Chapter 1: Stats Starts Here (p. 2)

- Why am I here?
- Am I going to come out alive?
- What can I hope to learn?
  - how to *understand* the numbers that make our world
  - how to organize things to *make decisions* about our world.

Things vary:

- people are different
- can't see everything or measure it all
- what we *can* measure might be inaccurate.

How do we make sense of an imperfect picture of an imperfect world?

Chapter 2: Data (p. 7)

Airlines monitored for safety and customer service. For each flight, carriers must report:

- flight number
- type of aircraft
- number of passengers
- how late the flight was (0=on time)
- any mechanical problems.

Who, what, why, where, when, how?

- Who (individuals)?
  - Flights
- What (variables)?
  - Flight#, aircraft, passengers, how late, problems
- Why these variables?
  - Monitoring aircraft safety
- When?
  - *Don't know*
- Where?
  - World-wide
- How?
  - Records from each flight when it lands

Identifier variables (identify individuals):

- flight number

Categorical and quantitative variables

- Things we measure are of different kinds:
  - "What is your favourite colour" could be "red", "green", "blue" etc.
  - "What level of disease do you have" could be "none", "moderate", "severe"
  - These are **categorical**: can only count how many individuals per category.
- What about "how tall are you"? Or "what temperature is it outside now?"
  - Answers to these are *numbers (with units)*: 5 ft 9 inches, 26 degrees C. **Quantitative** variables.

Kentucky Derby

Year	Winner	Margin	Jockey	Duration	Track condition
2004	Smarty Jones	2 ¾	S. Elliott	2:04.06	Sloppy
2005	Giacomo	½	M. Smith	2:02.75	Fast
2006	Barbaro	6 ½	E. Prado	2:01.36	Fast
2007	Street Sense	2 ¼	C. Borel	2:02.17	Fast
2008	Big Brown	4 ¾	K. Desormeaux	2:01.82	Fast
2009	Mine That Bird	6 ¾	C. Borel	2:02.66	
2010	Super Saver	2 ½	C. Borel	2:04.45	
2011	Animal Kingdom	2 ¾	J. Velazquez	2:02.04	
2012	I'll Have Another	1 ½	M. Gutierrez	2:01.83	
2013	Orb	2 ½	J. Rosario	2:02.89	
2014	California Chrome		V. Espinoza	2:03.66	

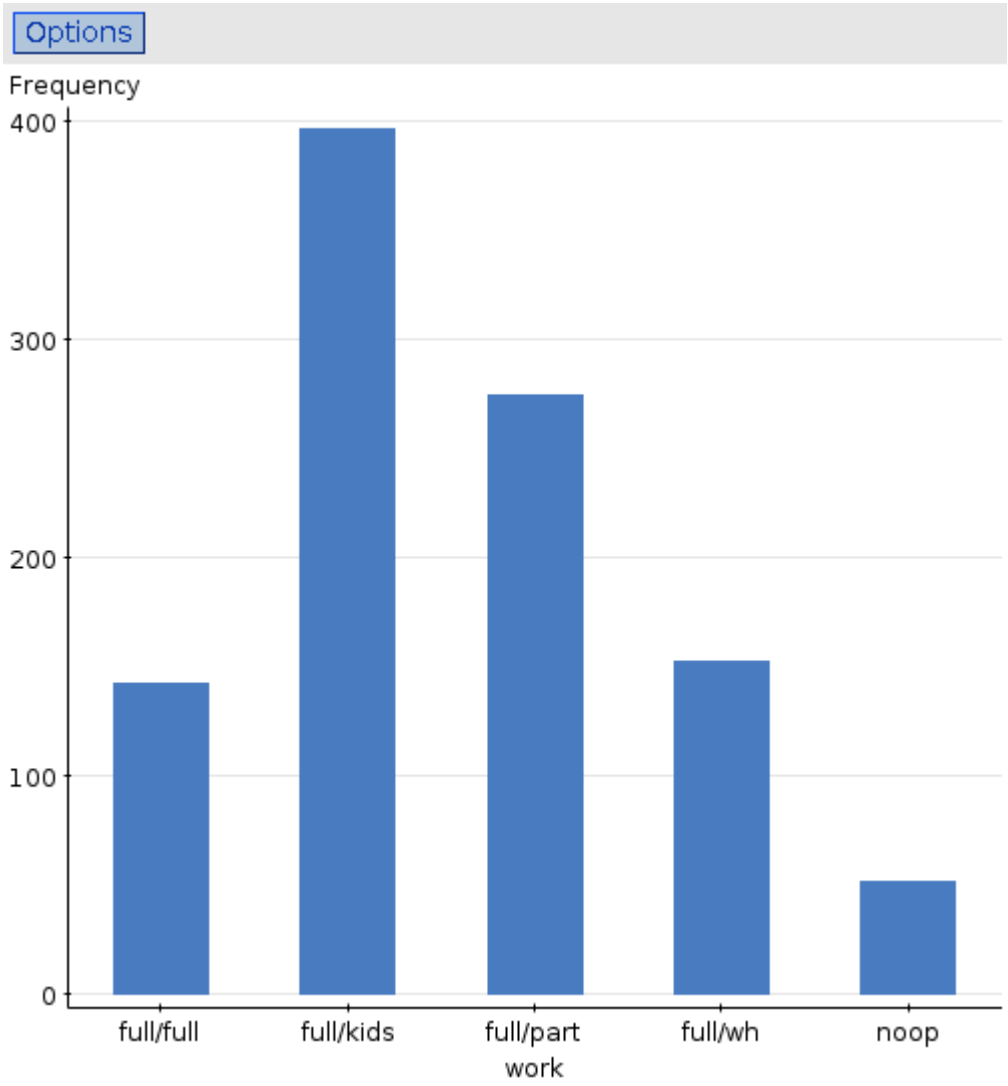
- Who (cases): winners of Kentucky Derby horse race
- What (variables):
  - year (quantitative, a number)
  - winning horse's name (categorical)
  - margin of victory (quantitative; horse lengths)
  - jockey's name (categorical)
  - time to run the race (quantitative; minutes and seconds)
  - track condition (categorical)
- How: obtained from textbook, Wikipedia
- Why: one of these
  - because it looked interesting!
  - To note any trends over time (eg. Are the winners running faster?)
  - to see whether certain jockeys have won the race multiple times
- When: for races 2004-2014; collected by me this morning!
- Where: just the Kentucky Derby horse race (not any others).

Chapter 3: Displaying and describing categorical data (p. 20)

In 1991 and again in 2001, a poll was taken of 1015 adults about their opinions on working parents. The question was "considering the needs of adults and children, what do you see as the ideal family in today's society?"

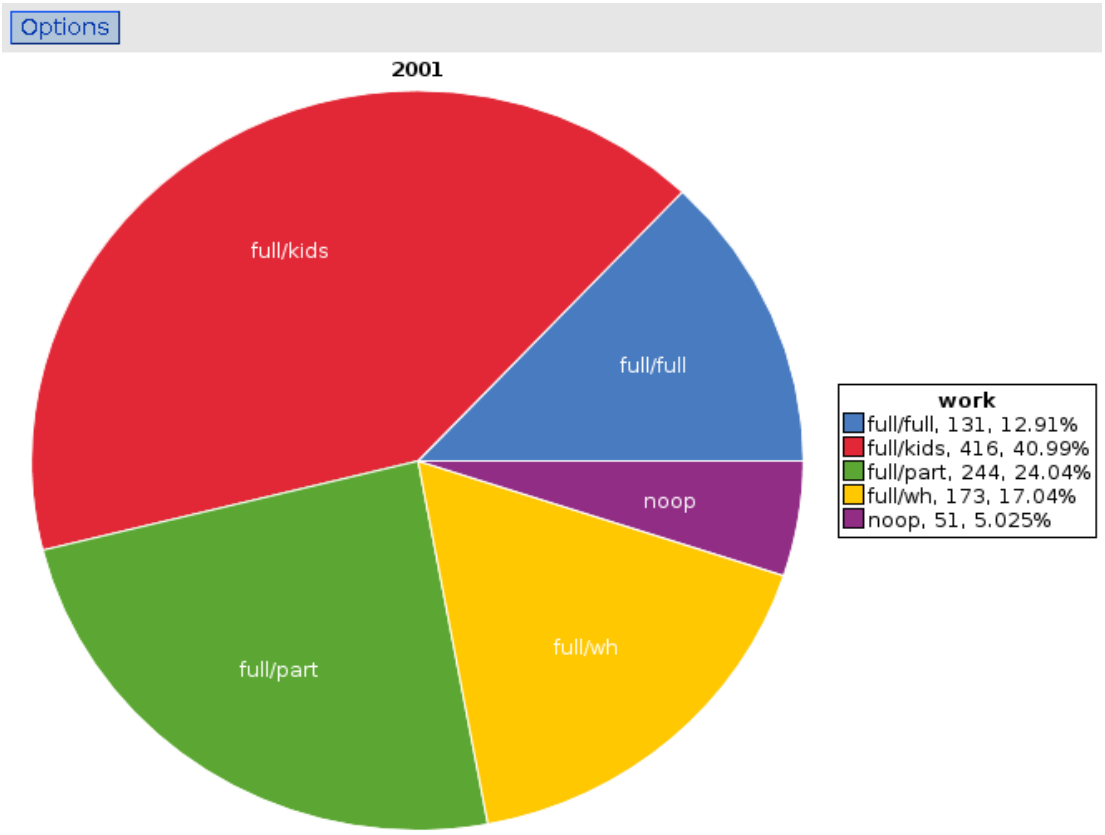
	1991	2001
Both work full time	142	131
One works full time, other part time	274	244
One works, other works at home	152	173
One works, other stays at home for kids	396	416
No opinion	51	51

The 1991 data, bar chart:



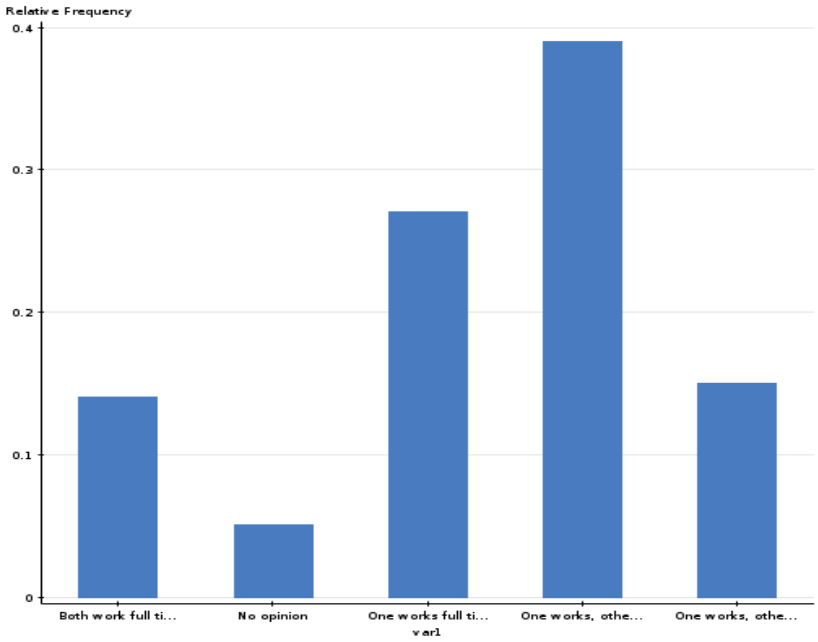
- more people think that having one parent stay home with kids is ideal
- but all four options chosen by reasonable number of people
- *StatCrunch: graphics, bar plot, with summary*

2001 data, pie chart

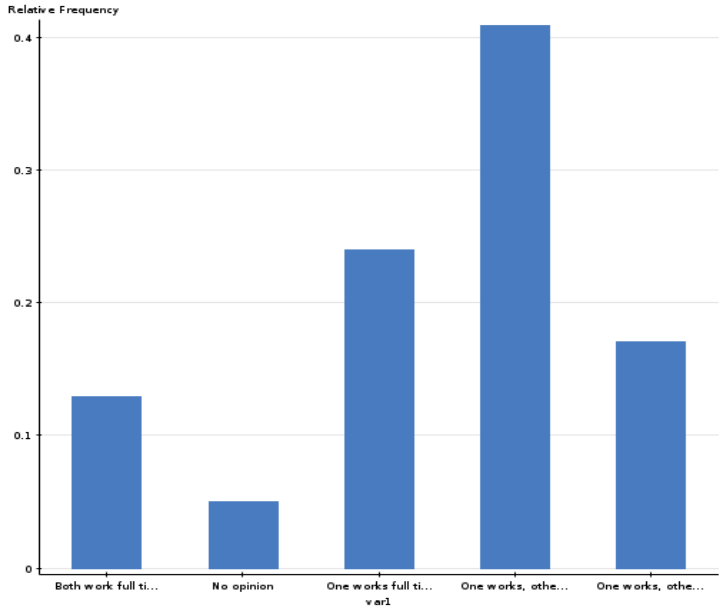


- Almost half of all people think that one parent ideally would stay at home with kids
- about a quarter of all people think that one parent working part time is ideal
- about equal numbers think that having both parents work full time or having one parent work at home ideal
- *StatCrunch: graphics, pie chart, with summary.*

– Bar charts



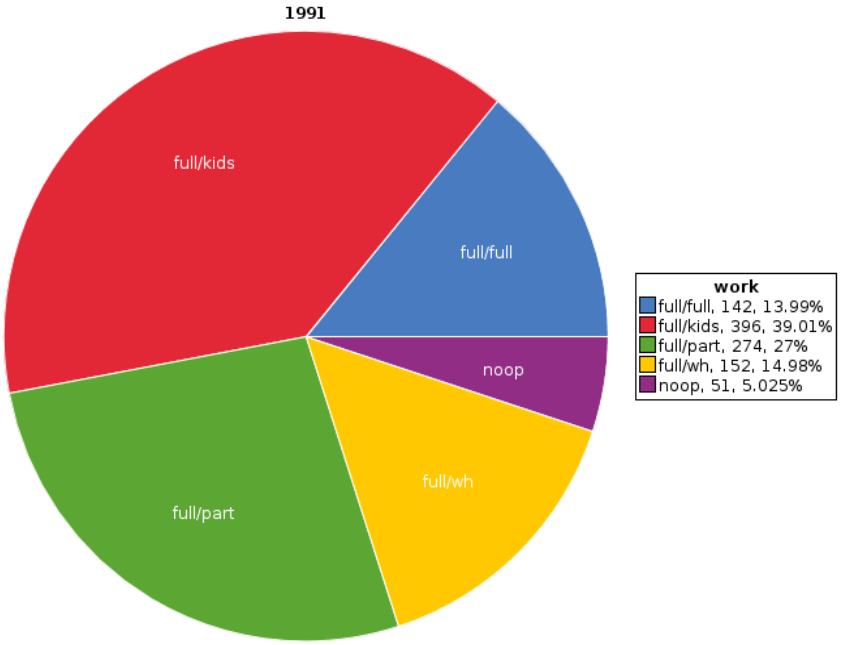
1991

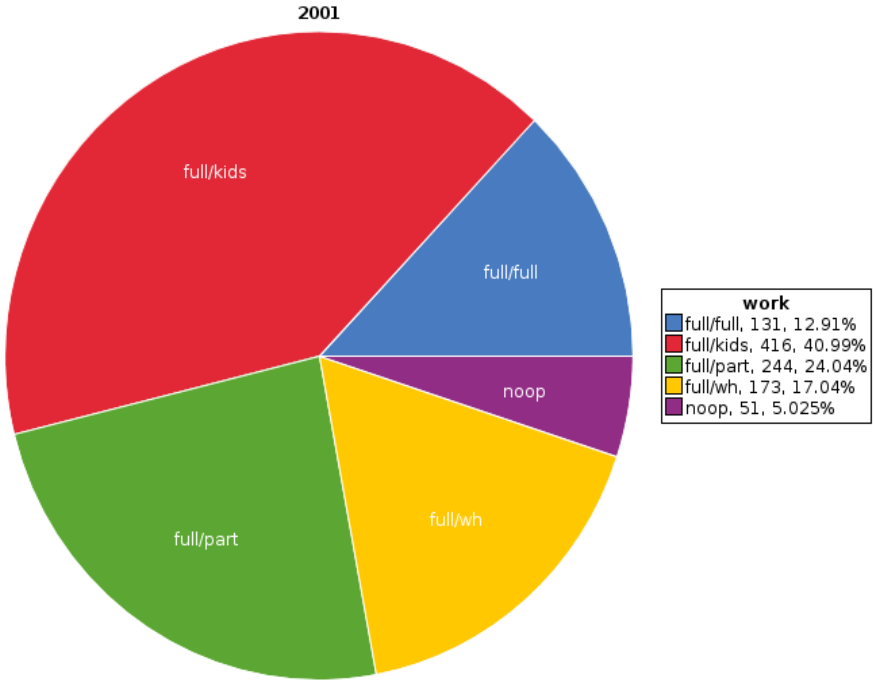


2001

Hard to see much difference.

Pie charts



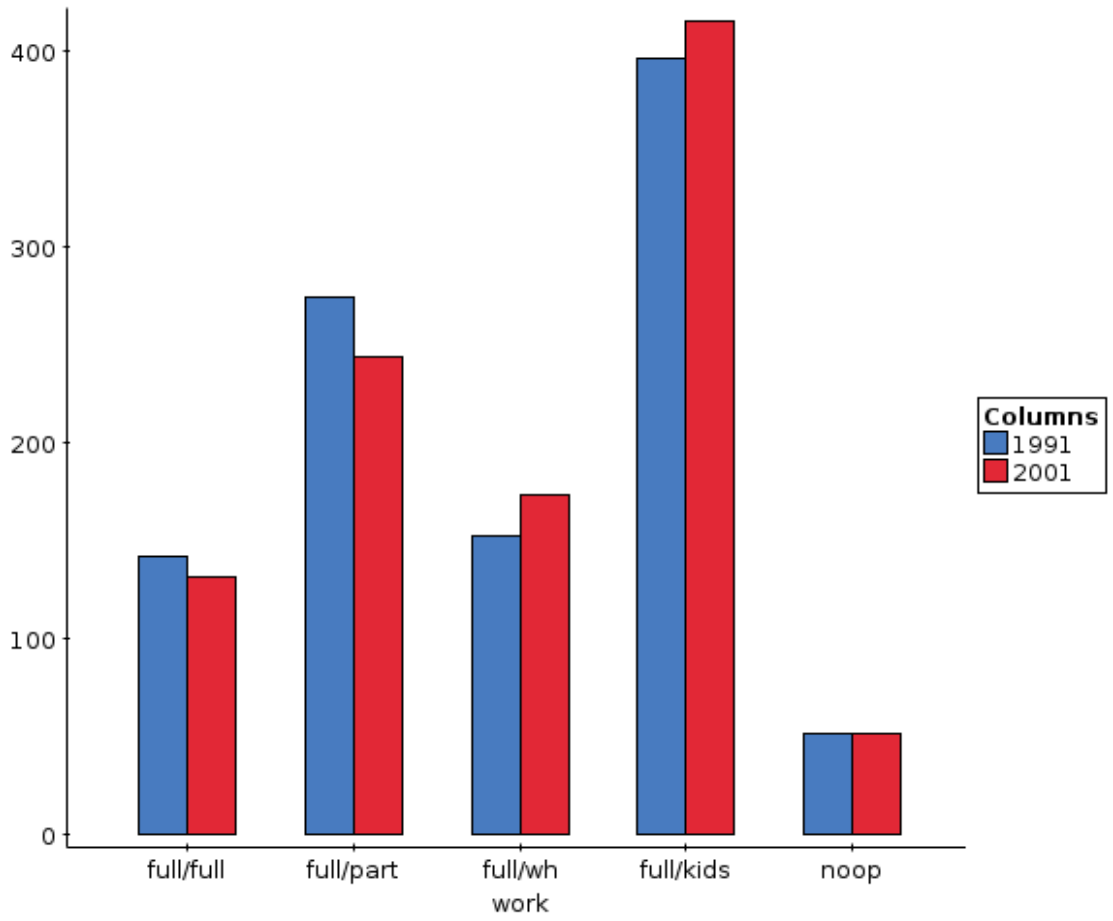


Again, not much difference apparent.

Ken's editorial on pie charts: **they should never be used.** Why? Because the eye judges angles *badly*, heights/distances much better (thus bar chart better than pie chart).

Better: make bar chart for each year, but put bars *side by side*:

- Slightly more had one parent at home, slightly fewer had both parents working outside the home. But differences *are* small.
- *StatCrunch: Graphics, Chart, Columns.*



Contingency tables: two (or more) categorical variables (p.24)

recall surveys on attitudes to child care, above:

	1991	2001
Both work full time	142	131
One works full time, other part time	274	244
One works, other works at home	152	173
One works, other stays at home for kids	396	416
No opinion	51	51

Reading a contingency table

- University records applications to professional schools:

	Accepted	Rejected	Total
Males	490	210	700
Females	280	220	500
Total	770	430	1200

- 280 of the applicants were females and accepted.
- How many of the applicants were males who were rejected?
  - 210
- How many females applied altogether?
  - 500

Percentages (p.25)

	Accepted	Rejected	Total
Males	490	210	700
Females	280	220	500
Total	770	430	1200

- More males than females applied (and more people accepted than not), so difficult to compare numbers.
- Compute *percentages.* One way: percent of total (divide everything by 1200).
- *joint distribution.*
- Stat, Tables, Contingency, With Summary.

Percent of total (p.26)
(take output from StatCrunch, copy, *paste special)*

	Accepted	Rejected	Total
Males	490 (40.83%)	210 (17.5%)	700 (58.33%)
Females	280 (23.33%)	220 (18.33%)	500 (41.67%)
Total	770 (64.17%)	430 (35.83%)	1200 (100%)

- 41% of all applicants were males who were accepted.
- The row marked Total is *marginal distribution of acceptance* (64% of all applicants were accepted)
- The column marked Total is *marginal distribution of gender (*42% of all applicants were female).

Conditional distribution

- Joint distribution is “out of everything”.
- Doesn't answer question “are more males than females accepted”?
- For that: *out of males*, what % accepted – *row percents.*
- *Statcrunch: as before, but display: row percent.*

	Accepted	Rejected	Total
Males	490 (70%)	210 (30%)	700 (100%)
Females	280 (56%)	220 (44%)	500 (100%)
Total	770 (64.17%)	430 (35.83%)	1200 (100%)

Table again:

	Accepted	Rejected	Total
Males	490 (70%)	210 (30%)	700 (100%)
Females	280 (56%)	220 (44%)	500 (100%)
Total	770 (64.17%)	430 (35.83%)	1200 (100%)

- See males and females both add up to 100%.
- 70% of male applicants accepted, but only 56% of female applicants.
- Discrimination?

Column percents

look like this:

	Accepted	Rejected	Total
Males	490 (63.64%)	210 (48.84%)	700 (58.33%)
Females	280 (36.36%)	220 (51.16%)	500 (41.67%)
Total	770 (100%)	430 (100%)	1200 (100%)

- 63% of people accepted were males.
- 51% of people rejected were females.
- Doesn't answer our question here.

Deciding between row and column percents

- Look for words “out of”: “out of females, what % accepted”.
- Look for “outcome”. Here, gender fixed, but acceptance or rejection *in columns* was outcome. So need *row* percents. (Ie. thing that is *not* outcome.)
- Whichever you use, getting *conditional distribution.* “**If** I look at females, what % are accepted” = conditional distribution of acceptance for females.

Another example: airline punctuality

	America West	Alaska	Total
On time	6438	3274	9712
Delayed	787	501	1288
Total	7225	3775	11000

- what is the outcome variable?
  - *On time/delayed in rows*
- do we want row or column percents?
  - *Column percents*
- which airline is more punctual (StatCrunch)?
  - *Data airline-delayed*
  - *Stat, Tables, Contingency, With Summary*
  - *select airlines as columns, appropriate Display*
  - *need to Paste Special, HTML*

Cell format
Count (Column percent)

	America West	Alaska	Total
On time	6438 (89.11%)	3274 (86.73%)	9712 (88.29%)
delayed	787 (10.89%)	501 (13.27%)	1288 (11.71%)
Total	7225 (100%)	3775 (100%)	11000 (100%)

Three categorical variables and Simpson's paradox

Professional schools example: also recorded acceptance and rejection separately for law school and business school:

Law	accepted	rejected	total
males	10	90	100
females	100	200	300
total	110	290	400

Business	accepted	rejected	total
males	480	120	600
females	180	20	200
total	660	140	800

What would be appropriate percents to find here, and what do we conclude? Think first about total for both schools.

Professional schools (data school-accept)

- both schools (row percents)

Contingency table results:
Rows: gender
Columns: None

Cell format
Count (Row percent)

	accepted-all	rejected-all	Total
males	490 (70%)	210 (30%)	700 (100%)
females	280 (56%)	220 (44%)	500 (100%)
Total	770 (64.17%)	430 (35.83%)	1200 (100%)

More males than females accepted.

- Law school:

Cell format
Count (Row percent)

	accepted-law	rejected-law	Total
males	10 (10%)	90 (90%)	100 (100%)
females	100 (33.33%)	200 (66.67%)	300 (100%)
Total	110 (27.5%)	290 (72.5%)	400 (100%)

- More females than males accepted
- different from total.

- Business school:

Cell format
Count (Row percent)

	accepted-bus	rejected-bus	Total
males	480 (80%)	120 (20%)	600 (100%)
females	180 (90%)	20 (10%)	200 (100%)
Total	660 (82.5%)	140 (17.5%)	800 (100%)

- More females than males accepted
- in contrast to total again.

Summary of results:

	Male % accepted	Female % accepted
Law school	10	33
Business school	80	90
Overall	**70**	**56**

- More females accepted at *each* school.
- More males accepted overall
- how is that possible???

Why we get this answer

- Look at where males tend to apply
  - is it easy to be accepted there?
    - *Business school; easy*
- Look at where females tend to apply
  - is it easy to be accepted there?
    - *Law school; difficult.*

- Acceptance depends mainly on *where you apply*, not on whether you are male or female.
- In fact, females have *larger* acceptance rate, other things being equal.
- If the *same* number of females applied to each school, the overall percent of females accepted would be higher.
- Original comparison of overall acceptance rates is apples vs. oranges, because it mixes up two different things (schools).

Actually, the airline example also contains a Simpson's paradox; the extra variable there is “airport”. Percent delayed:

Airport	Alaska	America West
Los Angeles	11.4	14.4
Phoenix	5.2	7.9
San Diego	8.6	14.5
San Francisco	16.9	28.7
Seattle	14.2	23.2
Total	13.3	10.9

- Alaska Airlines more often late overall,
- but *less* often late at every single airport (makes no sense!)
- Explanation: America West flies more often into Phoenix (easy to be on time), but Alaska Airline flies more often into San Francisco/Seattle (hard to be on time).
- Airport makes a difference, so *do not* calculate averages over airport. Look at airports separately.

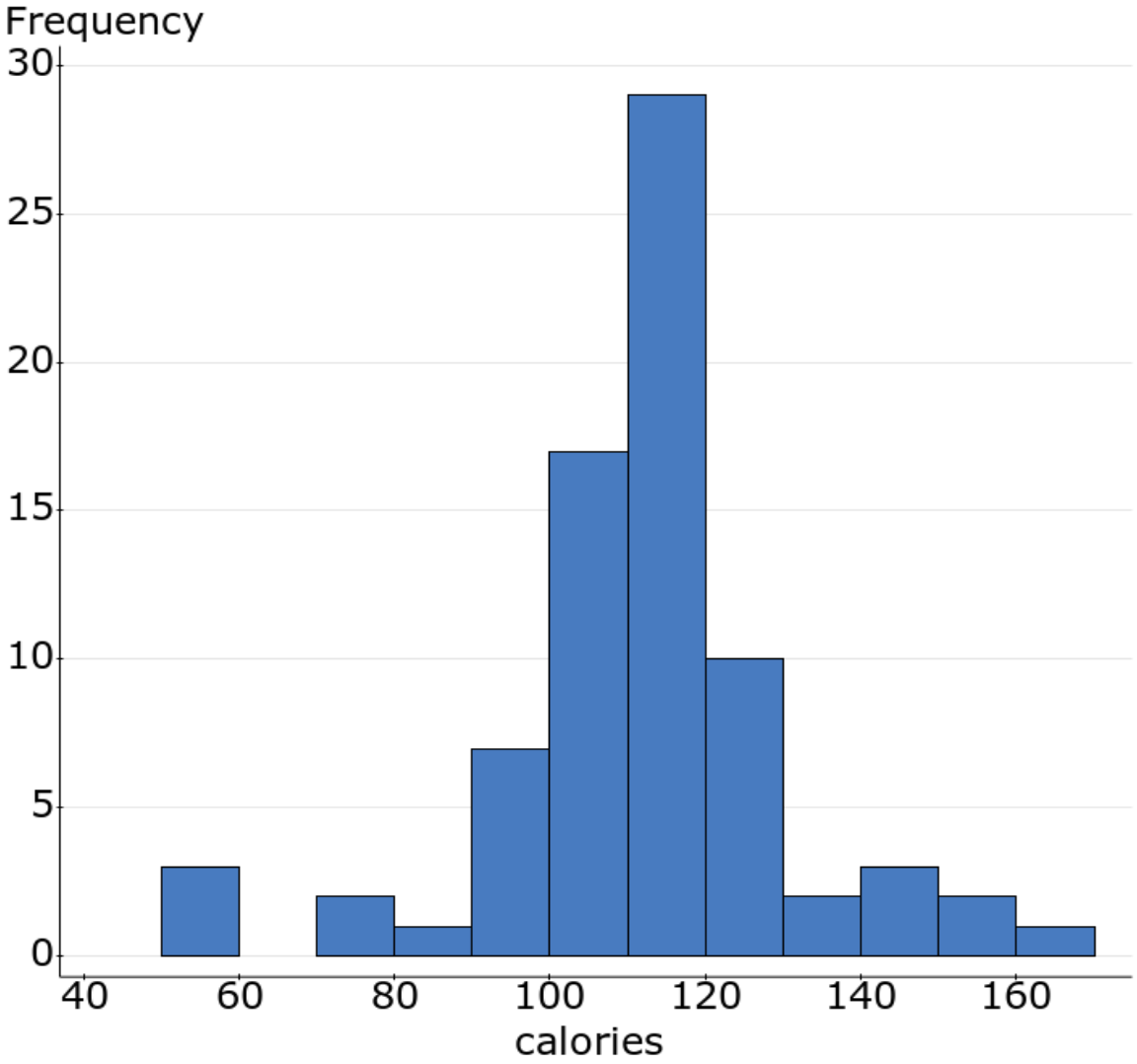
Chapter 4: Displaying and summarizing quantitative data (p. 49)

The breakfast cereal data

Study collected data on nutritional content per serving (and other things) of 77 different breakfast cereals, so that different cereals can be compared.

Mostly quantitative variables.

Histogram for calories per serving (p.50)



- most cereals have between 100 and 120 calories/serving.
- A few have a lot more or a lot less, but only a few.
- Shape *symmetric*: falls away from peak about same both sides.

Statcrunch

- Histogram: Graph, Histogram. Select Column (click on it), select Compute.
- Stemplot (in a moment): Graph, Stem and Leaf, Select Column, Compute.

Stemplot (p.51)

- Alternative to histogram.
- Divide data values into **stems** and **leaves**, plot separately.
- Eg. data 17, 19, 21, stems as 10s (leaves as 1s):
  - 17: stem 1, leaf 7.
  - 19: stem 1, leaf 9.
  - 21: stem 2, leaf 1.
- Plot all stems on left, add leaves to appropriate stem:

```
1: 7 9
2: 1
```

- You choose what units for stems.
- Cereal calories, StatCrunch used 10s for stems again:

Cereal calories stemplot

Variable: calories

Decimal point is 1 digit(s) to the right of the colon.

Low : 50, 50, 50

```
 7 : 00
 7 :
 8 : 0
 8 :
 9 : 0000000
 9 :
10 : 0000000000000000
10 :
11 : 00000000000000000000000000000
11 :
12 : 0000000000
12 :
13 : 00
13 :
14 : 000
```

High: 150, 150, 160

Variable: calories

Decimal point is 1 digit(s) to the right of the colon.

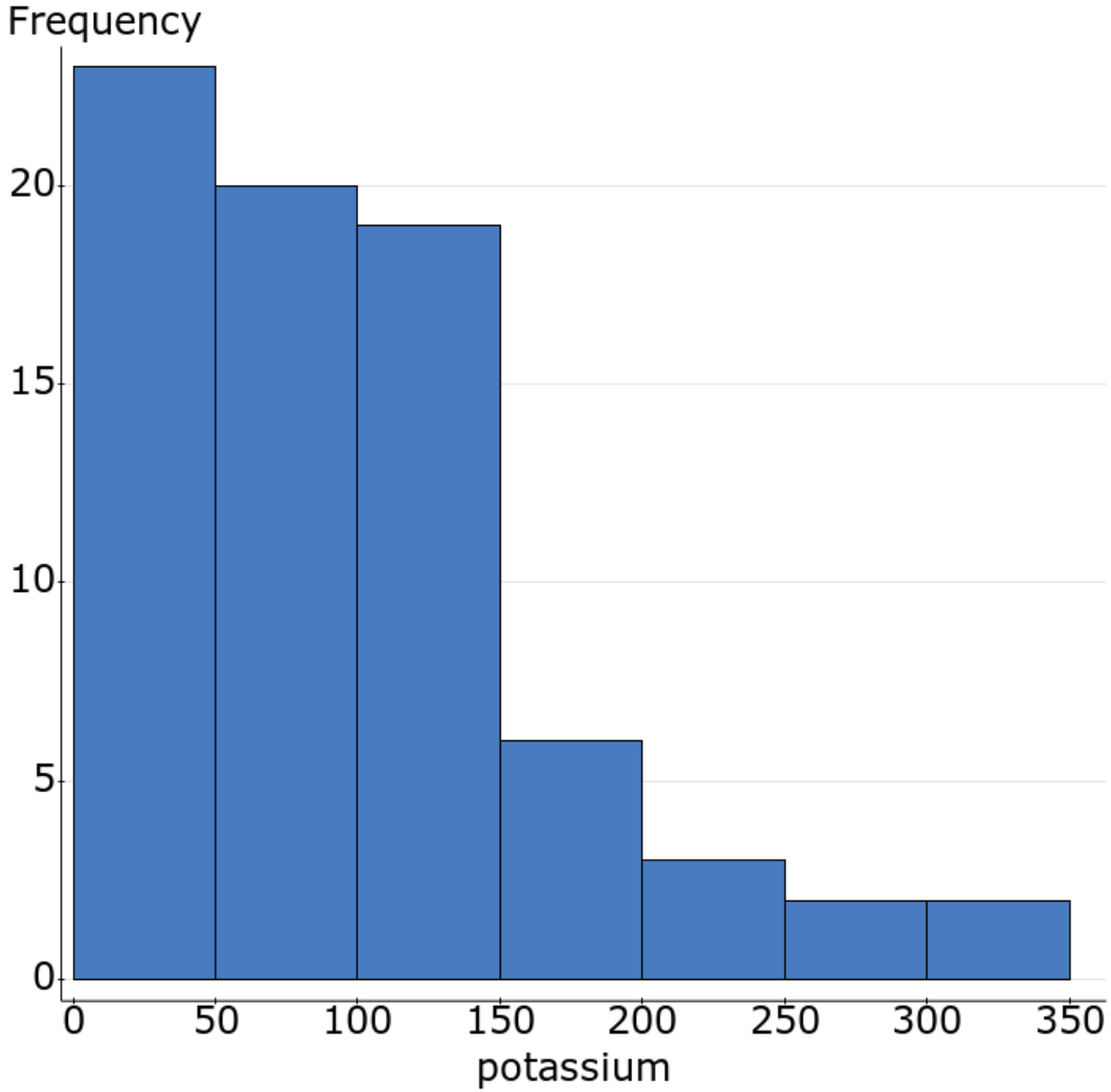
```
Low : 50, 50, 50

 7 : 00
 7 :
 8 : 0
 8 :
 9 : 0000000
 9 :
10 : 00000000000000000
10 :
11 : 0000000000000000000000000000
11 :
12 : 0000000000
12 :
13 : 00
13 :
14 : 000

High: 150, 150, 160
```

- same shape as histogram (turned on side)
- unusual values listed at top and bottom
- smallest value actually on plot is 70, largest 140.
- All leaves are 0:
  - actually only measured to nearest 10
  - couldn't see from histogram.

Cereal potassium data histogram:



- Peak on left with long straggle of values to right
- Has *right-skewed* shape.
- Unlike calories, which was symmetric.
- Distribution of values with long straggle (tail) to left would be *left-skewed*.

Potassium stem-and-leaf:

Variable: potassium

Decimal point is 2 digit(s) to the right of the colon.

```
0 : 22333333333444444444
0 : 55555666666778999999
1 : 000000011111111222233444
1 : 667799
2 : 034
2 : 68
3 : 23
```

- *Right-skewed* shape shows up as long straggle at *bottom* of picture.
- Highest value (end of last line) 330 (not 3.3, not 33, but 330).
- Lowest value (start of 1st line) 20.

The mean and median (measures of "centre"; p.57)

- mean: "average", add up values and divide by how many
- median: sort values into order, pick out middle one (or mean of 2 middle ones)

  data: 8, 12, 7, 5, 4

- mean (8+12+7+5+4)/5=7.2
- median: in order 4, 5, 7, 8, 12, so median=7
  - with values 4, 5, 7, 8, 9, 12, median would be (7+8)/2=7.5.
- *with n values, median is (n+1)/2-th value*
  - *n=5, median=6/2=3*
  - *n=6, median=7/2=3.5*

Mean and median from cereal data:

Summary statistics:

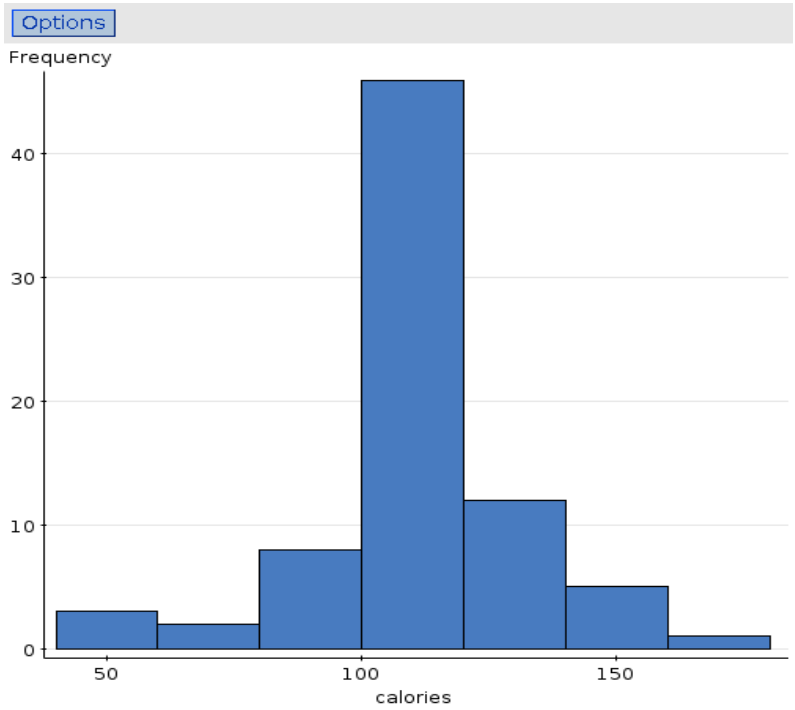
Column	Mean	Median
calories	106.88312	110

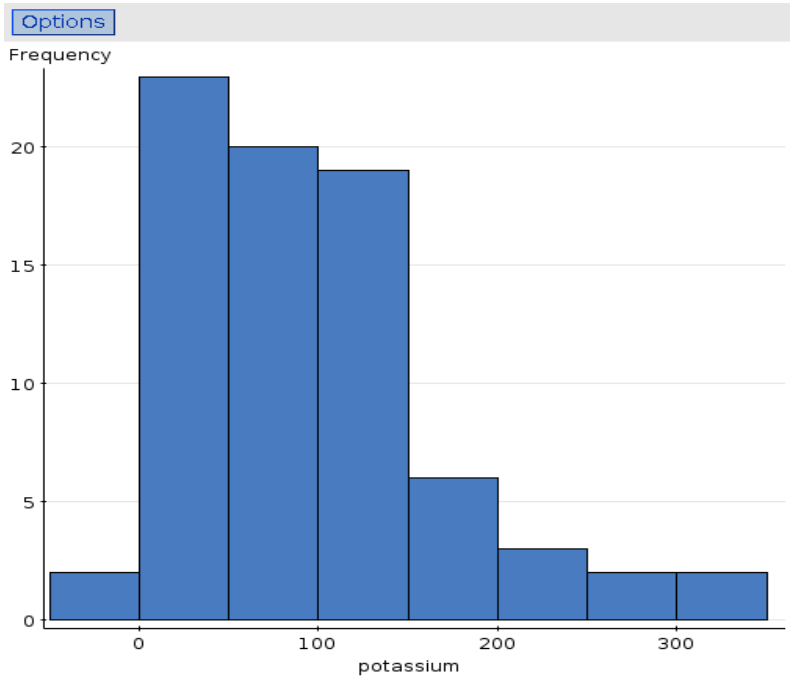
– Calories: median a little bigger than mean, but close together given nature of data

Summary statistics:

Column	Mean	Median
potassium	98.666667	90

– Potassium: mean bigger than median, because distribution right-skewed.

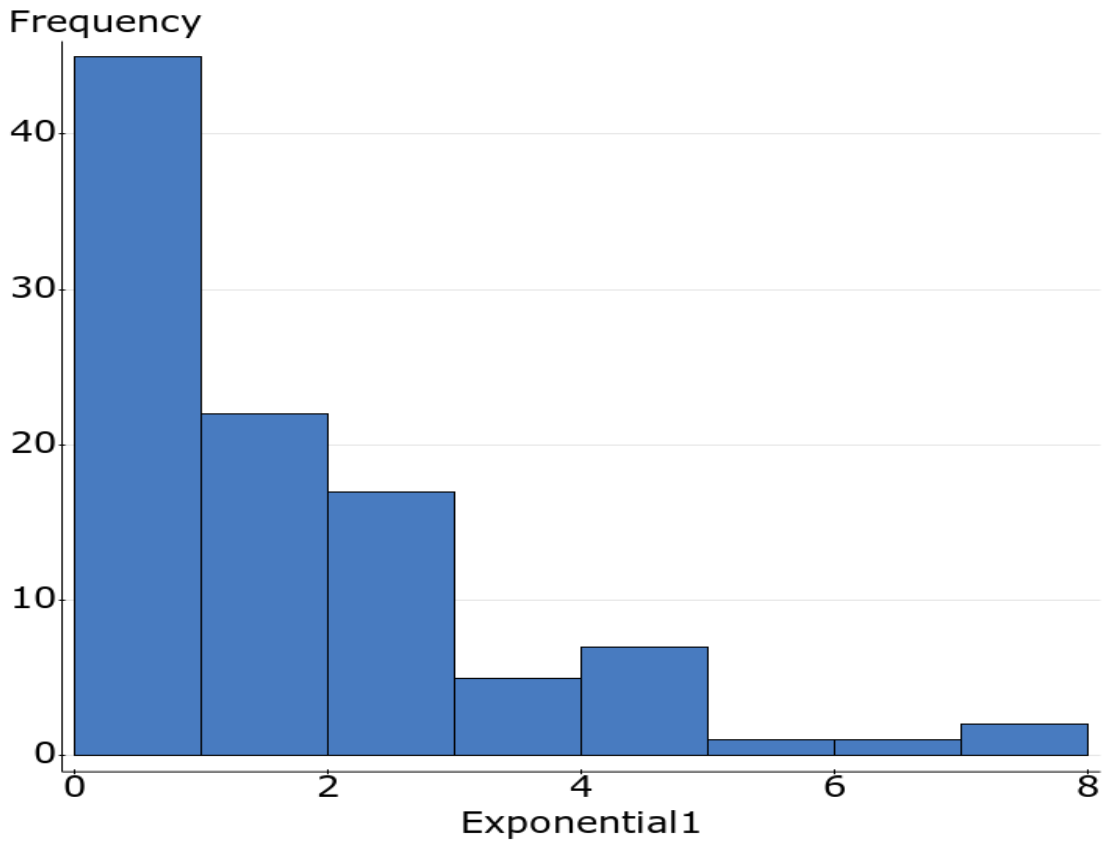




Another example:

Summary statistics:

Column	Mean	Median
Exponential1	1.6502398	1.1666597



- Very right-skewed
- Mean much bigger than median

With a lower or upper limit, there is "only one way" for a variable to go, especially if a lot of values close to the limit.

In the situations below, is there an upper or lower limit on the values of the variable? Which way would you expect the variable to be skewed?

- waiting time to be served at a bank
  - *lower limit 0; skewed right*
- number of employees in companies based in Scarborough
  - *lower limit 1; skewed right*
- scores on an easy quiz (marked out of 10)
  - *upper limit 10; skewed left*

Spread: interquartile range (p.61)

- 1st quartile Q_1 has ¼ of data values below it and ¾ above
- 3rd quartile Q_3 has ¾ of data values below it and ¼ above
- Find a quartile by taking lower (upper) half of data, and finding median of that half.text to be copied
- Interquartile range is IQR=Q_3-Q_1. Larger = more spread out.

Example: 2, 5, 7, 7, 9

- lower half 2, 5, 7 (include middle), so Q_1=5
- upper half 7, 7, 9 so Q_3= 7
- IQR=7-5=2
- *IQR not affected by extremely high or low values; same as median this way.*

Summary statistics:

Column	Q1	Q3	IQR
x	5	7	2

Standard deviation (SD; p.63): another measure of spread

Illustrate with example. Data as above, mean 6:

Data	Minus mean	Squared
2	-4	16
5	-1	1
7	1	1
7	1	1
9	3	9
Total	0	28

So **variance** is 28/(5-1)=7

and therefore SD is $\sqrt{7}$ = 2.65.

Summary statistics:

Column	Variance	Std. dev.
x	7	2.6457513

Fire up StatCrunch and enter these numbers into a column:

1, 2, 3, 4, 5.

- Find the mean and median. Are they the same? Would you expect them to be?
  - *Yes; yes (symmetric, no outliers)*
- Replace the number 5 with 10, and find the mean and median again. Are they still the same? If not, which is bigger?
  - *No; mean is bigger.*
- Now replace the 10 with 20. What has happened now? Do you think the mean or median is the better choice for the "centre"?
  - *Mean is even further bigger than median. Mean pulled upwards by outlier; use median.*

Optional extra: repeat for IQR and standard deviation.

Summary statistics:

Column	Mean	Median	Std. dev.	Q1	Q3	IQR
5	3	3	1.5811388	2	4	2
10	4	3	3.5355339	2	4	2
20	6	3	7.9056942	2	4	2

Chapter 5: understanding and comparing data (p. 88)

Data 10, 11, 14, 15, 17, 19, 21, 28, 35:

- why is median 17?
  - n=9 median is (9+1)/2=5th
- find Q_1 and Q_3
- Q1=median of 10,11,14,15,17 = 14
- Q3=median of 17, 19, 21, 28, 35=21
- find interquartile range
  - 21-14=7
- find 5-number summary min, Q_1, median, Q_3, max.
  - 10, 14, 17, 21, 35

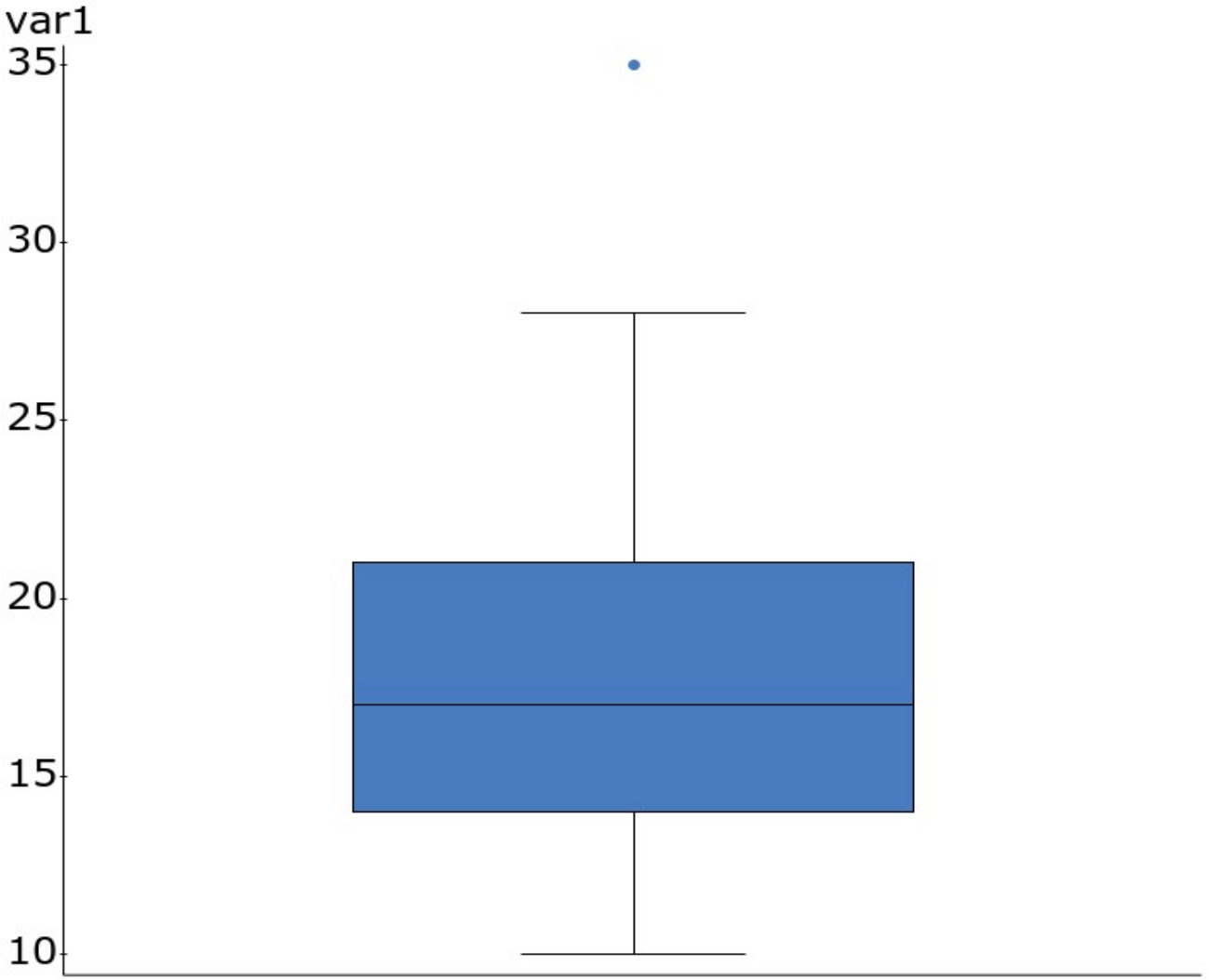
Boxplot (p.89)

Numbers from example above.

Box goes down the page, with scale on left.

- centre of box at **median (17)**
- top of box at $\mathbf{Q_3}$ (21)
- bottom of box at $\mathbf{Q_1}$ (14)
- calculate R=1.5 x IQR: 1.5(21-14)=10.5
  - *upper fence* at Q_3+R 21+10.5=31.5
  - *lower fence* at Q_1-R 14-10.5=3.5
- draw lines ("whiskers") connecting box to most extreme value within fences
- plot values outside fences individually. These are suspected outliers and deserve to be investigated.

StatCrunch boxplot (select “use fences to identify outliers”):



In boxplot above, why do whiskers go down to 10 and up to 28? Investigate below.

- What is lower fence?
  - 3.5
- What is smallest data value? Is it bigger (less extreme) than lower fence?
  - 10; yes
- How far down should lower whisker go?
  - 10 (it is bigger than the lower fence)
- What is upper fence?
  - 31.5
- What are 3 highest data values?
  - 21 28 35
- What is biggest data value that is *smaller (less extreme) than upper fence*?

- 28
- How far up should upper whisker go?
  - 28
- Are there any outliers?
  - Yes, 35.

Comparing distributions with histograms and boxplots (p.91)

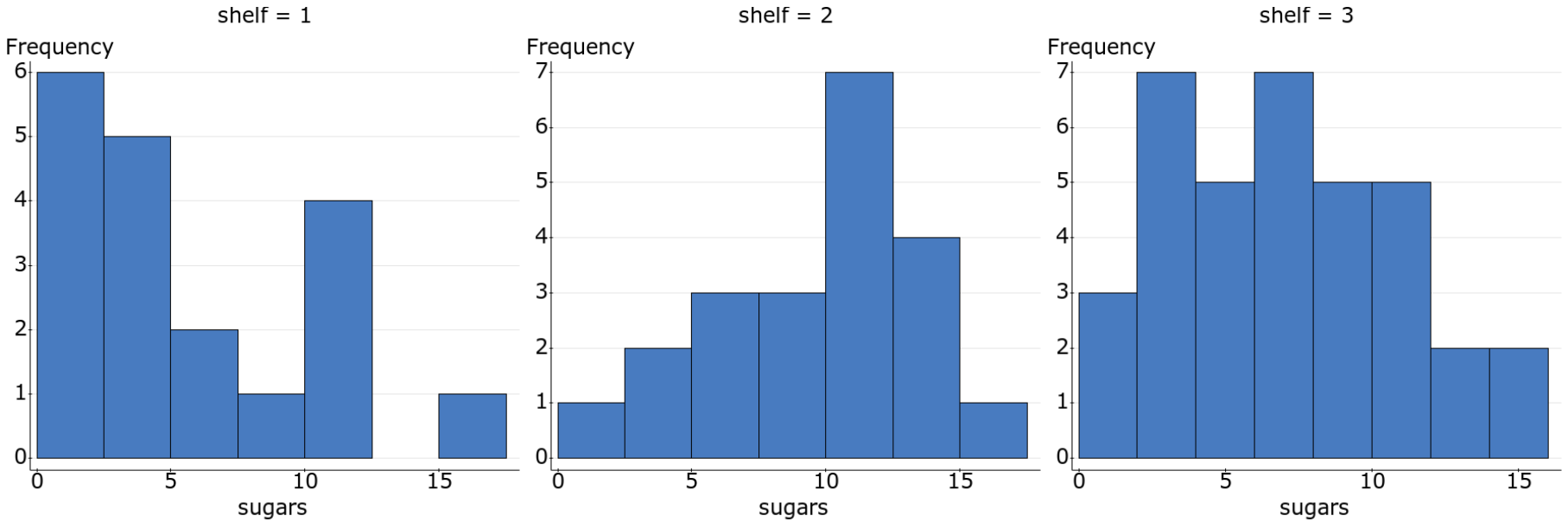
Cereals classified by shelf where found in grocery store:

- 1=top shelf
- 2=middle shelf
- 3=bottom shelf

Want to compare sugar/serving for shelves.

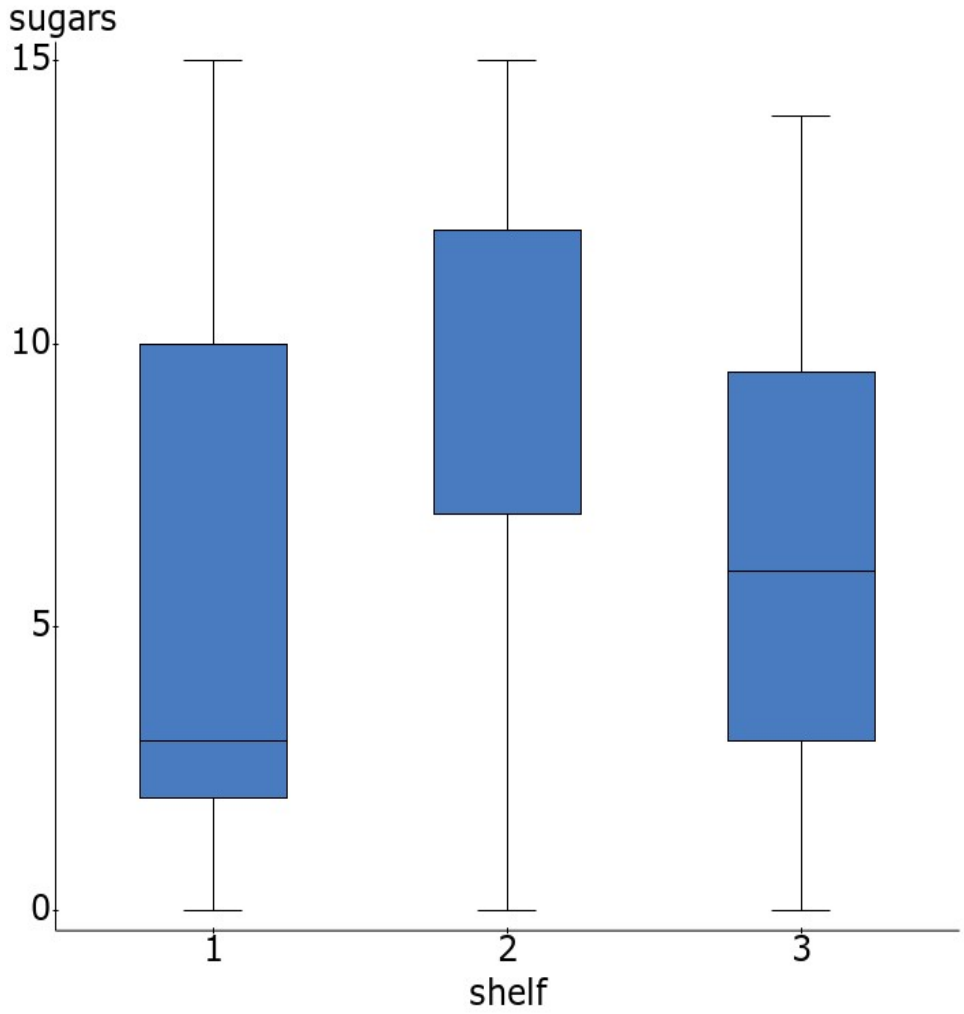
How about a histogram for each shelf, put results side by side?

(*Histogram of sugars, Group By shelf, at bottom Columns per page = 3)*



- so where are the most sugary cereals?
- maybe on shelf 2? Hard to decide.
- how about side-by-side *boxplots*? (Boxplot of sugars, group by shelf. Don't forget Use Fences to Identify Outliers! No need for Columns per Page this time: only one plot.)

Boxplots



Median definitely highest for shelf 2, lowest for shelf 1.

- Easier to see than on histograms.
- Bonus: shelf 1 sugar right-skewed, shelf 2 sugar left-skewed.
  - *shelf 1 boxplot has longer whisker above,*
  - *shelf 2 boxplot has longer whisker below.*
- where did median go for shelf 2 sugar? (see over)

… and why?

- Look at stemplot for shelf 2 sugar
- a lot of the cereals had sugars exactly 12.
- so Q3 and median for shelf 2 sugars are the same.
- 21 cereals: median 11th largest, Q3 6th largest.
- do the means tell the same story as the medians?
  - Yes; largest with largest, smallest with smallest
- does the skewness show up here as well?

Variable: sugars

Decimal point is 1 digit(s) to the right of the colon.

```
0 : 033
0 : 567999
1 : 12222223334
1 : 5
```

Summary statistics for sugars:
Group by: shelf

shelf	Q1	Median	Q3
1	2	3	10
2	7	12	12
3	3	6	9.5

Summary statistics for sugars:
Group by: shelf

shelf	Median	Mean
1	3	5.1052632
2	12	9.6190476
3	6	6.5277778

- Shelf 2 mean<median; shelf 1 mean>median, so yes.

Data set “audio” contains lengths (seconds) of audio files sampled from an iPod. Obtain a histogram and a boxplot of the track lengths.

- There are at least 2 (maybe 3) outliers. Are they reasonable track lengths?
  - Yes, eg, for classical music or “concept album”

Which summary of the track lengths do you prefer:

Summary statistics:

Column	Mean	Std. dev.
file length	354.1	307.94753

Summary statistics:

Column	Min	Q1	Median	Q3	Max
file length	46	188	267.5	398	1847

Why?

- 5-number summary (distribution skewed to right)

Chapter 6: The standard deviation as a ruler and the normal model (p. 121)

Which is the better exam score?

- 67 on exam A with mean 50 and SD 10
- 62 on exam B with mean 40 and SD 12?

What do you say to these:

- 67 is better because 67 > 62?
  - no, because mean is higher too.
- 62 is better because it is 22 marks above the mean and 67 is only 17 marks above the mean?
  - No, ignores different spreads.
- Or....?

Key: *z-scores*.

You look at StatCrunch report "location and spread under linear transformation":

http://www.statcrunch.com/5.0/viewreport.php?reportid=25026

Or click Explore, Reports, type title into box.

Summary:

- if you multiply/divide all data values by a constant, all measures of centre and spread multiplied/divided by that constant.
- if you add/subtract constant to all data values, measures of centre add/subtract that constant, but measures of spread *unchanged.*

When you calculate a z-score as

$$z = \frac{x - mean}{SD}$$

using the mean and SD **of x**, what are the mean and SD **of z**?

- First off, suppose x has mean 10 and SD 3.
- Then x-10 has mean 10-10=0 and SD 3.
- and z=(x-10)/3 has mean 0/3=0 and SD 3/3=1.
- *this actually works no matter what mean and SD x has.*
- Try it with x having mean -5 and SD 10, say.

No matter what mean and SD x has, z has mean 0, SD 1.

- Calculating a z-score sometimes called "standardizing". Above says why.

– Gives a basis for comparison for things with different means and sds.

Those exam scores above:

Which is the better exam score?

- 67 on an exam with mean 50 and SD 10
- 62 on an exam with mean 40 and SD 12?

Turn them into z-scores:

- 67 becomes (67-50)/10=1.70
- 62 becomes (62-40)/12=1.83

so the 62 is a (slightly) better performance, relative to the mean and SD.

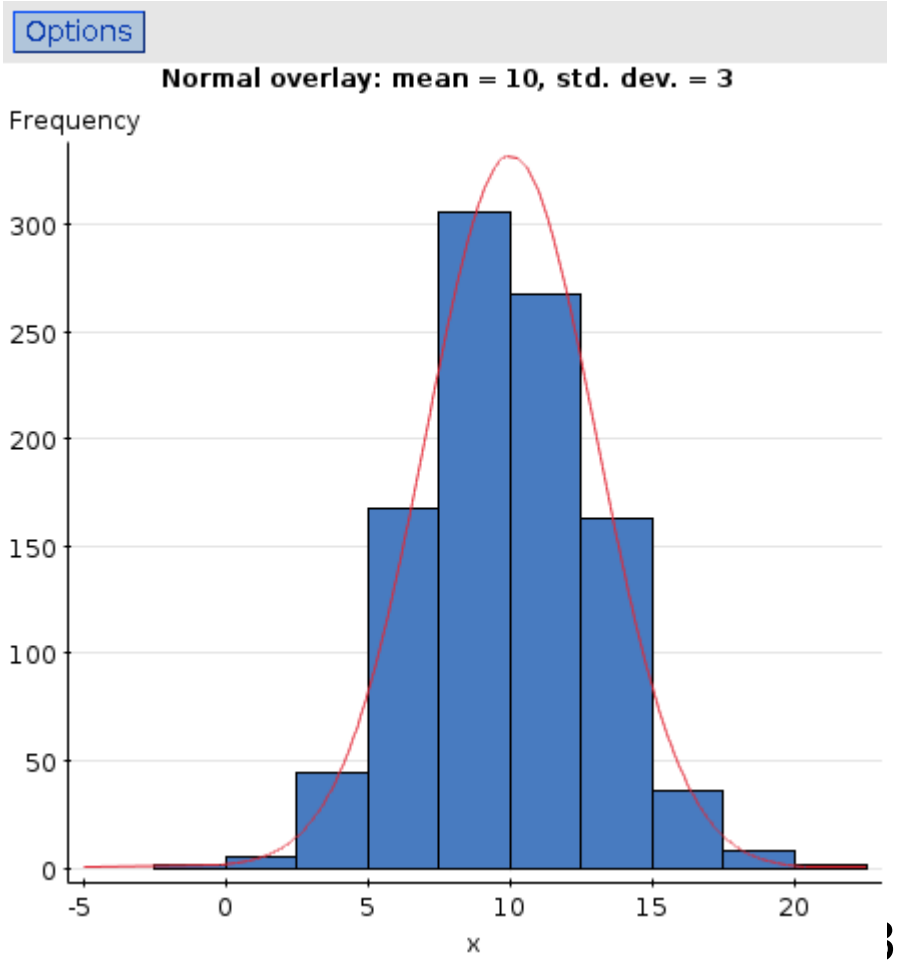
Density curves and the normal model (p.129)

How big might a z-score typically be?

To answer that, need *mathematical model* to describe what's going on.

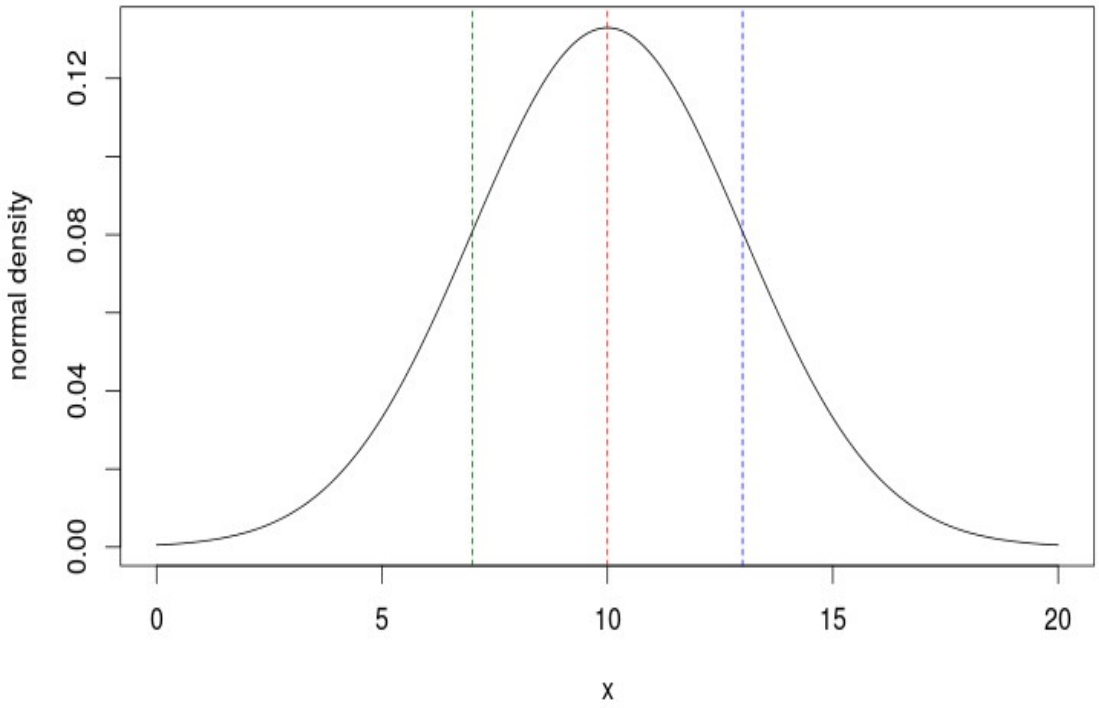
Here's one: often run into data with symmetric distribution and no outliers, like this:

Red curve is *normal distribution model.* Not a perfect match, but pretty close.



Mean and standard deviation on a normal distribution

- Mean (and median) at peak (10)
- for SD: look at where density function stops curving down and starts curving out. These are "shoulders": at 7 and 13.
- Distance from mean to a shoulder is the SD: 13-10=10-7=3.
- So mean is 10 and SD is 3.



Z values and Table Z (p.135)

How much of a normal distribution is less than a value, more than a value, between two values?

Use Table Z, pages 1047-8 in text:

- first calculate z
- then look up z in table, which gives you fraction *less.*
- Area under whole normal curve is 1.
- Sometimes easier to figure what you *don't* want.

Roma tomatoes have weights that have a normal distribution shape with mean 74 grams and SD 2.5 grams. What proportion of these tomatoes will weigh less than 70 grams?

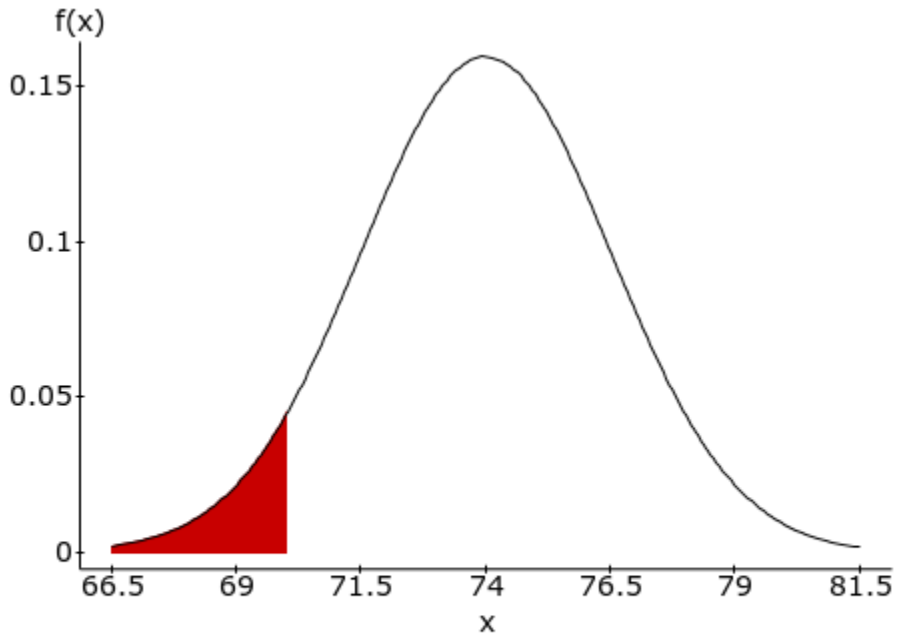
Z=(70-74)/2.5=-1.6_____;

look up in table Z to get 0.0548_____

StatCrunch equivalent on right (no need to standardize).

You should learn table and StatCrunch.

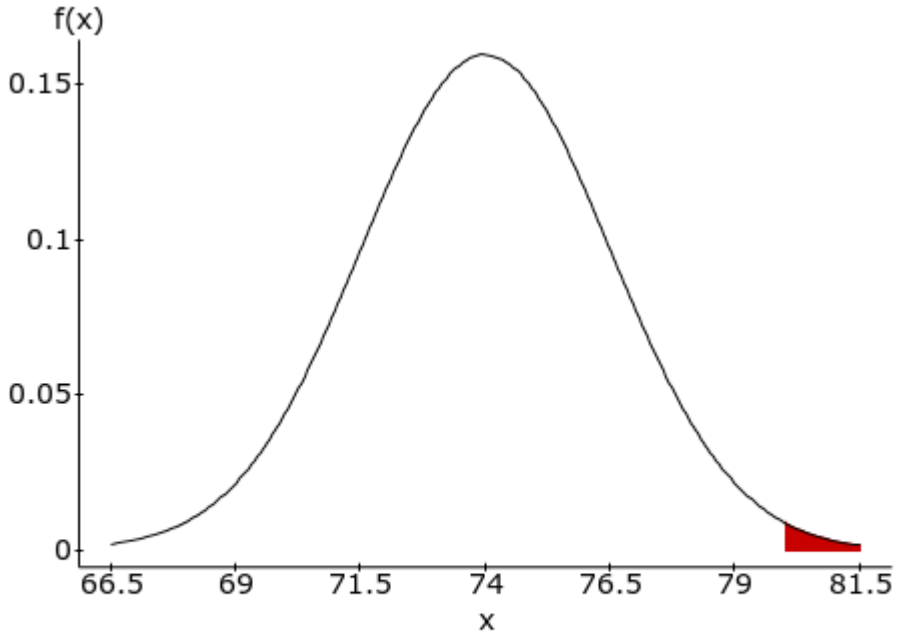
(On exam we give you Table Z.)



What proportion of the Roma tomatoes in the previous question will weigh more than 80 grams? (Mean 74, SD 2.5.)

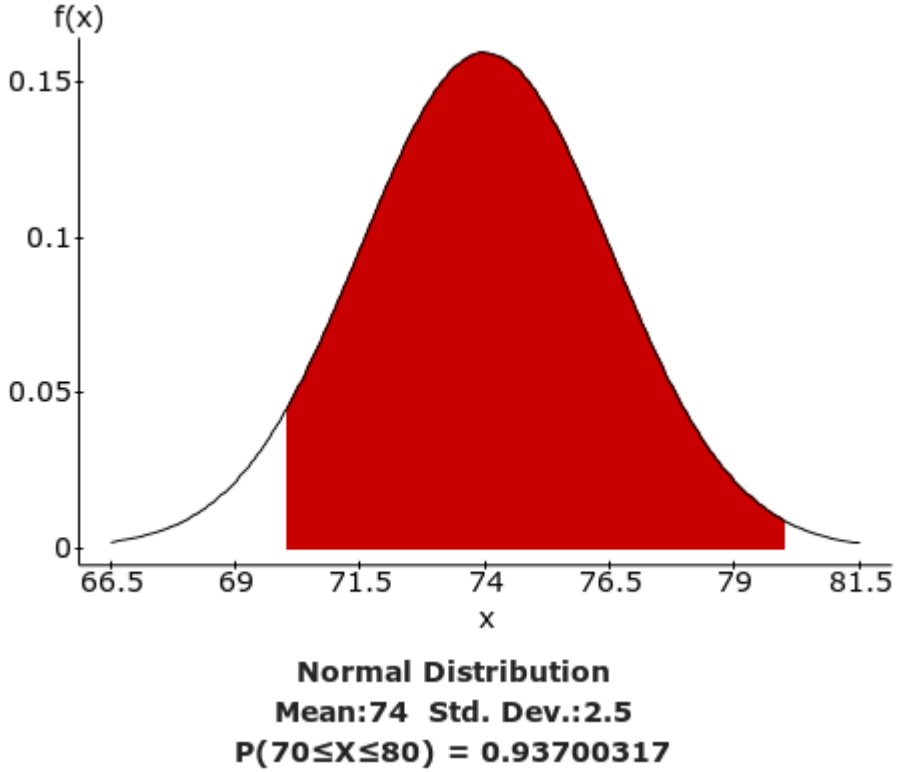
- z=(80-74)/2.5 = _2.4_____
- Table: _0.9918_____
- so 1-0.9918= 0.0082 more.

Draw picture (as at right): answer will be *very small.*



Normal Distribution
Mean:74 Std. Dev.:2.5
P(X≥80) = 0.00819754

What proportion of the Roma tomatoes of the previous two questions will weigh between 70 and 80 grams? (There are two ways to do this, both of which use the previous work.) StatCrunch first (gives nice picture):



Want all but the tiny white bits at the ends.

Way 1 (easier to understand)

- 0.0548 less than 70
- 0.0082 more than 80
- everything else between: 1-0.0548-0.0082=0.9370.

Way 2 (easier to do)

- 70 as z-score is -1.60, table gives 0.0548.
- 80 as z-score is 2.40, table gives 0.9918.
- Subtract: 0.9918-0.0548=0.9370.

These both check with StatCrunch.

What if z has 2 (or more) decimal places?

Example using Roma tomatoes again (mean 74, SD 2.5): proportion less than 77.4 grams?

- Z=(77.4-74)/2.5=1.36_____
- use *column* of table Z according to 2nd place, so z=1.36 gives _0.9131_____ (answer).
- If z has more than 2 decimals, round to 2, then use table.

Getting values from proportions (p.139)

- Use Table Z backwards to get z that goes with proportion less
- Turn z back into original scale.
- How? $z=(x-mean)/SD$, solve for x
- Gives $x=mean+(SD*z)$
- Examples coming up.

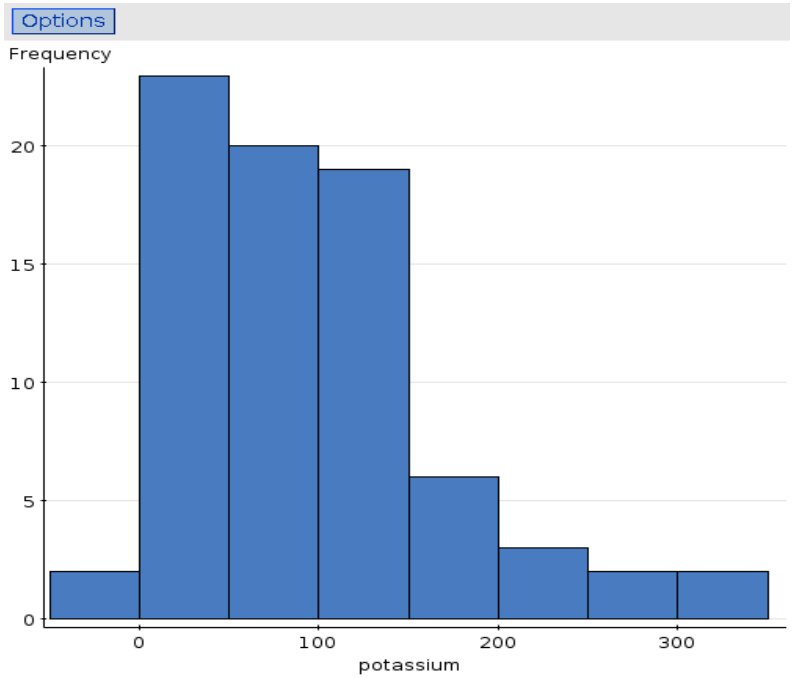
At-term newborn babies in Canada have weights that follow a normal distribution, with mean 3500 grams and SD 500 grams.

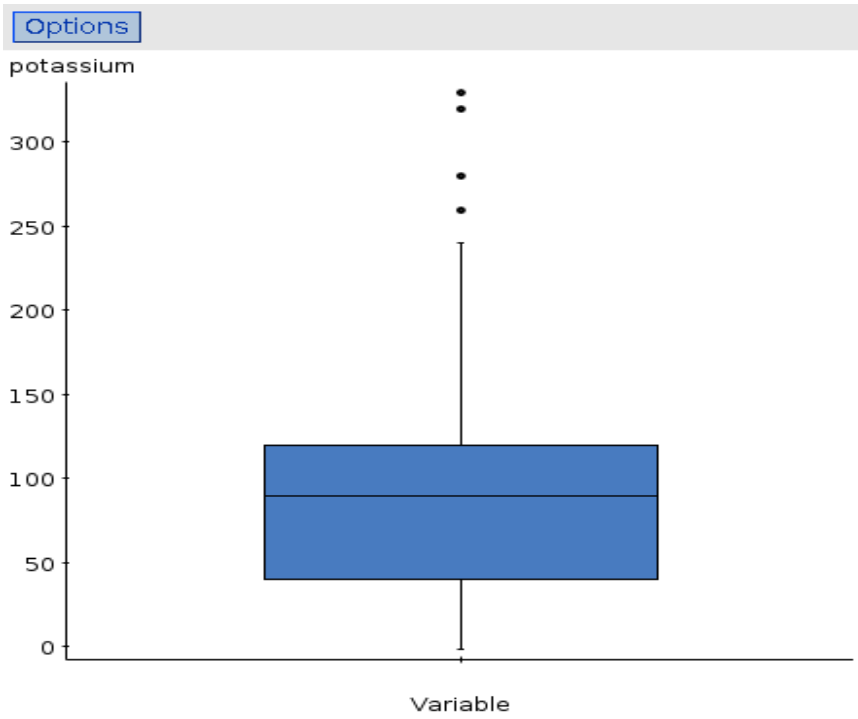
- A baby is defined as being "high birth weight" if it is in the top 2% of birth weights. What weight would make a baby "high birth weight"?
  - 2% more = 98% less = 0.9800 less
  - z = _2.05_____
  - weight = 3500_ + _2.05____*_500_= 4525_____ **grams** (or more)
- A baby is defined as being "very low birth weight" if it is in the bottom 0.1% of birth weights. What weight would make a baby "very low birth weight"?
  - 0.1% less = 0.0010 less
  - z = -3.09 (I picked the middle one, but any is good)
  - weight = 3500 + (-3.09)*500 = 1955_ **grams** (or less)

Is normal distribution is a good fit to data? (p.137)

Return to the cereal potassium data.

Histogram and boxplot:

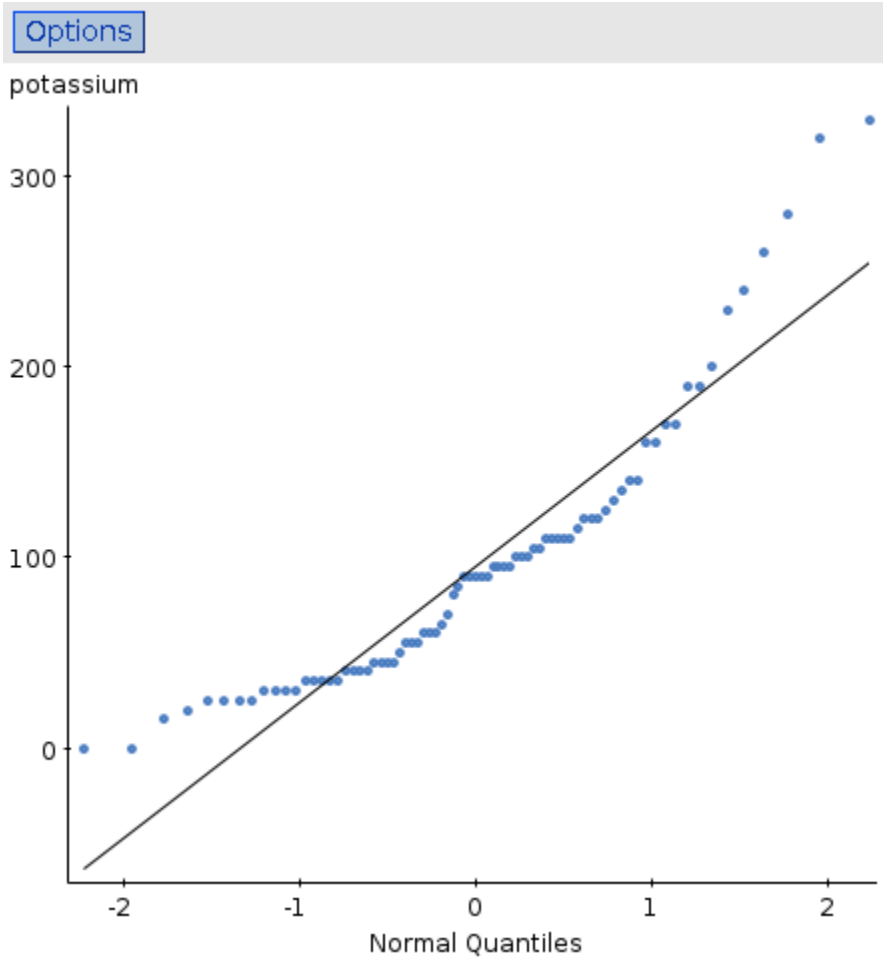




– distribution skewed right.

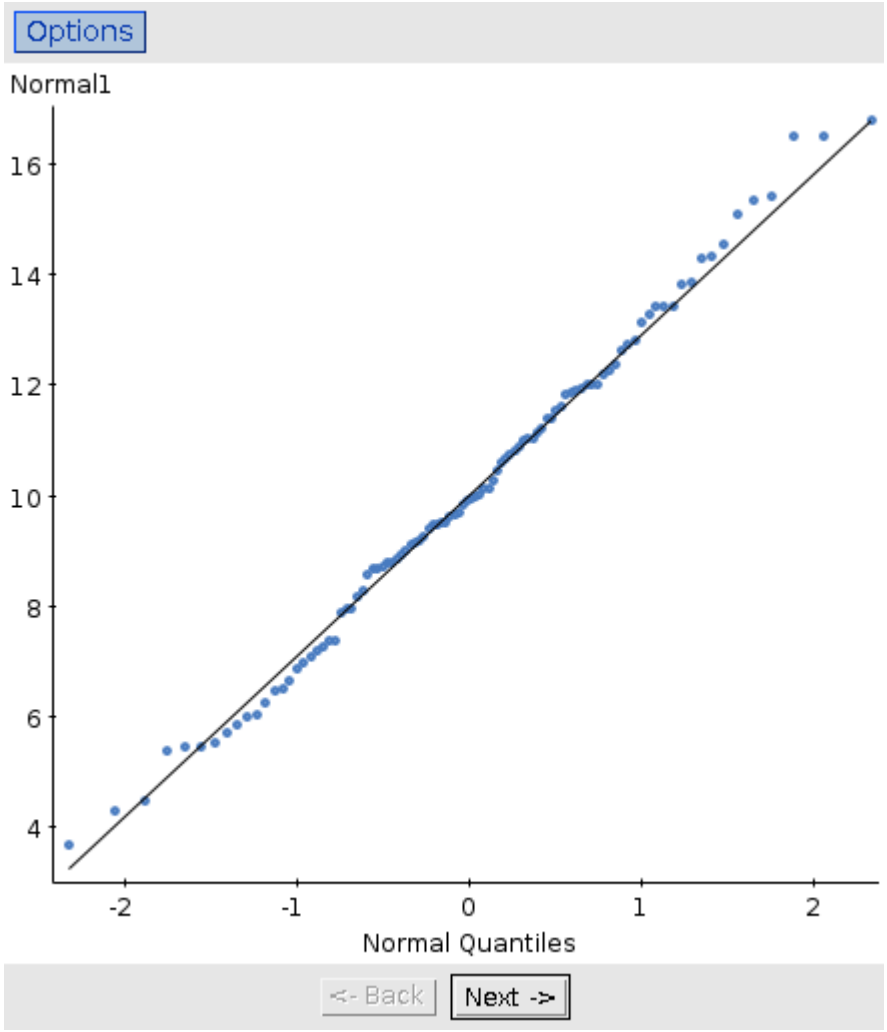
– look also at *normal probability plot* (QQ plot):

- if normal distribution OK, blue dots more or less follow central line (straight)
- Curve or other systematic deviation from line = not normal
- Here, not normal.
- low values too bunched together, high values too spread out: skewed to right.



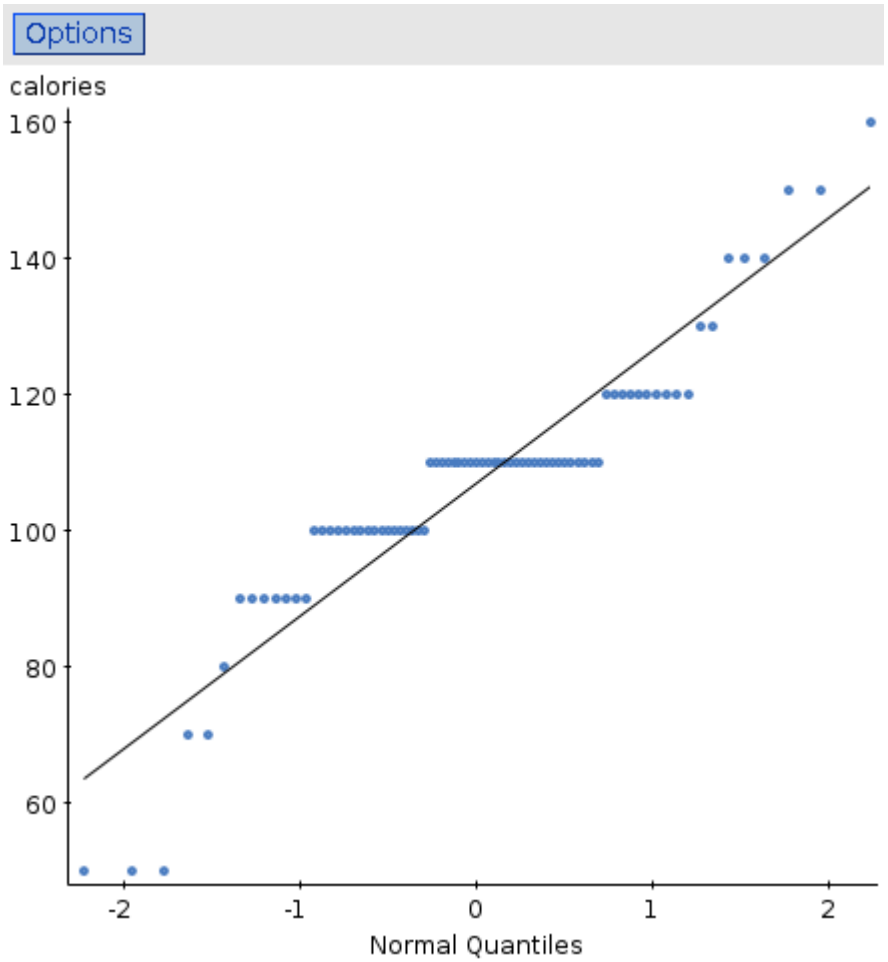
Actual normal data:

- not perfectly straight
- but no obvious outliers or curve
- normal distribution ok for these data.



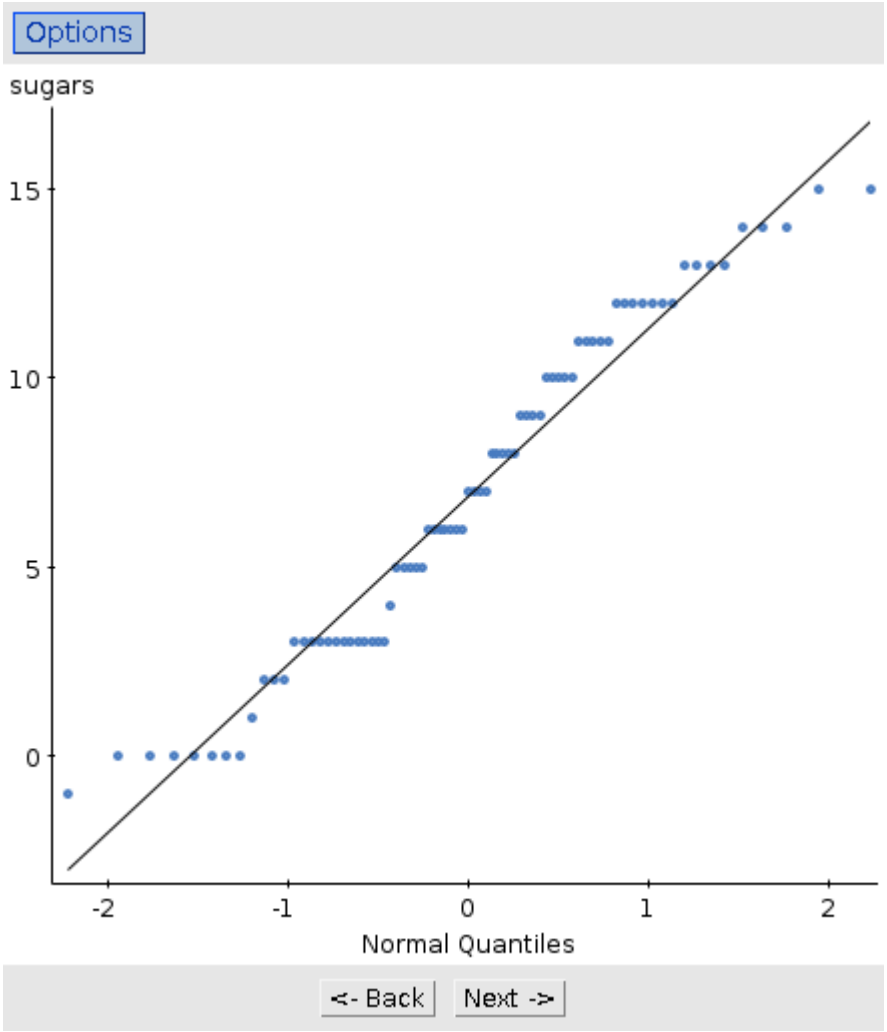
Cereal calorie data:

- horizontal blue dots: calories only measured to nearest 10
- high values maybe a little too high
- low values too low
- symmetric, but too many outliers for normal.



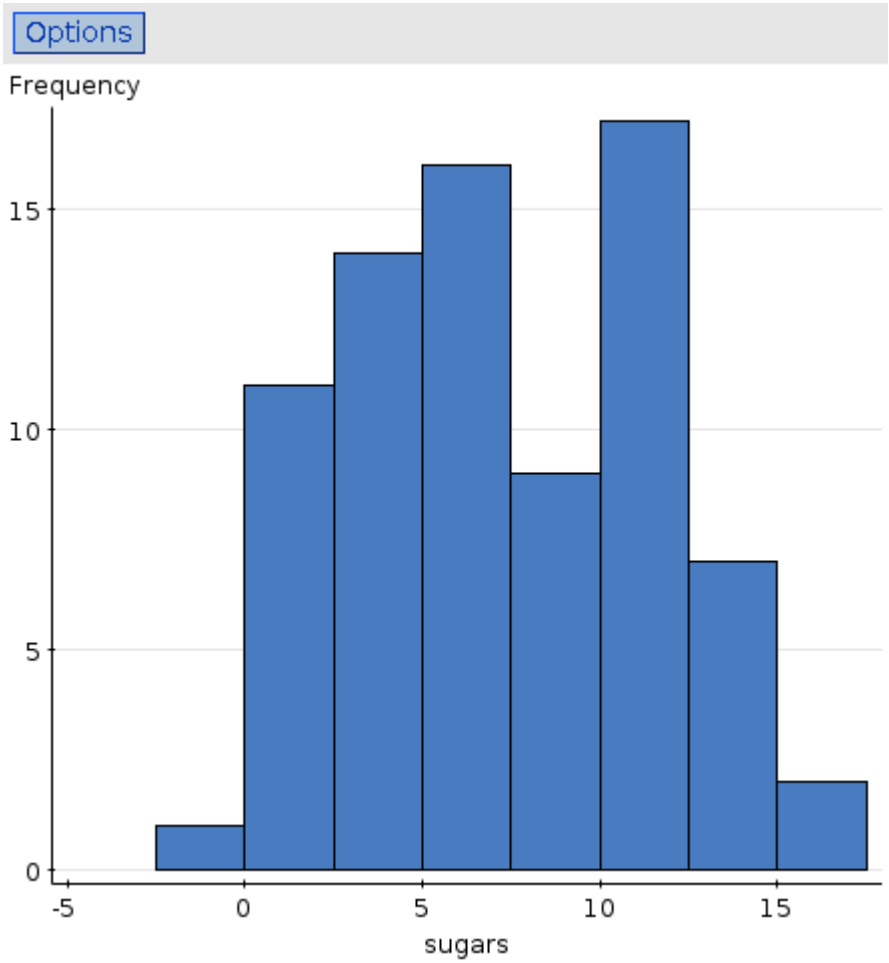
Cereal sugars:

- a lot of identical values (like calories)
- otherwise reasonably close to line
- normal probably not too bad.



Cereal sugars histogram:

- histogram has a “hole” between 7.5 and 10
- otherwise, not too far from normal.



68-95-99.7 rule (p. 133)

Sometimes can get a rough idea of normal proportions like this.

- about 68% of a normal distribution between mean +/- SD
- about 95% of a normal distribution between mean +/- 2 SD
- about 99.7% of a normal distribution between mean +/- 3 SD

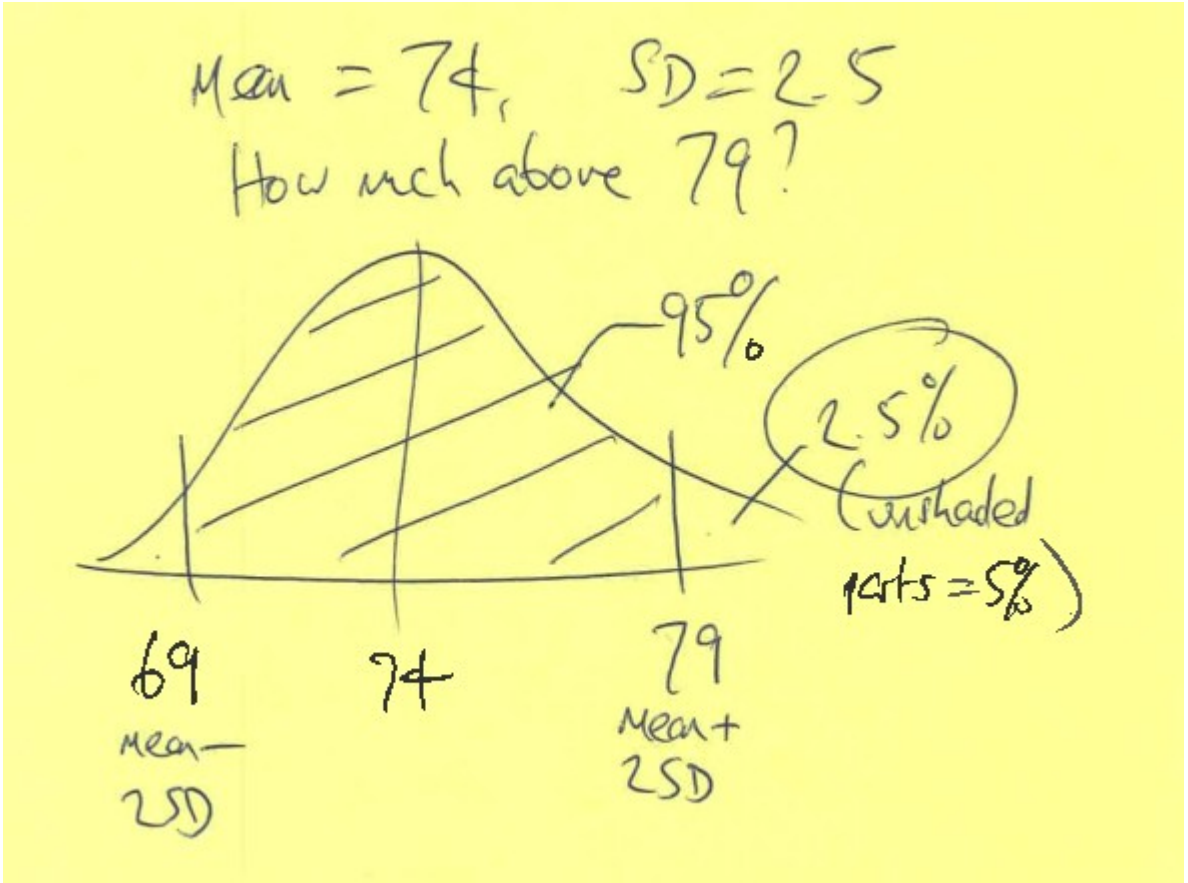
Recall weight of Roma tomatoes: mean 74, SD 2.5 (grams)

- what weights will about 95% of them be between?
  - *74-2*2.5=69 to 74+2*2.5=79*
- about what fraction of the weights will be between 71.5 and 76.5 grams?
  - *71.5 is mean minus ___1_ x SD, 76.5 is mean plus _1___ x SD, so answer is _68%_____.*

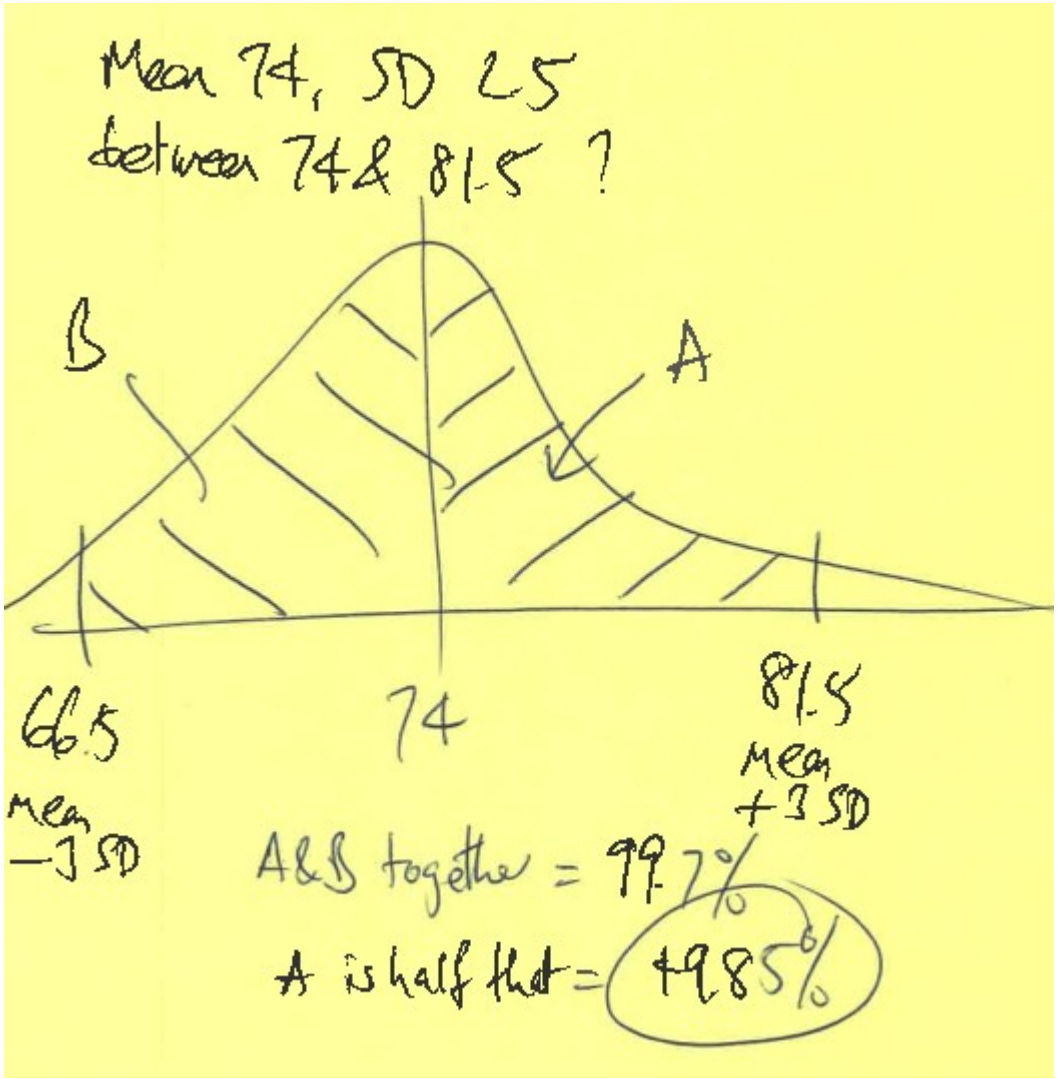
Variations

Again using mean 74, SD 2.5:

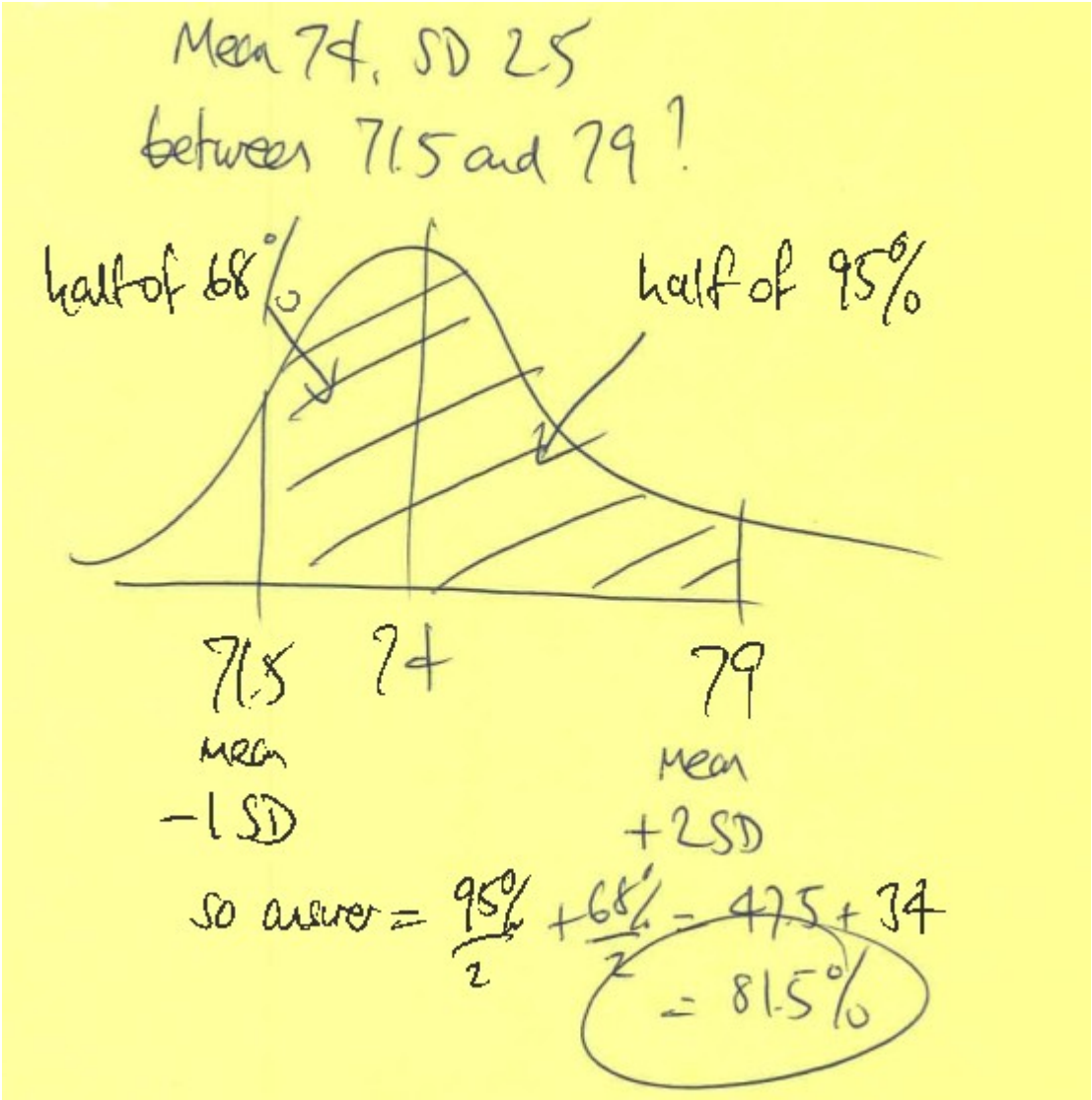
- about what fraction of weights will be more than 79 grams?
  - *95% between 69 and 79*
  - *5% beyond that*
  - *2.5% above 79*



- about what fraction will be between 74 and 81.5 grams?
  - *81.5 is __3_ Sds above mean, 66.5 is _3__ Sds below*
  - *Between 66.5 and 81.5: 99.7%*
  - *How much of that between 74 and 81.5?*
    - *half*
  - *How much of all tomatoes between 74 and 81.5? 49.85%*



- about what fraction will be between 71.5 and 79 grams?
  - *How much between 71.5 and mean? <u>Half of 68%: 34%</u>*
  - *How much between mean and 79? <u>half of 95%: 47.5%</u>*
  - *How much between 71.5 and 79?*
  - *<u>34+47.5=81.5%</u>*



When you don't have a normal table (skip)

Proportion of standard normal distribution less than z is approximately $0.5+z(4.4-z)/10$, 0.99 for 2.2<z<2.6, 1 beyond.

(this for z positive – if z<0, draw a picture and flip it around).

Roma tomatoes: mean 74, SD 2.5; proportion less than 77.4 gives z=1.36. Proportion less approximately

_____

Compare correct answer 0.9131. (Usually accurate to 2 decimals.)

Proportion less than 70: gives z=-1.60. Draw picture. Same as proportion more than 1.60. Proportion *less* than 1.60 approx:

_____

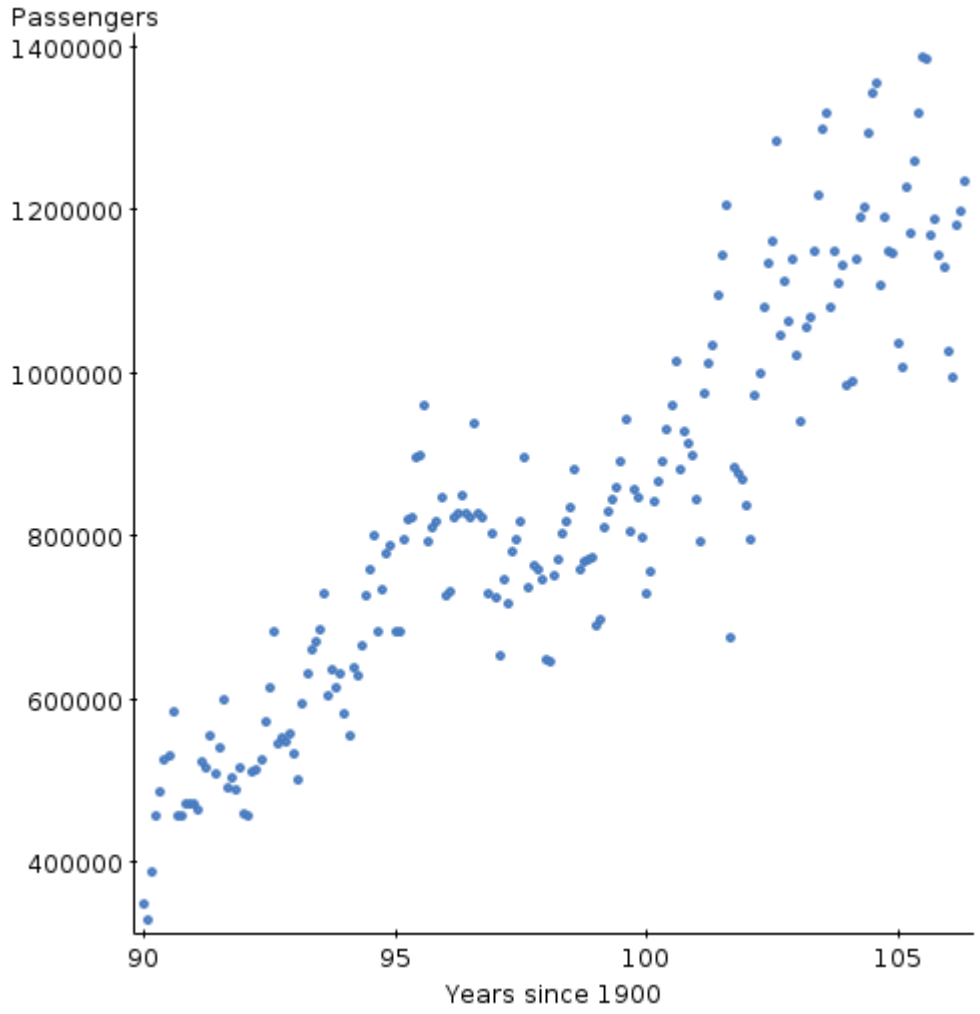
Proportion *more* than 1.60 = proportion less than -1.60

= _____ (compare exact 0.0548).

Chapter 7: Scatterplots, association and correlation (p. 168)

- Previously, single variables on their own.
- Or one or more categorical variables.
- Now look at **two quantitative variables.**
- First tool: scatterplot.
  - Plot values of two quantitative variables *against each other.*

The airport in Oakland, California recorded the number of passengers departing in each month from 1990 to 2006. Scatterplot of passengers against time:

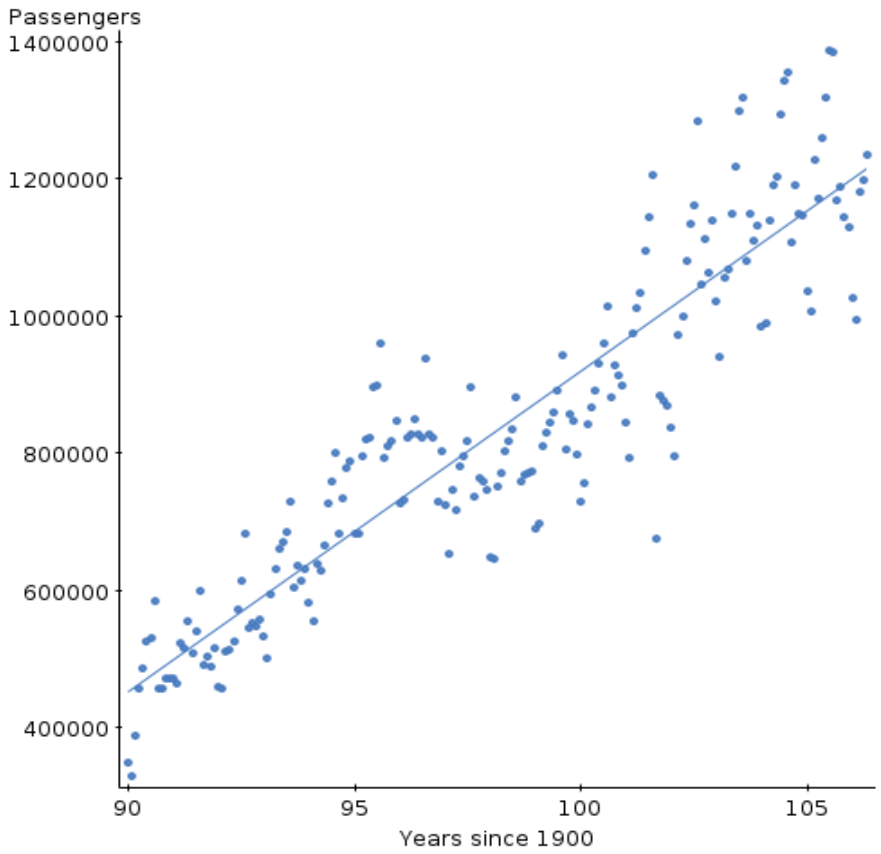


Talking about association (p. 170)

- Direction
  - Upward trend (positive), downward trend (negative)
- Form
  - Straight line, curve
- Strength
  - Strong (clear) relationship, moderate, weak

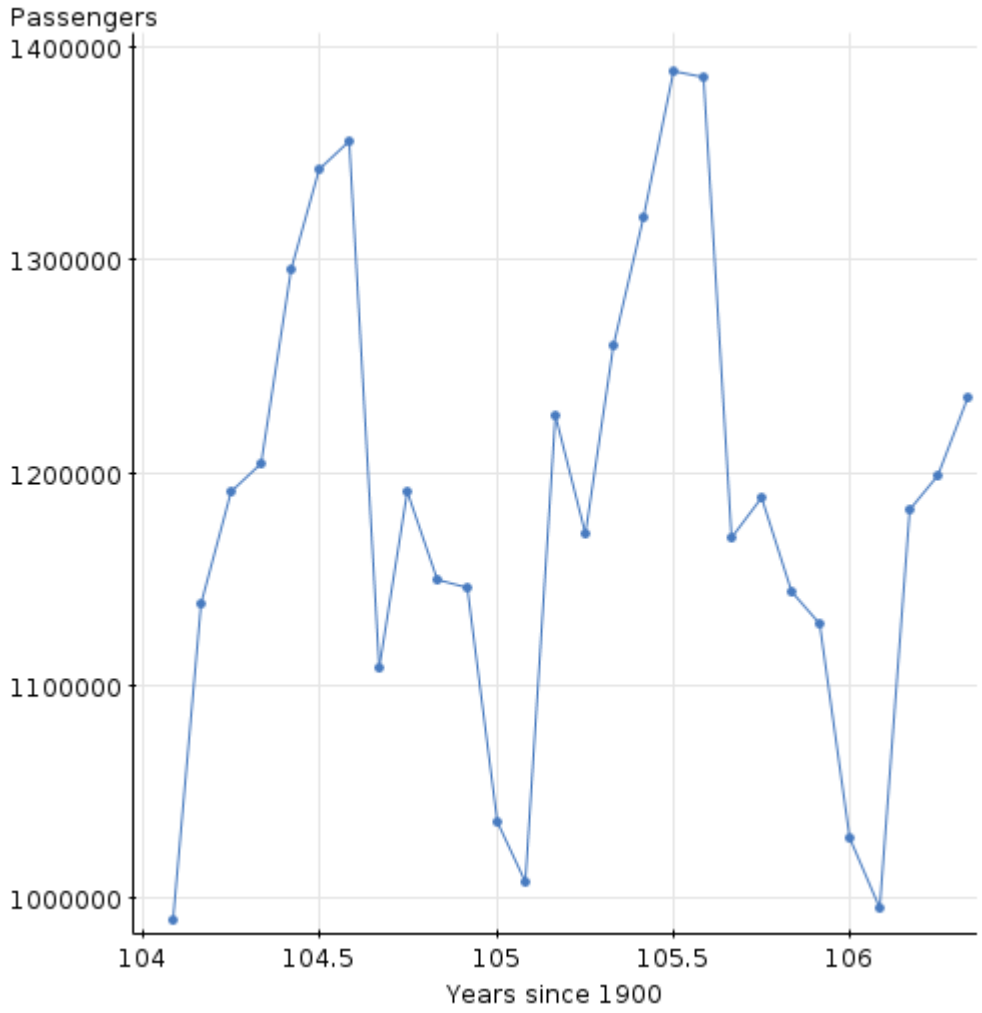
With line:

- _upward_ trend (direction)
- close to _line_ (line?)
- strength? _strong_



Now look at month-by-month for 2004-2006 (months joined by lines):

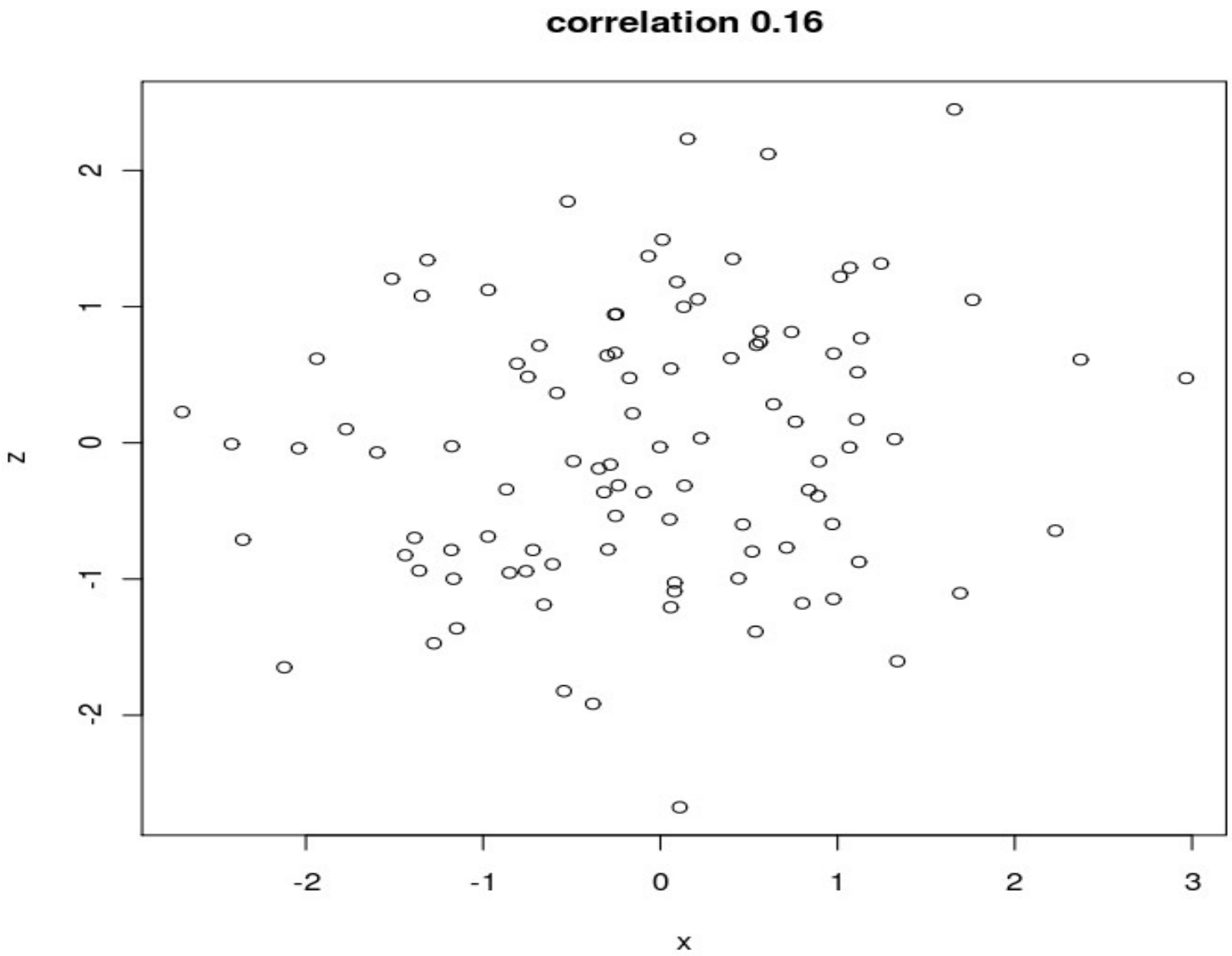
- what is happening? Why?
  - *Seasonal trends within year*
- why didn't we see it before?
  - *Mixed up with upward trend*
  - _____

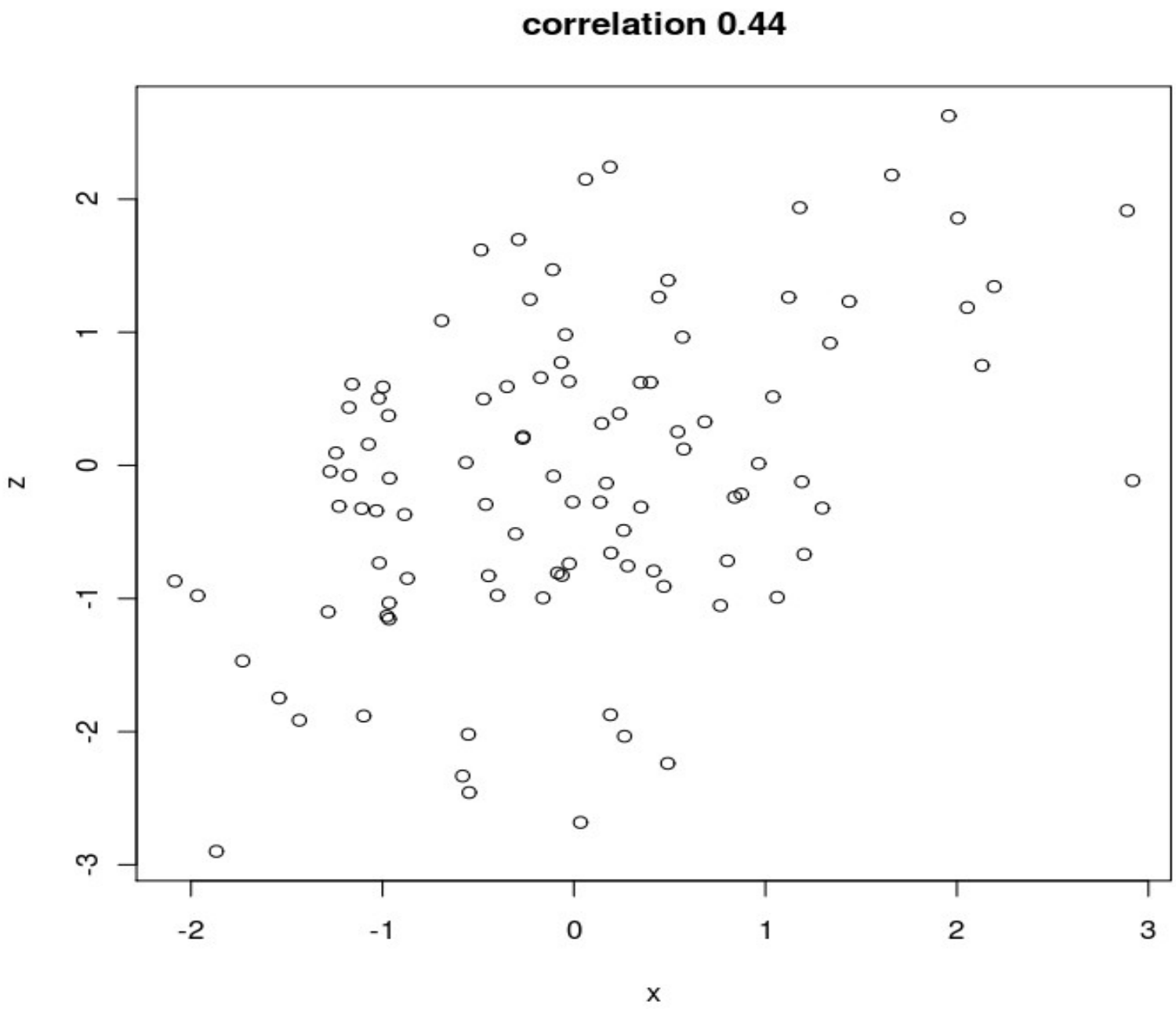


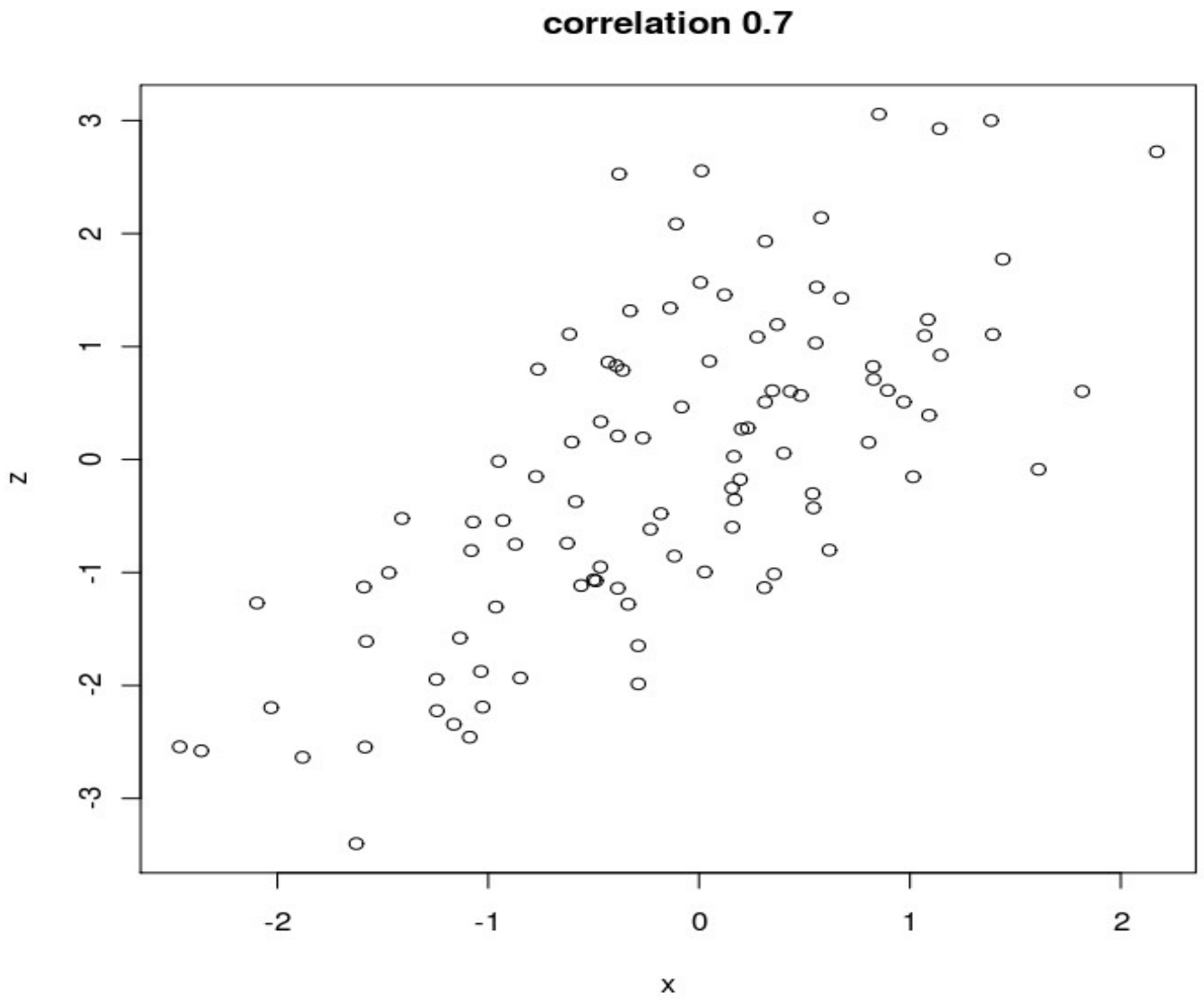
Correlation (p. 173)

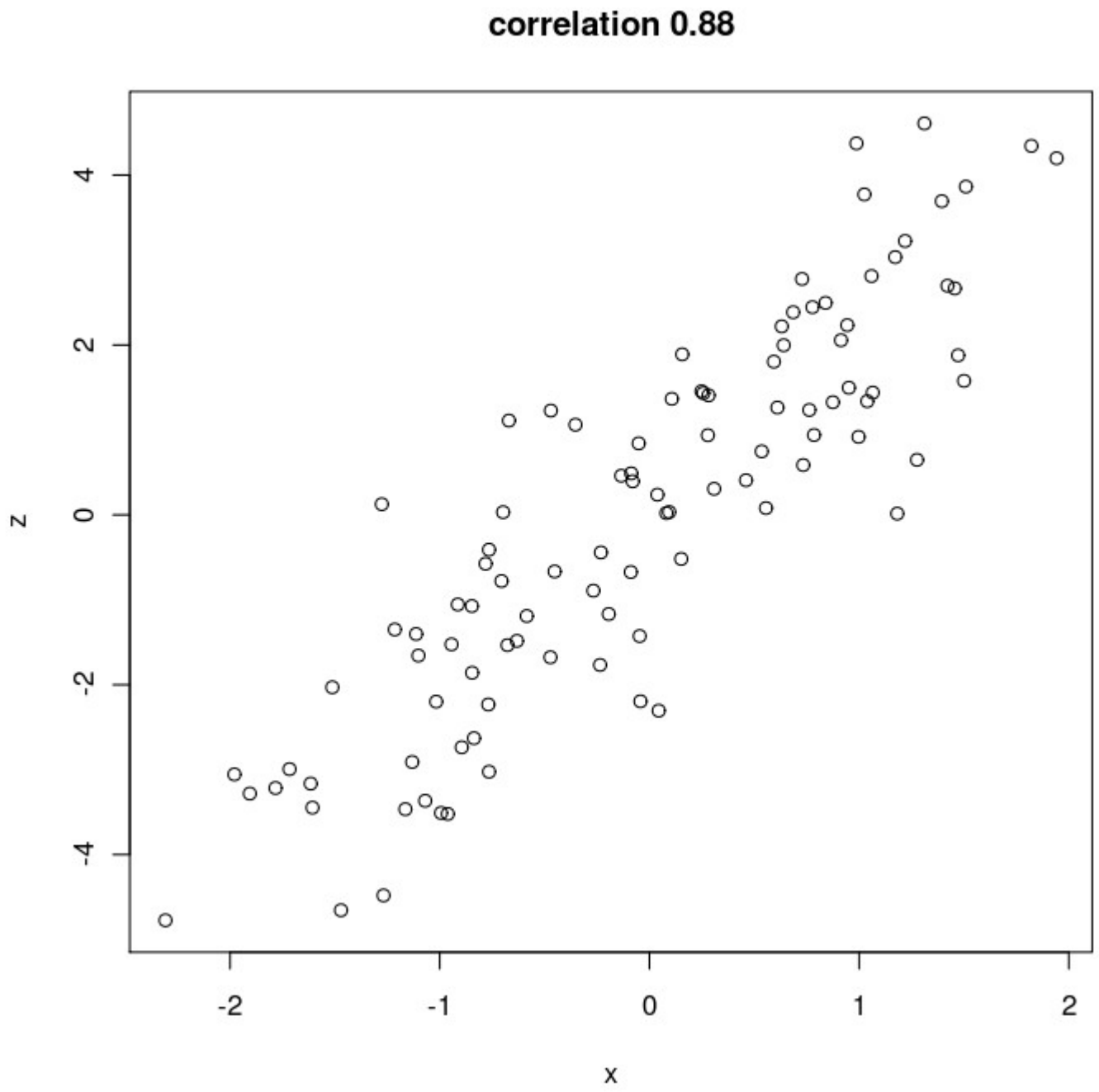
- **If the association is a line**, can calculate a number to describe how near a line the points are: *correlation (coefficient).*
- Number between -1 and 1:
  - 1 means perfect positive (uphill) association
  - 0 means no (linear) association at all
  - -1 means perfect negative (downhill) association
  - in between means in between

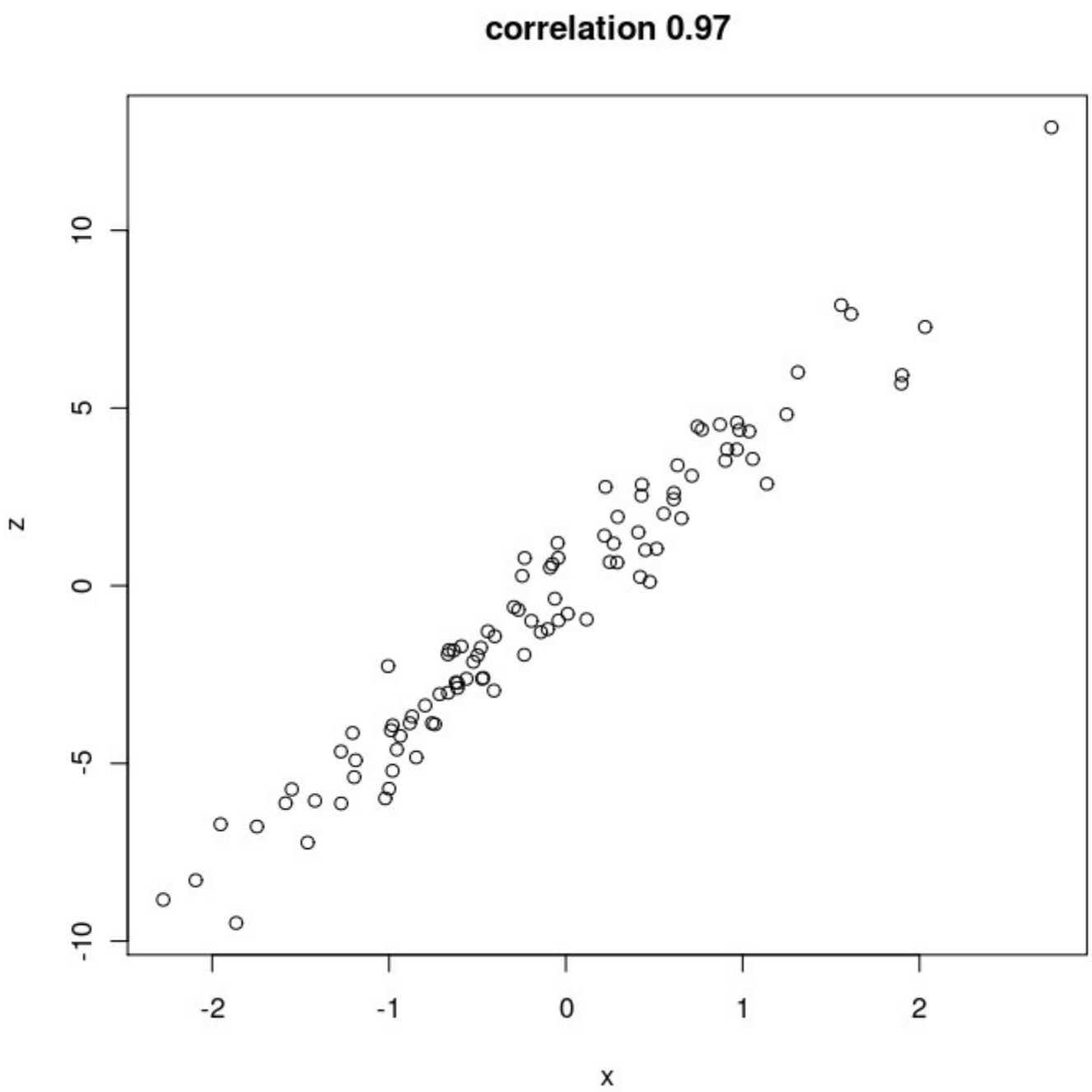
Some correlations

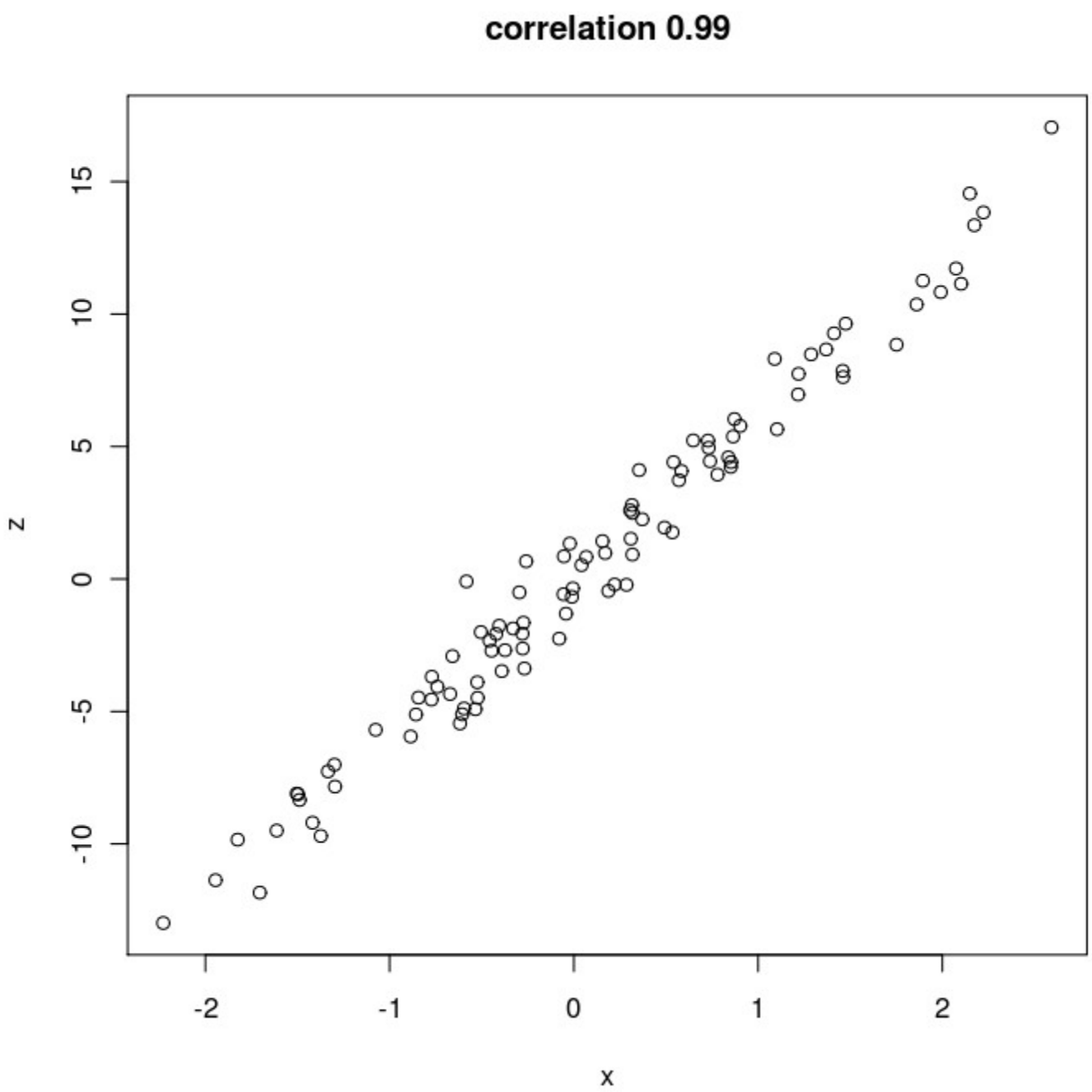


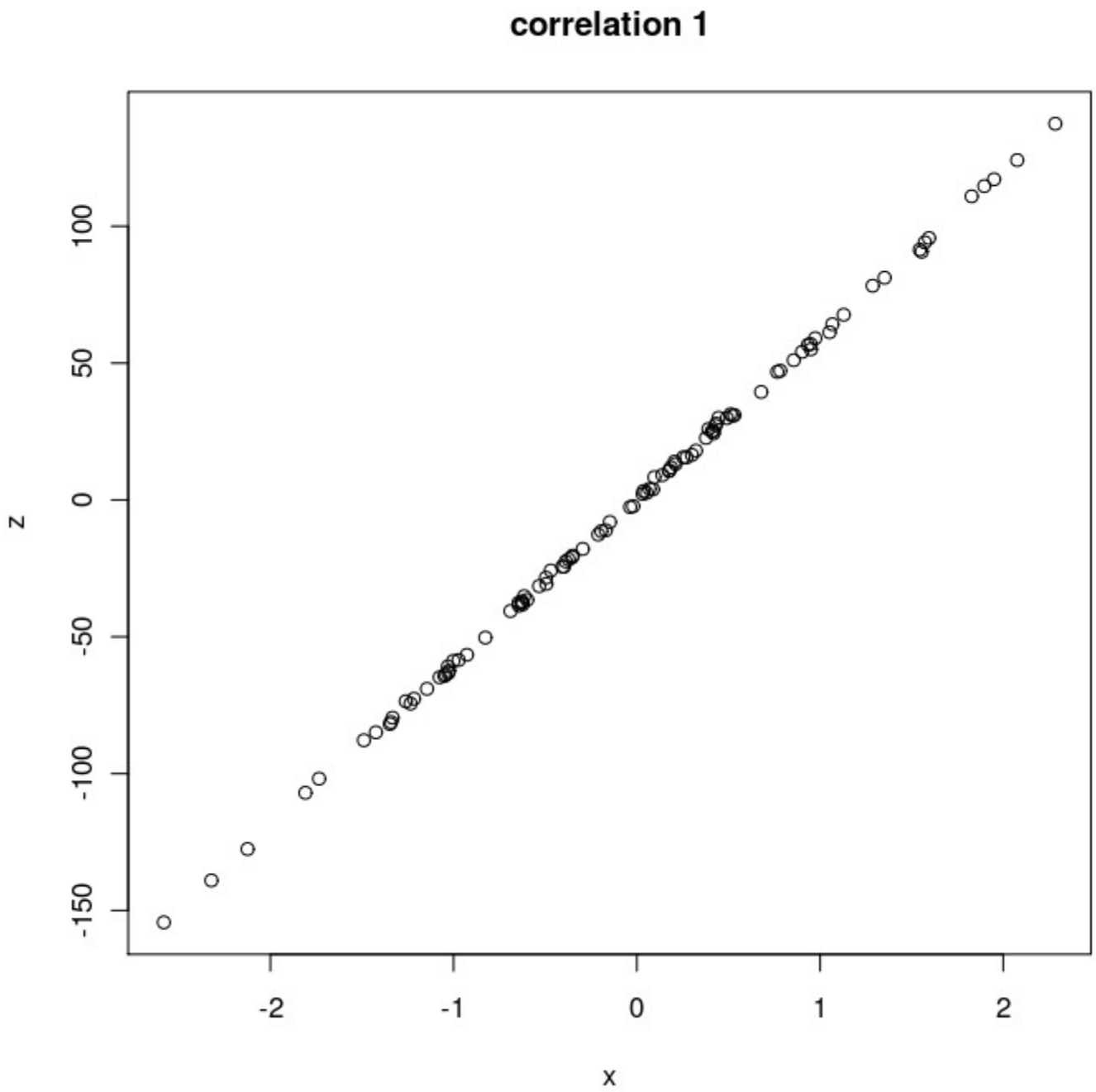

correlation 0.44
z
x
2
1
0
-1
-2
-3
-2
-1
0
1
2
3

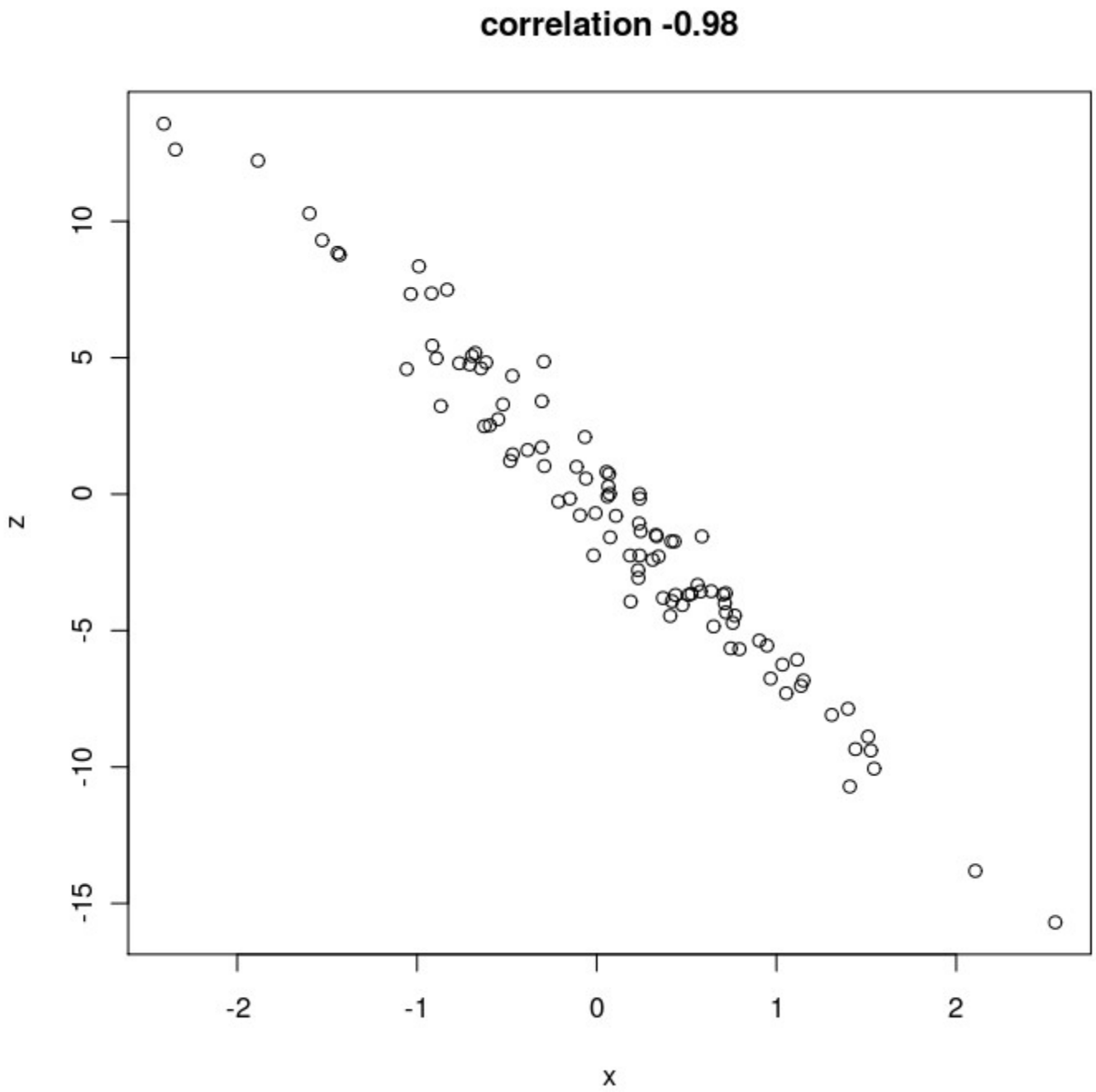

correlation 0.7
z
x
3
2
1
0
-1
-2
-3
-2
-1
0
1
2

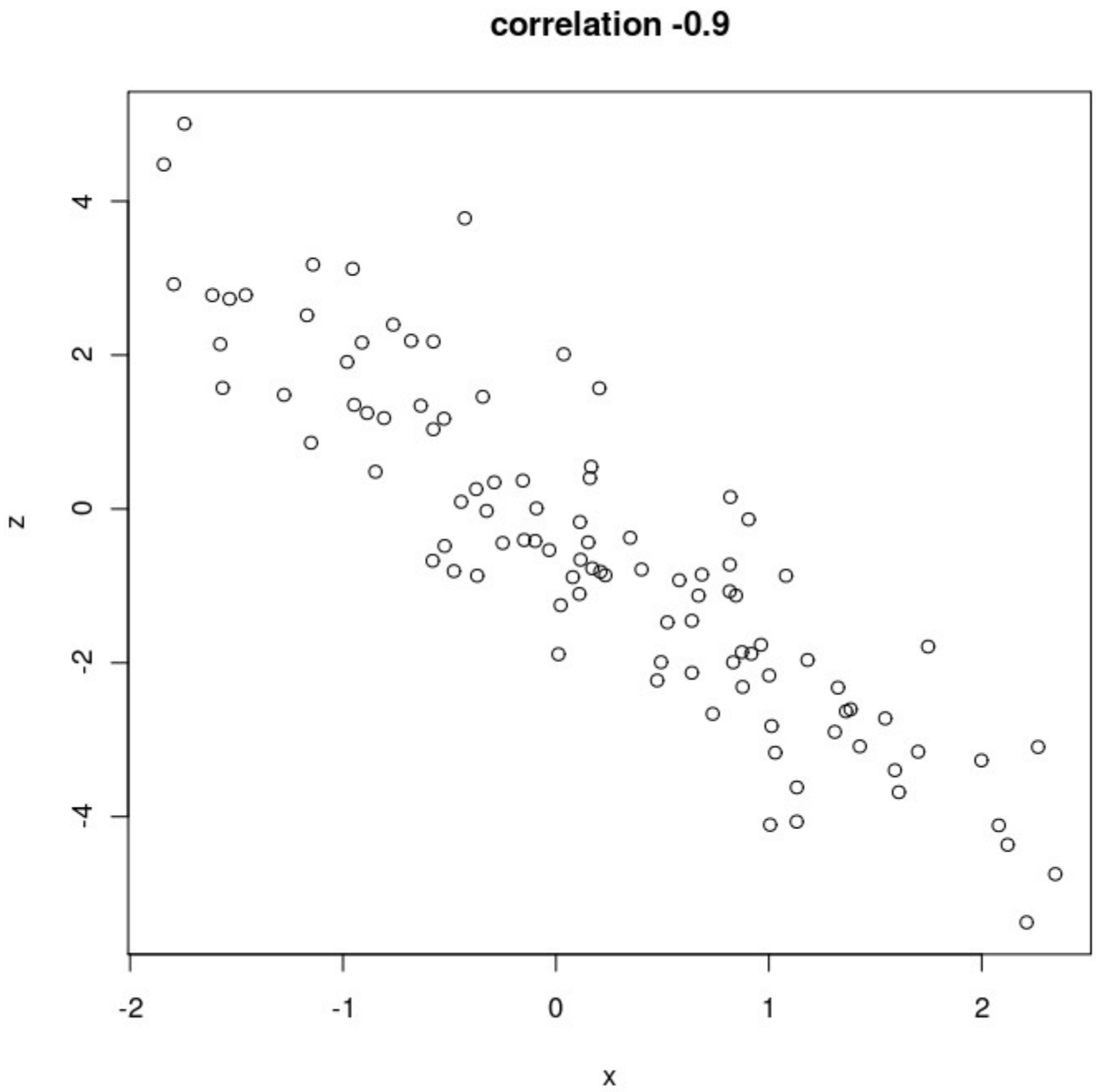

correlation 0.88
z
x
4
2
0
-2
-4
-2
-1
0
1
2

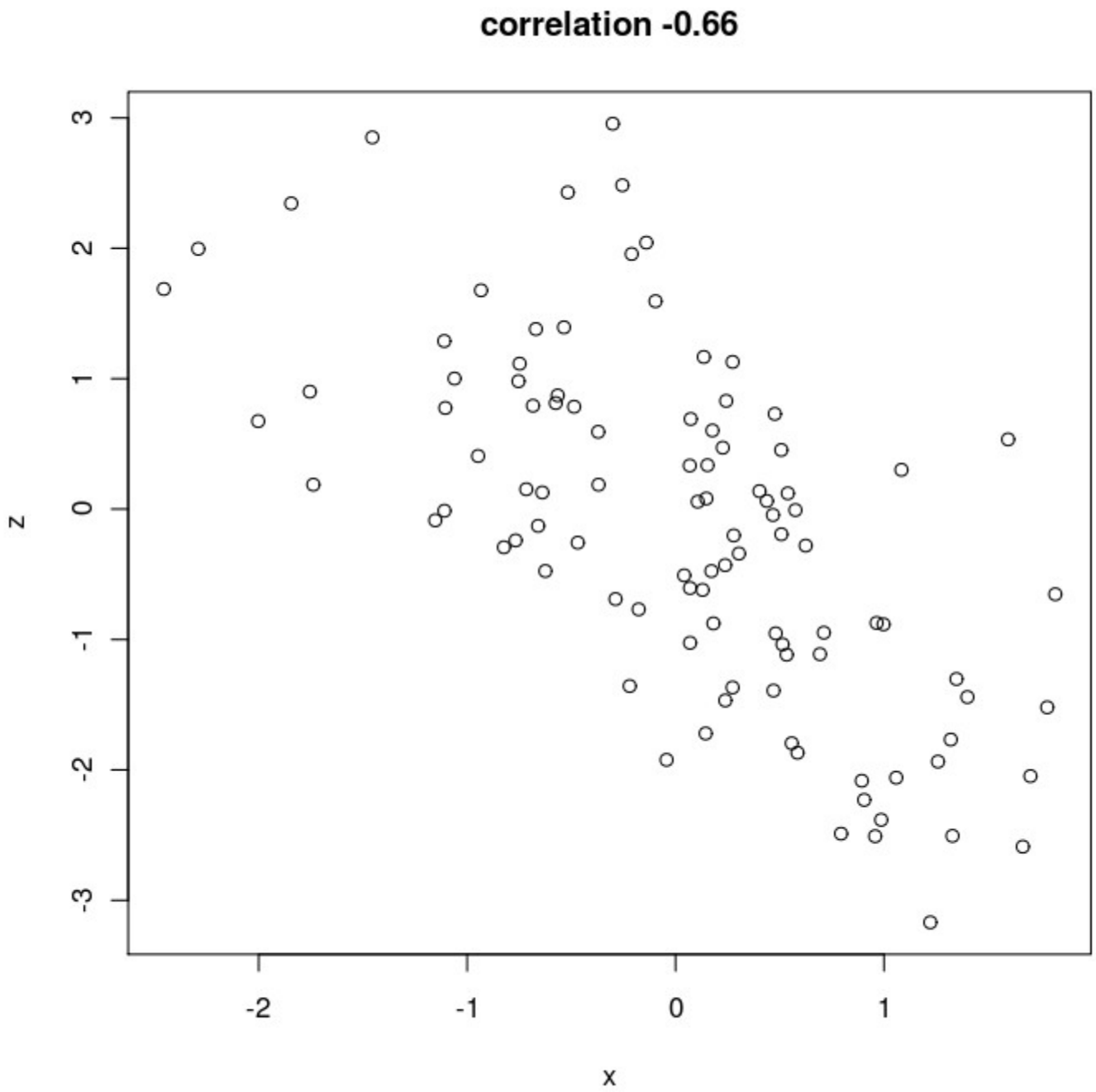

correlation 0.97
z
x
10
5
0
-5
-10
-2
-1
0
1
2

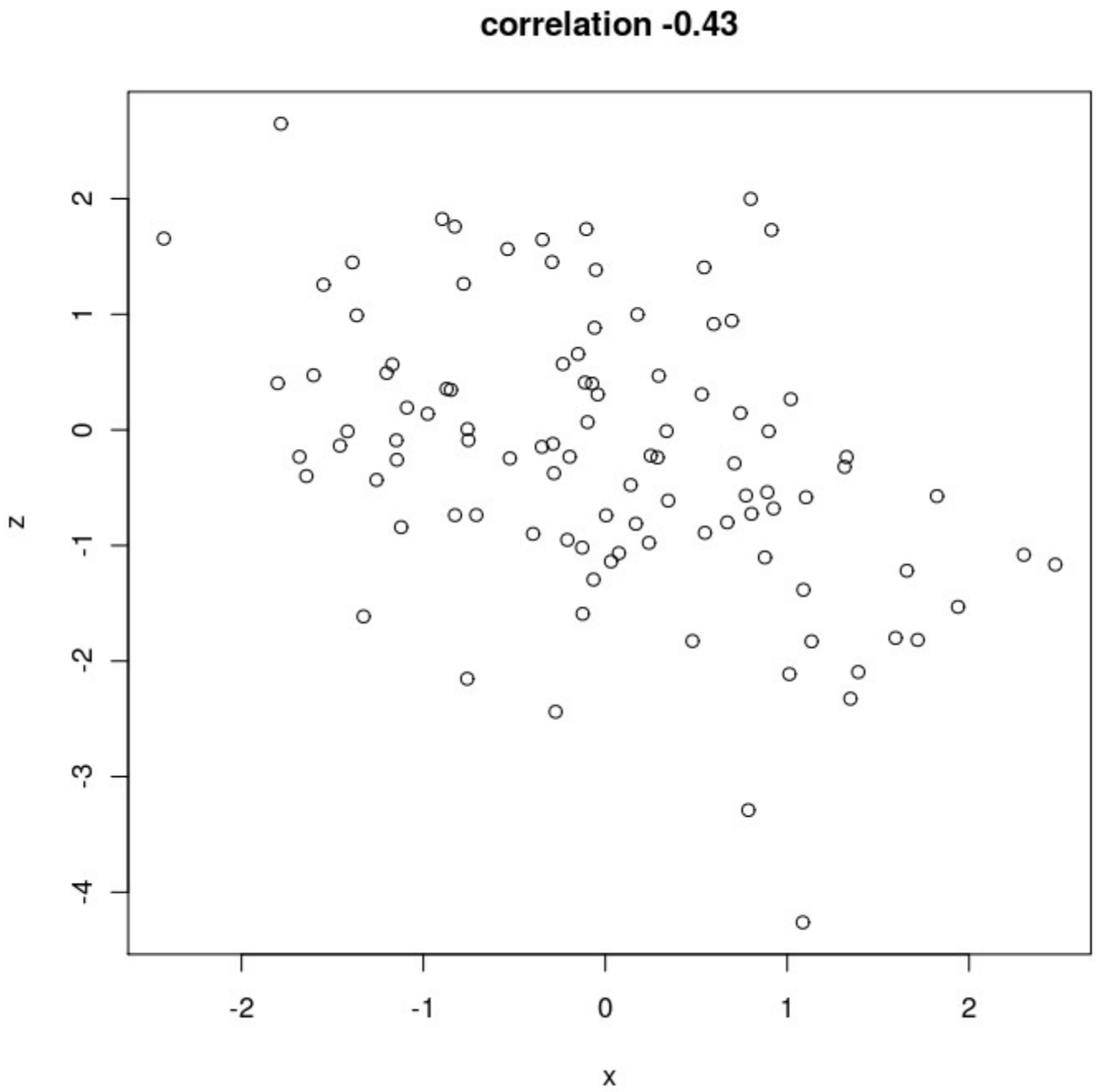

correlation 0.99
z
x
15
10
5
0
-5
-10
-2
-1
0
1
2

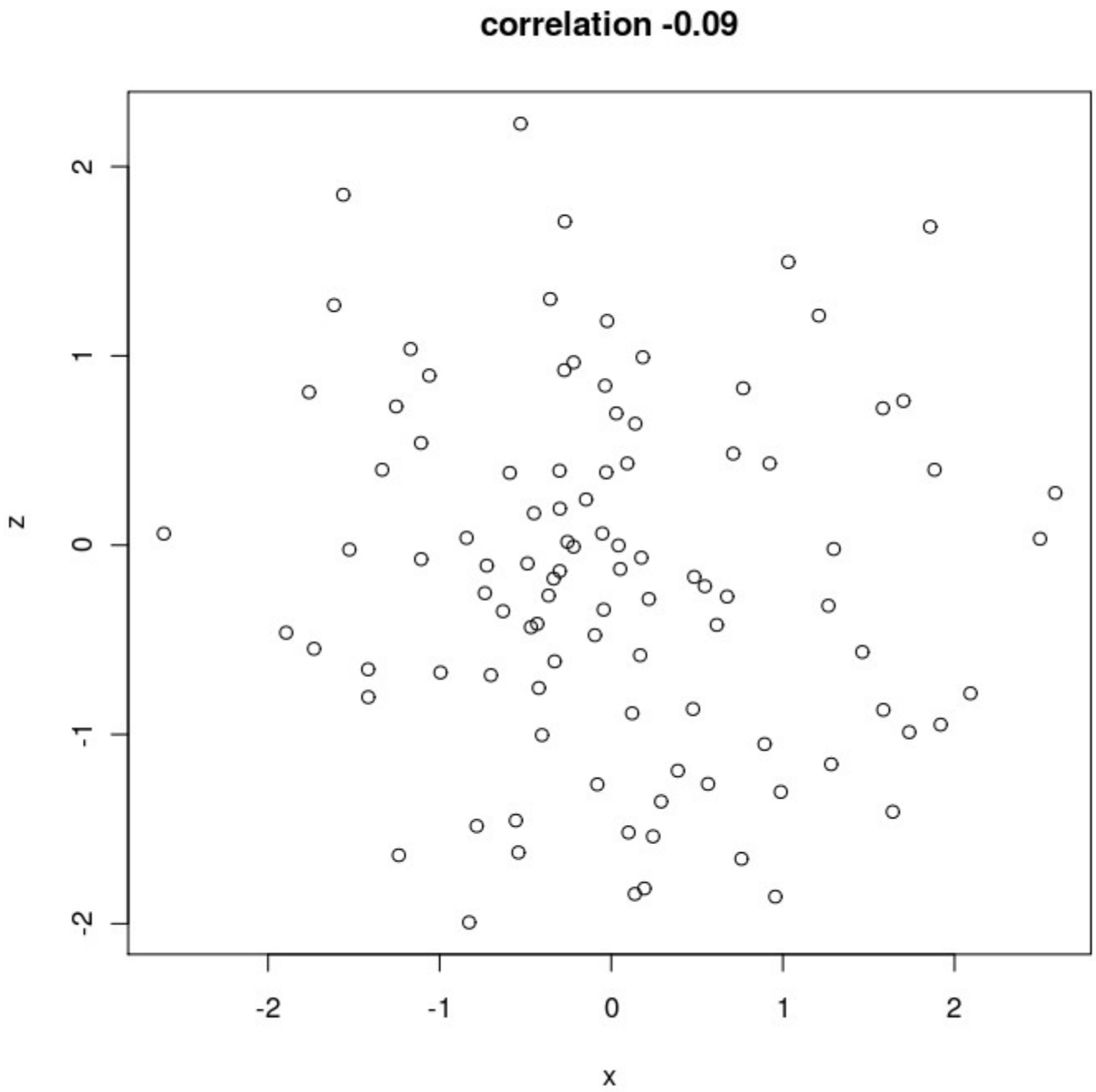

correlation 1
z
x
100
50
0
-50
-100
-150
-2
-1
0
1
2


correlation -0.98
z
10
5
0
-5
-10
-15
-2
-1
0
1
2
x


correlation -0.9
z
x
4
2
0
-2
-4
-2
-1
0
1
2

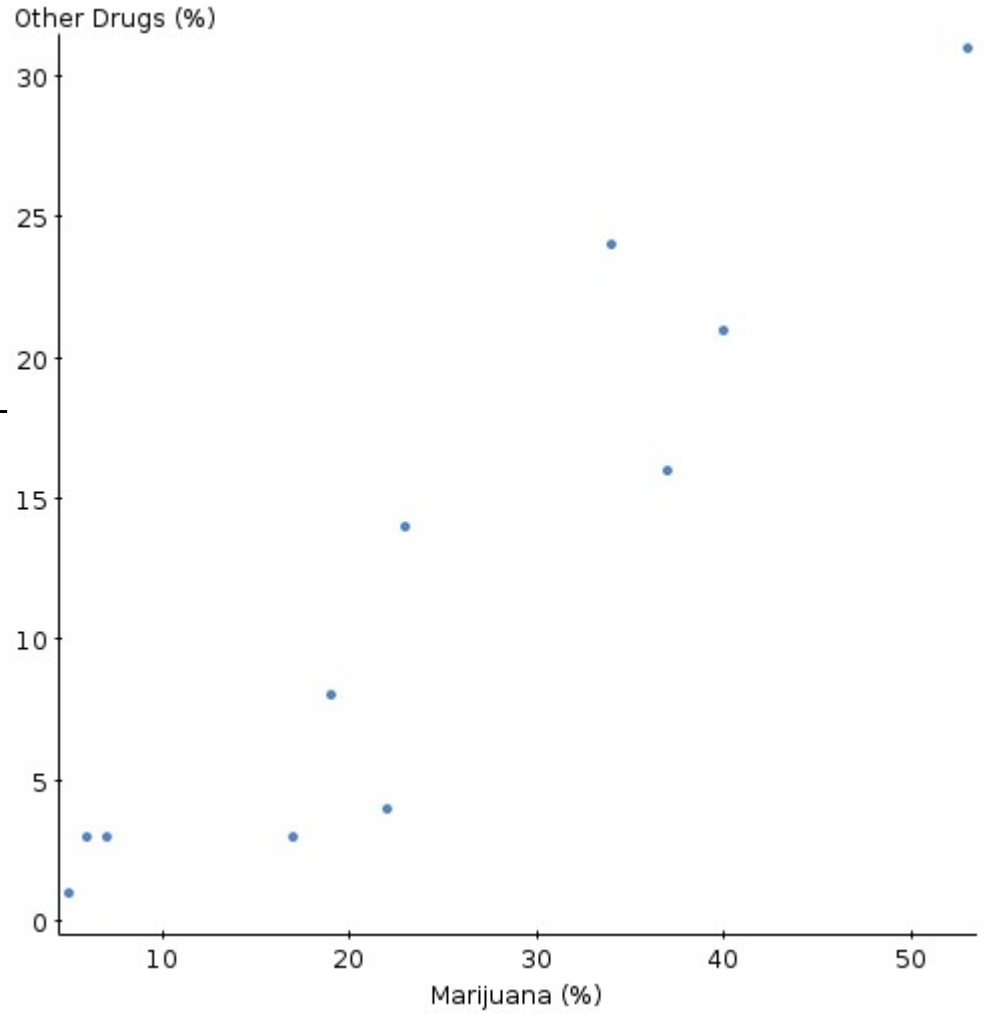

correlation -0.66
z
x
3
2
1
0
-1
-2
-3
-2
-1
0
1


correlation -0.43
z
x
2
1
0
-1
-2
-3
-4
-2
-1
0
1
2


correlation -0.09
z
x
2
1
0
-1
-2
-2
-1
0
1
2

Scatterplot of marijuana use vs. other drug use for different countries:

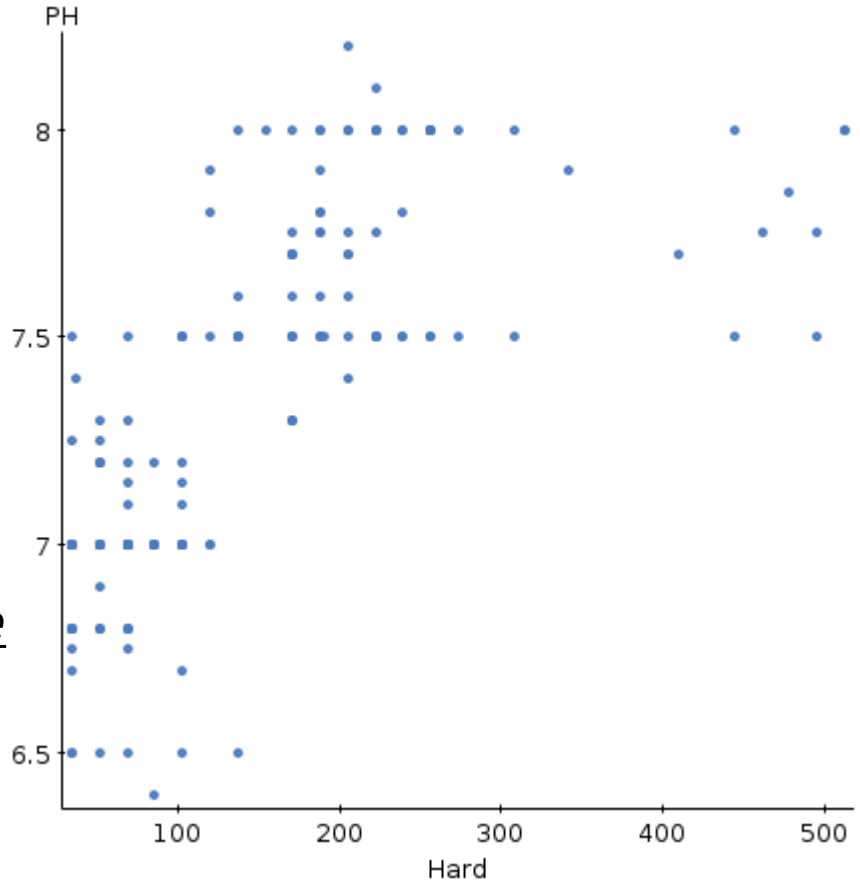
- Describe what you see on the scatterplot (form, direction, strength). Linear, upward,_____ moderate to strong_____
- Does it make sense to find a correlation here? Why, why not?
  - _yes, trend is straight



- Can you conclude that marijuana is a “gateway drug”: marijuana use leads to use of other drugs? Why or why not?
  - *no: correlation shows that relationship exists, but does not show cause and effect*

(My data set: “drug abuse”.)

In a study of streams in the Adirondack mountains, the following association was found between the pH of the water and its hardness:



- Describe the relationship (3 things).
  - *Form: curve; direction: upward; strength: moderate*
- Is it appropriate to use the correlation to summarize the relationship? Explain.
  - No: it's a curve not a line

Response and explanatory variables

- When you calculate a correlation, it doesn't matter which variable is x and which is y.
- Sometimes one variable is an "outcome" or *response*, and the other explains the outcome, an *explanatory variable.*
  - In that case, call the response y and plot it on the vertical axis of a scatterplot.

Chapter 8: Linear regression – finding the best line (p. 198)

In math, straight line relationship looks like

$$y = a + bx$$

where x and y are variables, and a and b are numbers that describe what kind of straight line you have.

- a = "intercept": value of y when x=0
- b = "slope": if you increase x by 1, how much do you increase y by?
  - *slope=2: increasing x by 1 increases y by 2*
  - *slope=-3: increasing x by 1 decreases y by 3*
  - slope could be negative, if line goes downhill.

If you know the intercept and slope, you know the straight line.

So aim: find intercept and slope of line that "best" describes data.

Straight line only *model*: won't be perfectly accurate. But may be useful.

Residuals

Go back to drug abuse data set.

Let y=other drug use, x=marijuana use. Let $\hat{y}$ be *predicted* other drug use.

Suppose model is $\hat{y} = -3 + 0.5x$.

For England, x=40, y=21, *prediction* $\hat{y} = -3 + (0.5)(40) = 17.$

England's actual drug use was 21, higher than predicted by the model: $\text{residual} = y - \hat{y} = 21 - 17 = 4.$

Idea: want "best" line to pass close to all the data, so want all residuals *close to 0*.

So decide how good a line is by working out all the residuals, *square* and add up. (Like variance). Best line is one with sum of squared residuals *smallest*. (*could try a bunch of candidate lines, work out all the residuals for each one, add up and compare. But can do better.)*

How to find the least squares line (p. 201)

- find means and SDs of both variables: $\bar{x}, \bar{y}, s_x, s_y$ and correlation r between them
- slope = $r\frac{s_y}{s_x}$
- intercept = $\bar{y} - (\text{slope} * \bar{x})$

Drug abuse example

Summary statistics (from StatCrunch)

Column	Mean	Std. dev.
Marijuana (%)	23.909091	15.552842
Other Drugs (%)	11.636364	10.239851

Correlation between Other Drugs (%) and Marijuana (%) is: 0.93410002

Calculate the slope and intercept of the least-squares regression line for predicting other drug use from marijuana use. (Answers: slope 0.615, intercept -3.066.)

– *slope=(0.934)*(10.24/15.55)=0.62*

– *intercept=11.64-23.91*0.615=-3.06*

– Check with StatCrunch. *Check.*

Prediction:

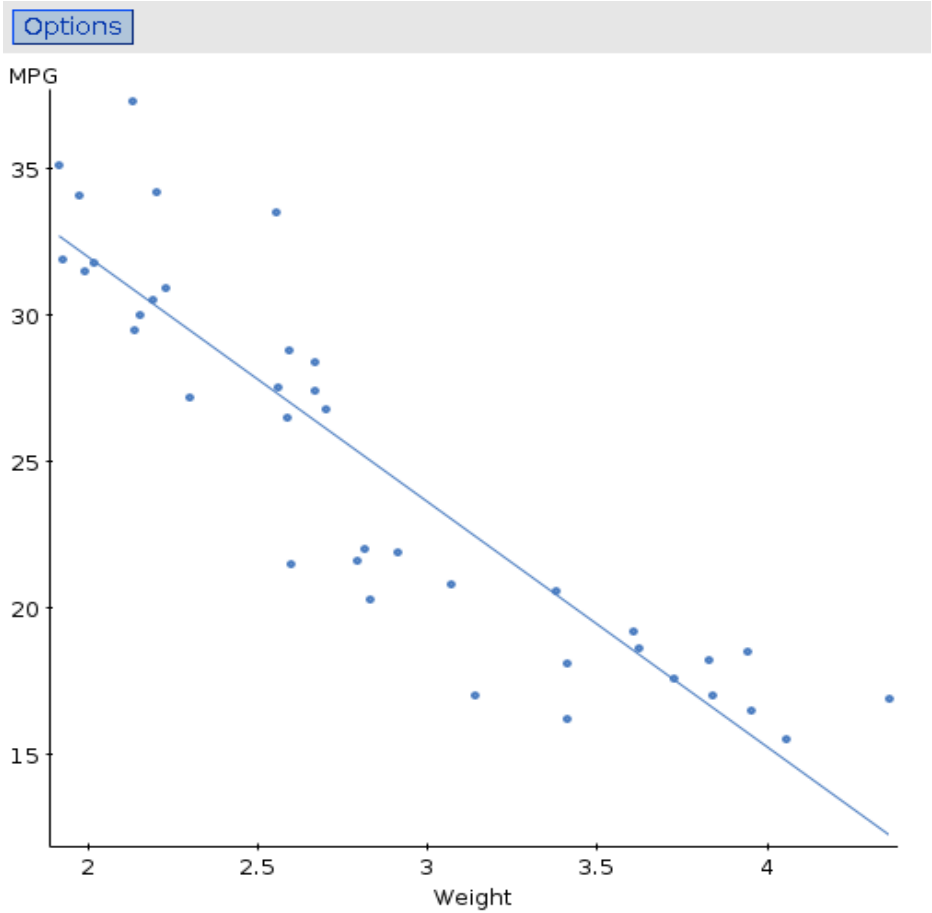
- Predict "other drug use" for England (x=40), which was actually 21. How close is the regression line?
  - regression line is $\hat{y} = -3.066 + 0.615\,x$
  - prediction is $\hat{y}$ = *-3.066+0.615*40=21.53*
  - actual value *21*
  - residual = *21-21.53=-0.53*

How good is the regression? (p.209)

- *If* the scatterplot looks straight, correlation will describe this.
- Also can use R-squared = correlation squared, between 0 and 1 (0 and 100%).
- R-squared generalizes to *multiple regression* where you have more than one x-variable.
- *Percent of variability explained by regression. The response variable (y) is higher or lower, but how much of that is because it depends on the explanatory variable (x)?*
- For drug abuse data, R-squared is 0.9341^2 =0.87. Pleasantly high.
- *Values of "other drug use" vary quite a bit, but most of that variability happens because it depends on marijuana use.*

Car data

- data on number of car models. Here predict gas mileage (miles per US gallon) from the weight of the car – expect heavier cars to have a worse (lower) mpg.
- *first: scatterplot.* Is association straight?
  - More or less a line
- correlation, -0.903 indicates *strong* relationship.



Regression for predicting MPG from weight (StatCrunch):

Simple linear regression results:

Dependent Variable: MPG
Independent Variable: Weight
MPG = 48.707495 - 8.3645999 Weight
Sample size: 38
R (correlation coefficient) = -0.90307083
R-sq = 0.81553692
Estimate of error standard deviation: 2.8508049

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Val
Intercept	48.707495	1.9536817	≠ 0	36	24.931131	<0.00
Slope	-8.3645999	0.66302032	≠ 0	36	-12.615903	<0.00

Note that R-squared, 0.816, pleasantly high.

You verify intercept and slope using the below (correlation is above):

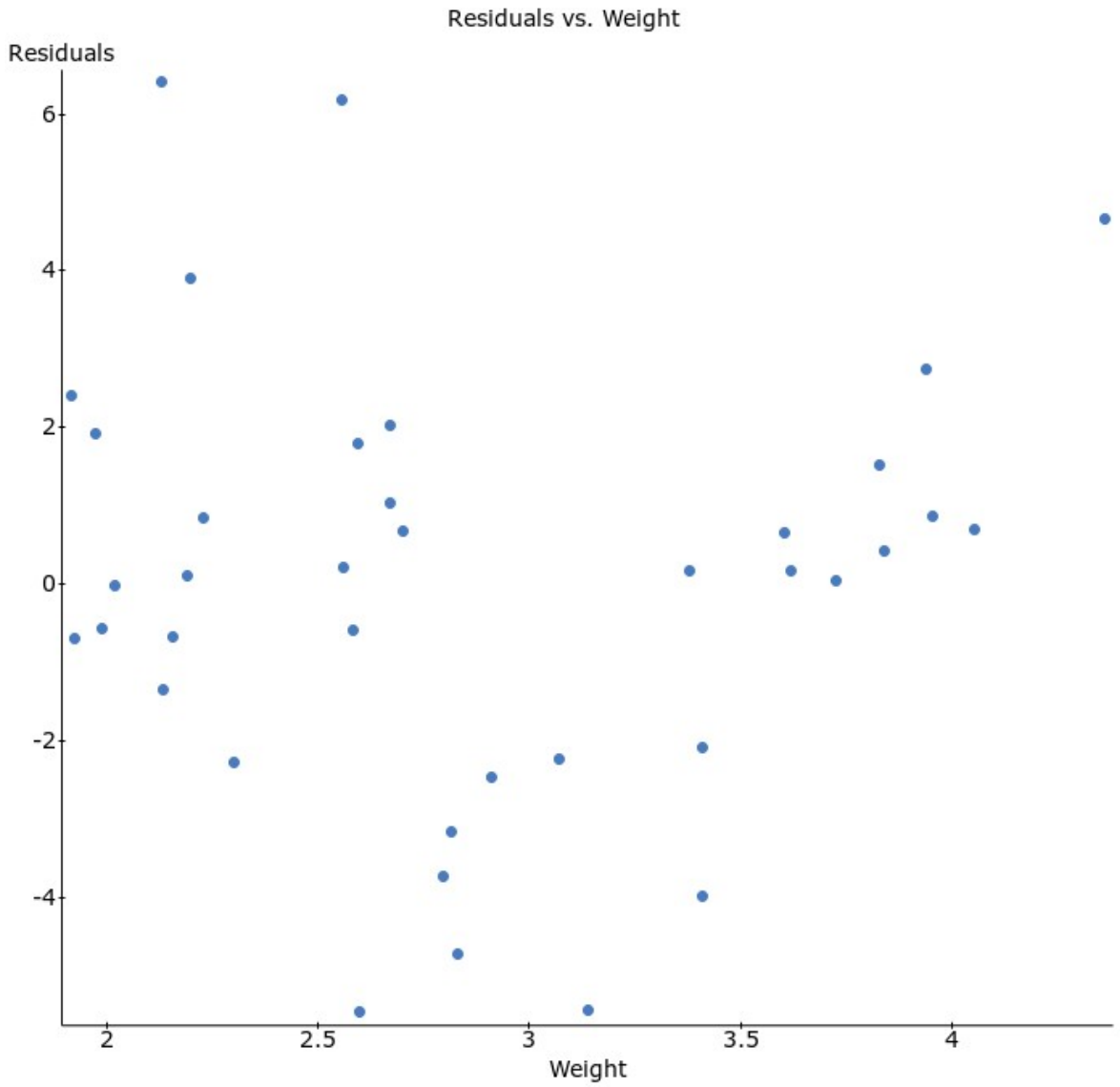
Summary statistics:

Column	Mean	Std. dev.
MPG	24.760526	6.5473138
Weight	2.8628947	0.70687041

- Intercept 48.7, slope -8.4.
- Predicted MPG for car weight 2.5 tons?
  - *48.7+2.5(-8.4)=27.7*
- Predicted MPG for car weight 6 tons?
  - *48.7+6(-8.4)=-1.7*
  - Why is this last one nonsense?
    - *Cannot be negative.*
- ***extrapolation (bad)***

Plot *residuals* against weight:

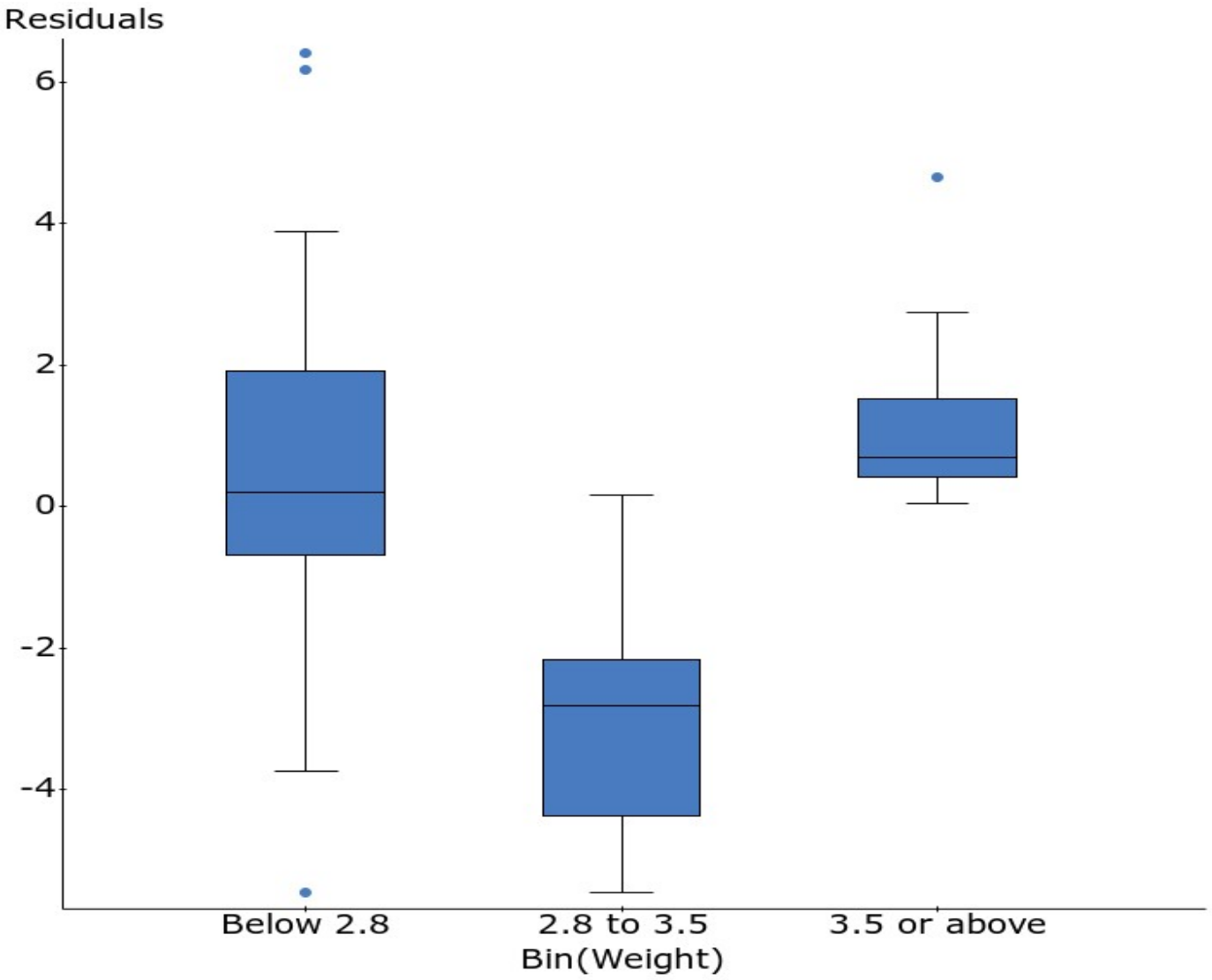
- down-and-up pattern
- residuals for weights around 3 tons are *negative*
- those for low & high weights mostly *positive*
- residual plot shows *curve*: **actual relationship is curved, not linear.**
- So predictions we made not completely trustworthy.



Not sure about that?

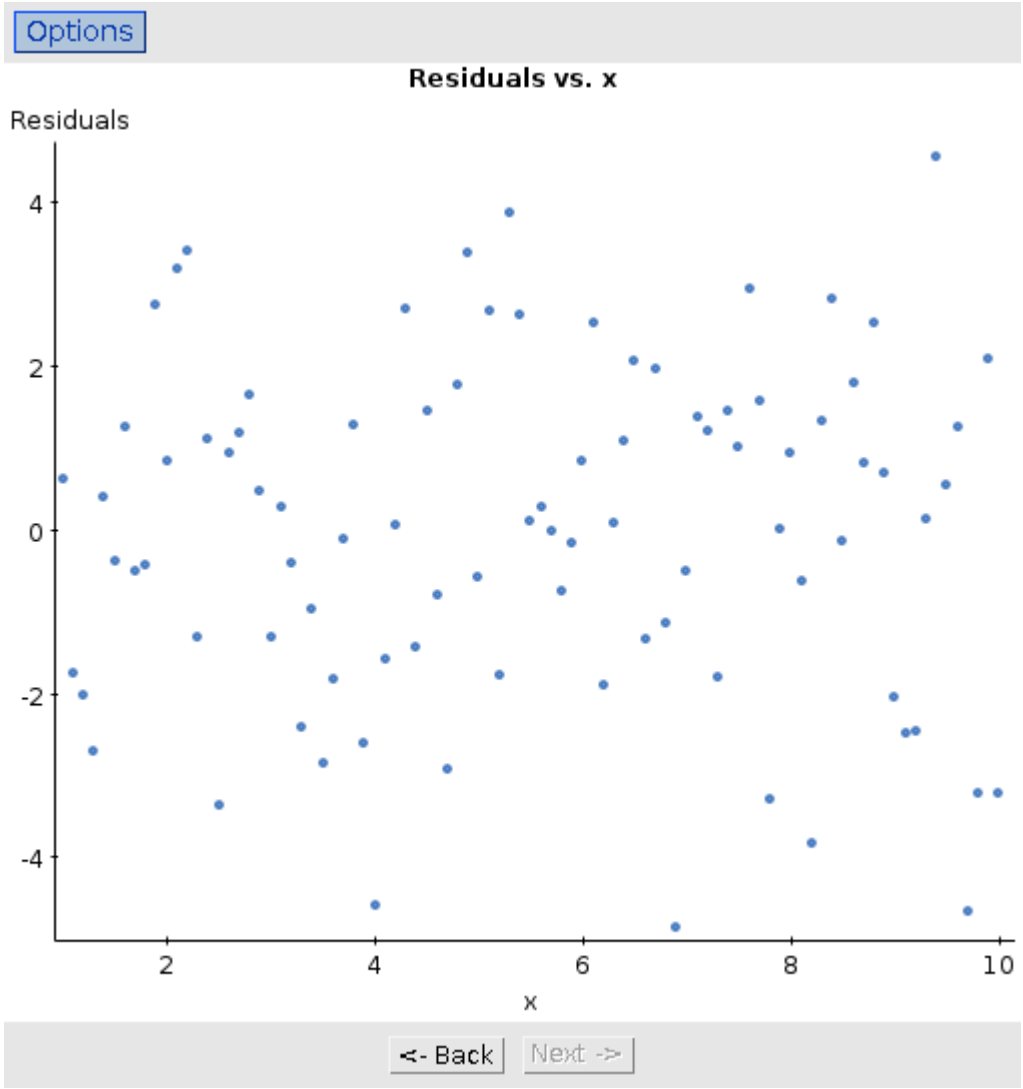
- Divide weights up into “light” (below 2.8), “medium” (up to 3.5), “heavy” (above 3.5).
- In StatCrunch: Data, Bin. Select variable Weight, Use Cut Points, enter in box as 2.8, 3.5 with comma.
- This starts from quantitative variable (Weight) and produces categorical (Bin(Weight)).
- To plot residuals by group (defined by Bin(Weight)), use *boxplot,* over.
- Conclusion: medium-weight cars tend to have *negative* residuals, heavy-weight cars tend to have *positive* ones.
- When the residuals *depend* on anything, we have a problem.

Boxplot of residuals by binned weight



Random rubbish

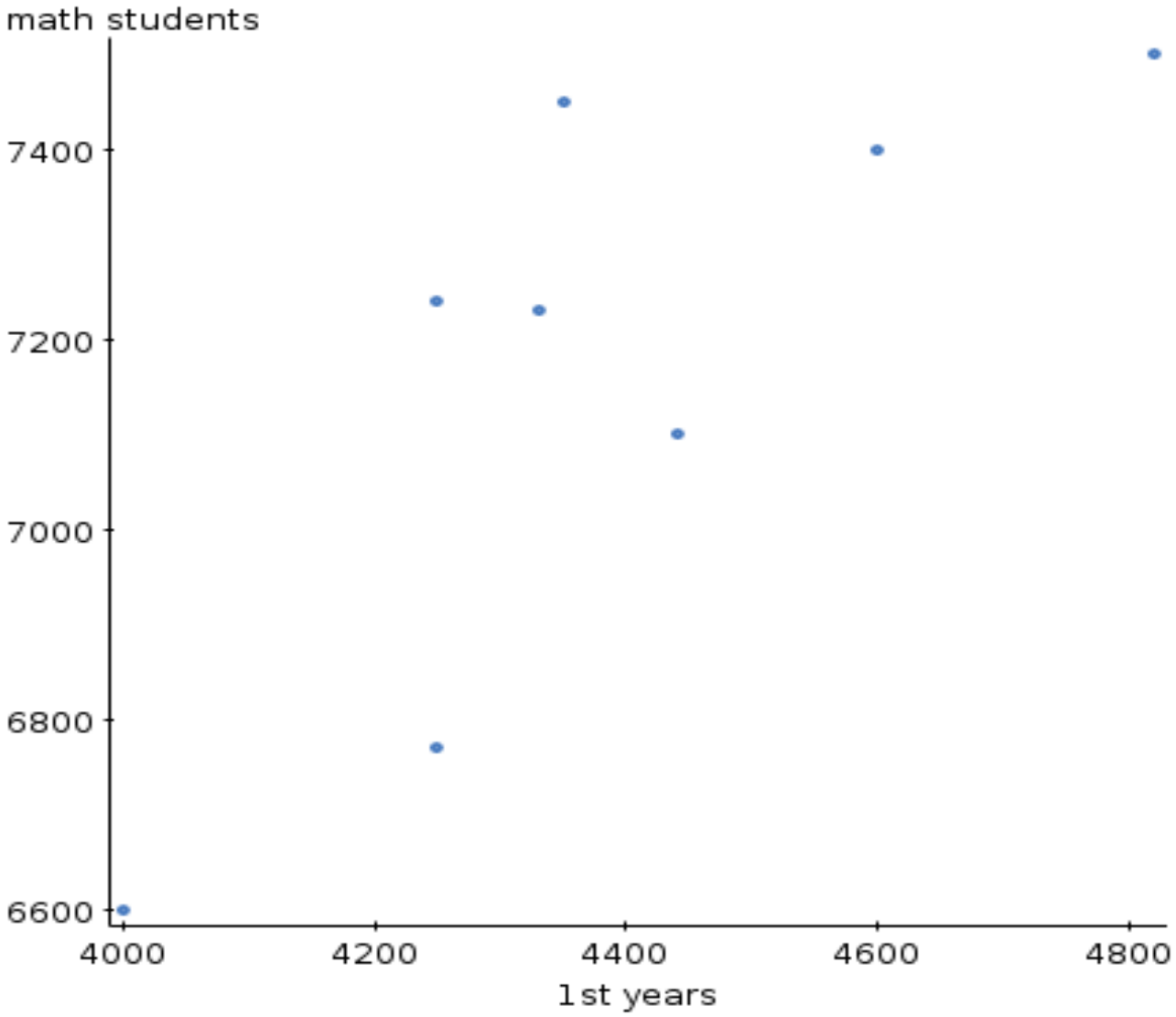
- This residual plot, from another regression, has no pattern whatever.
- The regression it came from has no problems.



Doing regression

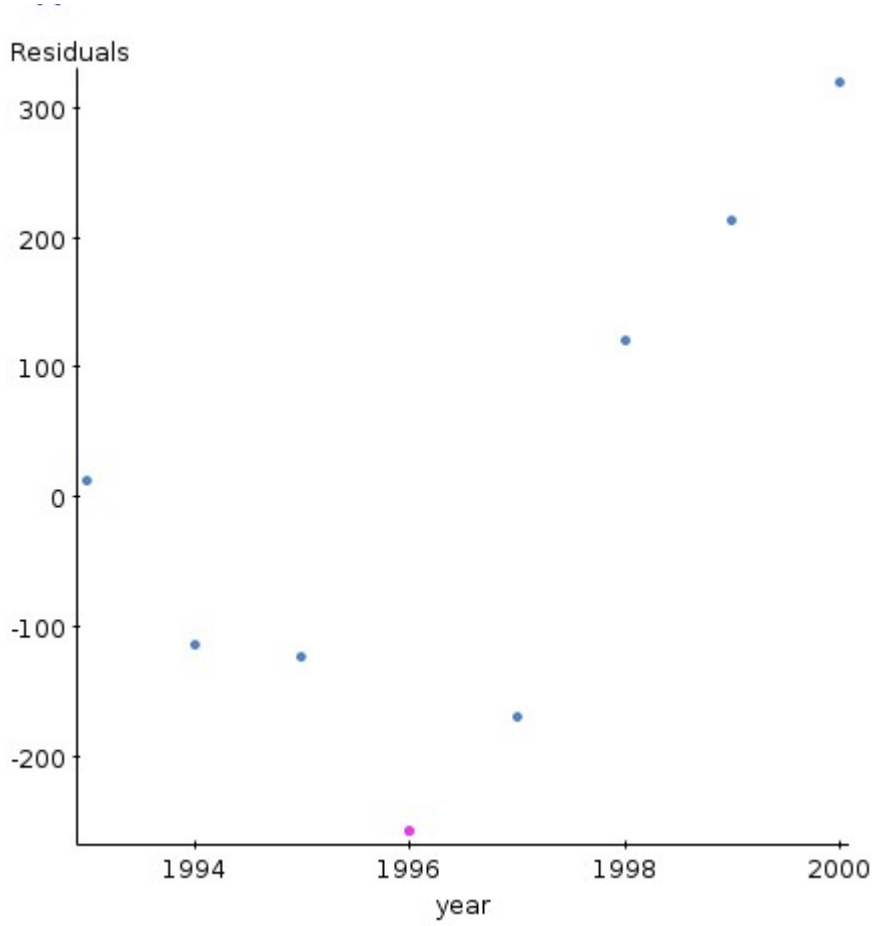
- start with a scatterplot
  - if it does not look like a straight line relationship, stop (see Chapter 10).
- otherwise, can calculate correlation and also intercept and slope of regression line
- check whether regression is OK by looking at plot of residuals against anything relevant
  - if not OK, do not use regression.
- Aim: want regression for which line OK, confirmed by looking at scatterplot and residual plot(s). Otherwise, cannot say anything useful.

At a certain university, for each of several years, the total number of first years is recorded, and also the number of students taking "elementary math courses". A scatterplot looks like this:



- Would you say that as the number of 1st years increases, the number of math students increases too?
  - *yes*
- Would you consider fitting a regression line here?
  - *I guess so*

– For the above data, a regression is done to predict the number of math students from the number of 1st years. The residuals are saved.
  – There is a third column in the data set, the year in which the students were counted. A plot of residuals against year looks like this:



– Does this suggest that the regression was satisfactory, or not? If not, why not?
  – *No: down and up (turns corner at 1996)*

– *New calculus requirements in 1997 (saw from residual plot that there was a problem)*

Correlation and causation (p. 179)

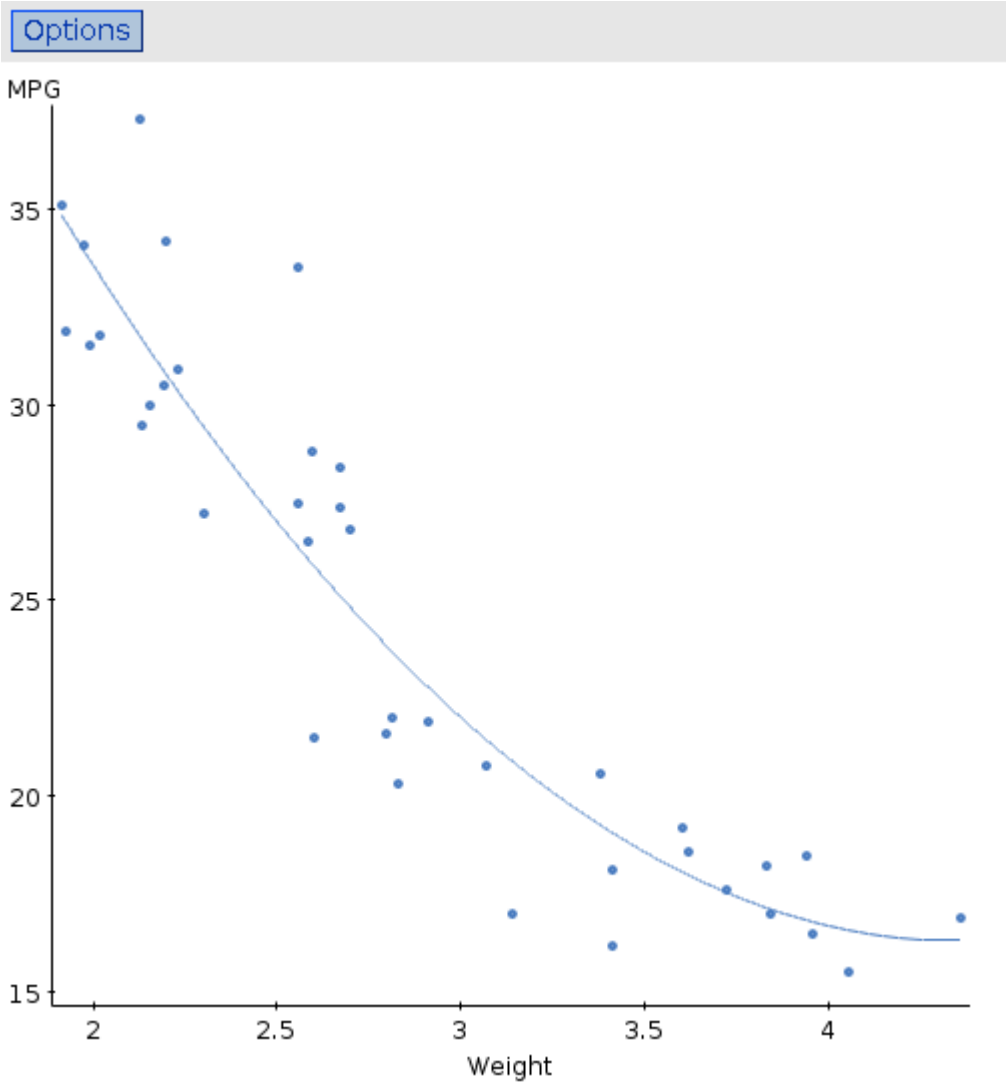
- high correlation between #sodas sold in year and #divorces, years 1950-2010. Does that mean that having more sodas makes you more likely to divorce?
  - *Over that time, population increased, so changes in both variables are caused by that*
- observe that smokers have higher blood pressure on average than non-smokers. Smoking causes higher blood pressure?
  - *Causation could be other way: high blood pressure causes smoking*
- high correlation between #teachers and #bars for cities in California. Teaching drives you to drink?
  - *Size of city controls number of teachers and number of bars*

- high correlation between amount of daily walking and quality of health for men aged over 65. Explanation?
  - *Maybe people warned about their health are walking more*
- positive relationship between #ski accidents and waiting time for ski lift for each day during one winter at ski resort. Having to wait a long time makes people impatient?
  - *More people at the resort will cause longer lines for the lift and more congestion on the slopes*
- Moral of the story:
  - *Correlation is not causation: cause and effect could be reversed, or there could be a 3rd variable driving the correlation*

Cars data again:

Scatterplot with fitted curve

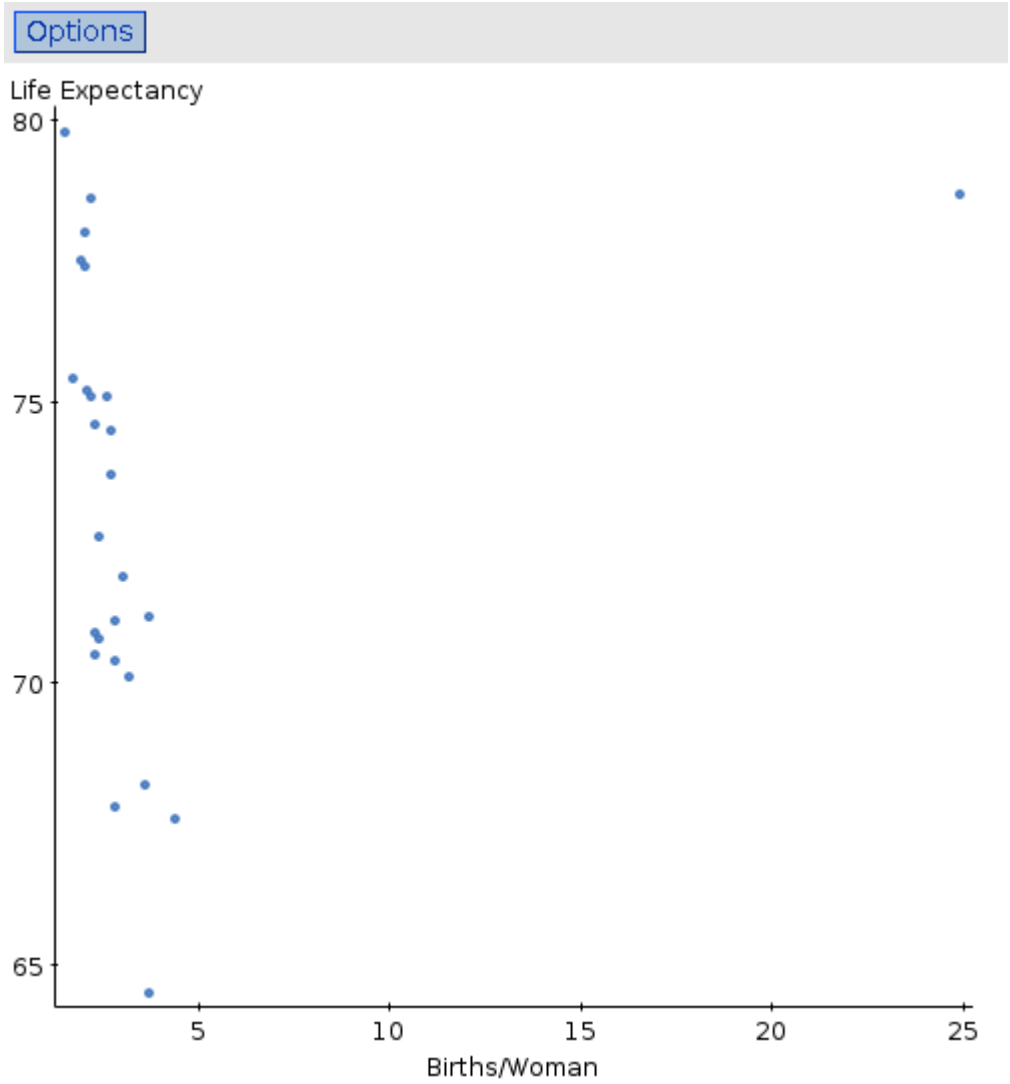
- curve *does* seem to go through points better
- as weight increases, MPG does not decrease so fast.



Life expectancy 2004

How does the life expectancy for a woman depend on number of children that woman has? Data for 26 countries ("life_ expectancy_2004"). Scatterplot:

- as births/woman increases, life expectancy decreases.
- *Huge* outlier top right.



Regression with and without the outlier:

R-sq = 0.028277527

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Va
Intercept	72.602542	0.99717777	≠ 0	24	72.808024	<0.0
Slope	0.15004325	0.17954001	≠ 0	24	0.83570926	0.4

Without (how to do?)

R-sq = 0.63304984

Parameter estimates:

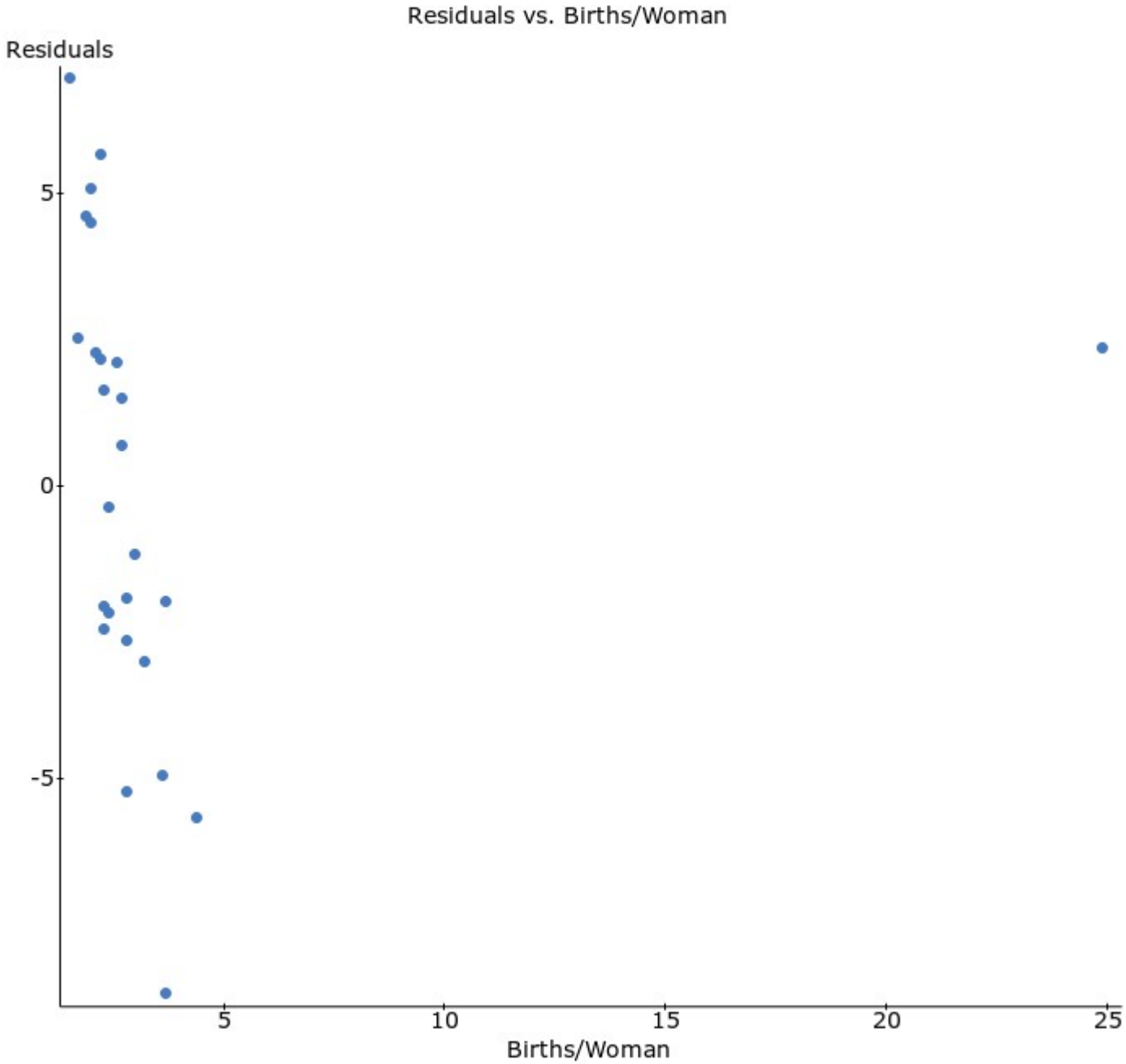
Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Valu
Intercept	84.497094	1.9019459	≠ 0	23	44.426655	<0.000
Slope	-4.4399288	0.70484996	≠ 0	23	-6.299112	<0.000

- *Completely* different!
- What has changed (two things)?
  - *Slope has become negative*
  - *R-squared has increased from almost zero to 63%*
- The outlier is *influential* because it has unusual value for explanatory variable.

How to do a regression omitting potential outliers

- Stat, Regression, Simple Linear.
- Fill in x as Births/Woman and y as Life Expectancy.
- What makes outlying point unusual? Its births/woman is way high.
- Want to select countries where births/woman reasonable, say less than 5.
- On Where box, click Build.
- Want to *include* countries where births/woman *less than 5*.
- Under Columns, click Births/Woman and click Add Column. Using blue boxes to left, click or type < and 5. Click “Okay”. Check that right thing now in Where box.
- Click Compute as usual.

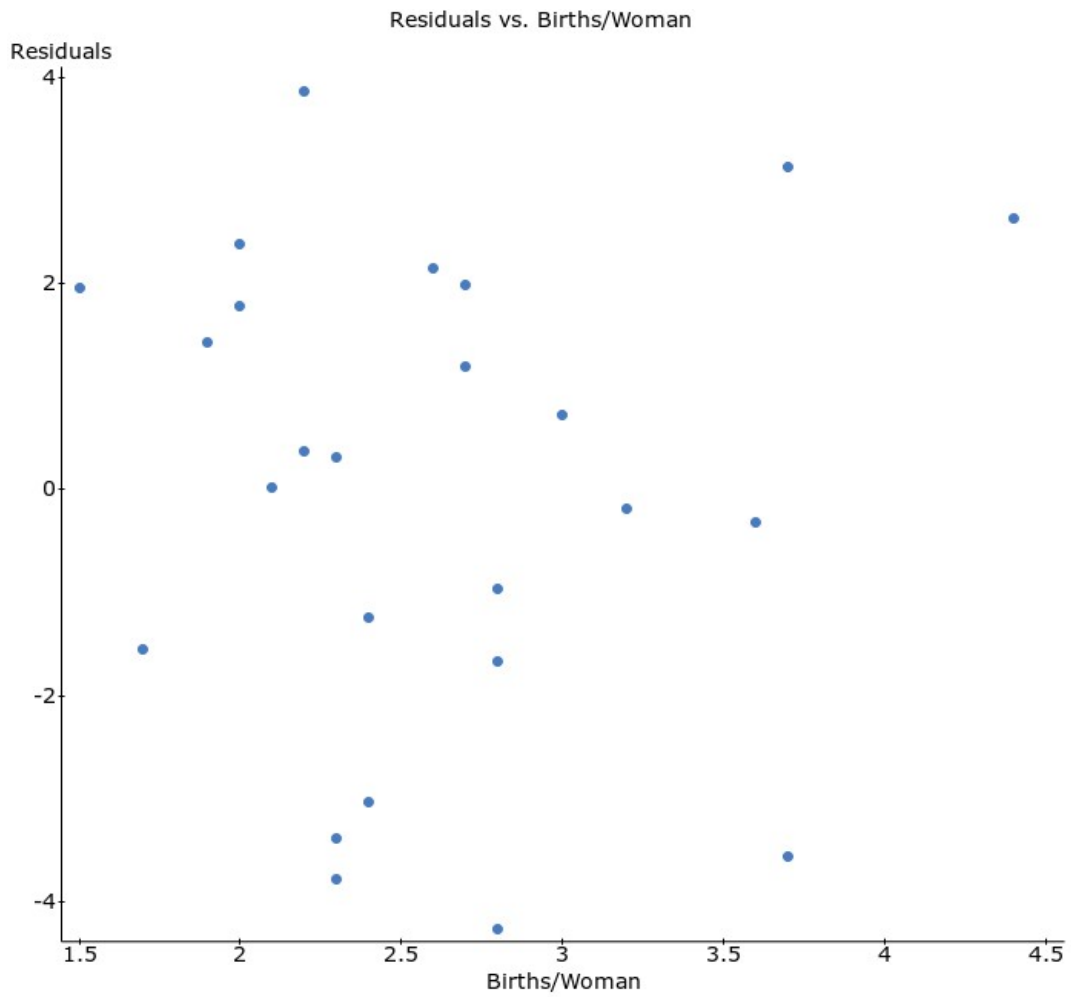
Residual plot from regression *with* outlier:



Do we trust the regression **without** the outlier?

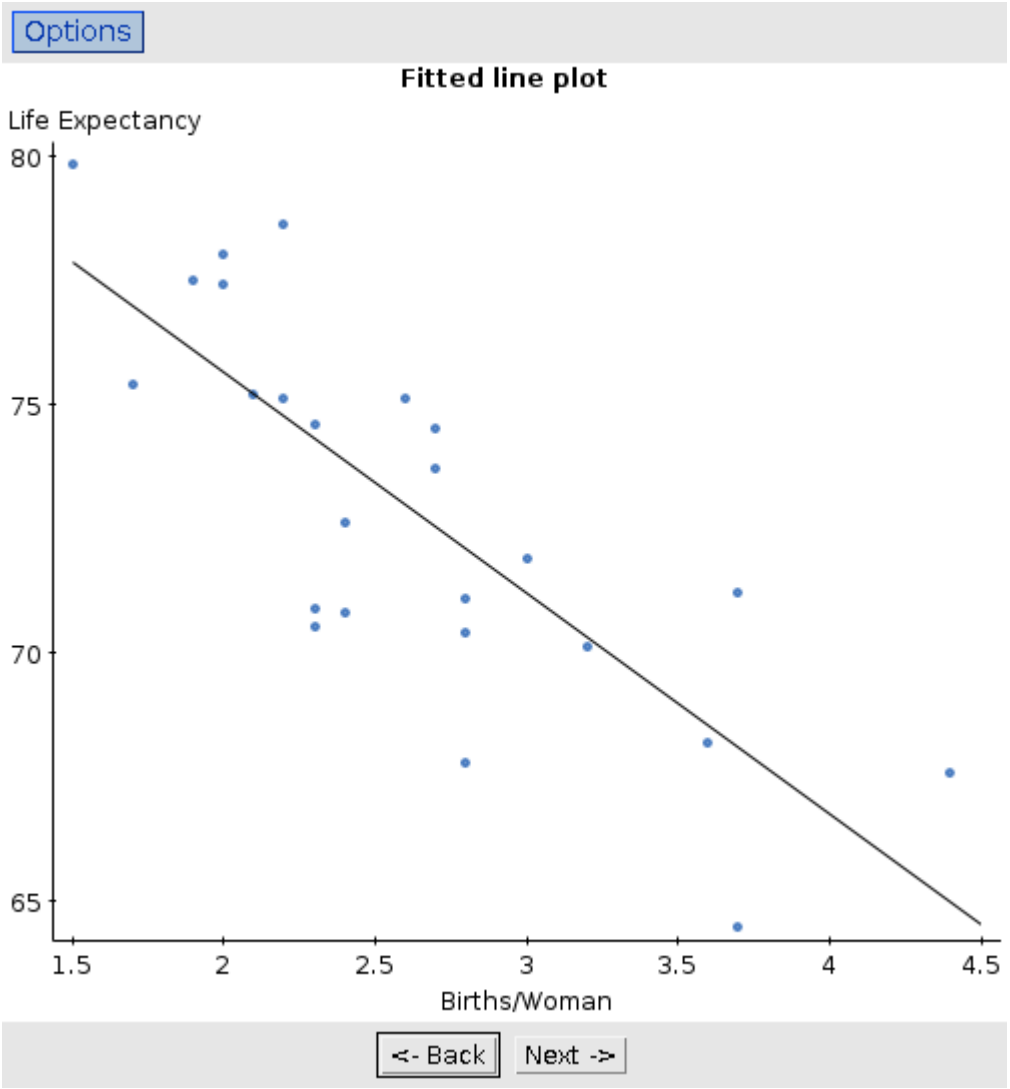
Residual plot:

Is it a random scatter of points? _*yes*_____



Fitted line plot (without outlier):

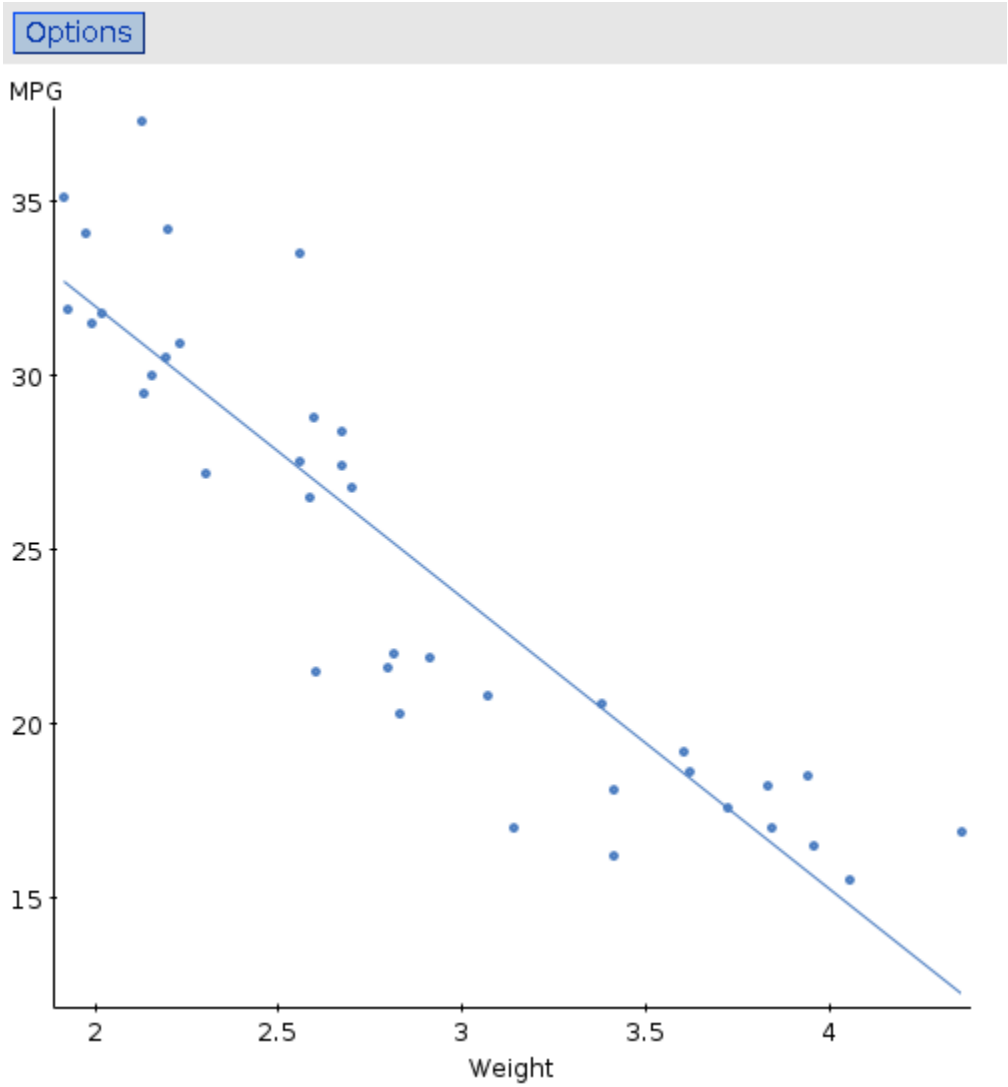
- Direction, form, strength?
  _*Downward; line; moderate to strong*_____
- What do you think Costa Rica's births/woman should have been (78.7 life exp, 24.9)?
  __*2.49*_____
- What do you think happened to Costa Rica's data?
  __*Somebody mis-transcribed it* _____



Correlation with subsets of data (p. 242)

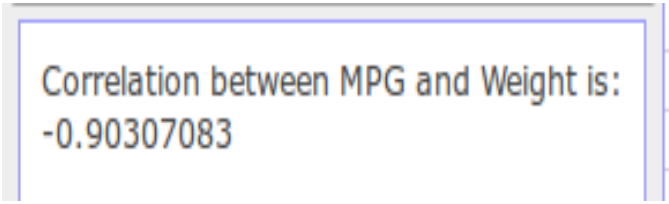
Return to cars MPG/weight:

- cars seem to divide into "good" gas mileage (above 25) and "bad" (below).
- What is correlation between MPG and weight for just the "good" cars, or just the "bad" cars? How does that compare to correlation for all cars?

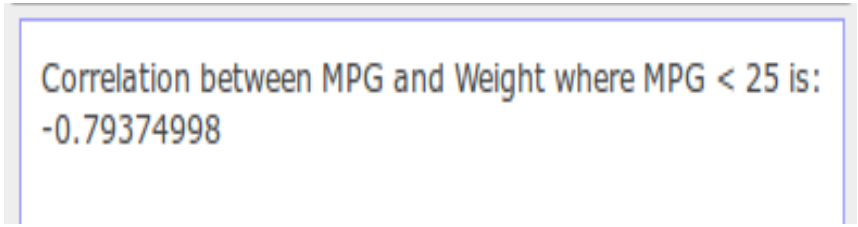


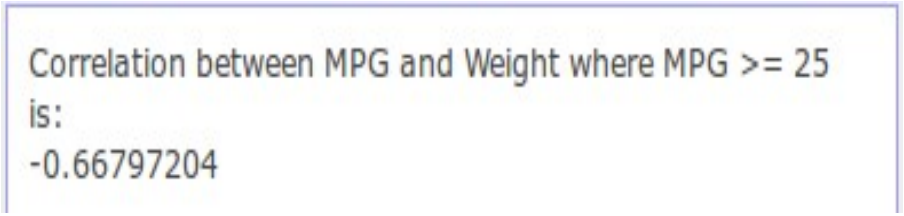
Correlations:

- correlation for just part of the data is *lower* (closer to 0) than for all the data.



- look at the data for just the low-MPG cars (in scatterplot above). How would you describe that MPG-weight relationship?
  *_Looks stronger than for all cars*_____



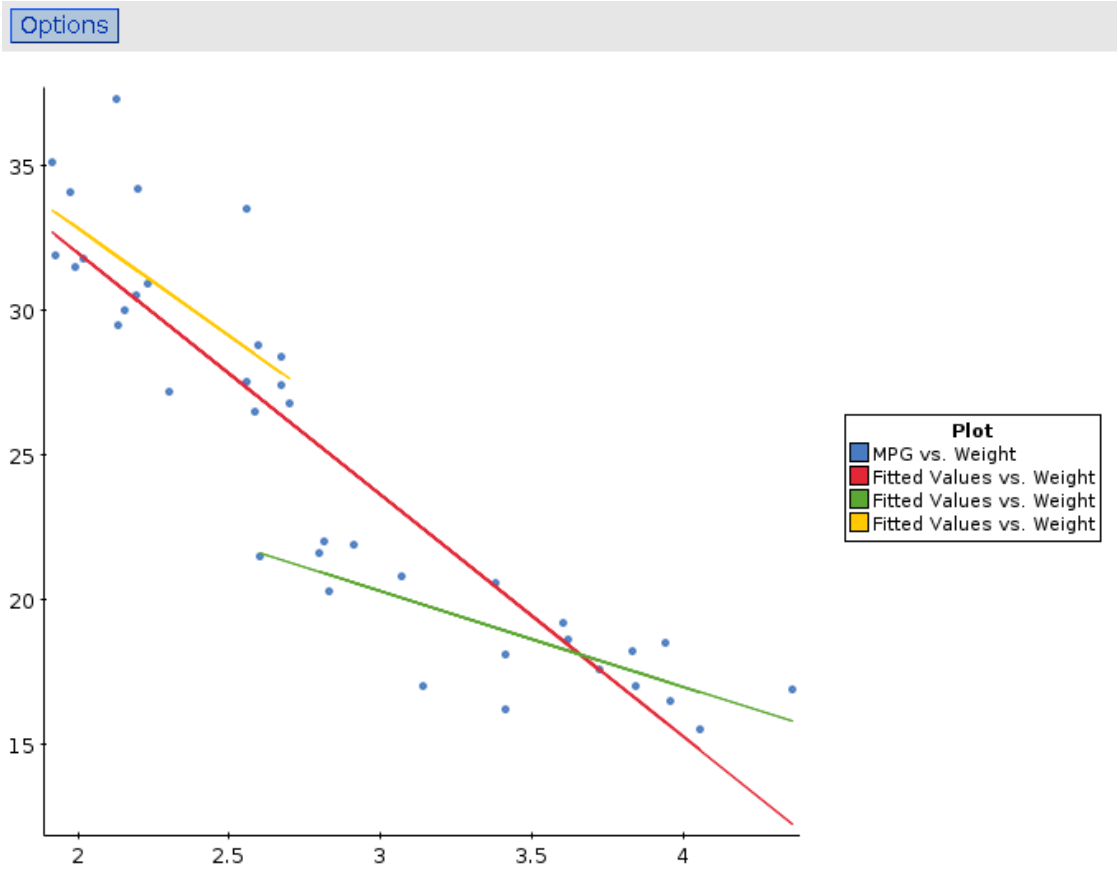


- Do the answers to the right surprise you? *_Yes; would have expected stronger correlation rather than weaker one*_____

Called the *restricted-range* problem: when one of the variables is restricted (you only look at some of the values), the correlation can be surprisingly low.

Scatter plot showing separate regression lines for each group

- the regression lines are quite different, which would make you expect to see a higher correlation *within* each group.
  - *but you don't!*



Chapter 10: Re-expressing data – Get it Straight! (p. 293)

- Take a simple function of the data (the response, in regression) to achieve:
  - make the distribution more symmetric
  - make spreads of several groups more similar
  - make a scatterplot more linear
  - make spread in a scatterplot same all the way along

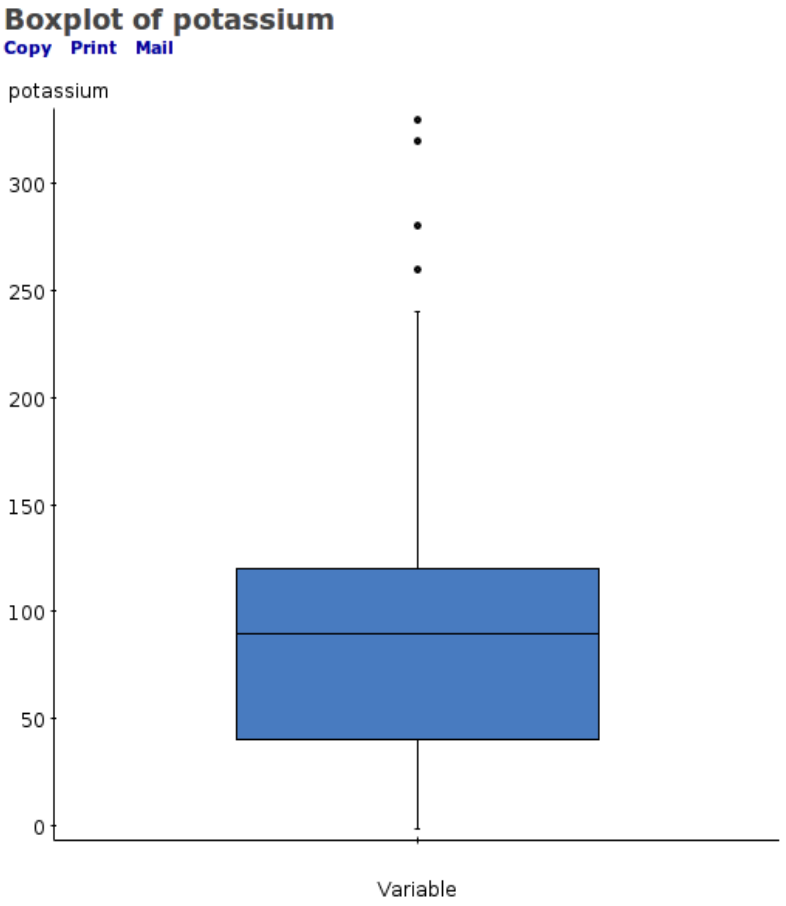
Stemplot and boxplot of cereal potassium data. What would you like to fix?

outliers

Variable: potassium

Decimal point is 2 digit(s) to the right of the colon.

```
0 : 002233333333444444444
0 : 55555666666778999999
1 : 0000000111111222233444
1 : 667799
2 : 034
2 : 68
3 : 23
```

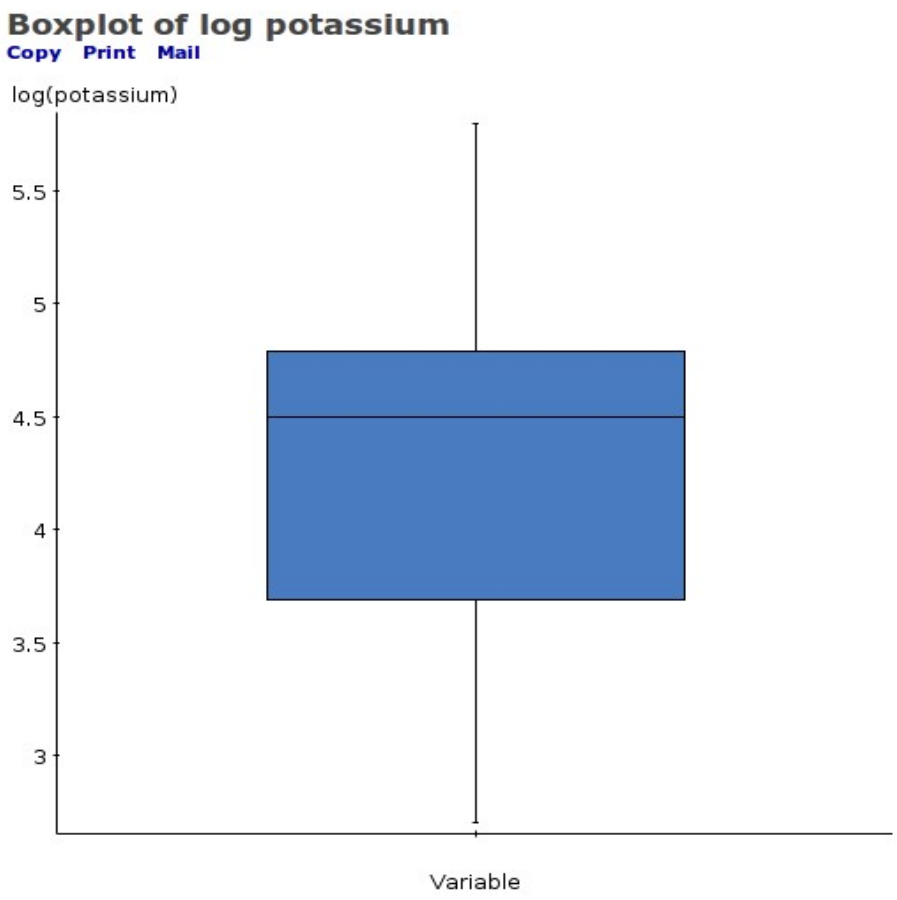


Try log of potassium:

Variable: log(potassium)

Decimal point is at the colon.

```
2 : 7
3 : 022224444
3 : 66666777788889
4 : 0001112244
4 : 55555666666677777777788889999
5 : 11112234
5 : 56688
```



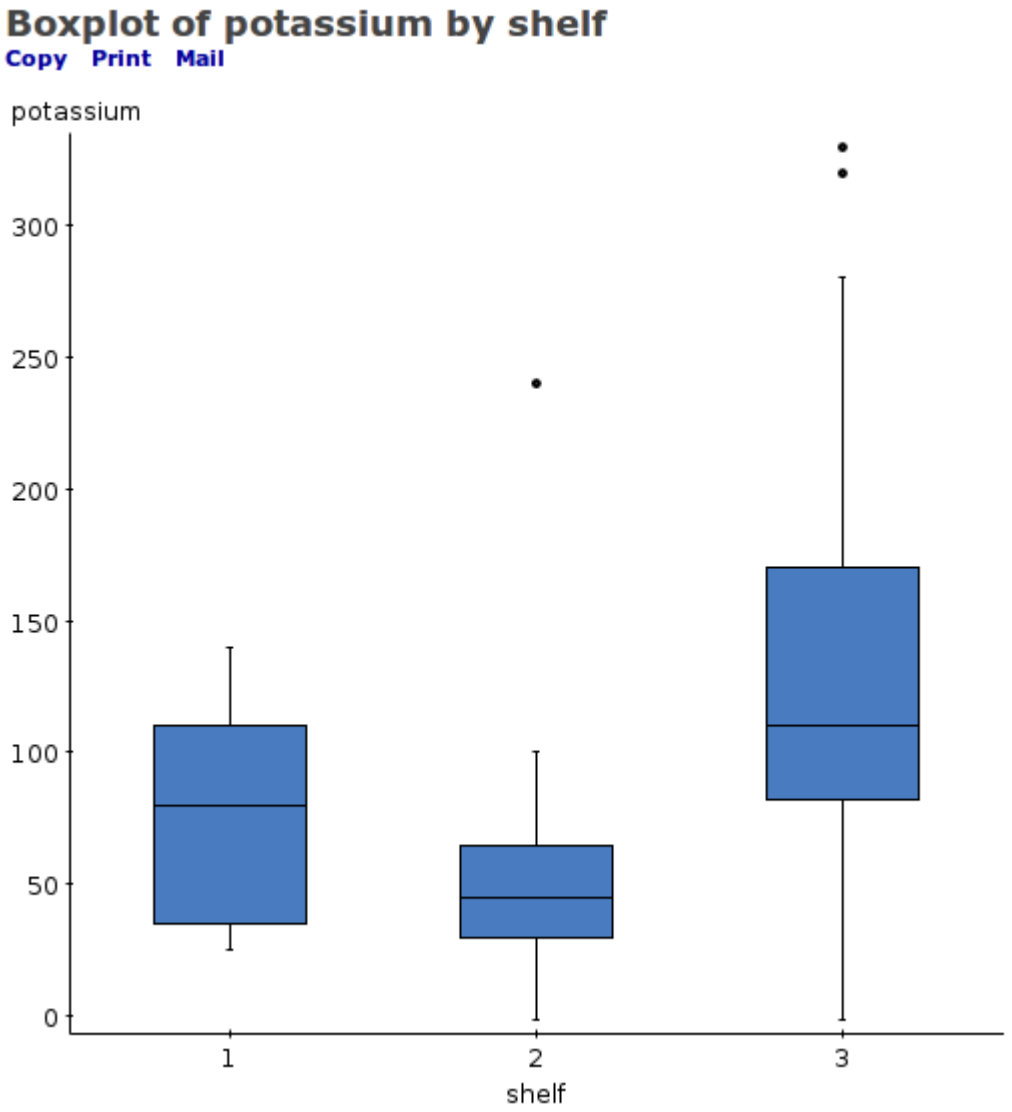
Is log of potassium more symmetric than potassium itself?

__*yes*_____

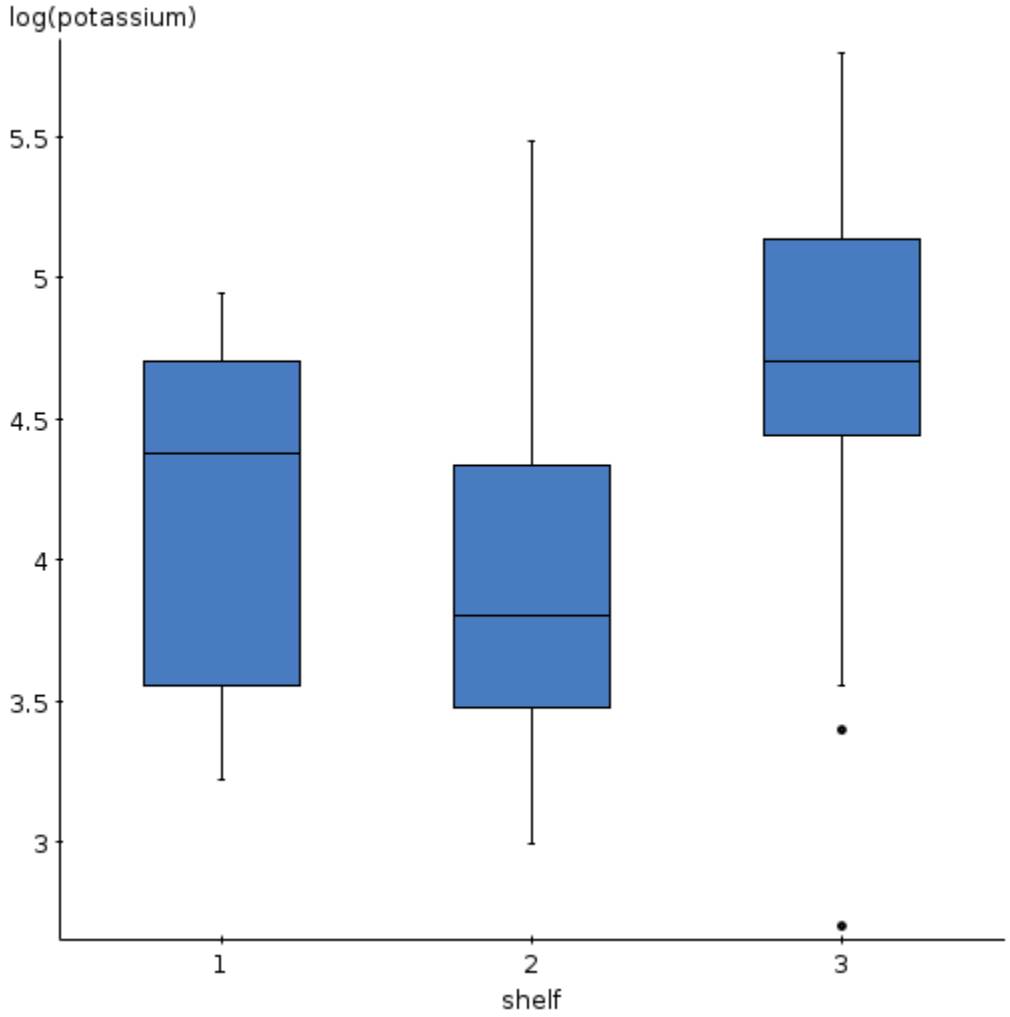
What about boxplots of potassium by shelf?

How does spread compare as centre gets larger?

___*It gets bigger*_____



Try using log of potassium values:

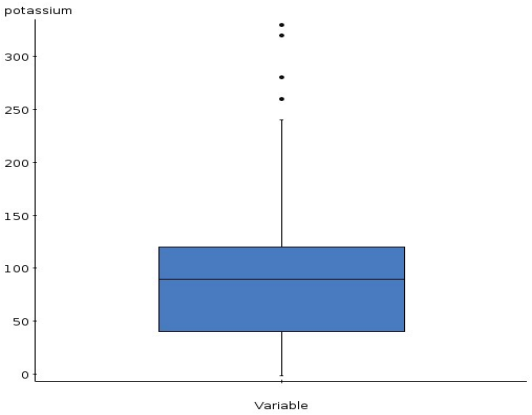


Spreads are more equal now, less dependent on centre.

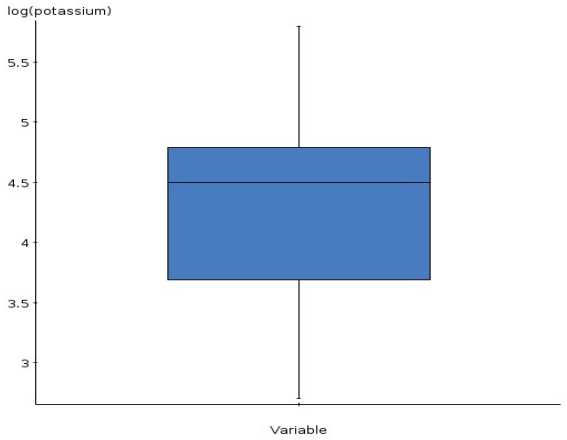
Ladder of powers (p. 269)

Power	Name	Notes
2	Square of values	Left skewed
1	Unchanged data	
0.5	Square root	Counted stuff
0	Logarithm	% change matters
-0.5	-1/square root	Rare
-1	-1/data	Ratio “wrong way up”

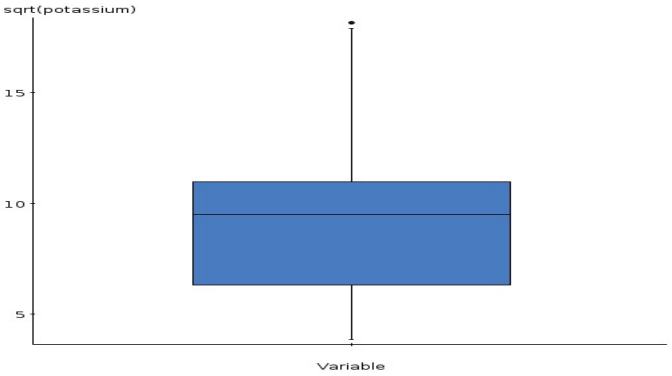
Cereal potassium data: log was good, but can we do better? Look at boxplots (read down to go down ladder)



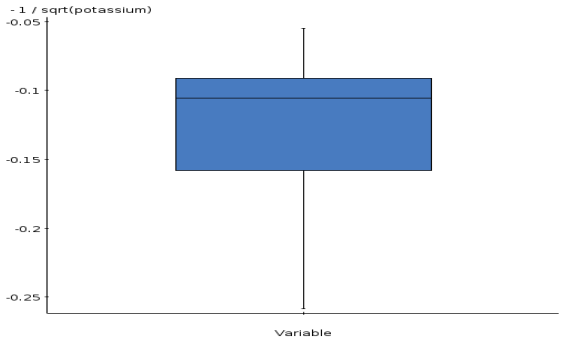
original data (1)



log (0)



square root (0.5)

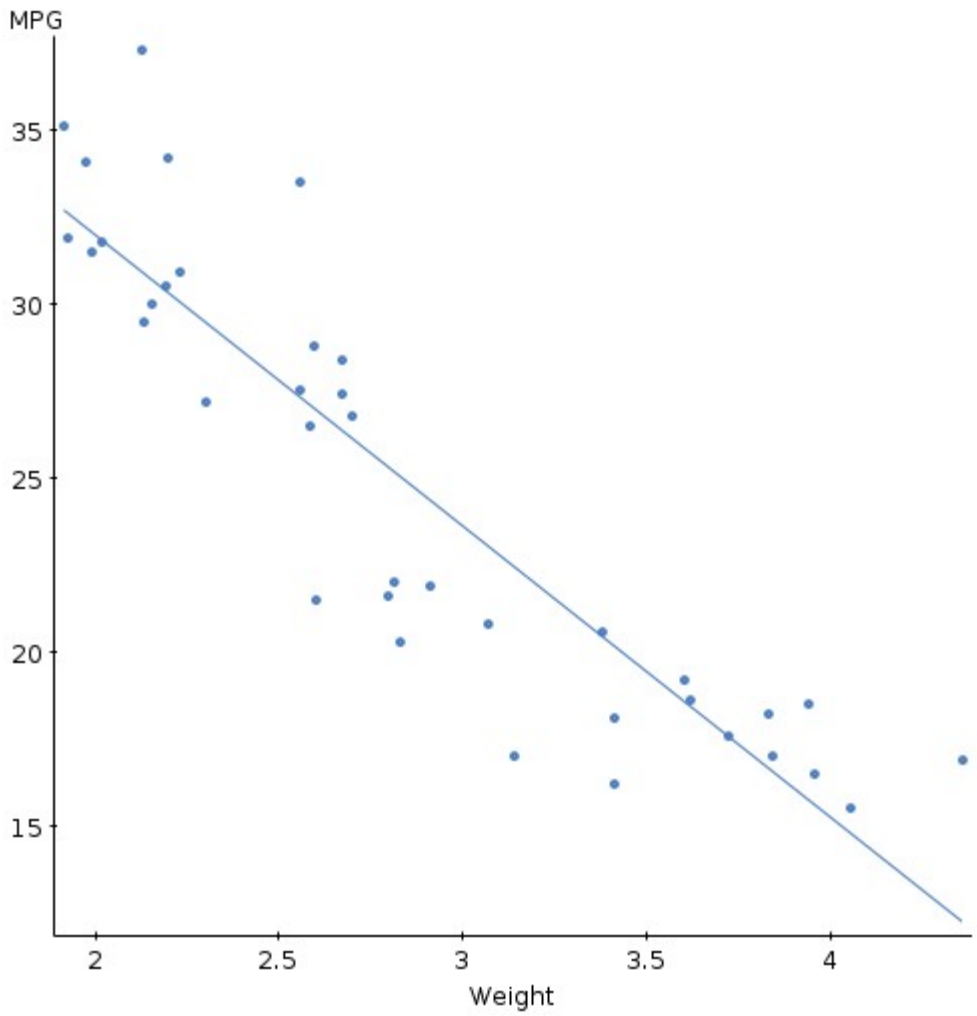


-1/sqrt (-0.5)

Where on the ladder of powers should we be to make the shape symmetric?

Power 0 or logarithm

Car MPG vs weight

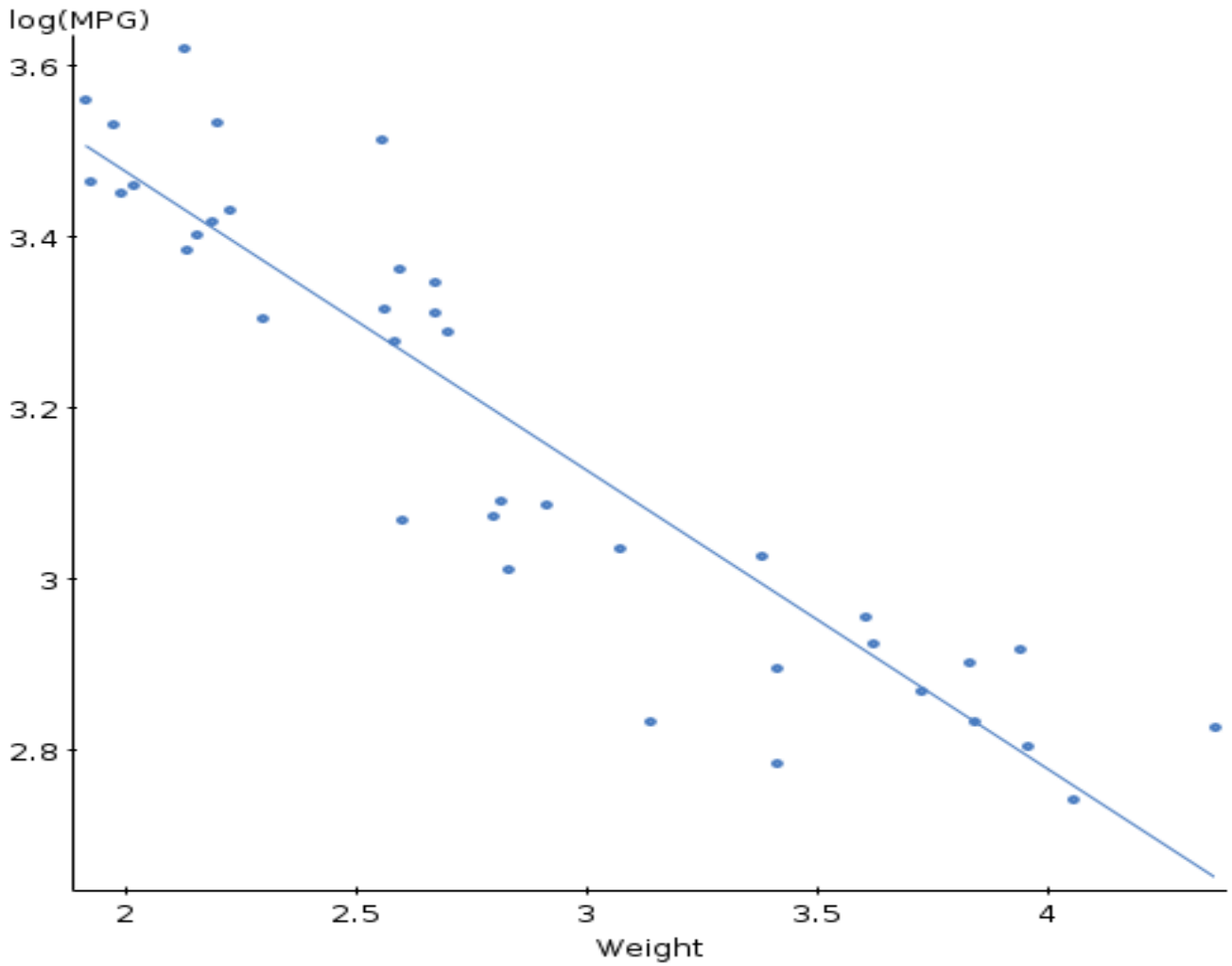


Regression of car mpg vs weight:

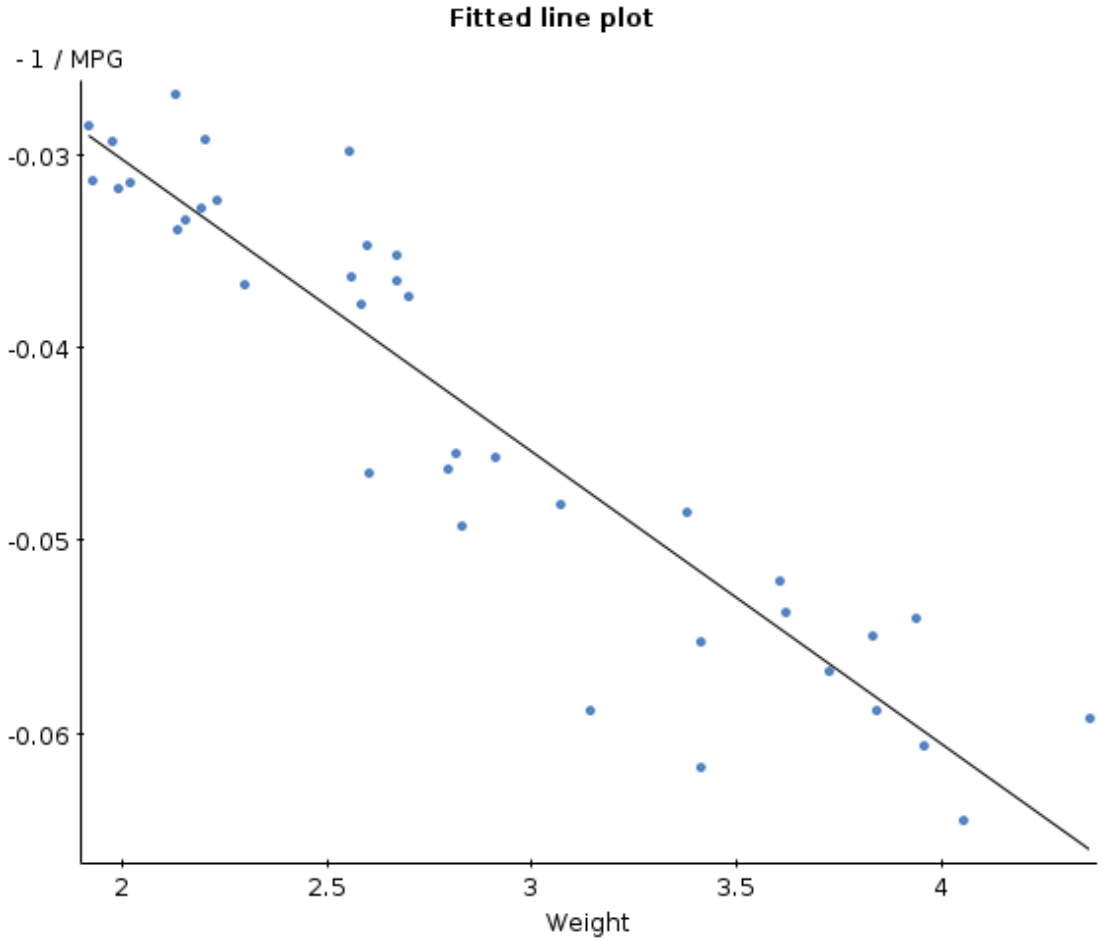
- Relationship looks curved (shows up on residual plot). Also logically mpg cannot go below zero.

- Try log of mpg.

log (power 0)



– Still curved (also on residual plot). Must go further. Try -1 power (negative reciprocal).



Which would you choose? Why?

_*We also tried power -2*

*and it was no better than power -1, so stick with power -1, "gallons per mile"*_____

– The original regression equation is

$\widehat{mpg}$ =48.7 - 8.365 Weight

For a car of weight 6 tons, predicted mpg was -0.89. Does this make sense?

– *_No, since MPG cannot be negative._____*

Using the -1 power

- Using the -1 power, fit again to get

$\widehat{-1/mpg}$ = 0 – 0.015 Weight

What is predicted mpg for car of weight 6 tons? Does this make some kind of sense?

- *Predicted negative gallons per mile is 0-(0.015)(6)=-0.090*
- *Solve -1/MPG=-0.090: MPG =1/0.090=11.11*
- *yes; plausible value for MPG.*

Chapter 11: Understanding Randomness (p.300)

- What does “random” mean?
- These games are random:
  - toss a fair coin, win if you get a head.
  - roll a fair six-sided die: win if you roll a 6.
  - play roulette at a casino: win if “red” comes up.
- In short term, *unpredictable*
- In long run, *predictable*:
  - coin: should win about ½ of the time.
  - die: should win about 1/6 of the time.
  - roulette: 18 of 38 numbers are red, so should win 18/38 of the time.
- Computer random numbers generated by non-random method (!) but look just like random numbers.

Random digit tables

- Easiest way to do simulation is with computer (eg. StatCrunch).
- Otherwise, can use tables of random digits 0-9: equally likely to be each digit, but next digit unpredictable, eg:

  24655 67663 61607 42295

  14635 62038 40528 12195

  85757 38452 76349 78850

Simulation: the dice game of 21

- Played with ordinary 6-sided die
- Each player keeps rolling die, totalling up spots, until:
  - the player decides to stop
  - the player's score goes over 21, in which case the player loses.
- Player with highest score of 21 or less is winner.

Suppose you are playing one opponent, who scores 18. How likely are you to win? (Strategy: keep rolling until you win or go over 21.) Use random number table to simulate games.

Use table thus: random digit of 1—6 is die roll, 7—9 and 0 ignored. Target: beat 18.

24655 67663 61607 42295

– 14635 62038 40528 12195

– 85757 38452 76349 78850

Trial 1		Trial 2		Trial 3	
Random	Total	Random	Total	Random	Total
2	2	6	6	6	6
4	6	7		1	7
6	12	6	12	6	13
5	17	6	18	0	
5	22	3	21	7	
	lose		win	4	17
				2	19 win

And so on. Here, won in 2 out of 3 simulations, but would do many more to get accurate view of how likely we are to win. I did 1000 simulations, and found the second player to have about a 72% chance of winning.

Here's a bigger table, all done by simulation:

First player scores	Second player wins
15	100%
16	95%
17	86%
18	72%
19	52%
20	29%
21	0%

A more interesting question is when the first player should stop. If she stops once she gets to 16, she cannot go over 21, but she is very likely to lose. If she aims for 20, she is likely to win if she gets 20 (or 21), but she is very likely to go over 21 trying.

(What looks like a good target for the first player to aim for?)

First player strategy

- Several strategies for 1st player: “stop at 16 or more”, “stop at 17 or more”,..., “stop at 20 or more”, “stop only at 21”.
- Do simulations of entire game, using each of these strategies for 1st player and “win or bust” strategy for 2nd. See how often each player wins. (2nd player automatically wins if first player goes over 21.)
- My results:

1st player's strategy	1st player wins	2nd player wins
Stop at 16 or more	29%	71%
Stop at 17 or more	39%	61%
Stop at 18 or more	48%	52%
Stop at 19 or more	50%	50%
Stop at 20 or more	45%	55%
Stop only at 21	29%	71%

A simulation is not *the* answer

- Because it's based on random numbers (and random numbers can be weird), a simulation won't be completely accurate.
- Later, see it's possible to get exact answers in some cases, or to use approximations that are more accurate than simulation.
- Simulation is cheap and (fairly) easy, so do lots of trials.
- With lots of trials, answers from simulation will be accurate enough to give a good idea.

The birthday month problem

Ask people one by one which month their birthday is in. How many people might you have to ask to find two people with their birthday in the same month?

To simulate:

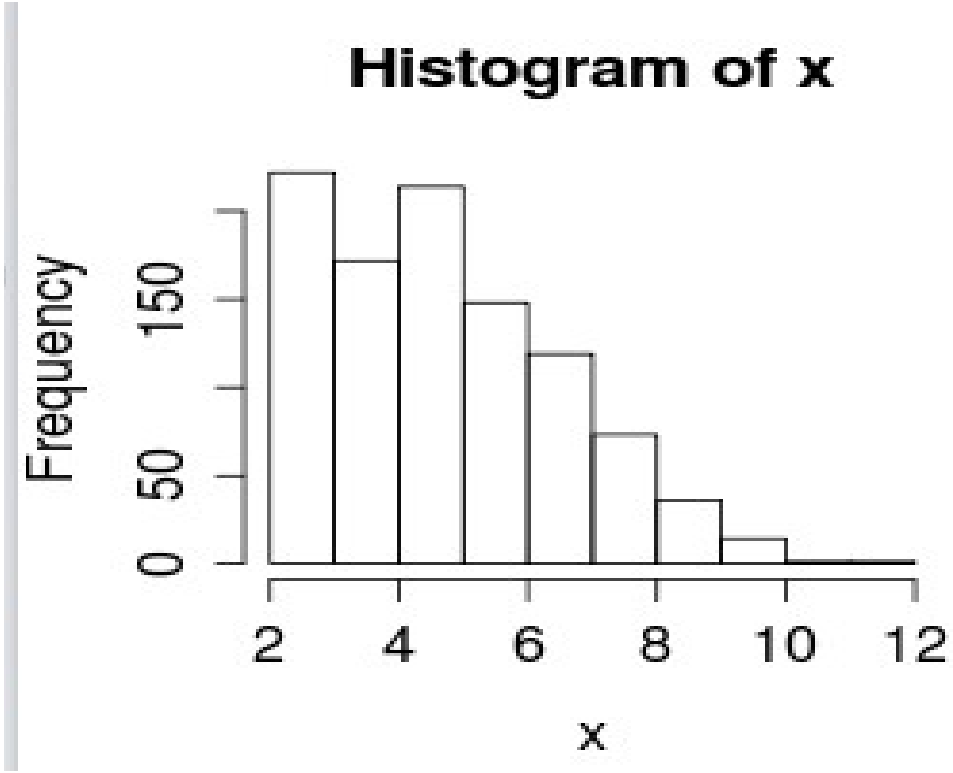
- arrange the months in a column
- sample, say, 20 months with replacement (2 people can have birthday in same month)
- count down columns until you find a repeated month.

Do by hand in StatCrunch.

- sample a bunch of months with replacement
- see how many simulated "people" are needed until a month appears for the second time

Summary of birthday-month results (from R)

- Having to ask a lot of people is possible but unlikely
- Distribution is skewed to right
- Most of the time, we will only have to ask 2 or 3 or 4 people (surprising?)



Chapter 12: Sample surveys (p.314)

See this kind of thing all the time:

- a survey asking "if there were an election for mayor tomorrow, which candidate would you vote for?"
- with results based on responses from maybe 1000 people, claimed to be accurate "to within 3 percentage points 19 times out of 20".

How is that done, and why?

What would happen if we tried to survey everybody?

Examine a part of the whole (p.315)

- Population = everyone we want to investigate.
- Need to use a *sample* that *represents* population (is like it in all important ways).
- Imagine radio call-in poll about highway tolls. What kind of people might call in? Is that likely to be a representative sample?
  - *People with strong opinions; not at all representative.*
- A sample that over-represents or under-represents some part of population called *biased.* Conclusions from biased sample cannot be trusted.

How might we select a representative sample?

1. Carefully select individuals to *match* the population in every way we can think of, eg. For highway tolls issue:
    - Males and females
    - the right mix of ages
    - right number of people living in each city/rural area
    - right mix of political opinions
    - etc, etc.

   – Difficult to do.

   – Might miss important way of matching population.

2. Select individuals **at random**.
    - Easy to do
    - A*pproximately* represents population in all ways, *including ones you didn't think of.*

Why does randomization work?

- Short term unpredictable, long term predictable
- *Cannot predict* which individuals are going to end up in sample
- With a large sample, sample will have approximately right proportion of males/females, urban/rural, old/young, etc., *and anything else we didn't think of*.

Do you have enough noodles in your soup?

- stir soup, take (random) sample. Does *that* have enough noodles?
- doesn't matter how much soup you're cooking, as long as you stir it (population size doesn't matter)
- the bigger your *sample* of soup, the better your estimate of how much noodles it has.
- but if you sample *too much* soup, none left for your guests!

Three keys for sampling (p.315--317):

1. Examine a part of the whole (sample)
2. Randomize (to obtain the sample)
3. It's the sample size (that is important).

Populations and parameters, samples and statistics (p. 318)

- Suppose we want to know what proportion of the population of the GTA are in favour of highway tolls.
- This is the *population parameter.* What we want, but unknown (except by asking everybody). Notation p .
- Take a sample, calculate proportion in favour in your sample. *Sample statistic.* Easy to figure out, but not the thing we want. Notation $\hat{p}$.
- Hope that sample statistic close to population parameter. If sample drawn using randomization, can work out how close (later).

How can we draw a sample?

- *Simple random sample* (p.319): put names of all members of population in "hat", shake, and draw out one by one without looking.
- Every member of population equally likely to be selected, *independently* of who else in sample.
- Every possible *sample* equally likely.

Need to have a list of whole population (sampling frame).

Drawing a simple random sample using random digit table

Suppose we have a population of 80 students, numbered 01—80, want a simple random sample of 6 of them. Use these random digits: 43623 33434 94776 15780 95603 64962 46971 95188.

43, 62, 33, 34 all ok

34 is a repeat: reject

94, 77 too big, reject

61, 57: ok

so sample is students numbered 43, 62, 33, 34, 61, 57.

Stratified sampling (p.321)

- Population is in groups that could be quite different from each other (in terms of what's being measured)
- Take a simple random sample from each group, and combine to make overall sample.
- Why is this good?
  - *Fair representation of all parts of population.*
  - *Therefore sample statistic should be closer to population parameter.*

Example of stratified sampling

Back to population of 100 students. 60 of them female, 40 male. Suppose issue is "do you plan to try belly-dancing in next year?" None of males will, but 50% of females will (thus 30% of whole population). Sample of size 10.

With simple random sample, might get a lot of females and over-estimate interest in belly-dancing. Eg. 8 females and 2 males, what is sample proportion likely to be?

50% of 8 + 0% of 2 = 4+0=4, 40%

Or might have 2 females and 8 males. What is sample proportion likely to be?

50% of 2 + 0% of 8 =1+0=10%

In stratified sample, *guaranteed* to have 6 females and 4 males in sample. So sample proportion should be close to 30%.

Drawing a stratified sample

Population of 100 students: 01-60 female, 61-99 and 00 male.

Use random digits: 18406 28903 75909 66389 28937 46983 49652 37406 .

Draw stratified sample of 6 females and 4 males.

18, 40 females (2 so far)

62, 89 male (2 so far)

03 female (3 so far)

75, 90 male (now 4 males, don't sample any more)

96, 63, 89 reject (would be more males)

28 (4^{th} female), 93, 74, 69, 83 (reject)

49, …, 23 (last 2 females).

Cluster and multistage sampling (p.322)

How would you randomly sample 100 words from the textbook?

- simple random sampling: number every single word (!) and then sample from them.
- easier: randomly sample 10 *pages* first, then randomly sample 10 words on each page. Why is that easier to do?
- not same as simple random sampling: if you select a particular page, other words on the same page *more likely* to be in sample.
- Called *cluster sampling,* or *two-stage sampling.*

Multistage sampling

- Often hierarchy of clusters eg. chapter – section – sentence – word, and could choose:
  - chapters
  - section within chosen chapter
  - sentence within chosen section
  - word within chosen sentence

Called *multistage sampling.* At each stage, choice made by simple random sampling.

Choose cluster/multistage sampling for *convenience*, choose stratified sampling for *accuracy*.

Things that can go wrong with sample surveys

- not getting Who you want (nonresponse): what to do?
  - _____
- getting the question(s) right (how to ask)
  - _____
- getting open-ended response
  - _____
- sampling volunteers (how might that happen?)
  - _____
- sampling badly but conveniently
  - *see above (Who)*
- undercoverage
  - *not being able to sample certain parts of population.*

– *Example:* _____

Chapter 13: Experiments and Observational Studies (p. 341)

How do you find out if exercise helps insomnia?

- look at a bunch of people, find out if they exercise and how much, ask them to rate their insomnia.
- Suppose the people who exercise more suffer less from insomnia. Can you conclude that people who suffer from insomnia should be recommended to exercise?

  - *No: people who exercise more might have other health benefits that help with insomnia.*

- This kind of study called *observational study.*
- Assesses association but not cause and effect (like correlation).
- Why not?
  - _____

Observational studies (p. 342)

- Commonly used
- May help identify variables that have an effect
  - but may not identify the most important ones.
  - How do we know there is nothing else that makes a difference?
- *Retrospective* study "looking back", like one above:
  - measure exercise and insomnia from historical records.
- *Prospective* study "looking forward":
  - identify subjects in advance, collect data as events happen.
- Are data from the past even accurate?

Which is better, retrospective or prospective?

- prospective: Outcome needs to be a common one (why?). Takes a long time.
- retrospective: easier to obtain enough data for rare events. Takes less time to do. More concerns about bias/confounding: how do we know past data is accurate/complete/relevant?

Experiments (p. 343)

How do we establish cause and effect?

- need to randomly choose some subjects and instruct them to exercise
- the other subjects are instructed *not* to exercise
- assess insomnia for all subjects.

Why is this better? How does it level out effects of other variables?

- *choosing two groups at random means evening out effects of all other variables*
- *if the groups end up unequal in terms of insomnia, it must have been the exercise that made the difference.*

Terminology

- People/animals/whatever participating in experiment called *experimental units / subjects.*
- Experimenter has at least one explanatory variable, a *factor*, to manipulate.
- At least one *response variable* measured.
- Specific values chosen for factor called *levels.*
- Combination of manipulated levels of factors called *treatment*.

Variation of exercise/insomnia experiment: add diet

- three kinds of exercise: none, moderate, strenuous
- two different diets: fruit/veg, “normal”
- factors are:
  - exercise, with 3 levels
  - diet, with 2 levels
- 3 x 2 = 6 treatments (6 combinations of 2 factors)
- divide subjects into 6 groups at random.

Principles of experimental design (p. 345)

1. Control
   - control experimental factors by design
   - control *all* other variables by randomization
2. Randomize
   - "control what you can, randomize the rest"
3. Replicate
   - get many measurements of response for each treatment.
4. Blocking
   - divide experimental units into groups of similar ones and sample appropriately (compare stratified sampling)

Blocking (p. 346)

Suppose you have 8 six-year-old girls and 2 ten-year-old girls who want to play soccer. How would you divide them into two teams?

- idea 1: use randomization to decide who goes on which team.
  - but: what if the two older girls end up on same team? Is that fair?
- idea 2: block A: the ten-year-old girls. Block B: the six-year-old girls. Choose at random one girl from block A and four from block B for each team.
- If some experimental units are different from rest, arrange in blocks so that units within block similar, in different blocks different. Then share out units from different blocks among different treatments.

How do we know a treatment is effective? (p. 349)

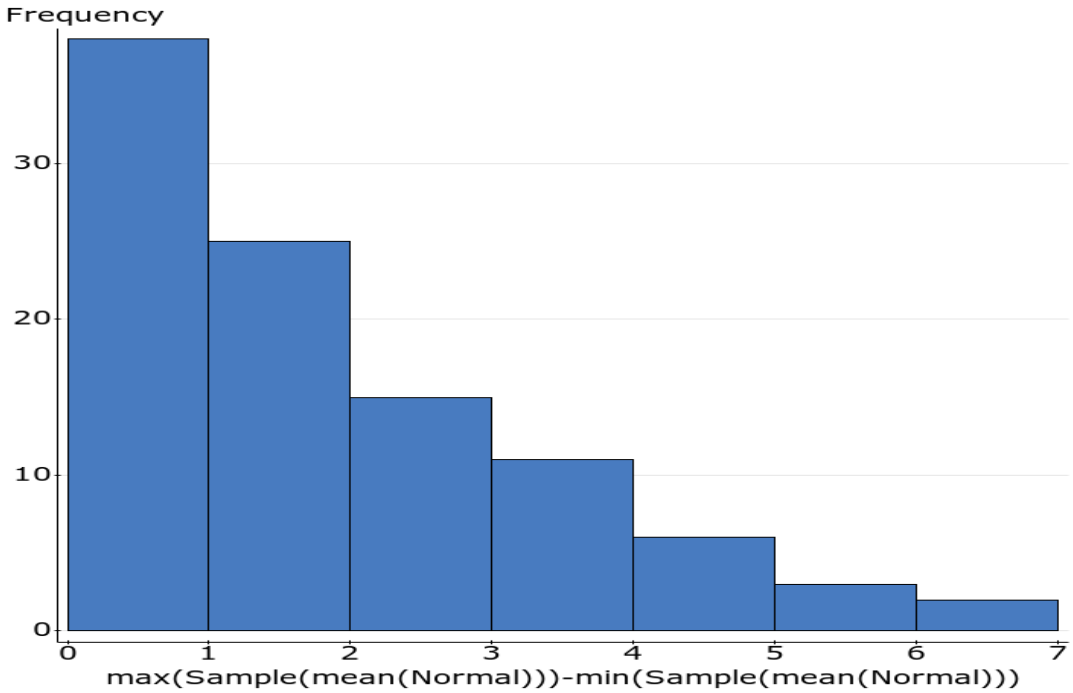
- If the treatments are equally good, will the means for each treatment group be *exactly* the same? Why or why not?
  - *Probably not; randomized groups won't be exactly same*
- If the mean for treatment 1 a lot bigger than mean for treatment 2, is that evidence of a difference between treatments? Why?
  - *Yes; the groups were (approx) equal by randomization, so outcomes should not be* very *unequal, if the treatment has no effect. Therefore we conclude the treatment does have an effect.*
- How big is "big"?
  - Suppose two treatments are *equally good.* How big a difference in treatment means might we see, just by chance (due to randomization)?

– simulation: 2 treatments, both mean 20, SD 5, 10 subjects for each. How far apart could the means be?

Statcrunch:

- Data, Simulate, Normal.
- Rows 10 (sample size 10), columns 100 (“many”).
- Fill in mean 20, SD 5 of normal data.
- Click Compute for Each Column, and type “mean(Normal)” into the box.
- Compute. Should get a column of 100 sample means.
- Want to pick 2 at random, see how different they are, repeat many times. Data, Sample, select column mean(Normal).
- Sample size 2, number of samples 100 (“many”).
- Click Sample With Replacement.
- Click Compute Statistic for Each Column, type or build max("Sample(mean(Normal))")-min("Sample(mean(Normal))"). Compute.
- Resulting column is kind of difference in means to expect. See

histogram below.



- *mean difference could be eg. 3 or 4 just by chance.*
- What does it mean if we see a difference bigger than that in our actual experiment?

The treatment really makes a difference

Placebos (p. 352)

- A placebo is a "fake" treatment designed to look like a real one.
- Why is that important?
  - Known that receiving *any* treatment will cause a subject to improve.
  - Want to show that the "real" treatment is not just effective, but better than a placebo. Then have evidence that the treatment is worth knowing about.
- Can also use current standard treatment to compare with.
- Subjects getting placebo/standard treatment called *control group.*

Blinding (p. 352)

- Suppose you participate in an experiment to see if a new herbal remedy for common cold really works.
- If you knew that you got the placebo, would that influence your recovery?
- If you knew that you got the herbal remedy, would that influence your recovery?
- Best if:
  - *you* don't know what you're getting
  - *the experimenter* doesn't know what you're getting: **Double-blind** so as not to bias results.
- In practice, design placebo to look just like herbal remedy, and label with eg. reference number so that no-one knows until after data analyzed which is which.

The best experiments (p.353)

are:

- randomized
- comparative
- double-blind
- placebo-controlled.

More factors (p. 355)

Recall (revised) insomnia experiment:

- three kinds of exercise: none, moderate, strenuous
- two different diets: fruit/veg, “normal”
- have subjects on all 6 combos of exercise/diet
- analysis tells us whether either (or both) of these variables have an effect on insomnia.

Suppose:

- group 1: all the subjects on no exercise were also on normal diet
- group 2: all the subjects on moderate/strenuous exercise were on the fruit/veg diet.

If group 2 comes out better, *cannot tell* whether exercise or diet deserve credit: exercise, diet *confounded (p. 356).*

Ethical experiments (p. 357)

Idea of *imposing* treatments on subjects might be questionable:

- what if study effects of smoking on lung disease?
- would have to prevent some subjects from smoking, and *make some subjects smoke* for duration of study (!!!)

There are some known unhealthy/dangerous things you cannot ask subjects to do. Also,

- giving a placebo when a best proven treatment is available *is not ethical.*
- subjects who receive placebo *must not be subject to serious harm* by so doing.

See **Declaration of Helsinki**, which governs experiments on human subjects:

www.wma.net/en/30publications/10policies/b3/index.html

Chapter 14: From randomness to probability (p. 376)

In chapter 11, thought about randomness:

- These games are random:
  - toss a fair coin, win if you get a head.
  - roll a fair six-sided die: win if you roll a 6.
  - play roulette at a casino: win if "red" comes up.
  - go into Tim Horton's, win if they have your favourite donut.
- In short term, *unpredictable*
- In long run, *predictable*:
  - coin: should win about ½ of the time.
  - die: should win about 1/6 of the time.
  - roulette: 18 of 38 numbers are red, so should win 18/38 of the time.

- Tim Horton's? Don't know, but there *is* a long run prob (of having your favourite donut).

Empirical probability (p. 377)

- Observe your game many times. (Reality/thought).
- Long-run fraction of times you win: *probability* of winning.
- Each single play you win or not, but probability guides actions.

Game: roll a (fair) die, win $3 if you roll a 6, lose $1 otherwise.

- prob. of winning 1/6 (prob. of losing 5/6)
- play game 6 times, expect to win once, lose 5 times
- total winnings 1 x $3 plus 5 x -$1, total -$2.
- *on average* expect to lose
- *but:* unpredictable in short term, might be lucky enough to win (eg if roll 6 first time).

Terminology

- collection of all things that can happen to you: *sample space.*
  - toss a coin once, sample space is?
  - roll a die once, sample space is?
- one of things in sample space called *outcome.*
  - one possible outcome of tossing a coin is a Head.
- one or more outcomes together called *event.*
  - roll a die, “even number of spots” an event, consisting of outcomes 2, 4, 6.
- each time we observe random event called *trial.*
  - rolling die to see how many spots come up is a trial.

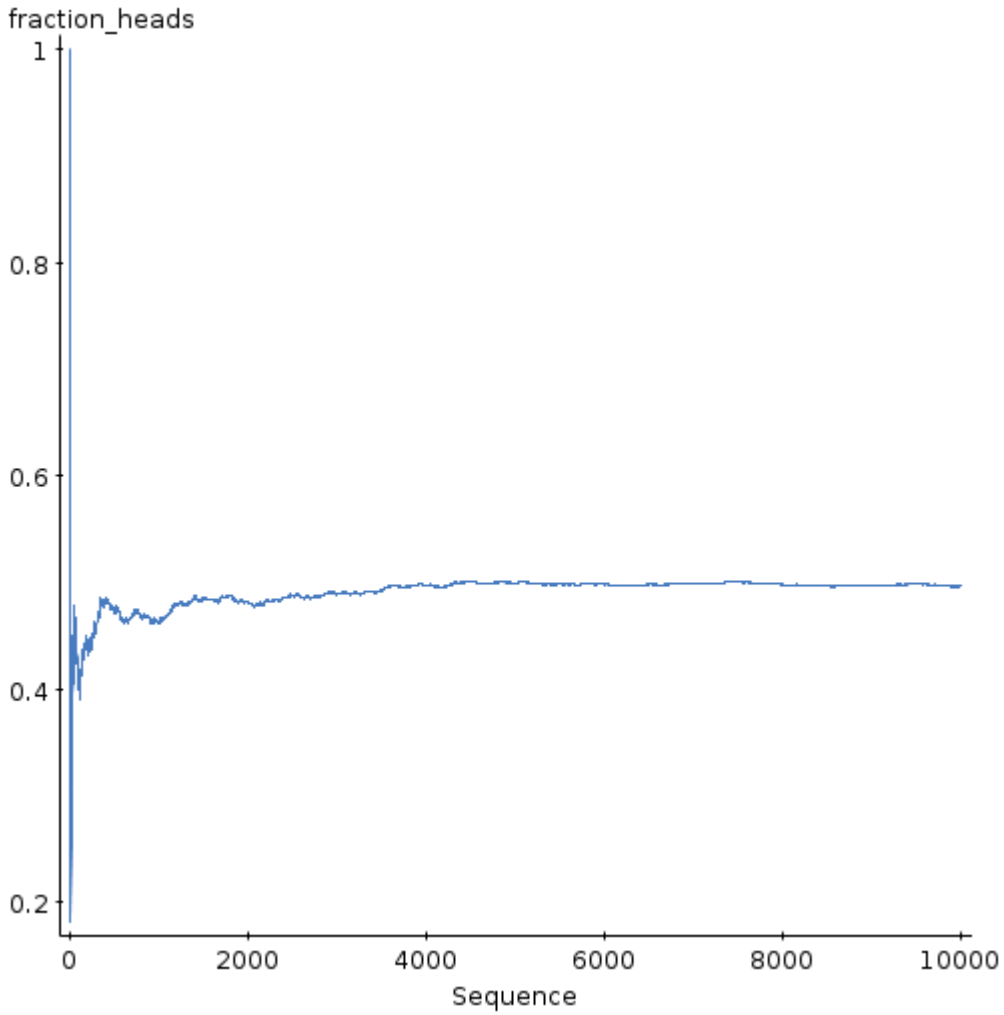
Law of Large Numbers (p.378)

- how do we know that the relative frequency of some outcome actually *will* settle down to one value?
- Law of Large Numbers (mathematical fact): relative frequency of some outcome has a limit as number of trials becomes large
- Or, with many trials, the relative frequency and probability will be approximately equal.

Look at my StatCrunch report "Law of Large Numbers for Probability".

Empirical probability of a Head getting close to 0.5

- Toss a coin 10,000 times
- Keep track of proportion of heads *so far*.
- That proportion appears to be getting closer to 0.5 as #tosses gets bigger.



"Law" of averages (p. 378)

- toss a coin, get 5 heads in a row: "due" a tail?
- play casino game, lose 6 times in a row. "Due" a win?
- *No!*
- that would require coin/casino to remember previous results!
- you have the same chance of winning every time, *independently* of what happened before.
- How does that square with law of large numbers?
  - law of large numbers says *nothing* about short-term.
  - short-term unpredictable, long-term predictable
  - if a short run of coin-tosses has a lot of heads, the long run that follows will have about 50% heads. The overall average dominated by long run, so will be about 50% heads also.

Theoretical probability (p. 380)

- Sometimes can argue (in a mathematical model) what probabilities should be:
  - toss a coin: two faces of a coin are just the same, so coin should be equally likely to land heads or tails, eg. P(H)=1/2.
  - roll a die: in theory it is a perfect cube, so each of the 6 faces equally likely to be uppermost: eg. P(6)=1/6.
  - Draw a card: ¼ of the cards are spades, but each *card* equally likely.
  - more generally, any time you have equally likely outcomes, prob. of event A is

$$P(A)=\frac{\text{outcomes in A}}{\text{total outcomes}}$$

Rolling red and green dice and getting 10 spots total

- Sample space, red die first:
  - S={(1,1),(1,2),...,(1,6),(2,1),...,(6,6)}
  - all 36 possibilities equally likely (each prob $\frac{1}{36}$)
- Which of those possibilities add up to 10?
  - *4 and 6; 5 and 5; 6 and 4;*
- How many of them are there?
  - *3 possibilities*
- So what is (theoretical) probability of total of 10?
  - *3/36=1/12*

Personal (subjective) probability (p.381)

- What is probability that it will rain tomorrow? How does weather forecaster get “40%”?
- Forecaster uses experience to say that in “similar situations” in past, it's rained about 40% of the time.
- Personal probabilities not based on long-run behaviour or equally likely events. So treat with caution.

Probability rules (p. 382)

1. A probability must be between 0 and 1 (inclusive).
2. Probability of the whole sample space is 1.
3. **Addition rule:** If events A and B have no outcomes in common (disjoint), prob. of either A or B is P(A)+P(B).

Roll a fair die once.

- S={1,2,3,4,5,6}, equally likely, so each prob is 1/6. Why would any other value be wrong?
  - *Only 1/6 six times will add up to 1*
- Let A be event "5 spots or more", A={5,6}, and B be event "2 spots or fewer", B={1,2}. What is prob that either A or B happens?
  - *P(A)=2/6; P(B)=2/6; P(either A or B)=2/6+2/6=4/6*

– Let A be event "5 spots or more"={5,6} as above, and C be event "even number of spots"={2,4,6}. Can you use the addition rule to find the prob of "either A or C"? What happens if you do? What is the correct answer?

try: P(either A or C): P(A)=2/6; P(B)=3/6; P(A or B)=2/6+3/6=5/6

But: "either A or C" means 2, 4, 5 or 6: P(A or B)=4/6?????

Why the difference? Outcome 6 in both events; counted twice; should not use addition rule; trust 4/6.

Probability of "not A"

To get the probability of an event A *not* occurring, written A^c , use rule $P(A^c)=1-P(A)$.

- for our A={5,6} above, $P(A^c)=1-2/6=4/6$.
- Does this make sense?
  - outcomes in "not A" are 1, 2, 3, 4
  - P(not A)= 4/6 ; check

General addition rule

- When two events A and B *not* disjoint (have outcomes in common), how to find P(either A or B or both)?
- found above that 1st addition rule gives answer that is too big. Fix up:
  - P(A or B or both)=P(A)+P(B)-P(both A and B)
- above: events A={5,6}, C=even number={2,4,6}
- Which outcomes make A and C both happen? What is P(both A and C)?
  6 ; 1/6
- Find P(A or C or both).
  - *P(A)+P(C)-P(A and C)=2/6+3/6-1/6=4/6 check.*

Note that if the two events are disjoint, P(A and B)=0.

Law and business school acceptance and rejection, again.

Law	acc.	rej.	total		Business	acc.	rej.	total
males	10	90	100		males	480	120	600
females	100	200	300		females	180	20	200
total	110	290	400		total	660	140	800

- How likely is an applicant to law school to be accepted overall?
  - 110/400=0.28
- How likely is a *female* applicant to law school to be accepted?
  - 100/300=0.33

Second answer is *conditional probability*: we *know* the applicant is female, so only look at females.

Notation:

- let F be event that applicant female
- let C be event that applicant aCcepted
- we found P(C|F): accepted *given* female (0.33)

Not the same as P(F|C): now I know the applicant was accepted. P(F|C)=100/110

How to find a conditional probability

$$P(B|A)=\frac{\text{P(both A and B)}}{P(A)}$$

- try with above example:
  - Want P(C|F)
  - P(both C and F)=100/400 (both female and accepted)
  - P(F)=300/400
  - so P(C|F)= $\frac{100/400}{300/400}=1/3$.

400's cancel out and answer is 100/300.

General multiplication rule

Turn conditional prob rule around:

$$P(\text{both A and B}) = P(A) \times P(B|A)$$

Prob that law school applicant is both male and rejected? Let A=male, B=rejected. Then

- P(A) = 100/400
- P(B|A) = 90/100
- so P(both A and B)= $\frac{100}{400} \times \frac{90}{100} = \frac{90}{400}$.
- also see from table: out of 400 applicants, 90 of them were males who were rejected.

Another example

Law	acc.	rej.	total		Business	acc.	rej.	total
males	10	90	100		males	480	120	600
females	100	200	300		females	180	20	200
total	110	290	400		total	660	140	800

- Randomly select *two* male applicants to law school. What is probability that they are *both* rejected?
  - *R1 event "1st one rejected"*
  - *R2 event "2nd one rejected"*
  - *P(R1 and R2)=P(R1) x P(R2|R1)=90/100 x 89/99*
  - *(we know the first male applicant was rejected)*

… and another:

Law	acc.	rej.	total		Business	acc.	rej.	total
males	10	90	100		males	480	120	600
females	100	200	300		females	180	20	200
total	110	290	400		total	660	140	800

For a randomly chosen applicant **to business school**, what is (a) probability that that person is either male or accepted? How does that relate to (b) the probability of being female *and* rejected at business school?

- (a) define M=male applicant to bus school, A=accepted at bus school; want P(M or A or both)= P(M)+P(A)-P(M and A)= 600/800+660/800-480/800=780/800

- (b) 20/800

- how is (b) related to previous answer?

(b) is "not (a)"

– If two events A, B are disjoint, they *cannot both happen.*
– Suppose A happens, then P(B|A) **must be 0**, whatever P(B) is.

– Suppose now C and D are independent events.
– Then P(D|C) **equals** P(D): knowing about C *makes no difference.*
– Also, then P(D and C)=P(D) x P(C) ("simple multiplication rule").
– Some examples on next page.

Examples of independence and disjointness

Suppose you are selected to take part in an opinion poll. Which of the following are independent, disjoint, or neither?

- A=your telephone number is randomly selected; B=you are not home when they call.
  - *independent*
- A=as selected subject, you complete the interview; B=as selected subject, you refuse to cooperate.
  - *disjoint*
- A=you are not home when they call at 11:00am; B=you are employed full time outside the home.
  - *Not independent.*

Turning conditional probabilities around

Suppose a restaurant has two (human) dishwashers. Alma washes 70% of the dishes, and breaks (on average) 1% of those. Kai washes 30% of the dishes, and breaks 3% of those. You are in the restaurant and hear a dish break at the sink. What is the probability that it was Kai? Answer over.

Even though Kai washes less than half of the dishes, when a dish breaks, it is more likely than not that Kai broke it.

Ken's easy way: make a contingency table (p. 405). The totals can be anything: I often choose 100 or 1000 for grand total.

	Dishwasher		
Dish breaks	Kai	Alma	Total
yes	9	7	16
No	291	693	984
Total	300	700	1000

*0.01*700=7; 0.03*300=9*
P(Kai|dish breaks)=9/16

Another one:

Three different airlines A, B, C operate night flights from LA to NY. On their night flights, airline A takes off late 40% of the time, B 50%, and C 70%. My travel agent books me on a night flight from LA to NY at random (equal prob. for the three airlines).

1. what is prob. that I'm on airline A and late taking off?
2. What is probability that I'm late taking off?
3. I was late taking off. What is the prob. that I was booked on airline A?

Make a table, pretending there are 300 flights altogether:

	Late taking off	On time	Total
Airline A	40	60	100
Airline B	50	50	100
Airline C	70	30	100
Total	160	140	300

1. 40/300

2. 160/300 =0.53

3. 40/160 =0.25

"At least one"

Suppose I buy a lottery ticket each week for 3 weeks. Each ticket has probability 0.1 of winning a prize each week, independently of other weeks. What is the probability that I win at least one prize?

P(not winning at all)=(1-0.1)x(1-0.1)x(1-0.1)=0.9*0.9*0.9=0.73

P(win at least once)=1-0.73=0.27

How does that probability change if I buy one ticket a week for 26 weeks?

P(no wins)=(1-0.1)^26=0.06 P(at least one win)=0.94

What if my winning chances are 0.05 for the first week, 0.10 for the second, 0.15 for the third week?

Same idea: P(no wins)=(1-0.05)(1-0.10)(1-0.15)=0.73
P(at least 1 win)=1-0.73=0.27

Binge drinking

44% of university students engage in binge drinking, 37% drink moderately, and 19% don't drink at all. Among binge drinkers, 17% have been involved in an alcohol-related car accident, among moderate drinkers, 9% have, and among non-drinkers, 0% have.

If a student has a car accident, what is the probability that they were a binge drinker? (over)

Make a table. Pretend 100 students altogether:

	Alcohol-related accident	Not	Total
Binge drinker	7.48	36.52	44
Moderate	3.33	33.67	37
Non-drinker	0	19	19
Total	10.81	89.19	100

44*.17=7.48; 37*.09=3.33

Know that a student was in an accident (1st column); out of all students in 1st column, how many were binge drinkers? 7.48/10.81=0.69.

Moral: only 44% of all students are binge drinkers, but those drinkers make up 69% of accidents.

Chapter 16: random variables (p. 422)

Sometimes events have numbers attached to them:

- count how many heads in 3 coin-tosses
- total number of spots when you roll 2 dice
- the most cards of the same suit in 13 cards from a deck
- how much you might claim in a year on a car insurance policy

These numbers called **random variables**.

Probability distributions

Each value of a random variable is an event, so each value has probability. List of values and probabilities called *probability model*.

Tossing 3 coins (this comes from *binomial distribution*, later):

# heads	0	1	2	3
Prob.	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Rolling two dice:

Spots	2	3	4	5	6	7	8	9	10	11	12
Prob.	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Combining values of random variable:

3 coins:

# heads	0	1	2	3
Prob.	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

- How likely are we to get two or more heads?
  - *add up probs:* *3/8+1/8=4/8=1/2*
- How likely to get at least one head? (2 ways)
  - *Use P(0 heads):* *1-1/8=7/8*
  - *or directly:* *3/8+3/8+1/8=7/8*
- What do all the probabilities add up to? Does this make sense?
  - *1/8+3/8+3/8+1/8=1: one of those # heads must happen*

The mean of a random variable (p. 423)

Here's a random variable, called X:

Value of X	2	3	4	5
Probability	0.1	0.2	0.4	0.3

- Mean *not* (2+3+4+5)/4=3.5 because 4 and 5 more likely than 2 or 3.
- Have to account for more likely values when adding up:
  - times each value by probability:
  - 2(0.1)+3(0.2)+4(0.4)+5(0.3)=0.2+0.6+1.6+1.5=3.9
  - (Weighted average, weights sum to 1.)

- Median of random variable is value of X where summed-up probabilities first exceed 0.5: 3 too small (total 0.1+0.2=0.3), 4 is right (0.1+0.2+0.4=0.7), so median 4.

- Mean a little smaller than median: left-skewed.

SD of a random variable (p. 425)

Another probability distribution:

Value of Y	3	4	5
Probability	0.1	0.8	0.1

- Mean is 4: why?
- Procedure for SD:
  - subtract mean from each possible value
  - square
  - *times each result by its probability*
  - add up: gives variance; square root to get SD
- Here: variance

$$(3-4)^2(0.1)+(4-4)^2(0.8)+(5-4)^2(0.1)=0.1+0+0.1=0.2.$$

– SD= $\sqrt{0.2}=0.45,$ small since Y most likely 4.

Linear changes to a random variable (p.427)

- What does it mean to add *a* to a random variable? Multiply it by *b*?
  - Take all the *values* and change them, while leaving the probabilities alone.
  - Here's Y, with mean 4 and SD 0.45:

Value of Y	3	4	5
Probability	0.1	0.8	0.1

2Y looks like this. You check that mean now 8, SD 0.9.

Value of 2Y	6	8	10
Probability	0.1	0.8	0.1

and Y+3 as below. You check that mean now 7, SD 0.45.

Value of Z	6	7	8
Probability	0.1	0.8	0.1

Summary

- If you add a constant to a random variable, what happens to its mean? SD?
  - Mean of (X+a) = mean of X, plus a
  - SD of (X+a) = SD of X
- If you multiply a random variable by a constant, what happens to its mean? SD?
  - Mean of bX = b times mean of X
  - SD of bX = b times SD of X.

Two (or more) random variables (p. 429)

Suppose X is #heads when tossing a coin 5 times, and Y is #spots when rolling a die once. What can we say about total "score" X+Y, which is a random variable too?

Given: X has mean 2.5, SD 1.12; Y mean 3.5, SD 1.71.

- *Probability distribution* of X+Y is difficult to figure out.
  - Example: P(X+Y=3)? Work out possibilities: X=2, Y=1; X=1, Y=2; X=0, Y=3. Find prob of each, add up.
- *Mean of* X+Y easy to figure out:
  - Mean of (X+Y) is mean of X + mean of Y.
  - here: mean total score = 2.5+3.5=6.

SD of X+Y

- *If X and Y are independent:*
  - variance of (X+Y) = variance of X + variance of Y.
  - two random variables are independent if knowing about one tells you about the other. Here, knowing coin toss result tells you nothing about die roll, so our X, Y independent.
  - X has SD 1.12, Y has SD 1.71.
  - In example, variance of total score = variance of X + variance of Y = $1.12^2+1.71^2$ = 4.18.
  - So SD of total score is $\sqrt{4.18}$ = 2.04.

Odd fact: SD of X-Y is same as SD of X+Y

- above example: difference in scores, coins minus die, has mean 2.5-3.5=-1, variance $1.12^2+1.71^2$, SD $\sqrt{4.18}$ = 2.04.
- Suggests: will score more on die than on coins on average, but large SD says will at least sometimes score more on coins.
- How often? Hard, but easiest by *simulation:*
  - coin score has *binomial* distribution, n=5, p=0.5
  - die score has *uniform* distribution, a=1, b=6.
  - Compute difference.
  - results: diff greater than 0 _280__ times out of 1000
- difference in scores isn't normal, but pretend it is:
  - for difference of 0, z=(0-(-1))/2.04=0.49, prob of greater than 0 is 1-0.6879=0.3121.

Continuous random variables (p. 433)

- So far: our random variables *discrete:* set of possible values, like 1,2,3,... , probability for each.
- Recall normal distribution: *any* decimal value possible, can't talk about probability of any *one* value, just eg. “less than 10”, “between 10 and 15”, “greater than 15”.
- Normal random variable example of *continuous.*
- Finding mean and SD of continuous random variable involves *calculus :-(*
- but if we are given mean/SD, work as above (example over).

Handling two normal distributions

Betty and Clara go for a swim every morning. The times it takes each of them to complete their swim have (independent) normal distributions. Betty has mean 10 minutes and SD 2 minutes, and Clara has mean 11 minutes and SD 1 minute. How likely is it that Clara will complete her swim first?

Let B be Betty's time, and C be Clara's. Then Clara will finish first if the random variable C-B is less than zero.

- What are the mean and SD of C-B?
  - *mean is 11-10=1; variance is* $1^2+2^2=5$ *so SD =* $\sqrt{5}$ *=2.24.*
- Turn 0 into a suitable z-score and find the answer
  - *z=(0-1)/2.24=-0.45, prob is 0.3264.*

How do you find SD of sum and difference if random variables are not independent?

- In this course, you don't.
- See p. 436—437 of text for gory details.

Chapter 17: Probability Models (p.445)

A certain coffee shop has a Roll Up the Rim to Win promotion. 15% of all cups win a prize:

- How many cups are you likely to have to buy before you get your first prize?
- How many prizes might you win if you buy 10 cups of coffee?

Reasonable to assume that each cup is a prizewinner or not, independently of other cups. Call act of rolling up the rim on one cup a "trial".

Often encounter:

- Two possible outcomes "success" and "failure"
- Prob p of success does not change.
- Trials are independent.

Called **Bernoulli trials.**

Are these Bernoulli trials?

- Tossing a coin, success is "heads".
  - *Defined success/failure, constant prob 0.5, independent: good*
- Rolling a die, success is getting a 6.
  - *Again good: P(success)=1/6, constant.*
- Joe buys 1 lottery ticket every week for a year; success is winning a prize.
  - *Success is "win a prize", prob is (about) same each week; one week doesn't affect next*
- A person keeps taking a driving test; success is passing.
  - *P(passing) should increase with each attempt: not constant, not a Bernoulli trial*

- Toss a coin 10 times; success is getting 2 heads in a row (eg. HTHHTTTHHH is 3 successes)
  - *P(success) depends on last toss (or, trials not independent)*
- Large population of people who agree or disagree with a statement; take simple random sample from population, success is "agree".
  - *Trials not quite independent (if you draw a success, there are fewer left to draw from), but with a large population, can act as if everything good.*

Recall:

A certain coffee shop has a Roll Up the Rim to Win promotion. 15% of all cups win a prize:

- How many cups are you likely to have to buy before you get your first prize?
- The cup you get is randomly chosen, so independence and constant probability ok.

To win the first prize on 1st cup: prob 0.15.

On 2^{nd} : fail then succeed: prob (1-0.15)(0.15)=0.1275

On 3^{rd}: fail 2x then succeed: prob (1-0.15)(1-0.15)(0.15)=0.1084

More calculation:

- Prob 0.56 of finding winner within first 5 cups.
- Prob 0.09 of *not* finding winner within first 15 cups.

Number of trials until 1st success: use **geometric** model (p.447) as above.

Mean number of trials to first success = 1/p, here 1/0.15 = 6.67.

Second question above:

A certain coffee shop has a Roll Up the Rim to Win promotion. 15% of all cups win a prize:

- How many prizes might you win if you buy 10 cups of coffee?

Interested in *number* of successes in *fixed* number of trials. This different, uses **binomial** model (p.449):

- fixed number of trials *n*
- fixed prob of success *p* on any one trial
- *variable* number of successes

Here, n=10, p=0.15.

Reminder:

- Bernoulli trials
  - Two possible outcomes “success” and “failure”
  - Prob p of success does not change.
  - Trials are independent.

Interested in *#trials until first success:* **Geometric model**

Interested in *#successes in fixed #trials:* **binomial model.**

Binomial table

The link to Statistical Tables on course website includes table of *binomial distribution probabilities.* (You get these on an exam.)

In here, find chance of exactly k successes in n trials with success prob p.

Rolling up the rim: 10 cups, P(winner)=0.10

(n=10,p=0.10):

Prob of 0 prizes (k=0) is 0.3487, 1 prize (exactly) (k=1) is 0.3874 2 prizes (exactly) (k=2) is 0.1937.

So chance of 2 prizes or less is 0.3487+0.3874+0.1937

Chance of 2 prizes or more: "not 0 or 1" 1-(0.3487+0.3874)

*("Not 2 prizes or less" is **3** prizes or more.)*

What if p>0.5 in binomial table?

Suppose n=8 and p=0.7. What is the probability of

- exactly 7 successes?
- 7 or more successes?

Idea: count failures instead of successes.

P(success)=0.7 means P(failure)=__1-0.7=0.3__

7 successes = __8-7=1__ failure(s)

so look up n=__8__, p=__0.3__, k=__1__ prob=__0.1977__ which is answer we want.

7 or more successes = __7__ or __8__ successes

P(failure)=__0.3__

7, 8 successes = __1__ or __0__ failures

prob we want is 0.0576+0.1977=0.2553

Mean and SD of binomial distribution (p.451)

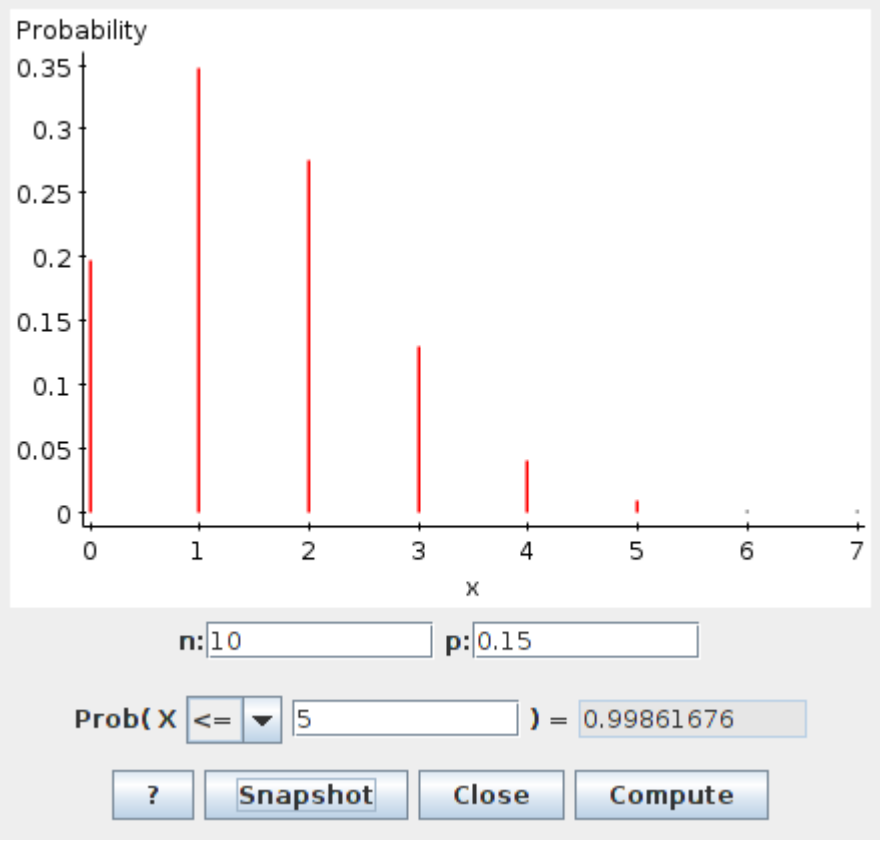
- Mean of number of successes in binomial distribution = *np.*
- SD of number of successes = $\sqrt{np(1-p)}$. (Derivation: math box p. 451.)

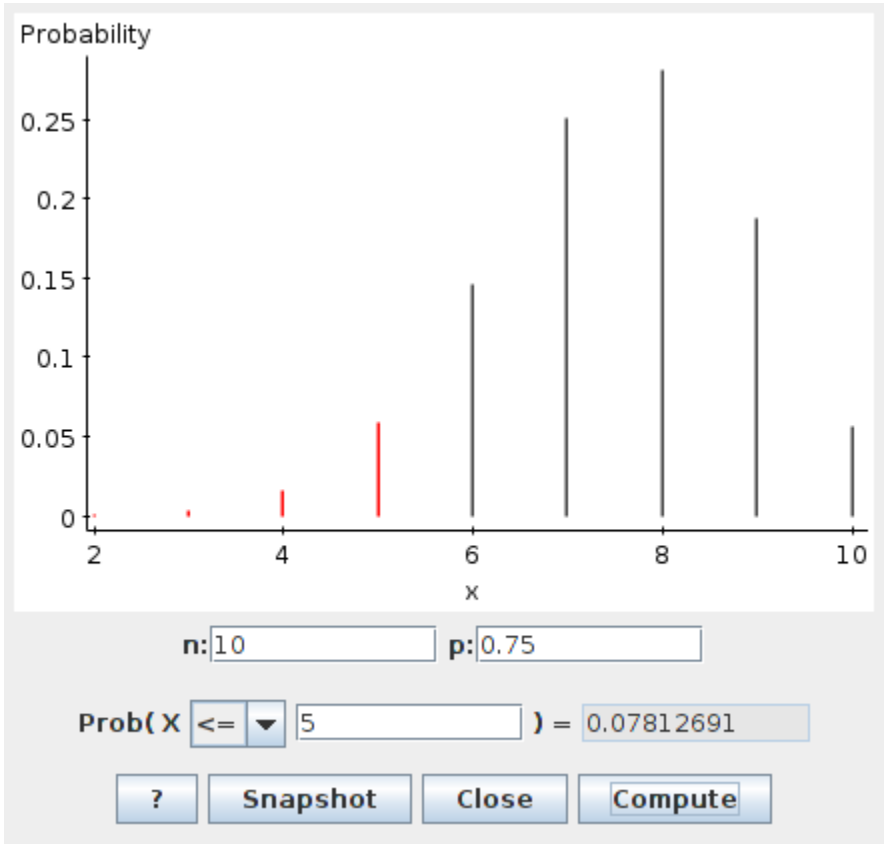
For our example, n=10, p=0.15, so

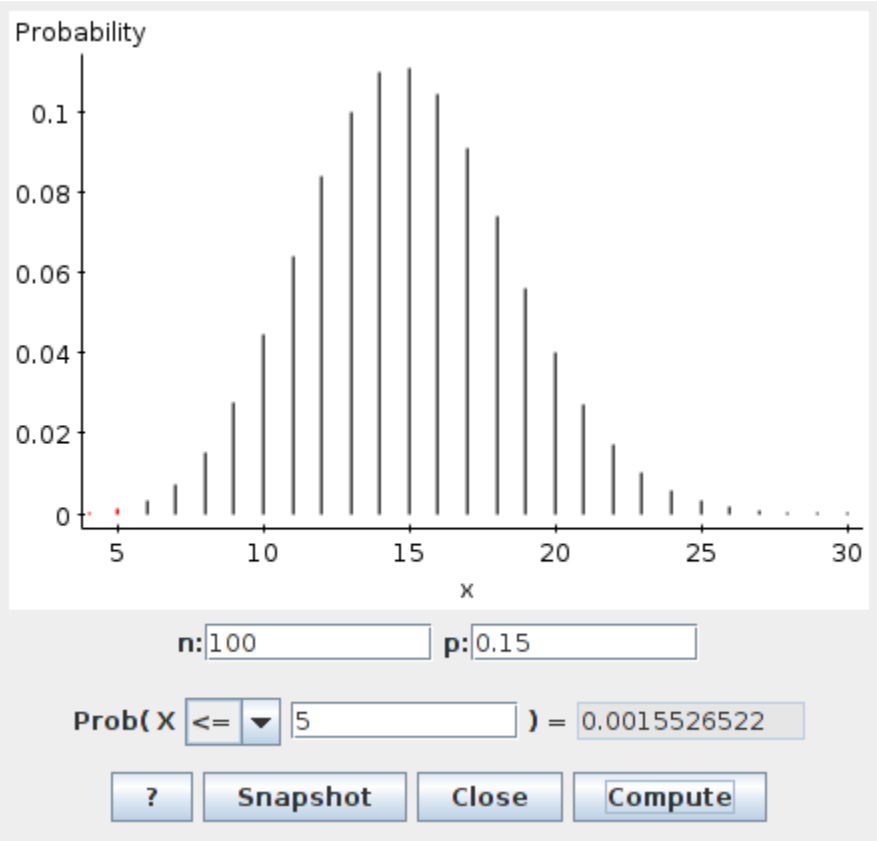
- mean = np = 1.5
- SD = $\sqrt{np(1-p)}$ = $\sqrt{10(0.15)(1-0.15)}$ = 1.13

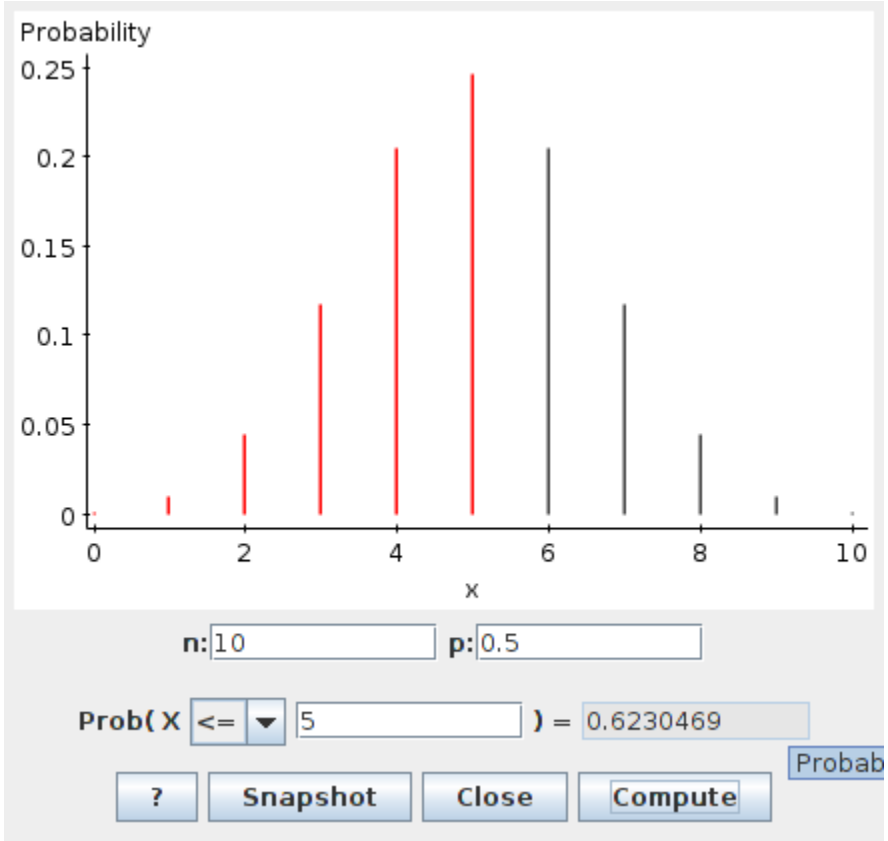
Could get up to 10 successes (though unlikely), so distribution of number of successes skewed to right. (Also, lower limit 0.)

Use StatCrunch to get probability histograms of binomial distributions (Stat, Calculators, Binomial):









– How does the shape depend on p?
  – *P<0.5 ? p>0.5 ? P = 0.5?*
  – *right-skewed; left-skewed; symmetric*

– What happens to the shape as n increases?

  – *shape becomes* _____*normal*_____

– What does this suggest to do if n is too large for the tables?

  – _____Normal approximation_____

If n too large for tables, try **normal approximation to binomial.**

Compute mean and SD of binomial, then pretend binomial actually normal:

P(10 or fewer prizes in 100 coffee cups, 15% of which are winners)?

- # prizes binomial n= 100 p= 0.15
- mean = 100*0.15=15,
- SD = $\sqrt{100(0.15)(0.85)}$ = 3.57
- for **10** prizes, z= (10-15)/3.57=-1.4
- prob of less is 0.0808
- exact answer (StatCrunch) 0.0994

Works if n large and p not too far from 0.5:

- rule of thumb $np \geq 10$ and $n(1-p) \geq 10$
- can relax this a bit if p close to 0.5. (p.454)
- for n=100, p=0.15: $np = 15 \geq 10$, $n(1-p) = 85 \geq 10$. OK.

Continuity correction (see p.455 and note 5 there)

- Know about what this is, but won't need to do it on exam.

Problem:

- binomial distribution *discrete*
- normal distribution *continuous*

so "10 or fewer" on binomial really means "anything *rounding to 10* or fewer" on normal

ie. less than *10.5* coffee cup winners:

z=(10.5-15)/3.57=-1.26, prob. 0.1038, much closer to exact answer 0.0994.

Compare "strictly less than 10 successes":

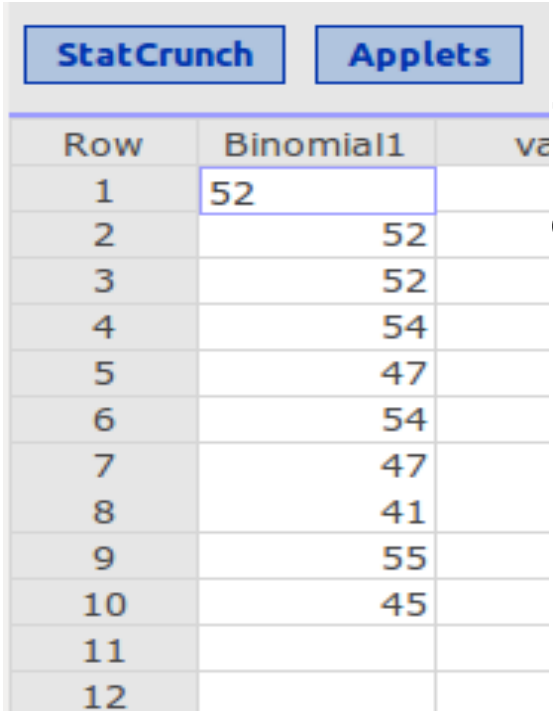
- exact (binomial) 0.0551
- straight normal approx 0.0808 as above

– with continuity correction use 9.5: z=(9.5-15)/3.57=-1.54, prob 0.0618.

Poisson model (p.455): we skip.

Chapter 18: Sampling Distribution Models (p. 473)

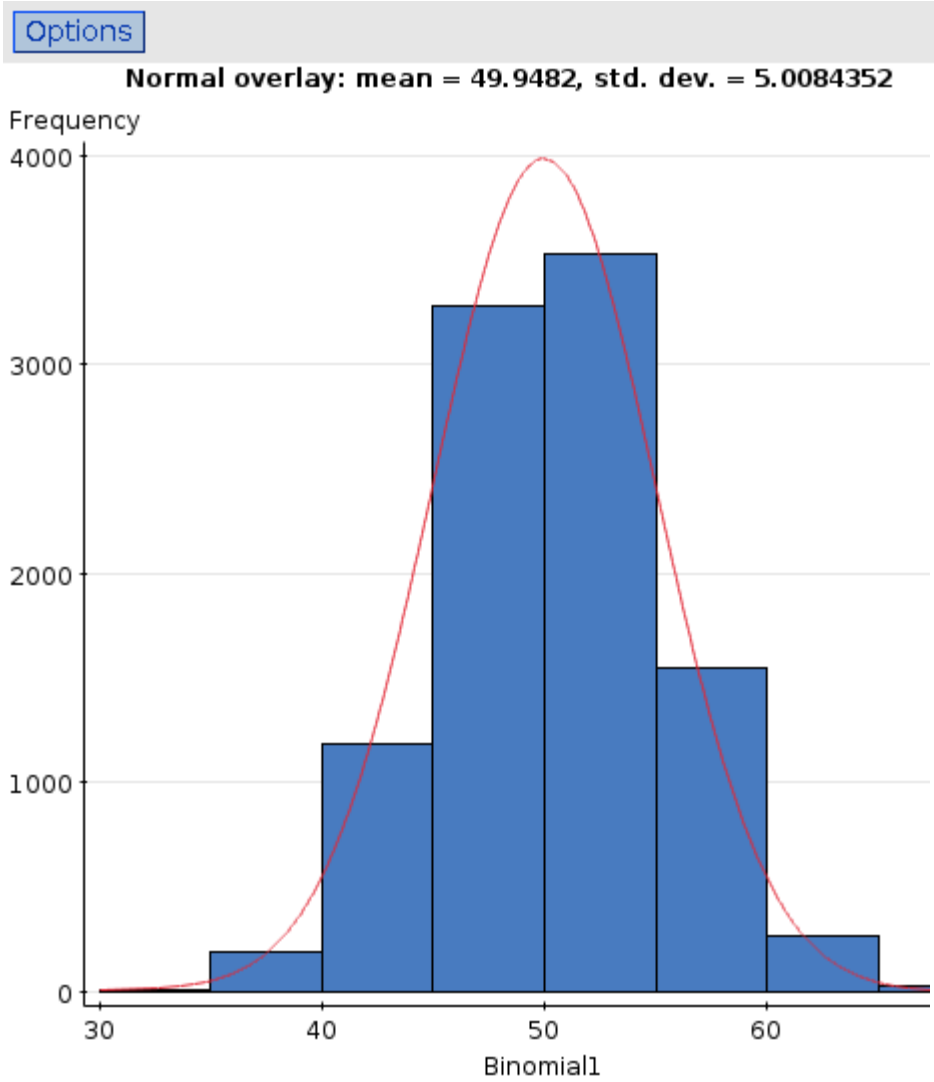
If you toss a fair coin 100 times, how many heads might you get?



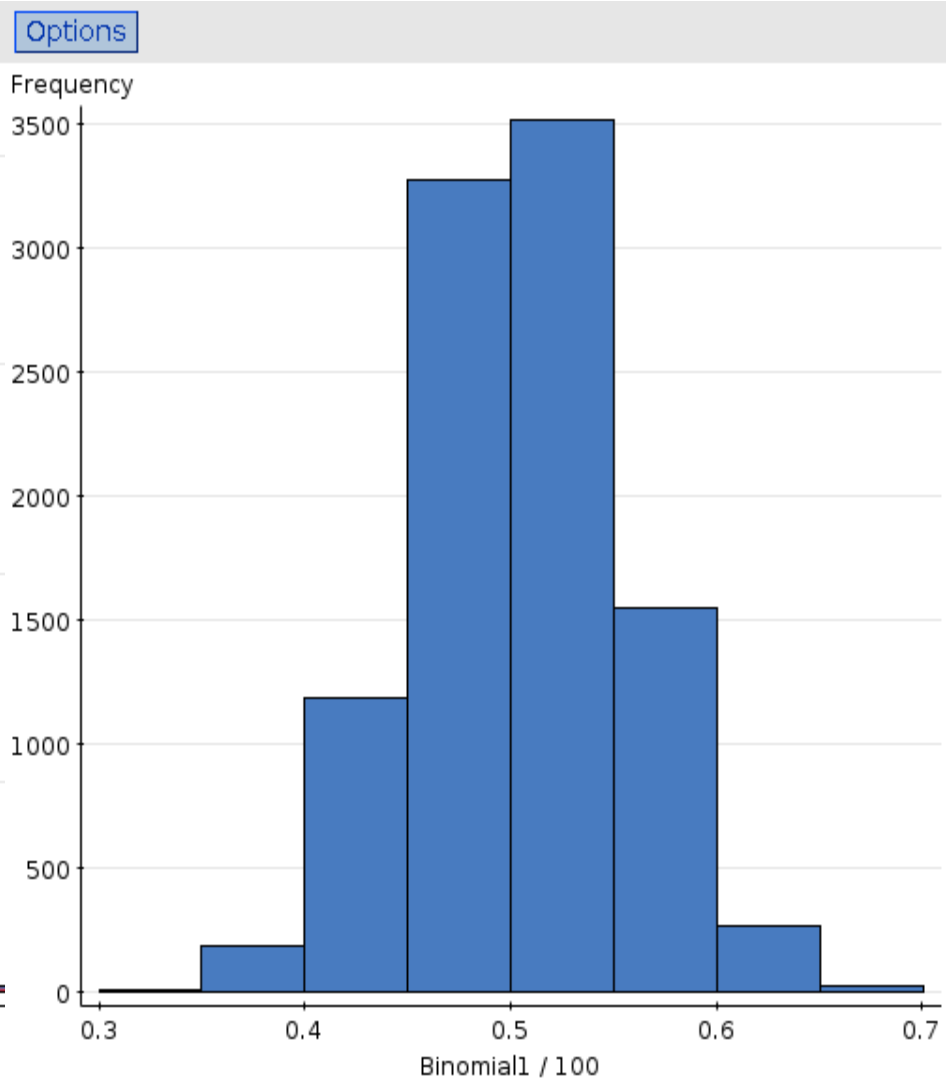
Row	Binomial1	va
1	52	
2	52	
3	52	
4	54	
5	47	
6	54	
7	47	
8	41	
9	55	
10	45	
11		
12		

Simulate with StatCrunch: Data, Simulate, Binomial, n=100, p=0.5. Here 10 rows, 1 column.

- number of heads not the same every time (sampling variability)
- usually between 40-60 heads
- usually *not* exactly 50 heads
- Do more simulations to get better idea.

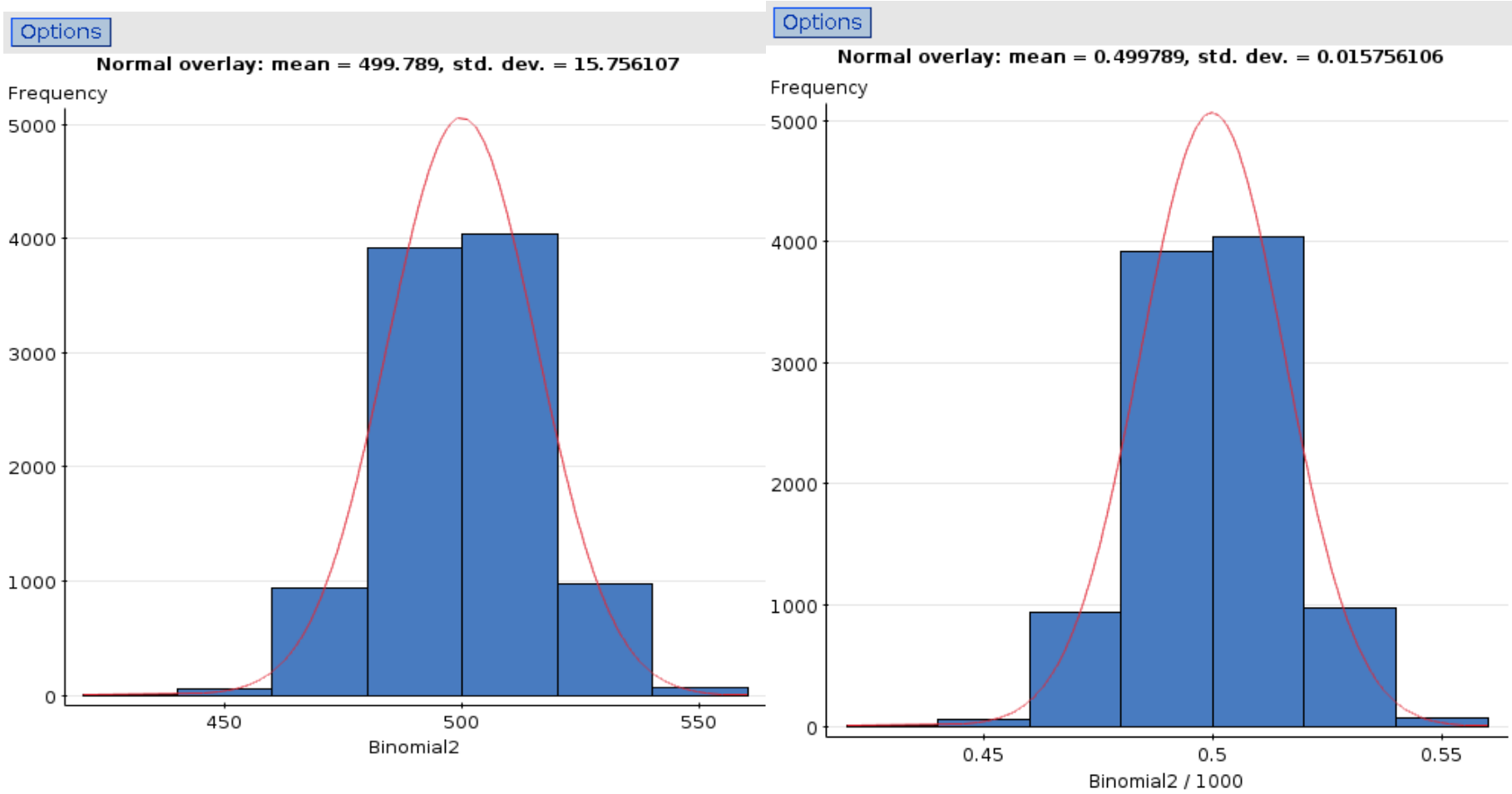


Number of heads in 100 tosses



Proportion of heads

What about 1000 tosses instead of 100?



– *proportion* of heads likely *closer* to 0.5

– *number* of heads might be *further from* half #tosses

- shapes for n=100, n=1000 both normal

Exact answers (p. 474):

- number of "successes" *and* proportion of successes both (approx) normal if np, n(1-p) both at least 10
- For *number* of successes:
  - mean np
  - SD $\sqrt{np(1-p)}$ (saw this last time)
- For *proportion* of successes:
  - mean p
  - SD $\sqrt{p(1-p)/n}$ (divide values for number by n)
- Our examples:

	n=100, p=0.5		**n=1000, p=0.5**	
	Number	Proportion	Number	Proportion
mean	50	0.5	500	0.5
SD	5	0.05	15.81	0.02

- Shows sample proportion closer to its mean for larger n.

Opinion poll:

- 1000 randomly sampled Canadians, 91% believe Canada's health care system better than US's.
- How accurate is that 91%?
- Sampling variability: another sample would contain different people, so its sample proportion $\hat{p}$ may not be 91%.
- Simple random sample, so #better binomial, n=1000, p=?
- *wwwAssume* (for lack of better) p=0.91.

- How far from 91% might another sample proportion be?
- SD is $\sqrt{(0.91)(0.09)/1000}$ = 0.0090.
- Based on this, how likely is a sample proportion over 95%?
  - Check: np=910, n(1-p)=90 both at least 10; normal OK.
  - z=(0.95-0.91)/0.0090=4.44; prob very close to 0.
- Most of the time, sample proportion between $0.91 \pm 2(0.0090)$: 0.892 to 0.928.
- Or, sample proportion unlikely to be more than 2% away from truth, with n=1000 and p near 0.91.

Sampling distribution for sample means (p. 482)

Lottery:

Winnings	-1	2	10
Probability	0.9	0.09	0.01

How much might you win per play if you play many times?

Mean winnings from 1 play is (-1)(0.9)+(2)(0.09)+(10)(0.01)=

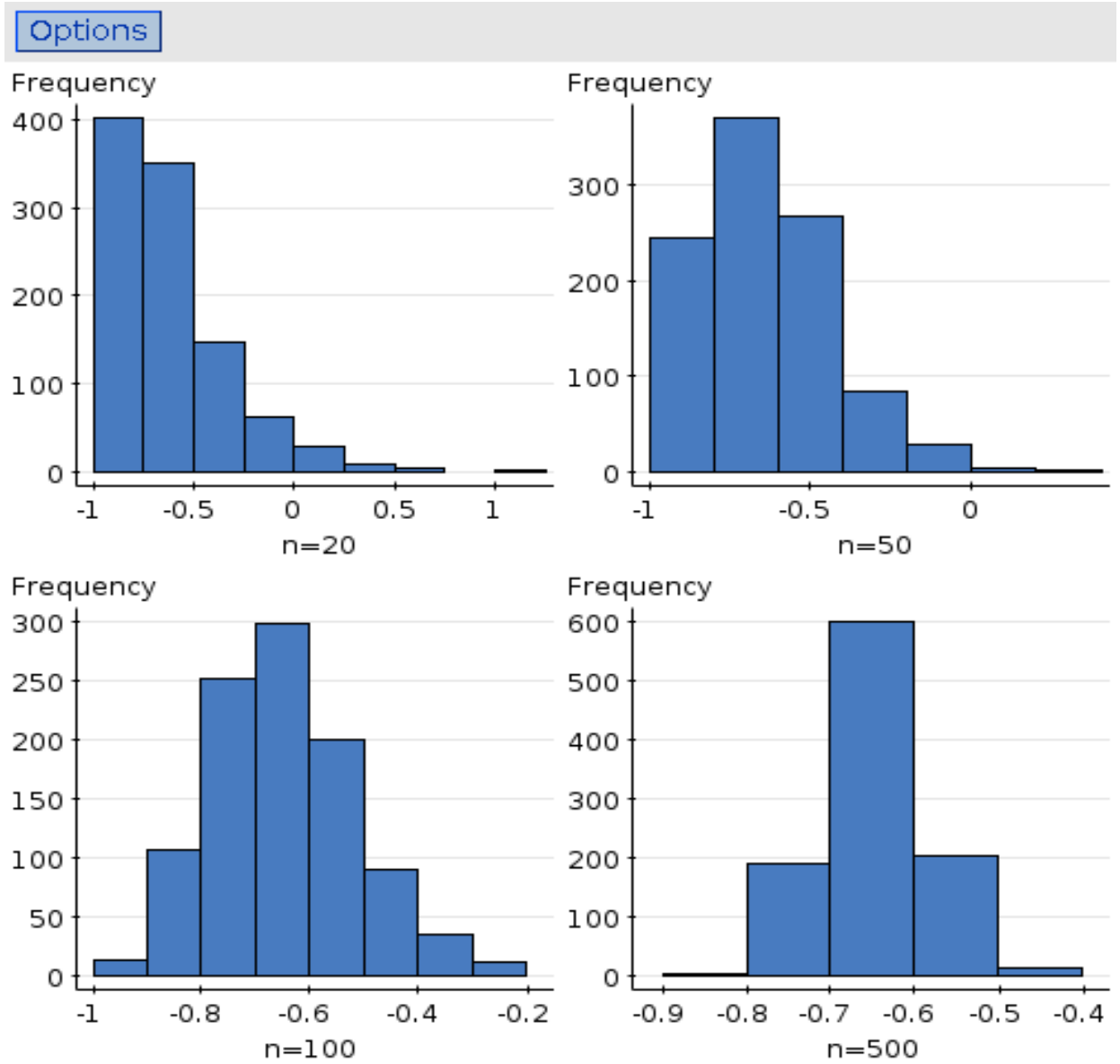
-0.62

(population mean).

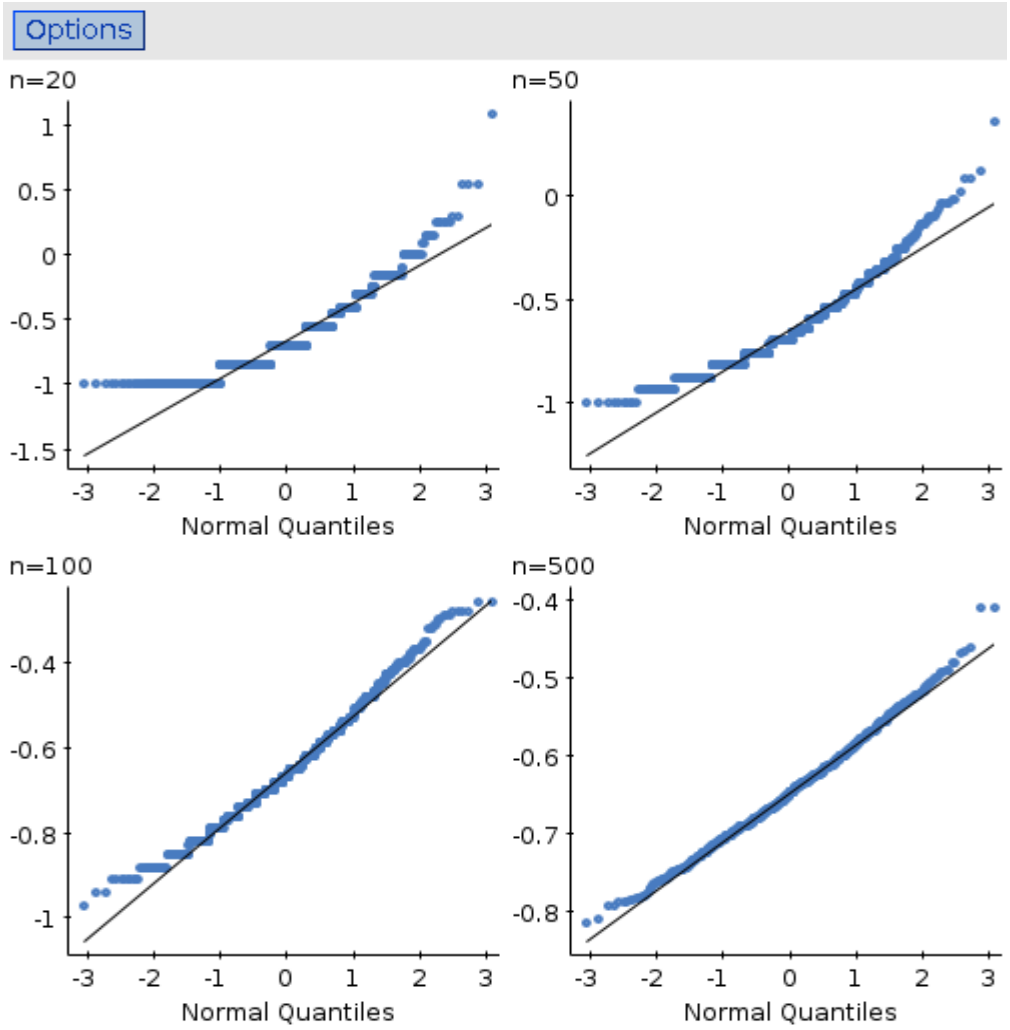
Law of large numbers: sample mean close to population mean for large enough sample. If you play 1000 times, you'll lose close to 0.62 per play.

What kind of sample means might you get for different sample sizes from this population? See below. (Recall: *population* skewed to right, because you have a small chance to win $10).

Sampling distributions of sample mean (per play) for various sample sizes:



Normal quantile plots:



Skewed to right but progressively less so as n gets bigger: more

and more normal.

Where did normal come from?

- Not the population
- Must be the large sample and fact we look at sample mean.

Remarkable fact:

- From *any* population, sampling distribution of sample mean is **approximately normal** if sample is large.
- **Central Limit Theorem** (p.484).

Even in our very extreme population, began to work at about n=100.

Usually ok with *much smaller* n (eg. n=30 often big enough).

Mean and SD of sampling distribution of sample mean (p.486)

- So sampling distribution of sample mean approx normal
- What are its mean and SD?
- Population has mean μ , SD σ
- Sampling distribution of sample mean has:
  - mean μ ,
  - SD $\frac{\sigma}{\sqrt{n}}$ (see Math Box p 486-487).
- As n gets larger, variability of sample mean gets less, so closer sample mean will be to population mean (law of large numbers again).
- Use this when want to know about *sample mean* might be.

Calculations for sample mean (p. 488):

A sample of size n=25 is drawn from a population with mean 40 and SD 10. What is prob that sample mean will be between 36 and 44? (Assume Central Limit Theorem applies.)

- Sampling dist of sample mean has mean $\mu = 40$ and SD $\sigma/\sqrt{n} = 10/\sqrt{25} = 2.$
- 36 gives $z = (36-40)/2 = -2$, 44 gives $z = (44-40)/2 = 2$.
- "Within 2 SDs of mean": prob is about 95% (0.9544).
- Most of the time, sample mean between 36 and 44.
- Most of the time, sample mean no more than 4 away from population mean.

Chapter 19: Confidence intervals for proportions (p. 504)

Recall: Opinion poll with 1000 randomly sampled Canadians, 91% believe Canada's health care system better than US's.

Sampling distribution of sample proportion has:

- mean p (unknown)
- SD $\sqrt{p(1-p)/n}$

For SD, use known n and best guess (91%) at p:

- SD approx $\sqrt{(0.91)(0.09)/1000}$ = 0.0090.
- Sampling distribution approx normal: $np \geq 10$, $n(1-p) \geq 10$
  About 95% of the time, sample proportion $\hat{p}$ should be inside

$$(p-2(0.0090),\ p+2(0.0090)) = p \pm 0.0180$$

that is, p and $\hat{p}$ should be less than 0.0180 apart.

We had $\hat{p}=0.91$, so, about 95% of the time p should be between 0.91-0.0180=0.892 and 0.91+0.0180=0.928.

Confidence interval

(0.892,0.928) called **95% confidence interval** for p.

What does "95% of the time" mean (p. 507)? *In 95% of all possible samples*. But different samples have different $\hat{p}$'s, and give different confidence intervals.

Eg. another sample, with n=1000, might have $\hat{p}=0.89$, giving 95% confidence interval for p of (0.870,0.910).

So *our confidence in procedure rather than an individual interval.*

Certainty and precision (p. 508)

We used 2*SD to get our 95% confidence interval. What if we use 1*SD or 3*SD?

Confidence level	Lower limit	Upper limit
68.0%	0.901	0.919
95.0%	0.892	0.928
99.7%	0.883	0.937

- if you want a shorter interval, you have to be less confident.
- if you want to be more confident in your interval, it has to be longer.
- no way to get a "certainly correct" interval – unless it is so long as to be meaningless – always the chance that your statement about p will be *wrong.*

Getting other confidence intervals for a proportion (p. 510)

How would we get a 90% interval? 80%?

- Sampling distribution approximated by *normal* (hence 1, 2, 3)
- Interval is $\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$
- with z* from normal table is value where
  - half the leftover is below -z*
  - half the leftover above z*
  - so amount of normal curve between -z* and z* is right %.

Find z* for 90% interval:

- leftover is 10%=0.1000
- half that is 5%=0.0500
- Table: z=-1.64 or -1.65 has 0.0500 less

– z=1.64 or 1.65 has 0.0500 more (0.9500 less).

– so z*=1.64 or 1.65.

Handy table:

Confidence level	z*
90%	1.645
95%	1.960
99%	2.576

– 1.96 a “more accurate” version of 2.

Example

A city ballot includes an initiative that would allow casino gambling. A poll of 1200 randomly chosen voters finds 53% in favour, while a second poll of 400 randomly chosen voters finds 54% in favour. In each case, find a 95% confidence interval for the proportion of all voters in favour.

First poll:

SD of $\hat{p}$ is _ $\sqrt{0.53*0.47/1200}$ *= 0.0144*

z* for 95% CI is *1.96*

95% CI margin of error = *1.96(0.0144)=0.028*

Interval from *0.53-0.028=0.502* to *0.53+0.028=0.558*

Second poll:

SD of $\hat{\bar{p}}$ is _ $\sqrt{0.54*0.46/400}$ *=0.0249*

z* for 95% CI is *1.96*

margin of error is *1.96*0.0249=0.049*

interval from *0.54-0.049=0.491* to *0.54+0.049=0.589*

(0.028 and 0.049 are margins of error for the confidence intervals)

- polls differ in % in favour (sampling variability)
- First poll allows conclusion that majority in favour
- Second poll gives less precise interval (smaller sample).

95% CI for second poll was 0.481 to 0.589. What would a 90% interval be?

SD of $\hat{p}$ was *0.0249*

z* for 90% CI is *1.645*

margin of error is *0.0249*1.645=0.041*

interval is from *0.54-0.041=0.499* to *0.54+0.041=0.581*

Compare with 95% CI.

Determining sample size (p. 514)

Suppose we plan a survey. How big a sample?

- margin of error $m = z^* \sqrt{\hat{p}(1-\hat{p})/n}$ determines how far CI goes up and down
- desired confidence level known: know z* (eg. 1.96)
- don't have a sample yet, but might have guess at p
- know how big we'd like margin of error to be (say m)
- then can solve for n:

$$n = \frac{z^{*2} p(1-p)}{m^2}$$

A study is to be carried out to estimate the proportion of all adults who have higher-than-normal levels of glucose in their blood. The aim is to have a margin of error on a 90% confidence interval of 4% (0.04). How many (randomly chosen) adults should be surveyed? A pilot study shows that the proportion is about 20%.

Use the formula. z*=1.645, m=0.04, p=0.20:

$$n=\frac{1.645^2(0.20)(1-0.20)}{0.04^2}$$ = *270.6*; sample *271* adults (round *up*).

Without a guess at p, use p=0.50 (worst case):

$$n=\frac{1.645^2(0.5)(1-0.5)}{0.04^2}$$ = *422.82*; sample size should be *423*.

It pays to have a guess at p!

Chapter 20: Testing Hypotheses about Proportions (p.530)

A newsletter reported that 90% of adults drink milk. A survey in a certain region found that 652 of 750 randomly chosen adults (86.93%) drink milk. Is that evidence that the 90% figure is not accurate for this region?

Difference between 86.93 and 90, but might be chance.

One approach: confidence interval.

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = 0.0123, so

95% CI is 0.845 to 0.893

99% CI is 0.838 to 0.901

so now what?

Better: hypothesis testing (p.530). Think about logic first by analogy.

Court of law

		Decision	
		Not guilty	Guilty
Truth	Innocent	Correct	Serious error
	Guilty	Error	Correct

- Truth (unknown)
- Decision (we hope reflects truth)
  - based on *evidence*: does it contradict accused being innocent?
  - "Presumption of innocence"
  - "Beyond all reasonable doubt".

Hypothesis testing

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	Correct	Type I error
	H_0 false	Type II error	Correct

- Null hypothesis H_0 is "presumption of innocence": given state of affairs is correct. *In milk example, H_0 is p=0.90 (newsletter correct).*
- Alternative hypothesis H_A is that H_0 is false ("guilty"). Need evidence (data) to be able to reject H_0 in favour of H_A . *In milk example, H_A is p not equal to 0.90 (newsletter wrong).*
- Hypotheses (p. 531): ask yourself "what do I need evidence for?" That's H_A .
- Milk example: trying to prove that 90% *not* correct for this region, so

- H_0: $p = 0.90$
- H_A: $p \neq 0.90$

How to assess whether we believe H_0?

- Assess the evidence
- Evidence is our data
- in particular: sample proportion $\hat{p}$.
- **P-value** (p. 533): probability of $\hat{p}$ as far or further from H_0 than the value observed.

In our data, H_0: $p = 0.90$ and $\hat{p} = 652/750 = 0.8693$. If H_0 true, value of $\hat{p}$ we might have observed approx normal, mean 0.90, SD $\sqrt{(0.90)(0.10)/750}$ = 0.0110.

Prob of observing $\hat{p}$ below 0.8693 (further away from null):

- $z = \frac{0.8693 - 0.9}{0.0110}$ = -2.79; prob (Table Z) 0.0026.

– Could have observed $\hat{p}$ *above* 0.90 too (same evidence against null), so P-value twice this, 0.0052.

"Beyond a reasonable doubt"

Is P-value of 0.0052 "beyond a reasonable doubt"? Says:

- a value of $\hat{p}$ like the one we observed very unlikely *if* $H_0: p=0.90$ *true*
- so either:
  - (a) we observed something very unlikely
  - (b) $H_0: p=0.90$ isn't true after all.
- when P-value so small, prefer to believe (b): **reject in favour of** $H_A: p \neq 0.90$ **.**

On the other hand, a P-value like 0.7789 not small. Says that if H_0 true, result we observed entirely possible. Have not proved that H_0 true, but **cannot reject** H_0 **.** Some people say "retain H_0", or "data consistent with H0". Ideas get at "we don't fervently

believe that H_0 is true, but we cannot prove it wrong”.

One-sided and two-sided tests (p. 538)

Our alternative was $H_A: p \neq 0.90$: *two-sided* since values of $\hat{p}$ too far above *or* below 0.90 could make us reject H_0 . Suppose now $\hat{p}$ had been 0.92. Then $z = (0.92 - 0.90)/0.0110$ = 1.82. P-value is prob of *above*, doubled: 2(1-0.9656)=0.0688.

Might have been looking for evidence that p was *smaller* than 0.90, ie. $H_A: p < 0.90$. Two parts to getting P-value:

- are we on correct side? $\hat{p} = 0.8693$ is, $\hat{p} = 0.92$ is not.
  - if on correct side, go to next step.
  - if not on correct side, stop and declare H_0 not rejected (as for $\hat{p} = 0.92$).
- P-value is prob of *less*. For $\hat{p} = 0.8693$ that is (z=-2.79) 0.0026.

Similar idea for other kind of one-sided alternative, like $H_A: p > 0.90$:

- on correct side? $\hat{p} = 0.8693$ is not, $\hat{p} = 0.92$ is.
  - if on correct side, go to next step
  - if not ($\hat{p} = 0.8693$), stop and declare H_0 not rejected.
- P-value is prob of *greater.* For $\hat{p} = 0.92$, that is (z=1.82) $1 - 0.9656$ = 0.0344.

When doing a test (p. 541):

- always state the P-value. Enables reader to draw own conclusion about truth or falsity of H_0 .
- follow up (particularly when you want to reject H_0) with a confidence interval. Enables reader to get idea of size of parameter (effect size) and hence whether result is *important* (makes practical difference) vs. *statistically significant (unlikely if null hypothesis is true)*

How small is small (for a P-value)?

- think about how plausible the alternative is
  - if alternative is implausible, need very strong evidence (very small P-value)
  - if alternative is plausible, weaker evidence (larger P-value) would do

– think about how costly or dangerous rejecting H_0 is
  – eg. if rejecting H_0 would mean rebuilding a factory

The rest of the way:

- chapter 21 today
- skip chapter 22
- chapter 23 (inference for means) on Friday
- chapter 24 (comparing two means) next Tuesday
- chapter 25 (paired data) next Friday
- then, done!

Chapter 21: More about Tests (p. 554)

- null hypothesis has to *give a parameter value* like $H_0: p = 0.7$.
- alternative has to say *what you are trying to prove* like $H_A: p \neq 0.7$. (two-sided)
- Kind of alternative you use depends on exactly what you want to prove:
  - is p different? (2-sided)
  - is p larger? (1-sided)
  - is p smaller? (1-sided)

Example: at a small computer peripherals company, only 60% of the hard drives produced pass all the performance tests the first time. Management recently invested a lot of resources into the production system. A sample of 100 hard drives produced recently found that 71 of them passed all the performance tests the first time. Did the investment help?

- let p = proportion of all hard drives produced recently that pass performance tests first time.
- Looking for evidence that investment *helped*, ie.

  H_A: _ $p>0.60$ _____.
- Null has to give value for p: $H_0: p=0.60$ _.
- SD of $\hat{p}=\sqrt{(0.60)(0.40)/100}$ = 0.0490.
- $100(0.6)=60$ and $100(0.4)=40$ both at least 10: normal approx OK. (Not usually a problem.)
- Sample gives $\hat{p}=?$ 71/100=0.71_____.

- Test statistic = z=(0.71-0.60)/0.0490=2.24
- On which side? Correct side-observed 0.60 successes or more
- P-value for z= 2.24 = prob of above (Table Z): 0.0125
- If p=0.60 correct, prob. 1-0.9875=0.0125 to have observed as high as $\hat{p}=0.71$
- so reject H_0 and conclude investment has helped.

Compare:

- prob of H_0 true, *if* observe data like this: *no.*
- prob of this kind of data, *if* H_0 true: *yes,* P-value.

Reason: H_0 either *is* true or false, so can't talk of its prob.

Previous example:

- $H_0: p = 0.6$
- $H_a: p > 0.6$

Suppose now $\hat{p} = 0.63$.

Then $z = (0.63 - 0.6)/0.0490$ = 0.61, with P-value 0.2709

Data not that unlikely if H_0 true, so cannot reject H_0 .

- have *not* proved that H_0 correct
- have only obtained the kind of data we *would* have seen, if H_0 were correct
  - so justified in acting *as if* H_0 were correct.

What if $\hat{p}=0.55$? *less* then 0.6, wrong side, don't reject H_0 .

Idea: if investment helpful, would have had *more* than 60% of hard drives work first time.

Alpha (p. 561)

- How to decide whether P-value small enough to reject H_0 ?
- Choose α (alpha) **ahead of time**:
  - if rejecting H_0 an important decision, choose small α (0.01)
  - if seeing whether any evidence at all, larger α (0.10)
  - "default" α 0.05.
- Reject H_0 if P-value less than the α you chose.

With $\alpha = 0.05$ in above examples:

- $\hat{p} = 0.71$: P-value 0.0125: *reject* H_0 *:* investment has helped
- $\hat{p} = 0.63$: P-value 0.2709: *do not reject* H_0 *:* no evidence that investment has helped.

With $\alpha = 0.01$, even $\hat{p} = 0.71$ (P-value 0.0125) not strong enough evidence to conclude that investment has helped.

Tests and Cis (p.565)

We do later (with inference for means).

Example of hypothesis test for proportion (ex. 20.28, edited)

A study of acid rain and trees in Hopkins Forest found that 25 of 100 (randomly selected) trees had some kind of damage from acid rain. This is different from the 15% quoted in a recent article. Is the 15% figure *wrong* for the Hopkins Forest? Use $\alpha = 0.05$.

- *Parameter:*
  - *p=proportion of all trees in Hopkins Forest damaged by acid rain*
- *Hypotheses:*
  - *alternative* $H_a: p \neq 0.15$ *(2-sided)*
  - *Null?* $H_{0L}\, p = 0.15$
- *SD of* $\hat{p}$ *?* $\sqrt{(0.15)(1-0.15)/100}$ *= 0.0357*
- *Test statistic: z=(25/100-0.15)/0.0357=2.8*

- *Probability of more extreme? 1-0.9974=0.0026*
- *P-value: 0.0026 x 2 = 0.0052*
- *Conclusion:*
  - reject H_0 *, conclude the article's value of p=0.15 of trees damaged by acid rain is wrong for the Hopkins Forest*
- *Follow up with 95% confidence interval: (0.175,0.343)*

Making errors (p. 567)

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	Correct	Type I error
	H_0 false	Type II error	Correct

- Type I error: reject H_0 when it is true
  - jury convicts innocent person
  - healthy person diagnosed with a disease
  - future patients get a useless treatment
- Type II error: fail to reject H_0 when it is false
  - jury fails to convict guilty person
  - sick person not diagnosed with disease
  - future patients do not get useful treatment

- Prob of type I error is α (eg. $\alpha = 0.05$)
- Prob of type II error called β (beta)
  - eg. $H_A : p \neq 0.3$ but need to know what p actually is to find β
  - if p far from 0.3, should be easy to reject H_0 (β small)
  - if p close to 0.3, could be hard to reject H_0 (β large)
- Prob of *not* making type II error called *power* ($1-\beta$). *(p. 568)*

Want to have power large enough to have decent chance to reject null if p "interestingly" different from H_0 value.

Example: suppose we have sample size 100. How likely are we to reject $H_0: p=0.3$ in favour of $H_A: p>0.3$ if p is really 0.4, using $\alpha=0.05$?

Simulate. Steps:

- simulate a bunch of binomials with n=100, p=0.4 (*true* distribution)
- turn each into sample proportion (divide by 100)
- calculate z for each (using $H_0: p=0.3$)
- calculate P-value (2-sided) for each using StatCrunch function "pnorm"
- count how many of those P-values < 0.05.

My simulation (some):

StatCrunch | Applets | Edit | Data | Stat | Graph | Help

Row	Binomial1	p-hat	z	P-value	reject	var6
1	43	0.43	2.8368326	0.0045563498	true	
2	53	0.53	5.0190115	5.1938048e-7	true	
3	42	0.42	2.6186147	0.008828761	true	
4	32	0.32	0.43643578	0.66252058	false	
5	42	0.42	2.6186147	0.008828761	true	
6	44	0.44	3.0550505	0.0022502266	true	
7	41	0.41	2.4003968	0.016377308	true	
8	34	0.34	0.87287156	0.38273309	false	
9	31	0.31	0.21821789	0.82725935	false	
10	40	0.4	2.1821789	0.029096332	true	
11	43	0.43	2.8368326	0.0045563498	true	
12	39	0.39	1.963961	0.049534613	true	
13	42	0.42	2.6186147	0.008828761	true	
14	42	0.42	2.6186147	0.008828761	true	
15	26	0.26	-0.8728715	0.38273309	false	
16	40	0.4	2.1821789	0.029096332	true	
17	37	0.37	1.5275252	0.12663046	false	
18	36	0.36	1.3093073	0.19043026	false	
19	37	0.37	1.5275252	0.12663046	false	
20	40	0.4	2.1821789	0.029096332	true	
21	45	0.45	3.2732684	0.0010631149	true	
22	43	0.43	2.8368326	0.0045563498	true	
23	37	0.37	1.5275252	0.12663046	false	
24	45	0.45	3.2732684	0.0010631149	true	

Results

- Stat – Tables – Frequency – select column “reject” -- compute frequency and relative frequency.

Frequency table results for reject:

reject	Frequency	Relative Frequency
false	378	0.378
true	622	0.622

My power about 0.62. If that's

too small, larger sample size needed.

Chapter 22: Comparing Two Proportions (p. 585)

2006/7 homicides:

	Shooting	Other	Total	%shooting
London	29	138	167	17.37
Toronto	72	82	154	46.75

- What can we say about the difference in proportion of homicides by shooting in the two cities (thinking of 2006/7 as a random sample of all years)?
- Best guess at difference is 0.4675-0.1737=0.2938.
- But how variable is that from year to year?

SE of difference in proportions (p.587)

- Let p_1 be proportion of homicides by shooting in Toronto
- Let p_2 be proportion of homicides by shooting in London
- $SE(\hat{p}_1)=\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}}$
- $SE(\hat{p}_2)=\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n}}$
- $SE(\hat{p}_1-\hat{p}_2)$? variance of difference is *sum* of variances:
- $SE(\hat{p}_1-\hat{p}_2)=\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}+\frac{\hat{p}_2(1-\hat{p}_2)}{n}}=D$, say.
- then CI for difference in proportions is $\hat{p}_1-\hat{p}_2\pm z^* D$

– with our numbers

$$D=\sqrt{\frac{(0.4675)(1-0.4675)}{154}+\frac{(0.1737)(1-0.1737)}{167}} = 0.0498;$$

– for a 95% CI z*=1.96, so margin of error $1.96(.0498) = 0.0976$,

– interval is $0.4675-0.1737\pm0.0976 = (0.2,0.39)$,

– so we think % of homicides in Toronto by shooting is between 20 and 39 percentage points higher than in London.

Note:

– same idea for CI as before, estimate +/- margin of error

– margin of error is z* times SE of estimate.

– So: figure out right estimate, right SE of estimate for what you need CI of.

Hypothesis test for difference in proportions (p.593)

- Null: $H_0: p_1 - p_2 = 0$ or $H_0: p_1 = p_2$
- Alternative: $H_A: p_1 - p_2 \neq 0$ or $p_1 \neq p_2$
  - (or $p_1 - p_2 < 0$ or $p_1 < p_2$)
  - (or $p_1 - p_2 > 0$ or $p_1 > p_2$)
- Now, act as if $p_1 = p_2 = p$, say, so can do better for $SE(\hat{p}_1 - \hat{p}_2)$
  - estimate p as *overall* proportion of successes and calculate

$$SE_{pooled}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} = D_p$$ (p="pooled")

  - then test statistic is $z = \frac{\hat{p}_1 - \hat{p}_2}{D_p}$
  - get P-value from normal distribution as before.

On our data:

$H_0: p_1 = p_2$

$H_A: p_1 \neq p_2$

	Shooting	Other	Total	%shooting
London	29	138	167	17.37
Toronto	72	82	154	46.75

– overall proportion of deaths by shooting is $\frac{29+72}{167+154}$ = 0.3146

– so $D_p = \sqrt{\frac{(0.3146)(1-0.3146)}{167} + \frac{0.3146(1-0.3146)}{154}}$ = 0.0519 (not much different from D=0.0498)

– so $z = \frac{0.4675 - 0.1737}{0.0519}$ = 5.66

– P-value=0.0000000151 (software). Conclude that shooting %'s are different.

Sample size for two proportions (p.598)

- Confidence interval for difference in proportions had margin of error $1.96(0.0498)$ =0.0976. How many homicides would we need to observe in each city to get this margin down to 0.07?
- Need to know p_1 and p_2
  - but have guesses $\hat{p}_1=0.4675$ and $\hat{p}_2=0.1737$
- Assume number of homicides observed in each city same: let $n=n_1=n_2$.
- Equate desired margin with formula, leaving n unknown:

$$0.07=1.96\sqrt{\frac{(0.4675)(1-0.4675)}{n}+\frac{(0.1737)(1-0.1737)}{n}}$$

- and solve for n: between 307 and 308, take 308. Need this many homicides *in each city.*
- If p's unknown, replace them by 0.5. In above case, would require an n of 392 in each city.
- Pays to have knowledge of p_1 , p_2 !

– Decreasing margin of error by a little can increase sample size required by a lot. (Cutting it in half: times sample size by *4).*

Power of hypothesis test

Again do by simulation. Have to know (or have guess at) both p's.

Suppose we'll do another study of % homicides by shooting for London and Toronto (using more recent data). We will look at 50 homicides for each city. How likely are we to be able to reject $H_0: p_1 = p_2$ in favour of $H_a: p_1 \neq p_2$ using $\alpha = 0.05$, when in fact $p_1 = 0.4675$ and $p_2 = 0.1737$?

- generate a bunch of simulated homicide deaths by shooting for each city:
  - Toronto binomial $n = 50, p = 0.4675$
  - London binomial $n = 50, p = 0.1737$
  - turn each into proportions out of 50
  - calculate D for each pair of simulated death proportions
  - calculate test statistic each time
  - get P-values

– count how many P-values are less than 0.05.

Some of my simulation:

StatCrunch | Edit | Data | Stat | Graphics | Help

Row	Toronto	London	p1-hat	p2-hat	D	z	P-value	reject
1	22	11	0.44	0.22	0.09143303	2.4061325	0.016122416	true
2	28	13	0.56	0.26	0.09368031	3.2023807	0.0013629679	true
3	20	11	0.4	0.22	0.09073037	1.9839002	0.047266964	true
4	25	7	0.5	0.14	0.08606974	4.182655	2.8812481E-5	true
5	19	5	0.38	0.1	0.08069696	3.4697711	5.2090193E-4	true
6	28	9	0.56	0.18	0.08876936	4.280756	1.8625937E-5	true
7	23	8	0.46	0.16	0.08749857	3.4286275	6.066417E-4	true
8	23	9	0.46	0.18	0.0889943>	3.146266	0.0016536951	true
9	23	10	0.46	0.2	0.09037699	2.8768384	0.004016811	true
10	26	5	0.52	0.1	0.08241359	5.0962467	3.4645342E-7	true
11	23	5	0.46	0.1	0.08226786	4.37595	1.2090487E-5	true
12	29	7	0.58	0.14	0.08532292	5.1568794	2.5109924E-7	true
13	22	9	0.44	0.18	0.08876936	2.9289384	0.0034012182	true
14	17	11	0.34	0.22	0.0889943>	1.3483998	0.17752986	false
15	27	10	0.54	0.2	0.09037699	3.7620196	1.6854685E-4	true
16	17	7	0.34	0.14	0.08304216	2.4084153	0.016021946	true

See a gray box a

Seems unlikely that we fail to reject the null.

Tabulate “reject”:

Options

Frequency table results for reject:

reject	Frequency	Relative Frequency
false	95	0.095
true	905	0.905

Rejected 905 times of 1000: power estimated at 0.905=90.5%. Even with small sample sizes still very likely to detect difference. True because %'s very different.

What a hypothesis test is

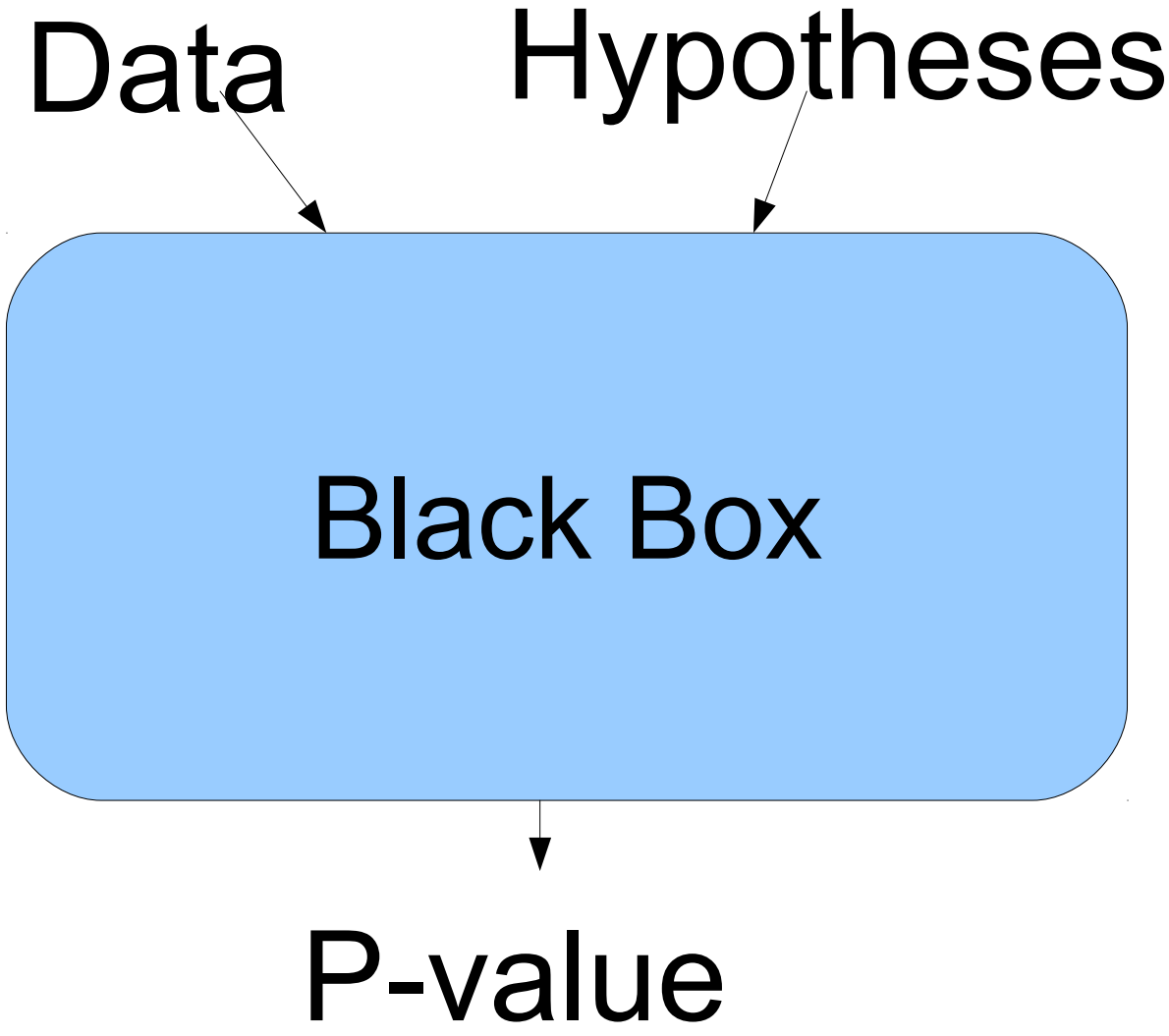


Table of tests and confidence intervals

Inference for	SD to use (for CI if different)	P-value from	For CI use
Proportion	$\sqrt{\frac{p(1-p)}{n}}$; $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	normal	z^*
Comparing proportions	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1}+\frac{\hat{p}(1-\hat{p})}{n_2}}$; $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	normal	z^*
Mean	$s/\sqrt{n}$	t, df n-1	t^*
Two means	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ $H_0: \mu_1-\mu_2=0$	t, df $min(n_1-1, n_2-1)$	t^*
Matched pairs	differences, $H_0: \mu=0,$ as for mean.		

Chapter 23: Inferences about means (p.617)

"Damaged by acid rain" or "Cause of homicide death" categorical variable. But what about eg. number of calories in a serving of yogurt? Might want to say something about mean calories/serving in *all* servings of a certain type of yogurt. That is quantitative variable. (exercise #36.)

Central limit theorem says:

- if you sample from a population with mean μ, SD σ, sampling distribution of $\bar{y}$ approximately
  - *normal*
  - mean μ
  - SD $\sigma/\sqrt{n}$

  if sample size n is large.

All very nice, but:

- sample size may not be large
- population SD σ almost certainly not known

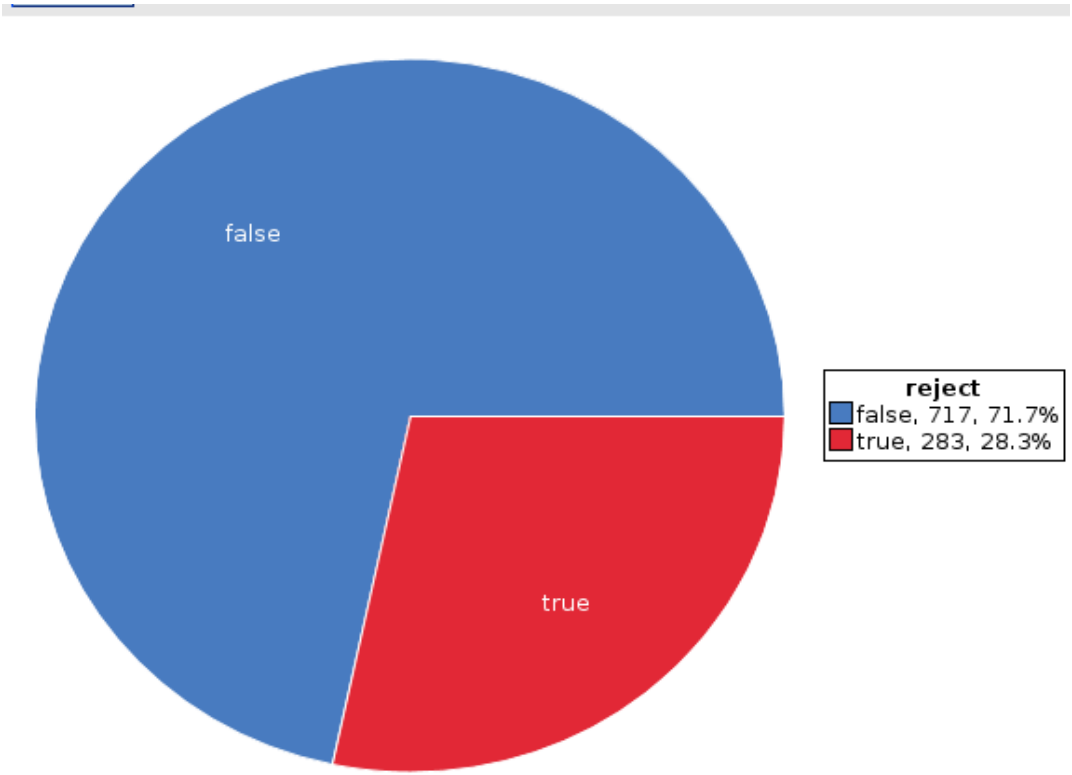
What then?

Investigate by simulation. Population normal, mean 30, SD 5, take samples of size 2. Pretend σ not known, so use sample SD's instead.

Look at t-sim data (on StatCrunch):

- generate normal population mean 30 SD 5
- take 1000 (“many”) samples of size 2, save in one column with column of which sample they belong to
- calculate mean of each sample
- calculate SD of each sample
- calculate z for each sample, using sample SD instead of population SD, and using correct mean 30
- calculate P-value for each z (2-sided alternative)
- for each P-value, note whether less than 0.05 (reject=true) or greater (reject=false).

– Since hypothesized mean of 30 is correct, proportion of (false) rejections should be around 0.05 (50 out of 1000).
– How many times did we reject?



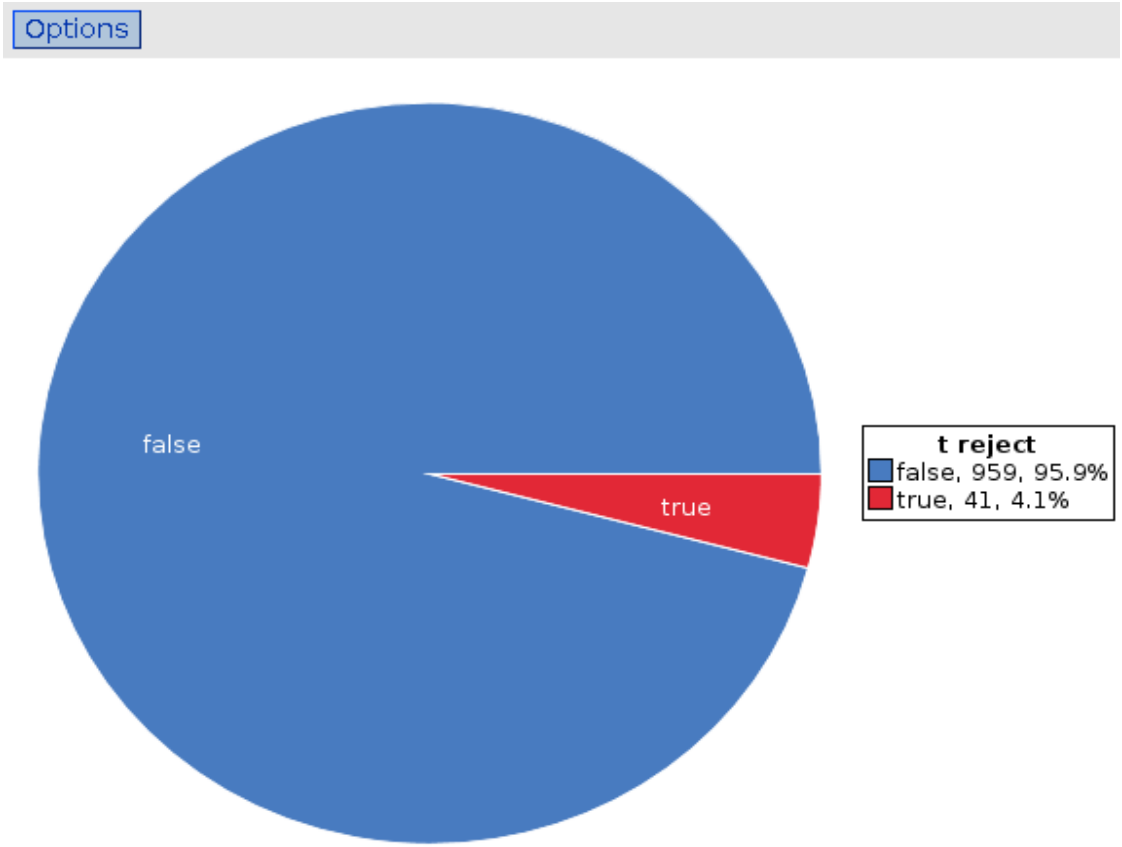
283 times, or 28.3%. *Way* too much!

Problem! How to fix this up?

- Issue: sample SD isn't same as population SD, might be far off if sample size small.
- Gosset (p. 620 of text) worked out what to do:
  - calculate test statistic using sample SD, call it t.
  - get P-value from table of **t-distribution** (Table T) with n-1 *degrees of freedom*.
- Example:
  - testing $H_0: \mu = 30$ vs. $H_A: \mu \neq 30$ (two-sided) and have n=10, $\bar{y} = 35$, s=5.
  - $t = (35 - 30)/(5/\sqrt{10})$ = 3.16.
  - Look up in table with 10-1=9 df:
    - 3.16 between 2.821 and 3.250
    - so P-value between 0.01 and 0.02
    - StatCrunch gives 2 x 0.0058 = 0.0116.

– Using $\alpha = 0.05$, reject null;
– conclude population mean *not* 30.

How does our simulation perform getting P-values from t-distribution with 1 df (sample size 2)?



Very close to 5% (wrong) rejections of the null. Good.

When will the t-test work?

- Theory based on *normal* population.
- With *large* samples:
  - central limit theorem will work regardless of actual shape of population.
  - sample SD will be close to population SD, so not knowing σ won't matter much.
- with *small* samples:
  - no central limit theorem to help
  - sample SD might be far from population SD
  - beware of outliers/skewness
- but: in most cases, t pretty good (*robust*).
- Draw a picture (eg. histogram) first!

Example

A diet guide claims that the average number of calories in a serving of vanilla yogurt is 120 or less. 14 brands of vanilla yogurt are randomly sampled; the sample mean is 157.9 and the sample SD is 44.8. Do these data offer evidence that the diet guide is wrong?

Let μ be the mean number of calories per serving of all brands of vanilla yogurt. Testing $H_0: \mu = 120$ against $H_A: \mu \neq 120$.

Test statistic $t = (157.9 - 120)/(44.8/\sqrt{14})$ = 3.17.
degrees of freedom 14-1=13
P-value less than 0.01 (off end of table)
Conclusion: reject null (mean is 120) in favour of alternative (mean not 120)

Confidence interval for the population mean (p.620)

Similar idea as for proportions:

- best guess at population mean is sample mean $\bar{y}$
- make interval that goes up and down depending on uncertainty in $\bar{x}$
- have to use t-distribution when using sample SD s.

So: CI is $\bar{y} \pm t^* \frac{s}{\sqrt{n}}$.

Yogurt data: n=14, $\bar{y} = 157.9$, s=44.8. With 13 df, $t^* = 2.160$ for 95% interval (look at *bottom* of Table T). Hence margin is $m = 2.160(44.8/\sqrt{14})$ = 25.9 and interval (132.0, 183.8).

Know that population mean almost certainly above 120 but don't know precisely what it is. (A lot of variability in data plus small sample.)

Test vs confidence interval (p. 632)

Another example: data n=30, $\bar{y}=40$, s=10. Test $H_0:\mu=35$ vs $H_A:\mu\neq35$ at $\alpha=0.05$: gives t=2.74, P-value close to 0.01. Reject null and conclude that $\mu\neq35$.

95% CI for μ: (36.3, 43.7). 35 *outside* this interval, once again conclude that 35 not plausible value for μ.

This works, if:

- test two-sided
- α for test and confidence level match up (eg. 0.05 and 95%).

P-value close to 0.01, so 35 should be right on edge of 99% CI, and is:

Options

99% confidence interval results:

μ : population mean

Mean	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
μ	40	1.8257419	29	34.967552	45.032448

Sample size and confidence intervals (p.633)

When planning a study, want to know what sample size to use.

Take margin of error $t^* \frac{s}{\sqrt{n}} = m$, say, solve for n to get

$$n = \left(t^* \frac{s}{m} \right)^2$$

But: don't know s, and don't know what df to use for t^* .

Use a guess at the standard deviation, and start with z^* (infinitely large df) instead of t^* . This will give a value of n. Use that to get df for t^* , and repeat.

Example

In our yogurt data above, (n=14, $\bar{y}=157.9$, s=44.8), the margin of error for a 95% CI was $m=25.9$. How big a sample would we need to reduce this margin of error to 20, if everything else stayed the same?

Step 1: don't know df, so use $z^*=1.96$ and calculate

$n=\left(1.96\frac{44.8}{20}\right)^2$ = 19.28.

Now n=19 approx, so use 18 df for t: $t^*=2.101$.

Recalculate:

$n=\left(2.101\frac{44.8}{20}\right)^2$ =22.15

Round *up* to be safe: a sample size of 23 should do it.

We know that sample size 14 gives $m = 25.9$, so another way:

- want to multiply margin of error by $20/25.9$ = 0.77, so *divide sample size by* 0.77^2 = *0.59.* That is, sample size should be $\frac{14}{0.77^2}$ = 23.48, almost as above.
- Not quite same as above because we didn't adjust for change in t^* in changing sample size.
- "Inverse square law": eg. if you want to cut margin of error in half, have to multiply sample size by *4*.

Sign Test (p.636)

What if we are not sure about using the t test, maybe because our data are skewed? We might also have doubts about basing test on *mean*: what about median?

Say $H_0: median = 50$. If H_0 true, each sample value equally likely to be above or below 50. Say n=10. Then number above is *binomial* with n=10, p=0.5, and use this to get P-value: **sign test**.

Makes no assumptions about data being normal.

Example

Yogurt data are: 160, 130, 200, 170, 220, 190, 230, 80, 120, 120, 80, 100, 140, 170.
Test $H_0: median = 120$ vs $H_A: median > 120$. 2 values are less than 120, 10 are greater, 2 exactly equal (throw away). On binomial with n=12, p=0.5, with X=number greater than 120:

- P(X=10) = 0.0161
- P(X=11) = 0.0029
- P(X=12) = 0.0002

Add these up to get P-value 0.0193. Again reject H_0 : the diet guide seems wrong.

P-value for sign test not as small as one for t-test (usually the case) but no assumption of (near-)normality.

Sign test *not as powerful as t test* if data normal, so when t test applies, should use it.

Power of t-test and sign test

Use yogurt example again.

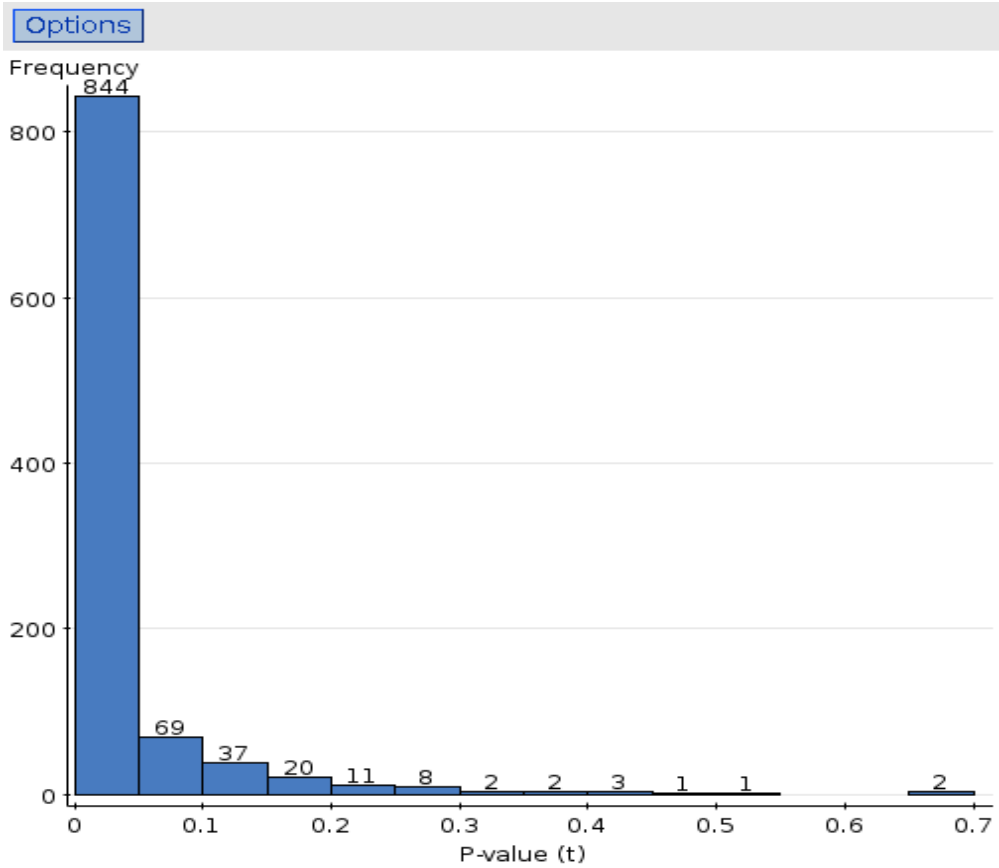
How likely are we to reject $H_0: \mu = 120$ in favour of a one-sided alternative ("greater") if distribution of yogurt calories per serving actually normal with mean 150, SD 50 and we take samples of size 20?

Do by simulation in StatCrunch:

- generate many values from Normal(150,50) to be our population
- generate many samples of size 20 from this population, saved in one column with sample ID
- calculate mean and SD of each sample
- calculate t statistic and P-value for each sample
- count number of values over 120 in each sample, and get P-value for it

– for t test and sign test, count number of rejections and compare.

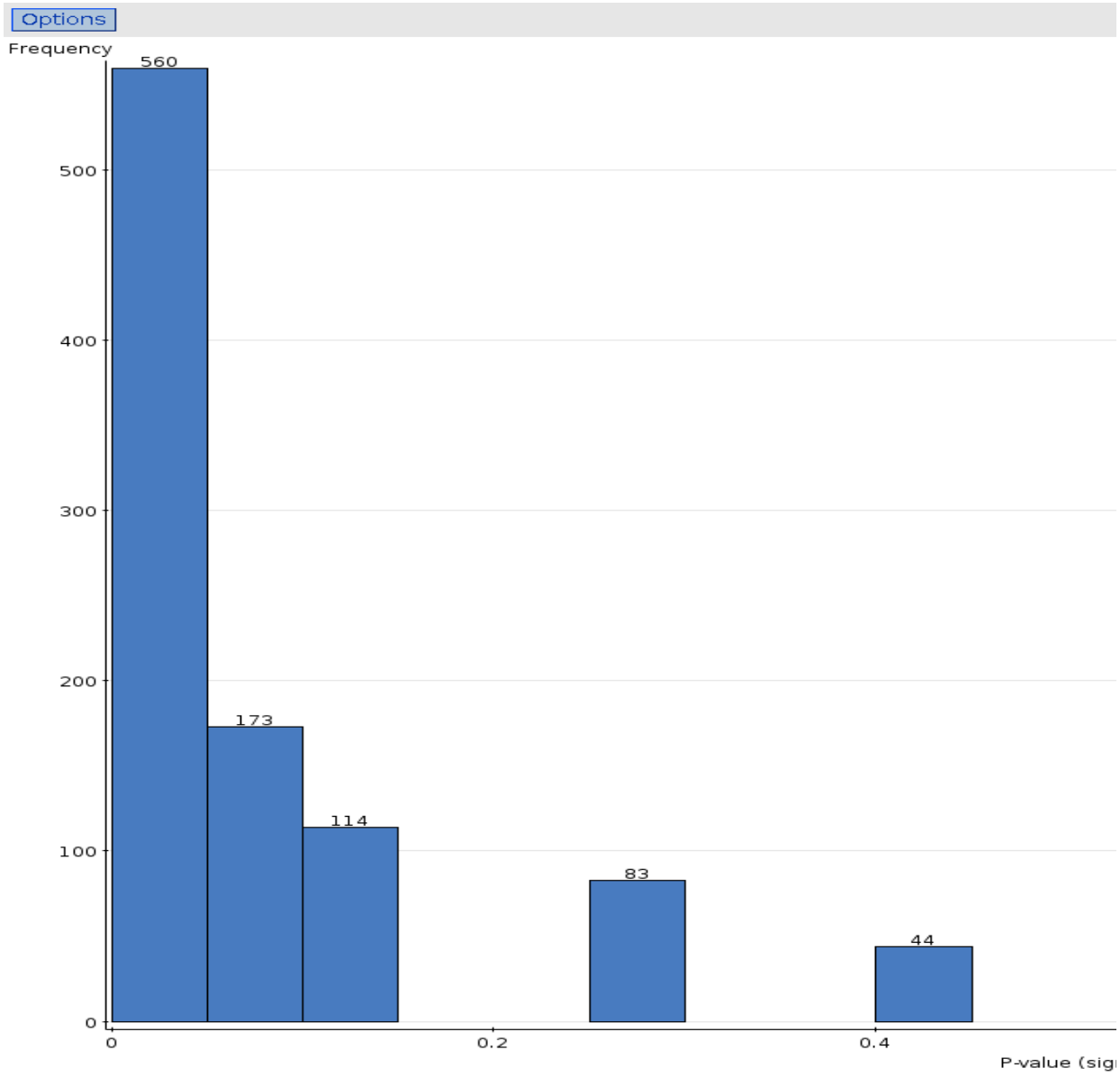
Results for t test:



844 P-values of 1000 were less than 0.05, so power about 84%.

Results for sign test:

- 560+173=733 P-values less than 0.05, so power around 73%.
- Less than power of t test (84%).
- In both cases, good chance of correctly rejecting null.



Chapter 24: Comparing two means (p.654)

- Seen that most useful results come from comparing two groups, eg. treatment vs control.
- How to make CI or test for *difference between two means*?

CI, generally, is
$estimate \pm t * SD(estimate)$. What are those two things here?

For estimate, use sample means, using 1 and 2 to indicate the two samples:

$\bar{y}_1 - \bar{y}_2$

Suppose group 1 has population SD σ_1 and group 2 has population SD σ_2 . Suppose also we have n_1 observations in group 1 and n_2 in group 2.

What is $SD(\bar{y}_1 - \bar{y}_2)$? Well, we know that

$$var(\bar{y}_1 - \bar{y}_2) = var(\bar{y}_1) + var(\bar{y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Now, we don't know σ_1 or σ_2, so we have to replace them with the sample Sds. That gives us this:

$$estimate = \bar{y}_1 - \bar{y}_2 \quad , \quad SD(estimate) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad .$$

A t distribution is approximately correct for this.
For df, base on smaller sample (or use scary formula at bottom of p. 657).

Example

In a study to compare differences in resting pulse rates between men and women, the sample of 28 men had a mean of 72.75 and an SD of 5.37; the sample of 24 women had a mean of 72.625 and an SD of 7.70. What is a 90% confidence interval for the difference in population means? (A boxplot is shown on p. 681, suggesting not much difference.)

Based on 24-1=23 df, $t^*=1.714$. Working with men minus women, the estimate for the difference in population means is $72.75-72.625$ = 0.075, and our standard deviation is $\sqrt{\frac{5.37^2}{28}+\frac{7.70^2}{24}}$ = 1.871. Thus the interval is $0.075\pm(1.714)(1.871)$ or (-3.13, 3.28). "No difference" *entirely* plausible.

Test to compare two means

Same story: $t=\frac{estimate-null}{SD(estimate)}$, compare with right t distribution.

The null is "no difference": $H_0:\mu_1-\mu_2=0$ or $H_0:\mu_1=\mu_2$ and H_A is here $H_A:\mu_1-\mu_2\neq 0$ or $H_A:\mu_1\neq\mu_2$. Can have one-sided alternative if needed.

Filling in everything gives

$$t=\frac{\bar{y}_1-\bar{y}_2-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

and get P-value from t with df based on smaller sample.

Example: 50 each of males and females had to place strangely-shaped pegs into matching holes. The number of pegs placed in one minute are as shown:

	Males	Females
Sample size	50	50
Sample mean	19.39	17.91
Sample SD	2.52	3.39

Is there evidence that the mean number of pegs is different for males and females?

Males=1, females=2: $H_0: \mu_1 = \mu_2, H_A: \mu_1 \neq \mu_2$.

$$t = \frac{(19.39 - 17.91)}{\sqrt{\frac{2.52^2}{50} + \frac{3.30^2}{50}}} = 2.52,$$

P-value (49 df, use 45) between 0.01 and 0.02.

At $\alpha = 0.05$ have evidence of difference in means (but not at

$\alpha = 0.01$).

One more example: Lianas are woody vines that grow in tropical rainforests. Researchers measured liana abundance (stems per hectare) in the central Amazon region of Brazil. Each area was classified as "near" (the edge of the rainforest, less than 100m away) or "far" from the edge of the rainforest. The researchers are looking to see whether liana abundance is higher near the edge of the rainforest.

Let 1 represent "near" and 2 represent "far". μ is population mean for each group.

Hypotheses are: $H_0: \mu_1 = \mu_2$, __ $H_a: \mu_1 > \mu_2$ _____.

Data: the 34 "near" liana areas have mean 438 and SD 125; the 34 "far" areas have mean 368 and SD 114.

$\bar{y}_1 - \bar{y}_2$ is *438-368=70*

SD of $\bar{y}_1 - \bar{y}_2$ is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ = *29.01*

Test statistic is t = *(70-0)/29.01=2.41*

Correct side or wrong side? *Correct side (diff in means +)*

Df = *34-1=33 (use 32)* ; P-value = *0.01-0.025*

Decision: reject null at alpha=0.05, conclude near has greater liana density than far

We can also find a 95% confidence interval for the difference in means:

With 33 df, t* = *2.037*

Margin of error *2.037 x (29.01)=59.09*

CI from *70-59.09=10.91* to *70+59.09=129.09*

.

STAB22 April 1, 2015

Today, I (Ken Butler) am pretending to be Srishta Chopra.

Chapter 25: Paired Samples and Blocks (p.688)

What about this?

Company institutes exercise break for its workers, assess workers' satisfaction before and after implementation of exercise program. Want to prove satisfaction *higher* afterwards.

Worker	1	2	3	4	5	6	7	8	9	10
Before	34	28	29	45	26	27	24	15	15	27
After	33	36	50	41	37	41	39	21	20	37

Two samples of 10 workers, use methods of last chapter?

NO! the same 10 workers were assessed before and after, so don't have two separate samples (would use 10 *different* workers before and after). **Paired data.**

Analysis: a piece of cake! Add a row to table:

Worker	1	2	3	4	5	6	7	8	9	10
Before	34	28	29	45	26	27	24	15	15	27
After	33	36	50	41	37	41	39	21	20	37
Difference	-1	8	21	-4	11	14	15	6	5	10

- For each worker, find differences after minus before.
  - Then have *one* sample of differences
- Test to see whether population mean diff could be zero.
  - 10 differences have mean 8.5 and SD 7.47
  - μ pop mean diff; $H_0: \mu = 0$, $H_A: \mu > 0$
  - $t = \frac{8.5 - 0}{7.47/\sqrt{10}}$ = 3.6, 9 df, P-value < 0.005.
  - **Reject** null hypothesis: conclude that mean satisfaction after *is* higher than before (for all workers).

- Or: 95% CI for population mean diff:
  - $8.5 \pm 2.262(7.47/\sqrt{10})$, 8.5 ± 3.54 = 4.96 to 12.04.

- Assumption: differences come (approx.) from normal distribution.
- Look at eg. histogram of differences.
- If differences not normal enough for you, do *sign test* on differences instead.

Matched pairs and two-sample (see box p.701)

The design of controls and instruments has a large effect on how easy they are to use. A sample of 25 right-handed students were asked to turn a knob a fixed distance (with their right hands). There were two identical knobs, one which turned clockwise and one counterclockwise. The times for each student for the two knobs are summarized below.

	n	Mean	SD
Clockwise	25	104.12	15.79
C-clockwise	25	117.44	27.26
difference	25	-13.32	22.94

Find a 95% CI to compare times. Is this matched pairs or two independent samples?

Each student gave two measurements, which we can pair up by student. So Matched Pairs.

- df = 25-1=24
- t* = *2.064*
- margin of error m=(2.064)(22.94/sqrt(25))=9.47
- CI -13.32 plus/minus 9.47 = from -22.69 to -3.85

Clockwise turns quicker than counter-clockwise ones on average.

Matched pairs and two-sample again

Physical fitness is related to personality. College faculty were divided into high-fitness and low-fitness groups. They were each given an "ego-strength" personality test, with the results summarized below:

	n	mean	SD
High-fitness	14	6.43	0.43
Low-fitness	14	4.64	0.69
Differences	14	1.79	0.73

Want to use a 95% CI to compare ego strengths of the high-fitness and low-fitness groups. Is this matched pairs or two independent samples?

Each faculty member contributes only one measurement, so there are 28 faculty members in the study altogether. This is two independent samples.

(How were the differences calculated anyway?)

t=2.160;* $6.43-4.64\pm2.160\sqrt{\frac{0.43^2}{14}+\frac{0.69^2}{14}}$ *. Sq root is 0.22; CI is from 1.31 to 2.27.*

Mean ego strength higher for high-fitness individuals than low-fitness.

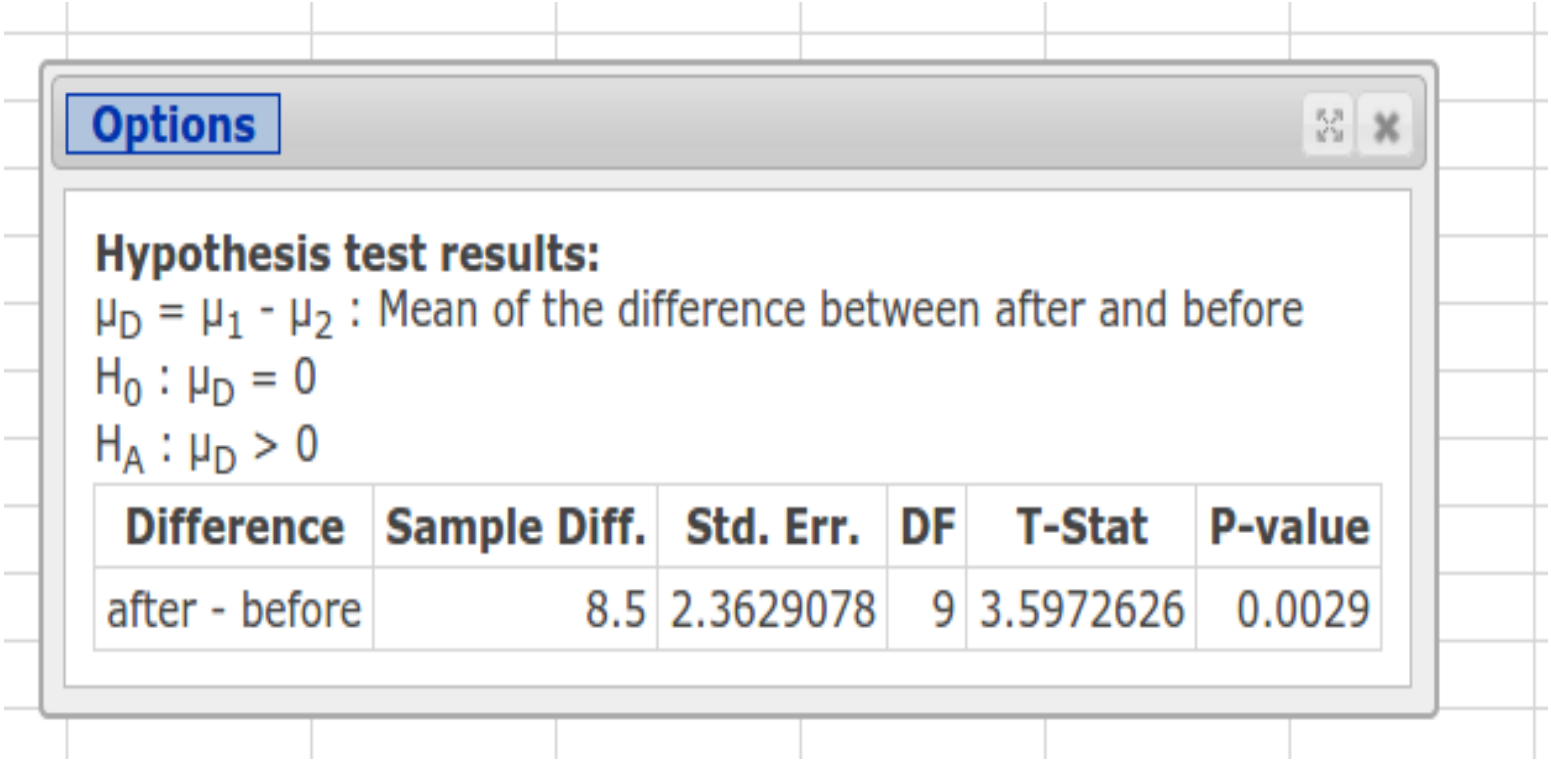
When the sample sizes are different, it must be two independent samples (no way of pairing them up), but when the sample sizes are the same, it could be either.

Effect of doing the wrong thing

Go back to workers' exercise program:

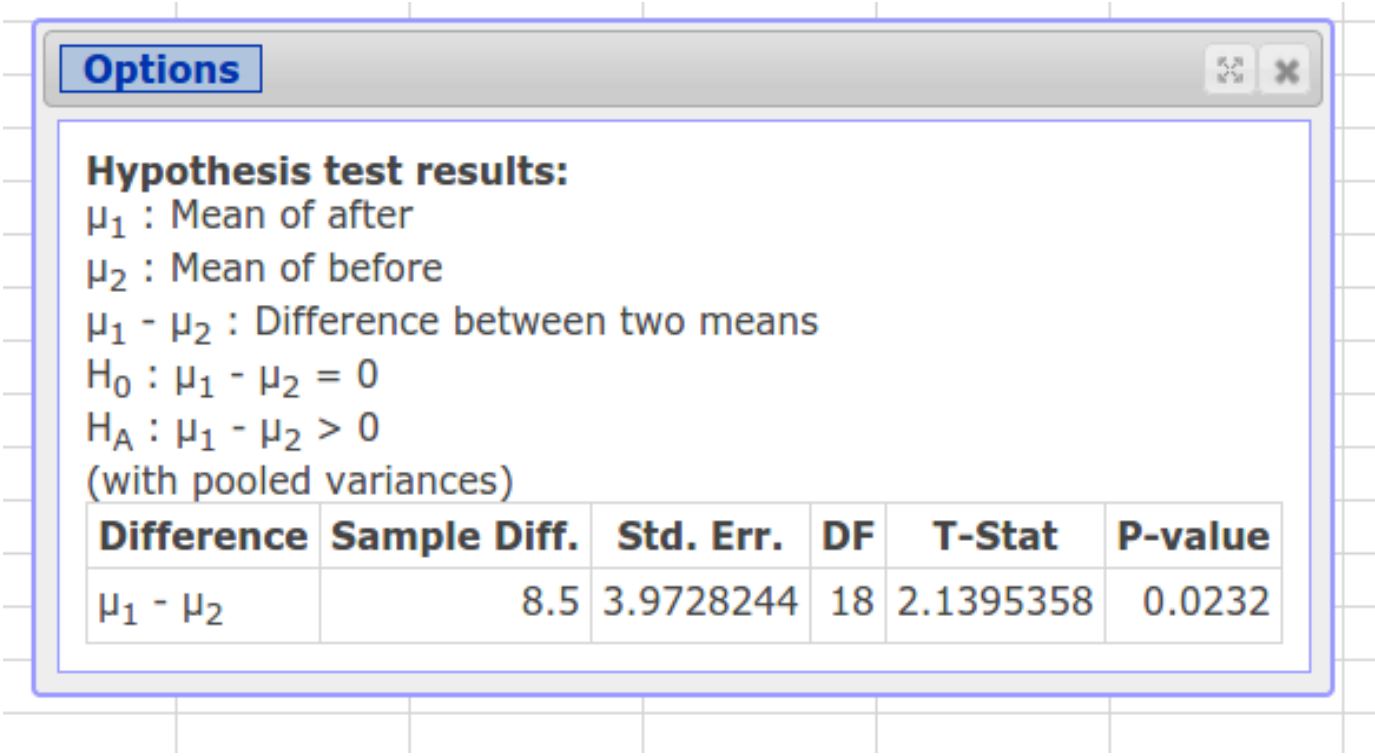
Worker	1	2	3	4	5	6	7	8	9	10
Before	34	28	29	45	26	27	24	15	15	27
After	33	36	50	41	37	41	39	21	20	37
Difference	-1	8	21	-4	11	14	15	6	5	10

Matched pairs



Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
after - before	8.5	2.3629078	9	3.5972626	0.0029

Two-sample: P-value bigger (less significant). Doing the wrong test gets you a less significant P-value.



Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	8.5	3.9728244	18	2.1395358	0.0232

Tests and CIs

Inference for	SD to use (for CI if different)	P-value from	For CI use
Proportion	$\sqrt{\frac{p(1-p)}{n}}$; $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	normal	z^*
Mean	$s/\sqrt{n}$	t, df n-1	t^*
Two means	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ $H_0: \mu_1-\mu_2=0$	t, df $min(n_1-1, n_2-1)$	t^*
Matched pairs	differences, $H_0: \mu=0,$ as for mean.		

We are done!

:-)