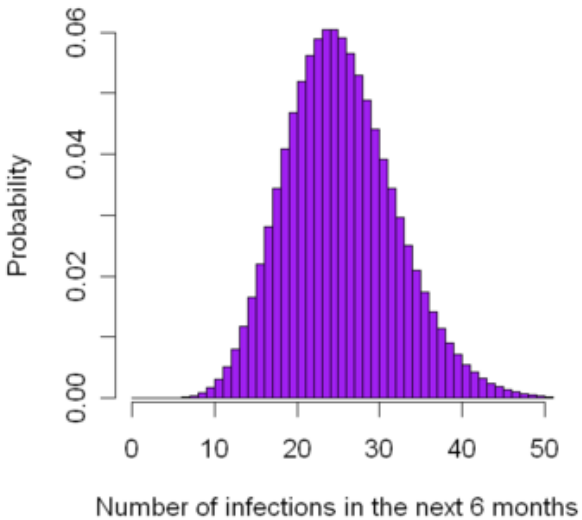
STAB22
Introduction to Statistics

Instructor: Ken Butler

butler@utsc.utoronto.ca

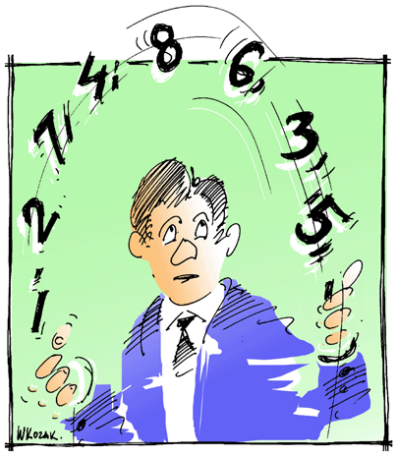
September 12, 2011

What Statistics is



- Data: collecting, organizing, interpreting.
- Understand world, choose how to act.
- Separate sense from nonsense.

In this course



- Learn about most important concepts and techniques in statistical work.
- Provides understanding of use of Statistics in your own field.
- Gain understanding from data.
- From data through analysis to conclusions.

What to expect

- Learn concepts, their interrelationships.
- Some calculation.
- Interpretation of software output.
- As little math as we can manage!

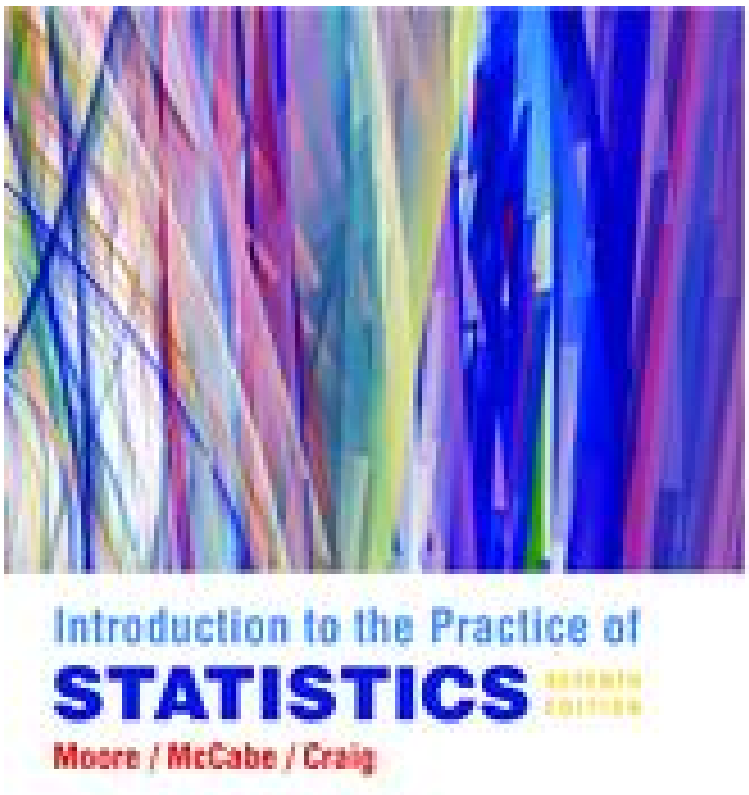
The instructor

- Ken Butler, e-mail `butler@utsc.utoronto.ca`.
- Office: IC471
- Office hours: after class (11:00-12:15 Tu & Fr)
- E-mail will always reach me.

This course

- 2 lectures a week.
- 1 tutorial a week, starting next week.
- Tutorials: help practice what you learned, also quiz.
- Assigned problems with full solutions (not graded).
- Midterm and final exams, multiple choice.

The text



- Introduction to the Practice of Statistics
- Moore, McCabe and Craig
- 7th ed., publ Freeman.

Software

- We use StatCrunch software (on web).
- Requires access code (I give you in class).
- Better tested than Excel, easy to learn.
- I show you what to do.
- Learning StatCrunch enables you to analyze realistic data.

Assessment

Item	Percent of grade
Quizzes	20%
Midterm exam	30%
Final exam	50%

Quizzes:

- in each tutorial except last.
- 10 quizzes, best 9 count for your grade.
- Intended to be straightforward if you are keeping up with material.

Usually 20%–25% A's in the course. Less than 5% of all students who complete the course work fail.

The exams

- *Multiple choice*, about 40 questions (midterm), 60 (final).
- Allowed "cheat sheets": 1 for midterm, 2 for final, but no other books/notes.
- You'll need a *calculator*:

Makeups

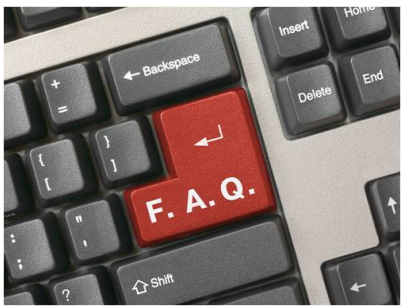
- There are no makeup exams or quizzes.
- With documentation:
  - missed quizzes ignored in calculation of average.
  - weight of missed midterm exam taken on final exam.
- Without documentation: any missed work scores 0.

What we cover

Chapters 1–8 of the text:

1. Looking at data — distributions
2. Looking at data — relationships
3. Producing data
4. Probability — the study of randomness
5. Sampling distributions
6. Introduction to inference
7. Inference for distributions
8. Inference for proportions

Frequently asked questions



Currently living at:

`http://www.utsc.utoronto.ca/~butler/b22/faq.html`

Check before you ask your instructor.

Lecture 1: §1.1 (a small part)

Coming up:

- What makes up a data set?
- What are cases?
- What are variables?
- What kinds of variables are there?

Questions

1. Here is a data set. What do the rows and columns of the table represent?

Car	MPG	Weight	Cyl.	HP	Country
Buick Estate Wagon	16.9	4.360	8	155	U.S.
VW Rabbit	31.9	1.925	4	71	Germany
Mercury Zephyr	20.8	3.070	6	85	U.S.
Fiat Strada	37.3	2.130	4	69	Italy
Mazda GLC	34.1	1.975	4	65	Japan
Saab 99 GLE	21.6	2.795	4	115	Sweden
BMW 320i	21.5	2.600	4	110	Germany

Car	MPG	Weight	Cyl.	HP	Country
Buick Estate Wagon	16.9	4.360	8	155	U.S.
VW Rabbit	31.9	1.925	4	71	Germany
Mercury Zephyr	20.8	3.070	6	85	U.S.
Fiat Strada	37.3	2.130	4	69	Italy
Mazda GLC	34.1	1.975	4	65	Japan
Saab 99 GLE	21.6	2.795	4	115	Sweden
BMW 320i	21.5	2.600	4	110	Germany

2. In the data as above:
   (a) What is the difference between the variable `weight` and the variable `country`?
   (b) What is the difference between the variable `weight` and the variable `cylinders`?
   (c) What is meant by the *distribution* of the variable `cylinders`?

3. What else might we want to know about a data set?

Lecture 1 summary

- What makes up a data set?
  - *Cases and variables*
  - *Why, who and what*
- What are cases?
  - *People, animals or things of interest*
- What are variables?
  - *Quantities measured/counted/classified for each case*
- What kinds of variables are there?
  - *Categorical: places each individual in category*
  - *Quantitative: something counted or measured for each case*

To read for next time

§1.1, Displaying Distributions with Pictures

- Categorical variables: Bar graphs and pie charts
- Quantitative variables: Stemplots
- Histograms

Also, acquaint yourself with StatCrunch by watching eg.

`http://www.youtube.com/watch?v=blEK8Gc2YnA`

and visit `statcrunch.com`, click Subscribe and Redeem Access Code. The access code will be given out in class.

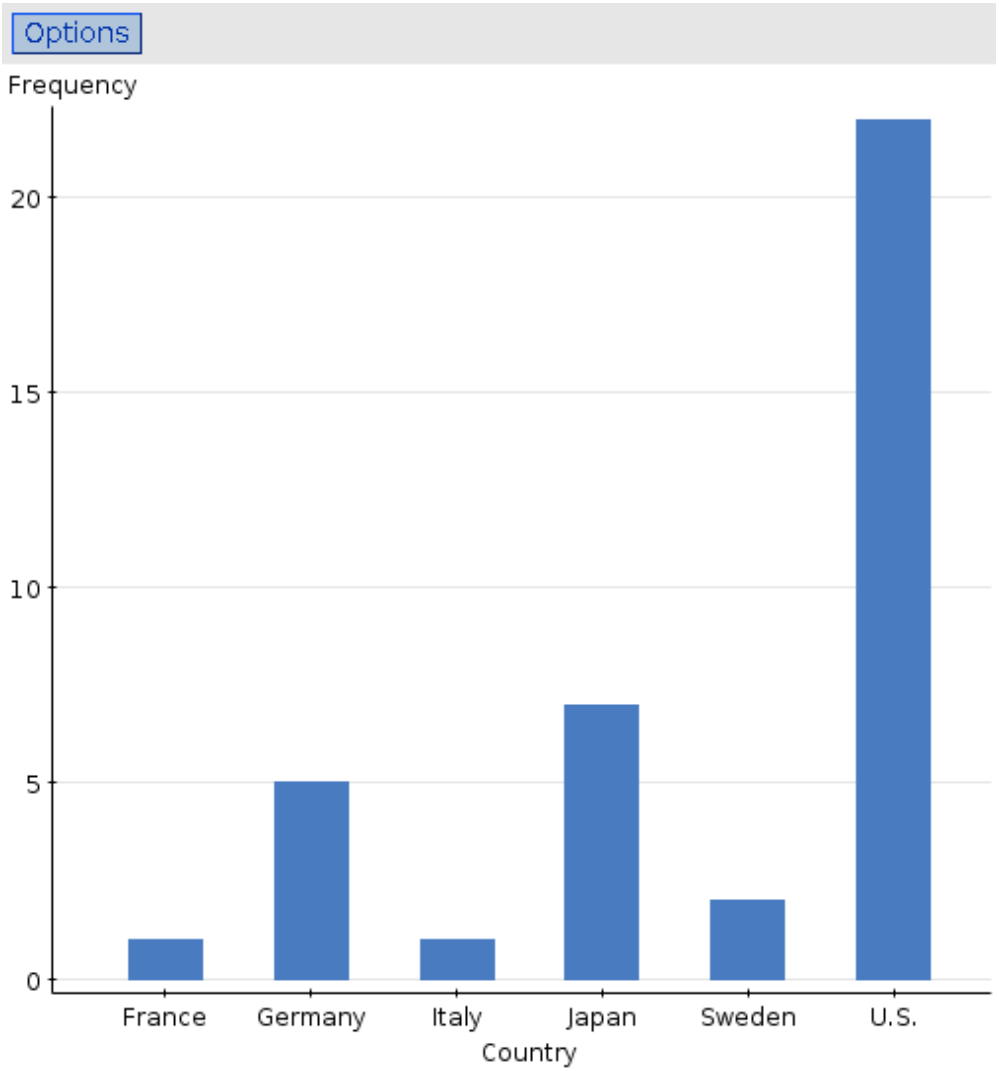
Lecture 2: §1.1

Coming up:

- What is exploratory data analysis?
- What is the distribution of a variable?
- How do we display categorical variables?
- How do we display quantitative variables?

Questions

1. Here's a **bar chart** of the categorical variable `country` for the cars data:

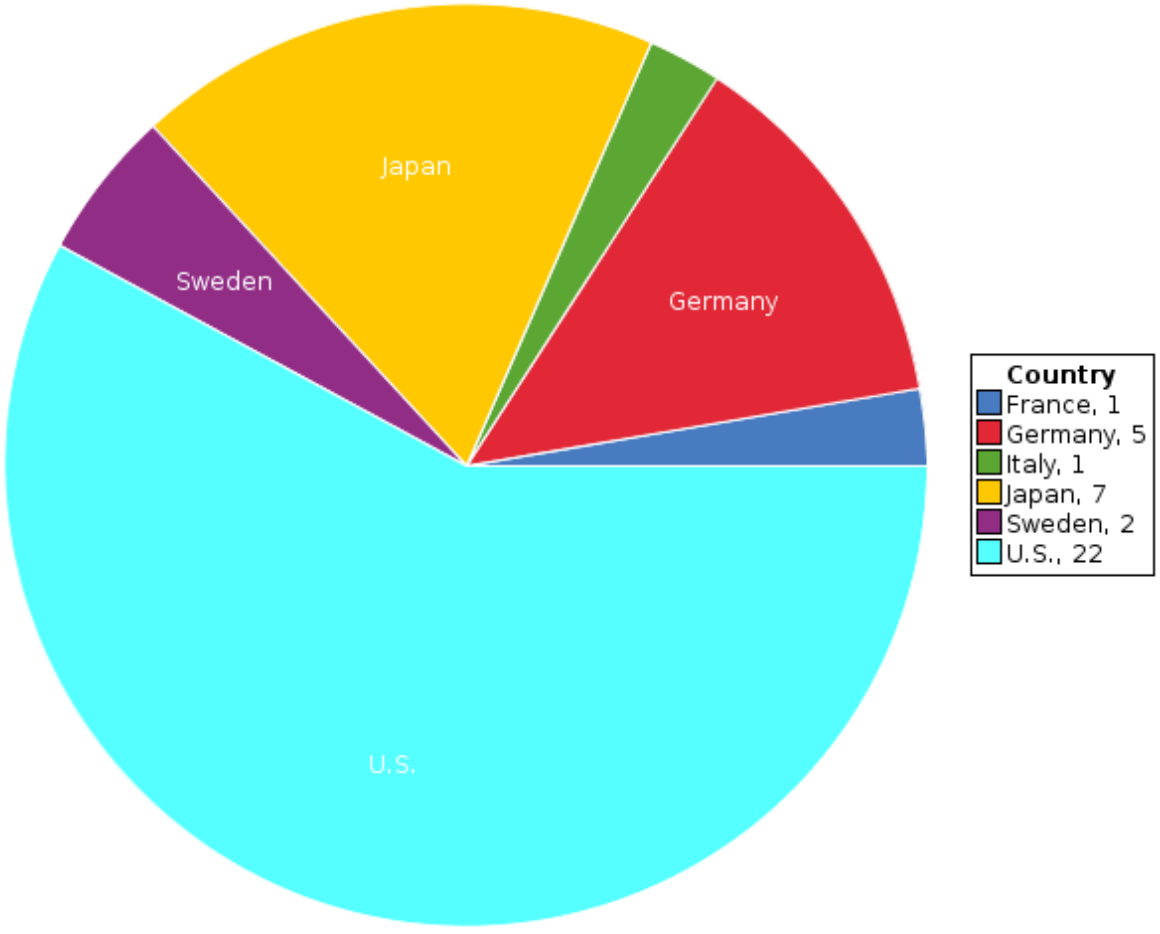


(a) What does the height of each bar represent?

(b) Which country has the most cars in the data set?

(c) Which countries have the fewest?

2. Below is a **pie chart** for `country`:



(a) Do more or less than a half of all cars in the data set come from the US?

(b) Do more or less than a quarter of all cars in the data set come from Japan?

3. Below is a stemplot for the quantitative variable `MPG`:

```
Variable: MPG
Decimal point is 1 digit(s) to the right of the colon.

1 : 667777888999
2 : 0112222
2 : 7777889
3 : 0011222444
3 : 57
```

   (a) How many cars in the data set have MPG of 35 or higher?
   (b) How many cars in the data set have MPG between 23 and 26 (inclusive)?
   (c) StatCrunch has two lines for each tens digit. How do you decide which units digit goes on which one?

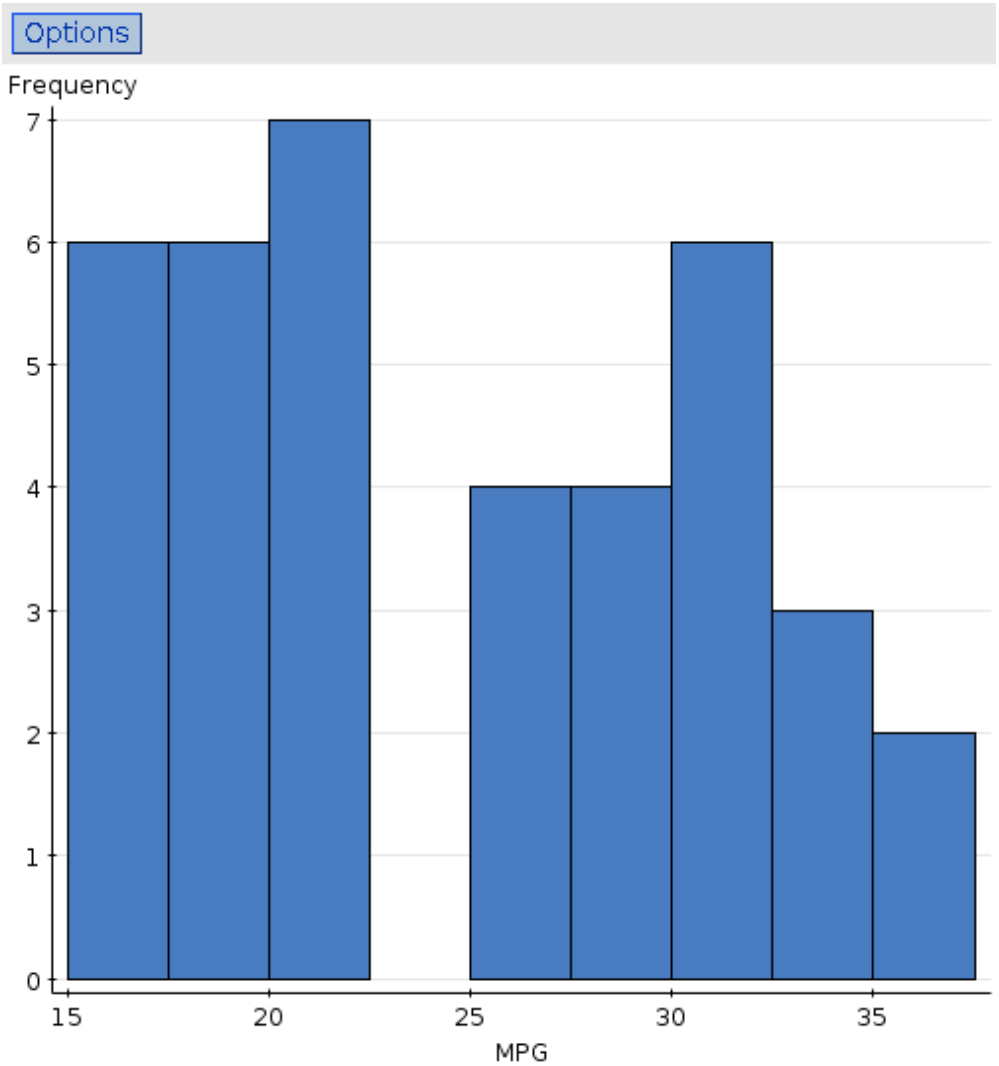
4. The first 10 MPG figures in the cars data are:

```
28.4  30.9  20.8  37.3  16.2  31.9  34.2  34.1  16.9  20.3
```

   (a) Truncate these values to 2 digits.
   (b) Draw a stemplot of these values using tens as stems and units as leaves. Arrange the leaves in order on each line.
   (c) Use your stemplot to count the number of car MPG figures that are 31 or more but (strictly) less than 37.
   (d) Explain why (a) enabled you to get the correct answer to (c) straight from the stemplot.

5. Using your answer to the previous question, split each stem into two parts. How would you split a stem into *five* parts?
6. What would happen if you used *hundreds* as stems and tens as leaves?

7. Below is a **histogram** for the MPGs of the cars data:



(a) How many cars have MPG between 20 and 22.5?

(b) What % of the 38 cars have MPG less than 20?

(c) How many cars have MPG between 25 and 30?

(d) Can you say how many cars have MPG between 26 and 30?

(e) What else can you learn from the histogram?

8. Open the cars data in StatCrunch.
   (a) Make a histogram of the car weights.
   (b) Make another histogram where the bars start at 1.5 and have width 0.2.
   (c) Compare the two histograms. Can you think of an advantage *and* a disadvantage to having more bars on a histogram?

Lecture 2 summary

- What is exploratory data analysis?
  - *Using graphs and numbers to describe variables in data set*
- What is the distribution of a variable?
  - *List of values taken by a variable and how often it takes each one*
- How do we display categorical variables?
  - *Bar graph*
  - *Pie chart* (for fractions of a whole)
- How do we display quantitative variables?
  - *Stemplot*
  - *Histogram*
  - *(later) boxplot*

To read for next time

§1.1, Displaying Distributions with Graphs

- Data analysis in action
- Examining distributions: shape, centre and spread
- Dealing with outliers

from §1.2:

- Boxplots

Also, practice using StatCrunch to reproduce the results of the last lecture.

Lecture 3: §1.1, the rest

Coming up:

- Describing a distribution by shape, centre and spread
- What are some different kinds of shape?
- What are outliers, and what do we do with them?
- What is a boxplot?

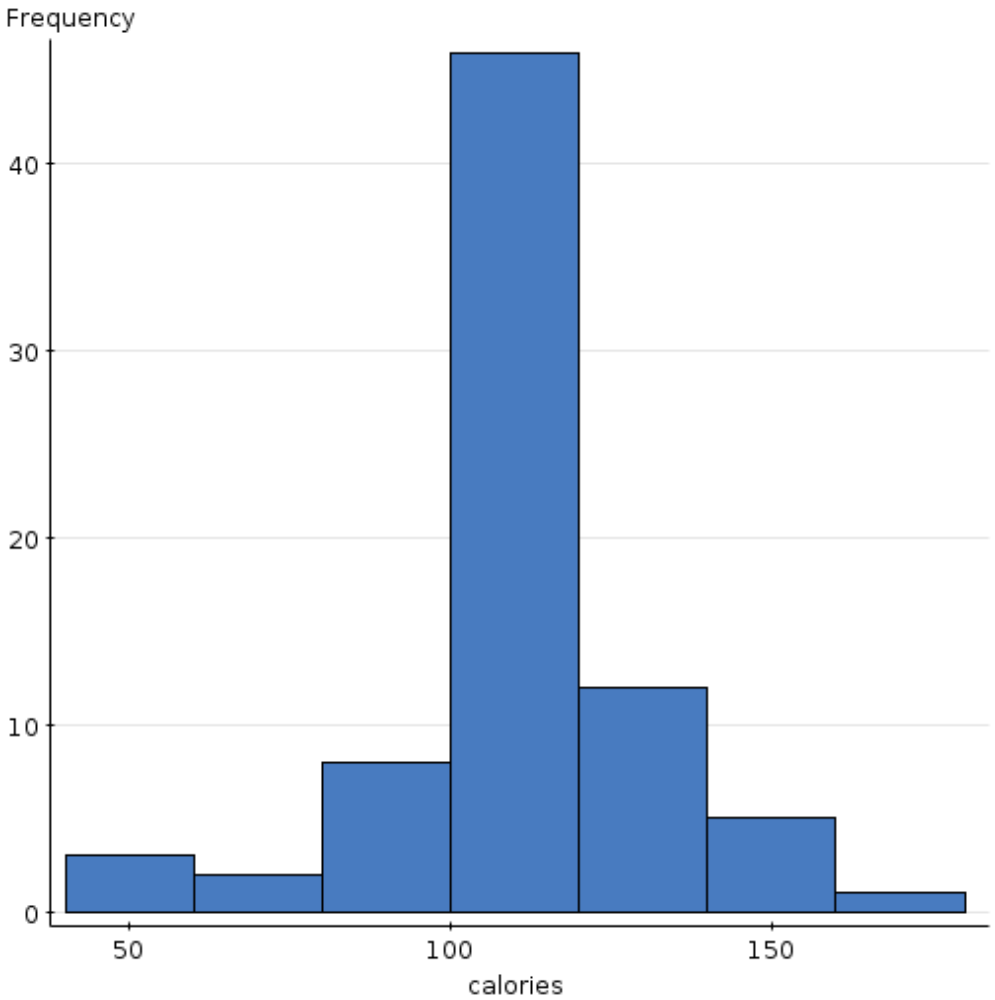
The breakfast cereal data

Different data: information on 77 kinds of breakfast cereal (individuals), lots of variables such as:

- calories per serving
- protein per serving
- fat per serving
- sodium per serving
- fibre per serving
- potassium per serving
- shelf on which found at supermarket (1, 2, 3)
- serving size (cups)

Questions

1. Below is a histogram of the calories per serving for the cereals.

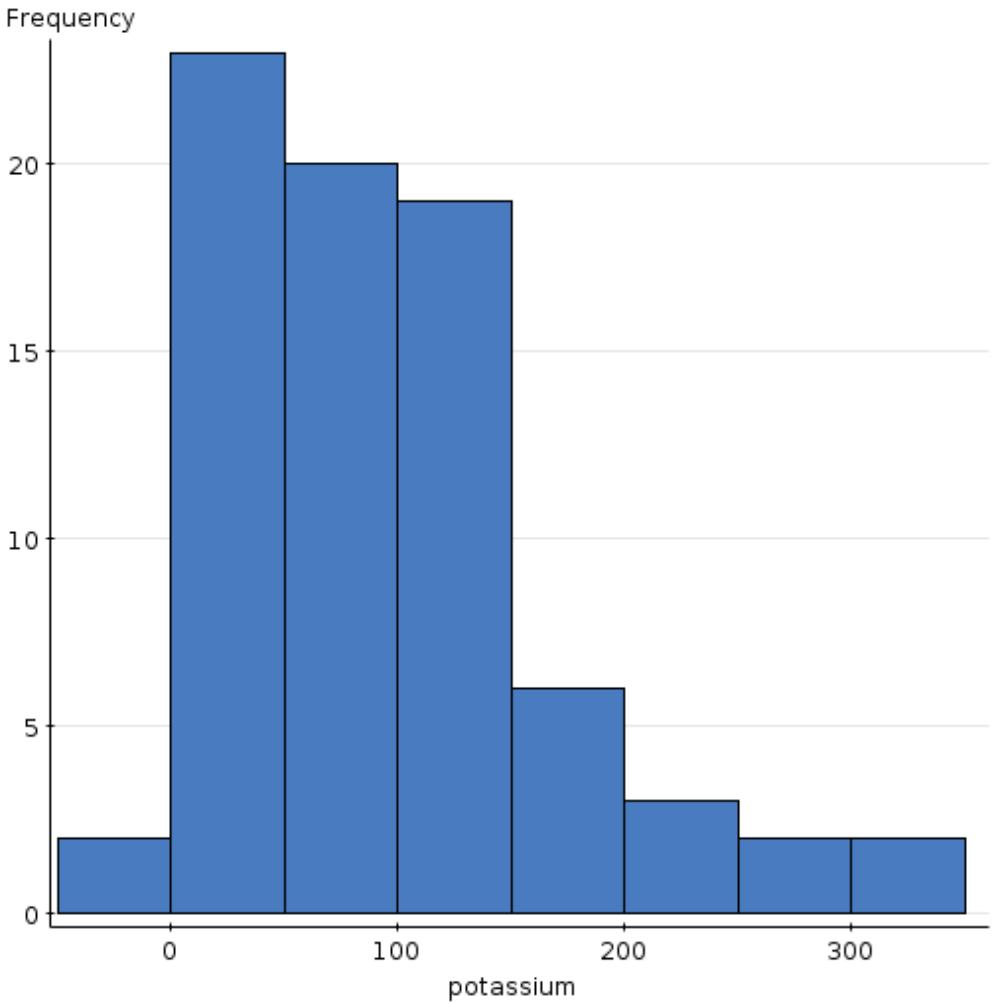


(a) Where would you say is the centre of the histogram?

(b) Would you say the spread is large or small (that is, are most of the values close to the centre or not)?

(c) Would you describe the shape as symmetric (same both sides of centre) or skewed (not)?

2. Below is a histogram of the potassium per serving for the cereals.



(a) Where would you say is the centre of the histogram?

(b) Would you say the spread is large or small compared to the calories histogram?

(c) Would you describe the shape as symmetric (same both sides of centre) or skewed (not)?

3. Below is a stemplot of the potassium per serving:

```
Stem and Leaf
Variable: potassium

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

0 : 00223333333444444444
0 : 55555666666778999999
1 : 00000001111111222233444
1 : 667799
2 : 034
2 : 68
3 : 23
```

(a) What is the highest potassium per serving in the data set?

(b) Use the stemplot to assess the centre, spread and shape of this distribution. Do you get similar conclusions as from the histogram?

(c) What extra information does the stemplot give you that the histogram does not?

4. Below is a stemplot of the calories per serving.

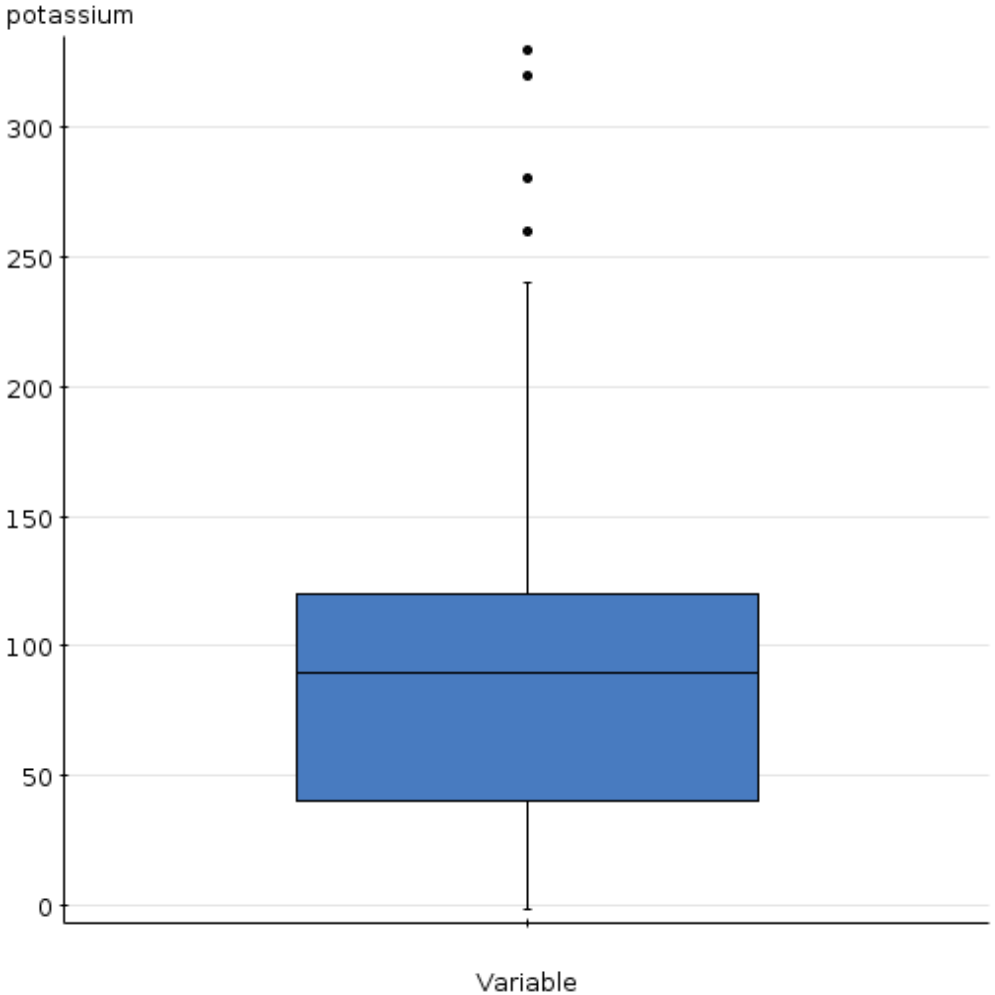
```
Variable: calories

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

Low : 50, 50, 50

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

High: 150, 150, 160
```

(a) Explain the "low" and "high" lines on StatCrunch's stemplot.

(b) Does the stemplot tell the same story as the histogram about the centre, spread and shape of the distribution?

(c) What do you learn about the measurement of the calorie content per serving from the stemplot that you didn't see on the histogram?

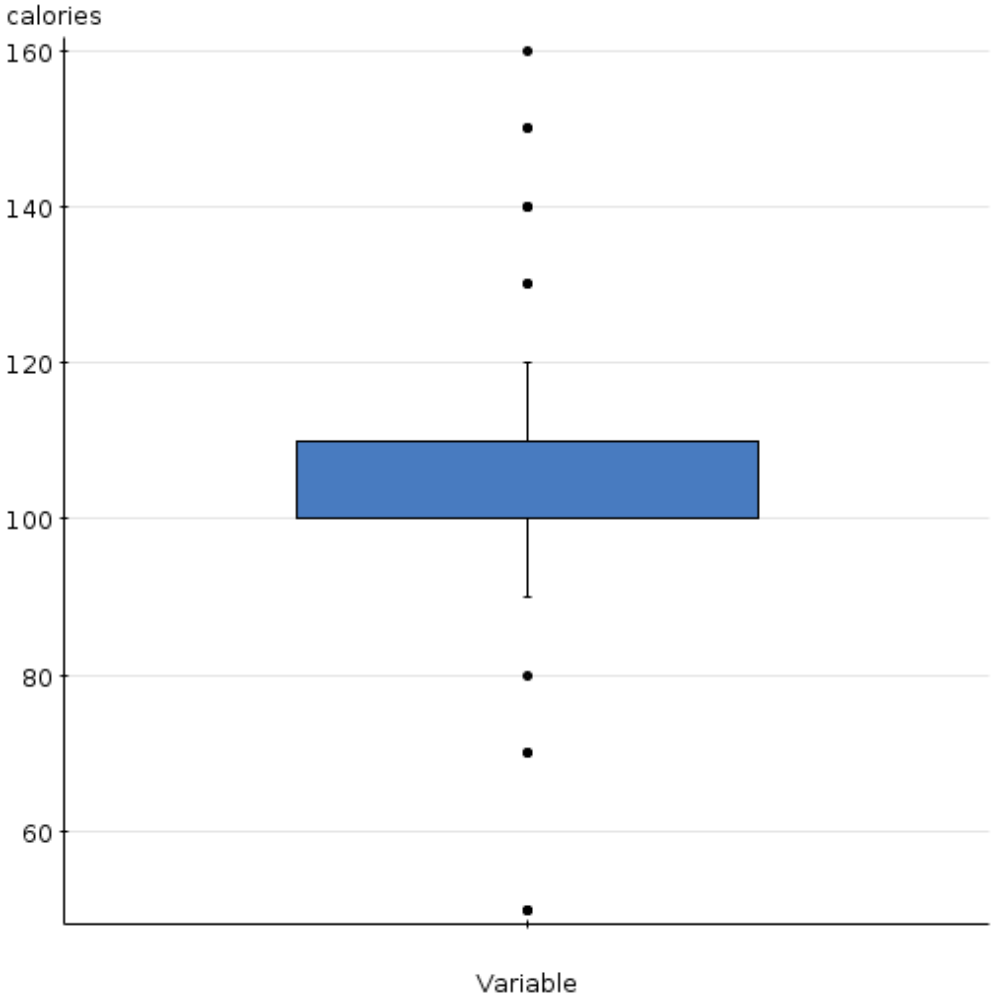
5. The potassium content values had centre just below 100, a large spread, and a right-skewed shape. A boxplot is shown below.



(a) Verify the statements in the question from the boxplot.

(b) What else does the boxplot tell you?

6. The calorie content values had a small spread, a symmetric shape, and outliers at both the top and bottom. A boxplot is shown below.



(a) What happened to the line across the middle of the box? (The centre is 110).

(b) How does this boxplot show the spread, shape and outliers for the calorie data?

7. Use StatCrunch to produce a histogram, stemplot and boxplot of the cereal fiber data. What do you conclude about the shape, centre and spread of the distribution? Are there any outliers? (When drawing the boxplot, be sure to check "use fences to identify outliers" before you click "create graph".)

Lecture 3: summary

- Describing a distribution by shape, centre and spread
  - *Centre: where "middle" of distribution is*
  - *Spread: whether values are close together or far apart*
  - *Shape: is distribution symmetric or skewed one way or the other?*
- What are some different kinds of shape?
  - *Symmetric: same both sides*
  - *Skewed right: the highest values seem unusually high*
  - *Skewed left: the lowest values seem unusually low*
- What are outliers, and what do we do with them?
  - *Values that seem "wrong" compared to rest of distribution*
  - *Try to explain them: look for cause*
- What is a boxplot?
  - *A plot designed to show:*
    - ★ *centre*
    - ★ *spread*
    - ★ *shape*
    - ★ *outliers*

To read for next time

§1.2:

- Describing distributions with numbers
- Centre: mean and median
- Mean vs. median
- Spread: quartiles and IQR
- 1.5 IQR rule for outliers
- Spread: standard deviation
- Choosing measures of centre and spread
- Changing scale of measurement

Lecture 4: §1.2

Coming up:

- How to measure centre of distribution?
- How to choose measure of centre?
- Using quartiles to measure spread
- What is the 5-number summary and how does it relate to boxplots?
- How to decide whether a value is an outlier
- What is the standard deviation?
- How to choose measures of centre and spread?
- What happens to measures of centre and spread if we make a linear transformation of the data?

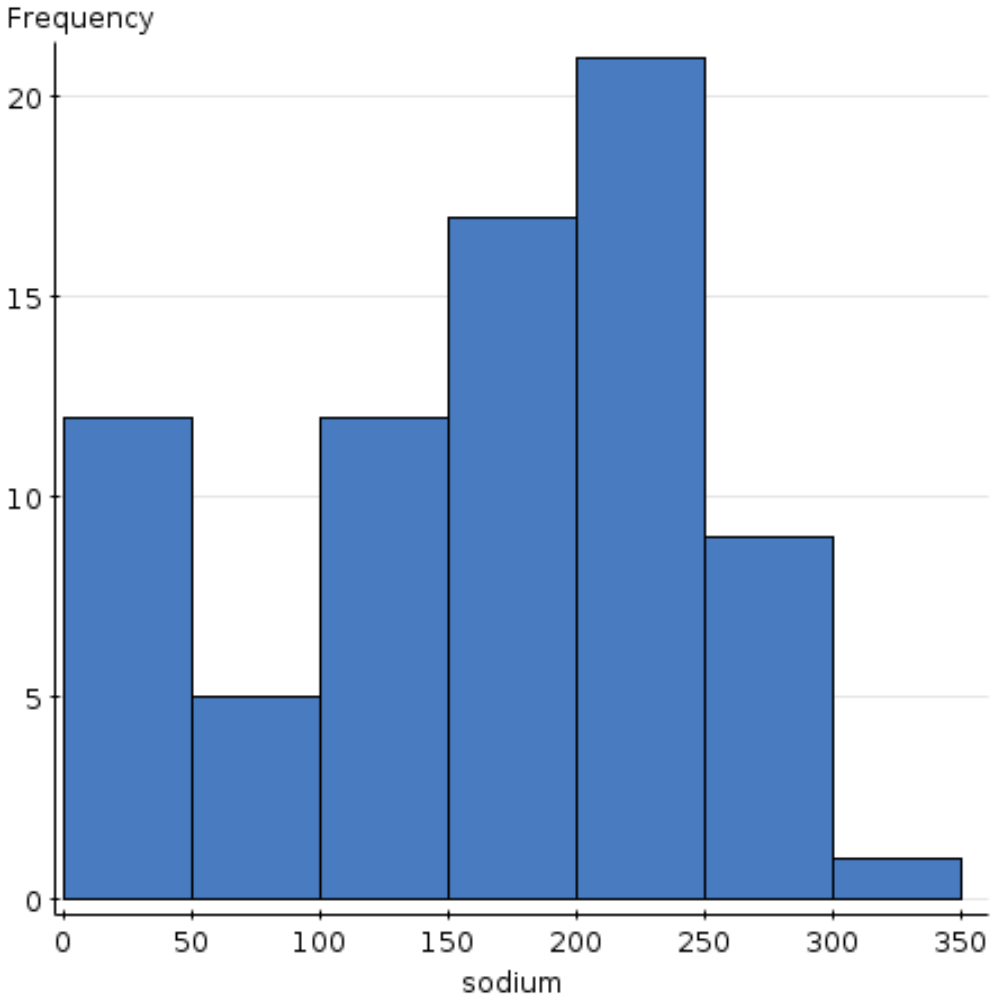
Questions

1. For the data 5, 3, 11, 1, 4:
   (a) Find the mean.
   (b) Find the median.
   (c) What would the median be if the data had been 5, 3, 11, 1, 4, 2?

2. For the cereal potassium data:
   (a) Use StatCrunch to find the mean and median.
   (b) Why do you think the mean is bigger than the median? (Hint: look at a graph of the data.)

3. Another variable is `carbo`, the amount of carbohydrates per serving. For this variable:
   (a) use StatCrunch to find the mean and median.
   (b) Do you think the distribution is symmetric or skewed? Now draw a boxplot. What do you think now?

4. A histogram of `sodium` is given below.



(a) Use the histogram to find the median sodium content of the cereals. (There are 77 cereals altogether.)

(b) Use StatCrunch to find the median, and verify that you were correct.

5. Can you make a general statement about:
   (a) When the mean and median will be similar.
   (b) When the mean will be bigger than the median.
   (c) When the mean will be smaller than the median.
   (d) When you should use the mean as a measure of centre.
   (e) When you should use the median as a measure of centre.

6. Find the quartiles Q1 and Q3 for the data sets below:
   (a) 10, 11, 13, 15, 18, 19, 21 (median 15).
   (b) What if data were 10, 11, 13, 15, 18, 19 (median 14).

7. Finding the five-number summary:
   (a) What five numbers make up the five-number summary of a variable?
   (b) For the two data sets in the previous question, write down the five-number summaries. (The data were (i) 10, 11, 13, 15, 18, 19, 21 and (ii) 10, 11, 13, 15, 18, 19.)
   (c) Use StatCrunch (and Stat, Summary Stats) to get the five-number summary for the cereal calories data.

8. Find the interquartile range for the data sets given below.
   (a) 9,9,9,9,9,9,9 ($Q1 = 9, Q3 = 9$)
   (b) Cereal calories data ($Q1 = 100, Q3 = 110$)
   (c) Cereal potassium data ($Q1 = 40, Q3 = 120$)
   (d) Compare the IQRs for `calories` and `potassium`. What does this tell you about the data?

9. Suppose a variable has a distribution with 80% of its values equal, but with extreme outliers at the upper and lower ends.
   (a) What would its inter-quartile range be?
   (b) How, in general, do you think the IQR responds to a small number of extreme values?

10. How do you determine if a data value is an outlier?
    (a) State the rule based on the IQR.
    (b) For the cereal calorie data, $Q1 = 100$, $Q3 = 110$. The lowest values are 50, 70, 80, 90; the highest are 120, 130, 140, 150, 160. Which of these are outliers? Which are the highest and lowest non-outliers?
    (c) Use StatCrunch to draw a boxplot of these data. Identify the outliers, and the highest and lowest non-outliers, on the boxplot. Do they agree with your calculations?

11. In the cars data, do cars with more cylinders in their engines generally have worse gas mileage? In StatCrunch, produce a boxplot of `MPG` grouped by `Cylinders` to assess this. What do you conclude? (This is called a side-by-side boxplot, and is useful for comparing distributions.)

12. Calculate the standard deviation of the numbers 4, 6, 8 (the mean is 6). When you are done, use StatCrunch to check your work.

13. Some questions about the standard deviation:
    (a) What is smallest possible SD?
    (b) Give a set of 3 numbers that has smallest possible SD.
    (c) Is there a largest possible SD?

14. In the cereal data, we found that the IQR for `potassium` was bigger than the IQR for `calories`. Use StatCrunch to find out whether the same is true for SD. Does your result make sense?

15. Start with set of numbers 10, 11, 12, 13, 15, 16, 17.
    (a) Using StatCrunch, find the mean, median, IQR and SD.
    (b) Change the 17 to something bigger. What happens to the mean and SD? What happens to the median and IQR?
    (c) Repeat the previous question, replacing 17 by something bigger still.
    (d) What does this tell you about when to use SD and when to use IQR?

16. Use StatCrunch to answer the questions below:
    (a) The numbers 2, 3, 4 have mean 3 and SD 1. Multiply all the values by 2 (to get 4, 6, 8). What are mean and SD now?
    (b) What appears to happen to mean and SD when you multiply by 2?
    (c) Take 2, 3, 4 and add 3 to all the values. What are mean and SD now?
    (d) What appears to happen to mean and SD when you add 3?

Lecture 4 summary

- How to measure centre of distribution?
  - *Mean or median*
- How to choose measure of centre?
  - *Mean if distribution symmetric with no outliers*
  - *Median if skewed shape or outliers*
- Using quartiles to measure spread
  - *Quartiles: Q1 has 1/4 of values below; Q3 has 3/4 of values below*
  - *IQR* $Q3 - Q1$ *gives idea of spread of data*
- What is the 5-number summary and how does it relate to boxplots?
  - *min, Q1, median, Q3, max*
  - *On boxplot, Q1, median and Q3 are bottom, middle and top of box*
- How to decide whether a value is an outlier
  - *Work out* $Q1 - 1.5 \times IQR$ *and* $Q3 + 1.5 \times IQR$
  - *Below 1st value or above 2nd considered outlier*

. . . Part 2

- What is the standard deviation?
  - *Measure of spread*
  - *Based on how far values tend to be from mean*
- How to choose measures of centre and spread?
  - *Symmetric, no outliers: mean and SD*
  - *Skewed or outliers: median and IQR*
- What happens to measures of centre and spread if we make a linear transformation of the data?
  - *Call transformation* $a + bx$
  - *Centre: apply same transformation to mean and median*
  - *Spread: multiply SD and IQR by* b

To read for next time

§1.3:

- Density curves and normal distributions
- Density curves
- Normal distributions
- Standardizing observations
- Normal distribution calculations
- Using Table A

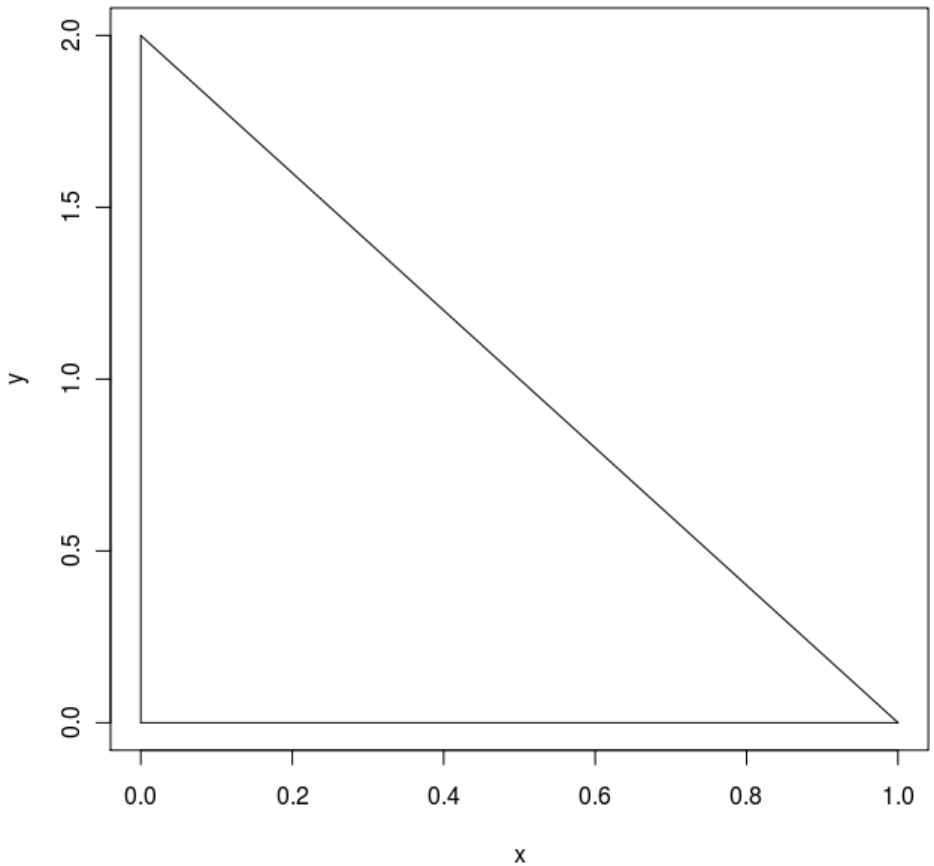
Lecture 5: §1.3

Coming up:

- What is a density curve?
- Where are the mean and median on a density curve?
- What kind of density curve is a normal distribution?
- How do you judge mean and SD on a normal distribution?
- How do you standardize a value, and what does it mean?
- What does Table A do, and how do you use it?

Questions

1. Below is a picture of a density curve. It is a triangle. The area of a triangle is half its base times its height.



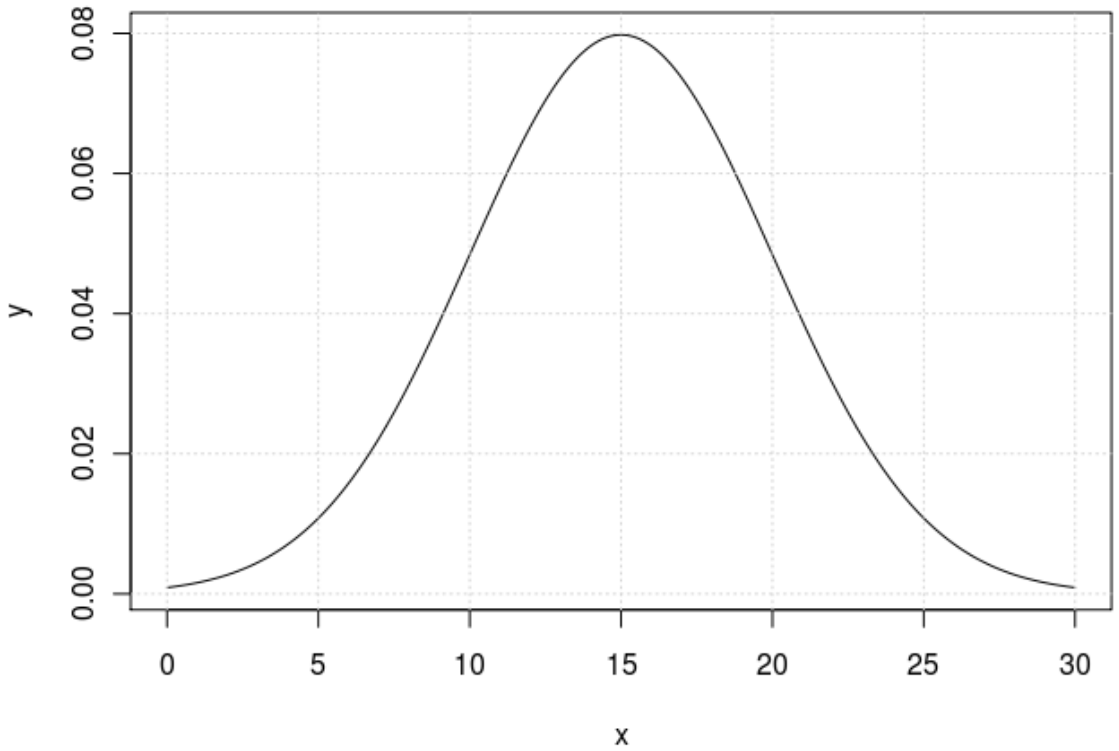
(a) Verify that this is a legitimate density curve.

(b) What proportion of the density curve lies to the right of 0.5? Why would you expect this to be less than a half?

(c) Verify that the median is about 0.3. (The height of the density curve at 0.3 is 1.4.)

(d) Do you think the mean would be bigger than 0.3, smaller than 0.3, or about equal to 0.3? Why?

2. What are mean and SD of the normal curve below?



3. What are three reasons why the normal distribution is useful?

4. The carbohydrate content per serving of the breakfast cereals has mean 14.6, SD 4.3, and approximately a normal shape. Corn Flakes' `carbo` content is 21.
   (a) Calculate the z score for Corn Flakes.
   (b) Your z-score should be positive. What does this tell you about the `carbo` content of Corn Flakes?

Note: this z-score means that only 7% of cereals have more carbos than Corn Flakes. We see later how this is worked out.

5. SAT scores have a normal distribution, mean 1026, SD 209. What proportion of students have scores 820 or less? Follow the steps below:
   (a) Obtain a z-score for 820.
   (b) Look the z-score up in Table A.
   (c) Does the table give you the proportion less, or greater? Do you want less, or greater? What, therefore, is the answer?

6. Cereal carbo content has a normal distribution with mean 14.6 and SD 4.3. Corn Flakes have carbo 21 per serving. We claimed that only about 7% of cereals had more carbos than this.
   (a) Verify this claim (by doing the calculation).
   (b) Use StatCrunch to verify the claim.

7. In SAT example, we want to find the proportion of students scoring between 700 and 800. (Mean 1026, SD 209.) Follow the steps below.
   (a) Draw a picture of the proportion you are trying to find.
   (b) Method 1:
      i. What proportion of students score less than 700?
      ii. What proportion of students score more than 800?
      iii. The two proportions you just found, plus the one you want, must add up to 1 (why?). What, therefore, is the proportion of students scoring between 700 and 800?
   (c) Method 2:
      i. What proportion of students score less than 700?
      ii. What proportion of students score less than 800?
      iii. Why can you find the answer you want by subtracting the answers you just found?
      iv. What, therefore, is the proportion of students scoring between 700 and 800?

Lecture 5: summary

- What is a density curve?
  - *A smooth curve that describes the overall pattern of a distribution*
  - *It has total area 1 under it.*
- Where are the mean and median on a density curve?
  - *Mean: balance point*
  - *Median: point that splits area in half*
- What kind of density curve is a normal distribution?
  - *Bell-shaped, symmetric, specified by its mean and SD*
- How do you judge mean and SD on a normal distribution?
  - *Mean: where the peak is*
  - *SD: distance from mean to "shoulders" of normal density*
- How do you standardize a value, and what does it mean?
  - $z = (x - \mu)/\sigma$*; how far above or below mean* x *is*
- What does Table A do, and how do you use it?
  - *Says what proportion of normal curve is less than value* z*.*
  - *Look up* z *value with 1st decimal place on left, 2nd decimal at top*

To read for next time:

§1.3:

- Inverse normal calculations
- The 68–95–99.7 rule
- Normal quantile plots

Lecture 6: §1.3

Coming up:

- How do you find a value from a proportion?
- What is the 68–95–99.7 rule and how does it work?
- How do you tell whether a normal distribution is reasonable for describing data?
- How can you approximate normal proportions without using Table A?

1. What SAT score do 10% (0.1000) of students score less than? The mean SAT score is 1026 and the SD of scores is 209. Do this by hand first, then check your answer with StatCrunch.

2. Scores on another test have a normal distribution with mean 200 and SD 50. What test score would get you into the top 3%? Again, check your answer with StatCrunch.

3. Explain carefully what the 68-95-99.7 rule says about normal distributions.

4. In a normal distribution with mean 10 and SD 3:
   (a) Proportion 68% of normal curve falls between which two values?
   (b) Proportion 95% of normal curve between which two values?
   (c) What % of values between 1 and 19?
   (d) What % of values between 10 and 13? (Draw picture.)
   (e) What % of values below 4? (Draw picture.)
   (f) What % of values between 4 and 13? (Use previous pictures.)

5. Cereal carbos have approximately a normal distribution. Using StatCrunch, carry out the following steps:
   (a) Select Data and Compute Expression. Into the Expression box, type

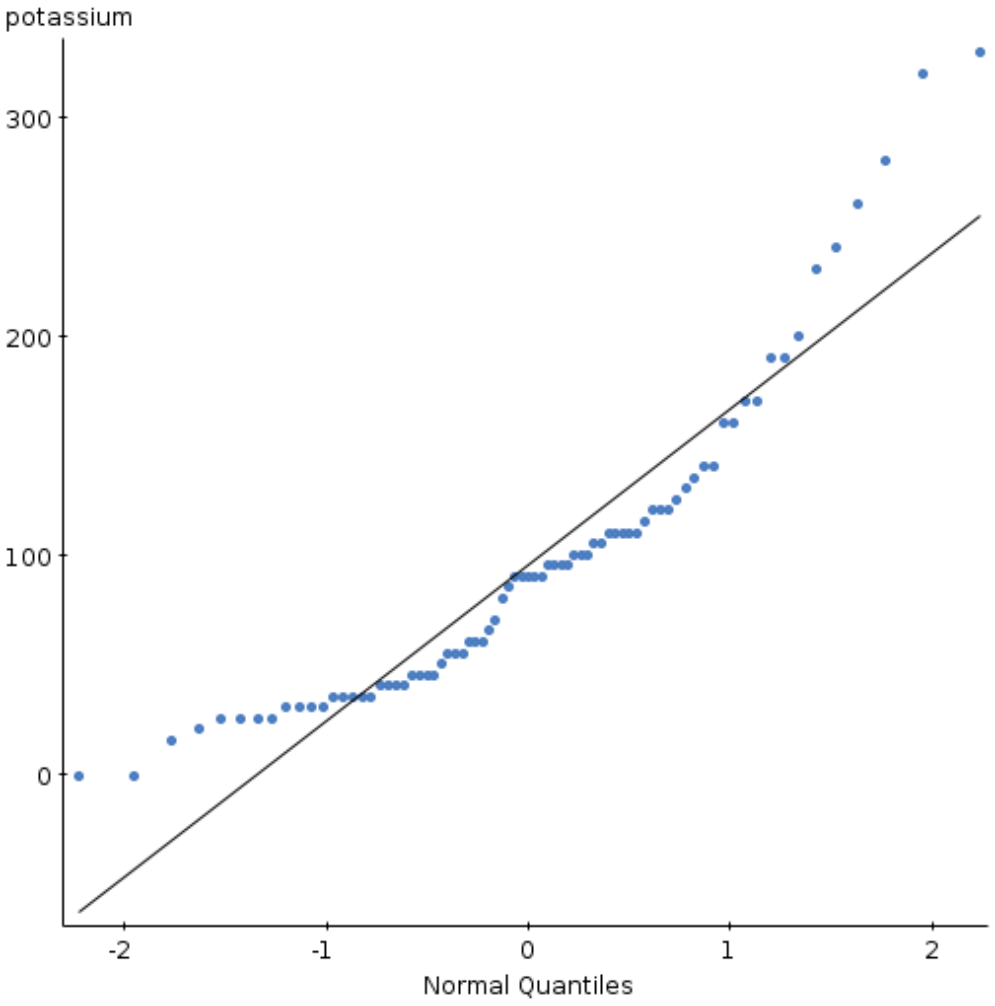
   ```
   (carbo > mean(carbo)-std(carbo))
   and (carbo < mean(carbo)+std(carbo))
   ```

   (or use Add Column and Add Function from the Expression Builder to save some typing). This figures out which `carbo` values are greater than $\mu - \sigma$ *and* also less than $\mu + \sigma$. At the bottom, give the new column the name "between". Click Compute. The new column will be `true` for those data values within 1 SD of the mean, and `false` otherwise.
   (b) Draw a pie chart of `between`. What % `carbo` values are within 1 SD of the mean? What % should be? Are these answers close?

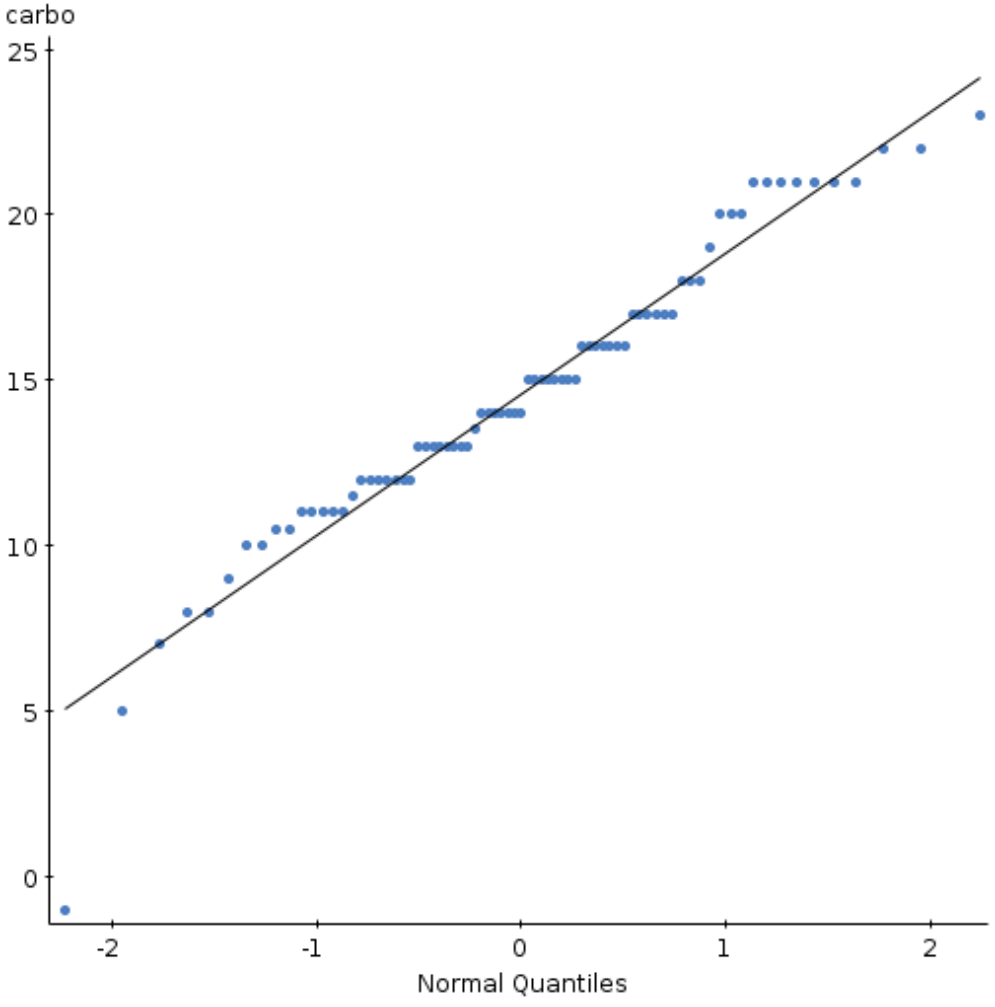
6. Using the ideas from the previous question, and again using StatCrunch:

   (a) Find out what % of the `potassium` values are within 1 SD of the mean.
   (b) Is this close to 68%?
   (c) The potassium values are skewed to the right. What does this say about using the 68–95–99.7 rule for non-normal distributions?

7. The variable `potassium` does not have a normal distribution. How can you tell this from the normal quantile plot below? How can you deduce that the shape of the distribution is skewed to the right?



8. We have been assuming that the variable `carbo` has a normal distribution. It has one value, -1, which is an error. Apart from that, does the plot below suggest that a normal distribution describes the data?



9. Use StatCrunch to draw a normal quantile plot for `calories`. In what way(s) does this variable fail to have a normal distribution (I can think of 2)? (Hint: in StatCrunch, the plot you want is called a QQ Plot.)

10. Proportion of standard normal curve *between 0 and z* approx

$$0.1z(4.4 - z)$$

for $0 \leq z \leq 2.2$ (0.49 for $2.2 \leq z \leq 2.6$, 0.50 for $z > 2.6$):

(a) Find the proportion of the normal distribution less than $z = 0.67$ using the formula, and using Table A.

(b) Repeat the last part for $z = 1.28$.

(c) How accurate is the approximation in each case? In the source, it is claimed to be accurate "to 2 decimal places".

(d) How would you use the formula to find the proportion of the normal distribution less than $z = -1.65$?

Lecture 6 summary

- How do you find a value from a proportion?
  - *Look up proportion in body of table to get z*
  - *Unstandardize to get value: $x = \mu + \sigma z$*
- What is the 68–95–99.7 rule and how does it work?
  - *On any normal distribution:*
    - ★ *68% between $\mu - \sigma$ and $\mu + \sigma$*
    - ★ *95% between $\mu - 2\sigma$ and $\mu + 2\sigma$*
    - ★ *99.7% between $\mu - 3\sigma$ and $\mu + 3\sigma$*
  - *Can get proportion from value(s) or values from proportion*
- How do you tell whether a normal distribution is reasonable for describing data?
  - *Use normal quantile (QQ) plot*
  - *If pattern straight, normal OK*
  - *If pattern curved, distribution skewed and normal not OK*
- How can you approximate normal proportions without Table A?
  - *Proportion between 0 and z approx. $0.1z(4.4 - z)$ for $0 \leq z \leq 2.2$ (about 2 decimal accuracy)*

To read for next time

§2.1, 2.2:

- Examining relationships
- Scatterplots
- Interpreting scatterplots
- Categorical explanatory variables
- The correlation r
- Properties of correlation

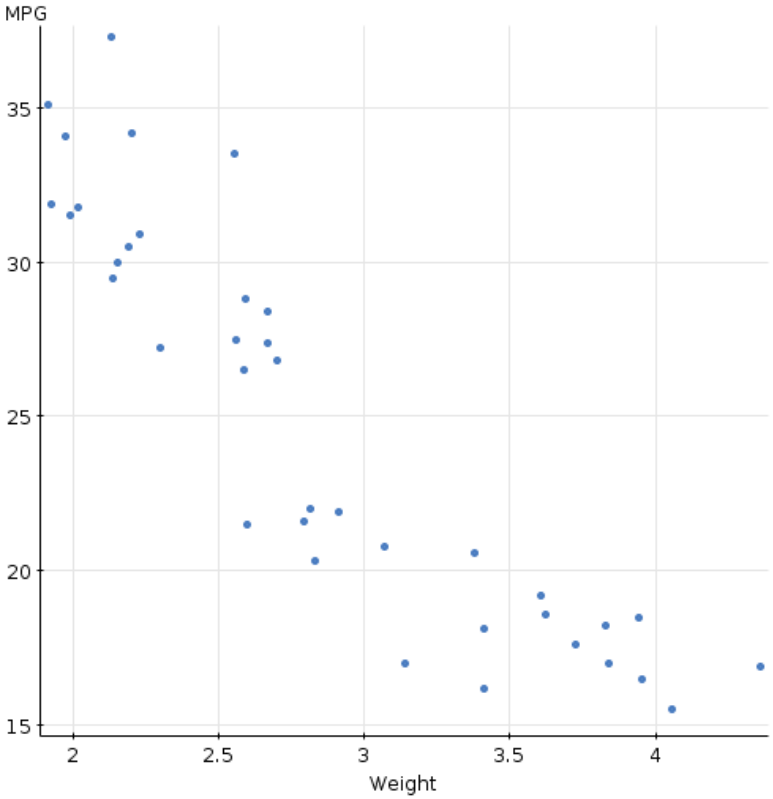
Lecture 7: §2.1, 2.2

Coming up:

- How do we decide whether there is any relationship between two variables?
- Which variable goes on which axis of our plot?
- How can we see the effect of a categorical variable on a scatterplot?
- How do we interpret a scatterplot?
- If the relationship is linear, how do we describe its strength?
- Can we use the correlation for any kind of relationship?
- What do different kinds of correlation look like?
- Should we be worried about outliers?

Questions

1. The picture below is a scatterplot of gas mileage vs. weight for the cars data.



(a) Based on what you know about cars, do you think a heavier car would usually have a better or worse gas mileage than a lighter one?

(b) Does the scatterplot support your answer to the previous part?

(c) Is this what statisticians would call a positive association or a negative one? Why?

2. Use StatCrunch to make a scatterplot of fiber (on the y axis) vs. potassium (on the x axis). Is there no association, a positive one or a negative one?

3. Edit your fiber-potassium scatterplot to show, in addition, the categorical variable `shelf`. How is the shelf on which each cereal is found related to fiber and potassium?

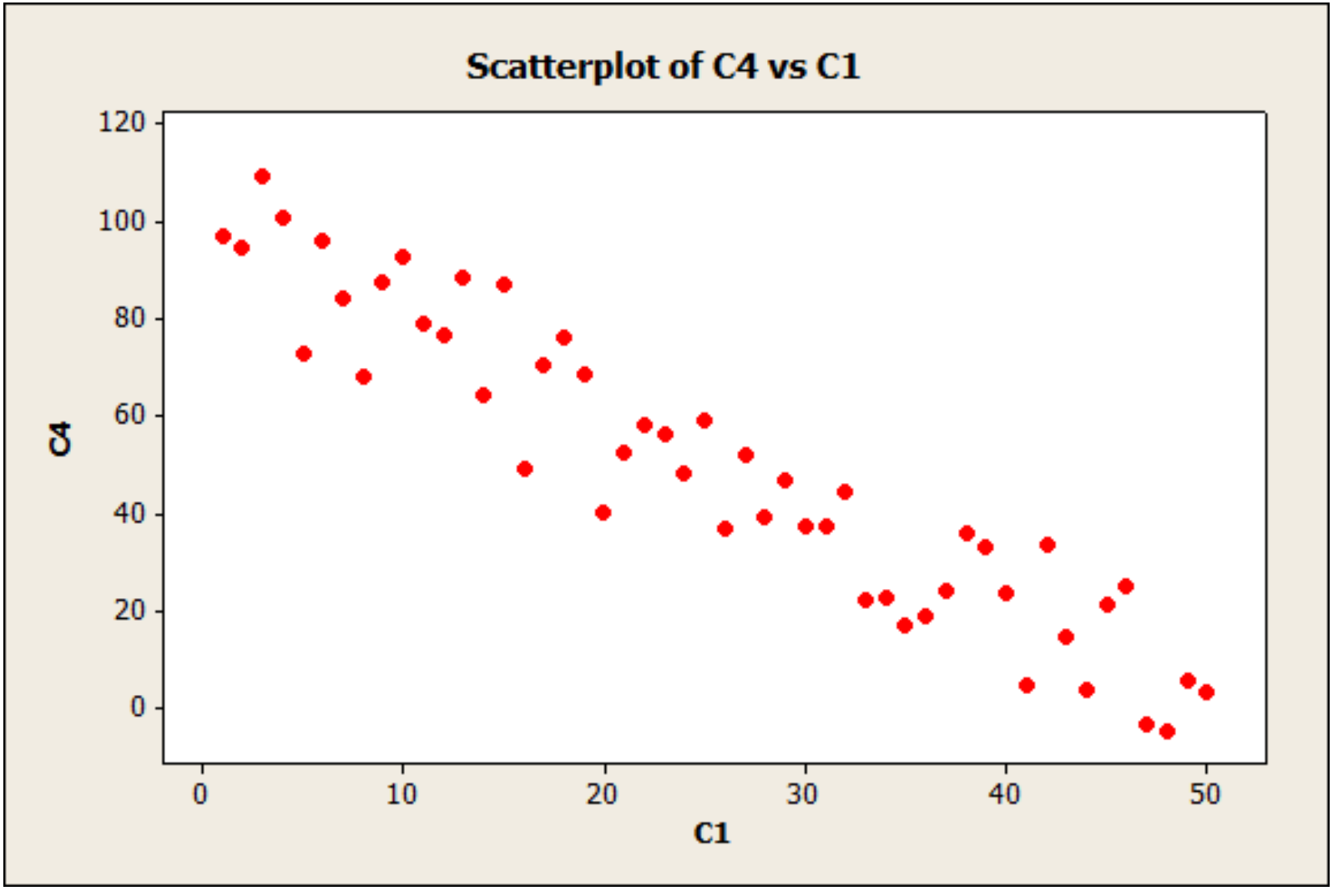
4. Use StatCrunch to make a scatterplot of calories (on the y axis) vs. potassium (on the x axis). What association do you see? (You can investigate this by clicking Options (on the graph) and Edit, going back to the dialog with Display at the top, and choosing something for Polynomial Order. 1 fits a straight line, and 2 fits a curve.)

5. Use StatCrunch to make a scatterplot of sodium vs. sugars. What do you see?

6. Three important aspects of an association are its *form*, *direction* and *strength*. What kinds of form, direction and strength do we need to look out for? (Exercise for you: assess form, direction and strength on the scatterplots we've just seen.)

7. How would you answer the two questions below?
   (a) You have data on students' midterm marks and final exam marks in a course. Which should be response and which explanatory, or neither?
   (b) You have data on students' high school English and high school math grades. Which should be response and which explanatory, or neither?

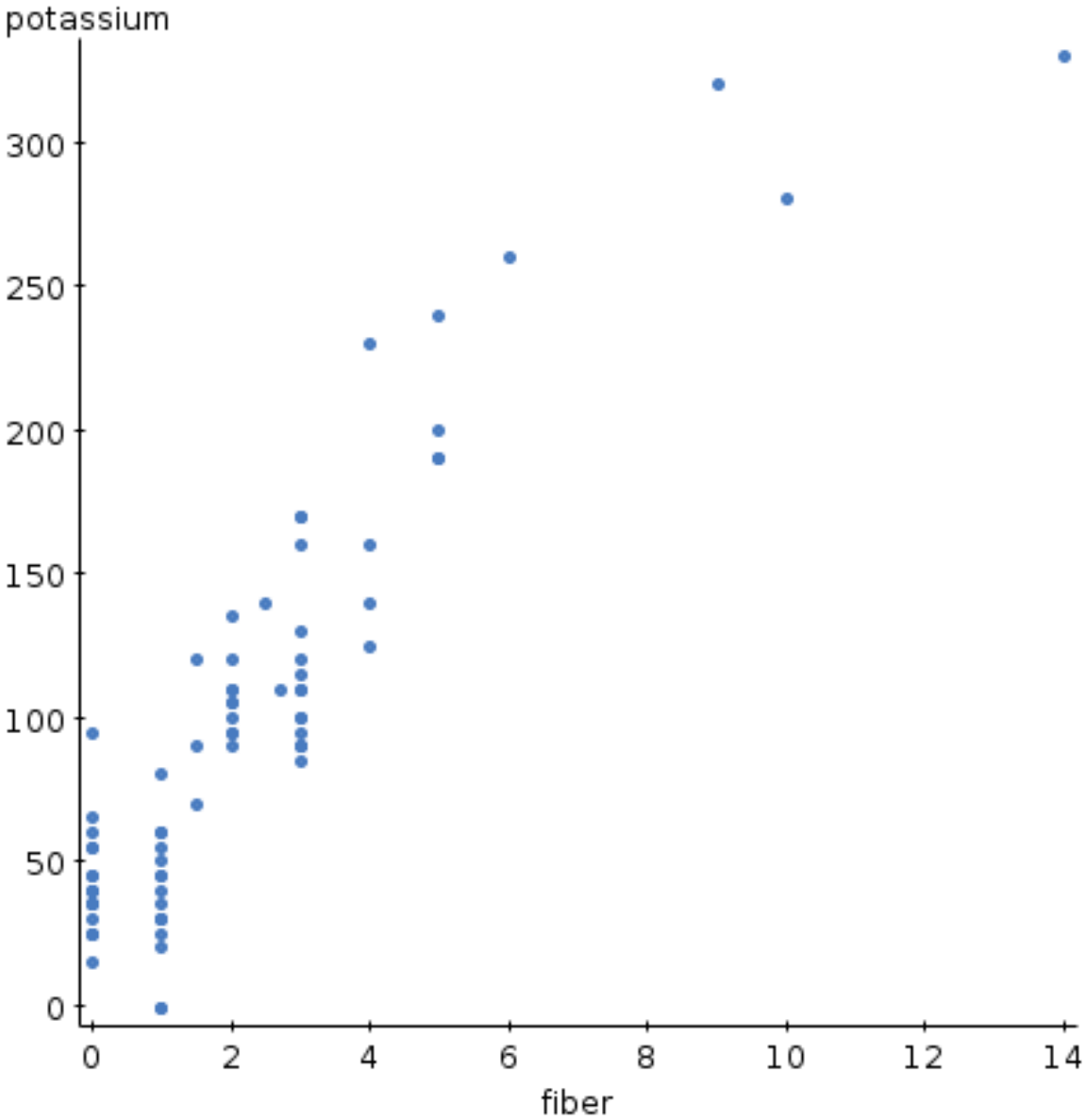
8. Look at the scatterplot below. Does it show a straight-line association?



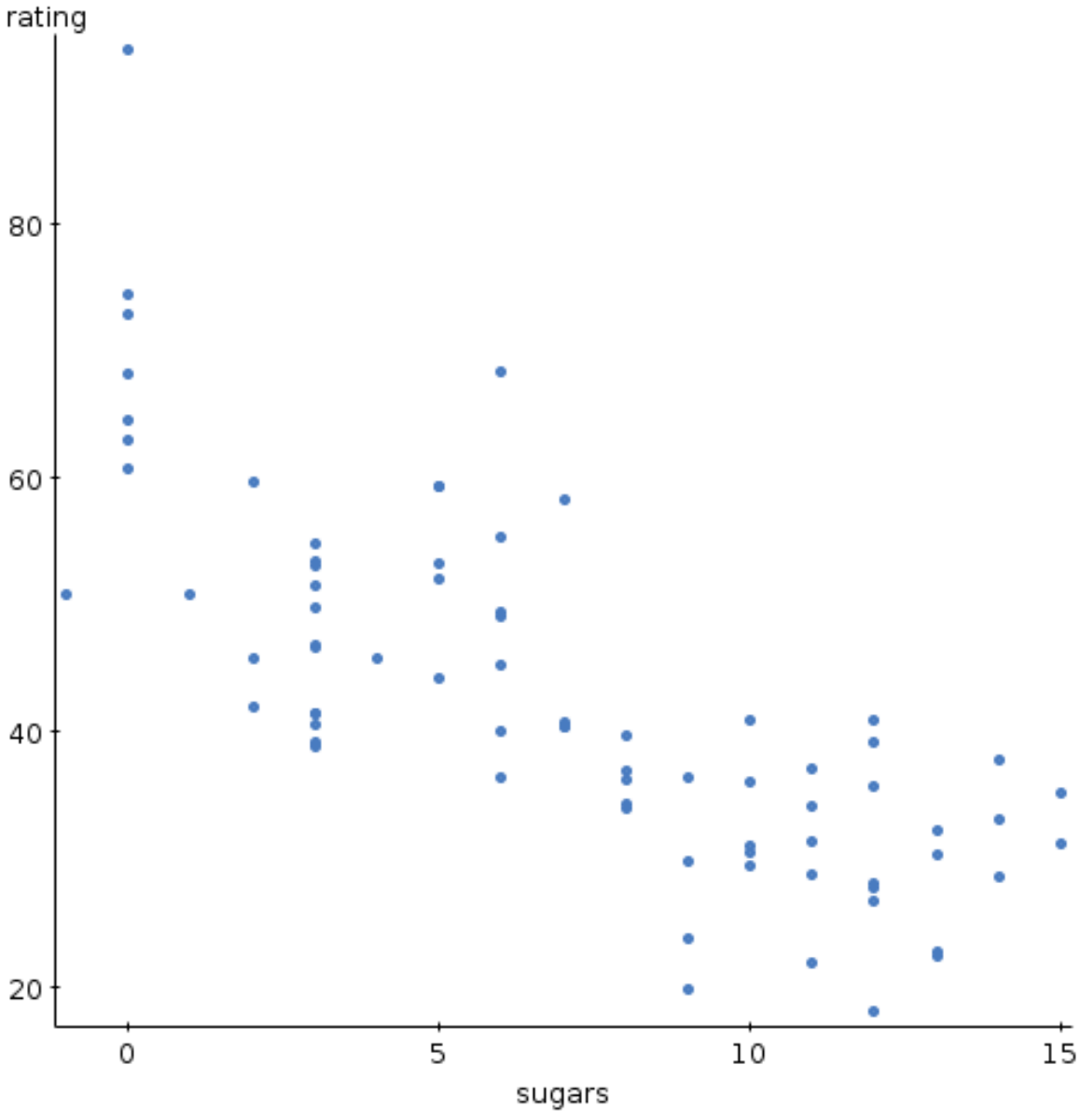
9. About the correlation:
   (a) What correlation goes with a scatterplot that shows no association?
   (b) What correlation goes with a scatterplot that shows a perfect stright-line association?
   (c) What correlation goes with an association that is a perfect curve?
   (d) What correlation goes witn a positive association?
   (e) What correlation goes with a negative association?

10. The correlation between variables x and y is very high, 0.95.
    (a) x has mean 10. Does the high correlation tell you that y must have mean 10 as well?
    (b) Does the high correlation tell you that y is predictable if you know x?

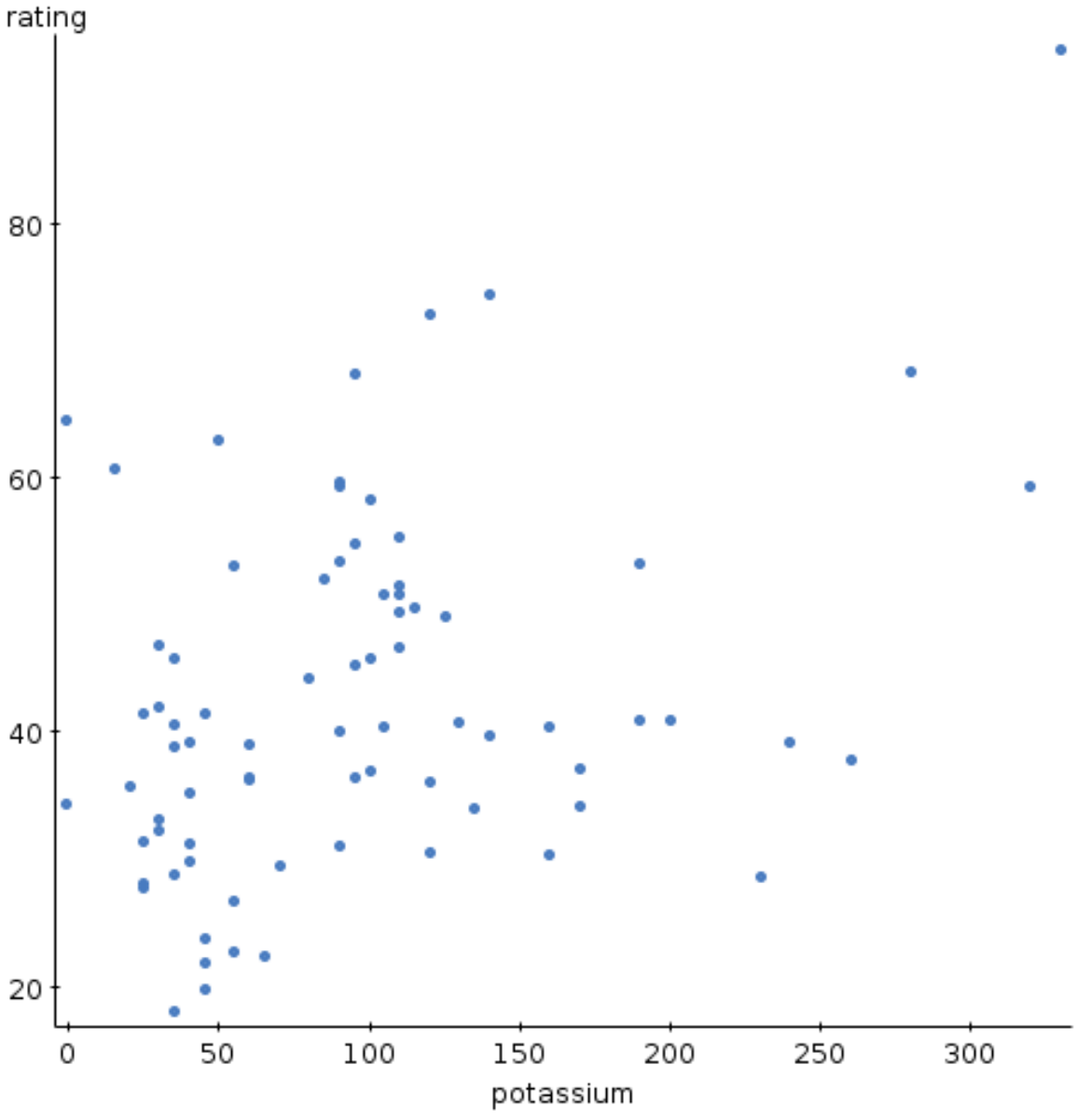
11 Guess the correlation:



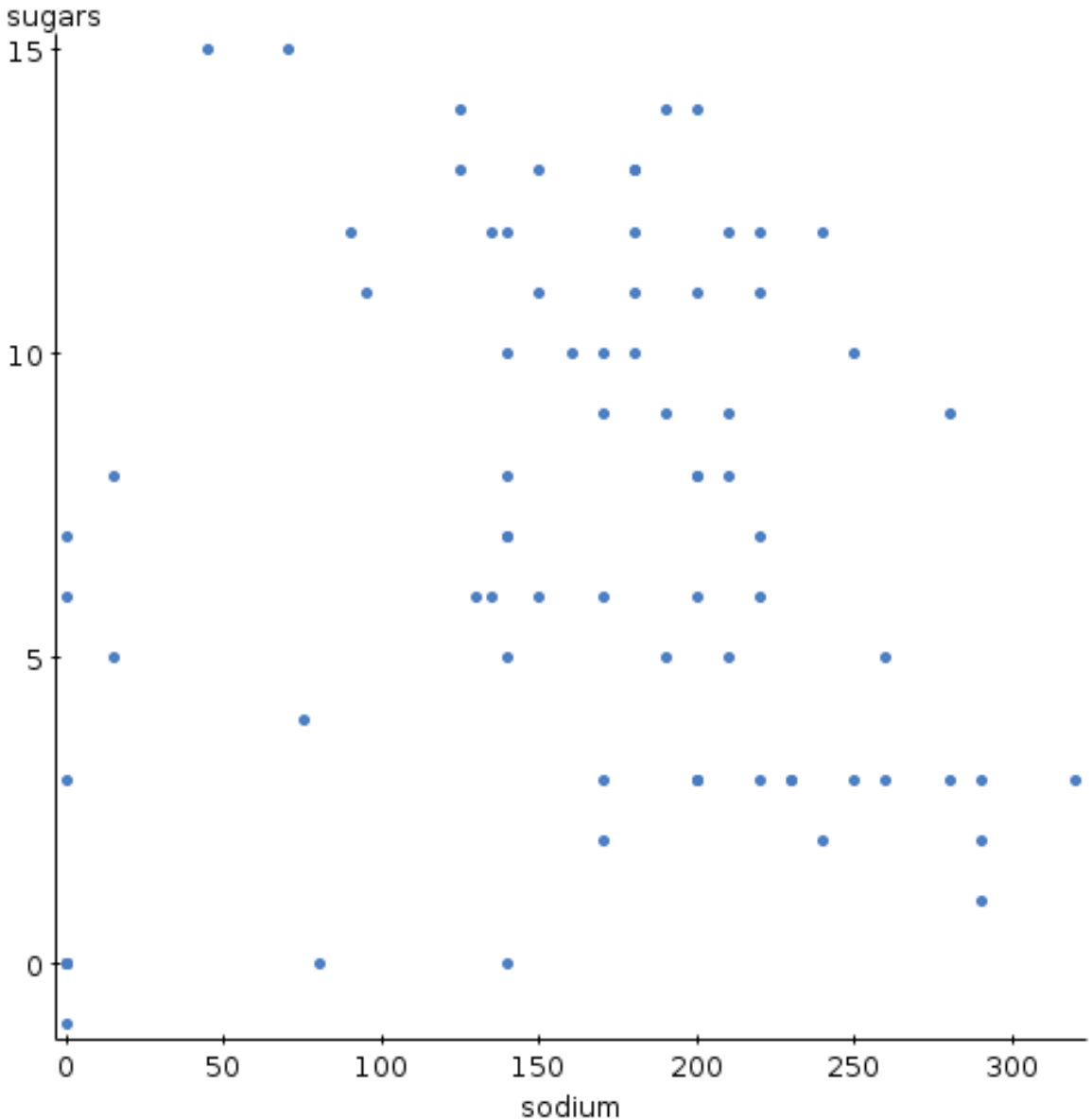
12 Guess the correlation:



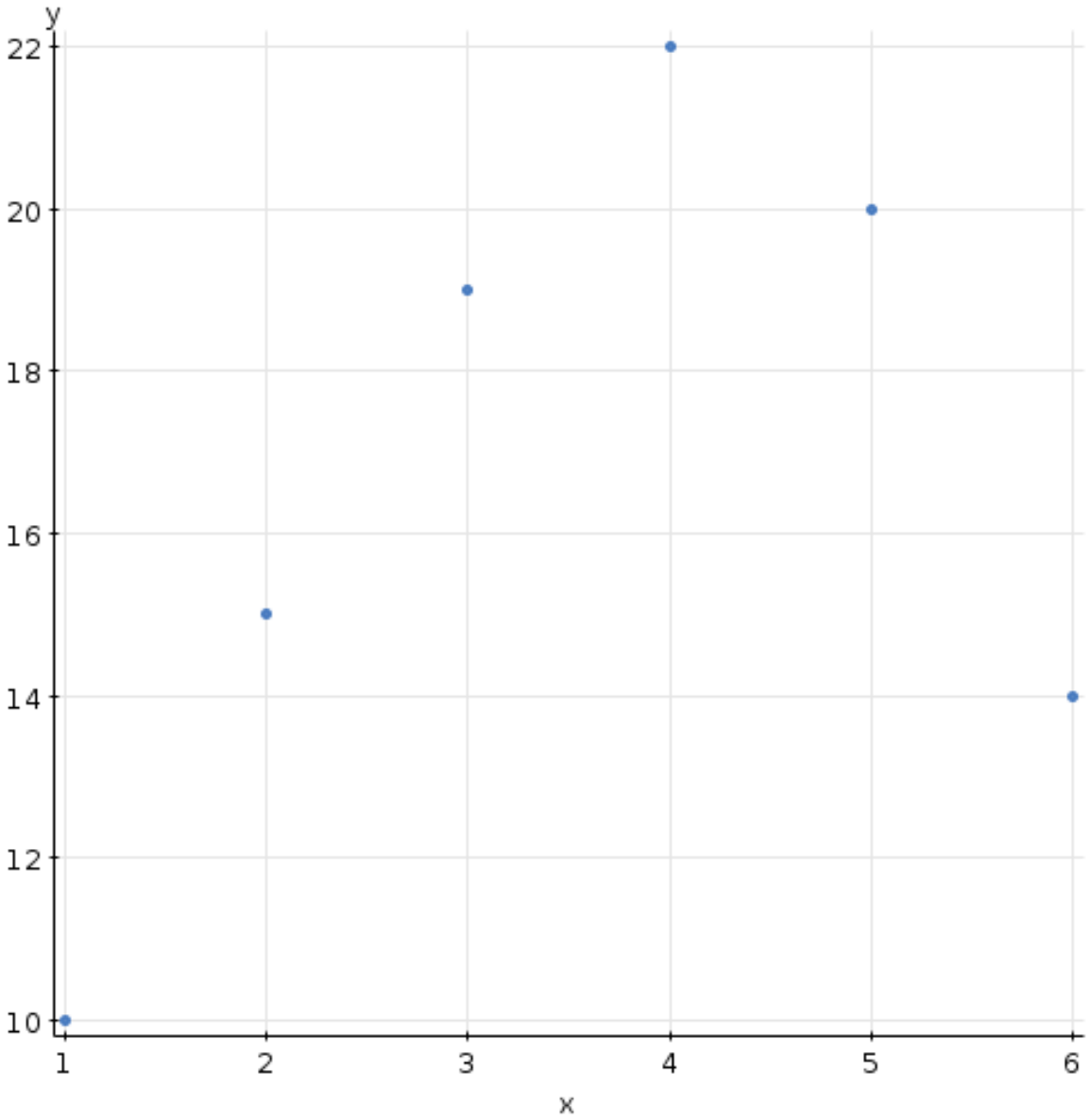
13 Guess the correlation:



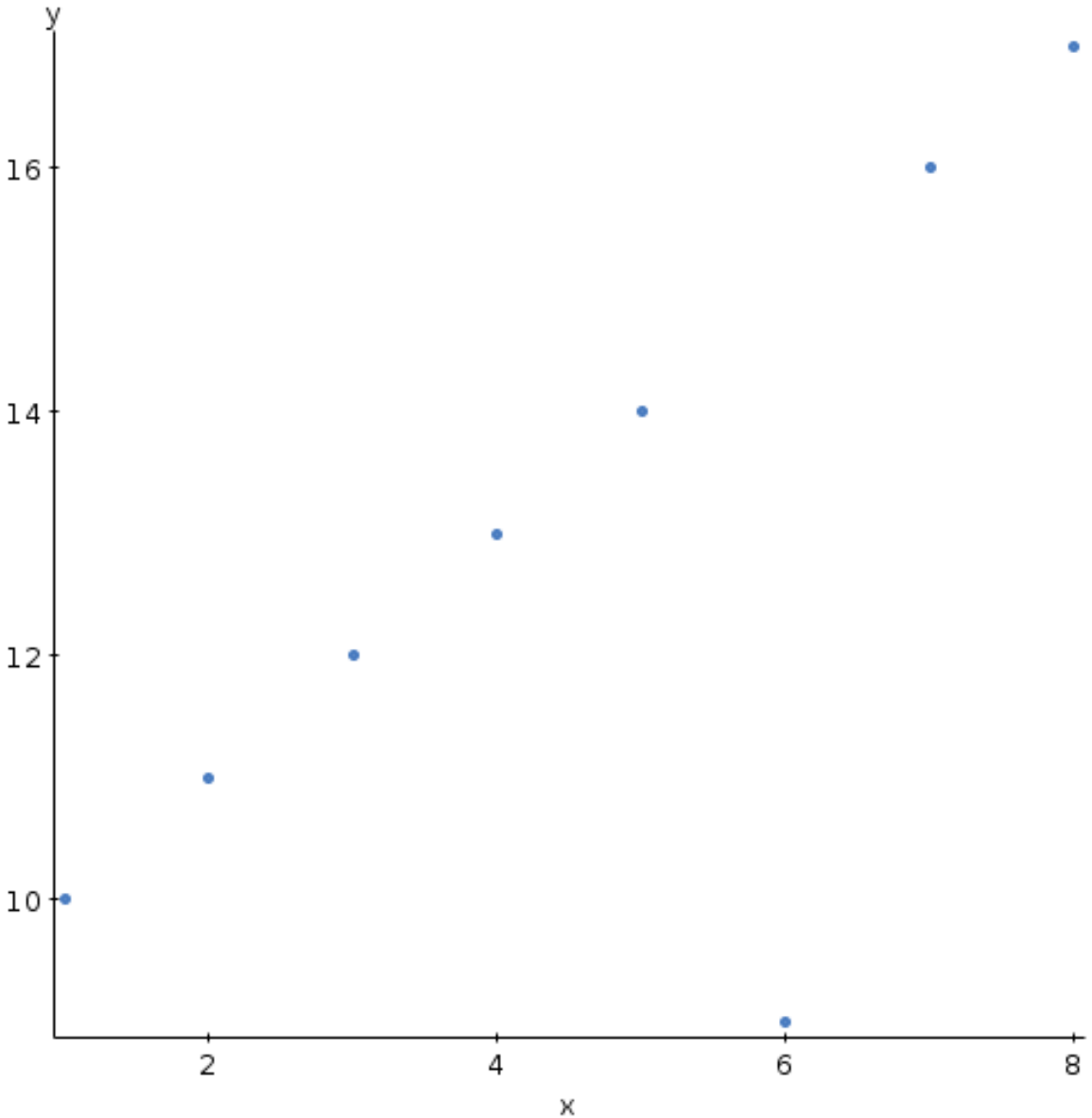
14 Guess the correlation:



15 Guess the correlation:



16 Guess the correlation:



Lecture 7 summary

- How do we decide whether there is any relationship between two variables?
  - *Look at a scatterplot*
- Which variable goes on which axis of our plot?
  - *Response, if there is one, on vertical y axis*
- How can we see the effect of a categorical variable on a scatterplot?
  - *Use different colours or symbols to identify the data from each category*
- How do we interpret a scatterplot?
  - *Is there anything there: does changing x have any effect on y?*
  - *Form: Is the relationship straight or curved?*
  - *Direction: Is the relationship upward or downward?*
  - *Strength: is it clear-cut or is there a lot of variability?*

. . . Part 2

- If the relationship is linear, how do we describe its strength?
  - *Use the correlation coefficient*
- Can we use the correlation for any kind of relationship?
  - *No, only for straight-line ones*
  - *Misleading if the relationship actually curved*
- What do different kinds of correlation look like?
  - *Look back at the pictures*
- Should we be worried about outliers?
  - Definitely: can seriously affect correlation.

To read for next time

§2.3:

- Least squares regression
- Fitting a line to data
- Prediction
- Least-squares regression
- Interpreting the least-squares regression line
- Facts about least-squares regression
- Regression and correlation
- Understanding R-squared

Lecture 8: §2.3

Coming up:

- What does the regression line do?
- What is the regression line for?
- What does the least-squares idea give us?
- How do you interpret the slope of a regression line?
- How do you interpret the intercept of a regression line?
- How do you calculate slope and intercept of regression line from data?
- What is extrapolation? Is it a good idea?
- How is the correlation related to regression?

Questions

1. This chapter is all about regression, correlation and scatterplots. What is each technique for, and which order should you use them in?

2. Consider the straight line $y = 2 + 3x$:
   (a) For each of $x = 0, 1, 2, 3$, work out what y is according to the straight line.
   (b) The intercept of this line is 2. Where do you see 2 in the calculations you just did?
   (c) The slope of this line is 3. Where do you see 3 in the calculations you just did?

3. Here is a small data set:

x	0	1	2
y	3	5	6

   (a) Sketch a scatter plot of the data. What kind of association do you see?
   (b) One candidate regression line for these data is $y = 6 - x$. Work out the three predicted values of y (for $x = 0, 1, 2$), then work out the three residuals, then square them and add them up. What do you get?
   (c) Repeat the previous part for the line $y = 4 + x$.
   (d) Which of the lines in (b) and (c) fits better? How can you tell?
   (e) How might you have guessed the answer to (d) without doing any calculations?
   (f) What would be a way of finding the overall best-fitting line for a particular data set?

4. For the data set of the previous problem, we will use StatCrunch to find the least-squares regression line for predicting y from x. The data were:

x	0	1	2
y	3	5	6

   (a) Start StatCrunch with an empty worksheet, and type the data into two columns, one for x and one for y. Name the columns suitably.
   (b) Select Stat, Regression, Simple Linear and fill in your explanatory and response variables.
   (c) What are the intercept and slope of your regression line?
   (d) The line found by StatCrunch has a sum of squared residuals of 0.17. Does that indicate a better fit than the two values you found in the previous question? How can you tell?
   (e) The line $y = 4 + x$ passes exactly through two of the data points, but the so-called best-fitting line doesn't pass through any of them. How do you explain this?

5. For the data set of the previous problems, x has mean 1 and SD 1, and y has mean 4.667 and SD 1.527. The correlation between x and y is 0.982. Use the appropriate formulas to calculate the slope and intercept of the regression line for predicting y from x. Are your answers close to those found by StatCrunch? (They may not be exactly the same because the values here have been rounded off.)

6. Why does y vary?
   (a) Enter the following data into StatCrunch:

x	1	2	3
y	4	10	15

   (b) Make a scatterplot of y vs. x. The values of y vary, but do they do so (i) because y depends on x, or (ii) just randomly?
   (c) Now enter these data into StatCrunch (keeping the data of (a) around):

x	1	2	3	4
y	9	11	4	8

   (d) Likewise, for the data of (c), draw a scatterplot of y vs x. Does y vary (i) because y depends on x, or (ii) just randomly?
   (e) Use StatCrunch to find R-squared for each dataset. Which one is higher? What does that tell you about the fit of a line to each data set?

7. Open up the car data in StatCrunch, and make a scatterplot of MPG against weight, adding the regression line to the plot.
   (a) Do you think R-squared is large or small?
   (b) The regression line is $MPG = 48.71 - 8.365$ weight. What is the predicted gas mileage for a car that weighs 6 tons? Is the answer sensible?
   (c) Do you think it is a good idea to trust the regression line for this kind of weight? Why, or why not?

8. For the car data:
   (a) Create a new variable "gallons per mile" (`gpm`) as $1/MPG$.
   (b) Make a scatter plot of `gpm` against `weight`. Does the association look like a straight line?
   (c) If sensible, use StatCrunch to fit the regression line and obtain a predicted `gpm` for a car weighing 6 tons. Is the result sensible (or at least not obviously nonsense)? What is that prediction in miles per gallon?

Lecture 8 summary

- What does the regression line do?
  - *Gives the "best" straight line describing relationship between response y and explanatory x*
- What is the regression line for?
  - *Predicting what y should be when x known*
- What does the least-squares idea give us?
  - *Definition of "best" for line*
  - *A way of finding intercept and slope of that line*
- How do you interpret the slope of a regression line?
  - *Average increase in y when x increases by 1*
- How do you interpret the intercept of a regression line?
  - *Average value of y when $x = 0$*

. . . Continued

- How do you calculate slope and intercept of regression line from data?
  - Start with data means $\bar{x}, \bar{y}$, SDs s_x, s_y and correlation r.
  - Slope $b_1 = rs_y/s_x$.
  - Intercept $b_0 = \bar{y} - b_1\bar{x}$.
- What is extrapolation? Is it a good idea?
  - *Predicting y from an x above or below your data*
  - *Don't know whether straight line still applies, so bad idea*
- How is the correlation related to regression?
  - *Correlation-squared is fraction of variability in y explained by regression of y on x.*
  - *Any remaining variability in y "random", unexplained.*

To read for next time

§2.4

- Cautions about correlation and regression
- Residuals
- Outliers and influential observations
- Beware the lurking variable
- Beware correlations based on averaged data
- The restricted-range problem

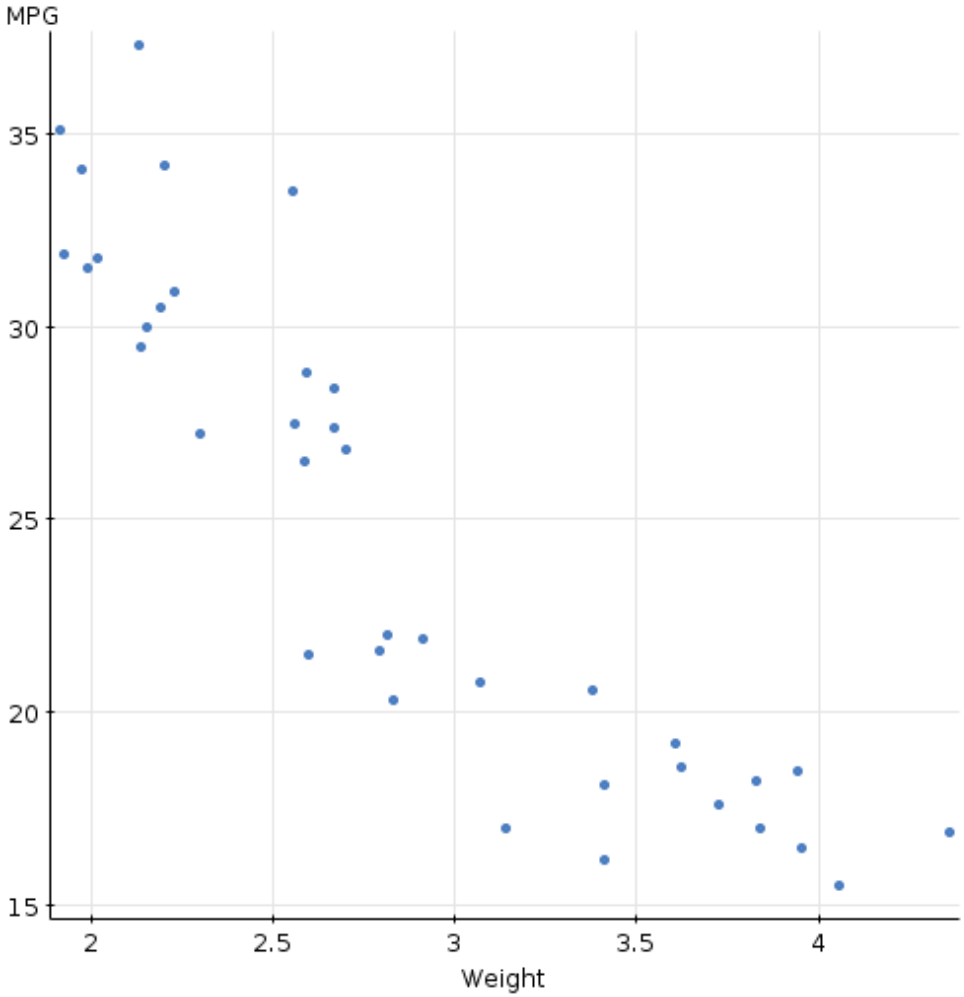
Lecture 9: §2.4

Coming up:

- What do the residuals tell us?
- How can we tell if a regression line is not fitting well?
- What are outliers and how do we detect them?
- What are influential observations and how do we detect them?
- What is a lurking variable and how might it affect our regression?
- If there is a high correlation, can we conclude that one of the variables is the cause of the other?
- What can happen if we base a correlation on summarized (averaged) data?
- What can happen if the data don't cover the full range of x and y?

Questions

1. The scatterplot for MPG vs. weight for the car data is shown below.

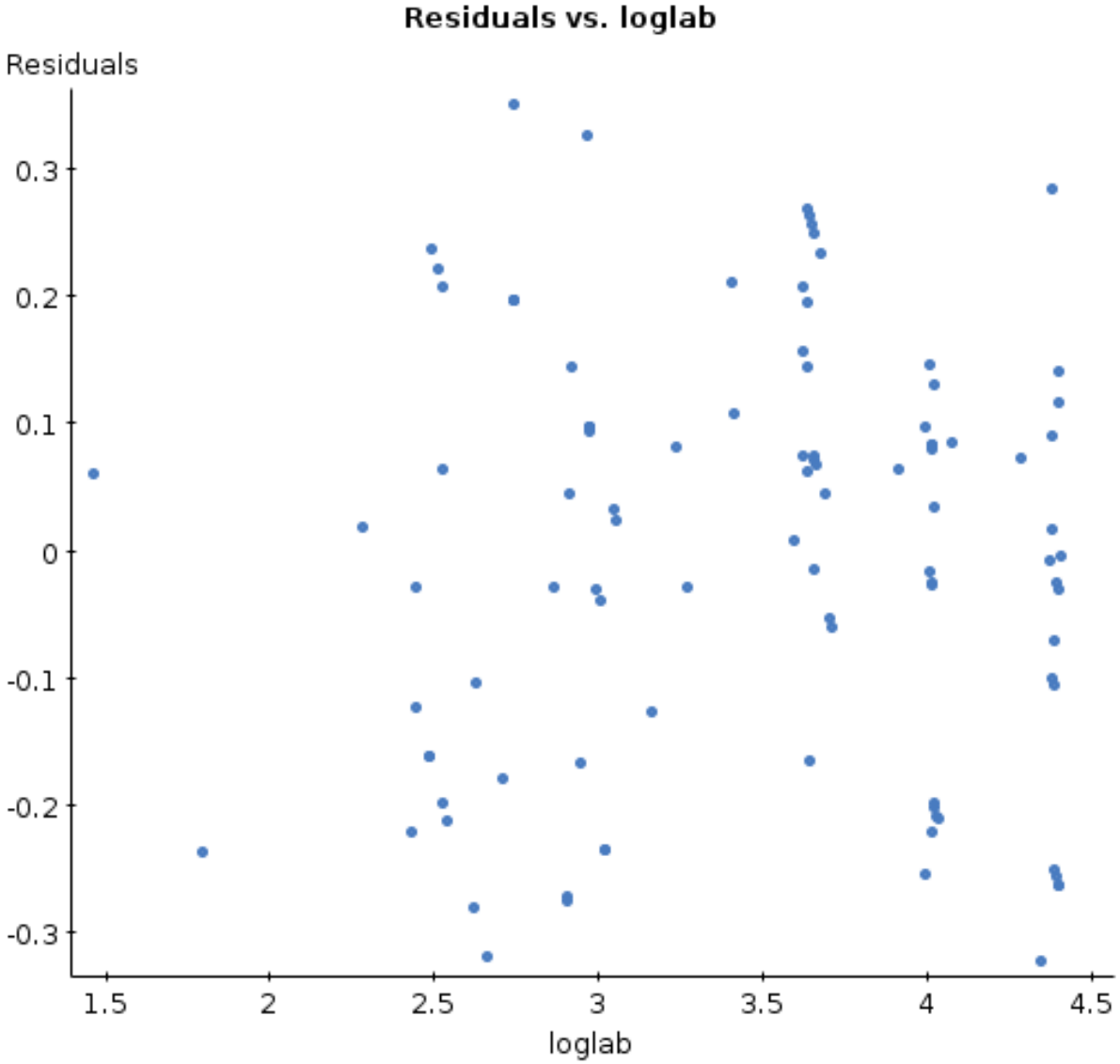


(a) There is a suggestion that the relationship (to the left) is curved (decreases more slowly for higher weights). Use StatCrunch to run the regression of MPG from weight and obtain a plot of the residuals (vs. weight). Is there a clearer suggestion of a curve?

(b) What do we do about curved relationships in this course?

2. The data set `labfield` contains lab and field measurements of depth of defects in an oil pipeline. Do the field measurements match the lab measurements reasonably well?
   (a) Open the dataset in StatCrunch. Make a scatterplot of `field` against `lab` and confirm that the association is linear (not obviously curved).
   (b) Run a regression and obtain a plot of residuals against `lab`. Do you see a problem? What is it? Can you explain it in a few words?
   (c) What might we do about the problem?

3. For the data of the last problem, I took logarithms of the lab and field measurements, and ran the regression again. My residual plot is below. Do you see any problems with it?



4. The `bloodsugar` data set contains two measures of blood sugar for 18 diabetics: FPG is measured by the diabetic at home, and HbA is measured by a doctor.
   (a) Read the data into StatCrunch and make a scatterplot (with FPG as the response variable). Plot the regression line on your scatterplot. Comment on the association. Are there any observations that don't seem to fit?
   (b) Obtain the regression line for predicting FPG from HbA. Note down the intercept, slope and R-squared.
   (c) Is there one observation that appears to be a long way off the line? Create a third variable in your dataset, called `omit`, and set it equal to 1 for this observation. Re-run the regression, omitting this point. What has happened to the intercept, slope and R-squared?
   (d) There is one observation a long way to the right. Re-run the regression omitting this point (put the other one back in). What has happened to intercept, slope and R-squared?
   (e) What can you say about the effect of observations with large residuals or unusual x-values?

5. The dataset `mathstudents` contains the records, for a university over several years, of the number of first years, and the total number of students taking elementary math courses.
   (a) Use StatCrunch to make a scatterplot of the number of math students (response) vs. number of 1st years. Would you be reasonably happy to fit a straight line here?
   (b) Do the regression and save the residuals.
   (c) Plot the residuals against year. Do you see any pattern? What appears to have happened?

6. The data set `corrwithmean` contains measurements for two variables x and y.
   (a) Open the data in StatCrunch, and make a scatterplot of y vs. x. What kind of association do you see?
   (b) What do you think the correlation between x and y is? Use StatCrunch to find the correlation.
   (c) Now compute the mean y for each value of x. This can be done by using Summary Statistics to find the mean value of y, grouped by the values of x. Save the results in the data table.
   (d) Now find the correlation between x and the mean values of y. Compare it with your answer from (b).
   (e) What does the correlation in (d) ignore? What, therefore, would you expect to happen to the correlation when it is based on averaged data rather then the original data?

7. Open the cars data set in StatCrunch:
   (a) Find the correlation between MPG and weight.
   (b) Now find the correlation between MPG and weight only for those cars that weigh over 3 tons by typing `Weight>3` into the Where box. How does it compare to your answer from (a)?
   (c) Now find the correlation between MPG and weight only for those cars that have MPG greater than 30, using the Where box. How does this answer compare to (a)?
   (d) What appears to happen when you calculate a correlation restricting the range of x or y?

Lecture 9 summary

- What do the residuals tell us?
  - *Whether an observation is close to or far from the regression line*
- How can we tell if a regression line is not fitting well?
  - *Look at plot of residuals (eg. against x values)*
- What are outliers and how do we detect them?
  - *Observations far from the regression line*
  - *Large (+ or −) residuals*
- What are influential observations and how do we detect them?
  - *Observations with great influence over where the line goes*
  - *Unusually high or low values of x (outliers in terms of x)*
- What is a lurking variable and how might it affect our regression?
  - *Variable we didn't include in regression but which has effect on relationships*
- If there is a high correlation, can we conclude that one of the variables is the cause of the other?
  - **No!**
  - *Apparent relationship might be caused by lurking variables*

. . . Continued

- What can happen if we base a correlation on summarized (averaged) data?
  - *Usually higher than if we had calculated from individuals*
  - *because individuals vary amongst themselves and that variability is lost*
- What can happen if the data don't cover the full range of x and y?
  - *Correlation can be lower than if we had observed the full range*
  - *because we could have observed higher and lower x and y that would have made the pattern clearer*

To read for next time

§2.5:

- Data analysis for two-way tables
- The two-way table
- Joint distribution
- Marginal distributions
- Describing relations in two-way tables
- Conditional distributions
- Simpson's paradox

Lecture 10: §2.5

Coming up:

- How can we summarize two categorical variables?
- What is the joint distribution?
- What is the marginal distribution? Does it say anything about the relationship between the two variables?
- What is the conditional distribution?
- How do we assess relationship between the two variables?
- What is Simpson's paradox? How does it happen? What can cause it?

Questions

1. I used StatCrunch to classify the cars by weight as "light" (under 2.5 tons), "medium" (2.5–3.5 tons), "heavy" (over 3.5 tons). This is called "binning" the variable `weight`. A contingency table of the cars classified by country of origin and binned weight is shown below.

Contingency Table: country by binned weight

Copy Print Mail

Contingency table results:
Rows: Country
Columns: Bin(Weight)

Cell format
Count (Total percent)

	Below 2.5	2.5 to 3.5	3.5 or above	Total
France	0 (0%)	1 (2.632%)	0 (0%)	1 (2.632%)
Germany	3 (7.895%)	2 (5.263%)	0 (0%)	5 (13.16%)
Italy	1 (2.632%)	0 (0%)	0 (0%)	1 (2.632%)
Japan	5 (13.16%)	2 (5.263%)	0 (0%)	7 (18.42%)
Sweden	0 (0%)	2 (5.263%)	0 (0%)	2 (5.263%)
U.S.	3 (7.895%)	10 (26.32%)	9 (23.68%)	22 (57.89%)
Total	12 (31.58%)	17 (44.74%)	9 (23.68%)	38 (100.00%)

(a) Why did I not use the original weight for the contingency table?

(b) How many cars are from Germany and weigh less than 2.5 tons?

(c) What percentage of all the cars are from the US and weigh more than 3.5 tons?

(d) How many cars weigh between 2.5 and 3.5 tons altogether?

(e) How many cars are from Japan altogether?

(f) Do you think there is an association between country and weight? How can you tell?

2. Using the contingency table for the previous question (reproduced below):

Contingency Table: country by binned weight

Copy Print Mail

Contingency table results:
Rows: Country
Columns: Bin(Weight)

Cell format
Count (Total percent)

	Below 2.5	2.5 to 3.5	3.5 or above	Total
France	0 (0%)	1 (2.632%)	0 (0%)	1 (2.632%)
Germany	3 (7.895%)	2 (5.263%)	0 (0%)	5 (13.16%)
Italy	1 (2.632%)	0 (0%)	0 (0%)	1 (2.632%)
Japan	5 (13.16%)	2 (5.263%)	0 (0%)	7 (18.42%)
Sweden	0 (0%)	2 (5.263%)	0 (0%)	2 (5.263%)
U.S.	3 (7.895%)	10 (26.32%)	9 (23.68%)	22 (57.89%)
Total	12 (31.58%)	17 (44.74%)	9 (23.68%)	38 (100.00%)

(a) Write the marginal distribution for weight.

(b) Write down the marginal distribution for country.

3. Using the contingency table for the previous question (below), find the conditional distribution of:

Contingency Table: country by binned weight

Copy Print Mail

Contingency table results:
Rows: Country
Columns: Bin(Weight)

Cell format
Count (Total percent)

	Below 2.5	2.5 to 3.5	3.5 or above	Total
France	0 (0%)	1 (2.632%)	0 (0%)	1 (2.632%)
Germany	3 (7.895%)	2 (5.263%)	0 (0%)	5 (13.16%)
Italy	1 (2.632%)	0 (0%)	0 (0%)	1 (2.632%)
Japan	5 (13.16%)	2 (5.263%)	0 (0%)	7 (18.42%)
Sweden	0 (0%)	2 (5.263%)	0 (0%)	2 (5.263%)
U.S.	3 (7.895%)	10 (26.32%)	9 (23.68%)	22 (57.89%)
Total	12 (31.58%)	17 (44.74%)	9 (23.68%)	38 (100.00%)

(a) weight for cars from Japan

(b) country for cars that weigh over 3.5 tons

(c) What does your answer to (b) tell you, in words?

(d) (optional extra) Use StatCrunch to check your calculation of the conditional distributions.

4. A university records applications to its professional schools:

	Accept	Reject	Total
Male	490	210	700
Female	280	220	500

(a) What percent of males and females are admitted overall?

(b) When you separate things out by school, the frequencies are these:

	Business				Law		
	Accept	Reject	Total		Accept	Reject	Total
Male	480	120	600	Male	10	90	100
Female	180	20	200	Female	100	200	300
Total	660	140	800	Total	110	290	400

What percentages of males and females are admitted to each school?

(c) How are your answers to (a) and (b) contradictory? This is Simpson's Paradox.

(d) What percentage of applications to each school are accepted?

(e) What percentage of applications to each school are female?

(f) Do your answers to (d) and (e) explain the paradox? How?

Lecture 10: summary

- How can we summarize two categorical variables?
  - *Two-way table with each cell counting number of individuals in that category combination*
- What is the joint distribution?
  - *Proportions (%'s) of individuals in each category combination*
- What is the marginal distribution? Does it say anything about the relationship between the two variables?
  - *Proportions in each category of one variable ignoring other*
  - *Says nothing about any relationship*
- What is the conditional distribution?
  - *For a particular category of one variable, proportions in each category of other*
- How do we assess relationship between the two variables?
  - *Compare the conditional distributions: if they are different, there is a relationship*
  - *Example: if conditional distribution for men different from conditional distribution for women, then there is relationship between gender and other categorical variable*

. . . Continued

- What is Simpson's paradox? How does it happen? What can cause it?
  - *Relationship that appears to go one way in one table and a different way in another table*
  - *Sometimes aggregating (summarizing) over a variable can distort the relationship*
  - *Caused by a lurking variable*

To read for next time

§2.6, §3.0

- The question of causation
- Explaining association
- Explaining association: causation
- Explaining association: common response
- Explaining association: confounding
- Establishing causation
- Chapter 3 introduction
- Anecdotal data
- Available data
- Sample surveys and experiments

Lecture 11: §2.6, §3.0

Coming up:

- Is a high correlation proof of cause and effect?
- What are some possible reasons for a high correlation?
- How do we make a convincing case for cause and effect?
- What is anecdotal evidence? Is it informative?
- What is available data? Is it informative?
- What are observational studies and (statistical) experiments? How do they differ? Is one likely to be more informative than the other?

1. For each of the following situations there is a high correlation between the two variables given. Is that high correlation due to cause and effect, common response or confounding? Why?
   (a) Mother's body mass index and daughter's body mass index.
   (b) Amount of saccharin in a rat's diet and number of tumours in the rat's bladder.
   (c) A student's SAT score and that student's 1st year college GPA.
   (d) Monthly flow of money into mutual funds and monthly rate of return on the stock market
   (e) Times per year a person attends religious services and how long the person lives
   (f) Number of years of education a person has and that person's income

2. What is a statistical experiment? How can it help us deduce that one variable is the cause of another? Why is an experiment sometimes impossible to do?

3. How can we establish cause and effect when we cannot do an experiment? Name five criteria that we should look for.

4. What do the statistical terms *individuals* and *population* mean?

5. Suppose we want to find out what causes (or prevents) a second heart attack, by looking at individuals who have already had one heart attack. It is proposed to record age, sex, and some diet and stress variables.
   (a) What are some advantages and disadvantages of looking at *everyone* who has had a heart attack?
   (b) Is it sensible to look at those heart attack victims that are convenient (eg. close to the researcher's hospital)? Why or why not?
   (c) Is it sensible to make an appeal for heart attack victims and to look at only those individuals who respond?
   (d) Is it sensible to look only at those individuals the researcher happens to know about?
   (e) Is it sensible to use a method of choosing individuals so that being chosen has nothing to do with any other possible variable?
   (f) What is the statistical term for the individuals that were chosen from the population?

6. Compare these two situations:

   A. Heart attack victims have a number of variables measured to see what is associated with a second heart attack.
   B. Two medications are being tested for effectiveness at stopping a second heart attack. The available heart attack victims are divided (at random) into two groups; the first group gets one medication and the second group the other. The groups are then compared to see which has fewer second heart attacks.

   (a) One of the situations above is an observational study, and one is an experiment. Which is which, and why?
   (b) Which situation will produce better evidence that something is a *cause* of a second heart attack?
   (c) Suppose you wanted to show that a high-cholesterol diet was a cause of second heart attacks. What kind of experiment could you run? Why would this be better than simply observing the cholesterol level of everyone's diet?

Lecture 11: summary

- Is a high correlation proof of cause and effect?
  - *No! Lurking variables can make a difference*
- What are some possible reasons for a high correlation?
  - *Cause and effect*
  - *Common response*
  - *Confounding*
- How do we make a convincing case for cause and effect?
  - *Experiment (§3.1), otherwise:*
  - *Strong association*
  - *Consistent association*
  - *Higher x goes with higher y*
  - *Alleged cause before effect*
  - *Alleged cause scientifically plausible*
  - *but these not as good as well-designed experiment*

. . . Continued

- What is anecdotal evidence? Is it informative?
  - *Haphazardly chosen cases that come to our attention (maybe because striking)*
  - *No because doesn't represent larger group of cases*
- What is available data? Is it informative?
  - *Data collected in past for another study*
  - *Can be informative (relevance, collection method)*
- What are observational studies and (statistical) experiments? How do they differ? Is one likely to be more informative than the other?
  - *Observational studies: record variables of interest but don't influence anything*
  - *Experiments: impose some treatment on individuals and observe response*
  - *In experiment, treatment imposed; in observational study, not*
  - *Sample survey: kind of observational study*

To read for next time

§3.1:

- Design of experiments
- Comparative experiments
- Randomization
- Randomized comparative experiments
- How to randomize
- Cautions about experimentation
- Matched pairs designs
- Block designs

Lecture 12: §3.1

Coming up:

- What are treatments, factors and levels?
- What are the basic principles of experimental design?
- Why is control necessary?
- Why use randomization?
- What is good about repetition?
- What are some other things we should think about with experiments?
- What are blocks and matched pairs, and why are they useful?

Questions

1. How would you distinguish a statistical experiment from an observational study? What can a statistical experiment achieve that an observational study cannot?

2. What is the statistical term for the following (in the context of a statistical experiment)?
   (a) The thing that is deliberately changed in the experiment.
   (b) One of the ways in which that thing is deliberately changed.
   (c) One of the variables that is deliberately changed.
   (d) The individuals that are part of the experiment.

3. How would you design an experiment to:
   (a) compare two specific diets and see which one is "more healthy" (assume you have a measure of "healthy" in mind)
   (b) compare two specific medications to see which one is more effective at preventing a second heart attack in people that have already had one heart attack.

4. Why is it a good idea to use randomization to assign individuals to treatments?

5. What is another important principle of experimental design?

6. In an experiment to study a new medication for the common cold, the subjects are divided at random into two groups. The first group is given the new medication, while the second is given a medication that looks exactly the same as the new medication, but contains no active ingredient.
   (a) What effect is being guarded against here?
   (b) What is the name for the second group of subjects?
   (c) Why is the above design better than giving the second group nothing at all? In particular, if the subjects in the first group recover from their colds much quicker than those in the second group, what can you say about the effectiveness of the new medication?
   (d) How could you make this study even better? What would be the name for such a study?

7. Line 101 of Table B is 19223 95034 05756 28713 96409 12531. Make random selection of 5 individuals from 1–30.

8. When using Table B, when choosing from 10 subjects, why is it more efficient to use 0 to represent the 10th subject than to select digits 2 at a time?

9. Using StatCrunch to do randomization for experiments.

10. In a randomized experiment:
    (a) What are two reasons that a treatment group might give better results than a control group?
    (b) What is it called when the results from the treatment group are better than those from the control group by more than you would expect by chance?

11. What do the following experiments have in common? What is the statistical name for this? Why might this be better than a randomized experiment, if you can do it?
    - In a study of training programs for running 800m, each subject timed before training program and after.
    - 10 subjects tested a new anti-inflammatory cream by treating one arm with the new cream and the other with a placebo, and measuring delayed-onset muscle soreness in each arm.
    - 44 sixth-graders were divided into 22 pairs, with the students in each pair having equal IQs. One student in each pair received "special training"; other did not. All students did IQ test at end.

12. Suppose you have 2 fifteen-year-old girls and 8 eight-year-old girls, and you want to set up a 5-a-side soccer game.
    (a) Suppose you use ordinary randomization (eg. Table B) to divide the players into two teams. What might happen? Is this desirable?
    (b) How might you do the randomization to produce a better result?
    (c) Suppose now that the girls on the first team receive extra coaching for 3 training sessions, and the girls on the second team do not. After those training sessions, all the girls are rated on their playing ability, to find out whether the coaching was helpful. Explain why your randomization of (b) would produce more trustworthy results.
    (d) Explain how you might proceed if you also had 4 twelve-year-old girls, and you want to divide the girls into two groups fairly.

13. Suppose you have a treatment and a control, and you are running a matched pairs experiment. Invent some data to show that even when there are considerable differences among the pairs, you can still detect even a small effect of treatment by comparing within pairs.

14. Suppose now you have 4 treatments A, B, C, D and 3 blocks 1, 2, 3. An experiment is run as "randomized blocks" with one observation per treatment for each block (and thus 12 observations altogether). Invent some data to show that even if the blocks are quite different, you can still detect differences between the treatments.

Lecture 12: summary

- What are treatments, factors and levels?
  - *Factors: categorical explanatory variables*
  - *Levels: categories of factors*
  - *Treatments: factor level combinations imposed on units/subjects in experiment*
- What are the basic principles of experimental design?
  - *Control, randomization and repetition*
- Why is control necessary?
  - *Don't want to confound treatment with other (eg. lurking) variables*
  - *Example: comparison with control group*
- Why use randomization?
  - *Allocates subjects/units to treatments without favoritism: no way to predict which unit ends up with which treatment*
  - *Produces treatment groups that are similar*
- What is good about repetition?
  - *Reduces role of chance variation*
  - *Can more easily see differences among treatments*

. . . continued

- What are some other things we should think about with experiments?
  - *Double-blind: no-one knows which treatment being given/received*
  - *Lack of realism: prevents generalization of results beyond experiment*
- What are blocks and matched pairs, and why are they useful?
  - *Units may differ in some way that affects response. Arrange units in groups called* **blocks** *and randomize within each one.*
  - *For comparing 2 treatments, can sometimes have 1 subject do both (randomize order), or group subjects into pairs and randomly assign 1 to each treatment. This called* **matched pairs**.
  - *Matched pairs like blocks with 2 measurements per block*

To read for next time

§3.2:

- Sampling design
- Simple random samples
- Stratified samples
- Multistage samples
- Cautions about sample surveys

Lecture 13: §3.2

Coming up:

- What is the name for all individuals we'd like to know about?
- What is the name for all individuals we actually do know about?
- What is a probability sampling design?
- What is a simple random sample?
- What is a stratified random sample, and why might you use it?
- What is a multistage random sample, and why might you use it?
- Is a voluntary response sample a good thing or a bad thing?
- What are some problems you might run into with sampling?

Questions

1. What, in statistical terms, is the population? What is the sample?

2. What is a good way, in general, to select a sample from a population? In particular, how is a simple random sample selected?

3. Use StatCrunch:
   (a) to select a simple random sample of 4 people out of 20.
   (b) to select a stratified random sample of 2 men and 2 women out of a population with 10 men and 10 women. When might you prefer to use a stratified sample over a simple random one?

4. Use StatCrunch to do the following:
   (a) Open up the data set `multi` which has 120 people working for a company in 20 different areas (6 people in each). The variable `areas` is a list of those 20 different areas, for use later.
   (b) Take a simple random sample of 12 employees from the whole company, carrying along their areas. How many different areas did you sample from?
   (c) Now take a simple random sample of 4 of the 20 areas. We are only going to sample employees from these areas.
   (d) Type the numbers 1 through 6 into an empty column. Generate 4 samples from this column.
   (e) Explain how you have now generated a multistage sample. List some of the employees in that sample.

5. Can you describe 4 differences between stratified sampling and multistage sampling? They relate to:
   (a) Accuracy and convenience
   (b) When they can/should be used
   (c) How they use simple random sampling
   (d) The characteristics of the sample drawn

6. Explain how you would take a systematic sample of 5 people out of a population of 100 people (numbered 1–100), and give an example. Is this the same as a simple random sample? Explain why or why not.

7. Give examples of the following, and explain how they can cause problems:
   (a) Undercoverage
   (b) Nonresponse
   (c) Response bias
   (d) Wording of question

Lecture 13: summary

- What is the name for all individuals we'd like to know about?
  - *Population*
- What is the name for all individuals we actually do know about?
  - *Sample*
- What is a probability sampling design?
  - *Any means of choosing a sample using randomization*
- What is a simple random sample?
  - *Choose individuals at random with equal chance, independently of which other individuals also in sample*
- What is a stratified random sample, and why might you use it?
  - *Divide population into strata, and take simple random sample within each one*
  - *When individuals in population differ in way important to response; groups of different individuals make strata*
- What is a multistage random sample, and why might you use it?
  - *Select successively smaller groups from population in stages*
  - *Convenience: easier to deal with smaller parts of population*

. . . continued

- Is a voluntary response sample a good thing or a bad thing?
  - *Bad!*
  - *Individuals choose themselves, often leads to bias*
  - *Better use a probability sample instead*
- What are some problems you might run into with sampling?
  - *Undercoverage*
  - *Nonresponse*
  - *Response bias*
  - *Poorly worded questions*

To read for next time

§3.3:

- Towards statistical inference
- Sampling variability
- Sampling distributions
- Bias and variability
- Sampling from large populations
- Why randomize?

We skip §3.4!

Lecture 14: §3.3

Coming up:

- What is a parameter? What is a statistic?
- Why do we look at samples?
- What is a sampling distribution of a statistic?
- What is sampling variability?
- What is bias?
- What can we guarantee about the sample mean for a simple random sample?

Questions

1. Suppose you wanted to work out the proportion of Japanese cars driven by all UTSC students (at least, the ones who have parking permits).
   (a) If you go to a particular parking lot on a particular day and work out what proportion of those cars that are Japanese, what is the name for this?
   (b) If you get a list from Parking Services of the make of every car driven by a student with a parking permit, and work out the proportion of these cars that are Japanese, what is the name for this?

2. Is a population parameter usually known or unknown? Fixed or random? What about a sample statistic? (What does "random" mean in this context?)

3. A teacher wants to estimate the average size of families in her city. She asks all the students in her class, "how many children are in your family, including yourself?", and calculates the sample mean. Is this a biased or unbiased estimator of the city's mean family size? (Hints: think about whether a family with 0 children can appear in the sample, and about what happens when a family has a lot of children.)

4. True or false? Explain.
   (a) A sample statistic is usually an estimator.
   (b) Bias is a good thing.
   (c) Small sampling variability is a good thing.
   (d) We can work out how close the sample mean will be to the population mean, for any kind of sampling method.
   (e) For simple random sampling, the sample mean is an unbiased estimator of the population mean, and has low sampling variability.
   (f) How much a statistic varies from one sample to another depends on how big the population is.

5. Suppose a college has 57% of its (2000) students being women. If we were to draw a sample of size 100 from this population, what % women might we end up with in our sample? Investigate by simulation using StatCrunch:
   (a) Make a column called `population` by stacking together 1140 1's (57% of 2000) and 860 0's.
   (b) Draw 1000 different samples of size 100 by: Data, Sample Columns; select your column of 0's and 1s; sample size 100, number of samples 1000; check Stacked with a Sample ID; click Sample Columns.
   (c) Calculate the sample mean for each sample (this is the % women in each sample): Stat, Summary Stats, Columns; select the column `Sample (Population)`; select Group By `sample`; click Next; unselect everything except Mean; click Store Output in Data Table; click Calculate.
   (d) Make a histogram of the last column `Mean`. This is the (simulated) sampling distribution of the proportion of women in samples of size 100.
   (e) How would you describe the centre, spread and shape of your sampling distribution?

Lecture 14 summary

- What is a parameter? What is a statistic?
  - *Parameter: some quantity about the population (want to know)*
  - *Statistic: some quantity about the sample (can calculate)*
- Why do we look at samples?
  - *Use sample statistic to make conclusion about population values (inference)*
- What is a sampling distribution of a statistic?
  - *If all possible samples taken, what values might statistic take?*
- What is sampling variability?
  - *With different sample, probably get different test statistic value*
  - *Look at spread of sampling distribution*
  - *If large sampling variability, cannot expect to estimate population parameter accurately*
- What is bias?
  - *Something about sampling procedure causing sample statistic to be systematically too high or low*
- What can we guarantee about the sample mean for a simple random sample?
  - *Unbiased estimate of population mean*

To read for next time

§4.1, §4.2:

- Randomness
- The language of probability
- Thinking about randomness
- The uses of probability
- Probability models
- Sample spaces
- Probability rules
- Assigning probabilities
- Independence and multiplication rule
- Applying the probability rules

Lecture 15: §4.1 and §4.2

Coming up:

- What is randomness?
- What is probability?
- What is a sample space?
- Outcomes and events
- Probability rules
- Where do probabilities come from?
- Independent and disjoint events
- More probability rules

Questions

1. For each of the games below, what can you say about whether you will win the next game? What can you say about the propotion of the next 100 games you will win?
    (a) Toss a (fair) coin, win if you get a Head.
    (b) Roll a (fair, six-sided) die: win if you roll a 6.
    (c) Play a casino game (eg. roulette, betting on Red).

2. What is the sample space in the examples below?
   (a) Toss a coin once.
   (b) Toss 4 coins, count the number of heads.
   (c) Choose a Canadian at random, note down the province of residence.

3. Suppose we toss one fair coin, so that the probabilities of heads and tails are both 0.5. Write down a suitable probability model.

4. Now suppose we toss two fair coins:
   (a) There are four outcomes. Write down the sample space.
   (b) Explain why "exactly one head" is an **event**. Which outcomes make up this event?
   (c) The four outcomes in the sample space each have probability 1/4. What is the probability of the event "exactly one head"?
   (d) What is the probability of the entire sample space? Does this make sense?

5. Suppose you roll a fair six-sided die, once.
   (a) The outcomes are equally likely. What is the probability of each one?
   (b) Find the probability of rolling either a 5 or a 6.

6. Return to tossing two fair coins. This time we are counting the number of heads we get in the two tosses.
    1. Using your work from above, which outcomes make up "exactly one head"? What is the probability of exactly one head?
    2. Which outcomes make up "exactly no heads" and "exactly 2 heads"? What is the probability of each?
    3. Is it true that equally likely outcomes lead to equally likely events?

7. Now we imagine tossing three fair coins, for which there are 8 equally likely outcomes:
   (a) What is the probability of HHT?
   (b) What outcomes make up the event "exactly two heads"? What is the probability of "exactly two heads"?
   (c) What outcomes make up the event "exactly three heads"? What is the probability of "exactly three heads"?
   (d) What outcomes make up the event "at least 2 heads"? What is the probability of "at least 2 heads"?

8. Finally, four fair coins. There are 16 equally likely outcomes, of which 6 have two heads and two tails:
   (a) What is the probability of getting exactly 2 heads?
   (b) Which is more likely: that you will get an equal number of heads and tails, or an unequal number?

9. The *Poisson distribution* (which we do not study further in this course) is often used as a model for "random" events in time, such as goals for a team in a hockey game.
    (a) Use StatCrunch to simulate 100 team scores in hockey games, using a mean of 3 goals per game. (Use Data and Simulate.)
    (b) What is the estimated probability of a team scoring no goals in a game? At least 6?

10. In the situations below (based on tossing 2 fair coins), are the events A and B: independent? overlapping? disjoint? How can you tell?
    (a) A : H on 1st coin, B : H on second coin.
    (b) A : HH, B : "head on 1st coin", so B= $\{HH, HT\}$.
    (c) A : HH, B : TT

11. In a certain lottery, you can win either prize A, prize B, or no prize. $P(A) = 0.01$, $P(0) = 0.90$.
    (a) What is $P(B)$?
    (b) What is the probability of winning some prize?
    (c) If 2 people play, what is prob that both win some prize? That neither wins a prize?

12. Suppose a Canadian has probability 0.2 of catching a cold in month of November. Take a simple random sample of 4 Canadians.
    (a) What is the prob that none of these 4 people catches a cold in November?
    (b) What is the prob that at least one of them does?

13. 51% of Canadians are women and 6% are over 75 years old. Also, 3.8% are both female *and* over 75 years old. Does this mean that, for a randomly chosen Canadian, the events "female" and "over 75" are independent? Does your answer make sense?

14. A person can have blood type O, A, B or AB. $P(O) = 0.45, P(A) = 0.40, P(AB) = 0.11$. A type B can receive a blood tranfusion from another type B or a type O. What is the prob. that a randomly chosen person will be a suitable donor?

15. A die is made biased so that $P(6) = 0.3, P(1) = 0.1$ and the other four faces are equally likely.
    (a) What is prob. of rolling a 2 with this die?
    (b) This biased die and a regular die are rolled together. What is the probability of a total of 11 spots?
    (c) For two regular dice, the prob. of 11 spots is 0.056. Was your answer in the previous part higher or lower than this? Does this make sense?

Lecture 15 summary

- What is randomness?
  - *short-term unpredictable, long-term predictable.*
- What is probability?
  - *Long-term proportion of times an event happens*
- What is a sample space?
  - *List of everything that might happen*
- Outcomes and events
  - *Outcome: something in sample space*
  - *Event: collection of events*
- Probability rules
  - event A: $P(A) \leq 0$ and $P(A) \leq 1$.
  - Sample space S: $P(S) = 1$.
  - $P(A^c) = 1 - P(A)$.

. . . continued

- Where do probabilities come from?
  - *n equally likely outcomes:* $1/n$ *for each outcome*
  - *Long-term observation of events*
  - *Mathematics*
- Independent and disjoint events
  - *Independent:* $P(A)$ *not affected by* B
  - *Disjoint: if* C *happens,* D *can not*
- More probability rules
  - If A, B independent, $P(A \text{ and } B) = P(A)P(B)$.
  - If C, D disjoint, $P(C \text{ or } D) = P(C) + P(D)$.

To read for next time

§4.3, 4.4:

- Random variables
- Discrete random variables
- Continuous random variables
- Normal distribution as probability distribution
- Means and variances of random variables
- The mean of a random variable
- Statistical estimation and the law of large numbers
- Thinking about the law of large numbers
- Rules for means
- Variance of a random variable
- Rules for variances and standard deviations

Lecture 16: §4.3 and §4.4

Coming up:

- What is a random variable?
- Discrete and continuous random variables
- Probability distributions and density curves
- How is normal distribution related to this?
- Mean and SD of random variable
- Rules for means and variances

Questions

1. Give some examples of random variables associated with tossing coins and rolling dice. In each case, what else do you need to complete the specification of a probability distribution?

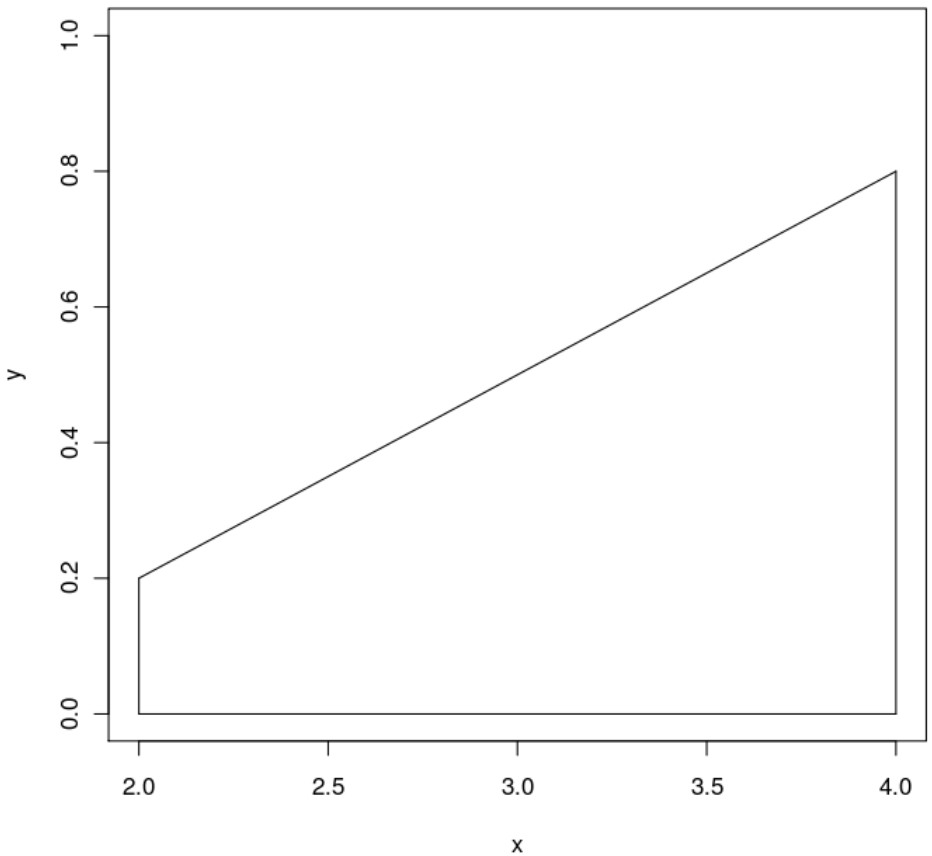
2. For this probability distribution:

Value of X	1	2	3
Probability	0.3	0.6	0.1

   (a) Find $P(X \geq 2)$.
   (b) Find the probability that X is either 1, 2 or 3. Does your answer make sense?

3. Toss 3 fair coins, count # heads (X):
   (a) Write down all 8 possible outcomes, the number of heads contained in each, and the probability of each.
   (b) Find $P(X = 2)$. What is this the probability of, in words?
   (c) Make a table of the probability distribution of X.

4. A continuous random variable X has a density function that is the shape of a trapezoid, as shown below. The area of a trapezoid is its base times the average of the heights at the two ends.



(a) Verify that this is indeed a density function.

(b) Would you guess that $P(X > 3)$ is more or less than 0.5? Explain.

(c) The density function at $x = 3$ is 0.5. Find $P(X > 3)$. (Drawing a picture might be helpful.) Was your guess correct?

5. For this distribution:

Value of X	1	2	3
Probability	0.3	0.6	0.1

find the mean of X. Does it make sense that the mean is less than 2?

6. The distribution of number of heads in 3 tosses of fair coin is:

Value of X	0	1	2	3
Probability	1/8	3/8	3/8	1/8

(a) What is the mean number of heads?
(b) Is this answer surprising?

7. Previously, we investigated the Poisson distribution with mean 3 as a model for goalscoring in hockey games. If we take samples from this distribution, how close might the sample mean be to 3? Use StatCrunch as below:
   (a) Generate a population of 1000 Poisson values by using Data, Simulate, Poisson with 1000 rows and 1 column.
   (b) Take 100 samples of size 5 from this population, using Data, Sample from Columns. Check "Stacked with Sample ID".
   (c) Calculate the mean for each sample using Stat, Summary Stats, and group by Sample, saving the results in a data table.
   (d) Make a stemplot of the sample means. Are they usually close to 3.
   (e) Repeat (b), (c) and (d) for samples of size 100.
   (f) Is the mean of a bigger sample usually closer to 3? What does that say about the informativeness of larger samples compared with smaller ones?

8. This distribution has mean 1:

Value of X	0	1	2
Probability	0.4	0.5	0.1

Find the variance and thus the standard deviation of X.

9. A random variable X has this distribution, with mean 0.9 and SD 0.54:

Value	0	1	2
Prob.	0.2	0.7	0.1

The distribution of the random variable $3X$ is obtained by multiplying all the values by 3 and leaving the probabilities unchanged.

   (a) Write down the distribution of $3X$.
   (b) Calculate the mean of $3X$. How does it compare to the mean of X?
   (c) Calculate the SD of $3X$. How does it compare to the SD of X?

10. A random variable X has this distribution, with mean 0.9 and SD 0.54:

Value	0	1	2
Prob.	0.2	0.7	0.1

The distribution of the random variable $X + 4$ is obtained by adding 4 to all the values and leaving the probabilities unchanged.

(a) Write down the distribution of $X + 4$.
(b) Calculate the mean of $X + 4$. How does it compare to the mean of X?
(c) Calculate the SD of $X + 4$. How does it compare to the SD of X?

11. X takes values 0 and 1 with $P(X = 0) = 0.3, P(X = 1) = 0.7$. Independently of X, Y takes values 0 and 1 with $P(Y = 0) = 0.4, P(Y = 1) = 0.6$. X has variance 0.21 and Y has variance 0.24. Let $Z = X + Y$.
    (a) Verify that $P(Z = 0) = 0.12$, $P(Z = 1) = 0.46$, $P(Z = 2) = 0.42$.
    (b) Z has mean 1.3. What is the variance of Z?
    (c) How could you have worked out the variance of Z without working out the three probabilities for Z?
    (d) What if you didn't know that X and Y were independent?

Lecture 16 summary

- What is a random variable?
  - *Each event produces a number.*
- Discrete and continuous random variables
  - *Discrete: possible values separate (eg. whole numbers)*
  - *Continuous: any value in an interval is possible*
- Probability distributions and density curves
  - *Possible values and probabilities*
  - *Prob. dist (discrete): show by probability histogram*
  - *Density curve (continuous): show by plot of density curve*
- How is normal distribution related to this?
  - *One type of continuous distribution*

. . . continued

- Mean and SD of random variable
  - *Mean* μ *sum of value times prob*
  - *variance* σ^2 *sum of* $(x-\mu)^2 p$
  - *SD* $\sigma = \sqrt{\sigma^2}$
- Rules for means and variances
  - Mean of $a + bX$: $a + b\mu$.
  - Variance of $a + bX$: $b^2\sigma^2$.
  - Mean of $X + Y$: $\mu_X + \mu_Y$.
  - Variance of $X + Y$: $\sigma_X^2 + \sigma_Y^2$, *if* X, Y independent.

To read for next time

§5.1:

- The sampling distribution of a sample mean
- The mean and standard deviation of $\bar{x}$
- The central limit theorem
- A few more facts

Lecture 17: §5.1

Coming up:

- Known population: sample mean might be?
- Sampling distribution of sample mean
- What if population exactly normal?
- What if sample large?

Questions

1. Start from a normal population with mean 20 and SD 5. Investigate, by simulation using StatCrunch, the sampling distributions of the sample mean:
   (a) for samples of size 10
   (b) for samples of size 50

   (hints: first generate 1000 values from a normal distribution using Data and Simulate Data to be your population, and then use Data and Sample Columns to generate your samples, storing the samples "stacked with a sample id". Then use Stat and Summary Stats, grouping by your samples, and saving the means into the data table. Finally, make a histogram or other picture of the sample mean.)
   What do you conclude about the sampling distributions as compared to the population?
   Also, what does the law of large numbers have to say here?

2. Start from a discrete population with five 0's, seven 1's, three 2's and one 6. We will investigate the sampling distribution of the sample mean:
   (a) for samples of size 5
   (b) for samples of size 100

   Note: the population is small, so instruct StatCrunch to select samples *with replacement*.

   In each case, investigate the shape of the sampling distribution by simulation using StatCrunch, and compare with the shape of the original population. What do you observe?

3. State carefully the Central Limit Theorem. How can you apply the Central Limit Theorem in practice?

4. A population has mean 100 and SD 15. You take a sample of size 50.
   (a) What, according to the theory, are the mean and SD of the sampling distribution of the sample mean?
   (b) The population has a somewhat skewed shape. What do you think the shape of the sampling distribution is? How do you know?
   (c) Find the (approximate) probability that a sample of size 50 from this population has a sample mean between 95 and 105.
   (d) What do you think would happen to the above probability if the sample were of size 200 rather than 50? Why?

5. A population has mean 60 and SD 5. Take a sample of size 100 (assume large enough for CLT to apply).
   (a) What is the (approx) probability that the sample mean will be bigger than 61?
   (b) An *individual value* from this population has probability 0.42 of being bigger than 61. Why is this value different to your answer from the previous question?

Lecture 17 summary

- Known population: sample mean might be?
  - *If SRS, sample mean has mean μ, SD $\sigma/\sqrt{n}$*
- Sampling distribution of sample mean
  - *Might be anything. But see below.*
- What if population exactly normal?
  - *Sample mean exactly normal also.*
- What if sample large?
  - *Sample mean approximately normal for any population, even a discrete one (central limit theorem)*

To read for next time

§5.2:

- Sampling distributions for counts and proportions
- The binomial distribution for sample counts
- Binomial distributions in statistical sampling
- Finding binomial probabilities
- Binomial mean and standard deviation
- Sample proportions
- Normal approximation for counts and proportions
- The continuity correction (just read enough of this to see what it does; we don't do calculations)

We also don't do the binomial formula.

Lecture 18: §5.2

Coming up:

- Population of successes/failures. Sampling distribution of sample success count?
- How to find probs from this distribution:
  - small sample
  - large sample
- Mean and SD of sampling distribution of:
  - sample count?
  - sample proportion?

Questions

1. A sequence of 20 trials produces "successes" and "failures". Consider the total number of successes in the 20 trials. What two other conditions must be true for that count to have a binomial distribution?

2. Does the binomial distribution apply in the cases below?
   (a) Joe buys 1 lottery ticket every week for a year; X = number of times he wins.
   (b) A person keeps taking a driving test until she passes; X = number of attempts.
   (c) Toss a coin 10 times; count occasions you get 2 heads in a row (eg. HTHHTTTHHH is 3 successes)
   (d) Large population of people who agree or disagree with a statement; take a simple random sample from population, count number in sample who agree.

3. Use StatCrunch to explore the shapes of these binomial distributions:
   (a) $n = 5, p = 0.5$
   (b) $n = 20, p = 0.9$
   (c) $n = 30, p = 0.2$
   (d) $n = 500, p = 0.4$

   Which distribution does the last one resemble?

4. 90% of the TVs produced in a factory come off the production line working properly the first time. For the questions below, use StatCrunch to draw a picture of the distribution and obtain rough answers.
   (a) Consider a simple random sample of 20 of these TVs. Is it likely that fewer than 85% (17) or more than 95% (19) of them will work first time?
   (b) Now consider a simple random sample of 200 of these TVs. Is it likely that fewer than 85% (170) or more than 95% (190) of them will work first time?
   (c) Does this suggest that if you have a larger sample, the sample proportion of working TVs will be typically be closer to the population proportion?
   (d) Does this suggest that if you have a larger sample, the sample *count* of working TVs will be closer to the value you'd expect?

5. Use Table C for the following, and check your answers with StatCrunch:
   (a) In a binomial distribution with $n = 9, p = 0.3$, what is the probability of exactly 2 successes?
   (b) In the same binomial distribution, what is the probability of 2 successes or less?
   (c) If $n = 10, p = 0.7$, what is the probability of exactly 6 successes?

6. Use the normal approximation to the binomial for this:
   (a) When $n = 100, p = 0.4$, what is the probability of 35 successes or fewer?
   (b) In the same binomial distribution as the previous part, what is the probability that the sample has a proportion of 0.35 successes or fewer. Why is the answer the same as for the previous part?
   (c) Find the exact answer to the first part using StatCrunch, and discuss why a continuity correction would make your approximate answer more accurate.

7. Suppose we want to find the probability of exactly 25 successes in a binomial distribution with $n = 30, p = 0.9$. What options do we have?

Summary of Lecture 18

- Population of successes/failures. Sampling distribution of sample success count?
  - *binomial n* sample size, *p* proportion of successes in *population*
  - *Requires fixed n, constant p, independent trials*
  - *OK (at least approx) with simple random sample*
- How to find probs from this distribution:
  - *small sample: use table C (if possible)*
  - *large sample: use normal approx if* $np \geq 10, n(1-p) \geq 10$
- Mean and SD of sampling distribution of:
  - *sample count: mean np, SD* $\sqrt{np(1-p)}$.
  - *sample proportion: mean p, SD* $\sqrt{p(1-p)/n}$.

To read for next time

§6.1:

- Overview of inference
- Estimating with confidence
- Statistical confidence
- Confidence intervals
- Confidence interval for a population mean
- How confidence intervals behave
- Choosing the sample size
- Some cautions

Lecture 19: §6.1

Coming up:

- How to estimate a parameter from 1 sample of data?
- Calculations: margin of error, z^*
- How to make the margin of error smaller
- How to choose sample size to get specified margin of error?
- When might confidence interval not work?

Questions

1. SAT-M scores have a roughly normal distribution with unknown mean μ and SD 100. 500 high-school seniors from California are sampled.
   (a) How likely is it that the sample mean will be within 9 points of μ (that is, between $\mu - 9$ and $\mu + 9$)?
   (b) Our sample had a sample mean of 461. Assuming that it is one of those samples whose sample mean is within 9 points of μ, what can you say about μ?
   (c) What would make you wrong in your statement about μ?

2. How do you get z^* for a confidence interval?
   (a) In particular, find z^* for an 80% CI.
   (b) Make a table of z^* values for 90%, 95% and 99% confidence levels.

3. A large hospital wants to estimate the average length of time patients stay in the hospital. Sample 90 records, find mean stay is 4.63 days; SD of length of stay known from previous data to be 3.7 days.
   (a) Find a 95% confidence interval for the population mean.
   (b) Find a 99% confidence interval for the population mean.
   (c) Which interval is bigger? Why?
   (d) Check your answers using StatCrunch.

4. In our confidence interval situation:
   (a) What is random and what is fixed?
   (b) If we take one sample and calculate our confidence interval for the population mean, what precisely do we mean by "95% confidence"?
   (c) Do we know for sure whether our calculated confidence interval actually contains μ or not?

5. In the hospital example, a 95% interval had margin of error 0.76 for $n = 90$, using $\sigma = 3.7$.
   (a) How many patient records are needed to reduce the margin of error to 0.5?
   (b) Do you find the answer surprising?
   (c) Check your answer in StatCrunch (assume that your sample mean is still 4.63, and check that the upper and lower limits of the interval are $2(0.5) = 1$ apart).

6. Some general questions about confidence intervals:
   (a) Can we use this procedure if we have eg. a stratified sample?
   (b) What if we have outliers in our data?
   (c) What if we have a small sample?
   (d) What if we don't know the population SD σ?

Lecture 19 summary

- How to estimate a parameter from 1 sample of data:
  - *Confidence interval*.
- Calculations: margin of error, z^*:
  - z^* from table, according to confidence level
  - $m = z^*\sigma/\sqrt{n}$.
- When might CI not work:
  - *do you have SRS?*
  - *is population too non-normal?*
  - *is σ known?*
  - *If any of above fail, CI won't work.*
- How to make the margin of error smaller:
  - *larger sample*
  - *smaller σ*
  - *smaller confidence level*
- How to choose sample size to get specified margin of error:
  - $n = (z^*\sigma/m)^2$.

To read for next time

§6.2:

- Tests of significance
- The reasoning of significance tests
- Stating hypotheses
- Test statistics
- P-values
- Statistical significance
- Tests for a population mean
- Two-sided significance tests and confidence intervals
- P-values versus fixed α

Lecture 20: §6.2

Coming up:

- How to decide whether population mean equals given value or not?
- How to write hypotheses?
- How to assess strength of evidence against null hypothesis?
- What does "statistically significant" mean?

Questions

1. Explain how the Canadian legal system tries to judge the innocence or guilt of an accused person by assessing the evidence presented in court.
2. Relate the above to judging whether a statement about a population mean ("null hypothesis") is true or false, based on evidence contained in sample mean.

3. Calcium levels in the blood of healthy young adults vary with mean 9.5 and SD 0.4 mg/dl. Suppose we conduct a study of pregnant women in rural Guatemala: is their blood calcium level different on average?
   (a) What are we trying to prove here? This is the alternative hypothesis. Write it in symbols.
   (b) If the pregnant women are in fact no different from young adults generally, what would be true about their mean calcium blood level? This is the null hypothesis. Write it in symbols.
   (c) Suppose we had been trying to prove that these women had a *higher* blood calcium level than healthy young adults generally. What would the alternative hypothesis have been then?

4. In the previous question, suppose the null hypothesis is true, and we intend to take a sample of 180 pregnant women.
   (a) What is the sampling distribution of the sample mean in that case?
   (b) The observed sample mean is 9.58. How likely are we to get a value this far or farther from 9.5, if the null hypothesis is true? (This is using $H_a : \mu \neq 9.5$). Your result is called the P-value for this test.
   (c) On the basis of this sample, do you think the population mean for Guatemalan pregnant women is 9.5, or not? Explain briefly.
   (d) How would your P-value change if we had $H_a : \mu > 9.5$?
   (e) How would your P-value change if we had $H_a : \mu < 9.5$?

5. What do you conclude in the following cases:
   (a) $\alpha = 0.05$, P-value 0.027.
   (b) $\alpha = 0.01$, P-value 0.027.
   (c) $\alpha = 0.10$, P-value 0.027.
   (d) $\alpha = 0.01$, P-value 0.007.
   (e) $\alpha = 0.05$, P-value 0.45.

6. In what kind of situation might you choose:
   (a) a small α like 0.01?
   (b) a large α like 0.10?

7. The General Health Questionnaire (GHQ) measures mental health (low score better). In general population, SD is $\sigma = 5$. Researcher wants to show that mean GHQ for all unemployed men exceeds 10. Sample of 49 unemployed men, sample mean 10.94.
   (a) Write down suitable null and alternative hypotheses.
   (b) Choose a value for α.
   (c) Calculate the test statistic and P-value for your hypotheses. What do you conclude?
   (d) Check your calculations using StatCrunch.

8. Summarize how you would calculate P-values for each combination of: the possible alternative hypothesis, and whether $z > 0$ or $z < 0$. Draw pictures in each case.

9. A sample of 24 male long-distance runners gave a sample mean weight of 61.8 kg. Assume that the SD of the weights of all male long-distance runners is 4.5 kg. Let μ be the mean weight of all male long-distance runners. Use StatCrunch for the following:
   (a) Calculate a 95% CI for μ.
   (b) Calculate a 99% CI for μ.
   (c) Obtain P-values for tests of the following H_0 values of μ, against a two-sided alternative: 59, 62, 64.
   (d) Display all your results in a table, and explain how your tests and confidence interval results are compatible.

Lecture 20 summary

- How to decide whether population mean equals given value or not?
  - *Use a test of significance*
- How to write hypotheses?
  - *Null H_0: given value is correct*
  - *Alternative H_a: it is wrong, eg. $H_a : \mu \neq 20$ (two-sided), or $H_a : \mu < 20$ or $H_a : \mu > 20$ (one-sided)*
- How to assess strength of evidence against null hypothesis?
  - *first, calculate test statistic*
  - *then, find P-value*
  - *Small P-value means strong evidence against H_0*
- What does "statistically significant" mean?
  - Evidence against H_0 is strong enough (P-value small enough) to reject H_0.
  - Choose α in advance (eg. $\alpha = 0.05$)
  - Result statistically significant if P-value $< \alpha$ (reject H_0 in favour of your H_a)
  - otherwise do not reject H_0.

To read for next time

§6.3, §6.4:

- Use and abuse of tests
- Choosing a level of significance
- What statistical significance does not mean
- Don't ignore lack of significance
- Statistical inference is not valid for all sets of data
- Beware of searching for significance
- Power and inference as a decision
- Power
- Increasing the power
- Inference as decision
- Two types of error
- Error probabilities
- The common practice of testing hypotheses

Lecture 21: §6.3 and 6.4

Coming up:

- Why quote a P-value instead of decision to reject or not?
- Must a result important in practice be statistically significant?
- Must a statistically significant result be practically important?
- What are some situations when you wouldn't do a test of significance?
- What decision errors can we make in testing?
- What does the power of a test measure?

Questions

1. P-values and decisions:
   (a) Suppose I do a test with $\alpha = 0.05$ and tell you that I rejected H_0. What does that tell you about my P-value?
   (b) Suppose that you wanted to do this test with $\alpha = 0.01$. What can you conclude from what I told you?
   (c) Suppose now I tell you that my P-value was 0.0234. What do you conclude now, at $\alpha = 0.01$?

2. Suppose we test $H_0 : \mu = 20$ vs. $H_a : \mu \neq 20$ using sample $n = 10,000$, $\sigma = 0.2$, $\alpha = 0.05$. The sample mean is $\bar{x} = 20.005$.
   (a) Would you guess that the sample mean is (statistically) significantly different from 20?
   (b) Calculate the z test statistic and P-value for this test. What do you conclude?
   (c) The researchers conducting the study tell you that the difference between 20 and 20.005 is not important in practice. Is this consistent with the result of your test? Why or why not?

3. This is similar to the previous question. This time, $H_0 : \mu = 20$, $H_a : \mu \neq 20$, $\bar{x} = 26$, $n = 2$ and $\sigma = 5$.
   (a) This time the sample mean is far from the null hypothesis. Would you expect to reject the null hypothesis?
   (b) Calculate the z-statistic and P-value. What do you conclude?
   (c) This time, the researchers tell you that the difference between 20 and 26 is definitely important in practice. Is this consistent with the result of your test? Why or why not?
   (d) Extra: use StatCrunch to calculate 95% confidence intervals for the population mean in this question and the last one. Does this help to explain what is going on?

4. Usually, we are searching for *small* P-values, so that we can reject our null hypothesis in favour of our alternative. Suppose we have some scientific theory that we test, based on an experiment with adequate sample size, and obtain a P-value that was not small. Can you think of two reasons why this finding might be interesting? Explain briefly.

5. An industrial process is supposed to produce output with mean 60. So each day, a sample is taken from the output and a test is conducted of $H_0 : \mu = 60$ vs. $H_a : \mu \neq 60$. The process manager thinks that if the null hypothesis is rejected, there must be a fault in the process. Perform a simulation in StatCrunch to see whether the process manager is right, as follows:
   (a) Generate a simulated population of 1000 normal values with mean 60 and SD 10.
   (b) Take 100 random samples of size 50 from this population, using Data and Sample Columns. Store the samples Stacked with a Sample ID, as we did in §5.1.
   (c) Carry out a test of the hypotheses given above for each sample. This is done by Stat, Z-statistics, 1-sample Z, With Data. Select the column with the random-sample data in it, fill in the population SD of 10, and Group By the sample number. Set up the test, click Calculate, and you'll see the results of the 100 tests.
   (d) These tests were all based on data where the null hypothesis is correct. What would you expect to see in the P-values? What do you see?
   (e) Do you think the process manager is right? Explain.

6. One of the concerns about SAT scores is that they are increasing over time. The historical mean is 450. Suppose we consider taking a sample of 500 students. What is the power of this test to detect an increase of 10 points in mean score? Use $\alpha = 0.05$ and assume $\sigma = 100$; proceed by simulation as below, using StatCrunch:
    1. Generate a large normally-distributed population with mean 460 (why?) and SD 100, using Data and Simulate. (A population of size 5000 is good.)
    2. Generate a large number (100 will do) of samples of size 500 from this population, using Data and Sample Columns. Save them "stacked with a sample ID".
    3. For each sample, test $H_0 : \mu = 450$ against $H_a : \mu > 450$ (using your different samples as "group by"). You'll get a table with 100 P-values in it. How many of those P-values are less than 0.05? This, out of 100, is the power of your test. (Alternatively, save the output as a Data Table, and then sort the P-values into order. This makes it easier to count how many of them are less than 0.05.)

7. In the situation of the previous question, roughly how large a sample should be used to obtain a power of 0.60 (60%)? Assess this by trying a few different sample sizes until you find one with about the right power. (A larger sample size should give a larger power, all else being equal.)

Lecture 21: summary

- Why quote a P-value instead of decision to reject or not?
  - *P-value expresses how strong evidence against H_0 is, but decision only says whether you thought it was strong enough or not.*
- Must a statistically significant result be practically important?
  - *No: with a very large sample, a tiny difference from H_0 can be statistically significant.*
- Must a result important in practice be statistically significant?
  - *No: with a small sample, the sample could be far from H_0 and yet not significant.*
- What are some situations when you wouldn't do a test of significance?
  - *when you don't have a SRS*
  - *when the data have outliers*
  - *when the data generating the hypotheses also being used to test them*
  - *when you are running many tests at once*

. . . Part 2

- What decision errors can we make in testing?
  - *rejecting H_0 when it is true (Type I error)*
  - *failing to reject H_0 when it is false (Type II error)*
- What does the power of a test measure?
  - *How likely a test is to reject a false H_0.*

Lecture 22: §7.1

Coming up:

- How to do inference for population mean when σ not known?
- How to get t^* for confidence interval?
- How to get P-value for test?
- How to analyze matched pairs data?
- What if the population is not normal?
- What to do if can't use t procedures?

Questions

1. Investigate whether using the known-σ procedure works when you *don't* know σ by simulation in StatCrunch, as follows.
   (a) Generate a population of 1000 values from a normal distribution with mean 0 and SD 1, using Data and Simulate Data.
   (b) Generate 100 samples of size 5 from this population, using Stat and Sample Columns. Store them "stacked with a sample ID".
   (c) Calculate a chapter 6-style confidence interval for each sample, by selecting Stat, Z-statistics, 1-sample Z, With Data. Enter your column of sampled values as your data, and Group By sample number. Select a 95% CI. Store the results in the data table.
   (d) This produces 100 confidence intervals. Count how many of them actually contain the population mean of 0. (Or investigate Data, Bin Column and Stat, Table, Contingency, With Data.
   (e) How many of your intervals actually contain 0? How many should contain 0? Does this mean that the Z procedure gives intervals that are too wide, too narrow or about right?

2. Instead of using z^* for a confidence interval, we use t^*, from table D. First find the df as $n-1$ (row), then find your confidence level along the bottom (column). This gives t^*.
   (a) Find t^* for the following intervals:
      (i) 95% CI with $n = 5$
      (ii) 99% CI with $n = 5$
      (iii) 95% CI with $n = 51$
      (iv) 95% CI with $n = 1001$
      (v) z^* for 95% CI.
   (b) What appears to happen as the confidence level gets bigger?
   (c) What appears to happen as the sample size gets bigger?

3. What are sensible t^* values to use for a 95% CI when $n = 48$, $n = 120$? What, in general, should you do if your df is not in the table?

4. When people buy bicycles, they often buy other accessories too (helmet, water bottle, etc.) A bike store took a random sample of 12 bike-buying customers and found that the sample mean amount spent on accessories was \$77.83 and the sample SD was \$33.51.
   (a) Find a 95% confidence interval for the mean sales of accessories.
   (b) Check your answer using StatCrunch.

5. As for confidence intervals, we do t tests by replacing the unknown σ by s, calling the test statistic t, and using Table D to get a P-value. What can you say about the P-value in each of the following cases?
   (a) $n = 10, t = 2.37, H_a : \mu > 10$
   (b) $n = 10, t = 2.37, H_a : \mu \neq 10$
   (c) $n = 20, t = 3.3, H_a : \mu \neq 10$
   (d) $n = 20, t = 5.1, H_a : \mu \neq 10$
   (e) $n = 20, t = 2.5, H_a : \mu < 10$

6. At a second bike shop, mean accessory sales per bike sold are \$90. Is there evidence that mean accessory sales for all bikes sold at the first bike shop are less than this?
   (a) Write down suitable hypotheses.
   (b) Use the sample data $\bar{x} = 77.83, s = 33.51, n = 12$ to obtain the test statistic.
   (c) Obtain a P-value (as accurately as Table D will let you). What is your conclusion?
   (d) Check your work with StatCrunch. Is the P-value you obtained from StatCrunch consistent with the one you got from Table D?

7. The t procedures are most useful for small samples. They are derived from populations having normal distributions. What if the populations are not normal?

8. Make a table summarizing how you use Table D to obtain a P-value for each of the 3 possible types of H_a and the 2 cases of whether t is positive or negative.

9. Back in Chapter 3, we learned about two different ways of doing a comparative study, matching and randomizing.
   (a) Briefly review these two ways.
   (b) How would matching fit into the 1-sample t framework that we just learned? (That is, how could we analyze a matched pairs experiment using methods that we already know?)

10. Sleep apnea is a condition in which a sleeping person stops breathing for a moment. It is especially serious in premature infants. A drug is tested on 13 premature infants, and the number of apneic episodes per hour recorded for each before and after the drug is given. Some results are shown below.

	n	mean	SD
Before	13	1.751	0.855
After	13	0.984	0.833
Difference	13	0.767	0.524

(a) Explain how this is a matched pairs experiment.
(b) Test whether the drug reduces the mean number of apneic episodes.
(c) Calculate a 90% confidence interval for the mean difference in apneic episodes per hour, before minus after. What is your confidence interval telling you about how effective the drug is?
(d) Check your results using StatCrunch.
(e) (optional) Get hold of the data, and check your answers again with StatCrunch.

11. Table 7.3 on page 421 of the text gives lengths in seconds of 50 audio files sampled from an iPod. We are interested in whether the "average" length of audio file is 240 seconds or not. Use StatCrunch to answer the following:
    (a) Why should you have doubts about using a t-test for these data? Hint: look at a picture.
    (b) One approach is to do a "transformation" of the data: take the logarithms (ln) of each value, and test whether the mean of the log-values could be $\ln(240) = 5.480$ or not:
        (i) Calculate a column of log file lengths using Data and Transform Data.
        (ii) Check that the new column is much less skewed.
        (iii) Do a 1-sample t-test on the new column, testing $H_0 : \mu = 5.480$ vs. $H_a : \mu \neq 5.480$. What do you conclude?

12. Another possibility, for the same data as above, is the "sign test". Let m be the population *median*. This uses the original data to test $H_0 : m = 240$ vs. $H_a : m \neq 240$:
    (a) If H_0 is true, what is the probability that an individual observation will be above 240? (Ignore the possibility that an observation could be exactly equal to 240.)
    (b) Out of the 50 values in the sample, what is the distribution of the number X of values bigger than 240, if H_0 is true?
    (c) In the data, 32 of the 50 values were bigger than 240. What is $P(X \geq 32)$ in your distribution above? Double it to obtain a P-value for your test. What do you conclude now?
    (d) (optional) Compare StatCrunch's built-in sign test (Stat, Nonparametrics, Sign Test). Do you get a similar P-value? Why is the P-value not exactly the same?

13. For the sleep apnea example, what is the power of the matched pairs test to detect a reduction of 0.5 apneic episode per hour? Assume that the differences have SD 0.6. Follow the steps below:
    (a) Generate a population of 1000 values from a normal distribution with mean 0.5 and SD 0.6 (the differences).
    (b) Obtain 100 samples each of size 13 from this population, "stacked with column ID".
    (c) For each sample, test $H_0 : \mu = 0$ vs. $H_a : \mu > 0$, obtaining a P-value. Save these P-values in the worksheet.
    (d) How many of the P-values are less than 0.05? What, then, is the (simulated) power of the test?

Note that the power of any t-test can be obtained in the same way, not just a matched-pairs one.

Lecture 22: summary

- How to do inference for population mean when σ not known?
  - *Replace σ by s and replace z by t.*
- How to get t^* for confidence interval?
  - *Use Table D (bottom row) with correct df.*
- How to get P-value for test?
  - *Use Table D (top row) to get approx answer.*
- How to analyze matched pairs data?
  - *Calculate differences and use t procedures on them.*
- What if the population is not normal?
  - *t-test doesn't depend much on population shape unless n very small.*
- What to do if can't use t procedures?
  - *Transform data to more normal shape*
  - *Use eg. sign test*

§7.2

Coming up:

- What if we have 2 independent samples?
- What df to use?
- Guidelines for use of these procedures?

Questions

1. Suppose population 1 has mean μ_1, pop. SD σ_1 and population 2 has mean μ_2, SD σ_2. Samples are taken from each population: the sample from population 1 has size n_1, sample mean $\bar{x}_1$, sample SD s_1, and the sample from population 2 has size n_2, sample mean $\bar{x}_2$, sample SD s_2. Assume that both populations are normal.
   (a) What is the sampling distribution of $\bar{x}_1$? $\bar{x}_2$?
   (b) What is the sampling distribution of $\bar{x}_1 - \bar{x}_2$?
   (c) What therefore is the distribution of $\bar{x}_1 - \bar{x}_2$, suitably standardized?
   (d) In practice σ_1^2 and σ_2^2 are not known, and are replaced by s_1^2 and s_2^2. What distribution would you guess the resulting statistic to have?

2. Using the results of the previous question, write down the formulas for the two-sample t confidence interval and test statistic.

3. A study assessed the effect of piano lessons on spatiotemporal reasoning.
Treatment group: 34 preschool children given piano lessons for 6 months, spatiotemporal reasoning measured.
Control group: 44 preschool children with no piano lessons, spatiotemporal reasoning measured at end.
Data:

Group	n	Mean	SD
Treatment	34	3.62	3.06
Control	44	0.39	2.42

   (a) Calculate the standard error of the difference in sample means.
   (b) Obtain a 95% confidence interval for the difference in population means. Do the piano lessons appear to be effective?
   (c) Is there evidence of a difference in mean spatiotemporal reasoning score between the two groups? Do a suitable test.
   (d) Use StatCrunch to do the last two parts, and compare with the answers you obtained. Why is there a difference?

4. A study of native butter clams measured the width and length of 4 randomly chosen clams:

Width	3.1	4.2	4.8	5.6
Length	4.1	5.5	6.6	7.0

   (a) Is a paired or two-sample t-test more appropriate? Explain.
   (b) The four differences, width minus length, have sample mean -1.375 and sample SD 0.33. Does this information enable you to calculate a 95% confidence interval for width minus length for all clams? If so, do it.
   (c) Enter the data into StatCrunch and do the appropriate test. If you have results to compare, compare them.

5. A comparison was made between the numbers of cigarettes smoked per day by randomly chosen females and males. The 33 females had a sample mean of 6.94 and sample SD of 7.47; the 19 males had a sample mean of 4.21 and sample SD of 5.57.

   (a) Is a paired or two-sample t test more appropriate? Explain.
   (b) Use the data, if possible, to assess the evidence that the mean numbers of cigarettes smoked per day is different for males and females. What do you conclude?
   (c) Do your analysis in StatCrunch, explaining differences from (b), if any.

6. In the situations below, would you be happy to use a two-sample t procedure, or not? If not, why not?
   (a) $n_1 = 50, n_2 = 30$, populations both normal.
   (b) $n_1 = 50, n_2 = 30$, populations mildly non-normal.
   (c) $n_1 = 80, n_2 = 2$, populations might not be normal.
   (d) $n_1 = 20, n_2 = 20$, population 1 skewed to left, population 2 skewed to right.
   (e) $n_1 = 20, n_2 = 20$, both populations skewed to right.

Lecture 23: summary

- What if we have 2 independent samples?
  - *Look at difference between means.*
  - *CI*:
$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
  - *Test statistic:*
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
- What df to use?
  - *Smaller sample size minus 1 (hand)*
  - *More accurate but complicated formula (software)*
- Guidelines for use of these procedures?
  - *As for 1-sample (based on smaller sample size)*
  - *Equal sample sizes better*

Lecture 24: §8.1, §8.2

Coming up:

- Inference about population proportion based on sample proportion?
- Sample size for margin of error of CI?
- When to use?
- How to compare two proportions?
- CI procedure?
- Testing procedure
- When to use?

Questions

1. A new snack food, marketed to students, is being tested. It will be profitable to sell if more than 10% of all students will buy it. A simple random sample of 500 students reveals that 71 of those students would buy the snack. Is this evidence that the new snack would be profitable when marketed to all students?
   (a) Write down suitable null and alternative hypotheses.
   (b) If the null hypothesis is true, what is the distribution of the number of students X in a sample of 500 that would buy the snack?
   (c) Does a normal approximation apply to this distribution?
   (d) What would be the appropriate mean and SD for the sampling distribution of $\hat{p} = X/500$?
   (e) Calculate a suitable test statistic and P-value to assess the sample evidence. What do you conclude?
   (f) Verify your analysis in StatCrunch.

2. In the situation of the previous question (simple random sample of 500 students, of whom 71 would buy the snack), make a 95% confidence interval for the proportion of *all* students who would buy the snack:
   (a) What is the distribution of the number of students out of 500 that would buy the snack? One of the parameters is unknown.
   (b) The margin of error for the confidence interval has this unknown parameter in it. What is your best guess at what the unknown parameter's value is?
   (c) Complete the calculation for the confidence interval.
   (d) Obtain the confidence interval with StatCrunch, and compare your answers.

3. In our example, how many students must we sample to get a margin of error of 0.02 or less in a 95% confidence interval.
   (a) Write down the formula for the margin of error of a general confidence interval for a proportion.
   (b) Write the formula in terms of n.
   (c) What is your best guess at p (use the data you have)? What sample size would this suggest for a 95% CI with a margin of error of 0.02 or less?
   (d) If you have no guess at p you can use $p = 0.5$. What effect does that have on the required sample size?

4. Suppose now that 12% of all students actually would buy the snack food. What is the power of the test above to reject $H_0 : p = 0.10$ in favour of $H_a : p > 0.10$ with $n = 500$? Follow the steps below, using StatCrunch:
   (a) Generate a population of 5000 binomial values using Data, Simulate, Binomial. Use $n = 1$ here. You'll get a column of 0s and 1s, with 1 meaning "would buy".
   (b) Obtain 100 random samples of 500 values each from this population, using Data and Sample Columns. Store them "stacked with a sample ID".
   (c) For each of your samples, test $H_0 : p = 0.10$ against $H_a : p > 10$, and save the results as a data table.
   (d) How many out of your 100 tests had a P-value less than 0.05? This is the power of your test.
   (e) Would you expect a higher or lower power if p were actually 0.15? Why?

5. Suppose a sample is taken from a binomial distribution with $n = n_1, p = p_1$, and suppose the normal approximation applies. What is the (approximate) sampling distribution of $\hat{p}_1 = X_1/n_1$, where X_1 is the number of successes?
6. Suppose, independently of the above, a second sample is drawn from a binomial distribution with $n = n_2, p = p_2$. Write down the (approximate) sampling distribution of $\hat{p}_2$.
7. We want to compare the two sample proportions. What is the (approximate) sampling distribution of $\hat{p}_1 - \hat{p}_2$? Note that this depends on the unknown p_1 and p_2.

8. When doing a confidence interval, we replace the unknown p_1 and p_2 under the square root by their sample estimates. Write down a formula for a confidence interval for $p_1 - p_2$.

9. In a study of syntax texts, are references to females more or less likely to refer to juveniles ("girl" vs. "woman") than references to males ("boy" vs. "man")? Data considered to be a random sample from all syntax texts.
Data: for females, $n_1 = 60$ of which $X_1 = 48$ were juvenile; for males, $n_2 = 132$ of which $X_2 = 52$ were juvenile.
Calculate a 95% confidence interval for the difference in proportions of juvenile references between male and female references.

10. The test of most interest when comparing two proportions has $H_0 : p_1 = p_2$.
    (a) Letting $p = p_1 = p_2$, what does the test statistic become?
    (b) How do you estimate p from the samples?

11. From the syntax text data, assess the evidence that the proportions of juvenile references are different between male and female references. Data $X_1 = 48, n_1 = 60, X_2 = 52, n_2 = 132$. (1 is female and 2 male).

Lecture 24: summary

- Inference about population proportion based on sample proportion?
  - *Use normal approx to binomial and methods of §5.2*
  - *CI:*

$$\hat{p} \pm z^* \sqrt{\frac{p(1-p)}{n}}$$

  - *Test statistic (p from H_0):*

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- Sample size for margin of error of CI?
  - $n = (z^*/m)^2 p^*(1-p^*)$
  - *or use* $p^* = 0.50$ *with no guess at* p
- When to use?
  - *Test and CI both OK when normal approx to binomial OK*
  - *CI a little more stringent*

. . . Part 2

- How to compare two proportions?
  - *Look at difference in sample proportions*
- CI procedure?
  - *Work out*

$$SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  - *CI is* $(\hat{p}_1 - \hat{p}_2) \pm z^* SE_D$.
- Testing procedure
  - *Work out*

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}; \qquad SE_{Dp} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

  - *test statistic*

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{Dp}}$$

- When to use?
  - *Again depends on normal approx to binomial.*