important to be aware of whether you are drawing with or without replacement. Drawing with replacement: Each draw has the same chances of getting each object. Drawing without replacement: The chances on the second, third, ..., draws change, depending on what happened before.

- Rolling two dice: You would draw with replacement. What you get on one roll does not affect the next roll.
- Being dealt a five-card hand: You would draw without replacement. You don't put the first card back before getting the second.
- Taking a random sample for a poll: You'd draw without replacement. There's no point interviewing the same person twice.

The chance that two things happen is the chance the first happens times the chance the second happens given the first happens. Specifically, draw two from the box  **with replacement**. What is the chance you get two 6's?

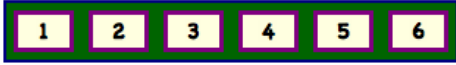
$$\left(\begin{array}{c} \text{Chance get 6 on} \\ \text{first draw and 6 on} \\ \text{second draw} \end{array} \right) = \left(\begin{array}{c} \text{Chance get} \\ \text{6 on first} \\ \text{draw} \end{array} \right) \times \left(\begin{array}{c} \text{Chance get 6 on} \\ \text{second draw} \\ \text{given} \\ \text{get 6 on first draw} \end{array} \right)$$

The key word in the last part of the equation is **given**. The part after that word has to be taken into account in order to determine what the box looks like on the second draw.

In particular, what the box looks like at that point depends on whether you are drawing with or without replacement.

Drawing with replacement

We wish to draw two tickets **with replacement** from the box



. What is the chance that both draws are 6?

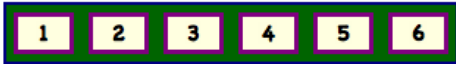
The multiplication formula is

$$\begin{pmatrix} \text{Chance get 6 on} \\ \text{first draw and 6 on} \\ \text{second draw} \end{pmatrix} = \begin{pmatrix} \text{Chance get} \\ \text{6 on first} \\ \text{draw} \end{pmatrix} \times \begin{pmatrix} \text{Chance get 6 on} \\ \text{second draw} \\ \text{given} \\ \text{get 6 on first draw} \end{pmatrix}$$

We know that

$$\text{Chance get 6 on first draw} = \frac{1}{6}.$$

Given we got a 6 on the first draw means we draw a 6, but then since we are drawing **with** replacement, we put it back before the second draw. So the box for the second draw looks the same:



. Then the chance of getting a 6 out of that box is again $1/6$:

$$\text{Chance get 6 on second draw given got 6 on first draw} = \frac{1}{6}.$$

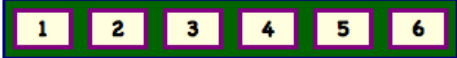
Multiply those chances:

$$\text{Chance get 6 on first draw and 6 on second draw} = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

In percentage terms: $100 \times \frac{1}{36} = 2.78\%$. A fairly small chance.¹

¹ Note: It is important to do the multiplication with the fractions, not the percentages. At the very end of the calculations is when you can find the percentage.


Drawing without replacement

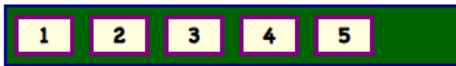
Now draw two from the box  **without replacement**. What is the chance you get two 6's? The formula is the same:

$$\left(\begin{array}{c} \text{Chance get 6 on} \\ \text{first draw and 6 on} \\ \text{second draw} \end{array} \right) = \left(\begin{array}{c} \text{Chance get} \\ \text{6 on first} \\ \text{draw} \end{array} \right) \times \left(\begin{array}{c} \text{Chance get 6 on} \\ \text{second draw} \\ \text{given} \\ \text{get 6 on first draw} \end{array} \right)$$

And as before,

$$\text{Chance get 6 on first draw} = \frac{1}{6}.$$

But now think what the box looks like **given** that the first draw was a 6. Since we are drawing **without replacement**, we've drawn the , but haven't put it back. So the box for the second draw is



Out of those five tickets, none are 6. Thus

$$\text{Chance get 6 on second draw given got 6 on first draw} = \frac{0}{5} = 0.$$

Multiply those chances:

$$\text{Chance get 6 on first draw and 6 on second draw} = \frac{1}{6} \times \frac{0}{5} = 0.$$

There's no chance to get two 6's if you do not replace the first one. Which makes sense.

So the chance of getting two 6's is 2.78% if drawing **with replacement**, and 0% if drawing **without replacement**.

? Again consider two draws from the box

1	2	3	4	5	6
---	---	---	---	---	---

. Now you are interested in the chance you get a 1 on the first draw, and a 2 on the second.

Draw the two **with replacement**. What is the chance of getting a 1 on the first draw?

What is the box for the second draw?

What is the chance of getting a 2 on the second draw **given** you got a 1 on the first draw?

What is the chance of getting a 1 on the first draw and a 2 on the second? (First find the answer as a fraction, then turn that into a percentage.)

How does it compare with the chance of getting two 6's drawing with replacement?

? Keep the same setup: Draw two from the box .

We are interested in the chance you get a 1 on the first draw, and a 2 on the second, now we draw the two **without replacement**.

What is the chance of getting a 1 on the first draw?

What is the box for the second draw?

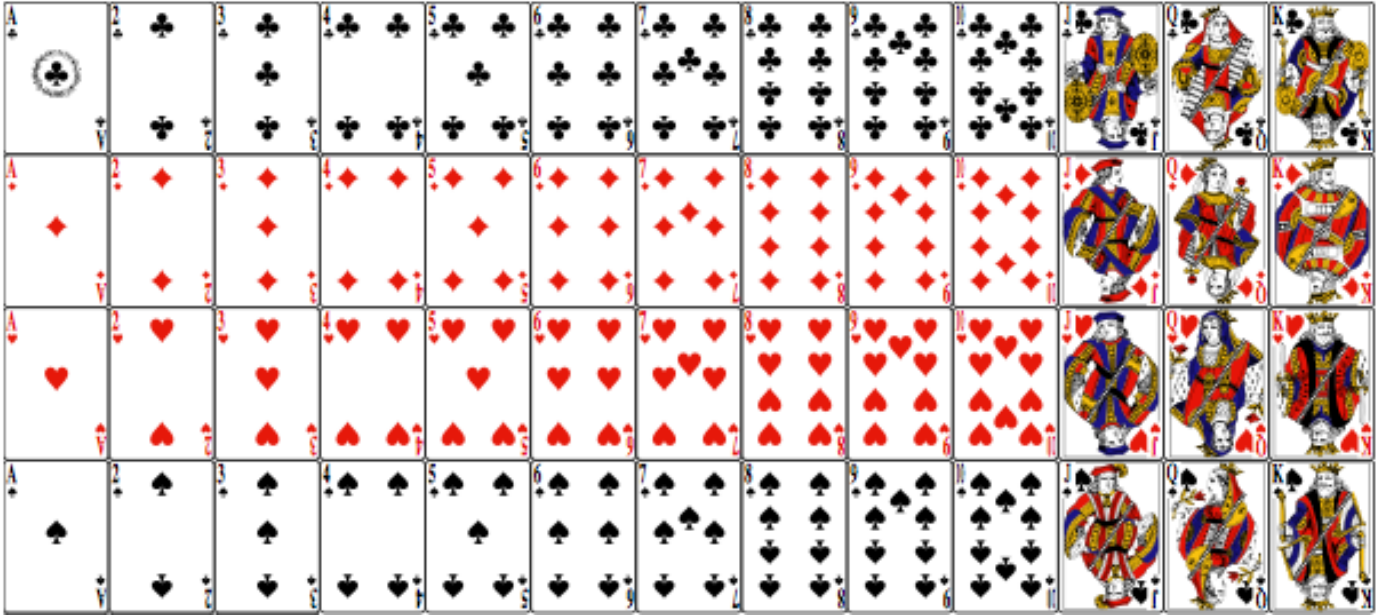
What is the chance of getting a 2 on the second draw given you got a 1 on the first draw?

What is the chance of getting a 1 on the first draw and a 2 on the second? (Again, get the answer as a fraction, then percentage.)

How does it compare with the chance of getting two 6's drawing without replacement?

Cards

If you are dealt two cards, what is the chance that both are hearts?
We are drawing two **without replacement** from



The formula is similar to before:

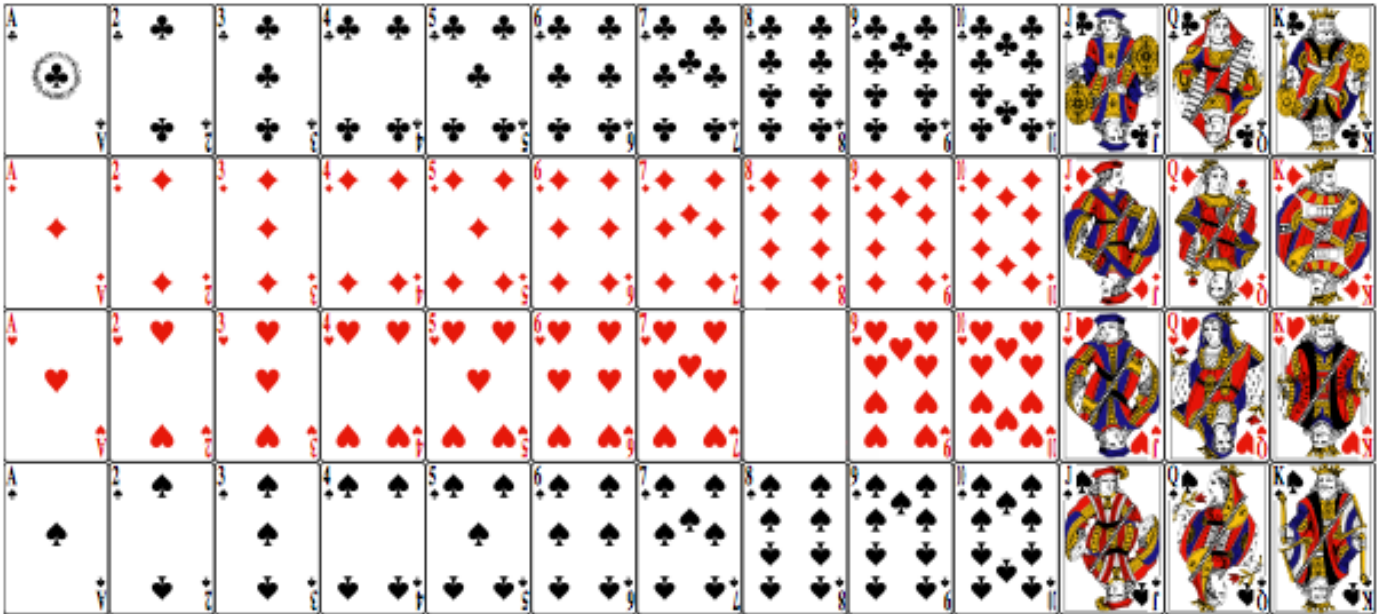
$$\left(\begin{array}{c} \text{Chance get heart} \\ \text{on first draw and} \\ \text{heart on second} \\ \text{draw} \end{array} \right) = \left(\begin{array}{c} \text{Chance get} \\ \text{heart on first} \\ \text{draw} \end{array} \right) \times \left(\begin{array}{c} \text{Chance get heart} \\ \text{on second draw} \\ \textbf{given} \text{ get heart on} \\ \text{first draw} \end{array} \right)$$

Since the box has 13 hearts and 52 cards total,

$$\text{Chance of heart on first draw} = \frac{13}{52} = \frac{1}{4}.$$

For the second draw, since we are drawing **without replacement**, we need to take into account that the first draw took out a heart and did not replace it. So we are drawing from a box with a heart missing.

The box without one of the hearts may then look like



The picture shows the 8 of hearts gone, but it could be any one of them. In any case, there are only 12 hearts left. Note also there are only 51 cards left as well. So

$$\text{Chance of heart on second given heart on first} = \frac{12}{51} = \frac{4}{17}.$$

Then recalling that the chance of getting a heart on the first draw was $1/4$:

$$\begin{aligned} \left(\begin{array}{l} \text{Chance get heart} \\ \text{on first draw and} \\ \text{heart on second} \\ \text{draw} \end{array} \right) &= \left(\begin{array}{l} \text{Chance get} \\ \text{heart on first} \\ \text{draw} \end{array} \right) \times \left(\begin{array}{l} \text{Chance get heart} \\ \text{on second draw} \\ \textbf{given} \text{ get heart on} \\ \text{first draw} \end{array} \right) \\ &= \frac{1}{4} \times \frac{4}{17} \\ &= \frac{1}{17} \\ &\rightarrow \text{Percentage: } 100 \times \frac{1}{17} = 5.88\% \end{aligned}$$

So about a 6% chance of getting two hearts with two draws.

Drawing five cards

What about the chance that in a five-card hand, all five are hearts? We draw five without replacement. Now we have a whole string of chances to multiply. For each draw, we have to figure out what the box is given the previous draws.

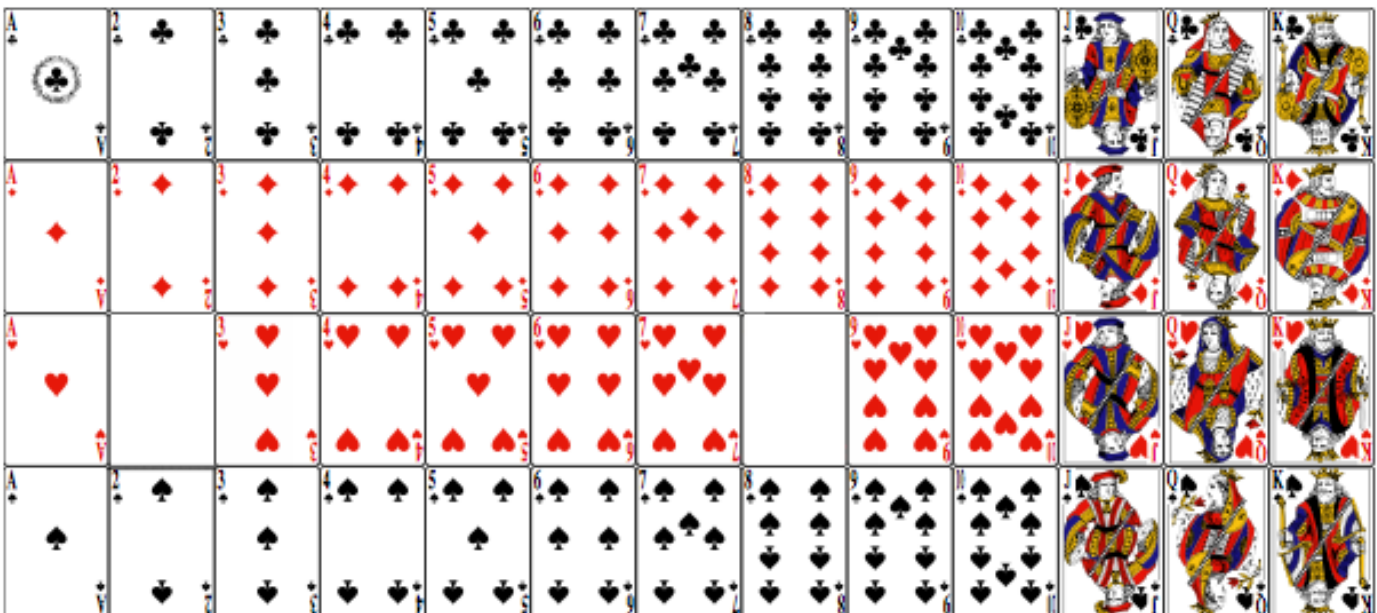
$$\begin{aligned} \text{Chance all five are hearts} = & (\text{Chance first is heart}) \times \\ & (\text{Chance second is heart given first is heart}) \times \\ & (\text{Chance third is heart given first two are hearts}) \times \\ & (\text{Chance fourth is heart given first three are hearts}) \times \\ & (\text{Chance fifth is heart given first four are hearts}) \end{aligned}$$

From before,

$$\text{Chance first is heart} = \frac{13}{52} \text{ and}$$

$$\text{Chance second is heart given first is heart} = \frac{12}{51}.$$

For the third draw, we have already taken two hearts from the box so there are only 11 hearts left, and only 50 cards left. For example, if the first two draws were the 2 and 8 of hearts, the box would look like



So, again, after drawing two hearts, there are 11 hearts left, out of 50 cards total left, so that

$$\text{Chance third is heart given first two are hearts} = \frac{11}{50}.$$

Each time you lose one heart and one card:

$$\text{Chance fourth is heart given first three are hearts} = \frac{10}{49}, \quad \text{and}$$

$$\text{Chance fifth is heart given first four are hearts} = \frac{9}{48}.$$

Put it all together:

$$\begin{aligned} \text{Chance all five are hearts} &= (\text{Chance first is heart}) \times \\ &\quad (\text{Chance second is heart given first is heart}) \times \\ &\quad (\text{Chance third is heart given first two are hearts}) \times \\ &\quad (\text{Chance fourth is heart given first three are hearts}) \times \\ &\quad (\text{Chance fifth is heart given first four are hearts}) \\ &= \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48} \\ &= 0.0004952. \end{aligned}$$

Turning the answers into a percentage:

$$100 \times 0.0004952 = 0.04952\%.$$

That is a very small chance, much less than 1 percent. Chances are about 5 out of ten thousand.

? Draw two cards from a regular deck without replacement. The goal is to find the chance both are aces. What is the chance the first is an ace?

What is in the box after the first draw, given that it was an ace? (How many tickets, and how many are aces?)

What is the chance of drawing an ace from that new box?

What is the chance both are aces, as a fraction? As a percentage?

? Now draw three without replacement. What is the chance all three are aces?

Now draw five without replacement. What is the chance all five are aces?

6.3 Conditional chance

Cards

Consider the question,

What is the chance of getting a heart on the second draw given you got a heart on the first draw?

The word “given” in the question means that it is asking for a **conditional** chance. A conditional chance is one for which the box may have been changed. The possible change is described by the “given” phrase. If we are drawing without replacement, then “given we got a heart on the first draw” means the box is missing one heart (as on page 177).

Other conditions are possible. In drawing one card, what is the chance of drawing a king **given** it is a face card? A face card is one with a face: A king, queen, or jack. Here are the face cards: There are 12 of them.

The phrase “given it is a face card” means you are drawing from the box with just the 12 face cards. Notice that 4 of them are kings.



So

$$\text{Chance of drawing a king given it is a face card} = \frac{4}{12} = \frac{1}{3}.$$

People

Here are 712 people:



We categorize them by gender and whether they are 72 inches and taller, or shorter than 72 inches.

	Shorter than 72 inches	72 inches or taller
Male	131	96
Female	478	7

So 7 women are 72 inches or taller, 131 men are shorter than 72 inches, etc.

If you choose one person from this box at random, what is the chance it will be a male? There are $131 + 96 = 227$ males, and 712 people all together. Thus

$$\text{Chance of drawing male} = \frac{227}{712} = 0.32.$$

So there is a 32% chance of drawing a male.

If you choose one person from this box at random, what is the chance the person will be 72 inches or taller? There are $96 + 7 = 103$ tall people, and again 712 people all together. Thus

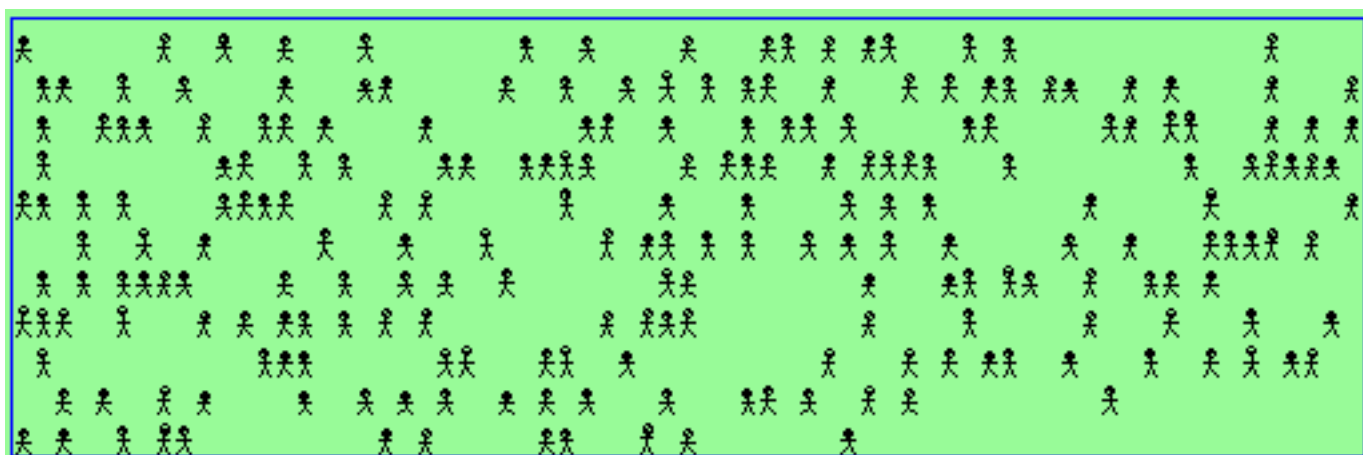
$$\text{Chance of drawing someone 72 inches or taller} = \frac{103}{712} = 0.14.$$

So a 14% chance of being that tall. Fairly small.

But if we choose a random male, what is the chance he is 72 inches or taller? This question is

What is the chance the person is 72 inches or taller **given** he is a male?

Now there is a new box to consider: “Given he is a male” means the box has just the 227 males in it:



Here is the table again:

	Shorter than 72 inches	72 inches or taller
Male	131	96
Female	478	7

Draw one from this new box. Since it just has males, the number of people 72 inches or taller is 96, and there are 227 males. So

$$\text{Chance 72 inches or taller given male} = \frac{96}{227} = 0.42.$$

Thus 42% of the men are tall, compared with only 14% of the overall class.

	Shorter than 72 inches	72 inches or taller
Male	131	96
Female	478	7

? What is the chance of being 72 inches or taller given female? First, how many females are there all together?

What does the box under consideration have in it now?

How many of the females are 72 inches or taller?

What is the chance (as a fraction) of drawing someone who is 72 inches or taller given that the person is female?

Give the answer in percentage terms.

Compare the percentage to that for the men. Which group has a higher percentage of people 72 inches or taller?

6.4 Independence

In statistics, two possibilities are **independent** if the chance the first happens is the same whether or not the second happens. Drawing from the box of 712 people, a possible question is

Is being 72 inches or taller independent of being male?



The two possibilities are

1. Being 72 inches or taller
2. Being male.

Independence here means that chance of drawing a tall person from the males is the same as drawing a tall person from the not-males (the females). That is, knowing the gender of a person does not give you any information about whether the person would be 72 inches and taller or not. [Intuitively, you should expect that being male and being tall are **not** independent. If you were told a random person was 72 inches or taller, you'd probably guess the person was male.]

The technical way to decide whether the two possibilities are independent or not is to find two conditional chances:

1. The chance of being 72 inches or taller **given** male;
2. The chance of being 72 inches or taller **given** not male.

From the previous two pages, we know these chances:

Chance of being 72 inches or taller given male = 42%;
Chance of being 72 inches or taller given not male = 1.4%.

They are not the same. So being tall is **not** independent of being male: Height and gender are **dependent**.

? For the following pairs of possibilities, use your intuition to decide whether you think they could be independent or not (circle your choice):

- The outside temperature; the month:
Could be independent Would guess they are not independent
- Gender; eye color:
Could be independent Would guess they are not independent
- Living in a city; being Democrat:
Could be independent Would guess they are not independent
- Get the polio vaccine; get polio:
Could be independent Would guess they are not independent
- Draw one card — Get a heart; get a red card:
Could be independent Would guess they are not independent
- Draw one card — Get an ace; get a heart:
Could be independent Would guess they are not independent
- Gender; being a freshman:
Could be independent Would guess they are not independent
- Year in college; weight:
Could be independent Would guess they are not independent
- Year in college; height:
Could be independent Would guess they are not independent

The rule for independence

To figure out whether two possibilities are independent, you have to figure out two conditional chances:

1. The chance that the first happens given the second **does** happen;
2. The chance that the first happens given the second **does not** happen.

If

$$\left(\begin{array}{l} \text{The chance that the first hap-} \\ \text{pens given the second } \mathbf{does} \\ \text{happen} \end{array} \right) = \left(\begin{array}{l} \text{The chance that the first hap-} \\ \text{pens given the second } \mathbf{does} \\ \mathbf{not} \text{ happen} \end{array} \right)$$

then

The two possibilities **are** independent.

If

$$\left(\begin{array}{l} \text{The chance that the first hap-} \\ \text{pens given the second } \mathbf{does} \\ \text{happen} \end{array} \right) \neq \left(\begin{array}{l} \text{The chance that the first hap-} \\ \text{pens given the second } \mathbf{does} \\ \mathbf{not} \text{ happen} \end{array} \right)$$

then

The two possibilities **are not** independent.

Freshman and male?

Is being a freshman independent of being male? The two possibilities are then

1. Being a freshman;
2. Being male.

We need a new table:

	Not freshman	Freshman
Male	152	75
Female	328	157

So the two conditional chances we need are

1. Chance of being a freshman given male;
2. Chance of being a freshman given not male.

We can figure those out:

$$\begin{aligned} \text{Chance of being a freshman given male} &= \frac{75}{152 + 75} \\ &= \frac{75}{227} = 0.33 \rightarrow 33\%; \end{aligned}$$

$$\begin{aligned} \text{Chance of being a freshman given not male} &= \frac{157}{328 + 157} \\ &= \frac{157}{485} = 0.32 \rightarrow 32\%. \end{aligned}$$

Those two chances are **not** equal. So, technically, we'd have to say "freshman" and "male" are **not independent**. (But they are very close to independent.)

? The members of a class were asked whether they were happy that day, and whether they had gotten enough sleep the night before. The questions is whether, for this class, being happy and getting enough sleep are independent or not.

	Got enough sleep	Didn't get enough sleep
Happy	15	5
Not happy	60	20

What are the two possibilities in this question?

What are the two conditional chances we need to calculate?

Calculate the two conditional chances.

Are the two conditional chances equal?

Is being happy and getting enough sleep independent for this class?

? The members of a second class were also asked whether they were happy that day, and whether they had gotten enough sleep the night before. The questions is whether, for this class, being happy and getting enough sleep are independent or not.

Other class	Got enough sleep	Didn't get enough sleep
Happy	40	10
Not happy	20	15

What are the two possibilities in this question?

What are the two conditional chances we need to calculate?

Calculate the two conditional chances.

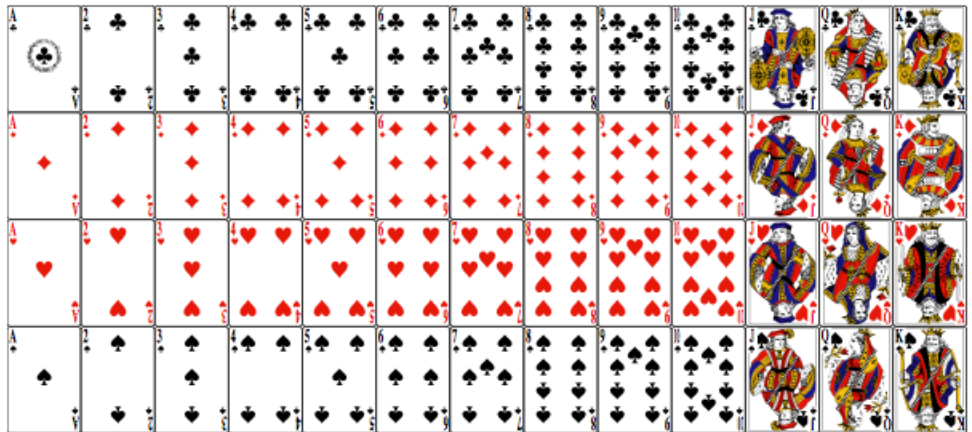
Are the two conditional chances equal?

Is being happy and getting enough sleep independent for this class?

Ace and heart?

Draw one card from the deck of 52 cards. Is getting an ace independent of getting a heart? The two possibilities are

1. Getting an ace;
2. Getting a heart.



The two conditional chances are

1. Chance of getting an ace **given** we got a heart.
2. Chance of getting an ace **given** we did not get a heart.

For the first chance, we are given that we got a heart, so we are choosing from just the 13 hearts. Among those, there is only one ace (the ace of hearts, of course). So

$$\text{Chance of ace given heart} = \frac{1}{13}.$$

For the second chance, we are given we did **not** get a heart. That means it was one of the other suits (clubs, diamonds, spades), of which there are 39 cards. Of those, three are aces. So

$$\text{Chance of ace given not heart} = \frac{3}{39} = \frac{1}{13}.$$

Those two conditional chances **are** equal, so we do have independence: Getting an ace **is** independent of getting a heart. That is, knowing whether a card is a heart or not does not help in guessing whether it is an ace or not.

Drawing with replacement

Draw two **with replacement** from

1	2	3	4	5	6
---	---	---	---	---	---

. Is getting a 6 on the first draw independent of getting a 6 on the second?

Here are the two possibilities:

1. Getting a 6 on the second draw;
2. Getting a 6 on the first draw.

Notice we put the second draw first: That is because it is easier to find what happens on the second draw given the first.

And the two conditional chances:

1. Chance of getting a 6 on the second draw **given** we got a 6 on the first draw;
2. Chance of getting a 6 on the second draw **given** we did not get a 6 on the first draw.

What does the box look like for the second draw given we got a 6 on the first draw? Since we are drawing **with replacement**, it has all the tickets in it. The 6 was replaced:

1	2	3	4	5	6
---	---	---	---	---	---

. So the chance of getting a 6 from that box is $1/6$:

$$\text{Chance 6 on second given 6 on first} = \frac{1}{6}.$$

But what if we drew a non-6 on the first draw. Say a 2. We put it back before the next draw, so again the box is

1	2	3	4	5	6
---	---	---	---	---	---


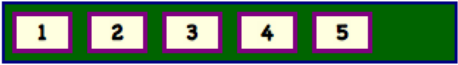
. No matter what we drew the first time, putting it back means we have the full box for the second draw.

$$\text{Chance 6 on second given not 6 on first} = \frac{1}{6}.$$


So it is $1/6$ either way: The draws **are** independent.

In fact, whenever you draw with replacement, the draws are independent.

Drawing without replacement

Now draw two **without replacement** from  Is getting a 6 on the first draw independent of getting a 6 on the second? What is the chance of a 6 on the second draw given a 6 on the first? Since we draw without replacement, after drawing the 6 first, we have the box  for the second draw. So

$$\text{Chance 6 on second given 6 on first} = \frac{0}{5} = 0.$$

What about the chance of a 6 on the second given a "not 6" on the first? The first draw must have been one of 1, 2, 3, 4, or 5. Whichever, it was not put back. So, e.g., if the first was a 4, the box for the second draw is  There are five tickets, one with a six. (The same as if 1, 2, 3, or 5 was chosen first.) So

$$\text{Chance 6 on second given not 6 on first} = \frac{1}{5}.$$

Si204nce

$$0 \neq \frac{1}{5},$$

the draws are **not independent**.

In fact, whenever you draw without replacement, the draws are not independent.

Independence: Drawing with or without replacement

When drawing several objects from a box,

- The draws **are** independent if you draw **with replacement**;
- The draws **are not** independent if you draw **without replacement**.

Because when you replace the object, what you get on one draw does not affect what you get on the next draw. If you do not replace it, what you get on one draw does affect what you get on the next draw.

Multiplying chances

If things are independent, then you can find the chance they all happen by multiplying their individual chances.

For example: Flip a fair coin 5 times independently. What is the chance you get five heads?

First, figure out the box: It has two tickets: Heads and Tails. We'll code these

Heads \rightarrow 1 and Tails \rightarrow 0 :



So the chance of flipping a Head is the chance of drawing a 1 from the box $= \frac{1}{2}$.

Flipping independently = drawing with replacement, thus the chance of getting HHHHH is

$$\begin{aligned}
 \text{Chance of five 1's} &= \text{Chance of 1 on first draw} \\
 &\quad \times \text{Chance of 1 on second draw} \\
 &\quad \times \text{Chance of 1 on third draw} \\
 &\quad \times \text{Chance of 1 on fourth draw} \\
 &\quad \times \text{Chance of 1 on fifth draw} \\
 &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \\
 &= \frac{1}{32}.
 \end{aligned}$$

That is about 3%.

A population

Imagine a population of 1000 people. 400 are male, 600 are female. Also, 100 of the people are left-handed. Thus if we pick one person at random

$$\text{Chance Male} = \frac{400}{1000} = \frac{2}{5},$$

and

$$\text{Chance Left-handed} = \frac{100}{1000} = \frac{1}{10}.$$

What is the chance that the person is male and left-handed? We need more information.

Now suppose we know that being male and being left-handed are **independent**. Then we can multiply the probabilities:

$$\begin{aligned} \left(\begin{array}{c} \text{Chance Male and} \\ \text{Left-handed} \end{array} \right) &= (\text{Chance Male}) \times (\text{Chance Left-handed}) \\ &= \frac{2}{5} \times \frac{1}{10} = \frac{2}{50} = \frac{1}{25}. \end{aligned}$$

$$\text{Percentage: } 100 \times \frac{1}{25} = 4\%.$$

So 4% of the people are left-handed males. The table for this population is

	Left-handed	Right-handed
Male	40	360
Female	60	540

Robbery

The text has an account of a purse-snatching. Eye-witnesses testified that the perps were an interracial couple, the woman blond with a ponytail, the man African American with a mustache and beard. They escaped in a yellow car. (All these features are alleged.) The prosecutor broke down the data, deciding the following probabilities were reasonable:

Feature	Chance
Yellow automobile	$\frac{1}{10}$
Man with mustache	$\frac{1}{4}$
Woman with ponytail	$\frac{1}{10}$
Woman with blond hair	$\frac{1}{3}$
Black man with beard	$\frac{1}{10}$
Interracial couple in car	$\frac{1}{1000}$

The couple on trial allegedly had all those features. The prosecutor then said that the chance all six of those features were present is found by multiplying the chance:

$$\begin{aligned}
 \text{Chance of all six features at once} &= \frac{1}{10} \times \frac{1}{4} \times \frac{1}{10} \times \frac{1}{3} \times \frac{1}{10} \times \frac{1}{1000} \\
 &= \frac{1}{12,000,000}.
 \end{aligned}$$

That's one out of twelve million. So they must have been the ones that committed the crime?

There are a number of problems with this argument. Where did those probabilities come from? And there are lots of people; 1 out of 12 million is low but its not impossible several couples have those features.

? But what was the worst crime statistically the prosecutor committed?

The worst crime statistically the prosecutor committed was

Multiplying the chances without having independence.

Here is the table again:

Feature	Chance
Yellow automobile	$\frac{1}{10}$
Man with mustache	$\frac{1}{4}$
Woman with ponytail	$\frac{1}{10}$
Woman with blond hair	$\frac{1}{3}$
Black man with beard	$\frac{1}{10}$
Interracial couple in car	$\frac{1}{1000}$

Having a beard is not independent of having a mustache.

Being an interracial couple is not independent of one person being black and one blond.

Being a couple in a car is not independent of having a yellow car.

? Is being blond independent of having a ponytail?

Moral

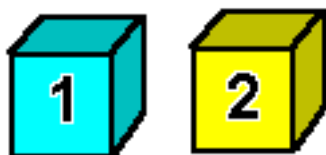
It is wrong to multiply the chances if you do not have independence.

6.5 Monty Hall

One of the boxes on the table below contains a new car!!!! You choose a box:



Suppose you chose box . Then there are two left on the table:



At least one of the boxes left on the table is empty (maybe both). Monty opens the empty box (randomly choosing one, if they are both empty). Suppose it is box #1 that he opens:



You get to choose whether to keep your box, or trade for the unopened one left (box #2) on the table.

? Should you keep? Trade? Or does it not matter?

Writing out the possibilities

If you pick the box with the car, then you should keep it. There's a

$$\frac{1}{3} = 33\%$$

chance you pick the car at first. So if you keep the box, you have a 33% chance of winning.

If you pick an empty box, then the car is in one of the two boxes on the table. Furthermore, Monty will open the empty box left on the table, so that the box that is left on the table has the car. So you should trade. There's a

$$\frac{2}{3} = 67\%$$

chance you pick an empty box at first. So if you trade, you have a 67% chance of winning.

We can write down all the possibilities in a table:

Box with Car	Box you choose	Box Monty opens	Box left on table	Box you get if you KEEP	Box you get if you TRADE
1	1	2 or 3	3 or 2	1 – WIN	3 or 2 – LOSE
1	2	3	1	2 – LOSE	1 – WIN
1	3	2	1	3 – LOSE	1 – WIN
2	1	3	2	1 – LOSE	2 – WIN
2	2	1 or 3	3 or 1	2 – WIN	3 or 1 – LOSE
2	3	1	2	3 – LOSE	2 – WIN
3	1	2	3	1 – LOSE	3 – WIN
3	2	1	3	2 – LOSE	
3	3	1 or 2	2 or 1		
Chance of winning				33%	67%

? Fill in the blanks (say what box you'd end up with, and whether you'd win or lose).

7.1 Drawing from a box

Many experiments can be modeled using the idea of drawing a number of objects from a box, then inspecting the draws. You have to decide

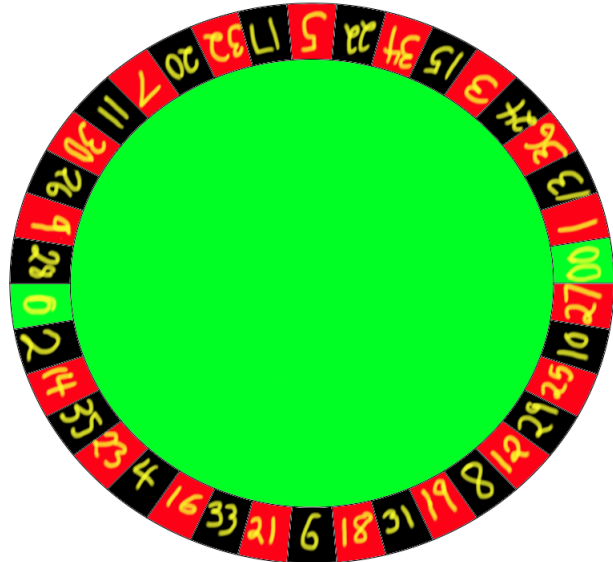
- What objects are in the box.
- How many draws to take.
- Whether the draws are taken with or without replacement.

Once you have the draws, you have to decide what to do with them. Some possibilities –

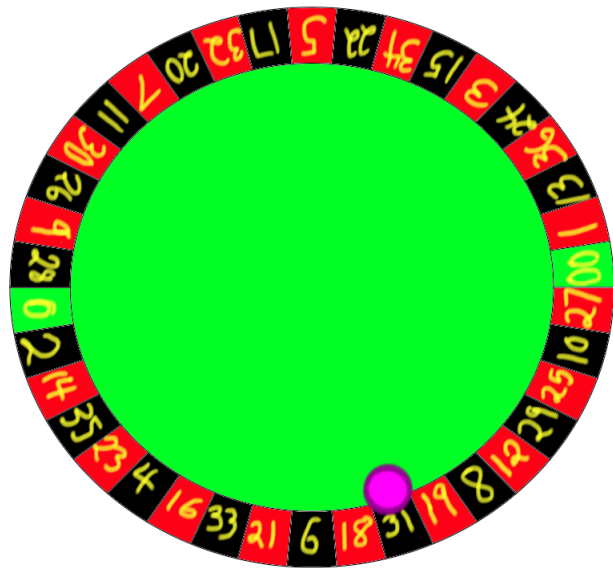
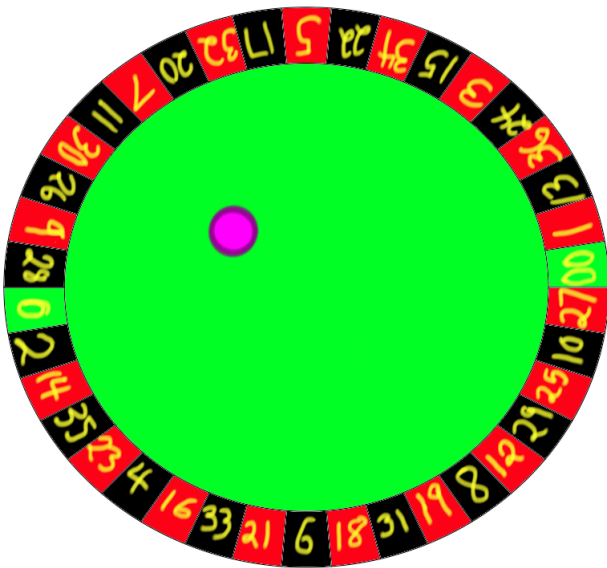
- Find the sum of the draws.
- Find the average of the draws.
- Find the percentage of something in the draws.

Roulette

The roulette wheel has the numbers from 1 to 36, as well as two zeroes: 0 and 00. The two zeroes are green. Of the other numbers half are red and half are black.




The wheel spins, and a ball gets knocked around.



Eventually the wheel stops spinning, and the ball settles in a slot of one of the numbers. Here it is on **31 Black**.

If you bet on 31, you win! 😊

Otherwise, you lose. 


Types of bets

There are a number of types of bet you can make. For example, if you want to bet on just the 7, you put your chip(s) on the 7 square. Then you win only if the 7 comes up.

		0 00		
1 to 18	1st 12	1	2	3
Even		4	5	6
		7	8	9
		10	11	12
Red	2nd 12	13	14	15
Black		16	17	18
		19	20	21
		22	23	24
Odd	3rd 12	25	26	27
19 to 36		28	29	30
		31	32	33
		34	35	36

If you want to bet on all the black numbers, then you put your chip where it says "Black." Then you win if any of the black numbers come up. There are 18 of them.

You can also bet on blocks of four numbers. For example, the numbers 16, 17, 19, 20 form a little square. You can bet on all four by placing your chip at the center of that square. You win if any of those four come up. There are many other possibilities.

		0 00		
1 to 18	1st 12	1	2	3
Even		4	5	6
		7	8	9
Red	2nd 12	10	11	12
		13	14	15
		16	17	18
 Black	2nd 12	19	20	21
22		23	24	
Odd		3rd 12	25	26
	28		29	30
	31		32	33
19 to 36	3rd 12	34	35	36

		0 00		
1 to 18	1st 12	1	2	3
Even		4	5	6
		7	8	9
		10	11	12
Red	2nd 12	13	14	15
Black		16	17	18
		19	20	21
		22	23	24
Odd	3rd 12	25	26	27
19 to 36		28	29	30
		31	32	33
		34	35	36

If you lose, you lose whatever you bet. If you win, how much you win depends on how many numbers you bet on. If you bet \$1 on a block of four numbers,

- You get \$8 if one of the 16, 17, 19 or 20 numbers comes up, 😊
- You get $-\$1$ if something else comes up. 😡

Notice the negative amount of money you get if you lose. That means you lose the \$1.

There are 38 numbers: 1, 2, ..., 36, and 0 and 00. So you'd have 4 chances to win, and $38 - 4 = 34$ chances to lose. Suppose you play this bet one time. What is the box? You want it to mimic how much money you end up with (plus or minus).

? Is the correct box

0	00	1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28
28	30	31	32	33	34	35	36		

or

-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	8	8	-1
8	8	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

?

Roulette Box Model

If you bet \$1 on a block of four numbers, you get \$8 if one of the 16, 17, 19 or 20 numbers comes up, you get $-\$1$ if something else comes up.

Suppose you play this bet one time. What is the box? You want it to mimic how much money you end up with (plus or minus). You have to translate each number into what you'd get if it comes up. So 16, 17, 19 and 20 translate to 8, everything else to -1. So the answer from the previous page is the second one:

-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	8	8	-1
8	8	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

Sum of the draws

If you play a particular bet in roulette a certain number of times, the total amount of money you end up with (plus or minus) is like the sum of that number of draws with replacement from a particular box.

Consider betting on the "1st 12" numbers, 1, 2, 3, ..., 12. Then

- You get \$2 if the roulette wheel ends up with 1, 2, 3, ..., or 12.
- You get $-\$1$ if it comes up anything else.

		0 00		
1 to 18	1st 12	1	2	3
		4	5	6
		7	8	9
Even		10	11	12
		13	14	15
		16	17	18
Red	2nd 12	19	20	21
		22	23	24
		25	26	27
Black		28	29	30
		31	32	33
		34	35	36
Odd	3rd 12			
19 to 36				

So there are 12 chances to get \$2's, and $38 - 12 = 26$ chances to get $-\$1$'s. To find the box representing what you get, you have to translate the numbers 1, 2, ..., 12 into "2," and everything else into -1 :

0	00	1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28
28	30	31	32	33	34	35	36		

→

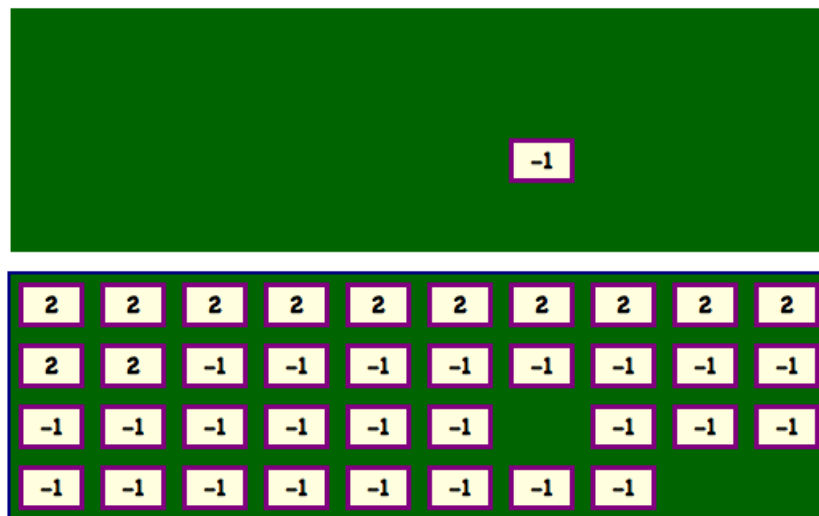
2	2	2	2	2	2	2	2	2	2
2	2	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

Note that the tickets are arranged so that the 2's come first. It doesn't matter the order in the box, just that there is the correct number of each type of ticket.

Draw 10 from the box

Now you play this bet 10 times, and see the total that you get. For the box model, that means you draw 10 tickets **with replacement** from the box. (Why “with replacement”?) Then you sum up the values on the tickets you drew.

So, the first draw may be a “−1:”



Write that down, then put the ticket back, and draw another, etc., until you’ve done it ten times. Here is a possible set of ten draws:

−1 2 −1 −1 2 −1 −1 −1 −1 −1

Add those up to get the **sum of the draws**, which is the amount of money you end up with (plus or minus):

$$\text{Sum of the draws} = -1 + 2 - 1 - 1 + 2 - 1 - 1 - 1 - 1 - 1 = -4.$$

Unfortunately, you get −\$4 (lose four dollars).

If you play another ten times, you’ll probably end up with a different amount.

The box again:

2	2	2	2	2	2	2	2	2	2
2	2	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

Now imagine that thirty people go to the roulette table. Ten people play this bet 10 times, ten people play it 100 times, and ten people play it 1000 times. What do they come away with? The table has their sums of draws. (So among the people who played ten times, one won \$5, one won \$2, one lost \$7, etc.)

Sum of 10 draws	Sum of 100 draws	Sum of 1000 draws
\$5, \$2, -\$7, -\$1, -\$1, -\$1, \$8, \$8, -\$4, \$11	\$5, \$11, -\$13, \$11, -\$19, -\$16, \$8, -\$13, \$5, \$8	-\$31, -\$73, -\$46, -\$67, -\$70, -\$127, -\$85, -\$115, \$14, -\$4

Some people won money, some lost. Add all those numbers to get -\$597. So these thirty people collectively lost almost \$600 to the casino.

What to notice:

- The more times people played, the larger the amounts they won or lost.
- The people who played 10 or 100 times were split about evenly between winners and losers.
- Almost all the people who played 1000 times lost money.
- The casino does well no matter what.

7.2 Expected values

Again, bet \$1 on the first twelve numbers. On any single bet, if you win, you get \$2, and if you lose, you get $-\$1$.

Thirty people go to the roulette table. Ten people play this bet 10 times, ten people play it 100 times, and ten people play it 1,000 times. The table shows their totals, i.e., their sums of draws.

Sum of 10 draws	Sum of 100 draws	Sum of 1,000 draws
\$5, \$2, $-\$7$, $-\$1$, $-\$1$, $-\$1$, \$8, \$8, $-\$4$, \$11	\$5, \$11, $-\$13$, \$11, $-\$19$, $-\$16$, \$8, $-\$13$, \$5, \$8	$-\$31$, $-\$73$, $-\$46$, $-\$67$, $-\$70$, $-\$127$, $-\$85$, $-\$115$, \$14, $-\$4$

Look more carefully at the sums:

- People who played 10 times: They seemed to end up about even (\$0) on average, plus or minus \$5 or \$10.
- People who played 100 times: They also seemed to end up about even, but plus or minus more like \$10 or \$15.
- People who played 1,000 times: They seemed end up with about $-\$50$ or $-\$60$ (losing) on average, plus or minus maybe another \$50.

We are interested first in figuring out what the average person would get. That average is called

The **expected value** of the sum of the draws.

The actual sums people get can vary widely from the expected value, just like people's weights can vary widely from the average weight.

The formula for the expected value

To find the expected value, you need to know what the box is, as well as how many draws you are taking. Start with drawing 10 from the box.

2	2	2	2	2	2	2	2	2	2
2	2	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

The formula is

$$\left(\begin{array}{c} \text{Expected value of} \\ \text{the sum of the} \\ \text{draws} \end{array} \right) = (\text{Number of draws}) \times (\text{Average in the box}).$$

The number of draws in this example is 10.

To find the “average in the box,” take the numbers written on the tickets in the box, and find their average, just as we did for data.

There are 38 tickets. We need to sum up the numbers on the tickets, then divide by 38. Be sure to add up all 12 of the 2’s, and all 26 of the -1’s:

$$\begin{aligned} \text{Average in the box} &= \frac{(2+2+2+2+2+2+2+2+2+2+2+2-1)}{38} \\ &= \frac{12 \times 2 + 26 \times (-1)}{38} = -\frac{2}{38} = -0.0526. \end{aligned}$$

So the average is -0.0526, or in dollars, about $-\$0.05 = -5$ cents.

What this means is that each time you play, you are expected to lose about a nickel. If you play 10 times, you expect to lose about 10 nickels, i.e., get $-\$0.50$. The more you play, the more you can expect to lose. Using the formula:

$$\begin{aligned} \left(\begin{array}{c} \text{Expected value of} \\ \text{the sum of the} \\ \text{draws} \end{array} \right) &= (\text{Number of draws}) \times (\text{Average in the box}) \\ &= 10 \times (-0.0526) = -0.526. \end{aligned}$$

That is about -53 cents. You expect to lose about 53 cents if you play the bet 10 times.

What about the expected value of the sum of 100 draws? The average in the box stays the same, but we replace the number of draws:

$$\begin{aligned} \left(\begin{array}{c} \text{Expected value of} \\ \text{the sum of the} \\ \text{draws} \end{array} \right) &= (\text{Number of draws}) \times (\text{Average in the box}) \\ &= 100 \times (-0.0526) = -5.26. \end{aligned}$$

That is about $-\$5.26$.

? Find the expected value of the sum of 1,000 draws.

What about the casino? How many draws were there all together? 11,100 draws. (Ten 10's, ten 100's and ten 1,000's). Find the expected total amount the 30 people together end up with.

To summarize:

	Sum of 10 draws	Sum of 100 draws	Sum of 1,000 draws
Actual sums	\$5, \$2, -\$7, -\$1, -\$1, -\$1, \$8, \$8, -\$4, \$11	\$5, \$11, -\$13, \$11, -\$19, -\$16, \$8, -\$13, \$5, \$8	-\$31, -\$73, -\$46, -\$67, -\$70, -\$127, -\$85, -\$115, \$14, -\$4
Expected Value	-\$0.526	-\$5.26	-\$52.60

The actual sums of the draws are around the expected values, plus or minus. You do not expect the actual sum of draws to be exactly the expected value. In fact, they can be quite variable. But on average they are near the expected values. Note that the more you play, the more you expect to lose.

For the casino, these people had 11,100 draws. So the expected amount for all 30 people together is

$$11100 \times (-0.0526) = -583.86.$$

They'd be expected to lose about \$584. It is good for the casino, anyway.

Plus or minus?

We don't just want to know the expected value, but also how variable the sums could be. For example, for 1,000 plays, one person lost \$127, and one won \$14, both numbers far from the expected value. So the amount people get is around

$$\text{Expected value} \pm \text{What?}$$

Finding the "what" is the next goal.

7.3 Standard errors

Stick with the roulette example, drawing with replacement from the box:

2	2	2	2	2	2	2	2	2	2
2	2	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

	Sum of 10 draws	Sum of 100 draws	Sum of 1,000 draws
Actual sums	\$5, \$2, -\$7, -\$1, -\$1, -\$1, \$8, \$8, -\$4, \$11	\$5, \$11, -\$13, \$11, -\$19, -\$16, \$8, -\$13, \$5, \$8	-\$31, -\$73, -\$46, -\$67, -\$70, -\$127, -\$85, -\$115, \$14, -\$4
Expected Value	-\$0.526	-\$5.26	-\$52.60
Off by about	$\pm 5?$ 10?	$\pm 10?$ 15?	$\pm 50?$

In the table, we see the expected values of the sums. Now we are interested in how far from the expected values the actual sums are likely to be. The last row estimates the \pm based on the amounts people got. Notice that the more draws, the bigger the \pm number.

The statistical term for this plus-or-minus number is the **standard error** of the sum of the draws. Next we calculate it.

The square root law

The standard error of the sum of the draws is related to the SD in the box and number of draws. The formula is called the square root law:

$$\left(\begin{array}{c} \text{Standard Error of} \\ \text{the sum of the} \\ \text{draws} \end{array} \right) = \sqrt{\text{Number of draws}} \times (\text{SD in the box}).$$

Before going into the actual calculations, compare this formula to that for the expected value:

$$\left(\begin{array}{c} \text{Expected Value of} \\ \text{the sum of the} \\ \text{draws} \end{array} \right) = (\text{Number of draws}) \times (\text{Average in the box}).$$

The more draws, the larger the expected value and the larger the standard error. But notice:

- The standard error has the **square root** of the number of draws in the formula.
- The standard error relies on the SD in the box. The next few pages we will explore that quantity.

SD in the box

Start with a simple box:

1	2	2	4	6
---	---	---	---	---

We find the SD in the box the usual way, as for data. (See page 76.)

Start with the average.

$$\text{Average in the box} = \frac{1 + 2 + 2 + 4 + 6}{5} = \frac{15}{5} = 3.$$

Then find the root mean square deviation: find the deviations, square them, find their average, and take the square root:

$$\begin{aligned} \text{SD in the box} &= \sqrt{\frac{(1-3)^2 + (2-3)^2 + (2-3)^2 + (4-3)^2 + (6-3)^2}{5}} \\ &= \sqrt{\frac{4 + 1 + 1 + 1 + 9}{5}} = \sqrt{\frac{16}{5}} = \sqrt{3.2} = 1.79. \end{aligned}$$

When the box has two types of tickets

This box has just 2's and -1's. There are twelve 2's and twenty-six -1's.

2	2	2	2	2	2	2	2	2	2
2	2	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

To find the SD the usual way, we first find the average. To start,

$$\begin{aligned} \text{Average in the box} &= \frac{2 + 2 + \cdots + 2 - 1 - 1 - \cdots - 1}{38} \\ &= \frac{12 \times 2 + 26 \times (-1)}{38} = \frac{24 - 26}{38} \\ &= \frac{-2}{38} = -0.0526. \end{aligned}$$

The "... " means there are a bunch of 2's and -1's.

Finding the SD in the box can be a bit tedious:

$$\begin{aligned} & \text{SD in the box} \\ &= \sqrt{\frac{(2-(-0.0526))^2 + \cdots + (2-(-0.0526))^2 + (-1-(-0.0526))^2 + \cdots + (-1-(-0.0526))^2}{38}} \end{aligned}$$

There are 12 of the $(2-(-0.0526))^2$'s and 26 of the $(-1-(-0.0526))^2$'s. Fortunately, when there are only two types of tickets in the box, there is an easier formula. The general formula for the SD in this kind of box is

$$\begin{aligned} & \text{SD in the box} \\ &= (\text{Bigger value} - \text{Smaller value}) \times \sqrt{\left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{bigger value} \end{array} \right) \times \left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{smaller value} \end{array} \right)}. \end{aligned}$$

For the box, there are 38 tickets total. The bigger value is 2, and there are twelve of them. The smaller value is -1, and there are 26 of them. Then

$$\left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{bigger value} \end{array} \right) = \frac{12}{38}, \quad \left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{smaller value} \end{array} \right) = \frac{26}{38}.$$

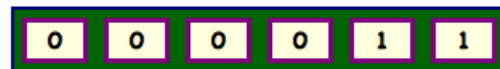
To find the SD in the box, use the formula:

$$\begin{aligned} & \text{SD in the box} \\ &= (\text{Bigger value} - \text{Smaller value}) \times \sqrt{\left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{bigger value} \end{array} \right) \times \left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{smaller value} \end{array} \right)} \\ &= (2 - (-1)) \times \sqrt{\frac{12}{38} \times \frac{26}{38}} = 3 \times \sqrt{0.2161} = 1.394 \end{aligned}$$

[Note: Be careful when you subtract a negative number: $2 - (-1) = 2 + 1 = 3$.]

Box with just 0's and 1's

Some boxes have just 0's and 1's, like this one:



There are 6 tickets total. The larger value is 1 and there are two of them. The smaller value is 0, and there are four of them. So

$$\left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{bigger value} \end{array} \right) = \frac{2}{6} = \frac{1}{3}, \quad \left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{smaller value} \end{array} \right) = \frac{4}{6} = \frac{2}{3}.$$

The formula for the SD is

$$\begin{aligned} & \text{SD in the box} \\ &= (\text{Bigger value} - \text{Smaller value}) \times \sqrt{\left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{bigger value} \end{array} \right) \times \left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with} \\ \text{smaller value} \end{array} \right)} \\ &= (1 - 0) \times \sqrt{\frac{1}{3} \times \frac{2}{3}} = 1 \times \sqrt{0.2222} = 0.471. \end{aligned}$$

The point here is that if there are just 0's and 1's, then

$$\text{Bigger value} - \text{Smaller value} = 1,$$

so the formula for the SD is just

$$\text{SD in the box} = \sqrt{\left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with 1} \end{array} \right) \times \left(\begin{array}{c} \text{Fraction of} \\ \text{tickets with 0} \end{array} \right)}.$$

Example

Back to the roulette example. First, find the SD in the box. The box has 38 tickets, twelve 2's and twenty-six -1 's. Because there are only two types of tickets, we can use the short-cut formula. Then from page 216,

$$\text{SD in the box} = 1.394.$$

What is the standard error of the sum of the draws if we draw 10 with replacement? We use the square root law, which says that

$$\begin{aligned} \left(\begin{array}{c} \text{Standard Error of} \\ \text{the sum of 10 draws} \end{array} \right) &= \sqrt{\text{Number of draws}} \times (\text{SD in the box}) \\ &= \sqrt{10} \times 1.394 \\ &= 3.162 \times 1.394 = 4.41 \end{aligned}$$

What about drawing 100 with replacement?

$$\begin{aligned} \left(\begin{array}{c} \text{Standard Error of} \\ \text{the sum of 100} \\ \text{draws} \end{array} \right) &= \sqrt{\text{Number of draws}} \times (\text{SD in the box}) \\ &= \sqrt{100} \times 1.394 \\ &= 10 \times 1.394 = 13.94 \end{aligned}$$

? Find the standard error of the sum of 1000 draws with replacement from that box.

The actual sums and the standard errors:

	Sum of 10 draws	Sum of 100 draws	Sum of 1,000 draws
Actual sums	\$5, \$2, -\$7, -\$1, -\$1, -\$1, \$8, \$8, -\$4, \$11	\$5, \$11, -\$13, \$11, -\$19, -\$16, \$8, -\$13, \$5, \$8	-\$31, -\$73, -\$46, -\$67, -\$70, -\$127, -\$85, -\$115, \$14, -\$4
Expected Value	-\$0.526	-\$5.26	-\$52.60
Off by about	$\pm 5?$ 10?	$\pm 10?$ 15?	$\pm 50?$
Standard Error	4.41	13.94	44.1

The calculated standard errors look reasonable given our estimated \pm numbers.

? Find the standard error of the sum of 11100 draws (the total number of draws for everyone combined.)

Rule of thumb

For data, we had the rule of thumb that about 68% of the data is between average \pm SD, and about 95% is between average ± 2 SD.

The expected value and standard error of the sum of the draws can be used in the rule of thumb for the actual sums of draws:

- Approximately 68% of the time, the sum of the draws will be between

Expected value \pm Standard error.

- Approximately 95% of the time, the sum of the draws will be between

Expected value $\pm 2 \times$ Standard error.

For the sum of 100 draws from the box, we found

Expected value = -5.26 , Standard error = 13.94 .

Then going \pm one standard error:

$$\begin{aligned}\text{Expected value} \pm \text{Standard error} &= (-5.26 - 13.94, -5.26 + 13.94) \\ &= (-19.20, 8.68).\end{aligned}$$

So about 68% of the time you play this game 100 times, you'd get between about $-\$19$ and $\$9$.

Going \pm two standard errors:

$$\begin{aligned}\text{Expected value} \pm 2 \times \text{Standard error} &= (-5.26 - 2 \times 13.94, \\ &\quad -5.26 + 2 \times 13.94) \\ &= (-33.14, 22.62).\end{aligned}$$


So about 95% of the time you play this game 100 times, you'd get between about $-\$33$ and $\$23$. So you may win some money, or lose even more.

Going plus or minus one standard error captures a lot of the actual sums. Going plus or minus two standard errors captures almost, but not quite, all actual sums.

	Sum of 10 draws	Sum of 100 draws	Sum of 1,000 draws
Actual Sums	\$5, \$2, -\$7, -\$1, -\$1, -\$1, \$8, \$8, -\$4, \$11	\$5, \$11, -\$13, \$11, -\$19, -\$16, \$8, -\$13, \$5, \$8	-\$31, -\$73, -\$46, -\$67, -\$70, -\$127, -\$85, -\$115, \$14, -\$4
Expected Value	-\$0.526	-\$5.26	-\$52.60
Standard Error	4.41	13.94	44.1
Expected Value \pm Standard Error	(-4.94, 3.88)	(-19.20, 8.68)	
Expected Value \pm 2 \times (Standard Error)	(-9.35, 8.29)	(-33.14, 22.62)	

? Fill in the two blanks for the sum of 1,000 draws.

Probability Histograms

Rolling a die is like drawing from the box . Imagine rolling two dice, which is the same as drawing two tickets from the box with replacement. (Draws are with replacement because whatever the first roll gives does not change the possibilities for the second roll.) What are you likely to get? Something between 2 and 12. For example, 20 times we rolled two dice, and found the sums:

6 2 3 4 7 7 11 6 9 2 3 6 8 3 6 6 2 4 7 4

The question is: What is the chance to get exactly 2 in rolling two dice? Or exactly 7?

Let's start by finding the chance that the sum is 2 when rolling the two dice. There is only one way to get that sum: The first draw is 1, and the second draw is 1.

First die's roll (draw) → ; Second die's roll (draw) → .

Since we are drawing with replacement, the draws are independent. And for each draw the chance is $\frac{1}{6}$.

$$\begin{aligned}
 (\text{Chance the sum is 2}) &= (\text{Chance the first draw is 1}) \\
 &\quad \times (\text{Chance the second draw is 1}) \\
 &= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} = 0.028.
 \end{aligned}$$

So there is a 2.8% chance you'd get exactly 2.

Draw two from the box again, but now find the chance that the sum of the draws is 3.

There are a couple of ways to get that sum: The first draw is 1, and the second draw is 2; the first is 2 and the second is 1.



OR



We figure the chance of each of those pairs:

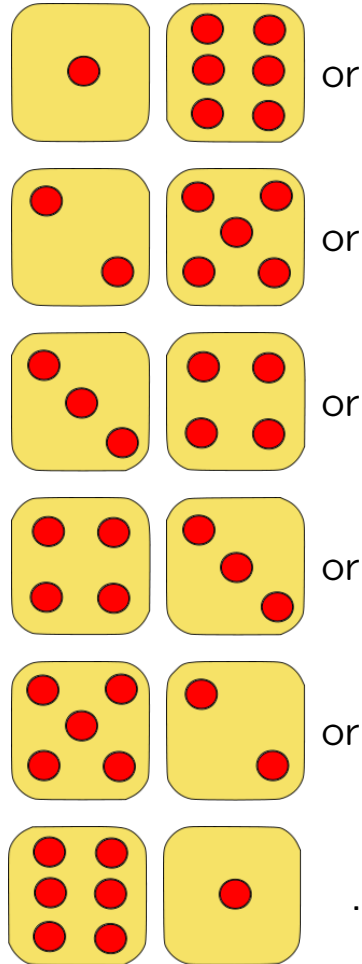
$$\begin{aligned} & (\text{Chance the first draw is 1}) \times (\text{Chance the second draw is 2}) \\ &= \frac{1}{6} \times \frac{1}{6} \\ &= \frac{1}{36} \text{ and} \end{aligned}$$

$$\begin{aligned} & (\text{Chance the first draw is 2}) \times (\text{Chance the second draw is 1}) \\ &= \frac{1}{6} \times \frac{1}{6} \\ &= \frac{1}{36} \end{aligned}$$

Since either of those pairs gives us what we want, we can add their chances:

$$(\text{Chance the sum is 3}) = \frac{1}{36} + \frac{1}{36} = \frac{2}{36} = 0.056 \rightarrow 5.6\%.$$

We could do the same for other possible sums. For example, the chance the sum of two draws is 7. There are several pairs that give you a sum of 7:



So there are six possible pairs that sum to 7. Each pair has chance $= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. Add up those six:

$$\begin{aligned}
 (\text{Chance the sum is 7}) &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\
 &= \frac{6}{36} = \frac{1}{6} = 0.167.
 \end{aligned}$$

So the percentage chance of rolling a 7 is 16.7%.

The table has the chances for (almost) all the possible sums of two draws. You can see that the chances are better to get sums in the middle. The sum of 7 is the most likely, 2 and 12 are least likely.

Sum of the Draws	Chance
2	2.8%
3	5.6%
4	
5	11.1%
6	13.9%
7	16.7%
8	13.9%
9	11.1%
10	8.3%
11	5.6%
12	2.8%

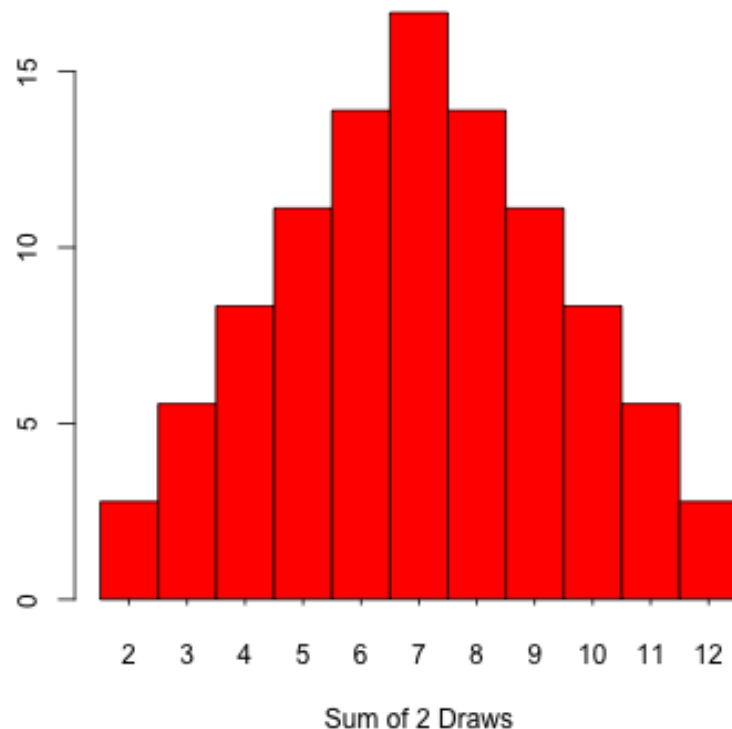
? Figure out the chance that the sum is 4. (First figure out which possible pairs will sum to 4. Then add up their chances.)

8.1 Probability histograms of the sums of the draws

We have values (2 to 12) and percentages, so we can make a histogram, just as for data. The histogram has chances as the percentages instead of percentage of the data, but it is the same idea. First we need intervals, where we use $\frac{1}{2}$'s.

Sum of Draws	Interval	Percentage	Base	Height
2	1.5 to 2.5	2.8%	1	2.8
3	2.5 to 3.5	5.6%	1	5.6
4	3.5 to 4.5	8.3%	1	8.3
5	4.5 to 5.5	11.1%	1	11.1
6	5.5 to 6.5	13.9%	1	13.9
7	6.5 to 7.5	16.7%	1	16.7
8	7.5 to 8.5	13.9%	1	13.9
9	8.5 to 9.5	11.1%	1	11.1
10	9.5 to 10.5	8.3%	1	8.3
11	10.5 to 11.5	5.6%	1	5.6
12	11.5 to 12.5	2.8%	1	2.8

Recall that the areas of the boxes in the histograms are the percentages, each one being base times height, so that $\text{Height} = \text{Percentage}/\text{Base}$. The bases are all 1. Now draw a histogram with those intervals and those heights. It looks quite a bit like a normal curve, though not exactly.



When can you add up chances?

When figuring out the chances of the sums of a two draws, we figured out the individual possible ways we could get the sum, then added up the chances of those possibilities. In general you can only add up chances like that if the possibilities do not overlap, that is, it is impossible for both to happen at the same time.

For example, consider drawing two tickets from

1	2	3	4	5	6
---	---	---	---	---	---

. Here are two possibilities for getting a sum of 3:

1. Get 1 on the first draw, and 2 on the second.
2. Get 2 on the first draw, and 1 on the second.

It is impossible that both can happen, i.e., if you get 1 on the first and 2 on the second, you cannot have gotten 2 on the first and 1 on the second. Then, you can add up the chances:

$$(\text{Chance the sum is 3}) = (\text{Chance of 1 on first draw and 2 on second}) \\ + (\text{Chance of 2 on first draw and 1 on second})$$

The rule is

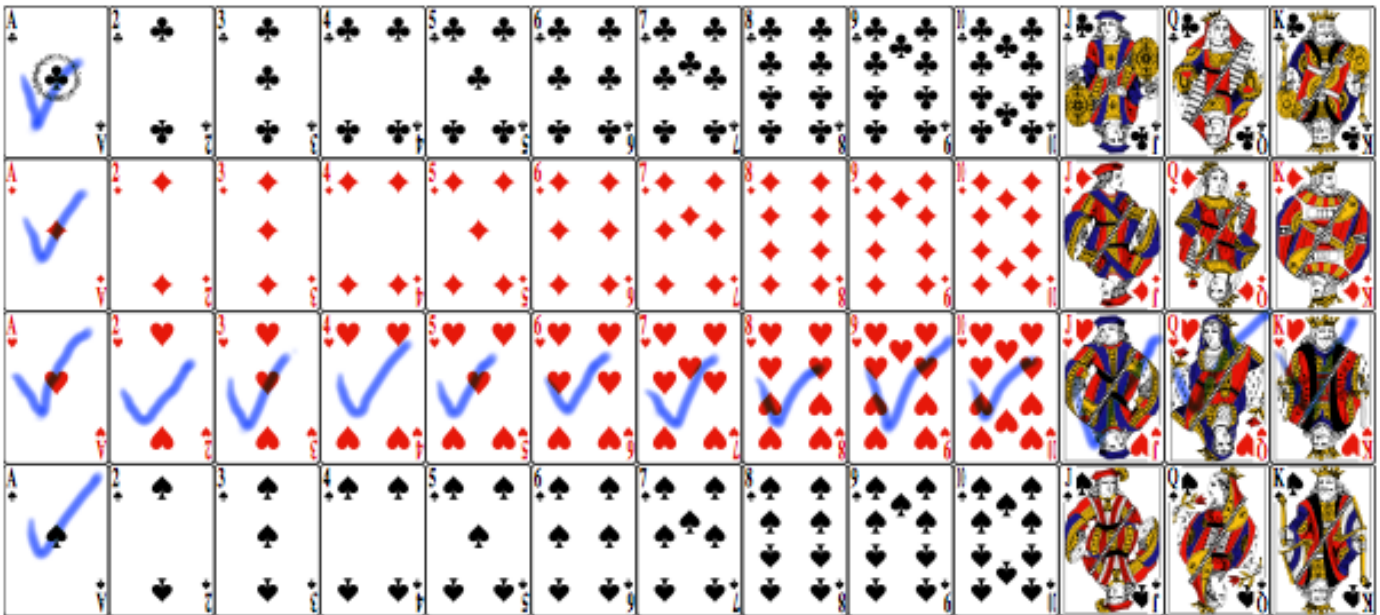
If the two possibilities cannot occur at the same time, then

$$\text{Chance one or the other happens} = (\text{Chance the first happens}) \\ + (\text{Chance the second happens}).$$

Cards. Suppose you draw one card, and you want to know the chance it is either a heart or an ace or both. The two possibilities are

1. It is a heart.
2. It is an ace.

The cards that satisfy the conditions (either a heart or ace or both) are checked:



You can count 16 of them. So, since there are 52 cards total,

$$\text{Chance of heart or ace or both} = \frac{16}{52} \rightarrow 30.77\%.$$

The chance of heart is $\frac{13}{52}$, and the chance of an ace is $\frac{4}{52}$. So if we add those chances:

$$(\text{Chance of heart}) + (\text{Chance of Ace}) = \frac{13}{52} + \frac{4}{52} \rightarrow 32.69\%.$$

That is different that 30.77%. What happened? There is overlap in the two possibilities, since you can get a heart and an ace on the same draw if you get the ace of hearts.

If both possibilities can happen at the same time, you cannot add up their chances.

The sum of three draws

What about the sum of more than two draws? To find the chances is similar, but there are more possibilities to write out. For example, here is a box for flipping a fair coin:



The heads is coded as a 1, the tails as a 0. The sum of three draws with replacement is like flipping the coin three times and counting the number of heads. What are the possible sums? You could get 0 heads, 1 head, 2 heads, or 3 heads. To figure out the chance of each of those possibilities, you have to write out the possibilities for three draws.

There is only one way to get 0 heads: All three draws are 0.

Sum of 3 draws	Possible Ways
0	0 0 0

Because the draws are with replacement, they are independent. So

$$(\text{Chance of 0 heads}) = (\text{Chance 0 on first}) \times (\text{Chance 0 on second}) \times (\text{Chance 0 on third})$$

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = 0.125 \rightarrow 12.5\%.$$

To get a sum of 1, you need to get one heads and two tails. The heads can be on the first, second, or third draw:

Sum of 3 draws	Possible Ways		
1	1	0	0
	0	1	0
	0	0	1

Each of those possibilities has the same chance of $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. There are three possibilities. So

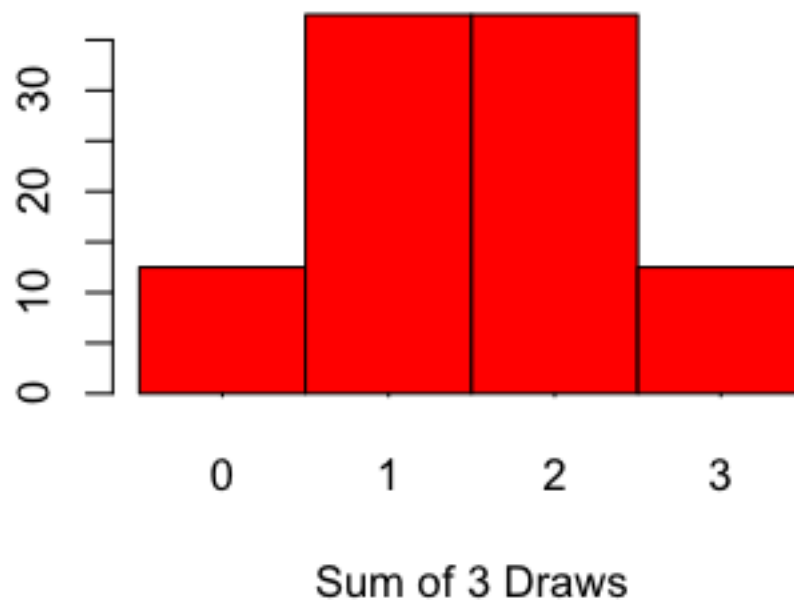
$$(\text{Chance of 1 heads}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} = 0.375 \rightarrow 37.5\%.$$

? Figure out the chances of getting a sum of 2, and a sum of 3.

The table for the chances for three draws is then

Sum of 3 draws	Possible ways	Chance
0	<div>0</div> <div>0</div> <div>0</div>	12.5%
1	<div>1</div> <div>0</div> <div>0</div> <hr/> <div>0</div> <div>1</div> <div>0</div> <hr/> <div>0</div> <div>0</div> <div>1</div>	37.5%
2	<div>1</div> <div>1</div> <div>0</div> <hr/> <div>1</div> <div>0</div> <div>1</div> <hr/> <div>0</div> <div>1</div> <div>1</div>	37.5%
3	<div>1</div> <div>1</div> <div>1</div>	12.5%

The histogram for this table is



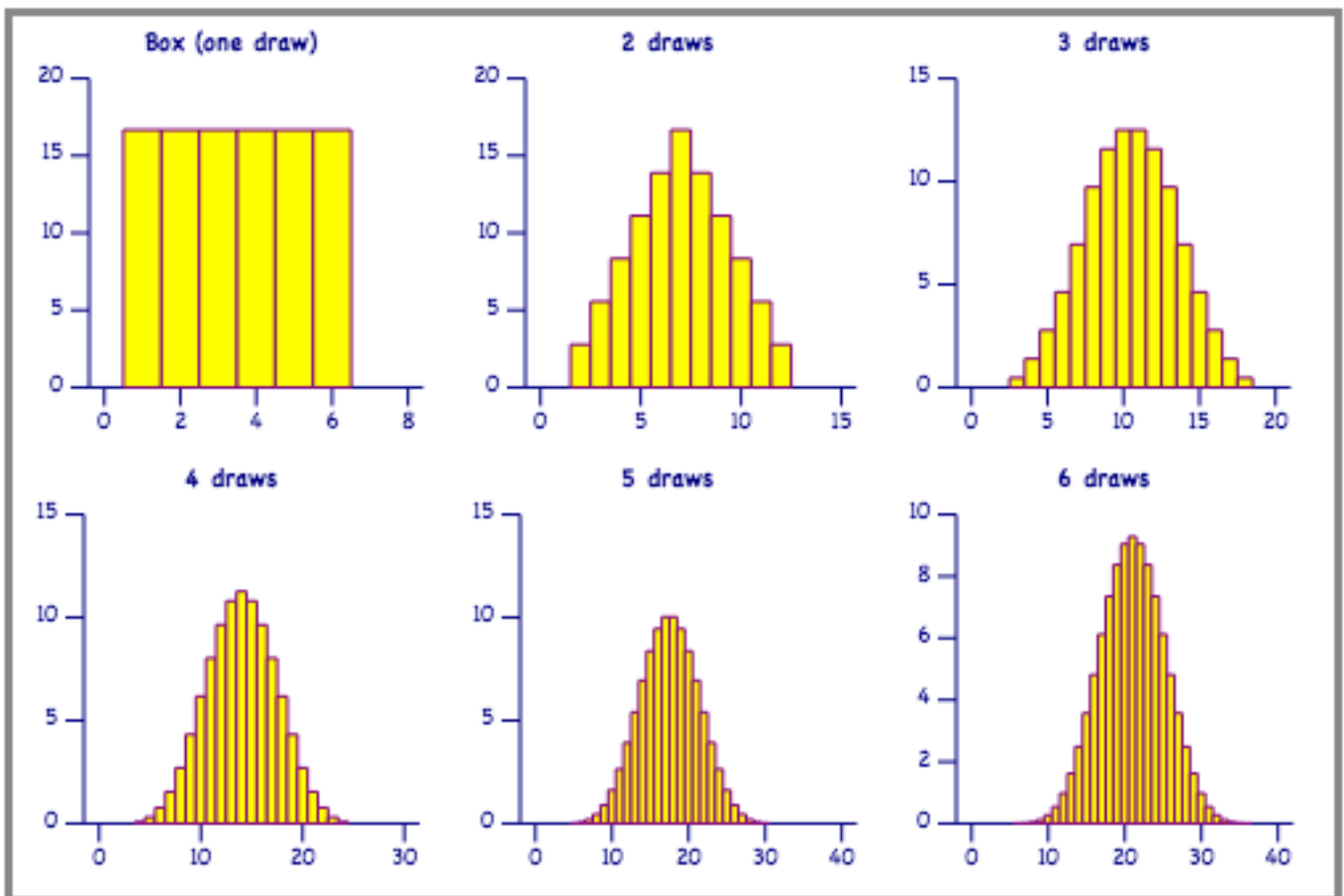
Dice

In principle, one can figure out the probability histogram for the sum of any number of draws from any box, but doing the calculations by hand can be very tedious. Computers can do the work quite quickly.

Go back to rolling dice, that is, drawing with replacement from the box

1	2	3	4	5	6
---	---	---	---	---	---

. Below are the probability histograms of the sums of 2 draws, and 3 draws, ..., and 6 draws. Also, the first one is the histogram for one draw, which is just the histogram for the tickets in the box.



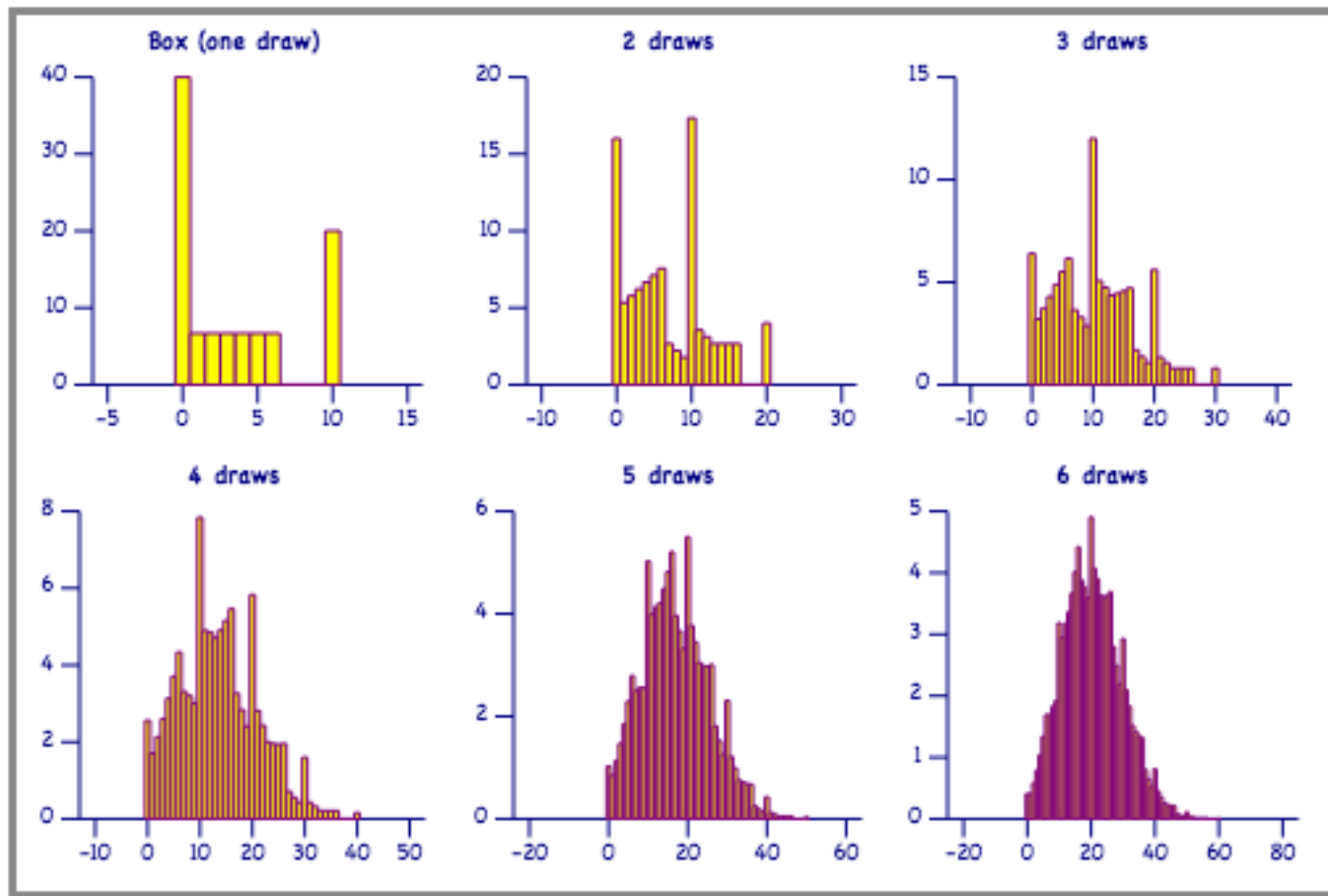
Notice that the more draws, the more the histogram looks like a normal curve. Even with 3 draws, the histogram looks fairly normal.

Another box

Consider the box

0, 0, 0, 0, 0, 0, 0, 1, 2, 3, 4, 5, 6, 10, 10, 10

Here are the histograms for the sums of the draws:

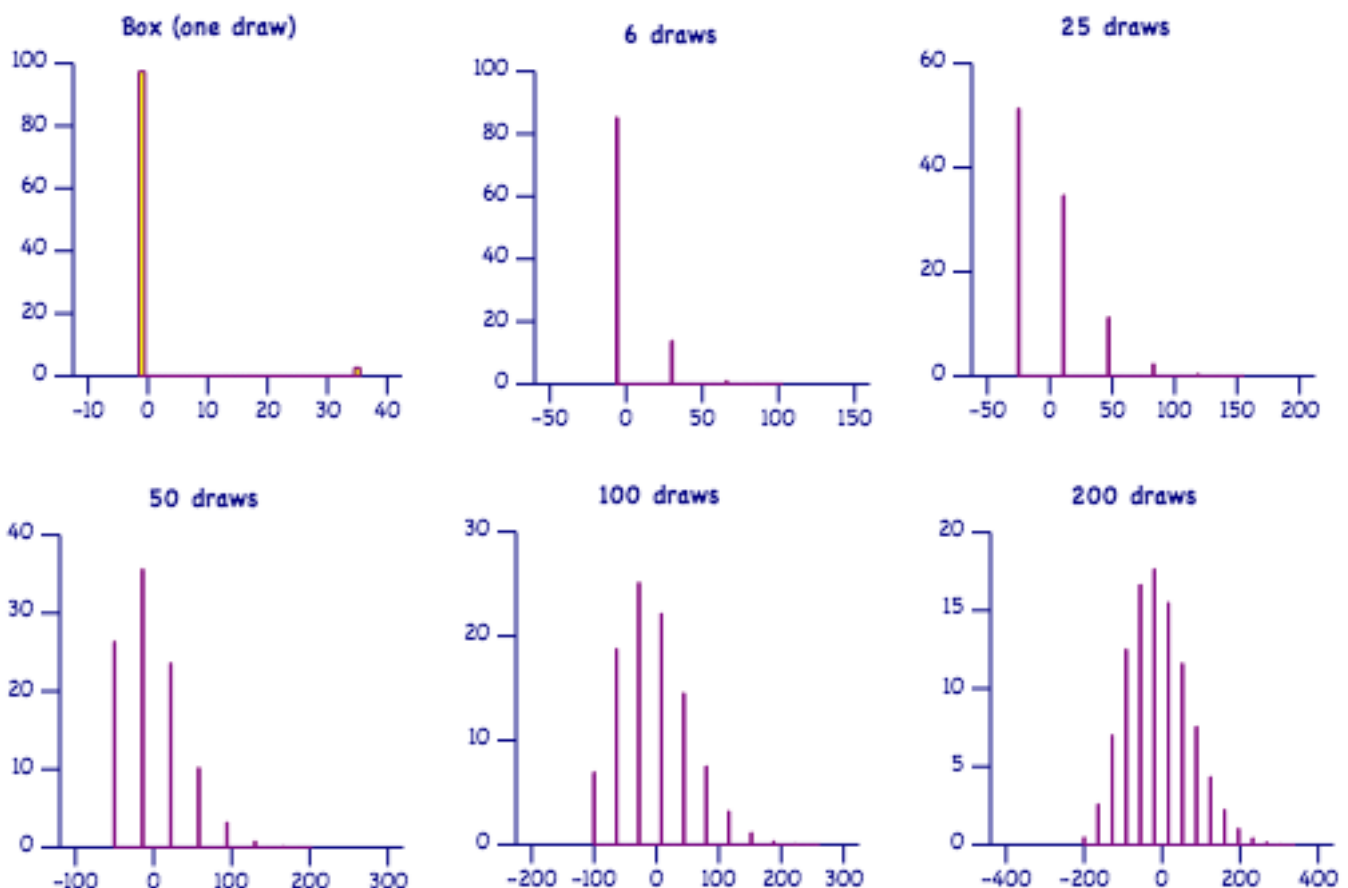


The histogram for the box is somewhat lopsided, and the first few probability histograms do not look very much like a normal curve. The histogram for 6 draws is somewhat normal looking, though is still a bit off.

Roulette

In betting \$1 on a single number in roulette, you get \$35 if your number comes up, but lose $-\$1$ if it does not. There is 1 chance to win, and 37 chances to lose, so the box has one 35 and thirty-seven -1 's.

35, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
 -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
 -1, -1, -1, -1, -1, -1, -1, -1, -1, -1



Now the box is very lopsided. The probability histogram for 6 draws is not at all like a normal curve. Nor is the histogram for 25 or even 50 draws. When we get to 100 draws, it is starting to look somewhat normal. The histogram for 200 draws is reasonably normal looking.

Moral

- The probability histogram of the sum of the draws with replacement from any box will look like a normal curve if you take enough draws.
- The more "lopsided" the histogram of the box, the more draws needed before the histogram looks like a normal curve.

Many boxes do not need more than 10 draws, but some need over 100.

8.2 Using the normal curve

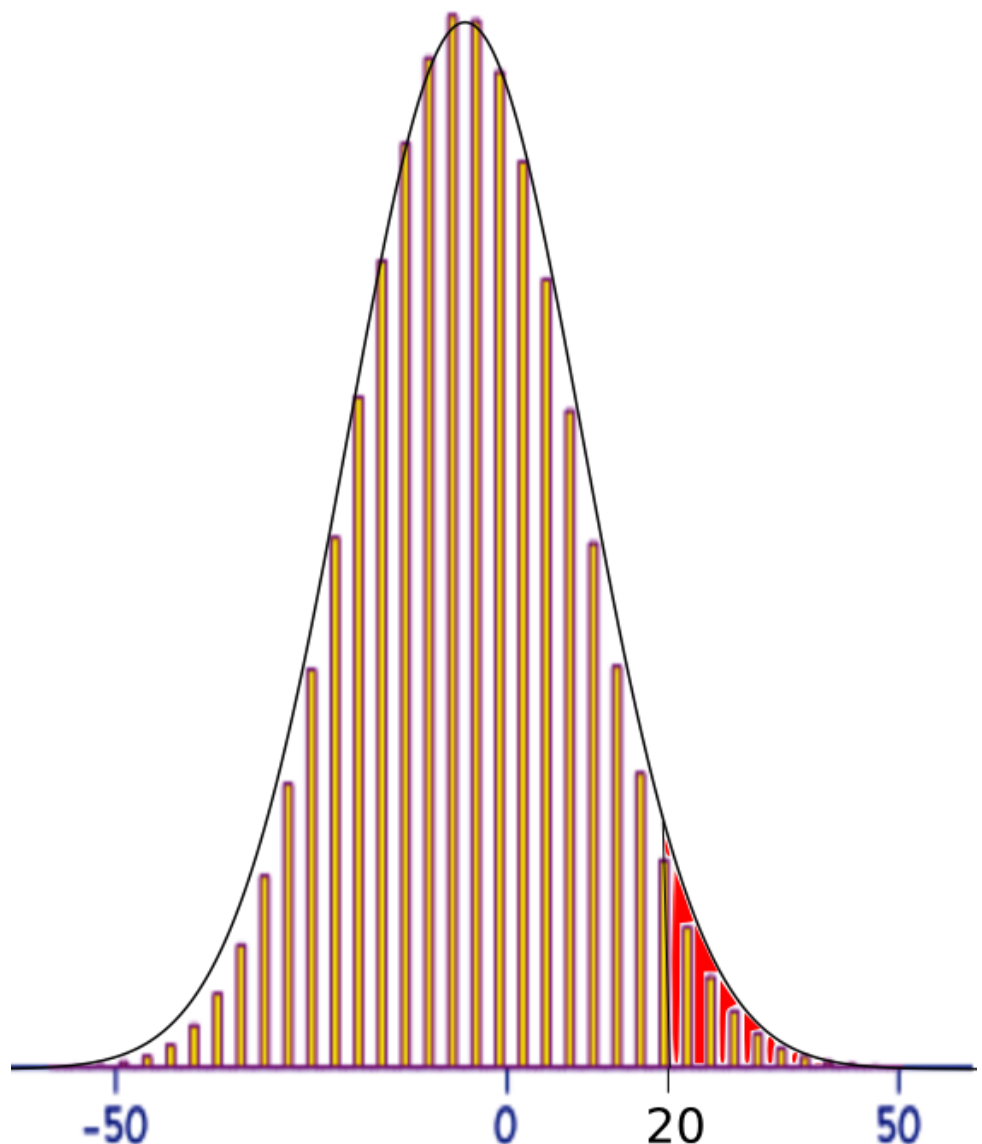
In roulette, betting \$1 on the first 12 numbers is like drawing a ticket from the box \rightarrow

2	2	2	2	2	2	2	2	2	2
2	2	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1		

Now we want to answer questions such as “What is the chance of winning \$20 or more when playing this bet 100 times?”

The results from playing this bet 100 times is like drawing 100 from the box with replacement.

The histogram is the probability histogram for the sum of the draws. Notice that it looks like a normal curve. We want to find the area to the right of \$20.



To use the normal curve to estimate the chance you end up with \$20 or more, we first need to change the value of \$20 into standard units, this time by subtracting the **expected value** and dividing by the **standard error**:

$$\text{Standard units} = \frac{\text{Value} - (\text{Expected value of the sum})}{\text{Standard error of the sum}}.$$

From previous work (page 219), we know that the average in the box is -0.0526 , and the SD in the box is 1.394 , so that

$$\text{Expected value of the sum} = 100 \times (-0.0526) = -5.26,$$

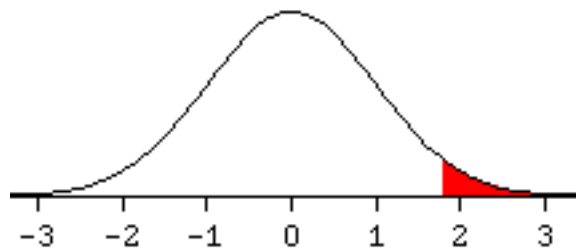
and

$$\text{Standard error of the sum} = \sqrt{100} \times 1.394 = 13.94.$$

Then the standard units are found using the formula

$$\begin{aligned} \text{Standard units} &= \frac{\text{Value} - (\text{Expected value of the sum})}{\text{Standard error of the sum}} \\ &= \frac{20 - (-5.26)}{13.94} = \frac{25.26}{13.94} = 1.81. \end{aligned}$$

So we want the area in the normal curve to the right of 1.81:



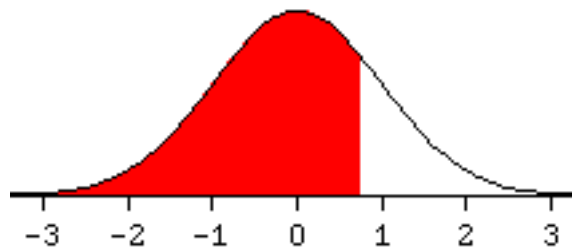
From the normal table, the area above 1.81 is about 3.59%¹. So there is a small chance that you'd win \$20 or more.

¹The area between ± 1.8 is 92.81%, so that answer is $(100 - 92.81)/2$.

What if you played that game 1,000 times? What is the chance you would lose \$20 or more? Now the expected value of the sum is -52.60, and the standard error of the sum is 44.10. Use the normal curve to estimate the chance you end up with -\$20 or less. Changing the value of -\$20 into standard units:

$$\begin{aligned}\text{Standard units} &= \frac{\text{Value} - (\text{Expected value of the sum})}{\text{Standard error of the sum}} \\ &= \frac{-20 - (-52.60)}{44.10} = \frac{32.60}{44.10} = 0.74.\end{aligned}$$

We want the area below 0.74 in the normal curve:



From the normal table, we get about 77.34%. There's a good chance you would lose at least \$20.

? Find the standard units for the value of \$0. What is the chance you end up with more than \$0?

Summary

If you take a large number of draws with replacement from a box, you can find the chances about the sum of the draws using the normal curve. You first need to find

- The expected value of the sum of draws (for which you need the average in the box):

$$\left(\begin{array}{l} \text{Expected} \\ \text{value of the} \\ \text{sum of the} \\ \text{draws} \end{array} \right) = (\text{Number of draws}) \times (\text{Average in the box}).$$

- The standard error of the sum of draws (for which you need the SD in the box):

$$\left(\begin{array}{l} \text{Standard error} \\ \text{of the sum of} \\ \text{the draws} \end{array} \right) = \sqrt{\text{Number of draws}} \times (\text{SD in the box}).$$

The standard units are then found by taking your value and calculating

$$\text{Standard units} = \frac{\text{Value} - (\text{Expected value of the sum})}{\text{Standard error of the sum}}.$$

You use the standard units to find the appropriate area under the normal curve.

Part III

What Can You Say about the Population?

9.1 *Literary Digest* Poll of 1936: Roosevelt (D) vs. Landon (R)

In 1936, the presidential race was between the Democrat incumbent Franklin Roosevelt, and the Republican challenger Alf Landon.

The *Literary Digest* was a popular magazine in the early part of the 20th century. They sent out 10,000,000 postcards asking people who they would vote for between Roosevelt and Landon. 2,400,000 cards came back.



The results:

	Roosevelt	Landon
<i>Literary Digest's</i> prediction	43%	57%
Actual result	62%	38%

Not only did they get the winner wrong, notice how far off the *Literary Digest* was: 43% for Roosevelt instead of 62%. Off by -19 percentage points. Why? They polled plenty of people.

Biased Sample

They sent out 10,000,000 postcards asking people who they would vote for between Roosevelt and Landon. 2,400,000 cards came back. Two possible sources of bias:

- **Selection bias.** To whom were the cards sent? To people on various lists, e.g., telephone lists, car owners, members of clubs, etc. They tended to be wealthier, who tended to be more Republican than the overall electorate. So the sample was not representative, but biased in favor of the Republican Landon.
- **Non-response bias.** Not everyone who got a card responded. In fact, only about 24% did. Are they different than the responders? They tend to be more middle class. (Is this bias for Roosevelt or Landon?) [In the polio study, the volunteers tended to be more affluent.]

There can be other sources of bias, e.g., in the way questions are asked. Notice: Even though the sample was very large, it was not a good one.



The Gallup poll



In 1936, George Gallup was just starting out. He took a sample of only 50,000 people. He predicted 56% for Roosevelt. Still -6 points off, but he got the winner.

Being a wise guy, he also took a sample of 3,000 from the same lists the *Literary Digest* used, and was able to guess how bad *Literary Digest's* prediction was, sampling just 3,000 instead of 2.4 million people.

	Roosevelt	Landon
<i>Literary Digest's</i> prediction	43%	57%
Gallup's prediction of <i>Literary Digest's</i> prediction	44%	56%
Gallup's prediction of election	56%	44%
Actual result	62%	38%

Quota sampling

Gallup used **quota sampling** from 1936 to 1948. Each interviewer had quotas, e.g., one might have to interview 6 men and 7 women. Of the men, 2 had to be black and 4 white. Or of the 7 women, 3 had to be middle income, 3 lower income, and 1 high income. The interviewer did have quotas, but otherwise could use discretion. There was still room for bias (e.g., people may tend to interview better dressed people within each group).

The Gallup Poll, 1936-1948

In 1948, it was Dewey (R) vs. Truman (D) (vs. Thurmond vs. Wallace). Gallup predicted that Truman would get 44% and Dewey 50%, so that Dewey would win. Truman actually won, 50% to 45%.



The next table has Gallup's results for four presidential elections, using quota sampling:

Year	Democrat	Republican	Gallup prediction of Democrat	Actual percentage for Democrat	Error
1936	Roosevelt	Landon	55.7	62.5	-6.8
1940	Roosevelt	Willkie	52.0	55.0	-3.0
1944	Roosevelt	Dewey	51.5	53.8	-2.3
1948	Truman	Dewey	44.5	49.5	-5.0

There seems to be some bias against Democrats. The standard size of these errors is about 4.6.

Bad sampling procedures

The quota sampling Gallup used was not terrible, because it did give him a more representative sample than the Literary Digest poll. But it still allowed bias to creep in, mainly because the interviewers were able to use their own judgement when choosing people to poll.

Here are some bad sampling procedures:

- Samples of convenience: Like the *Literary Digest* poll, the pollsters just take who is convenient. Has selection bias, plus non-response bias. Bad.
- Self-selected sampling: Polls where people themselves decide whether to be in the sample, for example, by calling or texting, or going to a favorite (or hated) website to vote. Bad.
- Quota sampling: Can be better, more representative, but still fairly bad.

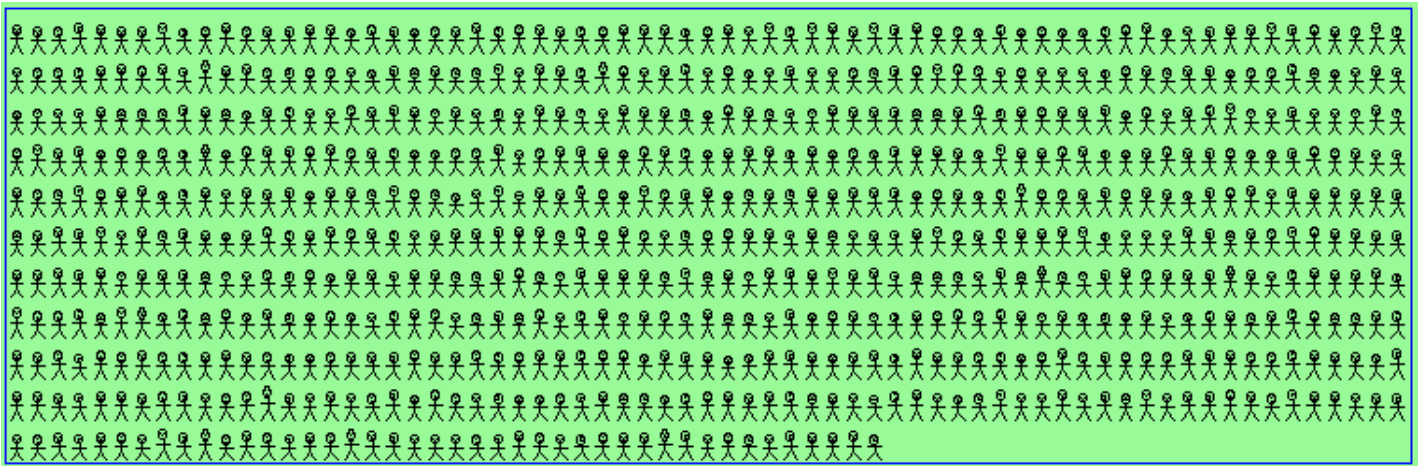
Good sampling procedures do not allow discretion on the part of interviewers or interviewees, and choose people fairly. The best are called probability samples.

9.2 Probability sampling

Simple random samples

The simplest type of probability sampling is called **simple random sampling**. You draw a number of people **without replacement** from the population. The box represents the population.

Box = Population



If you want a sample of 100 from this population, you draw 100 without replacement from the box. You draw without replacement because you do not want to interview the same person twice.

Multistage cluster samples

Simple random sampling is a good method, but can be impractical if you are trying to take a sample from the entire population of the United States, or even just Illinois. Another type of probability sample is called multistage cluster sampling. You need not know the exact details, but the idea follows:

- Randomly select a given number of towns in each region of the United States.
- Randomly select a given number of wards from each town selected.
- Randomly select a given number of precincts from each ward selected.
- Randomly select a given number of households from each precinct selected.
- Interview a specific person in the household.

The interviewer has no discretion: It is up to the random mechanism. It is a probability sample: Good!

The Gallup Poll, 1952-2012

From 1936 to 1948, Gallup used quota sampling, which showed some bias against Democrats. The standard size of these errors was about 4.6. From 1952 on, he used probability sampling. Here are the results up through 2012.

Year	Democrat	Republican	Gallup prediction of Democrat	Actual percentage for Democrat	Error
1952	Stevenson	Eisenhower	49.0	44.6	+4.4
1956	Stevenson	Eisenhower	40.5	42.2	-1.7
1960	Kennedy	Nixon	50.5	50.1	+0.4
1964	Johnson	Goldwater	64.0	61.3	+2.7
1968	Humphrey	Nixon	42.0	42.9	-0.9
1972	McGovern	Nixon	38.0	38.2	-0.2
1976	Carter	Ford	48.0	50.1	-2.1
1980	Carter	Reagan	44.0	41.0	+3.0
1984	Mondale	Reagan	41.0	40.8	+0.2
1988	Dukakis	Bush	44.0	46.1	-2.1
1992	Clinton	Bush	49.0	43.3	+5.7
1996	Clinton	Dole	52.0	49.2	+2.8
2000	Gore	Bush	46.0	48.4	-2.4
2004	Kerry	Bush	49.0	48.3	+0.7
2008	Obama	McCain	55.0	53.0	+2.0
2012	Obama	Romney	49.0	51.0	-2.0

There doesn't seem to be bias for or against Democrats (9 errors are +, 7 are -). The typical size of the errors is only 2.5. For the quota sampling it was 4.6. So probability sampling did much better than quota sampling. Though notice Gallup did get the result wrong in 2012.

9.3 Summary

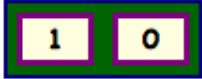
- Probability sampling: Good. Includes simple random sampling and more complicated methods.
 - Randomly decide who gets interviewed: No discretion on the part of the interviewer.
- Quota sampling: Fairly bad. Not a probability sample. Has bias.
- Sample of convenience or self-selected sample: Just plain bad. Not a probability sample. Has bias.

Estimating Population Percentages and Averages

Flipping a fair coin a number of times and counting the number of heads is like taking the sum of a number of draws with replacement from 1 0. Thirty people try flipping the coin. Ten people flip it 10 times, ten people flip it 100 times, and ten people flip it 1,000 times. How many heads do they get?

	# of heads in 10 flips	# of heads in 100 flips	# of heads in 1,000 flips
	Sum of 10 draws	Sum of 100 draws	Sum of 1,000 draws
Actual # of heads	5, 8, 3, 4, 8, 3, 4, 4, 7, 5	50, 52, 45, 50, 51, 51, 46, 45, 49, 42	501, 490, 503, 507, 514, 478, 516, 489, 513, 511
Expected # of heads	5	50	500
Errors	0, 3, -2, -1, 3, -2, -1, -1, 2, 0	0, 2, -5, 0, 1, 1, -4, -5, -1, -8	1, -10, 3, 7, 14, -22, 16, -11, 13, 11
Standard Error of sum	1.58	5	15.8

Not surprisingly, the more flips, the more heads you tend to get. The actual numbers of heads are reasonably close to their expected values. The errors are (Actual #) – (Expected value). The standard error gives an idea of the typical size of the errors. Notice that the more flips, the larger the errors tend to be.

? Draw with replacement from the box .

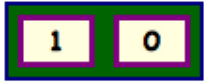
Find the average in the box, and the SD in the box.

Draw 10 with replacement. Find the expected value and standard error of the sum of the draws.

Draw 100 with replacement. Find the expected value and standard error of the sum of the draws.

Check that your answers agree with the table on the previous page.

10.1 Percentages

We are still drawing with replacement from . But now we look at the *percentage* of heads in the draws:

	10 Draws	100 Draws	1,000 Draws
Actual # of heads	5, 8, 3, 4, 8, 3, 4, 4, 7, 5	50, 52, 45, 50, 51, 51, 46, 45, 49, 42	501, 490, 503, 507, 514, 478, 516, 489, 513, 511
Actual % of heads	50%, 80%, 30%, 40%, 80%, 30%, 40%, 40%, 70%, 50%	50%, 52%, 45%, 50%, 51%, 51%, 46%, 45%, 49%, 42%	50.1%, 49.0%, 50.3%, 50.7%, 51.4%, 47.8%, 51.6%, 48.9%, 51.3%, 51.1%
For example	$\frac{8}{10} \times 100 = 80\%$	$\frac{45}{100} \times 100 = 45\%$	$\frac{516}{1000} \times 100 = 51.6\%$

The percentages average about 50%, which should be reasonable: If the coin is fair, about 50% of the flips should be heads. But you can also see that the actual percentages are not always that close to 50%, especially when there are only 10 draws.

Since we expect about 50% of the draws to be heads, we can think of the **expected value of the percentage** in the draws, which is the same as the percentage in the box:

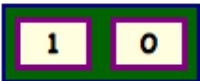
$$\left(\begin{array}{l} \text{Expected value of percentage of} \\ \text{1's in the draws} \end{array} \right) = \text{Percentage of 1's in the box.}$$

Since for this box, the percentage of 1's is 50%,

$$\left(\begin{array}{l} \text{Expected value of percentage of} \\ \text{1's in the draws} \end{array} \right) = 50\%$$

Notice that this percentage is the same no matter how many draws you take.

Errors

The percentage of 1's in the box  is 50%, so the expected value of the percentage in the draws is 50%. The next table also looks at the actual errors in percentage, where

$$\begin{aligned} \text{Error in percentage} &= \\ &\text{Actual percentage} - \text{Expected value of percentage.} \end{aligned}$$

So if the actual percentage is 50%, the error is 50%-50%=0%, and if the actual percentage is 30%, the error is 30%-50% = -20%. The table has the errors in percentage:

	% of heads in 10 flips	% of heads in 100 flips	% of heads in 1,000 flips
Actual % of heads	50%, 80%, 30%, 40%, 80%, 30%, 40%, 40%, 70%, 50%	50%, 52%, 45%, 50%, 51%, 51%, 46%, 45%, 49%, 42%	50.1%, 49.0%, 50.3%, 50.7%, 51.4%, 47.8%, 51.6%, 48.9%, 51.3%, 51.1%
Expected value of % of heads	50%	50%	50%
Error in %	0%, 30%, -20%, -10%, 30%, -20%, -10%, -10%, 20%, 0%	0%, 2%, -5%, 0%, 1%, 1%, -4%, -5%, -1%, -8%	0.1%, -1%, 0.3%, 0.7%, 1.4%, -2.2%, 1.6%, -1.1%, 1.3%, 1.1%

Notice that the more draws (flips), the smaller the error in percentage. So when taking just 10 draws, some errors are as high as 30%. For 100 draws, the errors are around 1% to 5% to 8%.

? About how large are the errors in percentages for 1000 draws?

The law of averages for percentages

We saw that the errors in percentage tended to be smaller when there were more draws. That effect is known as the **law of averages** for percentages:

When drawing with replacement from a box with just 0's and 1's, the **more draws** you take the **closer the percentage** in the draws is likely to be to the percentage in the box.

Recall from page 253 that the errors went up when looking at the sum of the draws:

When drawing with replacement from a box with just 0's and 1's, the **more draws** you take the **farther the sum** of the draws is likely to be from the expected value of the sum.

Warning

Suppose you are flipping a fair coin a number of times independently, and you get 10 tails in a row to start. Does the law of averages say that the next flip is more likely to be heads?

No.

The flips are independent. The 11th flip is still like a draw from



. The chance of heads is still $\frac{1}{2}$. Even if you get 100 tails in a row, if you really are flipping a fair coin independently, the next flip has a 50% chance of being heads. It is not as if the coin is due for a heads. What the law of averages says is that from now on, about 50% will be heads, and if you do enough flips, the first few, or many, draws are basically irrelevant.

The standard error of the percentage

Here again is the table with the errors in percentage when drawing with replacement from the box with one 0 and one 1:

	% of heads in 10 flips	% of heads in 100 flips	% of heads in 1,000 flips
Actual % of heads	50%, 80%, 30%, 40%, 80%, 30%, 40%, 40%, 70%, 50%	50%, 52%, 45%, 50%, 51%, 51%, 46%, 45%, 49%, 42%	50.1%, 49.0%, 50.3%, 50.7%, 51.4%, 47.8%, 51.6%, 48.9%, 51.3%, 51.1%
Expected value of % of heads	50%	50%	50%
Error in %	0%, 30%, -20%, -10%, 30%, -20%, -10%, -10%, 20%, 0%	0%, 2%, -5%, 0%, 1%, 1%, -4%, -5%, -1%, -8%	0.1%, -1%, 0.3%, 0.7%, 1.4%, -2.2%, 1.6%, -1.1%, 1.3%, 1.1%
Standard Error of %?	10? 20?	2? 3?	1?

The last row estimates the standard error of the percentage, that is, the typical size of the errors, ignoring sign. But there is a specific formula, which we develop next.

The formula

The idea is that the number of heads among the flips is the same as the sum of the draws from the box. So the percentage of heads in the flips is

$$\begin{aligned}\text{Percentage in the flips} &= \frac{\# \text{ of heads in draws}}{\# \text{ of flips}} \times 100 \\ &= \frac{\text{Sum of draws}}{\# \text{ of draws}} \times 100.\end{aligned}$$

Likewise, the **standard error of the percentage** in the draws takes the standard error of the sum of the draws, then turns it into a percentage:

$$\text{SE of percentage in the draws} = \frac{\text{SE of sum of the draws}}{\# \text{ of draws}} \times 100.$$

From before (page 214),

$$\text{SE of sum of the draws} = \sqrt{\# \text{ of draws}} \times (\text{SD in box}),$$

so

$$\text{SE of percentage in the draws} = \frac{\sqrt{\# \text{ of draws}} \times (\text{SD in box})}{\# \text{ of draws}} \times 100.$$

We can simplify the formula with a little math. Focus on the number of draws in the formula. Square roots work as follows:

$$\frac{\sqrt{\# \text{ of draws}}}{\# \text{ of draws}} = \frac{1}{\sqrt{\# \text{ of draws}}}.$$

That means we can simplify the standard error formula to be

$$\text{SE of percentage in the draws} = \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}} \times 100.$$

Since the number of draws is in the denominator, the more draws, the smaller the standard error of percentage.

Example

We have that

$$\text{SE of percentage in the draws} = \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}} \times 100.$$

The box is , so the SD in the box (see page 217) is

$$\text{SD in box} = (1 - 0) \sqrt{\frac{1}{2} \times \frac{1}{2}} = \frac{1}{2}.$$

Then for 10 draws,

$$\begin{aligned} \text{SE of percentage in the draws} &= \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}} \times 100 \\ &= \frac{1/2}{\sqrt{10}} \times 100 \\ &= 0.158 \times 100 = 15.8\%. \end{aligned}$$

? Find the SE of percentage for 100 draws, and for 1,000 draws.

Here is the table, including the standard errors:


	% of heads in 10 flips	% of heads in 100 flips	% of heads in 1,000 flips
Actual % of heads	50%, 80%, 30%, 40%, 80%, 30%, 40%, 40%, 70%, 50%	50%, 52%, 45%, 50%, 51%, 51%, 46%, 45%, 49%, 42%	50.1%, 49.0%, 50.3%, 50.7%, 51.4%, 47.8%, 51.6%, 48.9%, 51.3%, 51.1%
Error in %	0%, 30%, -20%, -10%, 30%, -20%, -10%, -10%, 20%, 0%	0%, 2%, -5%, 0%, 1%, 1%, -4%, -5%, -1%, -8%	0.1%, -1%, 0.3%, 0.7%, 1.4%, -2.2%, 1.6%, -1.1%, 1.3%, 1.1%
Standard error of %	15.8%	5%	1.58%

You can see that the standard error of percentages reflect fairly well the sizes of the actual errors in percentage.

Also, the more flips, the smaller the error in %. Which again is the **law of averages for percentages**.

When drawing with replacement from a box with just 0's and 1's, the **more draws** you take the **closer the percentage** in the draws is likely to be to the percentage in the box.

10.2 The law of average for any box

Now think of rolling a die a number of times, which is equivalent to drawing with replacement from the box . If we take 10 draws, a possible result is

2 5 2 2 5 3 5 1 5 4

The average of these draws is 3.4. Is that typical? What would we expect?

The expected value of the average

For any box, if you take a number of draws with replacement and find the average, you should expect this average to be somewhat close to the average in the box. In fact,

$$\left(\begin{array}{c} \text{Expected value of average in} \\ \text{the draws} \end{array} \right) = \text{Average in the box.}$$

For this box,

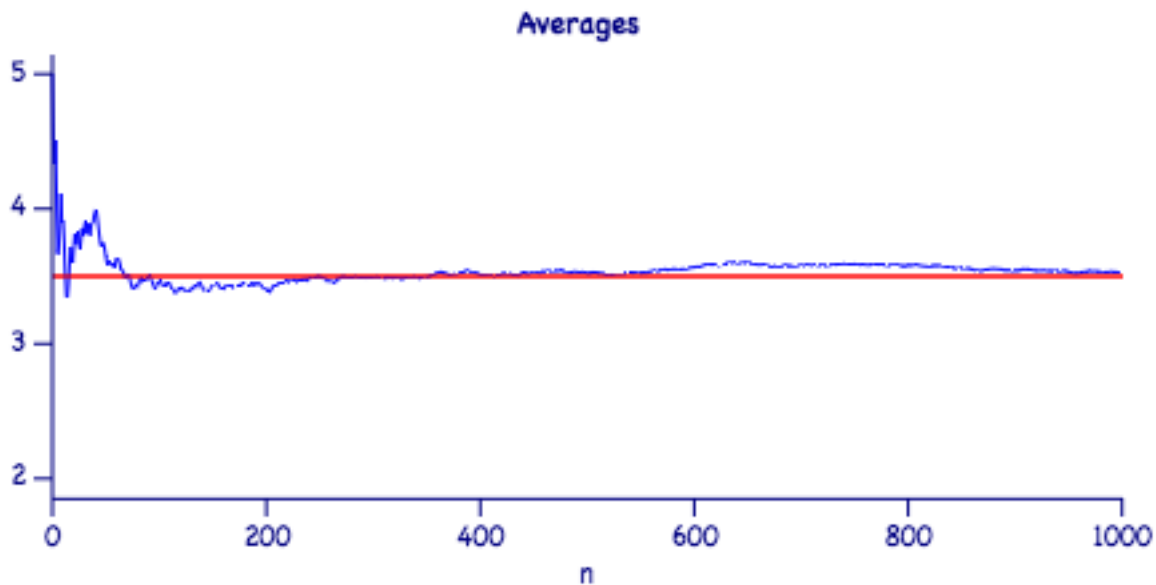
$$\text{Average in the box} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5.$$

In the example above, the average in the draws was 3.4, which is quite close to 3.5. The error is $3.4 - 3.5 = -.1$.

The above example, with a few more examples:

10 Draws	Average of draws	Average in box	Error
2 5 2 2 5 3 5 1 5 4	3.4	3.5	-0.1
1 1 6 6 2 2 4 4 6 5	3.7	3.5	0.2
1 1 2 4 1 3 1 4 4 5	2.6	3.5	-0.9
3 6 6 4 5 5 4 1 1 6	4.1	3.5	0.6
5 2 1 3 1 4 4 5 1 5	3.1	3.5	-0.4

What if we take more draws? The next plot graphs the average of the draws after 1 draw, 2 draws, 3 draws, ..., up to 1000 draws. The horizontal axis represents the numbers of draws, and the vertical axis has the values of the averages. The horizontal line indicates the average in the box, 3.5.



What we see is that for a few draws, the average wanders a good distance from the average in the box. For more draws, the average tends to become closer to the 3.5, but it does still wander around a bit.

In general, the more draws, the closer the average of the draws tends to be to the average in the box. This is again the **law of averages**, but for any box:

When drawing with replacement from any box, the **more draws** you take the **closer the average** in the draws is likely to be to the average in the box.

Summarizing the law of averages

When drawing with replacement from a box, the more draws you take

- The closer the percentage in the draws is likely to be to the percentage in the box (if the box has just 0's and 1's), and
- The closer the average in the draws is likely to be to the average in the box.
- What you get in the first few (or many) draws does not affect what happens later.

The standard error of the average of the draws

The law of averages says that the more draws you take with replacement from a box, the closer the average of the draws is likely to be to the average in the box. But how close is close? The standard error of the average of the draws gives a measure of how close. Remember:

$$\text{Average of the draws} = \frac{\text{Sum of the draws}}{\# \text{ of draws}}.$$

The formula for the standard error of the average when drawing with replacement is

$$\text{SE of average of the draws} = \frac{\text{SE of sum of the draws}}{\# \text{ of draws}}.$$

From page 214,

$$\text{SE of sum of the draws} = \sqrt{\# \text{ of draws}} \times (\text{SD in box}).$$

So

$$\text{SE of average of the draws} = \frac{\sqrt{\# \text{ of draws}} \times (\text{SD in box})}{\# \text{ of draws}}.$$

As we did for the standard error of the percentage, we can simplify the formula. Looking at the square roots:

$$\frac{\sqrt{\# \text{ of draws}}}{\# \text{ of draws}} = \frac{1}{\sqrt{\# \text{ of draws}}}.$$

That means we can simplify the standard error formula to be

$\text{SE of average of the draws} = \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}}.$
--

Example

We draw 10 with replacement from . Some possible results:

10 Draws	Average of Draws	Average in Box	Error
2 5 2 2 5 3 5 1 5 4	3.4	3.5	-0.1
1 1 6 6 2 2 4 4 6 5	3.7	3.5	0.2
1 1 2 4 1 3 1 4 4 5	2.6	3.5	-0.9
3 6 6 4 5 5 4 1 1 6	4.1	3.5	0.6
5 2 1 3 1 4 4 5 1 5	3.1	3.5	-0.4

To find the standard error of the average, we first need the SD in the box, which is 1.71^1 . Then

$$\text{SE of average of the draws} = \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}} = \frac{1.71}{\sqrt{10}} = \frac{1.71}{3.162} = 0.54.$$

This SE should seem reasonable, given the sizes of the errors are between 0.1 and 0.9.

? Find the standard error of the average when drawing 25 with replacement.

¹Check that this is correct.

Summary of standard errors

We have seen three formulas for the standard error when drawing with replacement, depending on what we are doing with the draws — finding a sum, a percentage, or an average:

$$\text{SE of the **sum** of the draws} = \sqrt{\# \text{ of draws}} \times (\text{SD in the box})$$

$$\text{SE of the **percentage** of 1's in the draws} = \frac{\text{SD in the box}}{\sqrt{\# \text{ of draws}}} \times 100$$

$$\text{SE of the **average** of the draws} = \frac{\text{SD in the box}}{\sqrt{\# \text{ of draws}}}$$

Warning

These standard errors are for drawing **with replacement**. The standard errors when drawing without replacement require some more work.

Standard errors when drawing without replacement

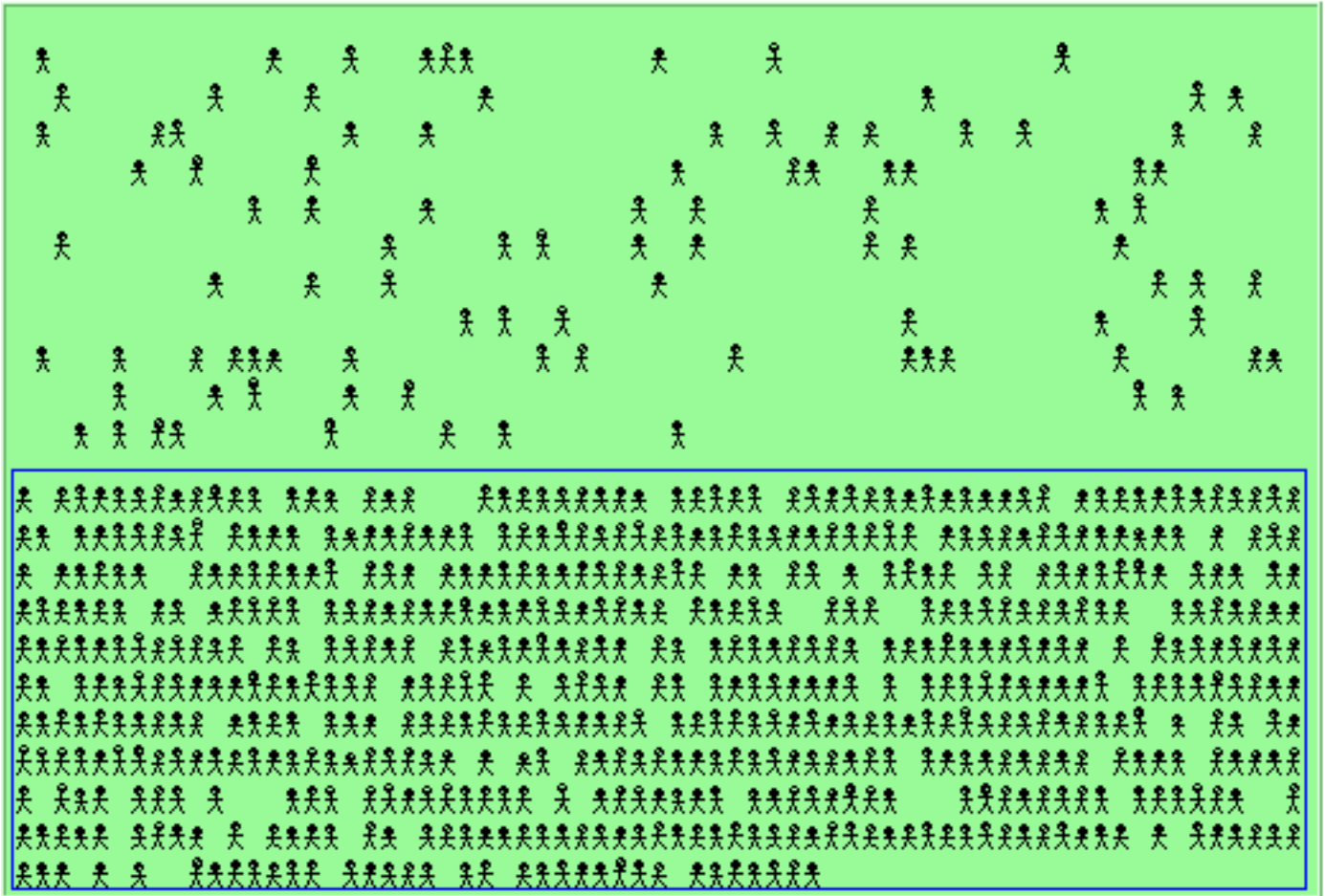
When doing survey sampling, such as a Gallup poll, we draw **without replacement**, because we do not need to ask the same person their opinion twice. So we cannot directly use the formulas from the previous page. But we still would like to know how accurate polls are.

Recall that taking a simple random sample of 100 people from a population is like taking 100 draws without replacement from the box containing the population:

Draws = Sample

Box = Population

In the box of students below, 68% of the people are female. In the sample, 61% are female. So the error in percent is $61\% - 68\% = -7\%$.



Correction Factor

To find the standard error for draws without replacement, we need a **correction factor**.

$$\left(\text{SE without replacement} \right) = (\text{SE with replacement}) \times (\text{Correction factor}).$$

The exact formula for the correction factor is

$$\text{Correction factor} = \sqrt{\frac{(\# \text{ Tickets in box}) - (\# \text{ Draws})}{(\# \text{ Tickets in box}) - 1}}.$$

Example

The box (population) has 712 tickets, 68% are 1's, and 32% are 0's. We take 100 draws (in the sample) without replacement.

Then to find the standard error of the percentage in the draws when drawing **without** replacement, we first find the standard error of the percentage when drawing **with** replacement. In the box, the fraction of 1's is 0.68, and the fraction of 0's is 0.32. So

$$\text{SD in box} = \sqrt{0.68 \times 0.32} = 0.47.$$

Then the SE of the percentage when making 100 draws with replacement uses the square root law for percentages:

$$\begin{aligned} \left(\begin{array}{c} \text{SE of percentage in draws} \\ \text{with replacement} \end{array} \right) &= \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}} \times 100 \\ &= \frac{0.47}{\sqrt{100}} \times 100 = 4.7\%. \end{aligned}$$

We are **not** done. We need to use the correction factor to get the standard error for drawing **without** replacement.

We have that

$$\left(\begin{array}{c} \text{SE of percentage in draws} \\ \text{with replacement} \end{array} \right) = 4.7\%.$$

The formula for the standard error **without** replacement is

$$\left(\begin{array}{c} \text{SE without} \\ \text{replacement} \end{array} \right) = (\text{SE with replacement}) \times (\text{Correction factor}),$$

where

$$\text{Correction factor} = \sqrt{\frac{(\# \text{ Tickets in box}) - (\# \text{ Draws})}{(\# \text{ Tickets in box}) - 1}}.$$

Our box has 712 tickets, and we are drawing 100 tickets. So

$$\text{Correction factor} = \sqrt{\frac{712 - 100}{712 - 1}} = 0.93.$$

Then to find the standard error when drawing without replacement, we multiply that by the Correction factor, which is 0.93:

$$\begin{aligned} \left(\begin{array}{c} \text{SE without} \\ \text{replacement} \end{array} \right) &= (\text{SE with replacement}) \times (\text{Correction factor}) \\ &= (4.7\%) \times 0.93 = 4.4\%. \end{aligned}$$

So the standard error when drawing without replacement is a little smaller than when drawing with replacement.

If the number of draws is a small fraction of the number of tickets in the box, then the correction factor is almost 1, so does not have much effect:

# of tickets in the box	# of draws	Correction factor
200	100	0.71
500	300	
1,000	100	0.95
10,000	100	0.995
10,000	500	0.97
35,000	1,000	0.986
100,000	1,000	0.995
1,000,000	10,000	

Most polling is done with large populations, e.g., the entire US, or Illinois, or Champaign-Urbana, or University of Illinois, so the correction factor doesn't need to be used. It is only when the sample is about $\frac{1}{20^{th}}$ or more of the size of the population that you have to use it.

For example, with 35,000 at the U of I: If you take a sample of 500 or 1,000, you need not use the correction factor. If you take a sample of 2,000 or more, you should use the correction factor.

If you take a sample from all of Champaign County, any practical sample size (a few hundreds or even thousands) will not need the correction factor.

If you (or Gallup) are taking a sample from the entire state, or the entire country, you never have to worry about it.

? Fill in the two blanks in the table.

The correction factor: What you need to know

You need to know when to use it:

- Never if you are drawing with replacement.
- When drawing without replacement:
 - It is always OK to use.
 - It is necessary if the number in the sample is about $\frac{1}{20^{th}}$ or more of the size of the population.
 - It is not necessary if the number in the sample is less than about $\frac{1}{20^{th}}$ of the size of the population.

You do not need to remember the formula for the correction factor, but you have to know what to do with it:

$$\left(\text{SE without replacement} \right) = (\text{SE with replacement}) \times (\text{Correction factor}).$$

10.3 Confidence intervals for percentages

In a May 2011 Gallup Pool, 53% of the respondents said they favored gay marriage. What does that mean? Gallup took a probability sample of 1,018 people from the population of US adults. There are over 200 million adults in the US. Why would people be interested in just the 1,018 interviewed? Did the interviewers call them just to be friendly?

The 1,018 people in the sample are representing the whole population. We are really interested in the percentage of the **entire US adult population** that supports gay marriage. We are not particularly interested in those sampled, though they may indeed be fine people.

So why not interview everyone in the US? There are too many.

Estimating the percentage in the population

Although we are really interested in the percentage of the 200 million in the US that supports gay marriage, we know only the percentage in the sample of 1,018 people. So we use the percentage in the sample to estimate the percentage in the population.

Percentage in sample \approx Percentage in population.

Is that a good estimate? How close is the percentage in the sample to the percentage in the population? That is what the standard error of the percentage measures.

Relate the situation to a box model. A simple random sample² is like drawing without replacement from a box, where

Box = Population, Draws = Sample.

We draw 1,018 without replacement from the box containing every adult in the US. So the question is

How close is the percentage in the draws to the percentage in the box?

²The Gallup poll is a probability sample, but not exactly a simple random sample. But we will approximate the results as if it were a simple random sample.

We know that the standard error of the percentage measures how close the percentage in the draws is likely to be to the percentage in the box. Recall the steps to find this standard error. First, remember that we are drawing **without replacement**, so we may need the correction factor:

$$\left(\text{SE without replacement}\right) = (\text{SE with replacement}) \times (\text{Correction factor}).$$

In this case, since the sample size (1,018) is a tiny fraction of the population size (over 100 million), we really do not need the correction factor.

We know also that for percentages,

$$\text{SE with replacement} = \frac{\text{SD in the box}}{\sqrt{\# \text{ of draws}}} \times 100.$$

The box has 0's and 1's, lots and lots of them. So for this box,

$$\text{SD in the box} = \sqrt{(\text{Fraction of 1's in box}) \times (\text{Fraction of 0's in box})}$$

But we don't know the fraction of 1's or 0's in the box, just in the draws. So we use the fractions in the draws. In the sample (= draws), 53% are 1's, so the fraction is 0.53. Also, 47% are 0's, so the fraction is 0.47. Then the SD is estimated

$$\begin{aligned} & \text{SD in the box} \\ &= \sqrt{(\text{Fraction of 1's in box}) \times (\text{Fraction of 0's in box})} \\ &\approx \sqrt{(\text{Fraction of 1's in draws}) \times (\text{Fraction of 0's in draws})} \\ &= \sqrt{0.53 \times 0.47} = 0.50. \end{aligned}$$

Then

$$\begin{aligned} \text{Standard error of percentage in the draws} &= \frac{\text{SD in the box}}{\sqrt{\# \text{ of draws}}} \times 100 \\ &\approx \frac{0.50}{\sqrt{1018}} \times 100 = 1.57\%. \end{aligned}$$

So the percentage in the sample is likely to be off by 1.57% from the percentage in the population.

Confidence interval for the Gallup poll

To recap, in the Gallup Poll, 53% of the respondents said they favored gay marriage. They took a probability sample of 1,018 people from the population of US adults. Using some approximations, we ended up with

SE of percentage in sample $\approx 1.57\%$.

That means that we expect the poll to be accurate to within $\pm 1.57\%$ or so.

More precisely, the normal curve says that

- About 68% of the time, the percentage in the sample will be within \pm (SE of percentage) of the percentage in the population.
- About 95% of the time, the percentage in the sample will be within $\pm 2 \times$ (SE of percentage) of the percentage in the population.

Going $\pm 2 \times$ (SE of percentage) from the sample percentage gives what is called an approximate **95% confidence interval for the population percentage**:

(Percentage in the sample) $\pm 2 \times$ (SE of percentage)

 is

$$\begin{aligned}
 53\% \pm 2 \times 1.57\% &\rightarrow 53\% \pm 3.14\% \\
 &\rightarrow (53\% - 3\%, 53\% + 3\%) \\
 &\rightarrow (50\%, 56\%).
 \end{aligned}$$

So we say

An approximate 95% confidence interval for the percentage of people in the population that support gay marriage is 50% to 56%.

Thus we can be fairly confident support in the population is somewhat over 50%, but not anything like 60% or more.

? In 2012, Gallup conducted a similar poll of 1024 people. In this sample, 50% said they favored gay marriage.

How many people are in the box?

How many draws are there?

Estimate the SD in the box.

Find the standard error of the percentage when drawing with replacement.

Find the standard error of percentage when drawing without replacement. (Do you need the correction factor?)

Find an approximate 95% confidence interval for the percentage in the population that favors gay marriage.

Is it plausible that about half the population favors gay marriage?

Is it plausible that under 45% favor gay marriage?

What is a confidence interval?

Now we have that (in 2011)

An approximate 95% confidence interval for the percentage of people in the population that support gay marriage is 50% to 56%.

A couple of important points:

- The confidence interval is **calculated** using the **sample**.
- The confidence interval is **for** the percentage in the **population**.

You do not need a confidence interval for the percentage in the sample, since you know the percentage in the sample.

You need it for the percentage in the population because you do not know the percentage in the population.

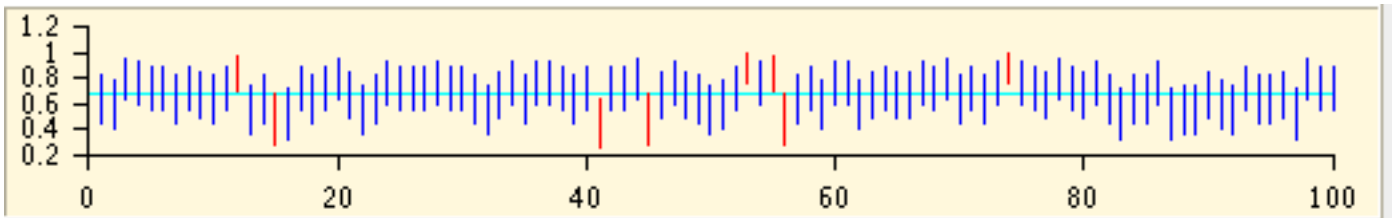
A 95% confidence interval, like the (50%, 56%) for the previous example, gives an idea of what the percentage in the population is. We do not know the percentage in the population, but we hope that it is in the confidence interval.

- The percentage in the population may or may not be in the confidence interval.
- We do not know whether it is or not.
- But about 95% of the time, the confidence interval will contain the percentage in the population.
- The confidence interval is random: Different draws will give different confidence intervals.

Confidence intervals are random

Confidence intervals are random. Sometimes they contain the population percentage, sometimes they don't. For an example, consider the population of 712 students, 68% percent of whom are women (the 1's), and 32% of whom are men (the 0's). If we take a sample of 25 from the population, and find a 95% confidence interval for the population percentage, the population percentage might be in the interval, or it might not.

In the plot, the left-most vertical line is the confidence interval based on a sample of 25 (in terms of proportions, not percentages). It goes from 0.46 to 0.82. The population proportion is 0.68, so this interval is good, 0.68 is in $(0.46, 0.82)$.



On the hand, the twelfth line from the left is based on another sample of 25. Now the interval $(0.70, 0.98)$. It does **not** cover the population proportion of 0.68. It is not a good interval.

The plot has 100 intervals. Notice they are random, that is, different samples give different confidence intervals. Also, some are correct, some are not. Among these, 92% are correct. If we took many many many such samples, approximately 95% would be correct.

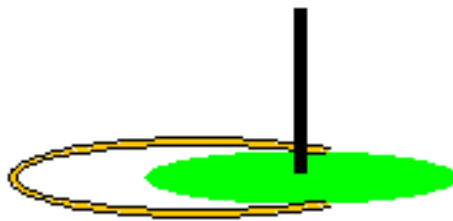
In practice, we only take one sample, and we do not know whether it is a good one or not. We hope so, and there is reason to hope, but we cannot be sure.

Horseshoes

Horseshoes provides an analogy to confidence intervals. You toss a horseshoe at a stake in the ground, trying to have the horseshoe capture the stake. The analogy to confidence intervals is:

- Horseshoe = Confidence interval
- Stake = Population percentage
- One toss = Confidence interval calculated from a sample

The confidence interval (horseshoe) is random, not the population percentage (stake). One toss may capture the stake, which means the confidence interval covers the population percentage (good):



Another toss may miss, meaning the confidence interval misses the population percentage (bad):



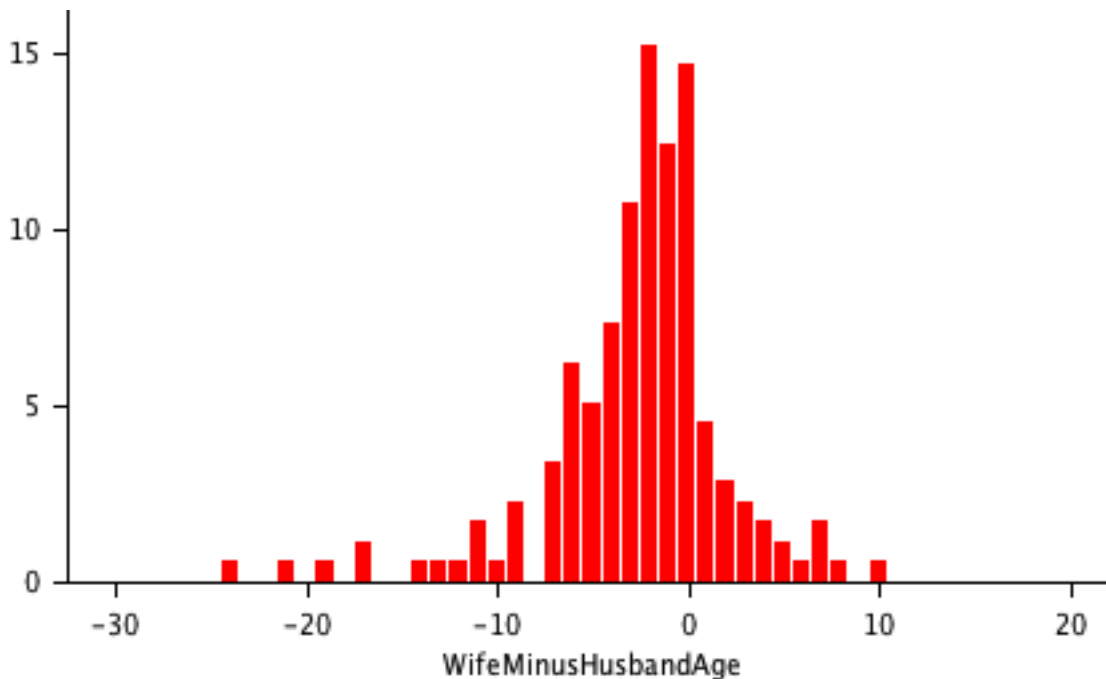
At least you hope you can capture the stake 95% of the time.

10.4 Confidence intervals for averages

Are husbands on average older than their wives? A simple random sample of 177 couples from Illinois was taken. We'll look at the difference in ages:

Wife's Age — Husband's Age

for each of the 177 couples. The histogram:



The average of these differences is -2.58, and the SD is 4.76. So for these 177 couples, the wives are on average 2.58 years younger than their husbands.

What can we say about the population of all couples in Illinois? Are wives on average about 2 or 3 years younger than their husbands?

The Box Model

The sample had 177 couples, selected from the population of Illinois. Since in general,

Box = Population

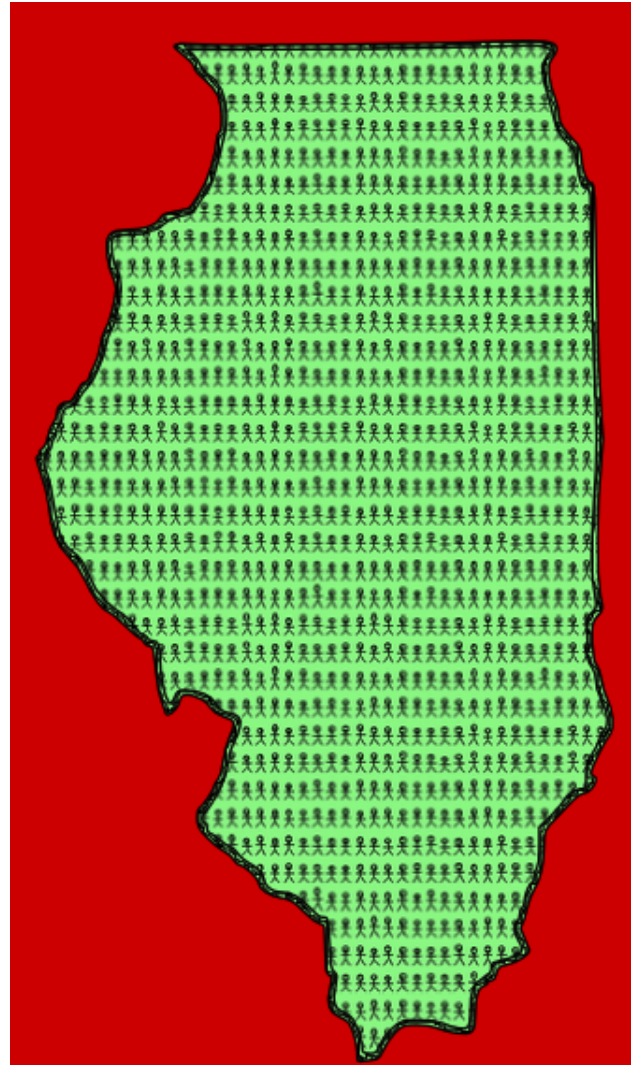
and

Draws = Sample,

the box must contain all millions of Illinois couples →

The draws are the 177 sampled values of

Wife's Age — Husband's Age.



Here are the draws:

```
-3 0 1 5 -2 -5 6 -13 -4 -6 0 0 -4 -11 -4 -3 -4 3 0 1 -6 0 -3 -1 -2 -7 -2
-10 -9 -4 -6 0 -6 8 -17 -2 3 -3 0 -6 -4 -2 -2 -2 -1 0 -5 -1 -3 -5 -2 2 -6
2 -1 0 -6 -3 2 -7 -11 1 0 -2 -3 -4 10 0 -4 -9 -2 -9 0 -1 -24 -2 7 -1 -5
-1 -5 -1 0 -4 4 0 -2 -21 0 -1 -7 -6 -2 -6 -7 0 -2 -1 -1 -1 -3 -5 -3 -4 1
-19 -2 -4 0 -3 0 -3 -12 3 -3 -2 -7 4 -11 -3 0 -4 -6 -7 -4 -2 0 0 -3 -3 -2
0 -1 -3 5 -9 -1 0 -2 -2 -2 1 -2 1 -6 -1 2 0 3 -5 -2 -5 0 7 1 -2 0 -1 1 -3
-5 -14 -1 -3 -1 -17 2 -1 7 -3 -1 -1 -2 -1 -2 4 -2
```

So some are positive and some negative. Then we can calculate:

Average of the draws = -2.58 , SD of the draws = 4.76 .

The confidence interval for the Illinois couples

The 177 draws had

$$\text{Average of the draws} = -2.58, \quad \text{SD of the draws} = 4.76.$$

But we are really interested in the average in the box = population. How close is the average in the draws likely to be to the average in the box? On page 267, we have that the standard error of the average of the draws when drawing **with replacement** is

$$\text{SE of average of draws} = \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}}.$$

Since we are drawing **without replacement**, we have to consider the correction factor. Do we need it? The number of draws is 177, and the number of tickets in the population is millions. So, no, we do not need the correction factor.

We do not know the box exactly, so we don't know the SD in the box. But we can estimate with the SD in the draws, which is 4.76. So

$$\text{SE of average of draws} \approx \frac{\text{SD in draws}}{\sqrt{\# \text{ of draws}}} = \frac{4.76}{\sqrt{177}} = 0.36.$$

An approximate 95% confidence interval for the average in the population is then

$$\boxed{\text{Average of the draws} \pm 2 \times (\text{SE of the average})}$$

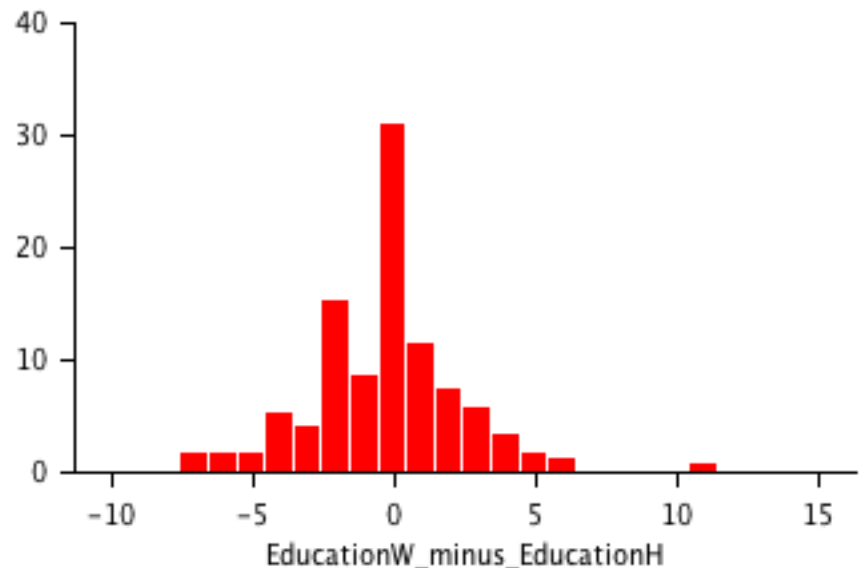
For the data, the interval is

$$\begin{aligned} -2.58 \pm 2(0.36) &\rightarrow -2.58 \pm 0.72 \\ &\rightarrow (-2.58 - 0.72, -2.58 + 0.72) \\ &= (-3.30, -1.86). \end{aligned}$$

We are about 95% confident that the average difference in ages of wives and husbands for the entire population of Illinois couples is -3.3 years to -1.86 years. So we are fairly confident that the wives are, on average, about 2 to 3 years younger than their husbands.

?

Here is the histogram for the differences in education for the 177 Illinois couples, where the values are wife's education minus husband's education. The average difference is -0.24 , and the SD is 2.6 . On average, in this sample, who has more education, the husbands or the wives?



Assuming this sample is a simple random sample from the population of Illinois couples, what is the estimated standard error of the average of the sample? (Do you need the correction factor?)

What is the approximate 95% confidence interval?

The confidence interval is for which average?

Based on this interval, would you say the wives surely have a lower educational level than their husbands, or it is plausible they have the same educational level, or it is certain they have the same educational level?

Summary

Now we have that an approximate 95% confidence interval for the average in the population is

$$\text{Average of the draws} \pm 2 \times (\text{SE of the average})$$

Important, as before:

- The confidence interval is **calculated** using the **sample**.
- The confidence interval is **for** the average in the **population**.

You do not need a confidence interval for the average in the sample, since you know the average in the sample. You need it for the average in the population because you do not know the average in the population.

Hypothesis Testing

Confidence intervals give you an idea of the percentage in the population (box), or the average in the population (box), when you do not know exactly what is in the box. The confidence interval for the average wife's age minus husband's age of Illinois couples was $(-3.30, -1.86)$. We are about 95% confident that on average, in Illinois, the wives are from 3.3 to 1.86 years younger than their husbands. A significance test answers a more specific yes-no question: Are wives, on average, younger than their husbands? It does look like they are.

The questions asked in significance testing are yes-no questions: The answer you want is "yes" or "no," not an interval. Examples:

- Are wives, on average, younger than their husbands?
- Is Obama's national approval rating 50%?
- Does the polio vaccine work?
- Is drinking diet sodas associated with obesity?
- Is a particular roulette wheel biased?
- Does the class have ESP?

11.1 The null hypothesis

The first step in performing a significance test is to set up the null hypothesis. The null hypothesis is what would be true if nothing were going on, that is, if the status quo remains in effect.



Examples

- In the 2012 presidential election, Gallup takes a poll of Obama vs. Romney. The null hypothesis: Nationally, they are tied.
- Testing the polio vaccine. The null hypothesis: The vaccine is no better than the placebo.
- Any kind of a test of a drug's or treatment's effectiveness. The null hypothesis: The drug is no better than the placebo.
- A casino is worried that one of its roulette wheel is biased. The null hypothesis: It is not biased. It is OK.

More examples:

- The draft lottery. Some people think the lottery process gives lower numbers to men born later in the year. The null hypothesis: The lottery is totally random. There is no bias for or against any birthdate.
- Is drinking diet sodas associated with obesity?
? What is the null hypothesis?
- Does the class have ESP?
? What is the null hypothesis?

In all the examples the null hypothesis is that there is nothing to get excited about.

Sometimes, you may want the null hypothesis to be true, for example, you are a chemical company and you hope that the null hypothesis that your **chemical does not cause cancer** is true.

Sometimes you may want the null hypothesis to be false, for example, you are a drug company and you have a new drug to fight impotence. You want the null hypothesis that **the drug has no effect** to be false.



Does caffeine harm memory?

Claire Henson, Bridget Rogers, and Nadia Reynolds, students at University of Illinois Laboratory High School (Uni High), conducted an experiment to see whether caffeine has a negative effect on short-term visual memory. Subjects were randomly chosen from each of three groups of the Uni High students: 9 chosen from the eighth graders, 10 from the tenth graders, and 9 from the twelfth graders. Each person was tested once after having caffeinated Coke, and once after having decaffeinated Coke. After each drink, the person was given ten seconds to try to memorize twenty small, common objects, then allowed a minute to write down as many as could be remembered.

The main question of interest is whether people remembered more objects after the Coke without caffeine than after the Coke with caffeine.

Focus on the twelfth graders. The question is whether caffeine hurts the performance of the entire twelfth grade (there are 57 of them), on average. The null hypothesis is that nothing is going on.

Null hypothesis: Caffeine does not affect memory in the population of twelfth graders.

One could possibly conduct the experiment on all 57 students, in which case you could see whether the average number of objects remembered with caffeine is the same as without caffeine. But since that may be impractical, the students took a simple random sample of 9 from the population.

So we have a box model:

Box = Population of 57 twelfth graders;
Draws = Sample of 9 drawn without replacement.

The data

The table gives the results for the nine students in the sample:

# objects remembered		
Without caffeine	With caffeine	Difference
7	7	0
8	6	2
9	6	3
11	7	4
5	5	0
9	4	5
9	7	2
11	8	3
10	9	1

Each row is for one student. The key numbers are the differences, which are how many more objects they remembered without caffeine than with:

$$\text{Difference} = (\# \text{ without caffeine}) - (\# \text{ with caffeine}).$$

Notice that all the differences are either zero or positive. So among the sample, on average people did better without caffeine. But again, we are not particularly interested in the nine people in the sample, but rather the 57 people in the population.

The question is then “What can we say about the population, based on the sample?”

11.2 The null box

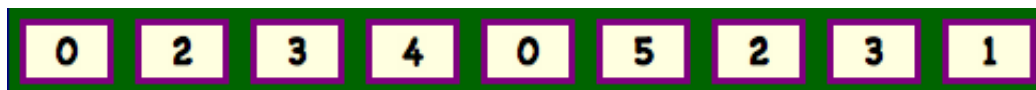
The box is the population — The 57 twelfth graders at Uni High. The box contains 57 tickets, and on each ticket is written the difference in number of objects remembered (without — with) that person would get. Before the experiment begins, we do not know exactly what is on the tickets:



The null hypothesis is that nothing is going on, i.e., the number of objects remembered is not affected by whether or not one had caffeine. So if the null hypothesis is true, the average in the box is zero:

Null hypothesis: The average in the box is 0.

The sample consists of the nine draws, drawing without replacement. The draws are the differences, from the previous page:



Key question: Could these draws have plausibly come from the null box? Do they look like a sample from a box with an average of zero? Maybe, but maybe not. There are no negatives among the draws, whereas in the null box the negatives and positives are supposed to cancel out to average zero.

11.3 The logic of hypothesis testing

Testing whether the hypothesis is true involves a specific logic. You start by imagining that the null hypothesis is indeed true. In the example, you would imagine that the average in the box is zero. (You pretend you haven't seen the data yet.)

Close your eyes and imagine that the null hypothesis is true.

The null hypothesis is true.
The null hypothesis is true.



Open your eyes and look at the data, that is, the draws.

0	2	3	4	0	5	2	3	1
---	---	---	---	---	---	---	---	---



What is your reaction? Does the data conflict with the null hypothesis?

If those draws look like they very well could have come from the null box, you are calm.

You then **fail to reject the null hypothesis**. The null hypothesis is plausible in light of the data.



On the other hand, if those draws do not look like they would come from the null box, you freak out a bit.

You then **reject the null hypothesis**. The null hypothesis is not plausible in light of the data.



The logical steps in the caffeine example

Each student in the sample was given a memory test. The data are the differences between the numbers of objects they remembered without caffeine and with caffeine.

- **Figure out what the box would be if the null hypothesis were true.**

The null hypothesis: Caffeine does not affect memory in the population.

If this null hypothesis is true, then the average in the box would be 0. So

Null box: The average of tickets = 0.

- **While assuming the null hypothesis is true, look at the draws.**

Here are the draws: 0, 2, 3, 4, 0, 5, 2, 3, 1.

The average of these draws is 2.22. (So on average, the twelfth graders in the sample remembered 2.22 more objects when they didn't have caffeine than when they did.)

Key question: Could these draws have plausibly come from the null box? Do they look like a sample from a box with an average of zero?

- If there is no conflict, i.e., if you could have probably gotten those draws from the null box, you **fail to reject the null hypothesis**.
- If there is a conflict, i.e., it is not likely you could have gotten those draws from the null box, you **reject the null hypothesis**.

How do you decide whether there is a conflict between the draws and the null box? In this example, there may indeed be a conflict. If the average in the box were really zero, you wouldn't expect all the draws to be zero or above. The *p-value*, which comes next, is a guide.

11.4 The p-value

The null box has an average of 0. The draws are 0, 2, 3, 4, 0, 5, 2, 3, 1, which has an average of 2.22. Is the 2.22 too far away from 0 for the box to be plausible?

To measure how much the data conflicts with the null box, we use what is called a **p-value**. The p-value is the chance of getting an average this far (2.22) or farther from the box average of zero:

$$\text{p-value} = \left(\begin{array}{l} \text{Chance that the average in the draws} \geq 2.22 \\ \text{when drawing from a box with average 0.} \end{array} \right)$$

We use the normal curve, so we have to put the average in the draws into standard units. The standard units we use are calculated **as if drawing from the null box**. We call the standard units Z :

$$\begin{aligned} Z &= \text{Standard units} \\ &= \frac{(\text{Average in the draws}) - \left(\begin{array}{l} \text{Expected Value of the} \\ \text{average from the null box} \end{array} \right)}{\text{Standard error of the average}}. \end{aligned}$$

The expected value of the average is just the average in the box, which is 0. The standard error is

$$\text{SE of the average} = \frac{\text{SD in the box}}{\sqrt{\# \text{ of draws}}} \times (\text{Correction factor}).$$

We need the correction factor because the number of draws, 9, is a reasonably large fraction of the number in the box, 57.

Since we do not know the SD in the box, we use the SD in the draws:

$$\text{SD of } \{0, 2, 3, 4, 0, 5, 2, 3, 1\} = 1.62.$$

The correction factor is 0.93, so

$$\begin{aligned} \text{SE of the average} &= \frac{\text{SD in the box}}{\sqrt{\# \text{ of draws}}} \times (\text{Correction factor}) \\ &\approx \frac{1.62}{\sqrt{9}} \times 0.93 \\ &= 0.50. \end{aligned}$$

Again, the Z is

$$Z = \frac{(\text{Average in the draws}) - \left(\begin{array}{l} \text{Expected Value of the} \\ \text{average from the null box} \end{array} \right)}{\text{Standard error of the average}}.$$

And we have that

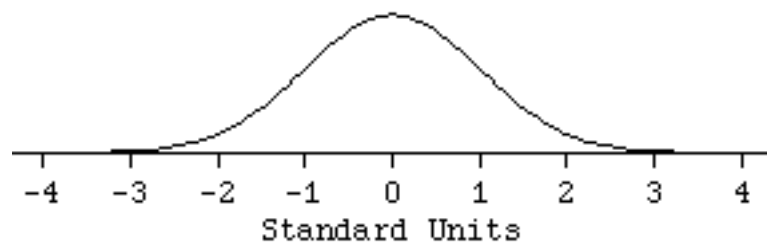
$$\begin{aligned} \text{Average in the draws} &= 2.22, \\ \text{Expected value of the average} &= 0, \\ \text{SE of the average} &= 0.50. \end{aligned}$$

Thus

$$\begin{aligned} Z &= \frac{2.22 - 0}{0.50} \\ &= 4.44. \end{aligned}$$

Is 4.4 a surprising value for something in the normal curve? Yes, you usually expect numbers in the ± 1 or ± 2 range. To measure how surprised you are, you find the area above the 4.44 in the normal curve, which is the p-value:

$$\begin{aligned} \text{p-value} &= \left(\begin{array}{l} \text{Chance that the average in the draws} \geq 2.22 \\ \text{when drawing from a box with average 0.} \end{array} \right) \\ &= \text{Area to the right of 4.44 in the normal curve} \\ &= 0 \end{aligned}$$



We do not even have to go to the normal table to see that there is basically nothing to the right of 4.44.

The logical steps, again

So now we have the p-value, we can review the steps:

- **Figure out what the box would be if the null hypothesis were true.**

Null box: The average of tickets = 0.

- **While assuming the null hypothesis is true, look at the draws.**

Here are the draws: 0, 2, 3, 4, 0, 5, 2, 3, 1.

The average of these draws is 2.22.

The p-value = 0. (Not likely!)



- ~~— If there is no conflict, i.e., if you could have~~
- ~~— probably gotten those draws from the null box, you~~
fail to reject the null hypothesis.
- If there is a conflict, i.e., it is not likely you could have gotten those draws from the null box, you **reject the null hypothesis.**

There **is** a conflict between the draws and the null box, since the chance of getting the draws is essentially 0. Thus we

Reject the null hypothesis,

and conclude that

Caffeine does affect memory for this population, the twelfth graders.

? The data below gives the numbers of objects remembered by 10 tenth graders at Uni High. The first columns shows the number after drinking Coke without caffeine, and the second after drinking Coke with caffeine. The difference is ($\#$ remembered with caffeine) – ($\#$ remembered without caffeine).

# objects remembered		
Without Caffeine	With Caffeine	Difference
6	3	3
9	11	-2
4	4	0
7	6	1
6	8	-2
7	6	1
6	8	-2
9	8	1
10	7	3
10	6	4

The data below is like what we saw before, but now they consist of a simple random sample from the population of 59 tenth graders, instead of twelve graders. We are interested in seeing whether caffeine affects memory for population of the tenth graders.

What is the null hypothesis?

What is the average in the null box?

? (continued) The average of the differences for the sample is 0.7. The SD is 2.1. What would the estimate of the standard error of the average of the draws be if drawing with replacement?

We are drawing without replacement. (Do we need the correction factor here? It is 0.919.) What is the SE of the average for drawing without replacement?

What is the expected value of the average for draws from the null box?

What is the Z?

What is the p-value?

? (continued) What is the p-value you found on the last page?

Is it surprising? Is it greater than or less than 5%?

Do you reject the null hypothesis, or fail to reject?

Which of the following statements is best?

1. In the population, tenth graders do the same with or without caffeine.
2. In the population, tenth graders do better without caffeine than with.
3. It is plausible that in the population, tenth graders do the same with or without caffeine.

For the twelfth graders, the same null hypothesis had a p-value of essentially 0%. Does it look like caffeine affects the tenth graders differently than the twelfth graders?

11.5 Summary



Once we have the null box and the draws, the significance test involved three steps:

- Step 1: Find the Z, which is the average in the draws turned into standard units:

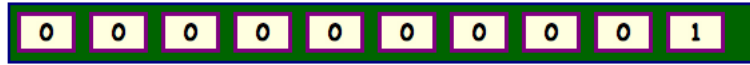
$$Z = \frac{(\text{Average in the draws}) - \left(\begin{array}{l} \text{Expected Value of the} \\ \text{average from the null box} \end{array} \right)}{\text{Standard error of the average}}.$$

- Step 2: Find the p-value, which is the area in the normal curve beyond the Z.
- Step 3: If the p-value is **too small, reject** the null hypothesis. If it is **not too small, you fail to reject** the null hypothesis.

But how small is too small? The common cutoff point is 5%, so if the p-value is less than 5%, you would reject the null hypothesis:

p-value > 5%	p-value < 5%
You could have probably gotten the draws given you drew from the null box.	It is not likely you could have gotten the draws given you drew from the null box.
	
Fail to reject the null hypothesis.	Reject the null hypothesis.

The box model is then to draw 2190 tickets **with replacement** from



The data we are using is the number of 1's among the draws, which is 235, that is,

$$\text{Sum of the draws} = 235.$$

We then need to turn that into standard units. Since we are using the sum of the draws, we need the expected value of the sum and the standard error of the sum.

The average in the null box is $\frac{1}{10}$, and there are 2190 draws, so

$$\begin{aligned} \left(\begin{array}{l} \text{Expected Value of the} \\ \text{sum of the draws} \end{array} \right) &= (\# \text{ of draws}) \times (\text{Average in the box}) \\ &= 2190 \times \frac{1}{10} = 219, \end{aligned}$$

as we said before.

To get the standard error of the sum (drawing with replacement), we use the usual steps:

$$\begin{aligned} \text{SD in the box} &= \sqrt{(\text{Fraction of 1's}) \times (\text{Fraction of 0's})} \\ &= \sqrt{\frac{1}{10} \times \frac{9}{10}} = 0.3. \end{aligned}$$

Then

$$\begin{aligned} \text{SE of the sum} &= \sqrt{\# \text{ of draws}} \times (\text{SD in the box}) \\ &= \sqrt{2190} \times 0.3 = 14.04. \end{aligned}$$

So now we have the important quantities:

$$\begin{aligned}\text{Sum of the draws} &= 235, \\ \text{Expected value of the sum from the null box} &= 219, \\ \text{SE of the sum} &= 14.04.\end{aligned}$$

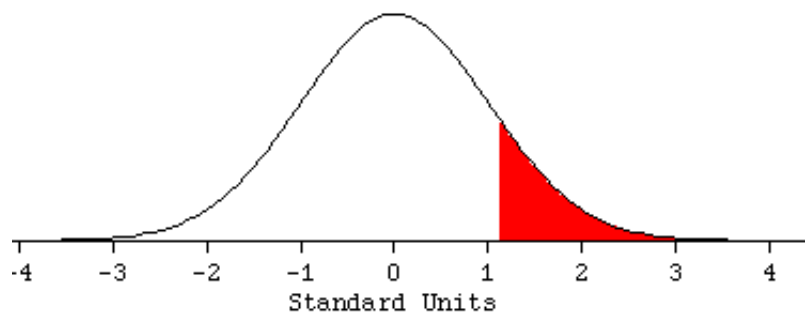
To find the p-value, we first have to find the Z, which is the sum of the draws put into standard units:

$$\begin{aligned}Z &= \frac{\text{Sum of the draws} - \text{Expected value of the sum}}{\text{SE of the sum}} \\ &= \frac{235 - 219}{14.04} = 1.14.\end{aligned}$$

(That value does not appear anything to get executed about.) The p-value is then the area in the normal curve above 1.14.

If we use the normal table, we'd find the area between ± 1.14 is about 74.99%, and the area above 1.14 is about

$$\frac{100 - 74.99}{2} = 12.5\%.$$



This p-value of 12.5% is not that small (larger than 5%), so we are OK with the null box. We **fail to reject** the null hypothesis, which means there is no reason to suspect the lotto is biased.



Summary of the calculations

In this example, we used the sum of the draws. So the significance test involved the following three steps:

- Step 1: Find the Z , which is the sum of the draws turned into standard units:

$$Z = \frac{(\text{Sum of the draws}) - \left(\begin{array}{l} \text{Expected Value of the} \\ \text{sum from the null box} \end{array} \right)}{\text{Standard error of the sum}}.$$

- Step 2: Find the p-value, which is the area in the normal curve beyond the Z .
- Step 3: If the p-value is too small, reject the null hypothesis. If it is not too small, you fail to reject the null hypothesis.

Summary of the logic

Start with the null hypothesis, which means nothing is going on.

Close your eyes, and imagine that the null hypothesis is true.

Open your eyes, and look at the data.

Keeping in mind that you are imagining that the null hypothesis is true, what is your reaction?

You could have probably gotten the draws given you drew from the null box.	It is not likely you could have gotten the draws given you drew from the null box.
Not surprised.	Surprised!
Fail to reject the null hypothesis.	Reject the null hypothesis.

Interpretation of “Fail to reject”

What does “Fail to reject the null hypothesis” mean?

- It does **not** mean we are sure the null hypothesis is true.
- It means either the null hypothesis is true, or it is close to being true, or we just do not have enough evidence to reject it.

For example, with the Illinois Lottery, the null hypothesis was that the lottery was totally random. We looked at the nines that came up in Pick 3 for 2011, and nothing seemed amiss. We failed to reject the null hypothesis. So either the lottery is totally random, or we did not have enough evidence to say it was biased. It still might be a little biased. Additionally, we could look at the 0's, 1's, ..., and 8's. Or look at other years. Or at the Pick 4 and other games. Then we might find evidence of bias. We do not know.

Criminal Court Cases

In a criminal court case, what does a verdict of “Not guilty” mean?

- It does **not** mean the accused is found innocent.
- It means either the accused is innocent, or that there is not enough evidence to convict.

So “fail to reject” is somewhat analogous to “find not guilty.” We never say “the null hypothesis is true” nor “the accused is innocent.”

Here is a table with the analogy:

Significance test	Court case
The null hypothesis	The accused is innocent
Statistician imagines the null hypothesis is true	Jury presumes the accused is innocent
Statistician looks at the data (draws)	Jury considers the evidence
Reaction?	
The data could plausibly have come from the null box: Fail to reject the null hypothesis	It is within reason that the evidence could be consistent with innocence: Not guilty!
No way the data could have come from the null box: Reject the null hypothesis	It is beyond a reasonable doubt that the evidence is not consistent with innocence: Guilty!

11.6 Student's t

Significance testing using the Z relies on a couple of approximations being reasonable:

- The probability histogram of the sum of the draws should be close to a normal curve.
- If we do not know what the SD in the box is, then we need that

$$\text{SD in the draws} \approx \text{SD in the box.}$$

These approximations work well if the number of draws is not too small, and the histogram of the box is not too lopsided. A more accurate approximation than the Z in certain situations uses what is called *Student's t*.

Student was the pseudonym of W. S. Gossett, who needed a better approximation for small experiments he was doing to make better Guinness beer.

He published a paper in 1908 presenting his "t," but his employers did not want to tip off their competitors they were using statistics, so he used the pseudonym.



When to use Student's t

Student's t is useful if

- The number of draws is less than 25;
- The histogram of the box looks somewhat like a normal curve; and
- The SD in the box is not known.

There are two small changes one makes in going from the Z to the t:

- Modify the standard error of the average of the draws a little.
- Use the Student's t table, rather than the normal table.

The box model here is that we draw a certain number with replacement from a box, or we draw without replacement but do not need the correction factor. Recall the Z for testing a null hypothesis about the average of the box:

$$Z = \frac{(\text{Average of the draws}) - (\text{Average in the null box})}{\text{SE of the average of the draws}}.$$

Student's t uses what we will call the SE^+ of the average, which is a modification of the usual standard error, making it a little larger:

$$\left(\begin{array}{c} \text{SE}^+ \text{ of the average} \\ \text{of the draws} \end{array} \right) = \left(\begin{array}{c} \text{SE of the average} \\ \text{of the draws} \end{array} \right) \times \sqrt{\frac{\# \text{ of draws}}{\# \text{ of draws} - 1}}.$$

Then Student's t is like the Z but using the SE^+ instead of the SE:

$$T = \frac{(\text{Average of the draws}) - (\text{Average in the null box})}{\text{SE}^+ \text{ of the average of the draws}}.$$

Example

The table below has the scores of five randomly selected students from a class of 107. The table contains their average scores on the midterm exams (three of them) and their scores on the final exam, as well as the difference between their final exam scores and the midterms. The question is whether for the class as a whole, the average on the final is lower than the average on the midterms. So the null hypothesis is that the average scores are the same in the population.

Midterms	Final	Final—Midterms
84.9	68	-16.9
72.13	60	-12.13
86.9	66	-20.9
68.81	42	-26.81
64.29	70	5.71

The box is the entire class. On each ticket is written the difference between the final and midterms. The null hypothesis is that the average difference **in the entire class** is zero:

Null hypothesis: Average in the box = 0.

We draw five without replacement. We can ignore the correction factor, since 5 is a small fraction of 107. The draws are the given in the last column of the table: -16.9, -12.13, -20.9, -26.81, 5.71. Then

Average of the draws = -14.2 , SD of the draws = 11.06.

The average of the sample is quite negative, meaning these five people did quite a bit worse on the final than the midterm. But what about the whole class? Is it a statistically significant difference?

We first need to find the standard error of the average. The SD of the draws is 11.06, and there are five draws, so (ignoring the correction factor):

$$\begin{aligned} \text{SE of the average of the draws} &= \frac{\text{SD in box}}{\sqrt{\# \text{ of draws}}} \\ &\approx \frac{\text{SD of draws}}{\sqrt{\# \text{ of draws}}} \\ &= \frac{11.06}{\sqrt{5}} = 4.95. \end{aligned}$$

For Student's t, we need to find the SE^+ :

$$\begin{aligned} \left(\text{SE}^+ \text{ of the average of the draws} \right) &= \left(\text{SE of the average of the draws} \right) \times \sqrt{\frac{\# \text{ of draws}}{\# \text{ of draws} - 1}} \\ &= 4.95 \times \sqrt{\frac{5}{5 - 1}} = 5.53. \end{aligned}$$

Then we use the SE^+ in the T:

$$\begin{aligned} T &= \frac{(\text{Average of the draws}) - (\text{Average in the null box})}{\text{SE}^+ \text{ of the average of the draws}} \\ &= \frac{-14.2 - 0}{5.53} = -2.57. \end{aligned}$$

Next, we need to find the p-value for this T.

Degrees of freedom

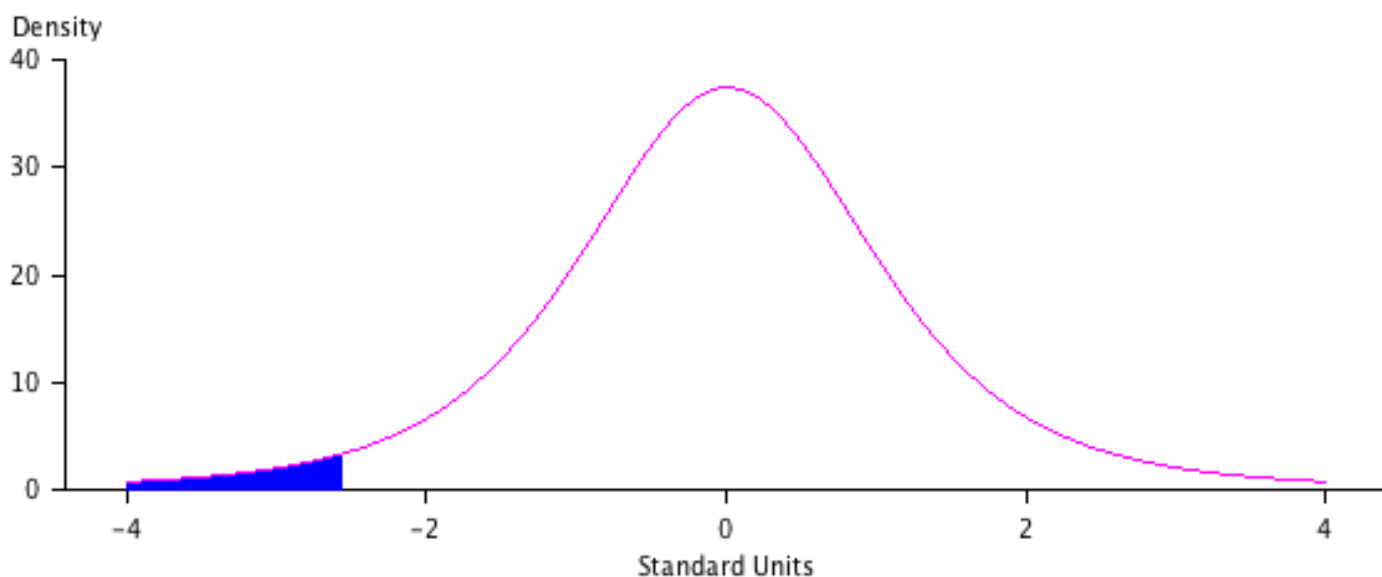
The $T = -2.57$, and we wish to find the p-value. The p-value is the area to the left of -2.57 , but in the **t-curve**, not the normal curve. In fact, there are many t-curves, one for each value of what are called *degrees of freedom*.

In this case, the degrees of freedom are just the number of draws minus one:

$$\text{Degrees of freedom} = (\# \text{ of draws}) - 1.$$

Since we have five draws, the degrees of freedom are four. Here is that curve:

Students t curve for 4 degrees of freedom



It looks quite a bit like the normal curve, but it is not exactly the same. The shaded area is the p-value. It looks small, but we use the t-table to find a numerical value.

Student's t table

Now we have the $T = -2.57$. We do not look at the normal table now, but the *t-table* in the book. The degrees of freedom indicate the row in the table. Since we have five draws, the degrees of freedom are four. Here is that row:

Degrees of freedom ↓	25%	10%	5%	2.5%	1%	0.5%	← p-value
4	0.74	1.53	2.13	2.78	3.75	4.60	← T

The top row gives the p-values, and the values in the second row are the T's. So if $T = 2.13$, the p-value is 5%, and if the $T = 2.78$, the p-value is 2.5%.

Our $T = -2.57$, but we first ignore the sign. Then since 2.57 is between 2.13 and 2.78, the p-value is between 2.5% and 5%. In any case, our p-value is less than 5%, so we can

reject the null hypothesis.

Recall that the null hypothesis is that the people do the same on average in the midterm and final.

Conclusion: In the population (class), we are fairly confident that people do worse on the final than on the midterms.

? It is generally claimed that the average body temperature for healthy human adults is 98.6° F. To test this claim a simple random sample of 9 adults are chosen. The population is large enough so that we do not need to use the correction factor. The average temperature of the 9 adults is only 98.3° F with a SD of 0.6° F. The question is whether the data on these 9 people is sufficient evidence to conclude that average body temperature among the population of all healthy adults is really lower than 98.6.

What is the null hypothesis? (Give it in terms of the average temperature in the [which: population or sample?].)

Find the SE of the average.

Since the sample size is small, we'll use Student's t. Find

$$\sqrt{\frac{\# \text{ of draws}}{\# \text{ of draws} - 1}}.$$

What is SE⁺?

What is the value of the T?

Below is part of the t table. In it, find the correct degrees of freedom. In the row corresponding to the correct degrees of freedom, the calculated T is between which two values?

The p-value is between which two p-values?

Degrees of freedom ↓	25%	10%	5%	2.5%	1%	0.5%	← p-value
5	0.73	1.48	2.02	2.57	3.36	4.03	← T
6	0.72	1.44	1.94	2.45	3.14	3.71	
7	0.71	1.41	1.89	2.36	3.00	3.50	
8	0.71	1.40	1.86	2.31	2.90	3.36	
9	0.70	1.38	1.83	2.26	2.82	3.25	

Do you reject the null hypothesis?

Are the data from the 9 people sufficient evidence to conclude that the average temperature of all healthy adults is really lower than 98.6?

Summary

Student's t is useful if

- The number of draws is less than 25;
- The histogram of the box looks somewhat like a normal curve; and
- The SD in the box is not known.

There are two small changes one makes in going from the Z to the t:

- Change the SE of the average of the draws to the SE⁺:

$$\text{SE}^+ \text{ of the average of the draws} = \left(\text{SE of the average of the draws} \right) \times \sqrt{\frac{\# \text{ of draws}}{\# \text{ of draws} - 1}}.$$

Then Student's t is

$$T = \frac{(\text{Average of the draws}) - (\text{Average in the null box})}{\text{SE}^+ \text{ of the average of the draws}}.$$

- Use the Student's t-table, rather than that for the normal curve, where

$$\text{Degrees of freedom} = (\# \text{ of draws}) - 1.$$

It is OK to use Student's t for any number of draws, if the other two conditions hold. The Z is at its best if

- The number of draws is more than 25 and the histogram of the box is not too lopsided; or
- The histogram of the box looks somewhat like a normal curve and the SD in the box is known.

If the number of draws is less than 25, and the histogram of the box is not at all like a normal curve, then neither the Z nor Student's t should be used. In future statistics courses, you may encounter what are called "nonparametric" procedures, which can be used in such situations.

11.7 Comparing two samples

Often, we wish to compare two groups, or two treatments. The null hypothesis would be that the two groups have the same average, or the drug and the placebo have the same effect.

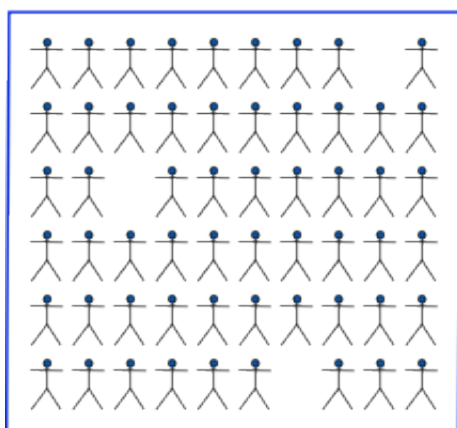
- Compare men and women on the number of shoes they own. Or on their BMI. Or on how they do on exams and assignments. For example, asked what was the fastest they've ever driven a car, the men said on average 97 MPH, and the women said on average 87 MPH. Is that a statistically significant difference?
- Compare the polio vaccine to the placebo. In the gold star study, the rate of polio among the vaccinated group was 28 per 100,000, and among the placebo group the rate was 71. Is that statistically significant? (Yes, it turns out.)
- Compare attitudes on gay marriage from 2011 to 2012. In the 2011 Gallup Poll, 53% of the respondents said they supported gay marriage. In 2012, support was 50%. Is that a statistically significant decrease?

“Statistically significant” means we reject the null hypothesis. If we fail to reject the null hypothesis, the result is not statistically significant.

A box model for two samples

Suppose we wish to compare two populations, say men and women. We could draw two independent samples, one sample from the population (box) of men, and another sample from the population of women.

Draw three from the men



Draw five from the women



The sample from the men is independent of the sample for the women if which men are chosen has nothing to do with which women are chosen. So it is not that the chosen men are the husbands of the chosen women, or the brothers, or the boyfriends. We just have two separate random samples.

Example

We have 227 men and 485 women in a STAT100 class. We are going to assume they are simple random samples from their respective populations of men and women in all STAT100 classes. (They are technically not a simple random sample, but we hope this assumption is close enough.) We are looking at the fastest these people have ever driven a car. For the men's sample, the average is 97 MPH and the SD is 22.8. For the women's sample, the average is 87MPH and the SD is 18.2.

	Men	Women
# of draws	227	485
Average of draws	97	87
SD of draws	22.8	18.2
Average of box (population)	?	?

For the people in the sample, we know that the men have driven faster on average than the women. That is not the question. The question is whether men drive faster in the population of **all** STAT100 students.

The null hypothesis is that nothing is going on: Men and women in the population have the same average speed. Null hypothesis:

$$\text{Average of the men's population} = \text{Average of the women's population.}$$

We look at the difference in the two sample averages, and see if that is significantly different from zero. So the Z is

$$Z = \frac{(\text{Average of men's draws}) - (\text{Average of women's draws})}{\text{SE of what?}}.$$

The SE for two independent samples

We are taking two samples, and finding two averages, so we also have two standard errors, one for the average of the men's draws, and an-

other for the average of the women's draws. In the Z,

$$Z = \frac{(\text{Average of men's draws}) - (\text{Average of women's draws})}{\text{SE of what?}},$$

we have the **difference** between the averages. We want the standard error of the difference:

$$Z = \frac{(\text{Average of men's draws}) - (\text{Average of women's draws})}{\text{SE of the difference in averages}}.$$

If the two samples are independent (drawn independently from two different boxes), then to find the standard error of the difference of their averages we

- Find the standard errors of the two averages.
- Square the two standard errors.
- Add up the squares.
- Take the square root of that sum.

The formula:

$$\left(\begin{array}{c} \text{SE of} \\ \text{difference} \\ \text{in averages} \end{array} \right) = \sqrt{\left(\begin{array}{c} \text{SE of average} \\ \text{of men's draws} \end{array} \right)^2 + \left(\begin{array}{c} \text{SE of average} \\ \text{of women's draws} \end{array} \right)^2}$$

Notice that the two squares are added, even though in the averages we subtract.

The table below contains our data. We need to find the standard error of the average for both groups, men and women.

For the men:

$$\left(\begin{array}{c} \text{SE of average} \\ \text{of men's draws} \end{array} \right) = \frac{\text{SD of men's draws}}{\sqrt{\# \text{ of men's draws}}} = \frac{22.8}{\sqrt{227}} = 1.51.$$

For the women:

$$\left(\begin{array}{c} \text{SE of average} \\ \text{of women's draws} \end{array} \right) = \frac{\text{SD of women's draws}}{\sqrt{\# \text{ of women's draws}}} = \frac{18.2}{\sqrt{485}} = 0.83.$$

	Men	Women
# of draws	227	485
Average of draws	97	87
SD of draws	22.8	18.2
SE of average of the draws	1.51	0.83
Average of box (population)	?	?

Then we combine the two standard errors to find the standard error of the difference in averages:

$$\begin{aligned} \left(\begin{array}{c} \text{SE of} \\ \text{difference} \\ \text{in averages} \end{array} \right) &= \sqrt{\left(\begin{array}{c} \text{SE of average} \\ \text{of men's draws} \end{array} \right)^2 + \left(\begin{array}{c} \text{SE of average of} \\ \text{women's draws} \end{array} \right)^2} \\ &= \sqrt{(1.51)^2 + (0.83)^2} \\ &= \sqrt{2.2801 + 0.6889} \\ &= \sqrt{2.969} = 1.72. \end{aligned}$$

Back to the

Null hypothesis: Average of the men's population = Average of the women's population.

We have that

Average of draws for the men = 97 MPH,
Average of draws for the women = 87 MPH,
SE of the difference in averages = 1.72 MPH.

The difference in averages is 10. Is that a statistically significant difference? We find the Z:

$$\begin{aligned} Z &= \frac{(\text{Average of men's draws}) - (\text{Average of women's draws})}{\text{SE of the difference}} \\ &= \frac{97 - 87}{1.72} = \frac{10}{1.72} = 5.8. \end{aligned}$$

Thus the Z is 5.8. !!!!

We do not even need to look at the normal curve to know that the p-value is zero:

Reject the null hypothesis that the men's and women's averages are equal.

We are then quite confident that the population of STAT100 men do, on average, have higher fastest speeds than women.



? Here are some variables:

- Gender
- Height
- Weight
- Shoe Size
- Shoe Number: How many pairs of shoes do you own?
- Year: What year are you in school?
- Pets: How many dogs plus cats have you and your family owned in your lifetime?
- Sibs: How many brothers and sisters do you have?
- Speed: What is the fastest you have ever driven a car, in miles per hour?
- Cash: Approximately how much cash do you have with you now, in dollars?
- Sleep: On average, how many hours of sleep do you get each night?
- Mother's Age
- Father's Age
- Random Number: Pick a random number between 1 and 10

For each, do you think men and women differ significantly? Do you think underclassmen and upperclassmen differ?

?

Do men and women differ on the number of pets they have owned? The table has the data. Imagine these samples as simple random samples of the men and the women in all STAT100 classes.

	Male	Female
#	227	485
Average	1.93	2.82
SD	2.45	3.88

Find the men's average minus the women's average in the sample. Who has more pets on average, men or women?

What is the null hypothesis. (Is it for the sample, or the population?)

Find the standard errors for the men and for the women.

Find the standard error for the difference in averages.

Find the Z, and the p-value.

What do you conclude?

Comparing two percentages

In the 2011 Gallup Poll, 53% of the respondents said they supported gay marriage. In 2012, support was 50%. The difference in percentages between 2012 and 2011 is $50\% - 53\% = -3\%$. Is that a statistically significant decrease?

The null hypothesis is that the percentages in the populations for the two years are the same.

Null hypothesis: Percentage in population that favors gay marriage in 2012 = Percentage in population that favors gay marriage in 2011.

To test the null hypothesis, we use the Z with the difference in percentages as the numerator:

$$Z = \frac{(\% \text{ in 2012 sample}) - (\% \text{ in 2011 sample})}{\text{SE of what?}}$$

Since we are looking at the difference in percentages, we want the standard error of that difference.

The process for finding the standard error of the difference in percentages for two independent samples is the same as for the difference in averages, except that we use the standard errors of percentages:

- Find the standard errors of the two percentages.
- Square the two standard errors.
- Add up the squares.
- Take the square root of that sum.

Or, as a formula:

$$\left(\begin{array}{c} \text{SE of difference} \\ \text{in percentages} \end{array} \right) = \sqrt{\left(\begin{array}{c} \text{SE of percentage} \\ \text{in 2012 sample} \end{array} \right)^2 + \left(\begin{array}{c} \text{SE of percentage} \\ \text{in 2011 sample} \end{array} \right)^2}$$

We can assume that the samples for the two years are independent. The population is so large, that there is no way that the same person would be taken in both samples. We previously looked at the 2011 poll. It had 1018 people in the sample (draws). In the 2012 sample, there are 1024 draws. To find the standard error for the 2012 percentage, we first need the SD for the population in 2012. In the sample = draws, 50% are 1's, so the fraction of 1's is 0.50, and so is the fraction of 0's. Then the SD is estimated by

$$\begin{aligned}\text{SD in the 2012 box} &= \sqrt{(\text{Fraction of 1's in box}) \times (\text{Fraction of 0's in box})} \\ &\approx \sqrt{(\text{Fraction of 1's in draws}) \times (\text{Fraction of 0's in draws})} \\ &= \sqrt{0.50 \times 0.50} = 0.50.\end{aligned}$$

The population is huge, so we do not need the correction factor. The standard error is then

$$\begin{aligned}\text{SE of percentage in 2012 draws} &= \frac{\text{SD in 2012 box}}{\sqrt{\# \text{ of draws in 2012}}} \times 100 \\ &\approx \frac{0.5}{\sqrt{1024}} \times 100 = 1.56\%.\end{aligned}$$

? In 2011, the sample had 1018 people, and the percentage who favored gay marriage was 53%. Find the SE of the percentage for 2011.

The calculations so far:

	2012	2011
# of draws	1024	1018
Percentage in the draws	50%	53%
SE of percentage in the draws	1.56%	1.57%
Percentage in box (population)	?	?

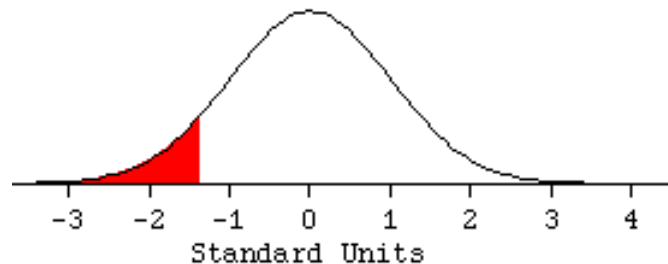
So the SE of the difference is

$$\begin{aligned} \left(\begin{array}{c} \text{SE of difference} \\ \text{in percentages} \end{array} \right) &= \sqrt{\left(\begin{array}{c} \text{SE of percentage} \\ \text{in 2012 sample} \end{array} \right)^2 + \left(\begin{array}{c} \text{SE of percentage} \\ \text{in 2011 sample} \end{array} \right)^2} \\ &= \sqrt{(1.56)^2 + (1.57)^2} = 2.21\%. \end{aligned}$$

Now we can calculate the Z:

$$\begin{aligned} Z &= \frac{(\% \text{ in 2012 sample}) - (\% \text{ in 2011 sample})}{\text{SE of difference in \% 's}} \\ &= \frac{50 - 53}{2.21} = \frac{-3}{2.21} = -1.36. \end{aligned}$$

The p-value is the area under the normal curve to the left of -1.36, which is about 8.85%.



This p-value is larger than 5%, so there's nothing to get excited about: So we fail to reject the null hypothesis.

Conclusion: There is not enough evidence to support the conclusion that support for gay marriage has declined.

The Z for two independent samples: Summary**Averages**

We have two populations, say A and B, and draw two samples independently, one sample from each population. We consider

Null hypothesis: The averages in the two populations are equal.

The Z is then

$$Z = \frac{(\text{Average of draws from box A}) - (\text{Average of draws from box B})}{\text{SE of the difference in averages}},$$

where

$$\left(\begin{array}{c} \text{SE of difference} \\ \text{in averages} \end{array} \right) = \sqrt{\left(\begin{array}{c} \text{SE of average of} \\ \text{draws from box A} \end{array} \right)^2 + \left(\begin{array}{c} \text{SE of average of} \\ \text{draws from box B} \end{array} \right)^2}.$$

Percentages

We have two populations, say A and B, and draw two samples independently, one sample from each population. The populations consist of 0's and 1's. We consider:

Null hypothesis: The percentages of 1's in the two populations are equal.

The Z is then

$$Z = \frac{(\text{Percentage in draws from box A}) - (\text{Percentage in draws from box B})}{\text{SE of the difference in percentages}},$$

where

$$\left(\begin{array}{c} \text{SE of difference} \\ \text{in percentages} \end{array} \right) = \sqrt{\left(\begin{array}{c} \text{SE of percentage} \\ \text{in draws from box A} \end{array} \right)^2 + \left(\begin{array}{c} \text{SE of percentage} \\ \text{in draws from box B} \end{array} \right)^2}.$$

Whenever doing a hypothesis test using a Z or T, check to see whether it is greater than or less than 2, ignoring sign. If it is somewhat greater than 2, you will probably reject the null hypothesis. If it is somewhat less than 2, you will likely fail to reject. If it is close to 2, then you may have to find the p-value to decide.

Hypothesis testing in randomized experiments

In the gold star experimental design for testing the polio vaccine, subjects (all volunteers) were randomly assigned to get either the vaccine or the placebo. We will use percentages, even though they will be very small numbers.

	# Studied	# cases	Rate per 100,000	Percentage
Vaccine	200,745	57	28	0.028%
Control	201,229	142	71	0.071%

Clearly, among the children in the samples, the vaccinated children had a lower rate of polio than those who received the placebo. Could that difference be just due to chance? That is, is the result statistically significant?

Even though we do not have two simple random samples from two populations, because of the randomization of the subjects to the two treatments, the formula for the Z that we used starting on page 324 is also valid here. (See Chapter 27, Section 3, in the text for an explanation.)

The boxes would be the population of all children, one box for those who would get the vaccine, one for those who would get the placebo. Then the SD's in the boxes can be approximated using the fractions of 1's and 0's in the samples, as below:

$$\text{SD in vaccinated box} = \sqrt{\frac{57}{200745} \times \frac{(200745 - 57)}{200745}} = 0.017,$$

$$\text{SD in placebo box} = \sqrt{\frac{71}{201229} \times \frac{(201229 - 71)}{201229}} = 0.019.$$

So far, we have

	# Studied	# cases	Percentage	SD
Vaccine	200,745	57	0.028%	0.017
Control	201,229	142	0.071%	0.019

Now we find the standard errors of percentage for the two groups:

$$\begin{aligned} \left(\begin{array}{c} \text{SE of percentage} \\ \text{in vaccinated sample} \end{array} \right) &= \frac{\text{SD in vaccinated box}}{\sqrt{\# \text{ of draws in vaccinated sample}}} \times 100 \\ &= \frac{0.017}{\sqrt{200745}} \times 100 = 0.0040\%, \end{aligned}$$

and

$$\begin{aligned} \left(\begin{array}{c} \text{SE of percentage} \\ \text{in control sample} \end{array} \right) &= \frac{\text{SD in placebo box}}{\sqrt{\# \text{ of draws in placebo sample}}} \times 100 \\ &= \frac{0.019}{\sqrt{201229}} \times 100 = 0.0042\%. \end{aligned}$$

Then the Z is as in the polling example,

$$Z = \frac{(\% \text{ in vaccinated sample}) - (\% \text{ in placebo sample})}{\text{SE of difference in \% 's}}.$$

The standard error of the difference uses the formula below:

$$\left(\begin{array}{c} \text{SE of difference} \\ \text{in percentages} \end{array} \right) = \sqrt{\left(\begin{array}{c} \text{SE of percentage} \\ \text{in vaccinated sample} \end{array} \right)^2 + \left(\begin{array}{c} \text{SE of percentage} \\ \text{in placebo sample} \end{array} \right)^2}$$

? Find the standard error of the difference in percentages.

Now we have

$$\begin{aligned}\% \text{ polio in vaccine sample} &= 0.028\%, \\ \% \text{ polio in placebo sample} &= 0.071\%, \\ \text{SE of difference in \%} &= 0.0058\%.\end{aligned}$$

Then

$$\begin{aligned}Z &= \frac{(\% \text{ in vaccinated sample}) - (\% \text{ in placebo sample})}{\text{SE of difference in \%s}} \\ &= \frac{0.028 - 0.071}{0.0058} = \frac{-0.043}{0.0058} = -7.4.\end{aligned}$$



That is a very large (in magnitude) Z , so the p -value is zero. There is no way that Z could happen by chance. We strongly **reject the null hypothesis** that the vaccine and placebo have equal effects.

We conclude that the vaccine really did work. (Which we already knew.)

11.8 Testing for independence

From back in Chapter 13, we saw that two possibilities were *independent* if the chance the first happens is the same whether or not the second happens.



Drawing from the box of 712 people: Is being 72 inches or taller independent of being male?

The two possibilities:

- Being 72 inches or taller;
- Being male.

We calculated the two conditional chances:

- Chance of being 72 inches or taller given male = 42%;
- Chance of being 72 inches or taller given female = 1.4%.

They are not the same. So being tall is **not** independent of being male in this sample: Height and gender are *dependent*.

How about being a freshman and being male? We need a table:

	Freshman	Not freshman
Male	75	152
Female	157	328

$$\begin{aligned}
 \text{Chance of being a freshman given male} &= \frac{75}{152 + 75} \\
 &= \frac{75}{227} = 0.33 \rightarrow 33\%;
 \end{aligned}$$

$$\begin{aligned}
 \text{Chance of being a freshman given female} &= \frac{157}{328 + 157} \\
 &= \frac{157}{485} = 0.32 \rightarrow 32\%.
 \end{aligned}$$

Those two chances are *almost* equal. So, technically, we'd have to say "freshman" and "male" are not independent in this sample, but they are very close to independent.

But now consider these 712 as a simple random sample from a larger population, the population of all STAT100 students. What can we say about the entire population? Are height and gender independent? I'd say no. Are year in school and gender independent? I'd guess maybe. We need a significance test, though, to be sure.

Box model

In order to perform a significance test, we need a box model. Since we are looking at two aspects of the people, year in school and gender, each ticket has two characteristics written on it. A small example, with just six tickets, is

Male Freshman	Male Non-Freshman	Male Non-Freshman
Female Freshman	Female Non-Freshman	Female Non-Freshman

Rather than having just a number on each ticket, we have both the gender and the year in school for that person.

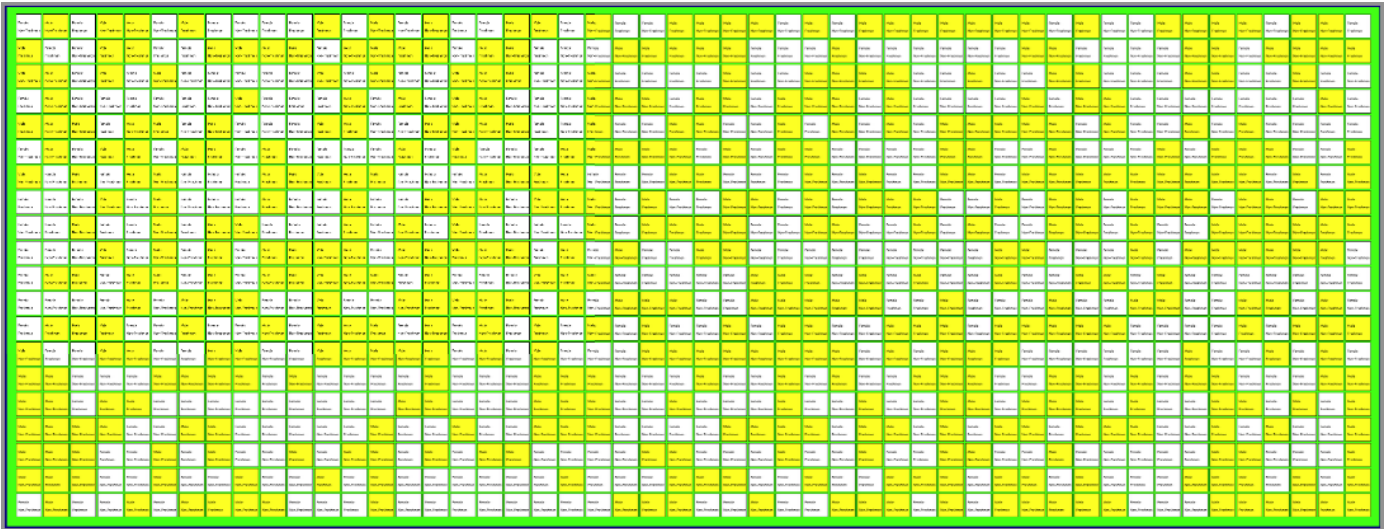
Are gender and year in school independent?

$$\text{Chance of being a freshman given male} = \frac{1}{3} = 0.33 \rightarrow 33\%;$$

$$\text{Chance of being a freshman given female} = \frac{1}{3} = 0.33 \rightarrow 33\%.$$

So for this box, year in school and gender **are** independent.

Here is a larger box, zoomed out so that you cannot really see what's on the tickets:



Now assume that the data consist of a sample of 712 draws from a very large population (box). From these draws, we find the following table:

	Freshman	Not freshman
Male	75	152
Female	157	328

In the sample (the draws), being male and being freshman are **not** independent, but what we want to know is

Are being a freshman and gender independent in the population?

To use a significance test, we formulate the null hypothesis as they are indeed independent:

Null hypothesis: Being a freshman and being male **are** independent in the population.

To see whether that null hypothesis is plausible in light of the sample, we figure out what we would expect the table to look like if the null hypothesis were true.

Expected values

For the significance tests using Z or Student's t, we started with the average (or sum, or percentage) of the draws, and then found what the expected value of the average (or sum, or percentage) would be *if the null hypothesis were true*.

The idea for testing independence is similar, but now we have four numbers to worry about, that is, the four numbers in the table: # of freshman males, # of freshman females, # non-freshman males, # of non-freshman females. We have

Null hypothesis: Being a freshman and being male are independent in the population.

We need to figure out what we would expect the table to look like if the null hypothesis were true.

Focus on the freshman males. If the null hypothesis were true, then we know that when drawing a ticket from the box, because of independence,

$$\left(\begin{array}{c} \text{Chance of freshman} \\ \text{and male} \end{array} \right) = (\text{Chance of freshman}) \times (\text{Chance of male}).$$

We also know that

$$\text{Chance of male} = \frac{\# \text{ of males in box}}{\text{Total } \# \text{ in box}}, \text{ and}$$

$$\text{Chance of freshman} = \frac{\# \text{ of freshman in box}}{\text{Total } \# \text{ in box}}.$$

The box is the population, so we do not know exactly what is in the box. But we can estimate the chances by using the draws (the sample):

$$\text{Chance of male} \approx \frac{\# \text{ of males in the draws}}{\text{Total } \# \text{ of draws}}, \text{ and}$$

$$\text{Chance of freshman} \approx \frac{\# \text{ of freshman in the draws}}{\text{Total } \# \text{ of draws}}.$$

The data consist of a sample of 712 draws from the box. To find

$$\text{Chance of male} \approx \frac{\# \text{ of males in the draws}}{\text{Total } \# \text{ of draws}},$$

we first need to find the total number of males, which is the sum of the freshman males and the non-freshman males, as in the table:

	Freshman	Not freshman	Total of row
Male	75	152	75+152=227
Female	157	328	157+328=485
Total of column	75+157=232	152+328 = 480	712

So

$$\begin{aligned} \text{Chance of male} &\approx \frac{\# \text{ of males in the draws}}{\text{Total } \# \text{ of draws}} \\ &= \frac{75 + 152}{712} \\ &= \frac{227}{712} (= 0.319). \end{aligned}$$

Similarly, the number of freshman is the sum of the number of freshman males and freshman females:

$$\begin{aligned} \text{Chance of freshman} &\approx \frac{\# \text{ of freshman in the draws}}{\text{Total } \# \text{ of draws}} \\ &= \frac{75 + 157}{712} \\ &= \frac{232}{712} (= 0.326). \end{aligned}$$

Then, under the null hypothesis that male and freshman are independent:

$$\begin{aligned} \left(\begin{array}{c} \text{Chance of freshman} \\ \text{and male} \end{array} \right) &= (\text{Chance of freshman}) \times (\text{Chance of male}) \\ &\approx \frac{227}{712} \times \frac{232}{712} (= 0.104). \end{aligned}$$

What we now need is the expected value of the number of freshman males in the sample. Counting the number of freshman males in the sample is like finding the sum of the draws from the box where freshman males have 1's, and everyone else has 0's. So, for example, in the little box, with only six tickets, we have

Male 1 Freshman	Male 0 Non-Freshman	Male 0 Non-Freshman
Female 0 Freshman	Female 0 Non-Freshman	Female 0 Non-Freshman

Our box has all STAT100 students, so has thousands of tickets. Then the average in the box is the same as the chance of drawing a 1, which we estimate to be 0.104.

The expected value of the sum of the 712 draws is then

$$\begin{aligned}
 \left(\begin{array}{c} \text{Expected number} \\ \text{of freshman males} \end{array} \right) &= (\# \text{ of draws}) \times (\text{Average in the box}) \\
 &= 712 \times (\text{Chance of freshman}) \\
 &\approx 712 \times \frac{227}{712} \times \frac{232}{712} (= 73.97).
 \end{aligned}$$

But notice that we can simplify this a bit, since there is one 712 in the numerator, and two in the denominators:

$$\left(\begin{array}{c} \text{Expected number} \\ \text{of freshman males} \end{array} \right) \approx \frac{227 \times 232}{712}.$$

Relate those numbers back to the table:

	Freshman	Not freshman	Total of row
Male	75	152	227
Female	157	328	485
Total of column	232	480	712

That is,

$$\left(\begin{array}{c} \text{Expected number} \\ \text{of freshman males} \end{array} \right) \approx \frac{\left(\begin{array}{c} \# \text{ of males} \\ \text{in the draws} \end{array} \right) \times \left(\begin{array}{c} \# \text{ of freshman} \\ \text{in the draws} \end{array} \right)}{\text{Total } \# \text{ of draws}}.$$

The expected table

We can find the expected number of people of each type using the formula

$$\text{Expected number} = \frac{(\text{Total of row}) \times (\text{Total of column})}{\text{Total \# of draws}}.$$

So for the four cells:

	Freshman	Not freshman	Total of row
Male	$\frac{227 \times 232}{712} = 73.97$	$\frac{227 \times 480}{712} = 153.03$	227
Female	$\frac{485 \times 232}{712} = 158.03$		485
Total of column	232	480	712

This table is called the **expected table**, being the expected values of the cells under the null hypothesis that male and freshman are independent.

? Fill in the blank cell. That is, find the expected number of people who are non-freshman females.

The observed table

The original table, of the draws, is called the **observed table**:

	Freshman	Not freshman	Total of row
Male	75	152	227
Female	157	328	485
Total of column	232	480	712

Recall that the question is whether the null hypothesis is plausible, that is, whether it is plausible that in the population, being male is independent of being a freshman. How do we test this hypothesis? What would lead us to doubt the null hypothesis?

Look again at the expected table:

	Freshman	Not freshman	Total of row
Male	73.97	153.03	227
Female	158.03	326.97	485
Total of column	232	480	712

If the null hypothesis is true, what we see in the data (the observed table) should be fairly close to what we expect from the null box (the expected table). Compare the entries in the tables:

(75 vs. 73.97), (152 vs. 153.03), (157 vs. 158.03), (328 vs. 326.97).

They are fairly close, off by a little over 1. If you were contemplating that the null hypothesis were true, then looked at the two tables, you would not be shocked.

The conclusion: **Fail to reject** the null hypothesis. It is plausible that being male and being freshman are independent in the population.



The chi-square statistic

The last page ended with us failing to reject the null hypothesis, but that significance test was not a formal one: We did not find a p-value. For future such tables, how can we tell whether the observed and expected tables are close enough or not?

Observed Table

	Freshman	Not freshman	Total of row
Male	75	152	227
Female	157	328	485
Total of column	232	480	712

Expected Table

	Freshman	Not freshman	Total of row
Male	73.97	153.03	227
Female	158.03	326.97	485
Total of column	232	480	712

We want to see how close these two tables are to each other. We cannot use just a Z (or a t), because there are four numbers to compare. Instead, we will calculate what is called a **chi-square statistic**. For each of the four numbers, we calculate the following:

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

For the freshman/male values, we have

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(75 - 73.97)^2}{73.97} = 0.0143.$$

We do similar calculations for the other three cells. (We do not bother with the totals rows and columns.)

Here are the tables we wish to compare again:

Observed Table

	Freshman	Not freshman	Total of row
Male	75	152	227
Female	157	328	485
Total of column	232	480	712

Expected Table

	Freshman	Not freshman	Total of row
Male	73.97	153.03	227
Female	158.03	326.97	485
Total of column	232	480	712

Then for each cell (not counting the totals), we calculate

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

Here are the results:

	Freshman	Not freshman
Male	$\frac{(75-73.97)^2}{73.97} = 0.0143$	$\frac{(152-153.03)^2}{153.03} = 0.0069$
Female	$\frac{(157-158.03)^2}{158.03} = 0.0067$	$\frac{(328-326.97)^2}{326.97} = 0.0032$

Now we sum those values to obtain the chi-square statistic:

$$\text{Chi-square} = 0.0143 + 0.0069 + 0.0067 + 0.0032 = 0.0311.$$

That is a small number, which means the two tables are fairly close to each other. But how close is close? That is, what is the p-value?

The chi-square table

To summarize so far — We have the null hypothesis,

Null hypothesis: Being male and being freshman are independent in the population.

We then figured out what the table would be if the null hypothesis were true, and compared it to the table based on the sample. The way we compared the two tables was by calculating the chi-square statistic. For our data, the statistic was

$$\text{Chi-square} = 0.0311.$$

Now we have to look at the **chi-square table** in the book. It also needs the degrees of freedom, which in this case is 1. (We will explain that number later.) The row from the chi-square table is

Degrees of freedom	99%	95%	90%	70%	50%	30%	10%	5%	1%
1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64

The top row has the p-values, and the second row has the chi-squares.

Our chi-square is 0.0311, which is between the 0.016, which has a p-value of 90%, and 0.15, which has a p-value of 70%. That means we have a p-value between 70% and 90%.

In any case, this p-value is very large, nowhere near as small as 5%. Thus,

We **fail to reject** the null hypothesis that being a freshman is independent of gender in the population.

Which is what we decided informally, anyway.

A larger table

A sample of 788 STAT100 students were asked their ethnicity and to rate their political attitudes, where 0 means very liberal, 10 means very conservative, and 5 is middle of the road. The table has the counts, where we grouped the responses:

	Liberal	Middle of the road	Conservative	Total
White	203	109	139	451
Black	40	19	17	76
Hispanic/Latino	28	19	2	49
Asian	66	62	44	172
Mixed/Other	16	14	10	40
Total	353	223	212	788

The question is whether the different ethnic groups have different attitudes in the population of all STAT100 students. The null hypothesis is that they have the same attitudes.

Null hypothesis: Ethnicity and political attitude are independent.

Even though this table has more rows and columns, the chi-square test works the same way as before. The steps:

1. Find the expected table.
2. Calculate the chi-square statistic.
3. Look up the chi-square statistic in the chi-square table to find the p-value.
4. Make a decision:
 - If the p-value $< 5\%$, reject the null hypothesis
 - If the p-value $> 5\%$, fail to reject the null hypothesis

Step 1 is to find the expected number under the null hypothesis of independence. We have to find

$$\text{Expected number} = \frac{(\text{Total of row}) \times (\text{Total of column})}{\text{Total total}}$$

for all $5 \times 3 = 15$ cells.

Expected Table

	Liberal	Middle of the road	Conservative	Total
White	202.03	127.63	121.34	451
Black	34.05	21.51	20.45	76
Hispanic/Latino	21.95	13.87	13.18	49
Asian	77.05	48.68	46.27	172
Mixed/Other	17.92	11.32		40
Total	353	223	212	788

The expected number of white liberals, for example, is then

$$\begin{aligned} \text{Expected number} &= \frac{(\text{Total of row}) \times (\text{Total of column})}{\text{Total total}} \\ &= \frac{451 \times 353}{788} = 202.03. \end{aligned}$$

That is very close to the observed count of 203.

? Fill in the blank in the table.

Step 2 is to find the chi-square statistic.

Observed Table

	Liberal	Middle of the road	Conservative	Total
White	203	109	139	451
Black	40	19	17	76
Hispanic/Latino	28	19	2	49
Asian	66	62	44	172
Mixed/Other	16	14	10	40
Total	353	223	212	788

Expected Table

	Liberal	Middle of the road	Conservative	Total
White	202.03	127.63	121.34	451
Black	34.05	21.51	20.45	76
Hispanic/Latino	21.95	13.87	13.18	49
Asian	77.05	48.68	46.27	172
Mixed/Other	17.92	11.32	10.76	40
Total	353	223	212	788

For each cell, we find

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}},$$

then add them up. So for Asian middle-of-the-roaders,

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(62 - 48.68)^2}{48.68} = 3.645.$$

Not particularly small. There are more than would be expected under independence.

Continuing step 2, the chi-square values for (almost) all of the cells are

	Liberal	Middle of the road	Conservative
White	0.005	2.719	2.570
Black	1.040	0.293	0.582
Hispanic/Latino	1.668	1.897	9.483
Asian		3.645	0.111
Mixed/Other	0.206	0.634	0.054

Add them up:

$$\text{Chi-square} = 0.005 + 2.719 + \cdots + 0.054 = 26.49.$$

? Find the value for the blank cell in the table.

The degrees of freedom

Step 3 is to look up the chi-square statistic in the chi-square table to find the p-value. We calculated the chi-square to be 26.49. To use the chi-square table we need to figure out the degrees of freedom, which are different than before.

In a chi-square test of independence, the degrees of freedom are given by the formula

$$\text{Degrees of freedom} = (\# \text{ of rows} - 1) \times (\# \text{ of columns} - 1).$$

We do **not** count the total row and total column. In the example, we have five rows (the ethnicities) and three classifications of political attitudes (conservative, middle of the road, and liberal). So

$$\begin{aligned} \text{Degrees of freedom} &= (\# \text{ of rows} - 1) \times (\# \text{ of columns} - 1) \\ &= (5 - 1) \times (3 - 1) = 4 \times 2 = 8. \end{aligned}$$

Now we have the chi-square value of 26.49, and eight degrees of freedom. Here is that row from the chi-square table:

Degrees of freedom	99%	95%	90%	70%	50%	30%	10%	5%	1%
8	1.65	2.73	3.49	5.53	7.34	9.52	13.36	15.51	20.09

Our chi-square value is larger than the largest value in the second row of the table. So the p-value is beyond the corresponding p-value of 1%, that is

$$\text{p-value} < 1\%.$$

So the chi-square statistic = 26.49, the degrees of freedom are 8, and the p-value is less than 1%. Step 4 is to decide whether the p-value is significant.



It is very significant, much less than 5%, so we can **reject** the null hypothesis.

Conclusion: We are fairly confident that ethnicity and political attitudes are **not** independent.

Comparing attitudes

Now that we have decided that ethnicity and political attitudes are not independent, we would like to see where the differences lie. The next table gives the percentages for the political attitudes for each of the ethnicities:

	Liberal	Middle of the road	Conservative	Total
White	45%	24%	31%	100%
Black	53%	25%	22%	100%
Hispanic/Latino	57%	39%	4%	100%
Asian	38%	36%	26%	100%
Mixed/Other	40%	35%	25%	100%

Whites appear relatively most conservative, with 31%.

Blacks and Hispanics are more liberal, and Hispanics very un-conservative (only 4% of hispanics say they are conservative).

Asians and the mixed/other are relatively more middle of the road.

? People were asked what sports they most like to participate in, and what sports they most liked to watch. The table below summarizes the data. For favorite sports to watch, it is divided into team sports & others. For sports to participate in, there are team sports, One/Two (which means one-on-one or two-on-two, like tennis), and individual sports, like aerobics and cycling. The question is whether there is independence between what people like to watch and what people like to do.

? (continued)

Fill in the various totals in the table:

Observed Table

	Participate			
Watch ↓	Team	One/Two	Individual	Total of row
Team	21	10	25	
Other	6	2	7	
Total of column				

Fill in the expected values (and check the totals are the same as above):

Expected Table

	Participate			
Watch ↓	Team	One/Two	Individual	Total of row
Team				
Other				
Total of column				

Fill in the values of $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$:

	Participate			
Watch ↓	Team	One/Two	Individual	Total of row
Team				
Other				
Total of column				

Find the sum of those values:

? (continued)

Here is a portion of the chi-square table:

Degrees of freedom ↓	50%	30%	10%	5%	1%	← p-value
1	0.45	1.07	2.71	3.84	6.63	← Chi-squares
2	1.39	2.41	4.61	5.99	9.21	
3	2.37	3.66	6.25	7.81	11.34	
4	3.36	4.88	7.78	9.49	13.28	
5	4.35	6.06	9.24	11.07	15.09	
6	5.35	7.23	10.64	12.59	16.81	

Find the correct degrees of freedom for the observed table (on the previous page).

The chi-square statistic you calculated on the previous page is between which two values in the row corresponding to the correct degrees of freedom? (If it is larger than all, or smaller than all, say that.)

The p-value is then between which two percentages? (If the chi-square is less than all the values in the row, the p-value is greater than 50%. If the chi-square is more than all the values, then the p-value is less than 1%.)

What do you conclude? (Do you reject the null hypothesis?) What can you say concerning independence of sports people like to play and those they like to watch?

A.1 Practice exam 1

Question 1

For one class, the teacher compared the average Final Exam scores of two groups of students: Those who followed the directions and filled out their Section Number on the Scantron form, and those students who left it blank. The average exam score for the 880 students who bubbled in their Section Number was significantly higher than the average exam score of the 220 students who did not.

- a) What kind of study is this? • Designed experiment • Observational study
- b) Do these data prove that bubbling in the Section Number cause students to do better on the Final Exam?
- c) Could there be a third factor that is a more likely explanation of why students who bubble in their Section Number are more likely to score higher on the Final Exam?
 - Yes, for example, serious students who follow directions are both more likely to do well on exams and more likely to bubble in their section numbers
 - Yes, there are more people who bubbled in their sections, hence they have more chance to do well
 - No, there is a placebo since the students did not know the teacher would collect the data

Question 2

Would students in Stat 100 learn better if they were allowed Formula Sheets during exams? To answer that question I did 2 studies.

Study A: I randomly assigned half the Stat 100 students to the Formula Sheet Group and half to the No Formula Sheet group.

Study B: I acted like a doctor and "prescribed" formula sheets to those students I thought really needed them and didn't prescribe them to those I thought would do well without them.

All students took the same exams and here are the results:

	Study A: Randomized		Study B: Non-randomized	
Group ↓	# of students	Average exam score	# of students	Average exam score
Formula sheet	500	80%	400	74%
No formula sheet	500	92%	600	94%

a) Both studies found that the No Formula group did much better than the Formula group, but the randomized design saw only a 12% difference whereas the Non-Randomized Design showed a 20% difference. What possible reason could account for that?

- In the non-randomized study, I chose the stronger students to be in the No Formula group and the weaker students to be in the Formula group, so the No Formula group did better both because they were stronger students to begin with and because not having a Formula sheet made them learn better

- In the non-random study, I showed how tailoring the study method to fit the student works better and therefore causes a more dramatic improvement

- In the non-randomized studies, more students were assigned to the Formula group than to the No-Formula group so that could account for the increased difference.

b) Which study is more likely to have third factors that explain the results?

- Study A ● Study B ● They're equally likely

c) Judging from both studies, would you conclude that there is good evidence for the following statements?

- Students learn better when they are allowed formula sheets based on their needs.

- Students seem to learn better when they are not allowed to rely on formula sheets.

Question 3

According to a recent Swedish study, playing golf regularly can extend your life. The death rates of 300,000 regular golf players were compared to 300,000 non-golfers. Those who played golf regularly lived, on average, 5 years longer than those who did not play golf.

a) What type of study is this?

- Non-randomized Controlled Experiment with Placebo
- Randomized Controlled Experiment
- Randomized Controlled Double Blind Experiment
- Observational Study with Controls

b) Could there be a third factor that explains why golfers lived longer?

- No, since the treatment and control groups are likely to be the same in all ways, except one group plays golf and the other doesn't
- No, since there were the same number of people in each group
- Yes, wealth could be a third factor
- Yes, golf courses are regularly sprayed with chemicals to maintain a lush, green lawn. Overexposure to these chemicals can be detrimental to health.

Question 4

Imagine Bart and Lisa Simpson going to the same college. Bart takes almost all easy courses and Lisa takes almost all very difficult courses. They both take a total of 100 hours. Here's a chart that shows how they did:

	Bart		Lisa	
	# Hours	GPA	# Hours	GPA
Easy courses	97	3.5	3	4.0
Difficult courses	3	1.0	97	3.4
Total	100	3.425	100	3.328

a) Who had the lower GPA for easy courses?

b) Who had the lower GPA for difficult courses?

c) Who had the lower total GPA?

- d) Why does the answer to part (c) seem to contradict the answers to parts (a) and (b)?
- There is no placebo
 - The calculations are wrong
 - It's Simpson's paradox
- e) Bart and Lisa both decide to take one more course. This time they'll both take the same exact course. Based on the results above, who would you expect to do better?

Question 5

130 people had their body temperature measured. The table below summarize the results by categorizing the data into four intervals:

Interval	Percentage
96 to 98	40%
98 to 98.5	23%
98.5 to 99	27%
99 to 101	10%

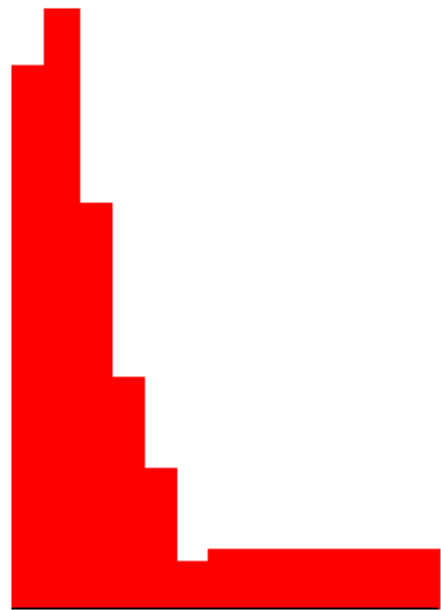
- a) You wish to draw a histogram of these data. What is the length of the base of the box that goes above the interval 96 to 98?
- b) What is the height of that box?

Question 6

In a class of 175 people, the ages ranged from 18 to 30. Here is a table of the data, used to create a histogram.

Interval	# of people
17.5 to 18.5	47
18.5 to 19.5	52
19.5 to 20.5	35
20.5 to 21.5	20
21.5 to 22.5	12
22.5 to 23.5	4
23.5 to 31.5	5

Which histogram is correct for these data?



18 19 20 21 22 23 24 25 26 27 28 29 30
Histogram A



18 19 20 21 22 23 24 25 26 27 28 29 30
Histogram B

Question 7

Here are the shoe sizes of five students:

5, 7, 8, 9, 7

a) What is the average?

b) What is the median?

c) The tallest man in the world was Robert Wadlow, from Alton, Illinois. His shoe size was 37. Suppose he joined the five students, so that there would be six people. Which would increase most in going from the five students to the six people, the average shoe size or the median shoe size?

Question 8

Here are the ages of five students:

19, 25, 20, 19, 22

a) What is the average of these five numbers?

b) Which set of the following numbers consists of the deviations, in some order?

● -1, -1, -2, -2, -4

● -1, 1, -2, -2, 4

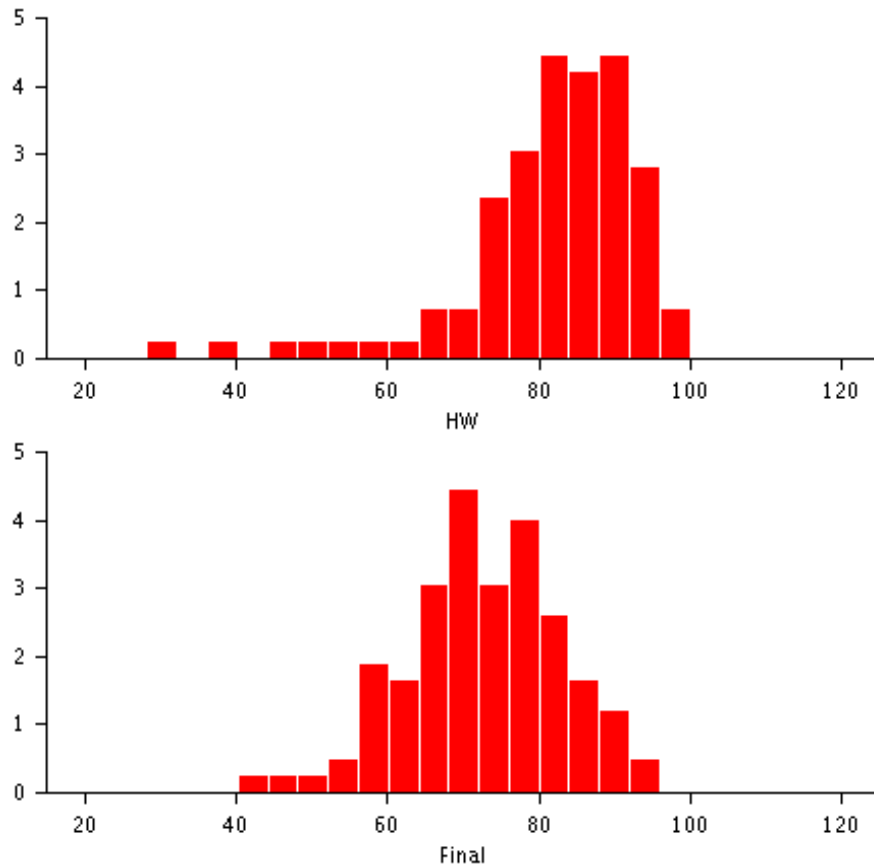
● 1, 1, 2, 2, 4

● -1, 1, 2, -2, -4

c) The sum of the squared deviations is 26. What is the SD?

Question 9

Below are histograms from a statistics class showing the students' scores on the homework (top histogram) and final exam (bottom histogram).

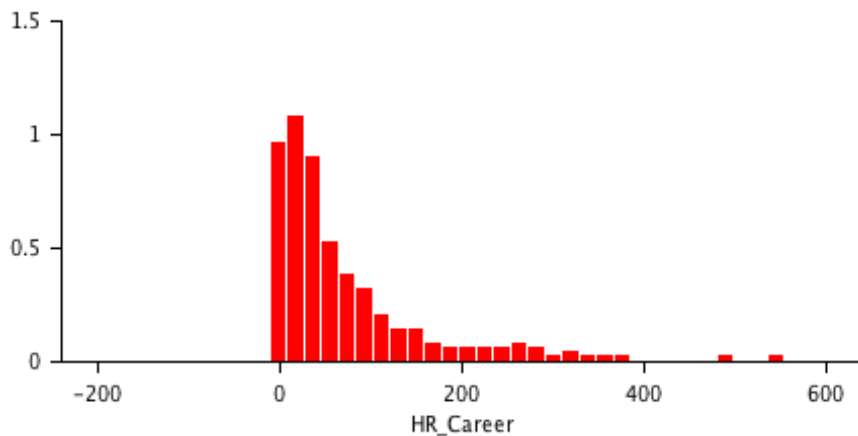


- a) The average score on the final exam is closest to
- 50 • 60 • 70 • 80 • 90
- b) The SD of the scores on the final exam is closest to
- 3 • 5 • 10 • 25
- c) Which is larger, the median for the homework scores or the median of the final exam scores?
- d) Which is larger, the median or the average of the homework scores?
- e) Which statement is best?
- The SD of the homework scores is about the same as the SD of the final scores

- The SD of the homework scores is about half the size of the SD of the final scores
- The SD of the homework scores is about one tenth the size of the SD of the final scores

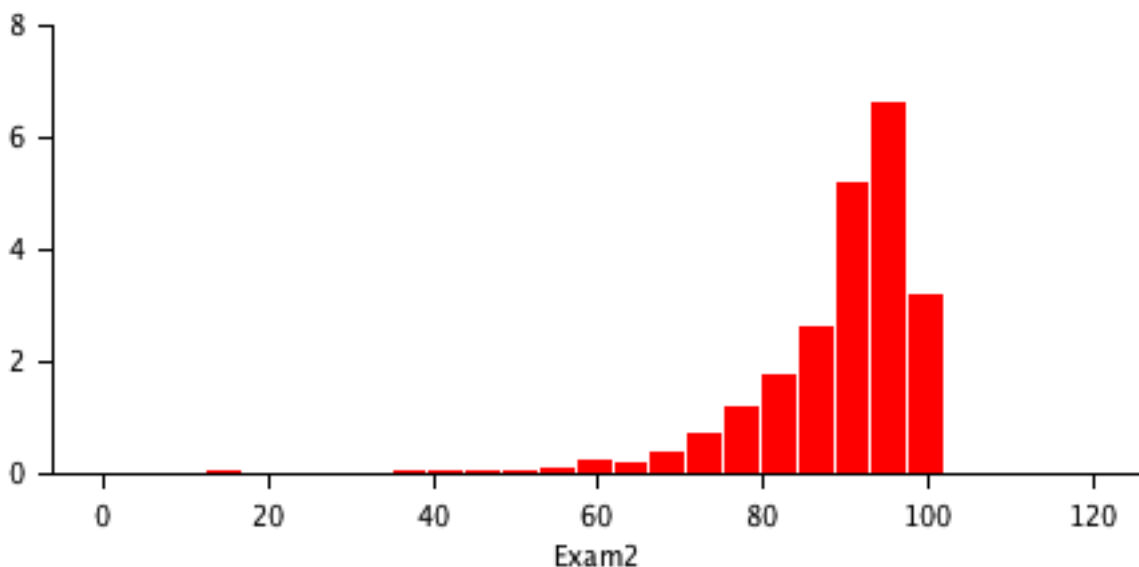
Question 10

- a) The histogram below shows the numbers of career home runs for a number of major league baseball players.



Which is larger, the median or the average?

- b) The next histogram is for the second exam in a statistics course:



Which is larger, the median or the average?

Question 11

The area under the normal curve between plus or minus 0.5 is 38.29%.

The area under the normal curve between plus or minus 1 is 68.27%.

The area under the normal curve between plus or minus 1.5 is 86.64%.

The area under the normal curve between plus or minus 2 is 95.44%.

Find the described areas in the normal curve:

a) The area between -2 and 1.5

b) The area between 0 and 1.5

c) The area above -2

d) The area above 0.5

Question 12

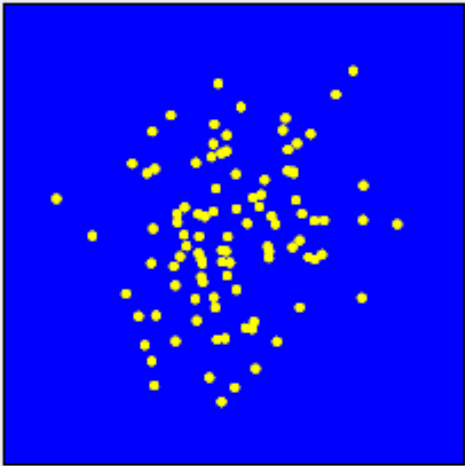
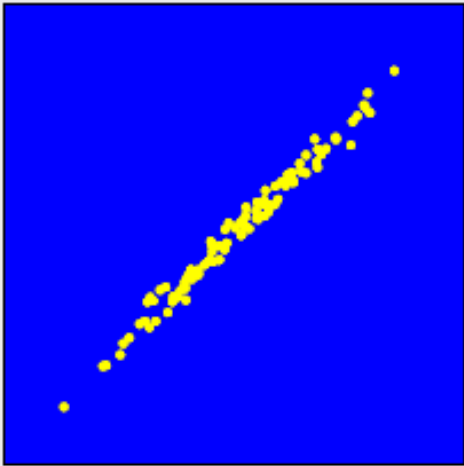
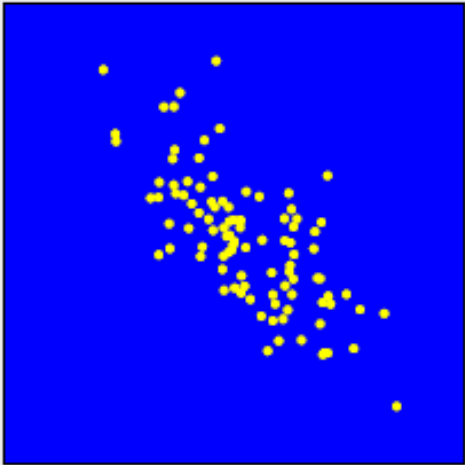
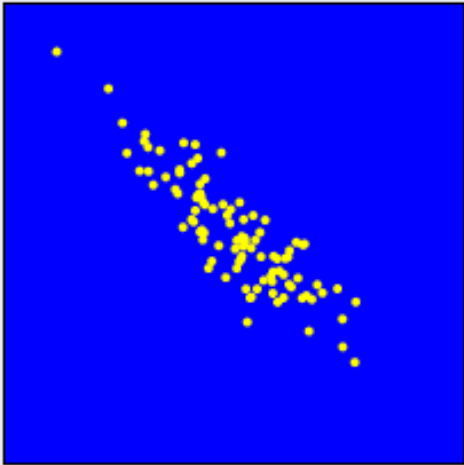
The histogram of the final scores in a course with a large number of students looked reasonably like a normal curve. The average score was 77 and the SD of the scores was 12. The grades were assigned on a straight scale, that is, people with scores 90 or above received an A, 80-90 a B, 70-80 a C, 60-70 a D, and below 60 an E. The scores were usually fractional, such as 89.76 or 55.23. One can then use the normal curve to approximate the percentages in each grade category.

a) Find the score 70 in standard units.

- b) According to the normal curve, is the percentage in the class who get a C or better more than 50% or less than 50%? (You can answer this without looking at the normal table. Just draw the normal curve and see what the area looks like.)
- c) Find the score 80 in standard units.
- d) According to the normal curve, is the percentage in the class who get a B or better more than 50% or less than 50%?

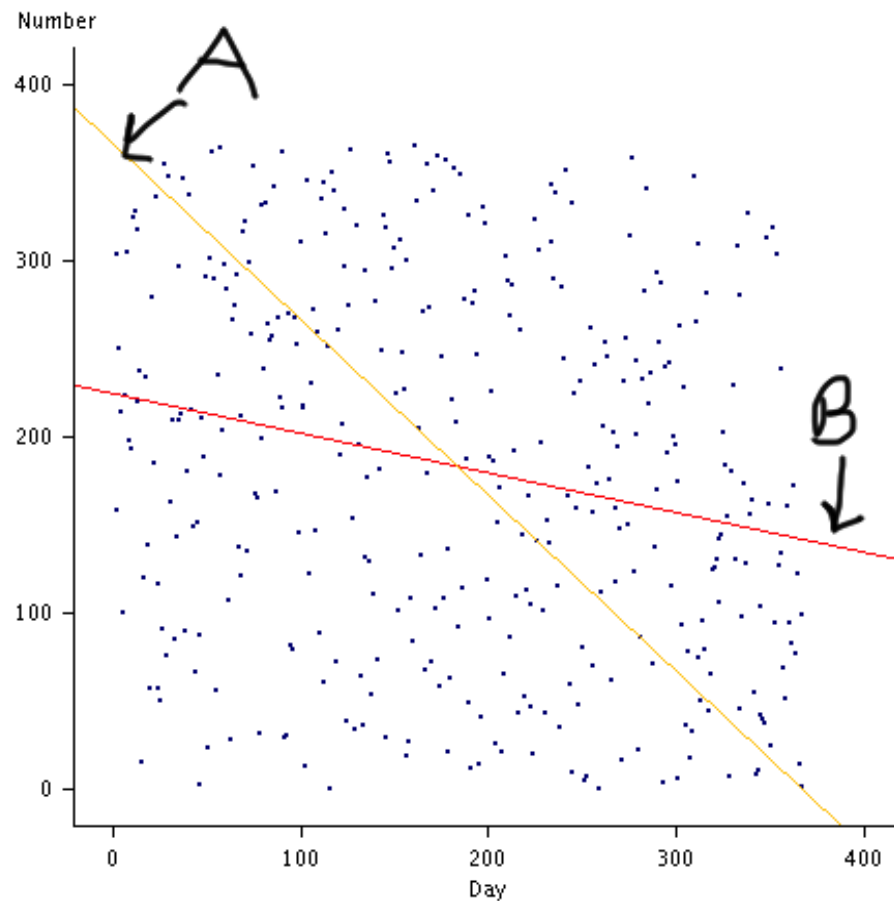
Question 13

Match the scatter plots with their correlation coefficients.

	<input type="radio"/> 0.99 <input type="radio"/> 0.04 <input type="radio"/> -0.71 <input type="radio"/> -0.91		<input type="radio"/> 0.99 <input type="radio"/> 0.04 <input type="radio"/> -0.71 <input type="radio"/> -0.91
	<input type="radio"/> 0.99 <input type="radio"/> 0.04 <input type="radio"/> -0.71 <input type="radio"/> -0.91		<input type="radio"/> 0.99 <input type="radio"/> 0.04 <input type="radio"/> -0.71 <input type="radio"/> -0.91

Question 14

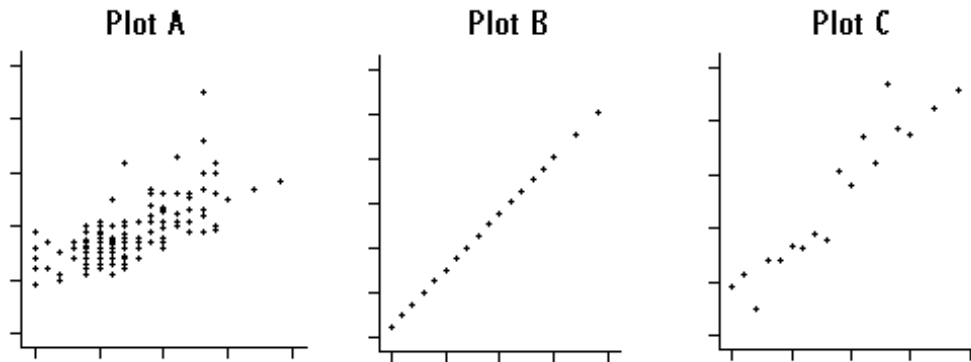
The next plot shows the results of the 1969 Draft Lottery, where the X is the day of the year (from 1 to 366), and the Y is the lottery number assigned to people with birthday corresponding to the X.



- a) One of the lines is the SD-Line and one is the regression line. Which line is the regression line?
- b) The average of the lottery numbers is 183.5, as is the average of the Day numbers. Which of the following statements is best supported by the scatter plot:
- On average, people with birthdays later in the year tended to have lower lottery numbers.
 - The regression effect explains why the regression line has negative slope.
 - The correlation coefficient is negative because of ecological correlations.

Question 15

Here are three scatter plots based on 126 students' data:



One plot has $X = \text{Height in inches}$ and $Y = \text{Weight in pounds}$ for the 126 people, one plot has $X = \text{Height in inches}$ and $Y = \text{Average Weight in pounds for everyone of height } X$, and one plot has $X = \text{Height in inches}$ and $Y = \text{Height in centimeters}$ for the 126 people.

- Which plot has the highest correlation coefficient? ● A ● B ● C
- Which plot has $X = \text{Height in inches}$ and $Y = \text{Average Weight in pounds for everyone of height } X$? ● A ● B ● C
- Which plot has $X = \text{Height in inches}$ and $Y = \text{Weight in pounds for the 126 people}$? ● A ● B ● C

For the plot with $X = \text{Height in inches}$ and $Y = \text{Weight in pounds}$ for the 126 people, the average height is 67 inches, the SD of the heights is 4 inches, the average weight is 145 pounds, the SD of the weights is 30 pounds, and the correlation coefficient is 0.7. The regression equation is

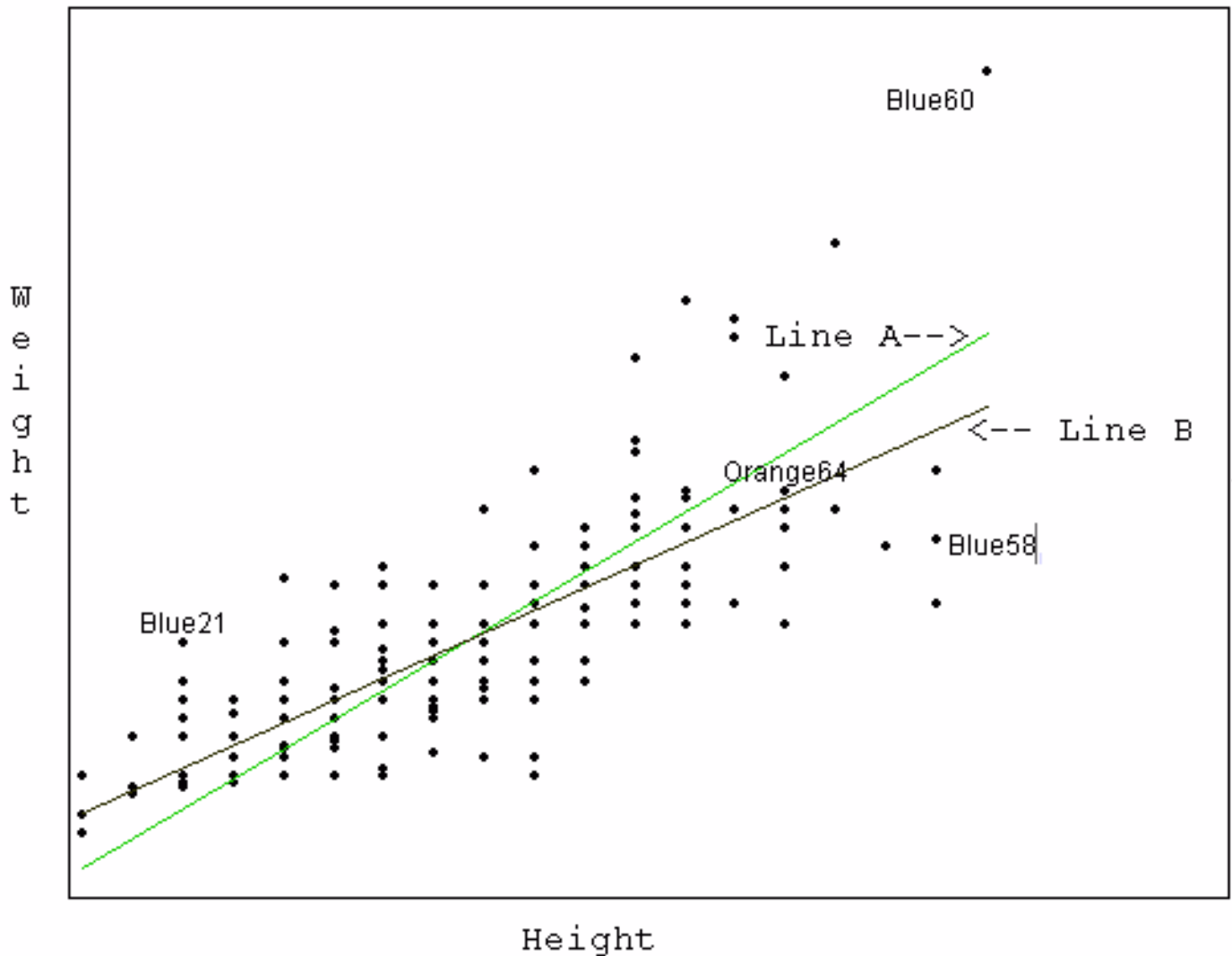
$$Y = -206.75 + 5.25 \times X.$$

- According to the regression line, the average weight of people who are 67 inches tall is
- According to the regression line, the average weight for people who are 71 inches tall is

- f) The correlation(s) of which of plots A and C are ecological correlations? (It could be one, both, or neither)

Question 16

The next scatter plot contains the heights in inches and weights in pounds of STAT 100 students.



- One of the lines is the SD-line, and one is the Regression Line. Which one is the Regression Line?
- Which of these two people weighs more? • Blue21 • Blue58
- Of the four people labeled on the plot, which exceeds the average weight estimated by the regression line for his/her height by the most?

- d) Of the four people labeled on the plot, which is under the average weight estimated by the regression line for his/her height by the most?
- e) Of the four people labeled on the plot, which is closest to the average weight estimated by the regression line for his/her height?
- f) The average of the errors is closest to • 0 • SD_{errors} • the slope • unable to determine from the information given

Question 17

Data was obtained over many days to see what the relationship is between how fast crickets chirp and the air temperature. The faster the chirping, the higher the temperature, in general. The SD of temperatures 6.7 degrees.

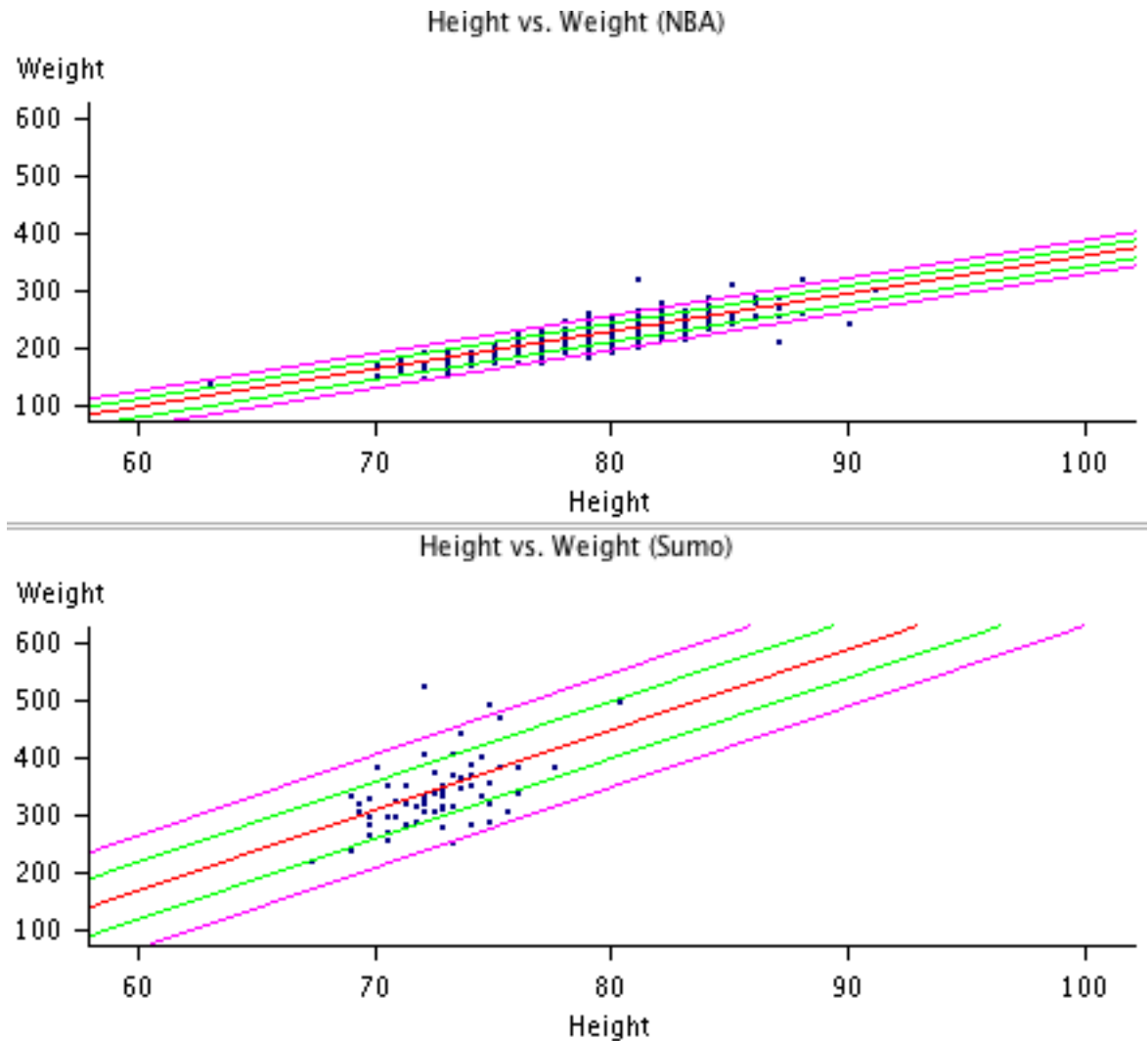
The regression line for the data with $X = \text{Cricket chirps per second}$ and $Y = \text{Air temperature (in Fahrenheit)}$ had an SD_{errors} of 3.7.

According to the regression line, when $X = 16$ chirps per second the average temperature was 77.7 degrees, and when $X = 17$ chirps per second, the average temperature was 81.1 degrees.

- a) What is the slope of the regression line?
- b) According to the regression line, what is the average temperature when $X = 18$ chirps per second?
- c) Consider just the days when $X = 17$. Using the normal curve, you want to estimate the percentage of days over 80 degrees. What would 79 be in standard units?
- d) Is the percentage above those standard units in the normal curve over 50% or under 50%? (You do not need a normal table for this question. It helps to sketch the picture.)

Question 18

Below are two plots showing $X = \text{height}$ and $Y = \text{weight}$ for samples of professional athletes. The top plot has basketball players, and the bottom plot has sumo wrestlers.



In each plot, the middle line is the regression line. The two outer lines are $\pm 2 \times SD_{\text{errors}}$ from the regression line, and the second and fourth lines are $\pm 1 \times SD_{\text{errors}}$ from the regression line.

- Which plot's regression line has the larger slope?
- The SD_{errors} for the bottom plot is closest to $\bullet 0 \bullet 10 \bullet 50 \bullet 100$
- The SD_{errors} for the top plot is less than or greater than the SD_{errors} for the top plot.

- d) About what percentage of points are between the two outer lines, in either plot?
- 5% • 50% • 68% • 95%

Question 19

In a sample of 100 girlfriend/boyfriend couples, looking at the numbers of music CD's these people owned, it was found that boyfriends with more than the average number of CD's tended to have girlfriends with fewer than average numbers of CD's. And the boyfriends with less than the average number of CD's tended to have girlfriends with more than average numbers of CD's. This is an example of what?

- Regression effect
- Ecological correlations
- Negative correlation
- Simpson's paradox

Question 20

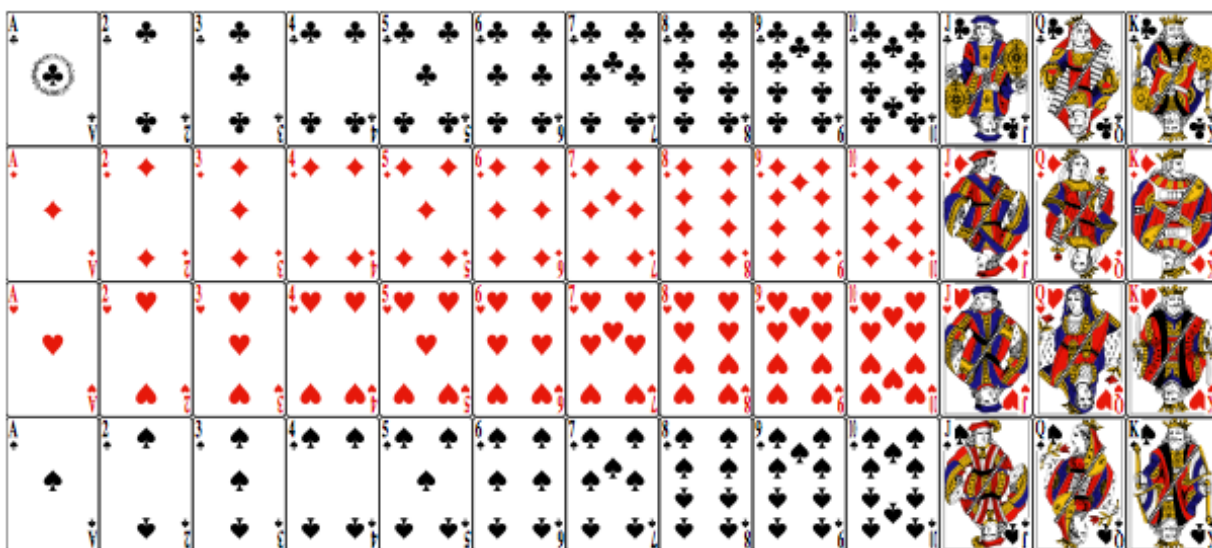
There are 20 M&M candies in a dish: 12 have peanuts, and 8 are plain (do not have peanuts). Of the peanut M&M's, 4 are green and 8 are red. Of the plain M&M's, 4 are green and 4 are red.

- a) If you randomly choose two M&M's (without replacement), the chance both are plain is closest to • 14.00% • 14.74% • 16.00% • 16.84%
- b) If you randomly choose one M&M, the chance you get a green one given you get a peanut one is closest to • 20% • 33.33% • 50% • 60% • 66.67%
- c) If you randomly choose one M&M, the chance you get a green one given you get a plain one is closest to • 20% • 33.33% • 50% • 60% • 66.67%
- d) If you randomly choose one M&M, is getting a green one independent of getting a peanut one?

- e) If you randomly choose one M&M, is getting a green one independent of getting a red one?

Question 21

Here is a picture of a regular 52-card deck of cards:



There are four suits (hearts, clubs, diamond, spades), each with 13 cards. There are four each of jacks, queens, and kings, one of each suit. The jacks, queens, and kings are all called face cards, so there are 12 face cards total.

- a) Suppose you intend to draw two cards without replacement, and the first card you draw is a face card. For the second draw, how many cards are left in the box?
- b) How many of those are faces cards?
- c) If you draw three cards **without** replacement, what is the chance that they are all face cards? (The answer is given as a product of fractions.)
- $4/52 * 3/52 * 2/52$
 - $13/52 * 12/51 * 11/50$
 - $12/52 * 11/51 * 10/50$
 - $12/52 * 11/52 * 10/52$
 - $12/52 * 12/51 * 12/50$

Question 22

Here is a table from a class of 250 students, where people are classified into gender (Male and Female) and whether they are left-handed or right-handed. So 14 of the men are left-handed, and 86 of the men are right-handed, out of a total of 100 men.

	Left-handed	Right-handed	Total
Male	14	86	100
Female	21	129	150
Total	35	215	250

- a) If you random pick someone from this class, what is the percentage chance that the person is male?
- b) What is the percentage chance that the person is left-handed given the person is a male?
- c) What is the percentage chance that the person is left-handed given the person is a female?
- d) Are getting a male and getting a left-hander independent?
- e) What is the percentage chance that the person is female given the person is left-handed?
- f) Is the chance of getting a female given you got a left-hander the same as the chance of getting a left-hander given you got a female?

Question 23

		0 00		
1 to 18	1st 12	1	2	3
		4	5	6
Even		7	8	9
		10	11	12
Red	2nd 12	13	14	15
		16	17	18
Black		19	20	21
		22	23	24
Odd	3rd 12	25	26	27
		28	29	30
19 to 36		31	32	33
		34	35	36

There are 38 numbers on a roulette wheel: The numbers from 1 to 36, as well as 0 and 00.

If you bet \$1 on the first twelve numbers, (1, 2, 3, ..., 12), you get \$2 if any one of those twelve numbers come up, and -\$1 if any other number comes up. Playing this bet **once** and seeing how much you end up with is like drawing from a box.

- How many tickets are in the box?
- How many draws do you make?
- What is written on the tickets?

Question 24

A large bowl has 366 capsules in it, one for each day of the year (including Feb. 29). You randomly draw four of the capsules out, without replacement. For each capsule that is the first of a month (i.e., Jan 1, Feb 1, ..., Dec 1), you get \$3. If it is any other day, you get nothing.

The amount of money you end up with is like the sum of the draws from a box.

- a) Do you draw the tickets with or without replacement?
- b) How many tickets are in the box?
- c) Which of the following is a possible result of the draws?
 - 1,0,0,1,1
 - 0,0,0,3
 - 0,3,6,9
 - 3,12,5,7
- d) Suppose three people play this game. (Each person draws four, and sums up the amount of money they get.) What is a possible set of winnings for the three?
 - 0,5,3
 - 1, 3, 2
 - 0,6,9
 - none of the above choices is possible

A.2 Practice exam 2

Question 1

A box contains two 3's and one 10, so it has $[3,3,10]$. Imagine drawing two tickets with replacement from this box, and finding the sum of the two draws.

- a) What is the smallest sum you can get?
- b) What is the largest sum you can get?
- c) There is another value you can get for the sum. What is it?

Question 2

A box has 20 tickets. There are 3 ones and 17 zeroes.

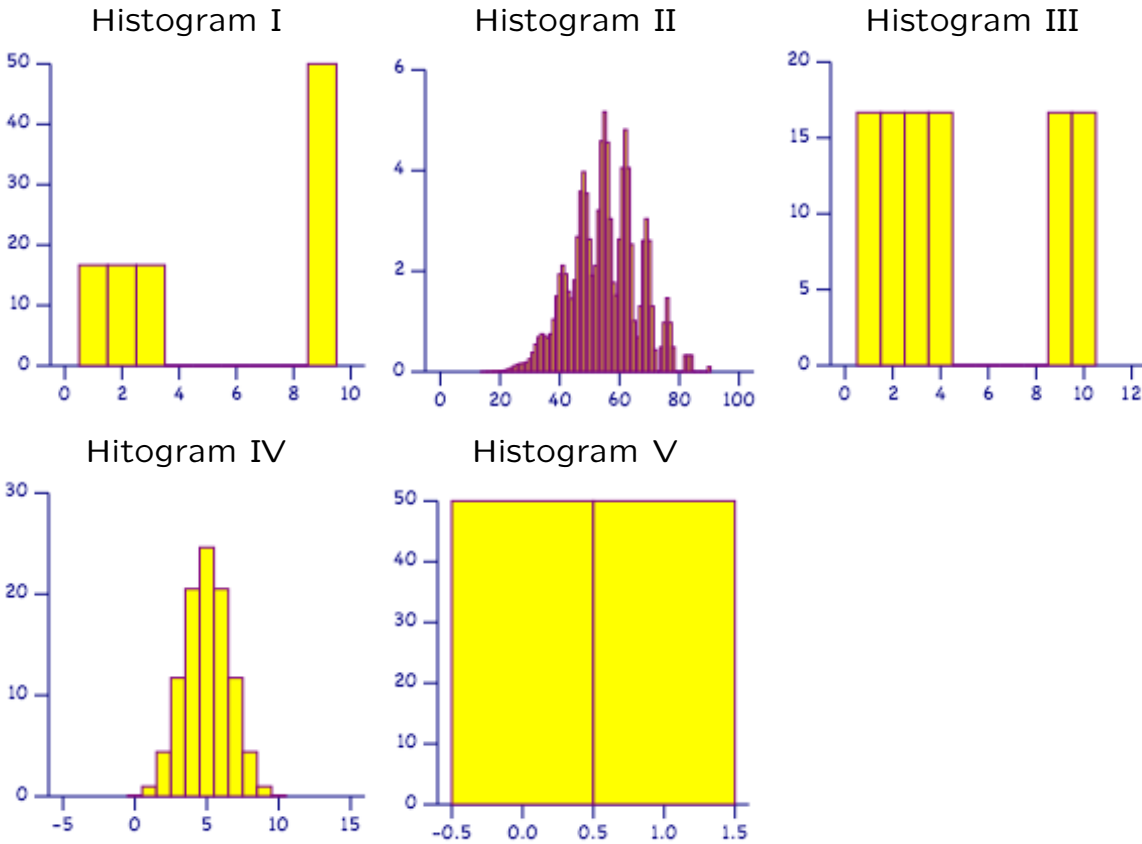
- a) What is the fraction of ones in the box?
- b) What is the SD in the box?

Question 3

This question is based on the following two boxes.

Box A	Box B
has one 1, one 2, one 3, and three 9's	has one 0 and one 1
<div><div>1</div><div>2</div><div>3</div><div>9</div><div>9</div><div>9</div></div>	<div><div>0</div><div>1</div></div>

Here are five histograms:



Match the histograms to the descriptions:

- a) The histogram for Box A
- b) The histogram for Box B
- c) The probability histogram of the sum of ten draws from Box A
- d) The probability histogram of the sum of ten draws from Box B
- e) None of the other choices

Question 4

In roulette, there are 38 possible numbers that could come up: 1 to 36, and 0 and 00. If you bet on the 5 numbers 0, 00, 1, 2, 3, you get \$6 if any of those come up, and get -\$1 if any other number comes up.

Playing this bet once and seeing how much you get is like drawing one ticket randomly from a box.

- a) How many tickets are in the box?
- b) What is written on the tickets?
- 5 tickets have 6 and the rest have -1.
 - The numbers 1 to 38, 0 and 00.
 - 5 tickets are ones and the rest zeroes
 - 6 tickets have 5 and the rest -1.
- c) The SD in the box is closest to ● 7 ● 0.34 ● 0.65 ● 1.69 ● 2.37 ● 30.39

Question 5

In roulette, consider betting \$1 on "Red." If red comes up, you get \$1, but if red does not come up, you get -\$1. The average in the corresponding box is $-.0526$, and the SD in the box is 1. Imagine playing this bet 400 times.

- a) The expected value of the sum of the draws is
- b) The standard error of the sum of the draws is
- c) Use the normal curve to estimate the chance that the sum of the draws is more than 0. The answer is closest to ● 15% ● 27% ● 73% ● 85% ● 100%
- d) Under which scenario would you be more likely to end up with the sum of the draws closer to the expected value? ● Drawing 1000 times ● Drawing 400 times

Question 6

A game at a county fair has a game where you can roll a twelve-sided die. The sides are various colors: 7 sides are blue, three are red, and two are white. Each side has the same chance of coming up.

Suppose you roll the die 25 times. Then the number of times blue comes up is like the sum of draws from a box.

- a) What is in this box?
- It has 25 tickets, some red, some white, and some blue, but we do not know exactly how many of each
 - It has 12 tickets, seven are 1, five are 0
 - It has 12 tickets, seven blue, three red, and two white
 - It has 25 tickets, with seven 1's and the rest 0's
- b) How many draws are there?
- c) The draws are ● with replacement ● without replacement.
- d) What is the expected value of the percentage of blues you get among the 25 rolls?
- e) The SD in the box is 0.49. What is the standard error of the percentage of blues you get among the rolls?
- f) If over 50% of you rolls are blue, then you get a prize. What is 50% in standard units?
- g) In the normal curve:
- The area between ± 1 is 68.27%.
 - The area between ± 2 is 95.44%.
 - The area between ± 3 is 99.73%.
 - The area between ± 4 is 99.9937%.

Use the normal curve to estimate the chance that you win a prize.

Question 7

The average weight of the undergraduate men at the U of I is 185 pounds, with a SD of 30 pounds. There are about 14,000 undergraduate men.

Suppose you take a simple random sample of 16 of them, and look at the average of that sample.

- a) What is the expected value of the sample average?
- b) In finding the standard error of the sample average, is it necessary to use the correction factor?
- c) What is the standard error of the average?

Question 8

- a) Someone has a Web site where one can vote in a poll. The poll for one week asked "Between the two cola giants in the U.S., Pepsi and Coke, which do you prefer?" The choices were "Pepsi", "Coke", "Like both about the same", "Don't like either". There were 386 people who voted. Of these, 34% liked Pepsi best, and 53% liked Coke best.

What kind of sample was this?

- Quota sample
 - Probability sample
 - Simple Random Sample
 - Multistage Cluster Sample
 - Self-selected sample
- b) True or false: A Quota sample is a type of probability sample.
 - c) A simple random sample is
 - like drawing tickets from a box with replacement
 - like drawing tickets from a box without replacement
 - the same as a quota sample
 - not a probability sample

- d) A researcher collected a sample of 100 people using quota sampling. Everyone in the sample responded. This sample is subject to
- Nonresponse bias
 - Selection bias
 - Both types of bias
 - Neither type of bias
- e) A researcher collected a sample of 100 people using simple random sampling. Eighty people in the sample responded. This sample is subject to
- Nonresponse bias
 - Selection bias
 - Both types of bias
 - Neither type of bias

Question 9

Suppose that 40% of the registered voters in Illinois are Democrats. Imagine taking a sample of 900 from the population of all registered voters in Illinois. (The SD in the box for this problem is 0.49.)

- a) How many tickets are in the box?
- 40% of 900
 - 900
 - 14,376
 - Many more than 14,376
- b) Suppose the sample of 900 people is a simple random sample. What is the standard error of the percentage of Democrats in the sample? (If it cannot be determined, then say that.)
- c) Now imagine taking a simple random sample of 900 from all the registered voters in the United States. Assuming the percentage of Democrats is 40%, the standard error of the percentage of Democrats in the sample from the entire U.S. is
- much smaller than
 - about the same as

- much larger than
the standard error of the percentage from the simple random sample from Illinois.
- d) Now suppose one takes a simple random sample of 900 from a small town that has only 900 registered voters. The standard error of the percentage of Democrats in the sample will be
- 0
 - About the same as that for the simple random sample from Illinois
 - About the same as that for the simple random sample from the U. S.
 - Much larger than the standard errors from Illinois or the U. S.

Question 10

A nationwide probability sample of 1501 people from the population of all US adults conducted by Pew Research Center found that 52% supported the legalization of marijuana.

Another poll was taken at a website that supports legalization of marijuana. The website is called Legalize.com. Anyone could vote who went to that site. About 10,000 people voted, and 92% said they supported legalizing marijuana.

- a) Based on these results, what do you think is the best guess of the percentage of adults in the US that support legalization of marijuana?
- 92%, because the Legalize.com had many more participants
 - 52%, because the Pew sample was a probability sample of the entire population
 - 52%, because it is closer to 50/50
 - 92%, because the people who voted at Legalize.com are more informed about the issue
- b) Based on the above data, can you conclude that a similar poll at the University of Illinois would also show 52% support?

Question 11

College A has 120 students, 30 of whom are Freshman. College B has 28000 students, 7000 of whom are Freshman. The University of California system has about 400,000 students and about 100,000 are Freshman. Imagine taking a simple random sample of 100 from each of College A, College B, and the University of California system. Focus on the percentages of Freshman in the samples. (The SD's in the relevant boxes are all the same, .43.)

- a) To find the the standard error for the percentage of Freshman in the sample from College A, should you use the correction factor?
- Yes
 - You do not need to use it, but it is ok if you do
 - No, you should not use it
- b) To find the the standard error for the percentage of Freshman in the sample from College B, should you use the correction factor?
- Yes
 - You do not need to use it, but it is ok if you do
 - No, you should not use it
- c) What is the standard error for the percentage of Freshman in the sample from College B?
- d) The standard error for the percentage of Freshman in the sample from the University of California system is the standard error for the percentage of Freshman in the sample from College B.
- much larger than
 - about the same as
 - much smaller than
- the standard error for the percentage of Freshman in the sample from College B.

Question 12

A simple random sample of 225 people from the University of Illinois undergraduates was taken, and the cholesterol level of each person in the sample was recorded. The average cholesterol in the sample was 172, and the SD was 20. Of interest is a confidence interval for the average.

There are about 28,000 UI undergraduates.

- a) What most closely describes the box?
- It has about 28000 tickets, with an average of 172
 - It has about 28000 tickets, with an average of 0
 - It has about 28000 tickets, but the exact average is unknown
 - It has 225 tickets, with an average of 172
 - It has 225 tickets, 20 of them ones, and 205 zero's.

b) What is the standard error of the average of the draws?

c) The 95% confidence interval is closest to

- (169.4, 174.6)
- (152, 192)
- (170.7, 173.3)
- (162, 182)
- It is impossible to find a confidence interval.

d) The confidence interval is for

- The average cholesterol in the sample of 225
- The average cholesterol of all UI undergraduates
- The average cholesterol of the entire nation

Question 13

A simple random sample of 1000 adults from Champaign-Urbana was obtained to estimate the percentage of adults who are registered to vote. (Assume Champaign-Urbana has 50000 adults.) The sample had 45% registered.

a) Which of the following most closely describes the actual box we are considering:

- The box has 450 ones and 550 zeroes;
- The box has 22500 ones and 27500 zeroes;
- The box has 1000 tickets, but we do not know how many are ones and how many are zeroes;
- The box has 50000 tickets, but we do not know how many are ones and how many are zeroes.

b) The standard error of the percentage in the sample is 1.6%. Find the 95% confidence interval for the percentage registered among the Champaign-Urbana adults:

- c) Of interest is whether 50% of the CU adults is registered. What can you say based on the above results?
- We are certain that the percentage of the adults in CU that are registered is 50%
 - We are fairly confident that the percentage of the adults in CU that are registered is less than 50%
 - It is plausible that the percentage of the adults in CU that are registered is 50%
 - We are certain the the percentage of the adults in CU that are registered is less than 50%

Question 14

In 1996, the National Opinion Research Center took a simple random sample of 196 people from the population of the United States. One question the people were asked was how many hours of television they watch per day. The average in the sample was 3.1, and the SD was 2.6.

- a) How many tickets are in the box for this situation?
- 196 ● 607.6 ● Between 196 and 1000 ● More than 1000
- b) The standard error of the average is closest to ● 0.186 ● 0.22 ● 14 ● 36.4 ● 43.4
- c) The 95% confidence interval is (2.73, 3.47). What is true?
- There is a 95% chance that the sample average is between 2.73 and 3.47.
 - We are 95% confident that the sample average is between 2.73 and 3.47.
 - We are 95% confident that the average of the whole country is between 2.73 and 3.47.
 - About 95% of the people in the sample watch television between 2.73 and 3.47 hours per day.

Question 15

In class, I conducted an ESP experiment. I randomly chose 16 cards from a regular deck, and the people in the class tried to guess the cards' suits. If there is no ESP, and everyone is guessing randomly, then the average total number correct the people should get is 4. For the 82 people in the class, the average number guessed correctly was 3.86, and the SD of the number guessed correctly was 1.61.

Consider the null hypothesis that people are guessing randomly, which translates to the experiment being like 82 draws with replacement from a null box. (The numbers on the tickets are the numbers possibly guessed correctly out of 16 cards.)

- a) What is the average of the null box? • 2 • 3.86 • 4 • 6 • It is unknown
- b) What is the standard error of the average of 82 draws with replacement from the box?
- c) What is the Z statistic? (Hint: It is between -1 and 1.)
- d) In the normal curve,
- The area between ± 1 is 68.27%.
 - The area between ± 2 is 95.44%.
 - The area between ± 3 is 99.73%.
 - The area between ± 4 is 99.9937%.
- Is the p-value greater than or less than 5%?
- e) Which of the following conclusions is best supported by the data.
- Reject the null hypothesis.
 - It is plausible that people are just guessing randomly.
 - We know beyond a reasonable doubt that the null hypothesis is true.
 - The class has ESP.

Question 16

A computer program is supposed to randomly generate the digits 1, 2, and 3 so that the chance of each is the same, 33.33%. The computer generated 3000 of these numbers, and there turned out to be 1083 1's. Is that too many 1's?

You now want to do a significance test to see whether the computer program is really working the way it should.

- a) The null box for this hypothesis contains
- Three tickets: one 1, one 2, and one 3.
 - 3000 tickets: 1000 1's, 1000 2's, and 1000 3's.
 - Three tickets: one 1 and two 0's.
 - 3000 tickets, 1083 with 1's, and the rest 0's.
 - 3000 tickets with 0's and 1's on them, but we do not know what the percentages of 0's and 1's are.
- b) The standard error for the number of 1's you should get in 3000 tries is about 26. What is the value of the Z? (Hint: It is over 3.)
- c) Is the p-value for the Z you calculated more or less than 5%?
- d) What should we conclude from the p-value?
- We are certain the null hypothesis is not true.
 - We are fairly confident that the computer program is not working the way it should.
 - We do not have enough evidence to conclude that there is anything wrong with the computer program.
 - We are sure that the computer program is working correctly.

Question 17

A simple random sample of 200 freshman was taken from the U of I, and an independent simple random sample of 500 upperclassmen. Among other questions, they were asked their weight and how many pets they had owned. Here is a table of the averages, as well as the standard error for the difference in averages:

	Freshman Average	Upperclassmen Average	SE of Difference in Averages
Weight	149	161	5.2
# Pets	3.50	3.42	0.3

- a) What is the difference in average weights, freshman minus upperclassmen?
- b) The question is whether that difference is statistically significant. The statistic for testing the null hypothesis that the average heights are equal in the two populations is $Z = 2.3$.

In the normal curve, the area between ± 2.3 is 97.86%. What is the p-value?

- c) Is the difference in the average heights statistically significant?
- d) What is the Z for testing the null hypothesis that the average number of pets is the same in the two populations?
- e) The p-value for this null hypothesis is greater than 5%. Is the difference in the average number of pets statistically significant?

Question 18

A simple random sample of 625 from the entire Illinois adult population showed that 47% were registered to vote. Another independent simple random sample of the same size from Iowa showed 51% registered.

The SD in the boxes for the two populations are essentially the same.

- a) What is your estimate of the SD in the Illinois box?
- b) What is the standard error (SE) of the percentage for the Illinois sample?

- c) The SE of the percentage for the Iowa sample is the same as for Illinois. What is the SE of the difference of the two sample percentages?
- d) What is the value of the Z for testing the null hypothesis that the percentages in the populations who are registered are the same in Illinois and Iowa?
- e) The p-value turns out to be over 5%. What do you do?
- Reject the null hypothesis
 - Fail to reject the null hypothesis
- f) What do you conclude?
- We are fairly confident that the percentage registered in Iowa is higher than in Illinois
 - We do not have enough evidence to say that there is a difference in the populations
 - The difference is statistically significant

Question 19

Polls in London and Berlin asked adults whether they exercised three times a week or not. The observed table is next:

	Exercise 3 or more times per week?	
	Yes	No
London	701	760
Berlin	544	245

The null hypothesis is that exercising and city are independent in the population. The expected table is then the following:

	Exercise 3 or more times per week?	
	Yes	No
London	808.42	652.58
Berlin	436.58	352.42

- a) Are the people in London more likely to exercise at least three times a week than expected under independence, or less likely?
- b) To find the chi-square statistic, we need the values of $(Obs - Exp)^2 / Exp$. What is the value for the London people who do exercise three or more times per week?
- c) The chi-square statistic is the sum of how many of those values?
- d) The chi-square statistic turns out to be 91.13. What are the degrees of freedom?
- e) Here is part of the chi-square table:

Degrees of freedom	99%	95%	90%	70%	50%	30%	10%	5%	1%	← p-value
1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64	
2	0.020	0.10	0.21	0.71	1.39	2.41	4.60	5.99	9.21	
3	0.12	0.35	0.58	1.42	2.37	3.67	6.25	7.82	11.34	
4	0.30	0.71	1.06	2.20	3.36	4.88	7.78	9.49	13.28	

What do you do? ● Fail to reject the null hypothesis. ● Reject the null hypothesis

Question 20

A simple random sample of 712 from the population of all STAT100 students was taken. People are classified into gender (Male and Female) and Year in School: Freshman, Sophomore, Junior and Senior or above. The observed table is below.

	Men	Women	Total
Freshman	75	157	232
Sophomore	104	205	309
Junior	21	69	90
Senior & above	27	54	81
Total	227	485	712

The null hypothesis is that year and gender are independent in the population.

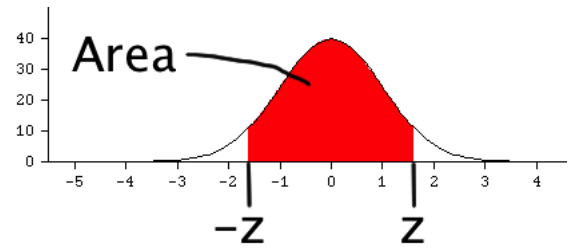
a) What is the value in the expected table for the freshman men?

b) The value of the chi-square statistic is 3.58. What are the degrees of freedom?

The p-value is closest to which of the following? (You can use the chart in the previous question.) ● 70% ● 30% ● 10% ● 5% ● 1%

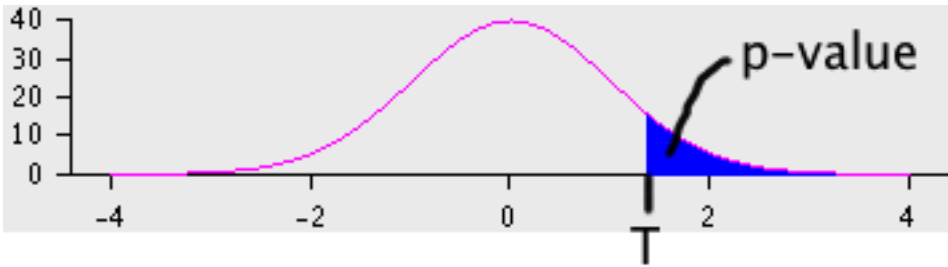
c) What do you do? ● Fail to reject the null hypothesis. ● Reject the null hypothesis

A.3 Normal table



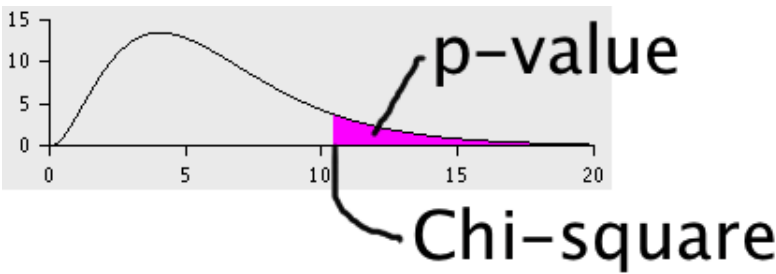
$z \downarrow$	Area between $\pm z$	$z \downarrow$	Area between $\pm z$	$z \downarrow$	Area between $\pm z$	$z \downarrow$	Area between $\pm z$
0.00	0.00	1.00	68.27	2.00	95.45	3.00	99.73
0.05	3.99	1.05	70.63	2.05	95.96	3.05	99.77
0.10	7.97	1.10	72.87	2.10	96.43	3.10	99.81
0.15	11.92	1.15	74.99	2.15	96.84	3.15	99.84
0.20	15.85	1.20	76.99	2.20	97.22	3.20	99.86
0.25	19.74	1.25	78.87	2.25	97.56	3.25	99.88
0.30	23.58	1.30	80.64	2.30	97.86	3.30	99.90
0.35	27.37	1.35	82.30	2.35	98.12	3.35	99.92
0.40	31.08	1.40	83.85	2.40	98.36	3.40	99.93
0.45	34.73	1.45	85.29	2.45	98.57	3.45	99.94
0.50	38.29	1.50	86.64	2.50	98.76	3.50	99.95
0.55	41.77	1.55	87.89	2.55	98.92	3.55	99.96
0.60	45.15	1.60	89.04	2.60	99.07	3.60	99.97
0.65	48.43	1.65	90.11	2.65	99.20	3.65	99.97
0.70	51.61	1.70	91.09	2.70	99.31	3.70	99.98
0.75	54.67	1.75	91.99	2.75	99.40	3.75	99.98
0.80	57.63	1.80	92.81	2.80	99.49	3.80	99.99
0.85	60.47	1.85	93.57	2.85	99.56	3.85	99.99
0.90	63.19	1.90	94.26	2.90	99.63	3.90	99.99
0.95	65.79	1.95	94.88	2.95	99.68	3.95	99.99

A.4 Student's t table



Degrees of freedom ↓	30%	10%	5%	2.5%	1%	0.5%	← p-value
1	0.73	3.08	6.31	12.71	31.82	63.66	← T
2	0.62	1.89	2.92	4.30	6.96	9.92	
3	0.58	1.64	2.35	3.18	4.54	5.84	
4	0.57	1.53	2.13	2.78	3.75	4.60	
5	0.56	1.48	2.02	2.57	3.36	4.03	
6	0.55	1.44	1.94	2.45	3.14	3.71	
7	0.55	1.41	1.89	2.36	3.00	3.50	
8	0.55	1.40	1.86	2.31	2.90	3.36	
9	0.54	1.38	1.83	2.26	2.82	3.25	
10	0.54	1.37	1.81	2.23	2.76	3.17	
11	0.54	1.36	1.80	2.20	2.72	3.11	
12	0.54	1.36	1.78	2.18	2.68	3.05	
13	0.54	1.35	1.77	2.16	2.65	3.01	
14	0.54	1.35	1.76	2.14	2.62	2.98	
15	0.54	1.34	1.75	2.13	2.60	2.95	
16	0.54	1.34	1.75	2.12	2.58	2.92	
17	0.53	1.33	1.74	2.11	2.57	2.90	
18	0.53	1.33	1.73	2.10	2.55	2.88	
19	0.53	1.33	1.73	2.09	2.54	2.86	
20	0.53	1.33	1.72	2.09	2.53	2.85	
21	0.53	1.32	1.72	2.08	2.52	2.83	
22	0.53	1.32	1.72	2.07	2.51	2.82	
23	0.53	1.32	1.71	2.07	2.50	2.81	
24	0.53	1.32	1.71	2.06	2.49	2.80	
25	0.53	1.32	1.71	2.06	2.49	2.79	

A.5 Chi-square table



Degrees of freedom ↓	30%	10%	5%	1%	0.1%	← p-value
1	1.07	2.71	3.84	6.63	10.83	← Chi-square
2	2.41	4.61	5.99	9.21	13.82	
3	3.66	6.25	7.81	11.34	16.27	
4	4.88	7.78	9.49	13.28	18.47	
5	6.06	9.24	11.07	15.09	20.52	
6	7.23	10.64	12.59	16.81	22.46	
7	8.38	12.02	14.07	18.48	24.32	
8	9.52	13.36	15.51	20.09	26.12	
9	10.66	14.68	16.92	21.67	27.88	
10	11.78	15.99	18.31	23.21	29.59	
11	12.90	17.28	19.68	24.72	31.26	
12	14.01	18.55	21.03	26.22	32.91	
13	15.12	19.81	22.36	27.69	34.53	
14	16.22	21.06	23.68	29.14	36.12	
15	17.32	22.31	25.00	30.58	37.70	
16	18.42	23.54	26.30	32.00	39.25	
17	19.51	24.77	27.59	33.41	40.79	
18	20.60	25.99	28.87	34.81	42.31	
19	21.69	27.20	30.14	36.19	43.82	
20	22.77	28.41	31.41	37.57	45.31	
21	23.86	29.62	32.67	38.93	46.80	
22	24.94	30.81	33.92	40.29	48.27	
23	26.02	32.01	35.17	41.64	49.73	
24	27.10	33.20	36.42	42.98	51.18	