

Chapter 1 – Looking at Data – Distributions

What Statistics is all about

- **Get information** from data (numbers, facts)
- Draw pictures and calculate numbers that **describe** the data
 - *can make a complicated list of numbers look very simple*
- Use data to **make decisions and conclusions** about the world.
 - *how do we know that a new drug actually works?*
 - *how do housing costs change over time?*

Data

Car magazine collected information about (38) models, like this:

Car	MPG	Weight	Cylinders	Horsepower	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

- each **row** describes *car* (**individual**)
- each **column** describes *feature of car* (**variable**)

Kinds of variables (chapter 1 intro)

Variables can be:

- **quantitative**: measured or counted (mpg, cylinders)
- **categorical**: name or group category (country)

Distribution of variable: list of values, how often each value occurs.

Data set

is: a number of variables for each of a number of individuals, plus a “story” of:

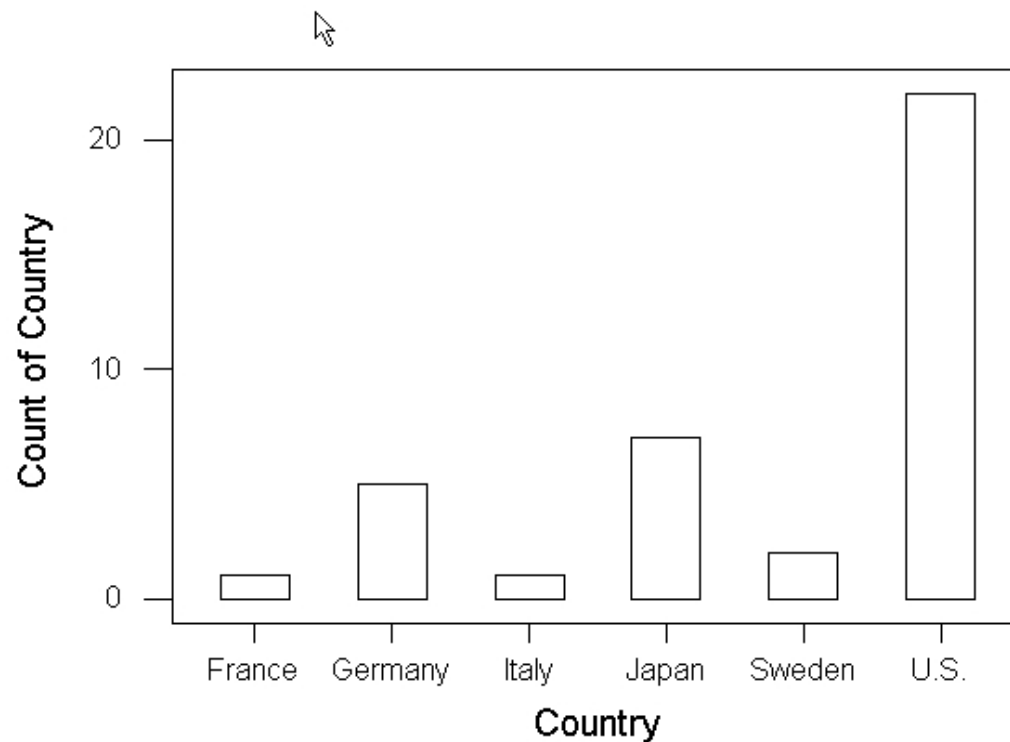
- **why** data were collected
- **who** is in there (which individuals)
- **what** variables measured and how

Can look at variables one at a time, or several together.

Displaying distributions with graphs (§1.1)

Bar chart

for categorical variables:



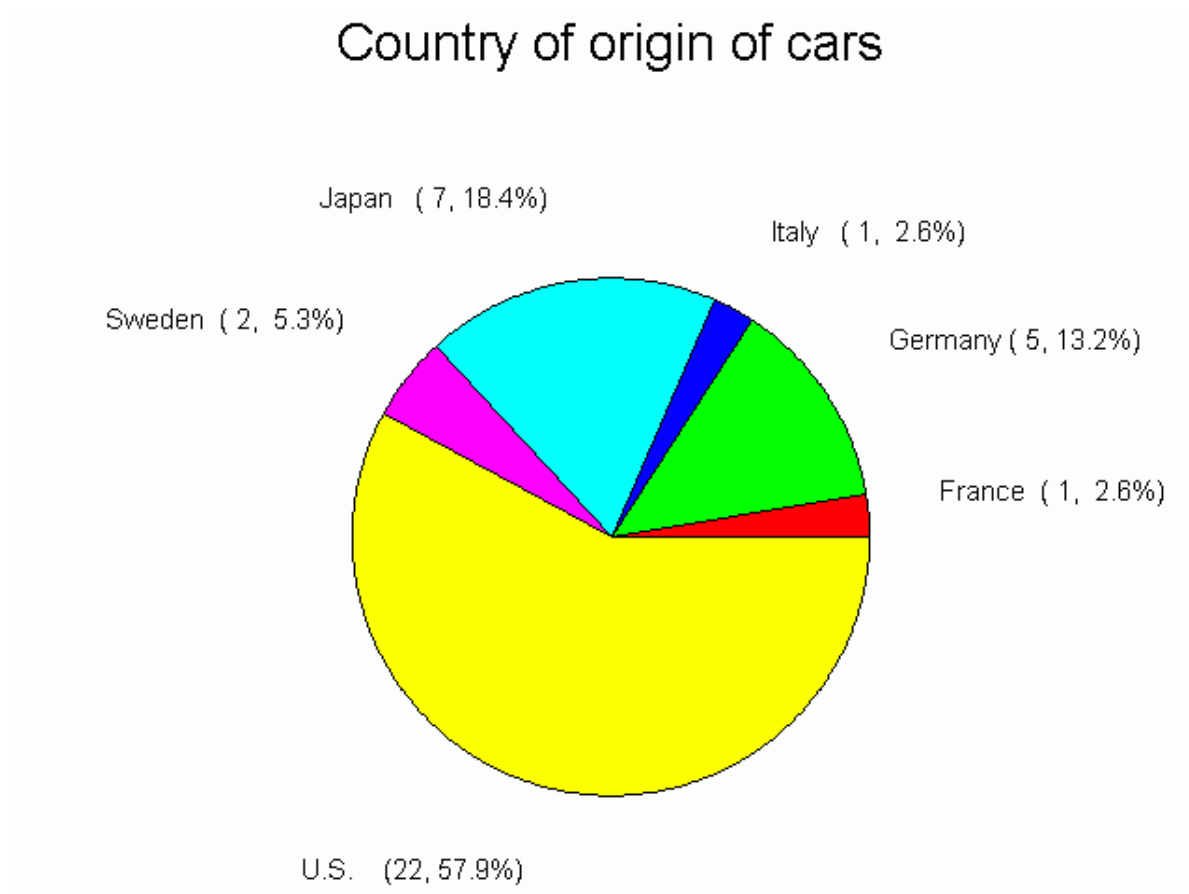
Shows cars from each country: most from US, fewest from France, Italy.

Very difficult to see from original list of numbers.

(In general: how many individuals in each category.)

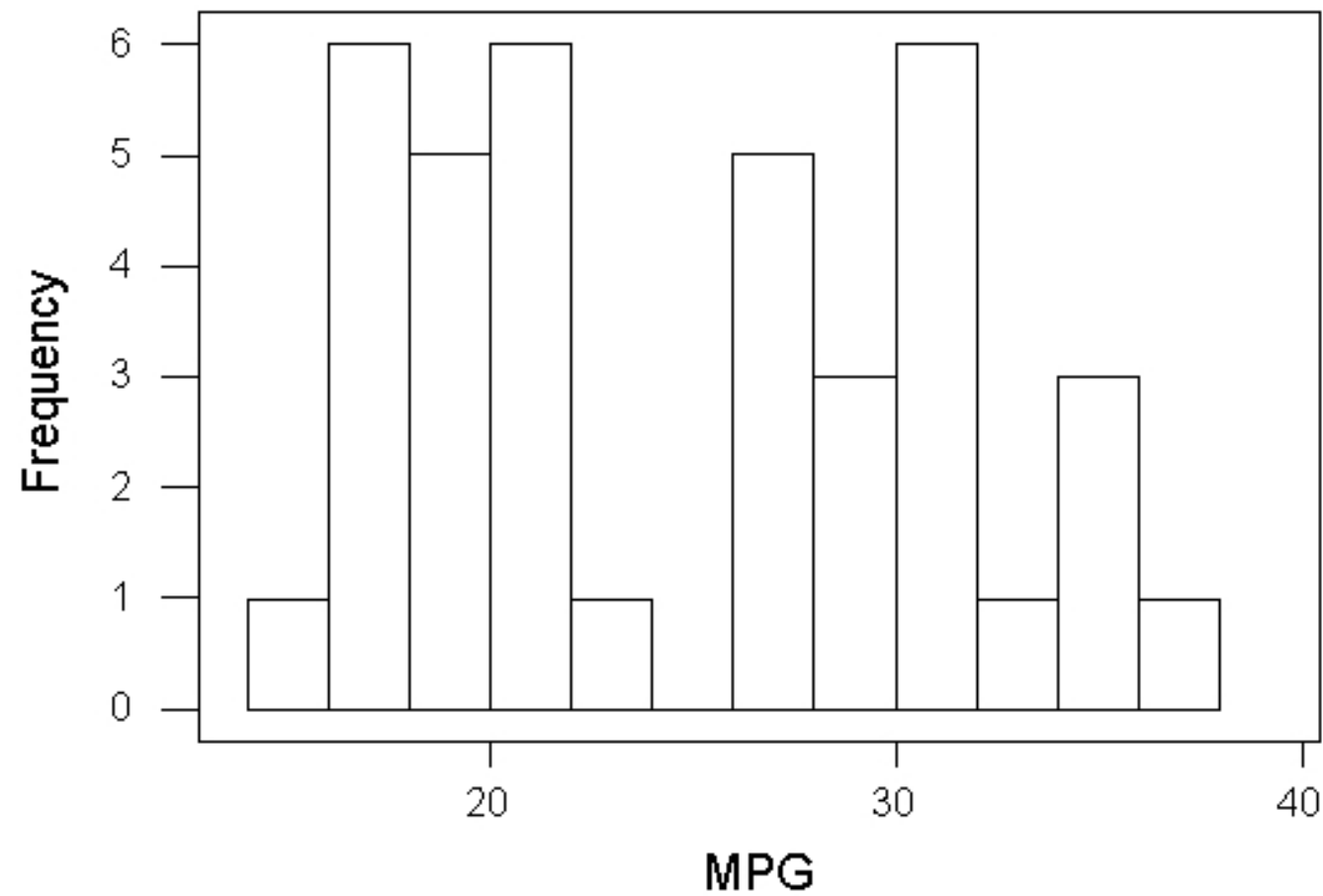
Pie chart

for data that are parts of a whole. Also true for country of cars:



For *quantitative* variables, can use **histogram**.

Example: miles per gallon.



Cars divide into 2 groups: those with MPG around 20, those with MPG around 30.

Again, hard to see without picture.

Alternative to histogram: **stemplot**. MPGs, rounded:

28	31	21	37	16	32	34	34
22	32	29	27	19	27	18	21
22	27	16	22	34	30	17	17
17	20	19	19	31	27	35	
22	32	18	17	18	28	30	

Make list of tens digits, write units on correct row (first 3 cars)

1			
2		8	1
3		1	

Continue until all cars done:

1		6	9	8	6	7	7	7	9	9	8	7	8			
2		8	1	2	7	9	7	7	1	2	7	2	0	7	2	8
3		1	7	2	4	4	2	4	0	1	5	2	0			

Optional last step: sort the leaves, lowest to highest, on each line:

1		6	6	7	7	7	7	8	8	8	9	9	9			
2		0	1	1	2	2	2	2	7	7	7	7	7	8	8	9
3		0	0	1	1	2	2	2	4	4	4	5	7			

Advantages: easy to do by hand; easy to see lowest (16) and highest (37) values

Disadvantage (here): hard to see shape, compared to histogram: too few “stems”.

Splitting stems

To get more lines on a stemplot (and better picture of shape of distribution), split each line into two parts: “leaves” 0–4 and leaves 5–9. For MPGs this gives:

```
1 | 6 6 7 7 7 7 8 8 8 9 9 9
2 | 0 1 1 2 2 2 2
2 | 7 7 7 7 8 8 9
3 | 0 0 1 1 2 2 2 4 4 4
3 | 5 7
```

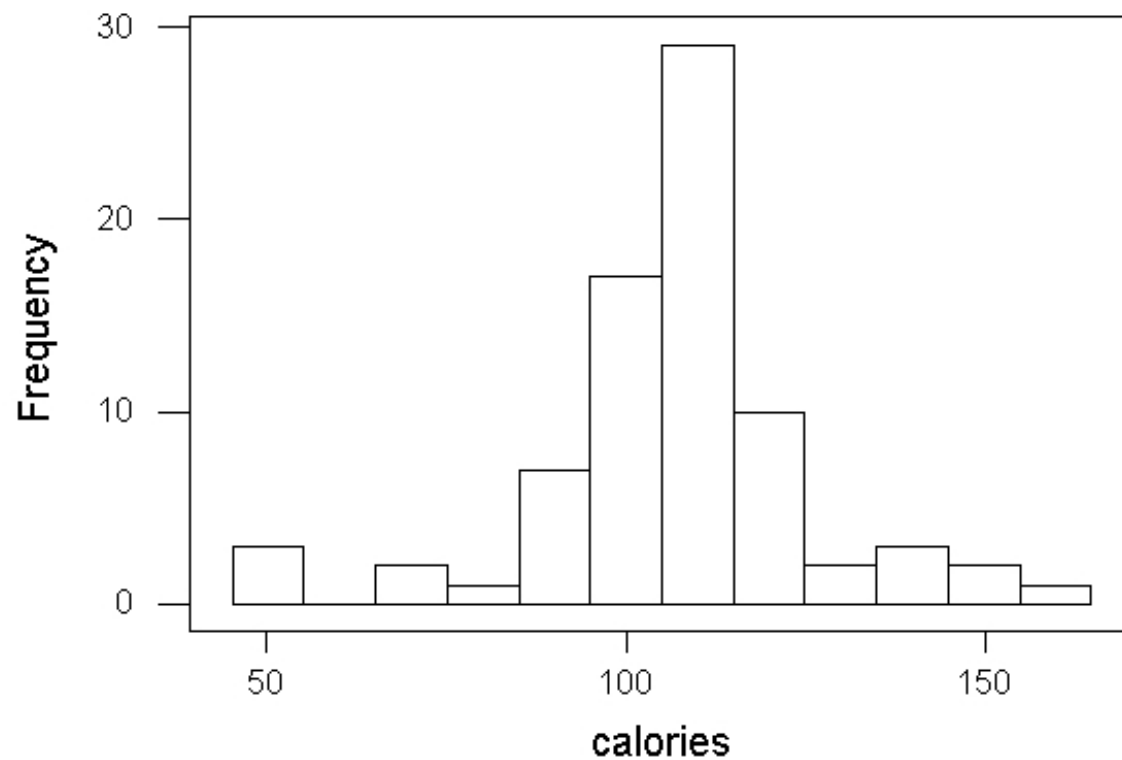
Like the histogram, shows many cars with MPGs in high teens, and in low 30s. Look carefully also to see “gap” between 22 and 27.

Drawing stemplots

- Separate each value into *stem* (all but rightmost digit) and *leaf* (rightmost digit).
 - If too many stems, *trim* one or more digits from right of values.
 - If too few stems, *split* in half.
 - Idea: want to show *shape* of distribution.
- Write stems in column, smallest at top, vertical line to right.
- Write each leaf in row to right of its stem.
- (optional) arrange leaves in increasing order.

Shape, centre and spread

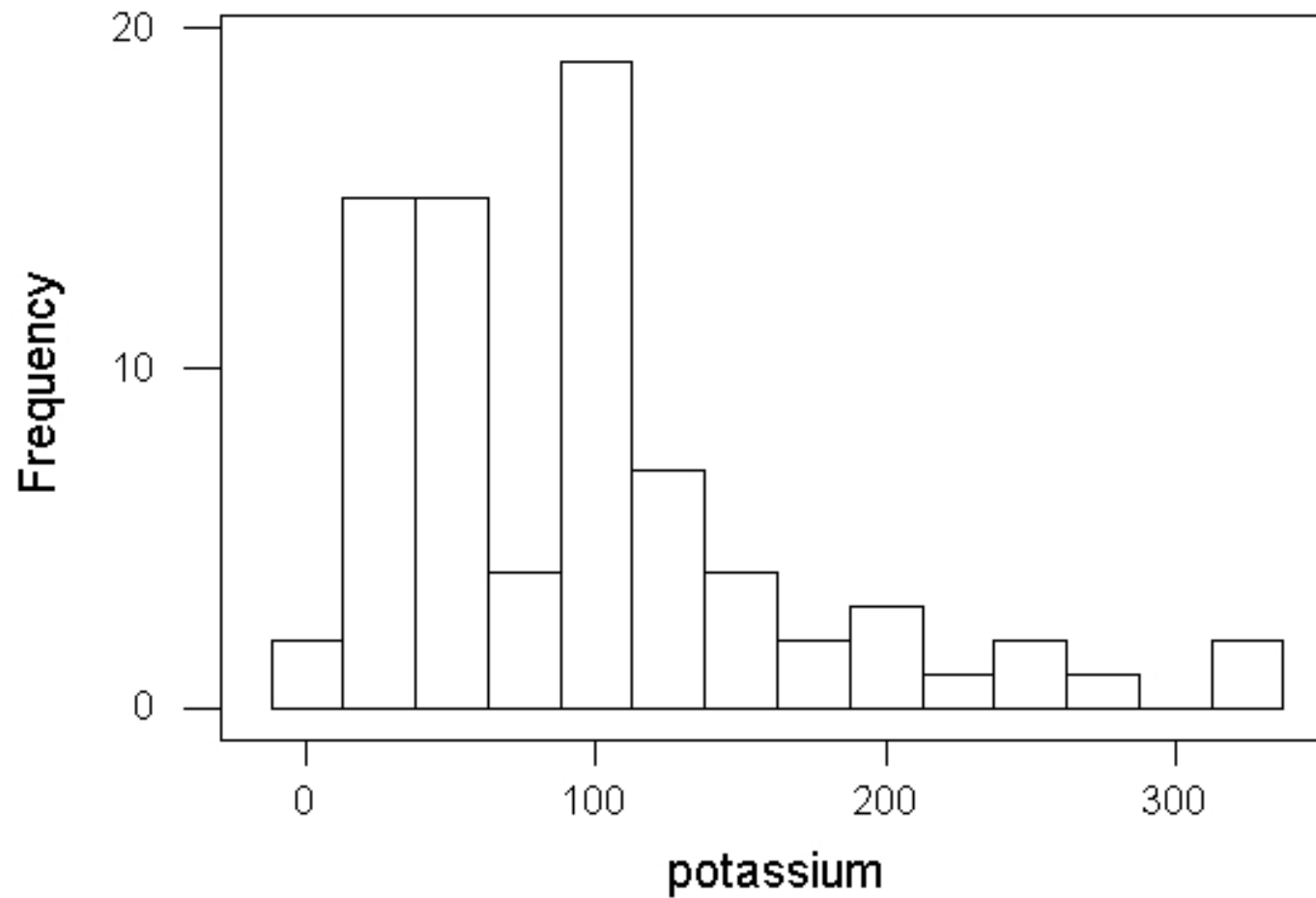
Different data: information on 77 kinds of cereal (individuals), lots of variables. Histogram of calorie content per serving:



Most cereals have around 110 calories/serving. Only a very few cereals have a lot more or fewer calories.

That is, **centre** of distribution of values around 110, **spread** fairly small. Also, **shape** fairly symmetric: picture goes “up” same way as “down”.

Compare histogram of potassium per serving:



Most cereals have 100 mg or less of potassium (difficult to say where exact centre is), but there is a lot of spread – from no potassium to over 300 mg.

Shape not symmetric: more cereals with unusually high potassium than with unusually low. Up quickly, down slowly. Called **skewed to right**.

Shape, centre and spread with stemplots

Stemplots show same kind of thing as histograms, so interpreted in same way. Stemplot of calorie content:

Stem-and-leaf of calories N = 77
Leaf Unit = 1.0

```

3      5 000
3      6
5      7 00
6      8 0
13     9 0000000
30    10 000000000000000000
(29)  11 00000000000000000000000000000000
18    12 00000000000
8      13 00
6      14 000
3      15 00
1      16 0

```

(calorie content rounded to nearest 10 in data).

Stemplot of potassium:

Stem-and-leaf of potassiu N = 77

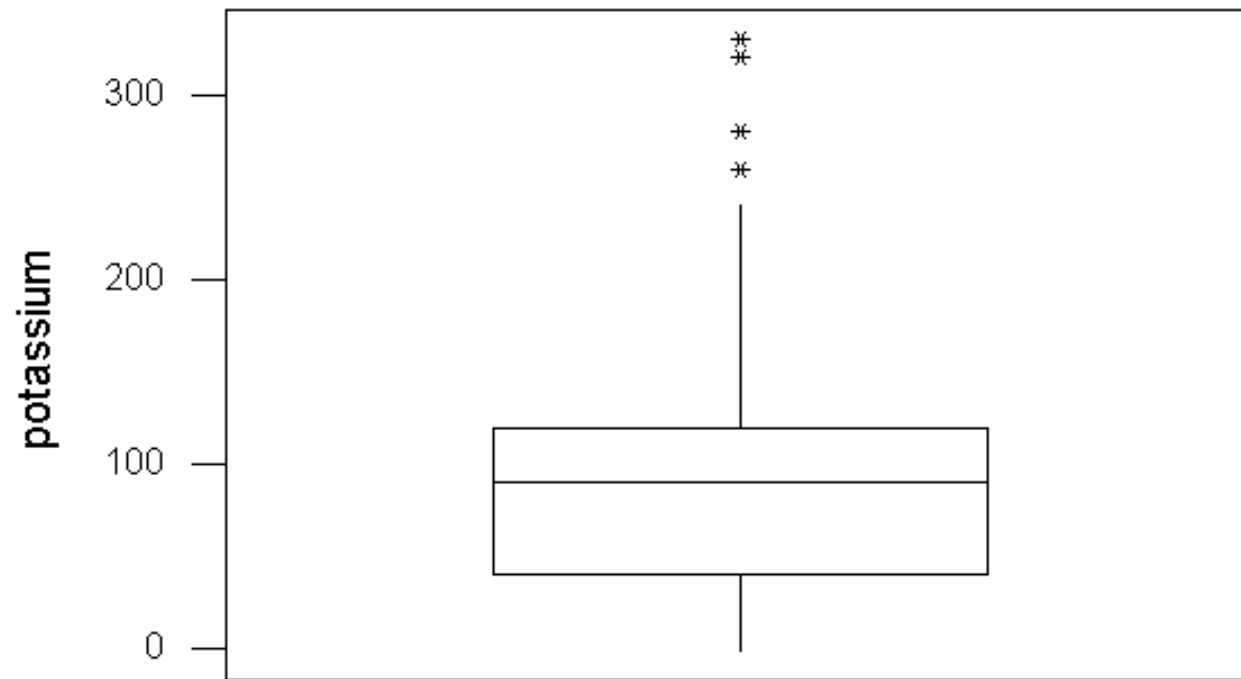
Leaf Unit = 10

3	0	001
17	0	22222333333333
29	0	444444445555
34	0	66667
(11)	0	889999999999
32	1	00000111111
21	1	222233
15	1	44
13	1	6677
9	1	99
7	2	0
6	2	3
5	2	4
4	2	6
3	2	8
2	3	
2	3	23
0	3	

(note splitting of stems, and long straggle towards bottom.)

Boxplots (from 1.2)

Boxplots designed to show shape, centre, spread (when “centre” makes sense). For potassium per serving:



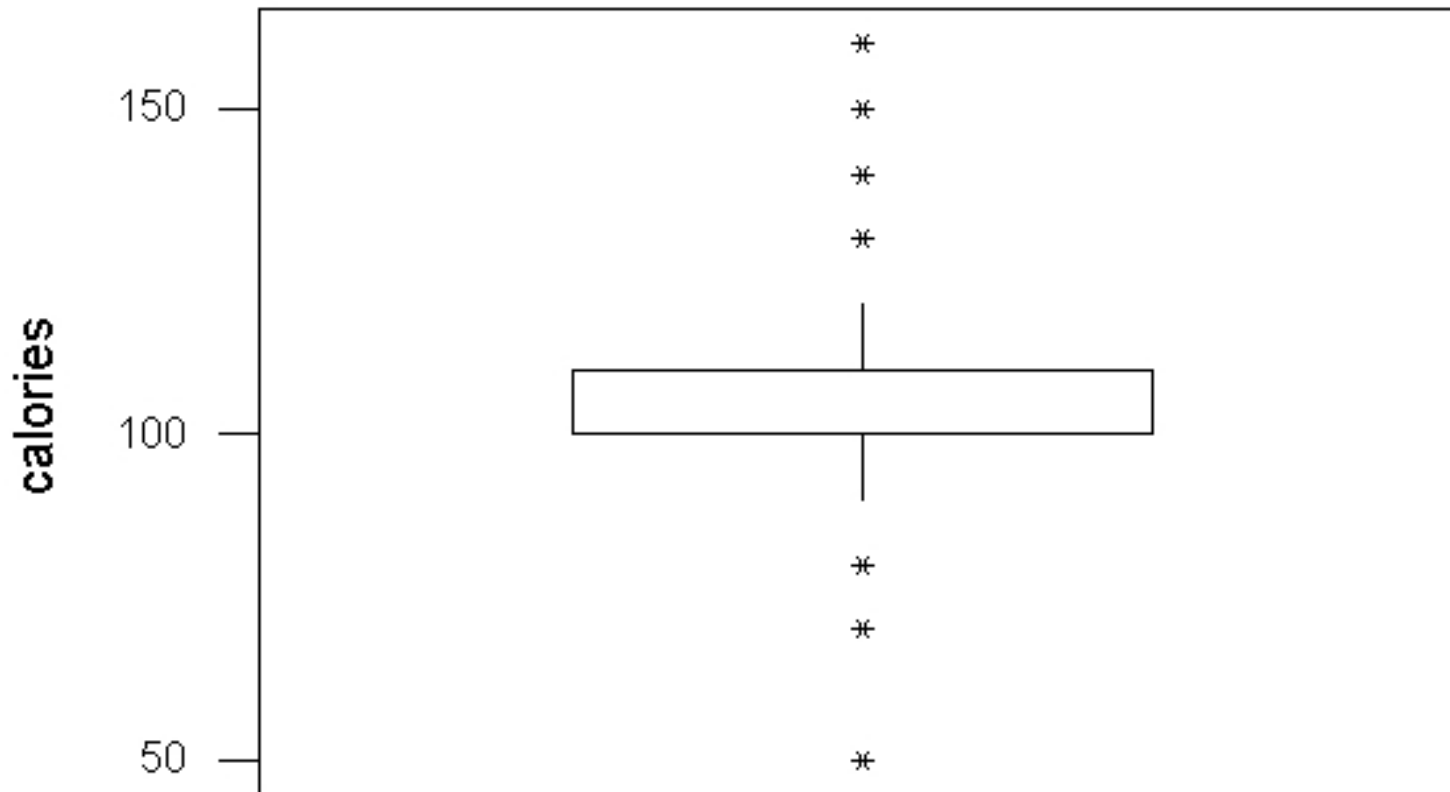
Parts of boxplot:

- rectangle with line across it. Line marks “centre”, top/bottom of rectangle show middle 50% of data. Typical potassium level just under 100, middle 50% between 40 and 120.
- vertical lines (“whiskers”) above/below rectangle show extent of “plausible” values (0 to 240).
- * marks individual “unusual” values (4, above 260).

Heights of rectangle and vertical lines show spread.

Comparative lengths of whiskers above & below show symmetry/skewness (long upper whisker – skewed to right).

Compare boxplot for calorie content:



Centre around 100, compressed box & whiskers show little spread.

Whiskers same length – symmetric shape. Many unusual values.

Describing distributions with numbers

(§1.2)

Picture often says much about distribution, but to convince others, may need *numbers* to describe.

Three major features of distribution: *shape, centre, spread*. Assess shape by picture; use numbers for centre and spread.

Centre: the mean

Centre says what “typical”, “average” value is.

Mean: add up values, divide by how many there are.

Example: running back on football team carries ball 5, 3, 11, 1 and 4 yards. Mean distance carried is

$$\frac{5 + 3 + 11 + 1 + 4}{5} = 4.8 \text{ yards.}$$

Cereal calorie content: from histogram, centre around 110. The mean for all 77 cereals is 106.88.

Centre: the median

Median: arrange values in order, pick out middle one.

Football running back: in order, 1, 3, 4, 5, 11. Median is 4 yards (compare mean 4.8 yards).

Cereal calorie content: median is 110 calories (compare mean 106.88).

If even number of data values, no single middle one. Use mean of middle 2. Example: median of 10, 11, 14, 20 is $(11 + 14)/2 = 12.5$.

Mean vs. median

Both measuring centre: why have both?

When distribution shape *symmetric*, mean and median close together. But if distribution *skewed*, or if unusually high or low values, can be quite different.

Running back: mean (4.8) bigger than median (4) because of one unusually long run (11 yards). Affects mean (adds to total) but only affects median by being bigger than 4.

So use *mean* if shape close to symmetric, *median* if shape skewed.

Spread: quartiles and interquartile range

Median splits data values in half (half above median, half below).

Quartiles: split data values $\frac{1}{4}$ — $\frac{3}{4}$. (Quarter below or quarter above). Calculation details in text section 1.2.

Cereal potassium content: 1st quartile is 40 (quarter of cereals have less potassium than this, three-quarters more). 3rd quartile is 120 (three-quarters less, one quarter more).

Interquartile range (IQR): difference between 3rd quartile and 1st quartile.

For potassium content, is $120 - 40 = 80$.

Larger IQR means data more spread out. (If all values same, IQR would be 0.)

Compare calorie content of cereals. 1st quartile 100, 3rd quartile 110, IQR $110 - 100 = 10$.

IQR for calorie content much smaller than for potassium, because values much less spread out.

Boxplot revisited: outliers

IQR measures spread of middle 50% of data. Also yardstick to decide whether value unusual compared to rest: an **outlier**.

Criterion: *more than $1.5 \times IQR$ above 3rd quartile, or more than $1.5 \times IQR$ below 1st quartile* considered outlier.

Boxplot (eg. for potassium): centre line at *median*, top and bottom of box at 1st and 3rd quartiles. Whiskers from box to most distant non-outlier; outliers plotted individually.

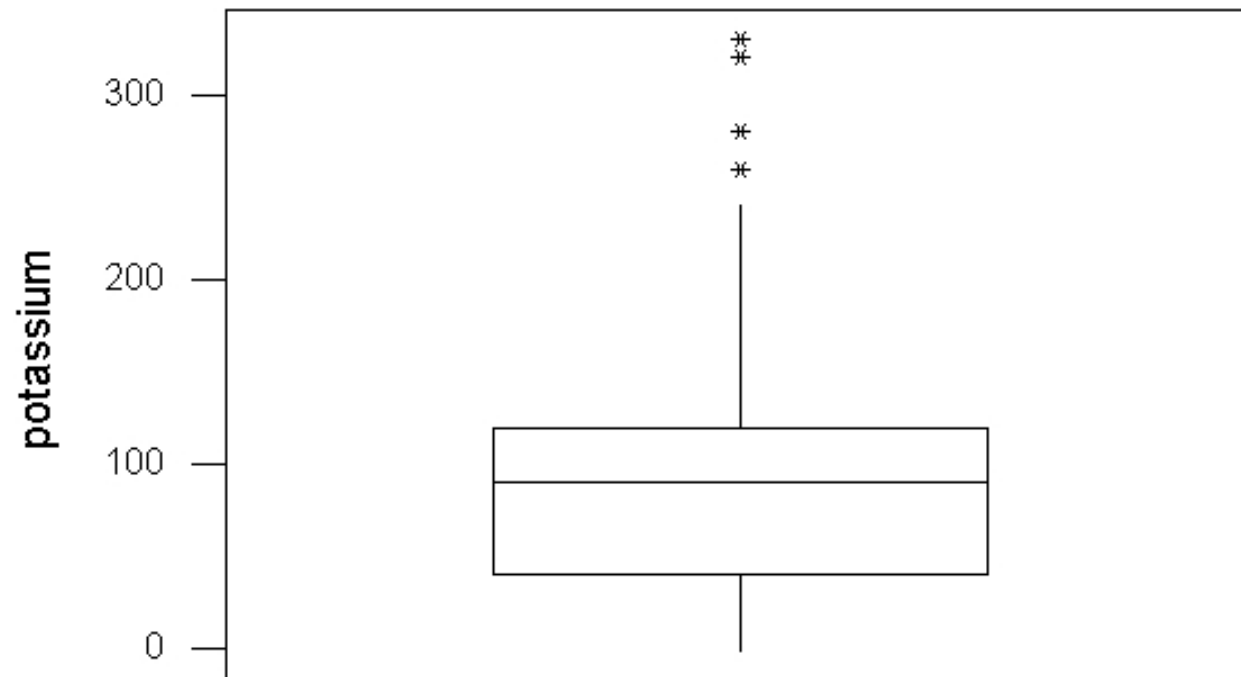
Check calculation for potassium: median 90, 1st quartile (Q1) 40, 3rd quartile (Q3) 120.

$$IQR = 120 - 40 = 80; \quad 1.5 \times IQR = (1.5)(80) = 120.$$

Therefore anything below $40 - 120 = -80$ is an outlier, also anything above $120 + 120 = 240$.

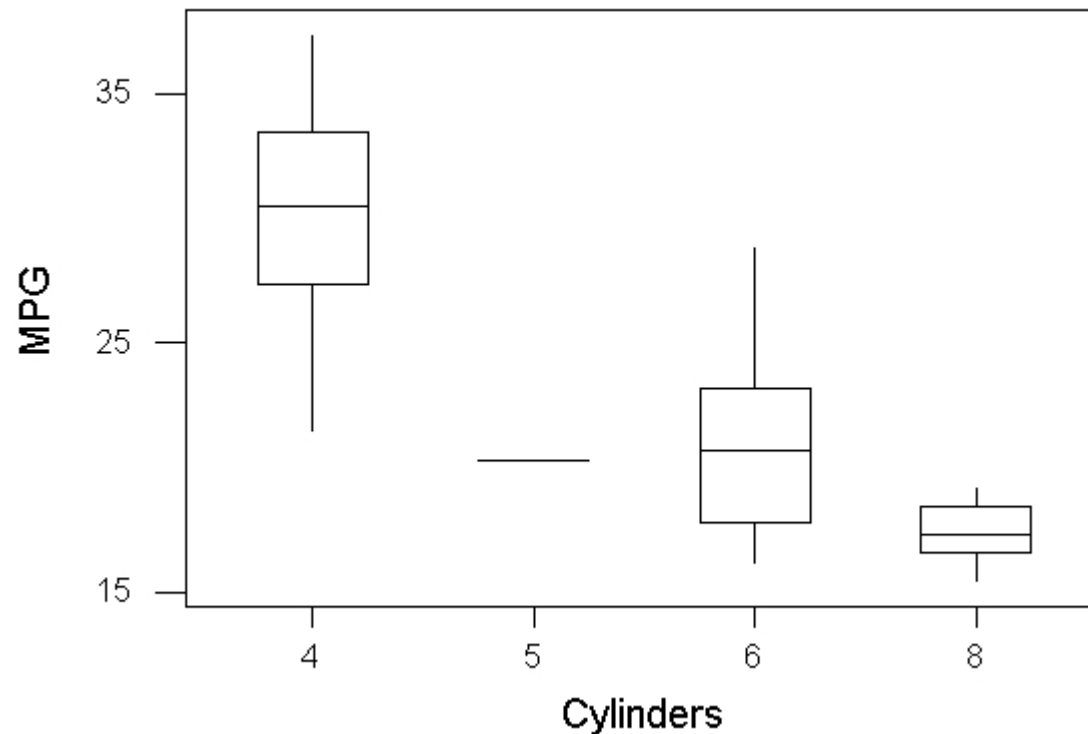
In data: smallest value 0 (no outliers at low end). 4 values above 240, one exactly equal. 240 is not outlier; is highest non-outlier, so extend whisker to 240, plot higher values individually.

Review boxplot:



Boxplots useful for comparison – plot side by side.

Cars: More cylinders means worse gas mileage? Boxplot for cars with 4, 6, ... cylinders:



Yes! 8-cylinder cars less spread (all bad!).

Standard deviation

Median, IQR use only middle of distribution. Mean uses all values; spread that does same?

Standard deviation (SD) based on how far away from mean each value is. If many values far from mean (or some very far), SD large. If all values close, SD small.

Value minus mean $+$ if value above mean, $-$ if value below. So calculation squares each value first (makes $+$), then takes square root at end. (Details in section 1.2 of text.)

Examples from cereal data:

Variable	IQR	SD
Calories	10	19.48
Potassium	80	71.29

IQR, SD numbers not directly comparable, but pattern same:
calories less spread-out than potassium content.

SD vs. IQR

SD, like mean, uses all data values. So can be badly affected by outliers or skewed shape.

So use SD when would use mean: shape symmetric, no outliers.
Otherwise use median and IQR.

Changing units of measurement

Could measure eg. heights in feet or metres, temperatures in Celsius or Fahrenheit. Conclusions shouldn't be affected by choice.

Common kinds of change:

- multiplying each value by the same number b
- adding the same number a to each value

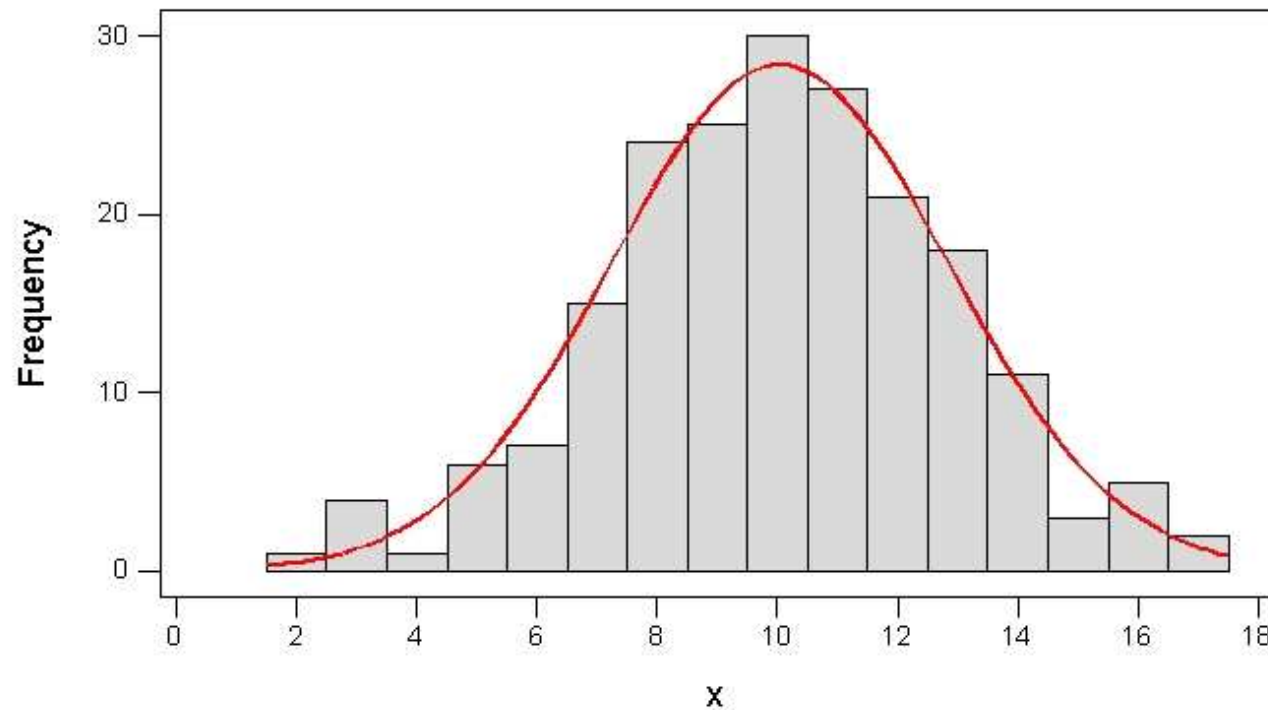
Changes numbers on plots, but *no effect* on shape.

Effects of these changes on quantitative descriptions predictable:

- multiplying by b multiplies mean, median, SD and IQR by b
- adding a adds a to mean and median, but leaves SD and IQR unchanged.

Makes sense: adding a number changes centre, but “shifts distribution along” without changing its spread.

The normal distributions (§1.3)



Histogram has fairly symmetric shape. Smooth curve superimposed passes (more or less) through top of histogram bars.

Curve called a **normal density function**. Might guess that histogram looks jagged because of randomness of data; “underlying” picture is smooth.

Normal density function is **mathematical model** of process producing data.

If histogram with bars matching normal density function, data said to have **normal distribution**.

Normal distribution described completely by *mean* and *standard deviation*.

Notation with Greek letters: mean μ (mu), SD σ (sigma).

Picture: $\mu = 10, \sigma = 3$.

Mean at peak of curve (obvious “centre”).

σ distance from peak to “shoulder” (same both sides).

If σ smaller, curve is more peaked; if σ larger, curve lower and flatter.

Why normal distribution is useful

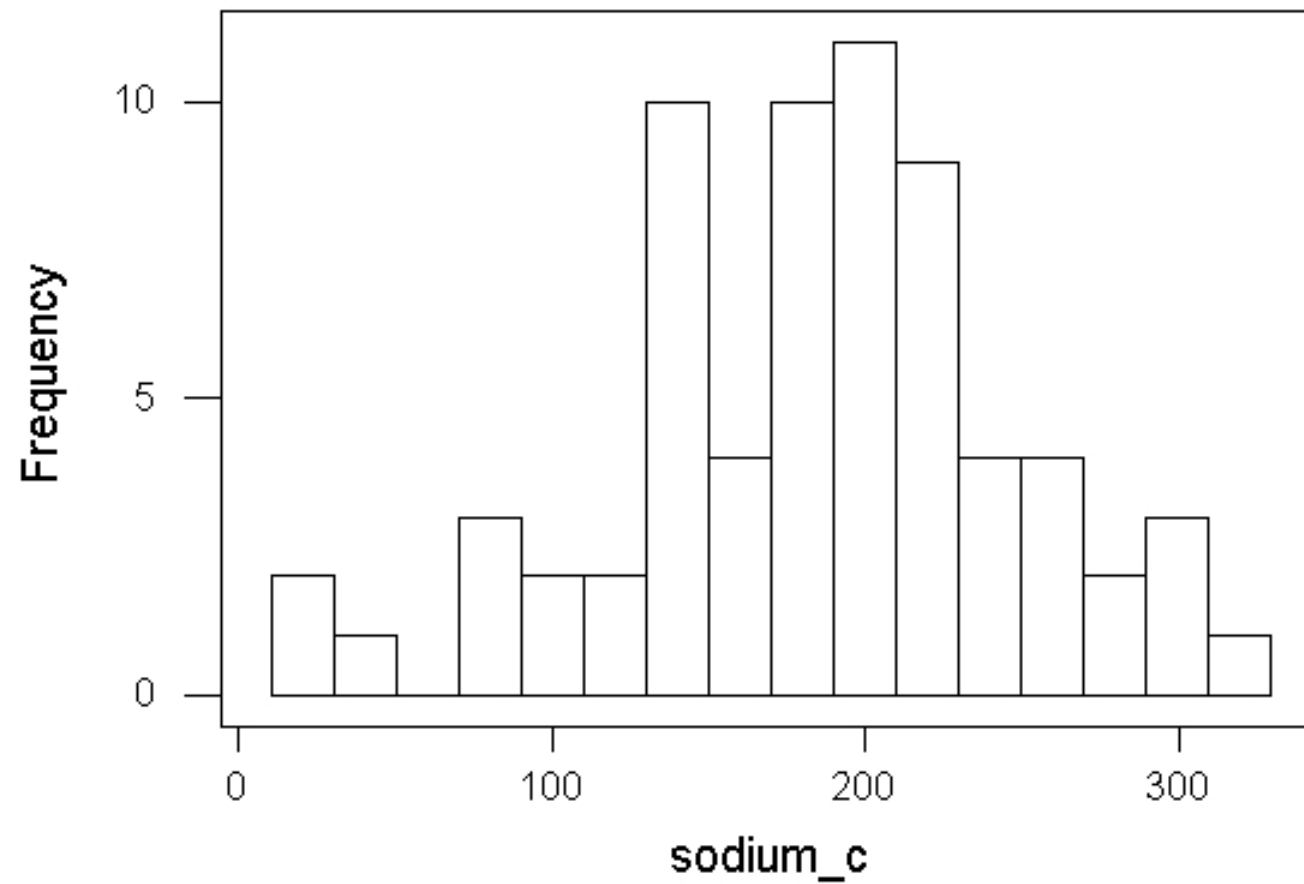
- distribution of *real data* sometimes
- distribution of *chance outcomes*
- central to *statistical inference* (drawing conclusions from data)

68–95–99.7

On *any* normal distribution (whatever mean and SD):

- 68% of values fall within σ of the mean
- 95% of values fall within 2σ of the mean
- 99.7% of values fall within 3σ of the mean

Cereal data: sodium content levels look normal-distribution-ish:



Try rule. Mean is 180.8, SD 64.0, 68 values.

Low	High	Count	%	Rule
116.8	244.8	50	74%	68%
52.8	308.8	64	94%	95%
-11.2	372.8	68	100%	99.7%

Pretty close!

Standardizing and z -scores

Can find out how unusual a value from a normal distribution is.

For example, Cheerios have 290 mg sodium/serving. Subtract mean, divide SD to get

$$z = \frac{290 - 180.8}{64.0} = 1.71.$$

Higher than average sodium; in fact, notably higher relative to SD.

Process called **standardizing**; result called z -**score**.

(Fact: if normal distribution correct, only 4% of cereals higher in sodium.)

Normal distribution calculations

Can be done with Minitab (see later) or Table A in text.

Example 1.28 in text: SAT scores have normal distribution, mean 1026, SD 209. What proportion of students have scores 820 or bigger?

Steps: value to z -score to proportion.

Value here 820. Then

$$z = \frac{820 - 1026}{209} = -0.99.$$

Look up -0.99 in Table A: 0.1611. *Table gives you “less”*, so proportion 0.1611 of students have SAT scores less than 820, and proportion $1 - 0.1611 = 0.8389$ have scores greater than 820.

Proportion between

To get the proportion between two values, turn *both* values into z -scores, look them *both* up in Table A, then take difference.

In previous example, proportion of students scoring between 720 and 820?

820 corresponds to $z = -0.99$ and proportion 0.1611. For 720,

$$z = \frac{720 - 1026}{209} = -1.46.$$

In Table A, $z = -1.46$ goes with proportion 0.0721.

So proportion of students scoring between 720 and 820 is $0.1611 - 0.0721 = 0.0890$.

Getting value from proportion

If you have a proportion and you want to get a value, reverse above procedure.

What SAT score do 10% (0.1000) of students score less than?

Use Table A backwards first: look up 0.1000 in *body* of table to get $z = -1.28$ (approx).

Then “unstandardize” to get SAT score of $1026 + (-1.28)(209) = 758$ (rounding).

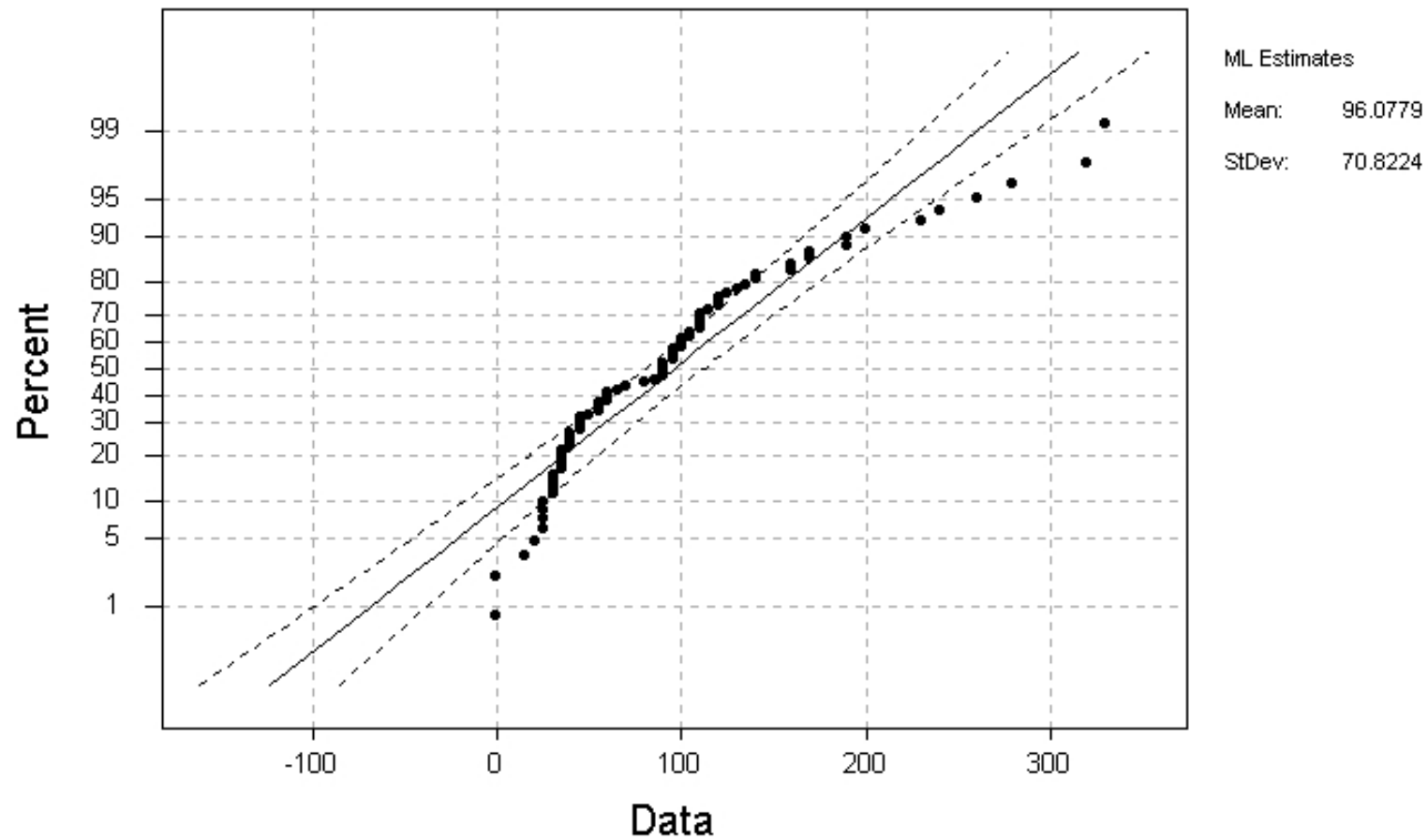
10% of students will score less than 758, and the other 90% will score more.

Normal quantile plot

To do any calculations with a normal distribution, need to know that shape is correct.

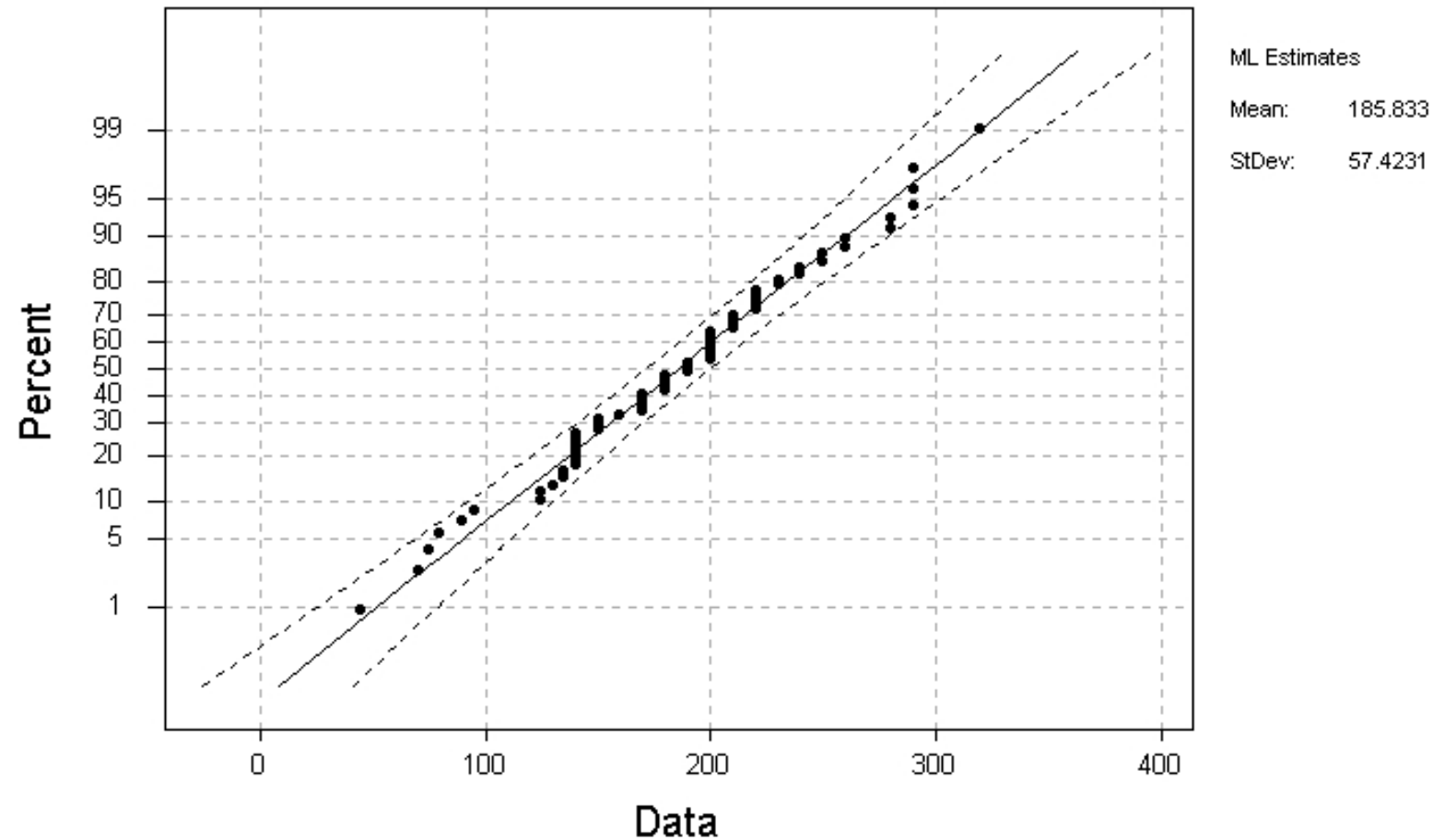
Example: cereal data, potassium content values had skewed distribution (so normal no good).

Normal quantile plot should show straight line:



The curve indicates a skewed distribution.

Compare sodium content values.



The pattern of dots wiggles, but follows the central line well.

Introduction to Minitab

Introduction

Minitab is software designed especially for doing Statistics. Comes bundled with textbook (with manual), also can be accessed in Windows labs on campus.

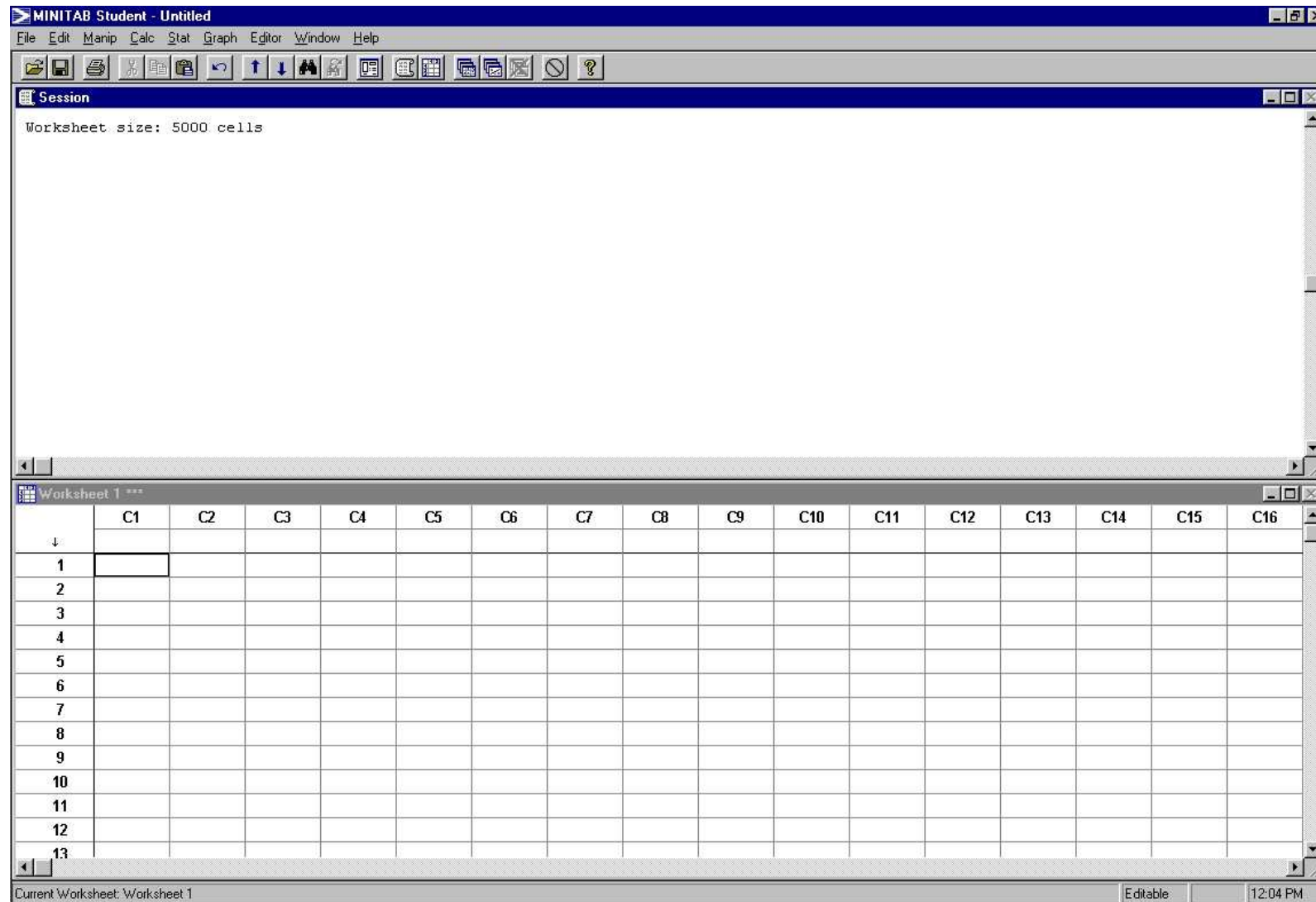
We use Minitab in this course because:

- it is not difficult to learn (in our opinion)
- it has been well tested over the years of its existence (you can trust its calculations)
- it does all the statistical analyses you are likely to need.

Spreadsheet software (like Excel) is not to be trusted for Statistics!

Startup: Start button, Programs, Minitab.

Startup



Bottom window for data (variables in columns and individuals in rows). Each variable can have a name (use boxes directly below C1, C2), above 1st row.

Results of calculations appear in the top window (called the Session window).

The menus at the top let you reorganize your data set (Manip), do calculations with it (Calc and Stat) and draw pictures (Graph).

Getting data in

- For textbook data sets: find on disk (in PCDataSets, MiniTab), try double-clicking or right-click and Open With Minitab.
- Windows copy and paste. Works from text file (eg. web page) or spreadsheet.

If variables have names, put the cursor in the cell above row 1.
Else put it in row 1, col 1.

- Type data in yourself. Works as on a spreadsheet. Enter moves cursor down to the next cell, or click cell to use next.
- Other: Minitab Manual I.7.

Saving, opening and creating worksheets

When you have your data set in Minitab, it's good to **save** it to re-use later. Select File, Save Current Worksheet As; give it a name, and save it to a suitable place (eg. your floppy disk).

If you make more changes, **save** them using Save Current Worksheet (over-writes previous version).

To **open** a previously-saved worksheet, use File, Open Worksheet. Or, to start Minitab with a worksheet opened, use Windows Explorer to find the worksheet, then double-click it.

Minitab allows multiple worksheets. To create a **new** one, select File, New, Minitab Worksheet. Click on the title bar of a worksheet to make it “active” (the sheet with the data for your work).

Graphs

Minitab does barplots, histograms, boxplots and quantile plots. In Graph menu; fill in boxes.

Cars worksheet: **bar plot** for country — Graph, Chart. Across to X column in top mini-sheet, click. Variables in left; double-click Country, click OK. Barplot pops up in own window.

Histogram for MPG: Graph, Histogram, double-click MPG in left, OK.

Boxplot for MPG: Graph, Boxplot, double-click MPG, click OK.

Boxplot for MPG by number of cylinders: as above, with MPG already selected. Click X column next to MPG, double-click Cylinders, click OK.

Normal quantile plot (for MPG): Graph, Probability Plot, double-click MPG. Minitab calculates some stuff then draws the plot.

Graphs can be **copied/pasted**. Inside graph window, right-click, select Copy Graph. Can then paste into eg. Word. Same with Session window.

Describing distributions with numbers

Easiest: Stat, Basic Statistics, Display Descriptive Statistics.

Double-click variable (MPG say), look at Session window. N is number of values. Get mean, median, SD (in 1st row), 1st quartile and 3rd quartile (in 2nd) so can calculate IQR.

Normal distributions

Can calculate what fraction of a normal distribution lies above or below a value, or between two values.

In cereal data set, sodium content was approx. normal with mean 180.8, SD 64.0. What fraction of cereals have less than 130 mg of sodium?

Calc, Probability Distributions, Normal. Select Cumulative Probability, fill in mean, SD. Click Input Constant, type in 130. Click OK.

In Session window, see that 0.2137 (21%) of cereals have less than 130 mg of sodium.

What fraction have more than 200 mg sodium? Repeat, change 130 to 200. Session window shows that 0.6179 of cereals have *less* than 200 mg sodium, so rest must have more: $1 - 0.6179 = 0.3821$.

What fraction between 160 and 240? Repeat above for 160 (get 0.3726), 240 (get 0.8225). 0.8225 of cereals less than 240, but 0.3726 less than 160 as well. In between:
 $0.8225 - 0.3726 = 0.4499$, about 45%.

Other way around: how much sodium do the top 5% of cereals have?

95% = 0.95 of cereals have *less* sodium than this value. Calc, Probability Distributions, Normal. Click Inverse Cumulative Probability, change Input Constant to 0.95. 95% of cereals have less than 286 mg sodium (Session window), so 5% have more.

Minitab command language

Minitab can be run using menus, but also by typing commands.

Menus are easier, but command language quicker when you know it.

To turn on, select Editor, Enable Command Language. See MTB > in Session window.

Now try (from menus) Stat, Basic Statistics, Display Descriptive Statistics. Above output see `Describe 'MPG' ..` This is command version of menu action. Try typing `describe 'mpg'`, Enter – get same output. (Uppercase/lowercase, final period not important.)

Chapter 2: Looking at Data – Relationships

Introduction

Returning to car data: cars usually cheap to run or powerful but not both. So might expect eg. heavier car to be more powerful, so less gas-efficient – variables “weight” and “MPG” **associated**.

New idea: so far, consider variables one at a time, but now have to consider 2 variables *together*.

Relationships between 2 variables might also be affected by other variables – potential for confusion.

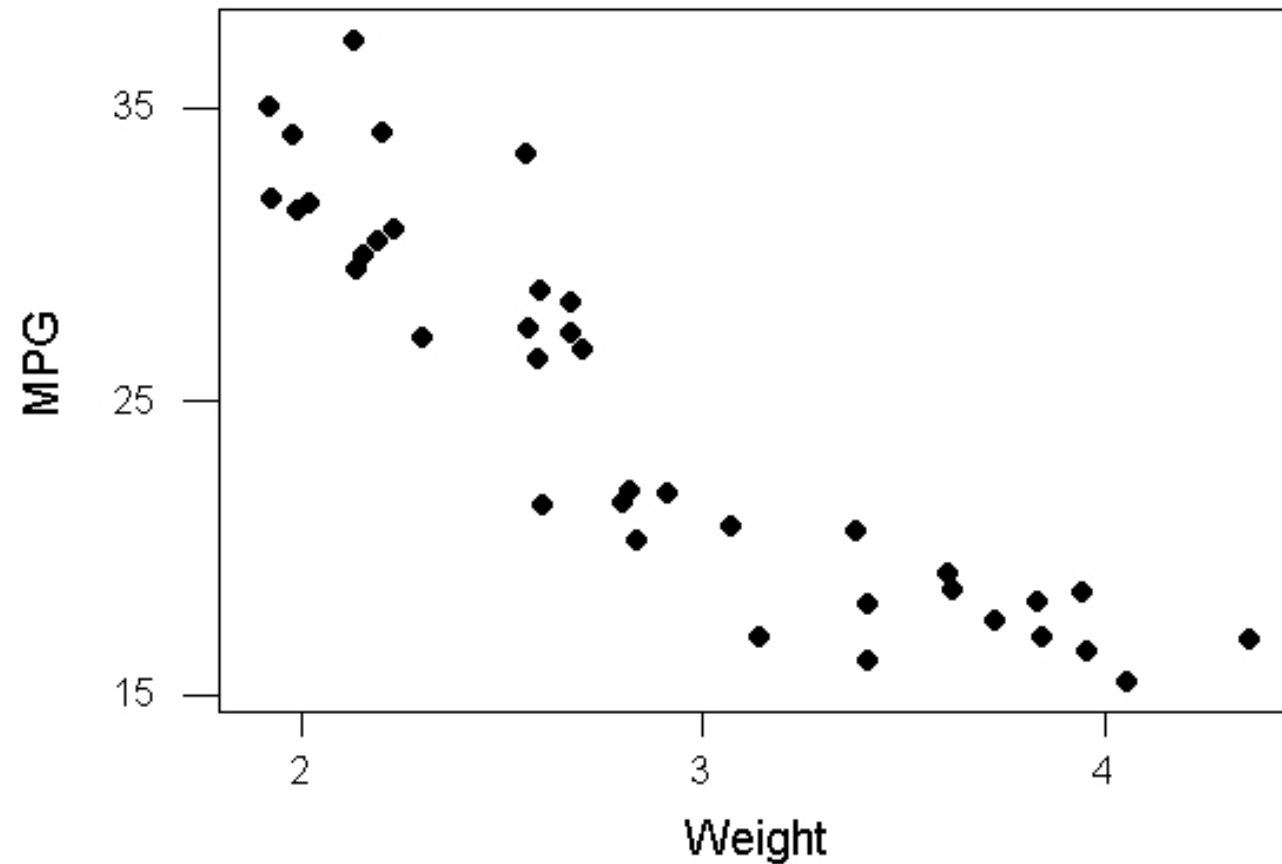
Start with quantitative variables.

Scatterplots (§2.1)

A good graphical display of quantitative variables is the **scatterplot**. For each individual, plot the values for the two variables on an x - y graph.

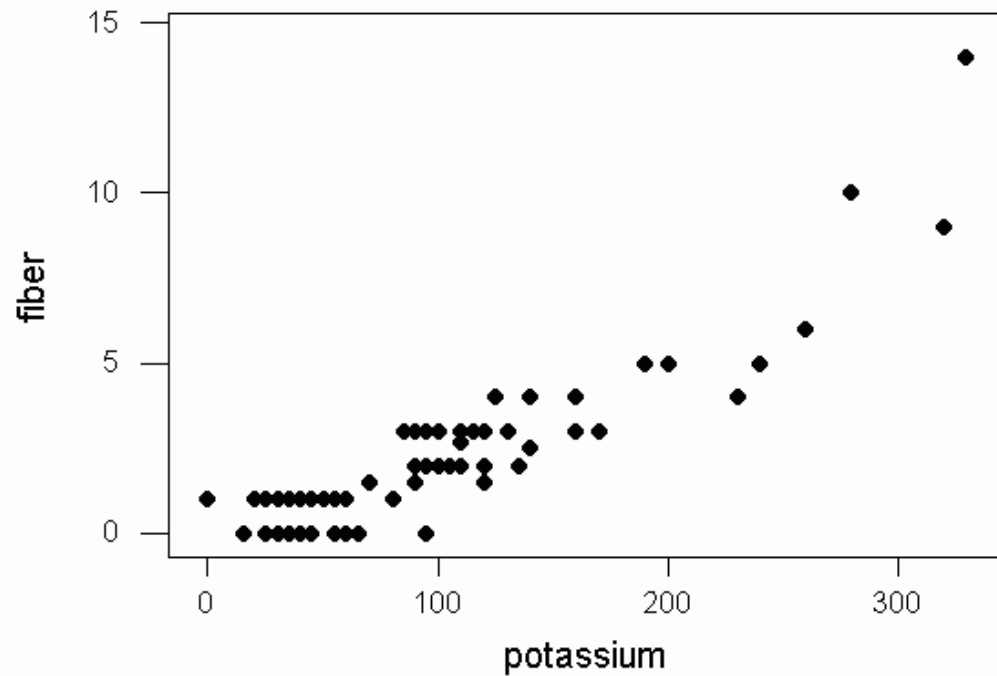
Car data: MPG and weight.

Get a scatterplot in Minitab from Graph, Plot then selecting MPG, weight by double-clicking.



Heavier cars generally have worse MPG (though exceptions).
Called a **negative association**.

Cereals: scatterplot for potassium vs. fibre:



Cereals with more potassium generally have more fibre. **Positive association.**

Other points

- Statistical associations are *general tendencies*, not ironclad rules. Almost always exceptions to trend. (What works “on average”.)
- Many studies have one variable that is *outcome* (final result). Eg. MPG is result of design, construction and driving of car. Called **response variable**. Other variables called **explanatory variables**: “explain” how response changes.
- In a scatterplot, put response variable on vertical (y) axis.

Summary

- With two related variables, need to look at both at once.
- Scatterplot is good display for quantitative variables.
- In scatterplot, can have:
 - positive association (big on one variable means big on other)
 - negative association (big on one is small on other)
 - curved association
 - no apparent association
- Statistical associations usually general trend with individual exceptions.

Correlation (§2.2)

First step in assessing 2-variable relationship is scatterplot. Can see if *is* association, and what kind (straight line, curve).

If association looks like straight line, describe strength of relationship using number called **correlation**.

Properties of correlation

- Number between -1 and 1 . 1 means perfectly straight *upward* trend; -1 means perfectly straight *downward* trend, 0 means no trend at all.
- Measures how “predictable” one variable is if you know other.
- Only works for straight-line associations (misleading otherwise).
- Has no units of measurement (pure number), so same for any measurement units of variables (km per litre instead of MPG).
- Based on mean and SD, so can be badly affected by outliers.

Examples of correlation

In Minitab, Stat, Basic Statistics, Correlation. Select 2 (or more) variables, click OK.

For fibre and potassium in cereals:

Correlation of fiber and potassium = 0.903, P-Value = 0.000

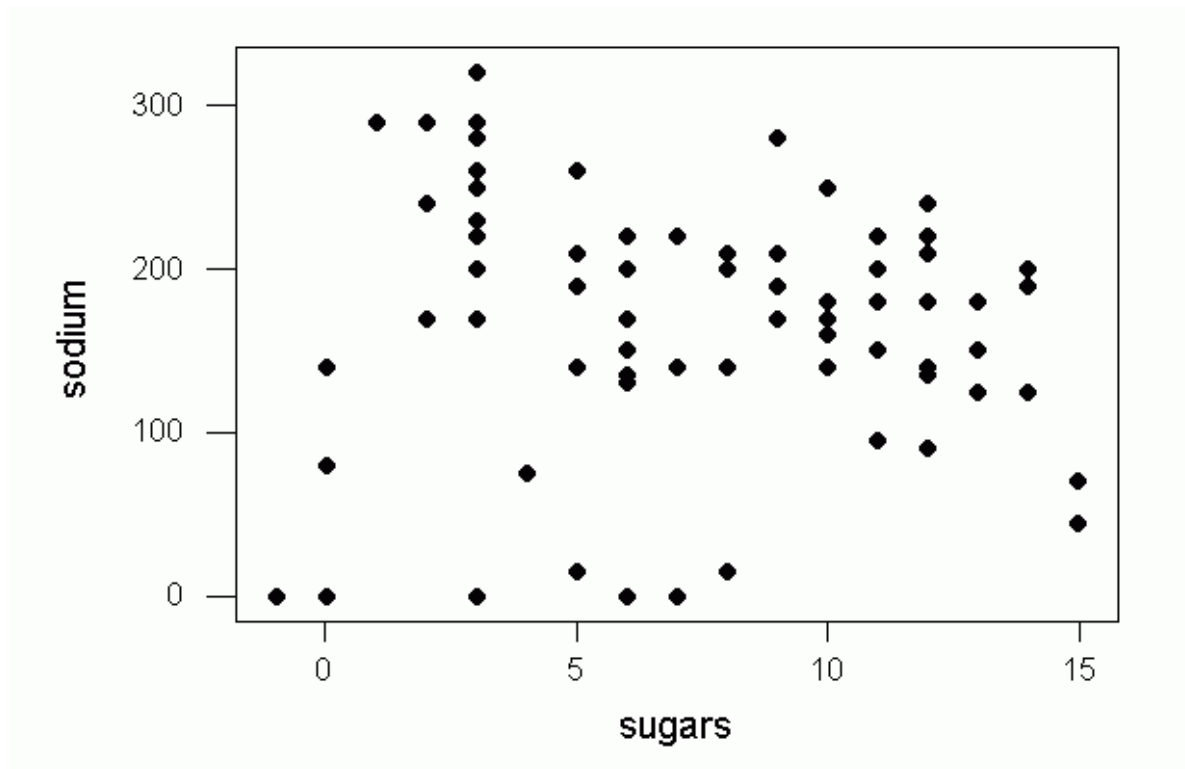
Correlation high and positive. Fibre predictable from potassium:
high potassium means high fibre. (Compare scatterplot.)

Cars: high weight meant low MPG. Correlation shows this too:

Correlation of Weight and MPG = -0.903 , P-Value = 0.000

Identical numerical values of 0.903 coincidence!

Cereals: Sodium and sugars scatterplot and correlation:



Correlation of sodium and sugars = 0.101, P-Value = 0.380

Almost no association here at all.

Summary

- Correlation is number summarizing extent and kind of relationship.
- Correlation $+1$ shows perfect positive association.
- Correlation -1 shows perfect negative association.
- Correlation 0 shows no (straight-line) association.
- Doesn't depend on units of measurement.
- Only works for straight-line relationships (not curves).

Regression (§2.3)

Correlation: “*is there* a straight-line association?”

Regression: “*what is* that straight line?”

When line known, can use it to predict value of response variable from a value of explanatory variable – given a car’s weight, can predict its MPG.

Straight lines in mathematics

Mathematical equation for straight line association between x and y is:

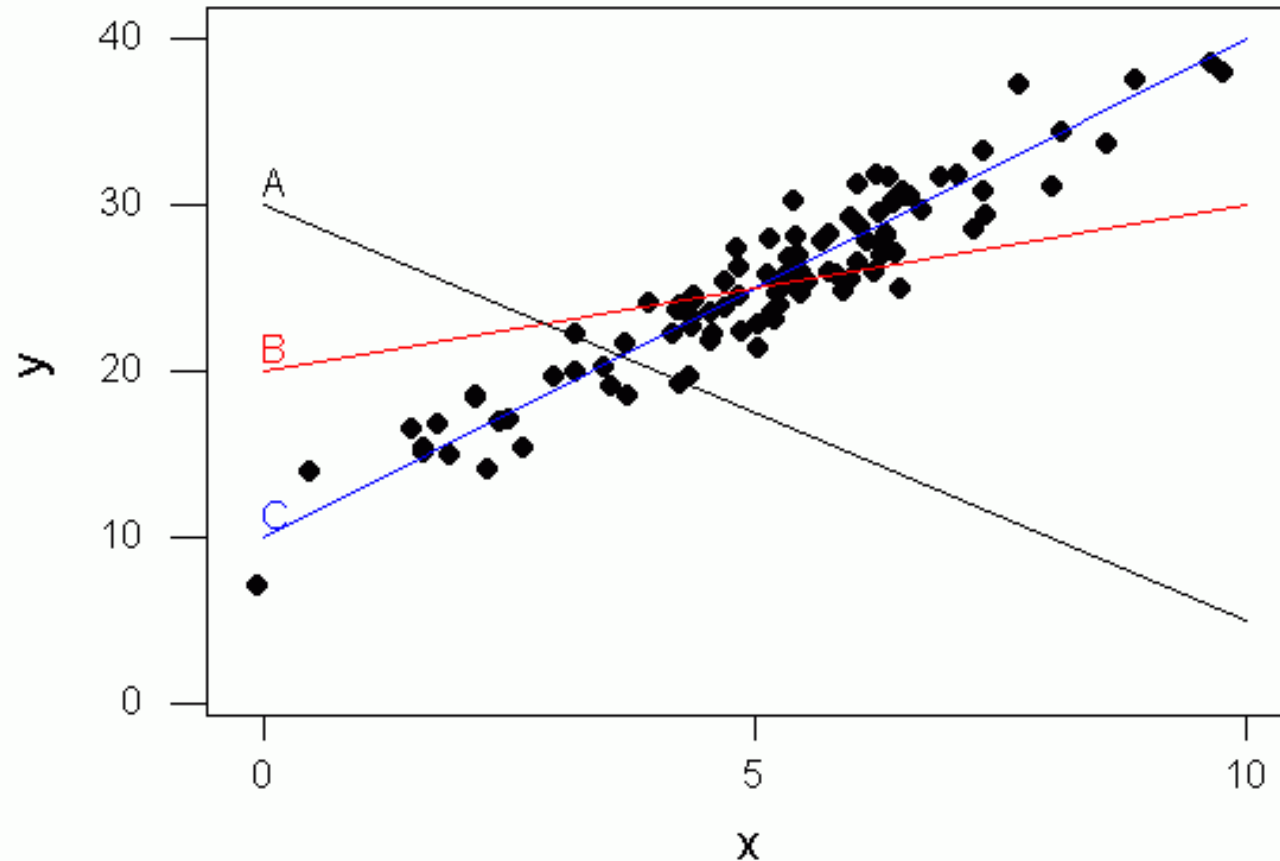
$$y = a + bx$$

a is value of y when $x = 0$, called **intercept**.

b says how much y changes when x increases by 1, called **slope**.

So choosing a line means choosing its intercept and slope.

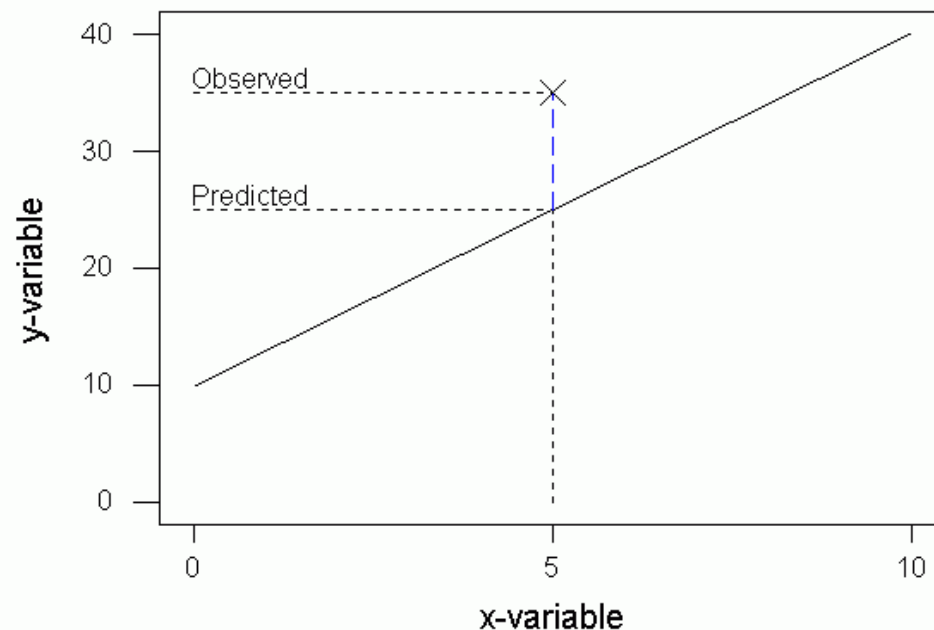
Choosing a line “through” the data



Line A all wrong; B not steep enough; C about right.

Least squares

Want line going “closest” to data with smallest “error”.



Observed: $x = 5, y = 35$. But line predicts $y = 25$ when $x = 5$, so error in using line is $35 - 25 = 10$ (height of dashed line).

This error called **residual**.

Each individual has x , y , predicted y ; for any particular line, work out residuals. Residual $+$ if observed $>$ predicted, $-$ if $<$.

Can't combine residuals by adding up, but can *square first* to make positive and then add up.

For a good line, all residuals small, sum of squared residuals small.

For a bad line, some residuals large, sum of squared residuals large.

Choose line with *smallest sum of squared residuals*. Called **least squares regression line**.

Return to MPG and weight for cars. Correlation was negative: larger weight means smaller MPG in general.

Regression line in Minitab: Stat, Regression, Regression.

Response is MPG (trying to predict), Predictor is weight (predicting MPG from it). Click OK:

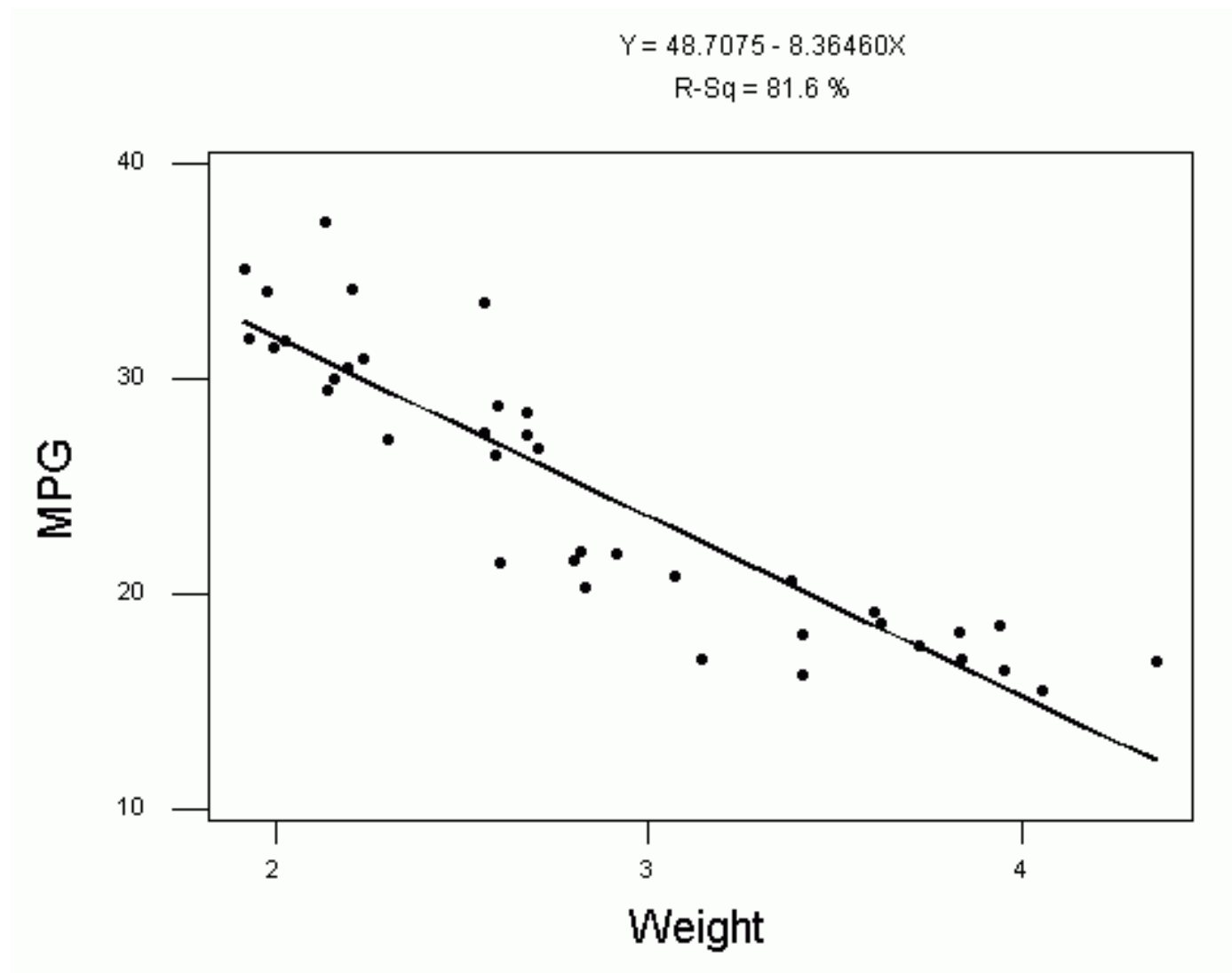
The regression equation is

$$\text{MPG} = 48.7 - 8.36 \text{ Weight}$$

plus other stuff.

Each increase in weight by 1 ton associated with *decrease* of 8 MPG.

Scatterplot with regression line:

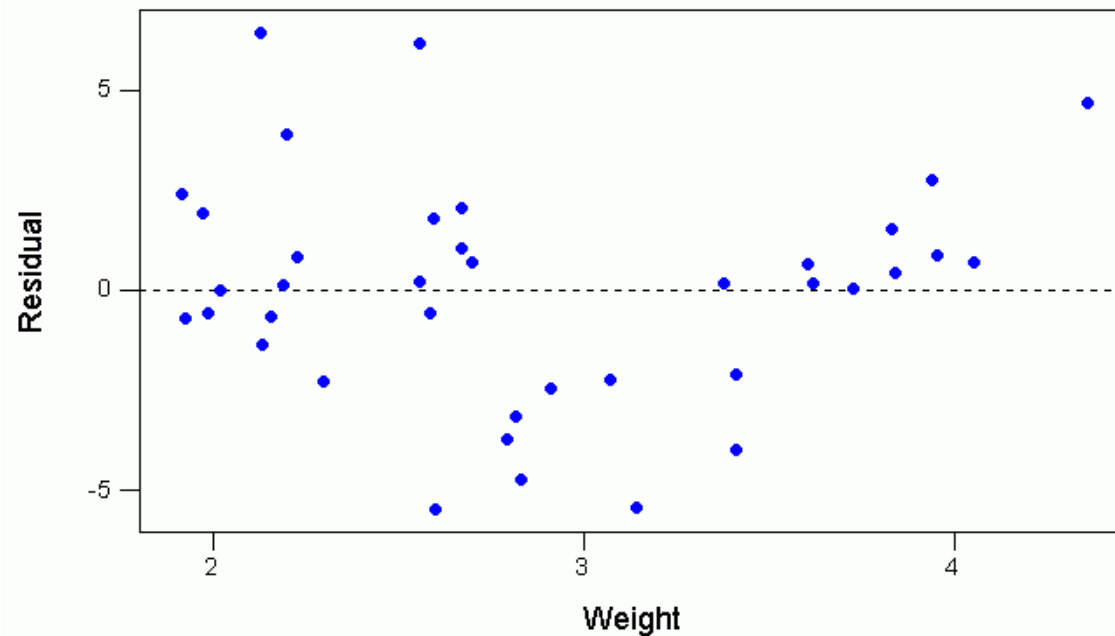


Line provides simple description of data.

But maybe *too* simple. Cars weight around 3 tons below line (worse MPG than predicted by line), lightest and heaviest cars above line (better MPG than predicted).

Residual plots (§2.4)

A good way to tell whether a regression is working properly is to plot the residuals against explanatory. For predicting MPG from weight, get down-and-up pattern.



Minitab: in Regression dialogue, click Graphs, in “Residuals vs. variables” box select explanatory (weight).

Aim: all association between variables summarized in *regression*.

Should be nothing left over, so residual plot should have *no pattern*.

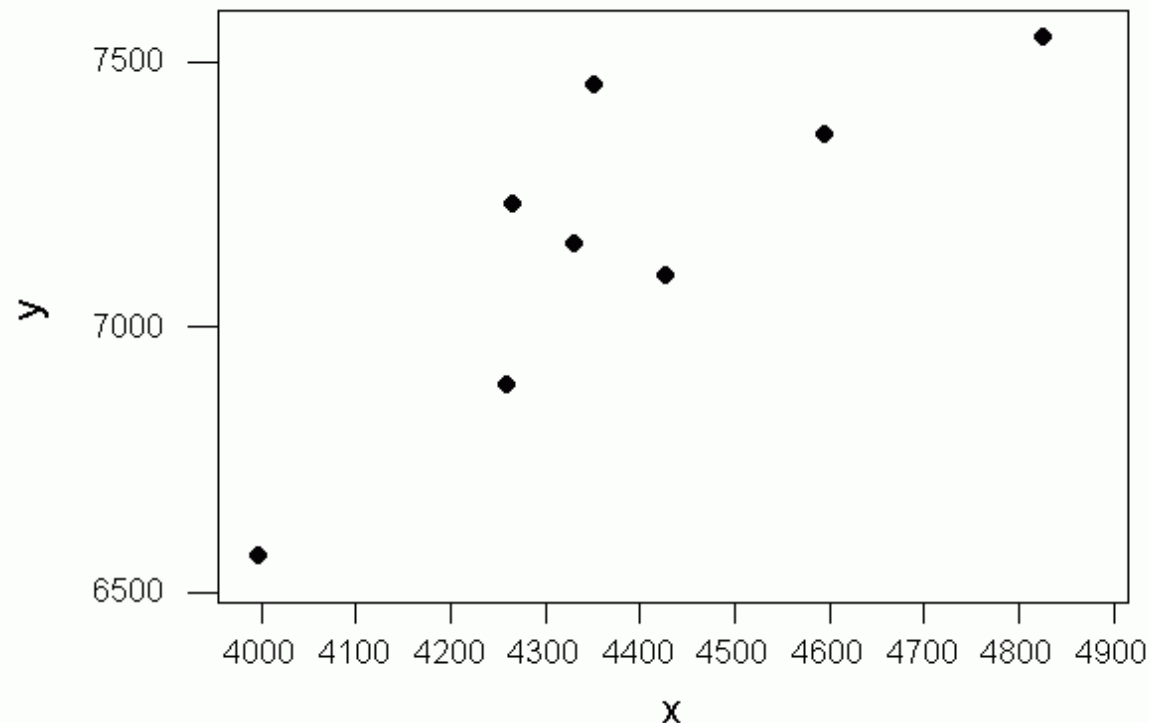
If there is a pattern, something wrong:

- curved pattern (as here and fig 2.28(b)) shows curved not linear association
- fan-out pattern (fig 2.28(a) p. 172) shows increasing variation about line as x increases – predictions more precise for smaller x .

Lurking variables

If *important variable missed* from regression, can be misled.

Predicting number of students in elementary university math courses (y) from total number of 1st years (x).



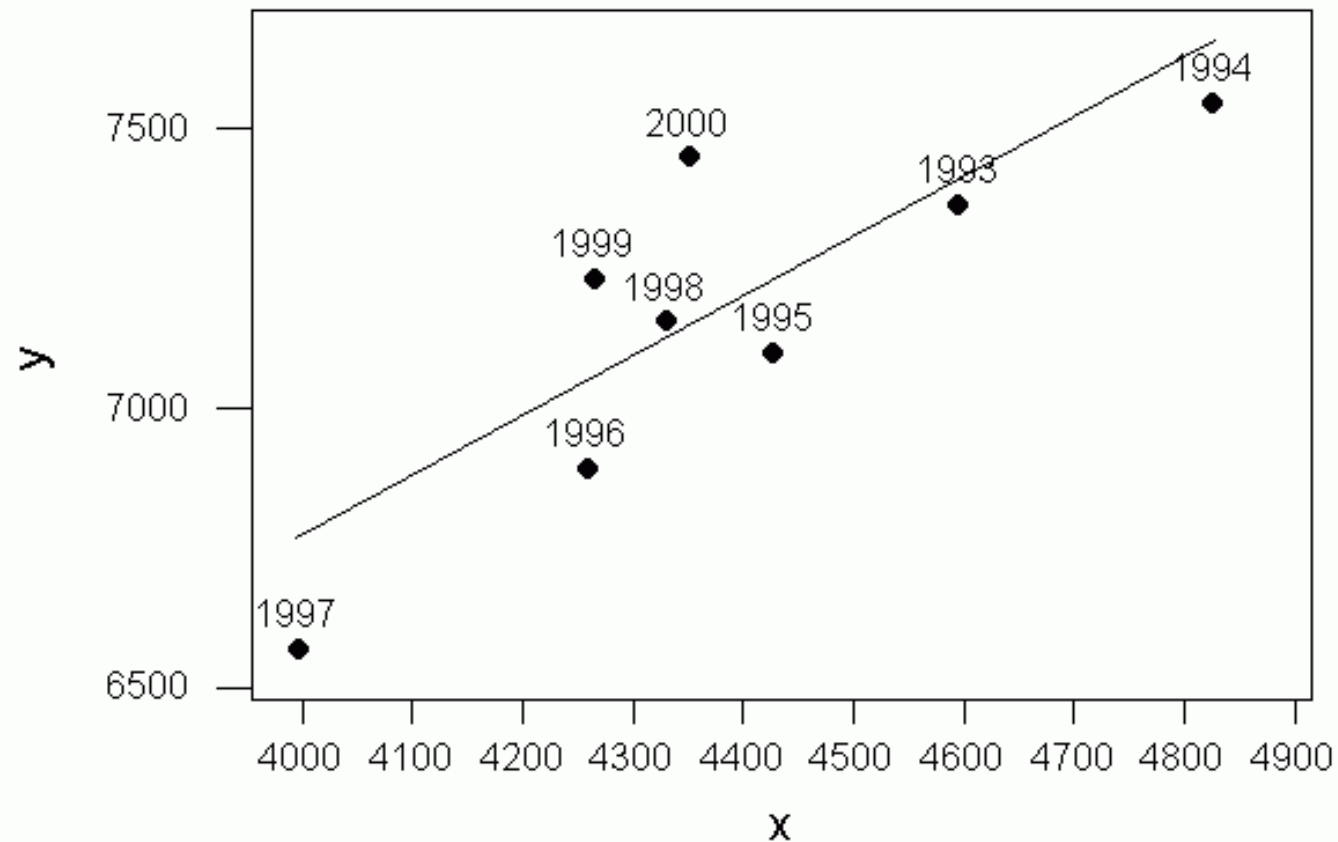
Upward trend (correlation 0.831).

Regression line:

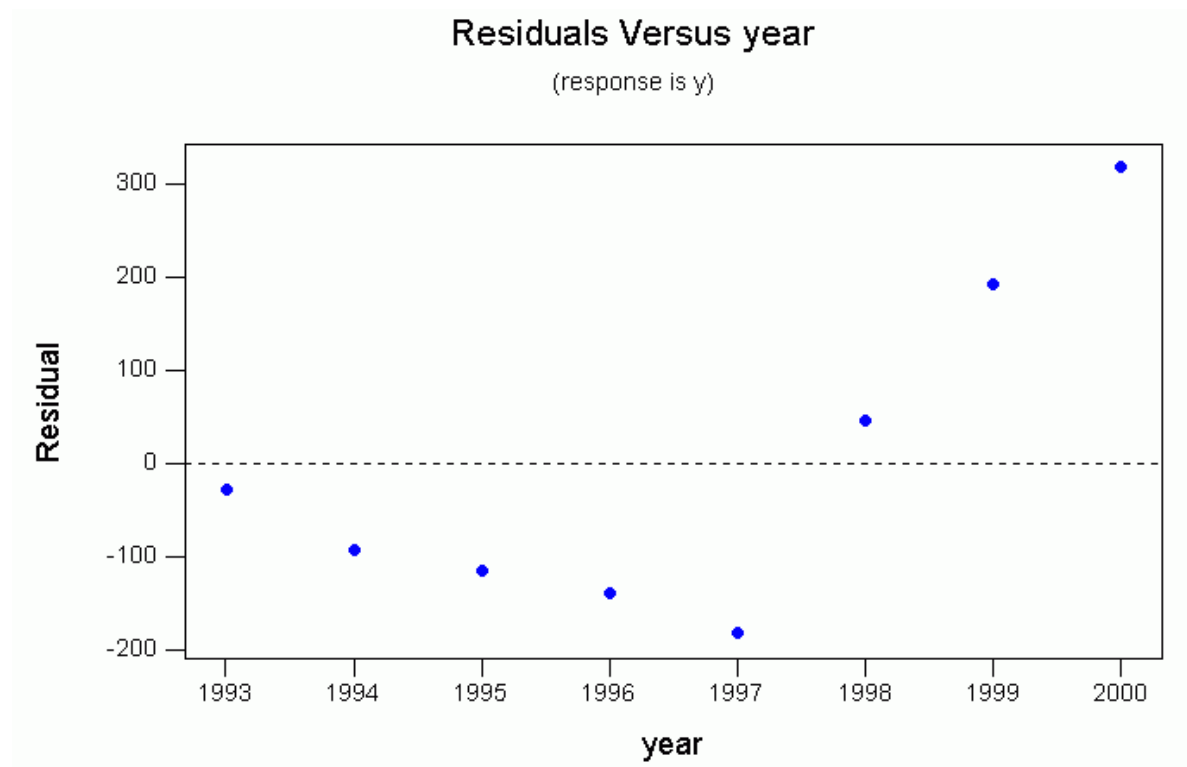
The regression equation is

$$y = 2493 + 1.07 x$$

Fit reasonably good, but plot line with years:



1st 5 years all below line, last 3 all above. Plot residuals against time to show more clearly:



Change in 1998: a dept required another math course.

Should not use pre-1998 data to estimate future enrollment.

Summary

- Correlation: does a straight line describe data?
- Regression: which straight line describes data?
- Choose line through data by least squares idea.
- Intercept is y when $x = 0$; slope is increase in y going with x increase of 1.
- Look at residuals to decide whether regression line useful.
- Beware of lurking variables that can distort apparent relationship.

Correlation does not imply causation (§2.5)

In science, find out how world works: want to know that changes in one variable *cause* changes in another.

But correlation and regression only show *association*: that the two variables change together. *Does not show whether either one causes the other.*

High correlation can be caused by lurking variable. Example: for countries, measure TV sets per person and average life expectancy.

High positive correlation, but TV ownership not *cause* of long life.
Cause for both variables is standard of living.

Establishing causation

Science: do *experiment* in which other lurking variables controlled (in lab). Gives most convincing evidence of cause and effect.

But in real world, impossible to experiment. To see whether smoking causes lung cancer, would have to choose 2 similar groups of people, make 1 group smoke, measure lung cancer rates. Not ethical!

How to convince others of cause and effect without experiment?

- **Strong association** (smokers suffer more from lung cancer than others)
- **Consistent association** (many different studies get same result)
- **Larger effect** with larger exposure (heavier smokers get more lung cancer)
- **Cause before effect** (lung cancer comes after starting smoking)
- **Cause is plausible** (experiments on animals show cigarette smoke causes cancer).

Summary

Correlation cannot prove that one variable is *cause* of another, but if:

- association strong and consistent
- larger effect associated with larger value of explanatory variable
- supposed cause comes before effect and is scientifically plausible

then have convincing evidence of cause.

Chapter 3: Producing data — Finding out what we want to know

What can we learn?

So far: have data, asked *What do we see?*

Use graphs, numbers to *describe*.

But this only a start. Ask *what can we learn?* Needs more work with:

- calculations and reasoning
- collecting data to make calculations/reasoning work

Aim here: collecting data well.

Finding available data (§3.1)

Question of interest, eg.:

- how are Canadians' eating habits changing?
- social/economic backgrounds of college students
- relative safety of flying vs. driving

Anecdotal evidence comes from own/others' experiences. But may not be typical, or remember untypical cases (plane crash killing 200). May not represent whole phenomenon.

Need to get data telling *whole story*. Visit library or Internet: much data (not gathered especially for us but helpful eg. for first question).

From Statistics Canada website:

- Canadians eating more pasta/bakery/cereals, less red meat than in past years
- Milk consumption steady (after falling in previous years), but now more low-fat milk.
- Consumption of cream has increased (because consumption of coffee has increased?)

Reliable? Probably yes. Detailed surveys of representative samples of Canadians gathered for information (not by group with axe to grind).

Producing data

Producing new data expensive, but often only good route to clear answers.

Ideal: **census**. Measure *every* individual of interest, whole **population**. Slow, impossibly expensive.

In practice: **sample**. Choose small collection of individuals to “represent” population, draw conclusions about population based on sample.

Samples are kind of *observational study*. Aim to collect data without changing anything, and conclude whatever possible.

Compare statistical **experiment**: conditions deliberately changed to see what happens. Better than observational study, because can observe effect on response by changing right thing (get cause/effect).

Example: to find what Canadians eat, observational study (sample) fine.

But to find whether one diet healthier than another, need experiment. Get 2 groups with same mix of age/sex/lifestyle; one group gets diet 1, other gets diet 2. Groups started similar, so differences at end caused by diets.

Summary

- Ask: what can we learn?
- Using available data
- Sample from large population
- Observational study: observe only
- Statistical experiment: deliberately change conditions
- In experiment: difference at end can mean factor modified *made* difference.

Design of experiments (§3.2)

- Individuals on which experiment done: **experimental units**, or **subjects** if people.
- Specific experimental condition applied to units: **treatment**.

Eg. comparing success rates of surgical procedure at 2 hospitals.

Could just observe, but success differences might be result of patient differences.

Experiment: control which patient (subject) goes to which hospital (treatment). Share out more/less critical cases among hospitals; difference in success rates then evidence of difference between hospitals.

Success rate here response; hospital is (categorical) explanatory variable, called **factor**. 2 specific hospitals called **levels** of factor.

Placebo effect and control groups

Many subjects in an experiment respond favorably even if treatment is ineffective (eg. patients getting pill with no active ingredient).

Called **placebo effect**, and **placebo** is treatment designed to do nothing except look like real treatment.

Group of subjects getting placebo called **control group**.

Important to have control group because *controls* for psychological effect of receiving “treatment” at all. Better to assess *one* treatment by comparing *two* groups, treatment and control.

Randomization

How to assign experimental units to treatments? Need to make treatment groups “similar”.

Can try *matching* on all relevant variables. Helpful, but if variables forgotten, can give dissimilar groups.

Easier: use **random** assignment. Ensures that group assignment not related to anything else. Groups usually similar in terms of anything important.

Using Minitab to do randomization

Suppose we have 30 subjects, want 15 in treatment group and 15 in control group. Effectively “draw names from hat”.

Give each subject a number, 1–30. Calc, Make Patterned Data, Simple Set of Numbers. Store in: type C1, First Value 1, Last Value 30. Click OK.

Now select 15 subjects for treatment group: Calc, Random Data, Sample from Columns. In box, sample 15 rows, click From box, select C1, type C2 in other box. I got:

26	25	8	27	18
2	22	19	9	20
14	4	11	24	23

These in Treatment group, rest in Control. (Different if repeated.)

For more than two groups, easier to “shuffle” whole list of 30 (sample all 30 rows). Then pick eg. first 10 for Treatment 1, next 10 for Treatment 2, last 10 for Control.

Conclusions from randomized experiments

Groups in randomized experiment not *exactly* same, though will be very similar. So any differences afterwards could be:

- chance: the groups weren't quite the same to begin with
- treatment effect: one of the treatments really is better.

Later, learn how big chance differences can be. Observed difference bigger than chance: must be treatment effect. Such difference called **statistically significant**.

Randomization at work

In many diseases, success of treatment depends on age of patient. So want treatment, control groups to have similar mean age.

Worksheet `trtcontrol` has ages for 30 people. Mean age is 54.

Select 15 people for treatment group, calculate mean age. In Sample from Columns, selecting 2 input and 2 output columns carries age info along. My results:

Mean of C5 = 50.733

Mean of C5 = 54.400

Mean of C5 = 56.667

Mean of C5 = 56.200

Mean of C5 = 47.600

Mean of C5 = 55.267

Mean of C5 = 55.600

Mean of C5 = 53.067

Mean age of treatment group generally very close to 54. Taking more people and larger groups would give even better results.

Cautions about experimentation

Experiments always conducted under artificial conditions. Setting of experiment may not reflect real world – **lack of realism**.

Example: TV commercials tried out in “focus groups” where people invited to watch and vote on several alternative commercials for same product. Very unlike TV-watching, so best focus-group commercial may not be best TV commercial.

Also, people administering and receiving treatment may have prejudices. (As patient, what if you knew you had placebo?)

Important to design experiment **double-blind** so no-one in running of experiment can be biased. (Each subject gets something identical with code number attached.)

Matching and blocking

Often measure each subject *before* experiment as well as after – gives baseline for comparison (eg. learning or exercise tasks).

Example of **matched pairs** – 2 measurements per subject.

“Case-control study” for cause of disease: for each “case” with disease, select similar person without disease (“control”), see whether supposed cause more common in cases than controls.

Idea: compare *within* pairs.

Another example: comparing effectiveness of pesticides in farming.
But soil type, fertility differ by location. So apparent effectiveness depends on pesticide *and* location.

Idea: divide locations into **blocks** – in each block, soil type, fertility same. Then randomize pesticides into locations *within* each block.
Then each block has fair comparison of pesticides A, B, C, D:

	Location			
	1	2	3	4
Low fertility	C	A	B	D
Mod fertility	C	B	D	A
High fertility	B	C	A	D

Summary

- Individuals in experiment called experimental units or subjects.
- Condition being modified called treatment.
- May be effect of “nothing” called placebo effect; use control group to provide comparison with “nothing”.
- Assign units (subjects) to control/treatment groups by randomizing: groups “similar on average”.
- Big enough observed difference can be effect of treatment.
- Experiments can lack realism or lack blinding.
- Before-after studies can isolate effect of treatment.
- Units very different: test each treatment on mixture (blocking).

Sampling (§3.3)

When can't experiment, can observe, but impractical to observe everyone. Jargon:

- Entire group of individuals of interest called **population**.
- Part of population actually examined called **sample**.

How to choose sample?

Bad: *voluntary response*, eg. polls on radio stations. People choose themselves to be in sample, usually because of interest in issue discussed. Results biased towards strong opinions.

Good: use *randomization*. No connection between being in sample and issue being addressed.

Simple random sample

Simplest: put names (population) in hat, draw out some (sample), say n .

This is **simple random sample**. Each set of n individuals equally likely to be sample actually selected.

In Minitab: same idea as for randomizing treatment group. Make list of population, use Calc, Random Data, Sample from Columns to select individuals for sample.

Stratified sampling

A population often has dissimilar groups in it. Eg. opinions of Ontario people different from those in Québec, West, Maritimes.

Simple random sample might under-represent some groups by chance. Idea: take simple random sample *from each group*.

Combined **stratified random sample** represents all groups.

Multistage sampling

Drawing simple random sample requires list of population – big if sampling all of Canada!

Also, SRS of Canada would have small numbers of people in widely scattered places – impractical.

- Divide Canada into areas (eg. electoral ridings), select random sample of ridings
- divide selected ridings into smaller areas (city blocks etc), select random sample
- select random sample of people in smaller areas

Easier process in stages; final sample less scattered.

Sample survey problems

Requires list of population/subpopulations. If list incomplete, sample biased because of **undercoverage**.

Individual chosen for sample may refuse to take part (**nonresponse**). Nonresponders may be different from responders; if so, causes bias in results. (Eg. calling homes in working hours.)

Questions on sample survey need careful, neutral **wording**:

- 44% think America spends too much on *welfare*
- 13% think America spends too much on *assistance to the poor*.

Good statistical design only part of successful sampling.

Towards statistical inference (§3.4)

So what is it possible to learn about a population when all we have is a sample?

Answer: quite a lot, as long as it's a *random* sample.

Really the purpose of statistics: from a sample, making **statistical inference** about population.

Because of randomness in sampling, won't get exactly right answer, but hope to be close (and to know how close).

Imagine mean height of all adult women in Canada. Some number of inches, but can only know by measuring everyone. But draw random sample of 100 Canadian women; mean height of those women can be calculated.

- **Parameter** is number describing *population*. Fixed but *unknown*.
- **Statistic** is number describing *sample*. Can be *calculated*, but different in different samples.

Mean height of all Canadian women is *parameter*, mean height of women in sample is *statistic*. Take different sample, get different statistic.

Sampling variability

Random sampling eliminates bias/favoritism (no control over which individuals end up in sample).

But different samples have different sample statistics – **sampling variability**. Hence sample gives imperfect answer. But maybe answer close enough to be useful.

Going from (one) sample to population difficult. But going from population to sample(s) can be easy.

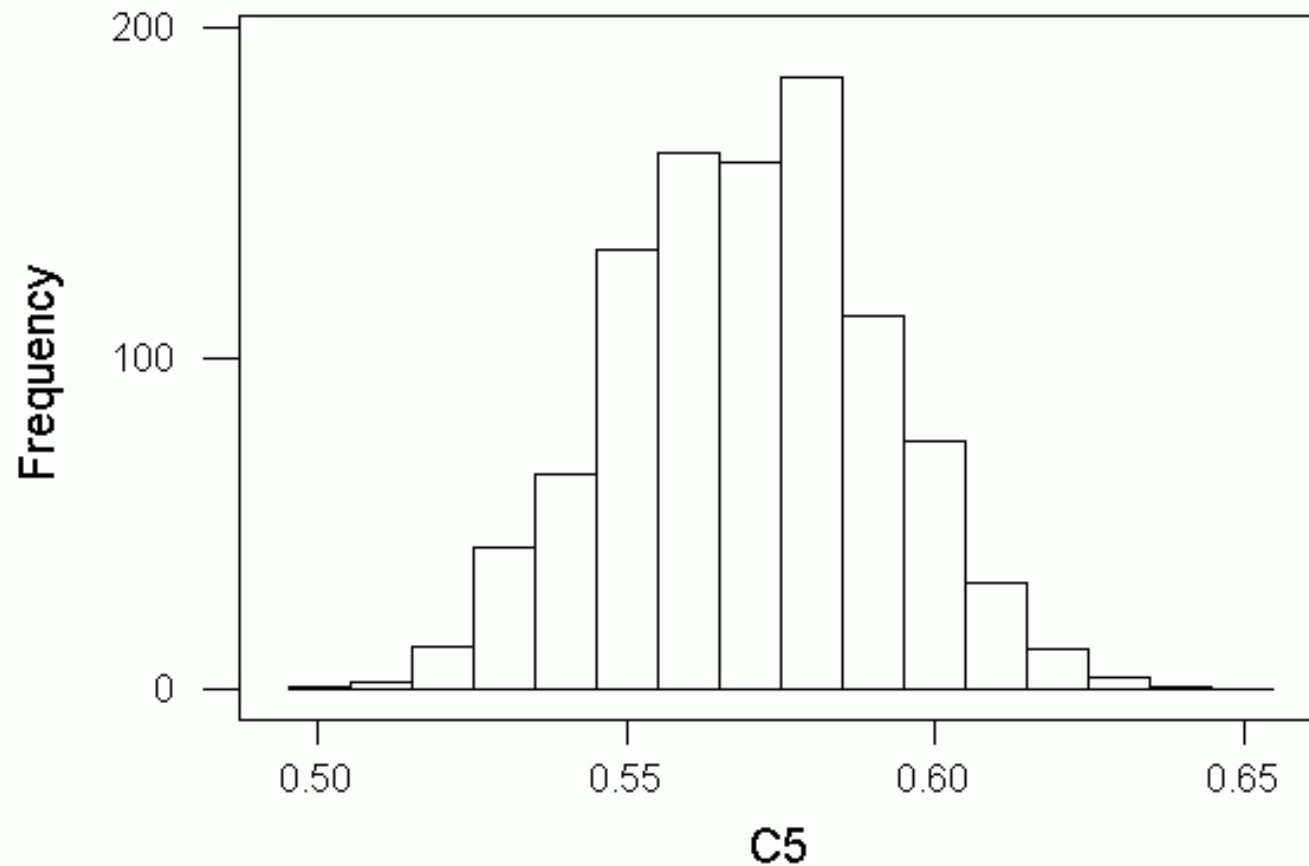
Simulation

Many possible samples from most populations (number of possible samples of 10 people from 100 is number with 14 digits). But looking at a lot of samples gives good enough idea.

Steps:

- pick a population to sample from
- generate a random sample from the population, calculate sample statistic
- repeat previous step many times
- make histogram of results

Example: 57% of students at college female. In samples of 500 students, what proportion of females might we get?



- Centre near 0.57.
- Spread *small* (50–65% women, values close to 57% commoner than extreme values).
- Shape close to *normal*.

Collection of possible sample proportions called **sampling distribution** – shows what sample statistic might be, if population parameter known.

Bias and variability

Statistic usually used to estimate population parameter (sample proportion to estimate population proportion). In this role, called **estimator**.

Sampling distribution of estimator desirably:

- centre around the population parameter (**unbiased**)
- has small spread (low **variability**).

Fact: provided sample small part of population, sampling distribution *does not depend on size of population*.

Benefits of randomization

Possible to work out sampling distributions by mathematics as well as simulation. Only true for random samples – take sample any other way, have no idea of sample-population relationship.

With random samples, can reduce variability by taking larger sample. With very large random sample, sample statistic very close to population parameter.

Sampling distribution shows how close sample, population quantities might be by going from population to sample. Later, use same idea to go from (one) sample to population.

Summary

- Choose sample by randomization (no connection between sample and variable studied)
- Simple random sample (each individual has same chance to be in sample, independent of others)
- Stratified, multistage sampling when simple random sample inconvenient
- Conclusions only as good as sampling method: beware of undercoverage, nonresponse, biased wording.

- Want to learn about population parameter, only have sample statistic. But with random sampling, can know how close we may be to “right answer”, even though different samples will give different results.
- Reasoning from sample to population hard, but if we know population, can figure out what kind of samples might come from it.
- Can understand effect of looking at many different samples by simulation.
- If sample is not drawn randomly, we are stuck!

Probability: The Study of Randomness

Randomness (§4.1)

Toss a coin – can't predict whether outcome will be Heads or Tails.

Roll a die – can't predict whether you'll roll a 1 or something else.

Play roulette (or other casino game) – can't predict whether you'll win or lose.

In each case, individual outcome cannot be predicted, but pattern emerges:

After many coin tosses, you will observe about 50% heads and 50% tails.

After many die rolls, about $\frac{1}{6}$ of the rolls will be 1.

When individual outcomes are uncertain, but a pattern emerges when many outcomes are observed, phenomenon called **random**.

Proportion of times an outcome is observed over the long run called **probability** of that outcome ($\frac{1}{2}$ for Heads, $\frac{1}{6}$ for 1 on die). Write as $P(H) = \frac{1}{2}$, $P(1) = \frac{1}{6}$.

Probability is a number between 0 and 1. 0 means outcome *impossible*, 1 means outcome *certain*.

Probability models: some jargon (§4.2)

To describe a random phenomenon, need two things:

- List of all possible outcomes (**sample space**, written S).
- Probability for each outcome.

These two things together called **probability model**.

Examples of sample spaces:

- flip a coin once. $S = \{H, T\}$.
- flip 5 coins, count number of heads. $S = \{0, 1, 2, 3, 4\}$.
- choose a Canadian at random, note down province of residence.

$$S = \{\text{BC, Alberta, } \dots, \text{Ontario, } \dots, \text{Newfoundland}\}.$$

Events

An **event** is a collection of outcomes.

Example: toss 2 coins. Sample space is

$S = \{HH, HT, TH, TT\}$. “Exactly one head” is an event, consisting of outcomes HT and TH.

An event has a probability, which you get by *adding* the probabilities of its constituent outcomes.

In this example, if the coin is fair (unbiased), any outcome has probability $\frac{1}{4}$, so prob. of “exactly one head” is $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

The whole sample space is an event. Prob. of S here is

$$P(S) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = 1.$$

Where do probabilities come from?

$P(S) = 1$, so probs for all the outcomes must add up to 1.

Can either:

- calculate prob. for each outcome
- assume that each outcome is *equally likely*; then probs are all same (and must add up to 1). In 2-coin example above, were 4 equally likely outcomes, so prob. of any one is $\frac{1}{4}$. In general, for n outcomes, prob. is $1/n$.

The event “not A ”

Returning to the two-fair-coin example, suppose $A = \{HH\}$. A contains 1 of 4 outcomes (equally likely), so $P(A) = \frac{1}{4}$.

The event “not A ”, written A^c , contains all the other outcomes. So here $A^c = \{HT, TH, TT\}$. Then $P(A^c) = \frac{3}{4} = 1 - \frac{1}{4}$.

General rule: for any event A , $P(A^c) = 1 - P(A)$.

Independent and disjoint events

Imagine that you toss 2 coins. Let A be event “get H on 1st coin”, B be event “get H on 2nd coin”. Knowing that A happens doesn’t make B any more or less likely: A and B are *independent* events.

Again imagine tossing 2 coins. Let A be event HH, let B be event “head on 1st coin”, so $B = \{HH, HT\}$. If A happens, B is certain to happen, so A and B *not* independent events. A and B have outcome HH in common, so are *overlapping* events.

Now let C be event TT. If A happens, C cannot happen (coin cannot show both heads and tails), so A and C not independent. But have no outcomes in common; called *disjoint* events.

Finding probabilities from independent and disjoint events

For general events A and B ,

- if A and B are independent, **multiplication rule** says

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

- if A and B are disjoint, **addition rule** says

$$P(\text{either } A \text{ or } B) = P(A) + P(B)$$

To use these rules: first, are events independent? If they are, can use multiplication rule. If not, are they disjoint? If they are, can use addition rule.

Examples

In a certain lottery, you can win either prize A, prize B, or no prize.

$$P(A) = 0.01, P(0) = 0.90.$$

To find $P(B)$, 3 probs have to add up to 1, so

$$P(B) = 1 - 0.01 - 0.90 = 0.09.$$

$P(\text{some prize}) = P(A \text{ or } B)$. A and B not independent, but are disjoint (can't win more than 1 prize from 1 ticket), so can use addition rule.

$$P(\text{some prize}) = P(A) + P(B) = 0.01 + 0.09 = 0.10. \text{ (Or:}$$

“some prize” means “not no prize”, which gives answer as

$$1 - 0.90 = 0.10.)$$

Suppose two people play this lottery. Prob. that both win some prize? Neither wins a prize?

If they pick numbers separately, events independent, so can use multiplication rule.

$$P(\text{both win some prize}) = 0.10 \times 0.10 = 0.01$$

$$P(\text{neither wins}) = 0.90 \times 0.90 = 0.81.$$

These two probs don't add up to 1 because one person could win and the other not.

Random variables (§4.3)

Suppose you:

- toss 4 coins and count the number of heads
- roll 2 dice and count the total number of spots

In each case, each outcome gives you a number. (In 1st case, don't care *which* came up heads, just *how many*).

This number called a **random variable**, labelled X .

Each value of random variable has probability. $P(X = 3)$ is sum of probs of all outcomes leading to X being 3.

Example

Value of X	1	2	3
Probability	0.3	0.6	0.1

Is example of *probability distribution* (list of values and probs).

$$P(X \geq 2) = P(X = 2) + P(X = 3) = 0.6 + 0.1 = 0.7.$$

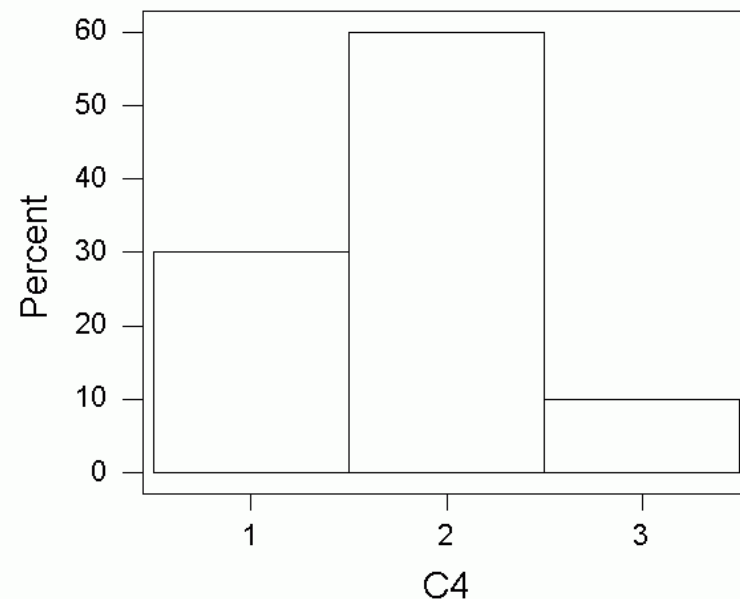
Note also that

$$P(X = 1, 2 \text{ or } 3) = 0.3 + 0.6 + 0.1 = 1,$$

that is, all the probs together add up to 1.

Probability histogram

Can make a plot: above each value of X , draw a bar whose height is the probability. Looks like a histogram. For above distribution:



Highest bar above 2, showing that $X = 2$ has highest prob.

Continuous random variables

So far, can list possible values and their probs: **discrete random variable**. But what if *all* values in an interval (say between 0 and 1) are possible?

Probability histogram becomes **density curve** (compare §1.3).

Probability of getting value in any interval is *area* under density curve (recall normal distribution: Table A gives area).

Mean of a random variable (§4.4)

Data distribution has mean and SD you can calculate. Likewise, probability distribution (and thus random variable) has a mean and SD. But the way to calculate them is different.

For the mean of a random variable, multiply each value by its probability, and add them up. To recycle our example:

Value of X	1	2	3
Probability	0.3	0.6	0.1

Mean of X is $(1)(0.3) + (2)(0.6) + (3)(0.1) = 1.8$. Often use symbol μ (μ_X) to represent mean (of X).

Here, $\mu_X < 2$ because $P(X = 1) > P(X = 3)$.

Statistical estimation and law of large numbers

Suppose we want to know mean height of all adults in Canada. Idea from Chapter 3: take sample, look at sample mean. But how close is that to mean height of *all* adults?

Height of randomly chosen Canadian adult is random variable H . H has a mean μ . Hope that sample mean \bar{x} will be close to μ , especially with large sample.

Law of large numbers says this more carefully. Decide how close you want to be to μ (say within 1 inch), and how likely you want to be that close (say 90%). Then, with large enough sample, will be as close as you want with specified probability.

Supports idea of using sample mean as estimate of population mean.

Variance and SD of a random variable

To calculate the SD of a random variable, first find variance, using this recipe:

1. Take each value of the random variable and subtract the mean.
2. Square each result.
3. Multiply by the probability and add up.

To use our example again:

Value of X	1	2	3
Probability	0.3	0.6	0.1

The mean was 1.8, so for step 1, the values are $1 - 1.8 = -0.8$, $2 - 1.8 = 0.2$, $3 - 1.8 = 1.2$.

Square them to get $(-0.8)^2 = 0.64$, $0.2^2 = 0.04$, $1.2^2 = 1.44$.

Times by probabilities and add up to get

$$(0.64)(0.3) + (0.04)(0.6) + (1.44)(0.1) = 0.36.$$

The SD is the square root of this, so SD of X is $\sqrt{0.36} = 0.6$.

Like the SD of a data distribution, the SD of a random variable measures its spread (whether values far from the mean are likely or unlikely).

Rules for means, SDs and variances

- Multiplying X by a constant multiplies the mean by same constant.
- Adding a constant to X adds the same constant to the mean.
- Multiplying X by a constant multiplies the SD by the same constant.
- Adding a constant to X doesn't change the SD.
- For two random variables X and Y , mean of $X + Y$ is sum of means of X and Y .
- For two *independent* random variables X, Y , *variance* of $X + Y$ is sum of variances of X and Y .

Suppose the temperature in a city has mean 7 and SD 10 (degrees C). What are the mean and SD in degrees F?

Get degrees F by multiplying by 1.8 and adding 32.

So mean in degrees F is $(7)(1.8) + 32 = 44.6$. But for SD, adding makes no difference, so SD in degrees F is $(10)(1.8) = 18$.

Now let X be sales of cars at a dealership: X has mean 20 and SD 5. Y is sales of trucks, with mean 8 and SD 4.

Total sales have mean $20 + 8 = 28$. If X and Y independent, variance of total sales is $5^2 + 4^2 = 41$, so SD is $\sqrt{41} = 6.4$. But X and Y may rise and fall together (eg. with economic conditions).

Chapter 5: Sampling distributions

Introduction

Statistical inference: from sample to population.

Simulation (mathematics): from population to possible samples (sampling distribution).

Look at some more sampling distributions. First simulate, then use exact mathematical results. Along way, learn why operating big casino is money-maker even though casino games based on chance and randomness.

Sampling distributions for counts and proportions (§5.1)

Sample surveys often ask questions with “yes/no” type answers, like “Do you consider that this university is well run?”.

Industrial quality control: samples of TVs are taken from production line and tested. Each TV “satisfactory” or not.

In each case, “successes” and “failures” in the sample tell something about the populations.

Population of “all students at UTSC” actually exists (registrar’s records). Population of “all TVs produced at factory” harder to grasp. But often justified in treating sampled TVs as random sample from hypothetical population.

Many possible sampling distributions (corresponding to different kinds of statistic we might measure). Nature of sampling distribution depends on kind of population and sampling procedure.

Sample counts and proportions

Suppose we sample 20 TVs from our production line, and find that 17 of them work satisfactorily. Just want to know what % of *all* TVs would work, so doesn't matter *which ones* in sample are OK.

17 here is **sample count** of satisfactory TVs out of 20 sampled.

Sample proportion is sample count as fraction of whole: here $17/20 = 0.85$.

Binomial distribution

Sampling distribution of sample count depends on sampling method, but common scenario is:

- *Fixed number n* of observations.
- Observations fall into one of two categories (“success” and “failure”, alternatively “heads” and “tails”)
- Observations all *independent* (knowing one observation to be success doesn’t affect chance of others being success)
- Chance of success for each observation *always same* (p).

Simple example: tossing coin 10 times; # heads has binomial distribution.

Binomial distribution in sampling

Counts of successes in sample often described by binomial distribution. True provided sample is small fraction of population.

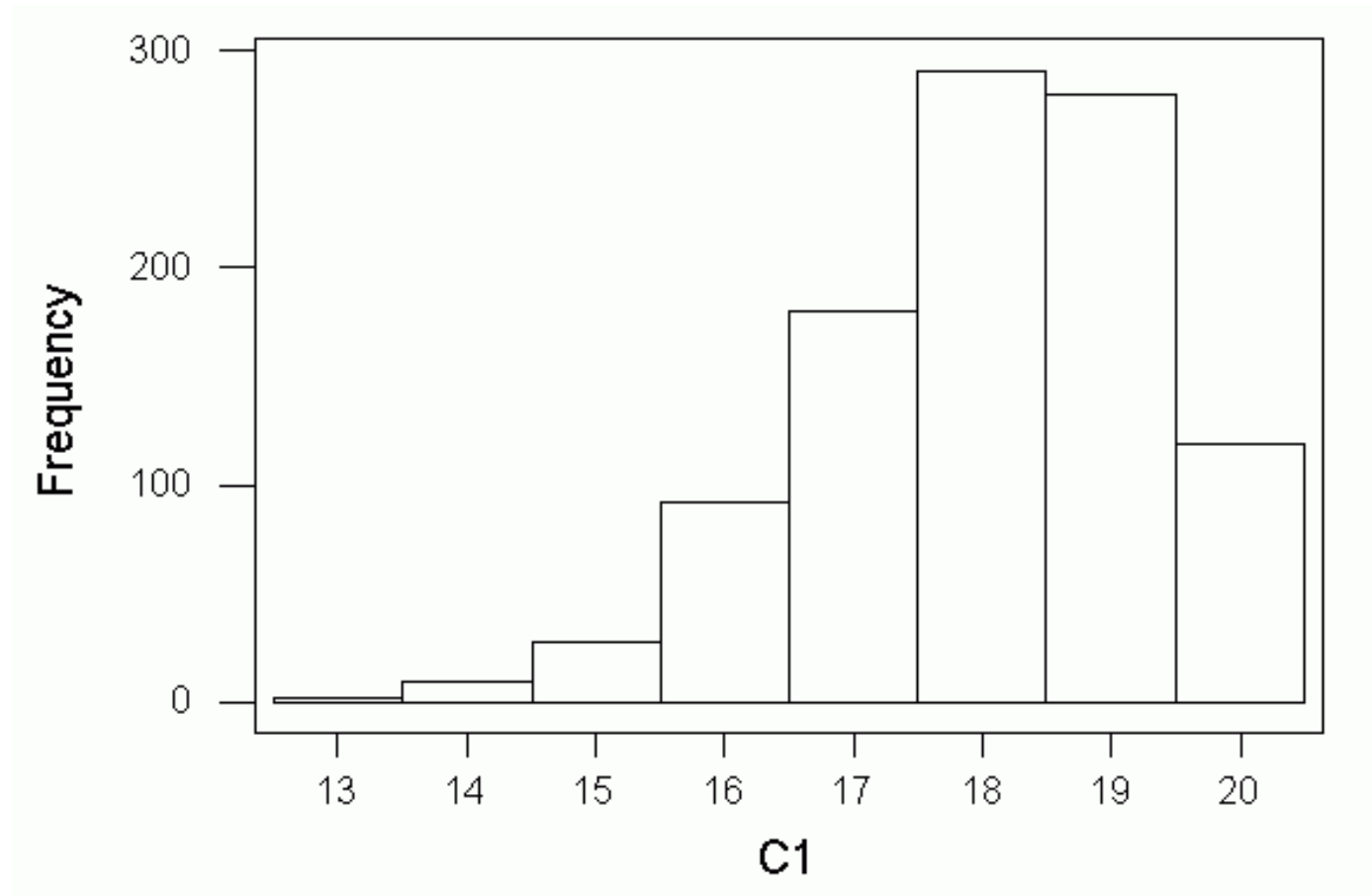
(If sample most of population, might get almost all population successes before finished sampling, then know we have failures to come.)

Simulating binomial

TVs: suppose 90% of all TVs produced in past satisfactory. Treat as population proportion $p = 0.90$. Sample $n = 20$ TVs; how many might be satisfactory?

Simulate. Minitab: Calc, Random Data, Binomial. Generate 1000 rows of data (“many samples”), store in C1, Number of Trials 20, Probability of Success 0.90.

Histogram of results:

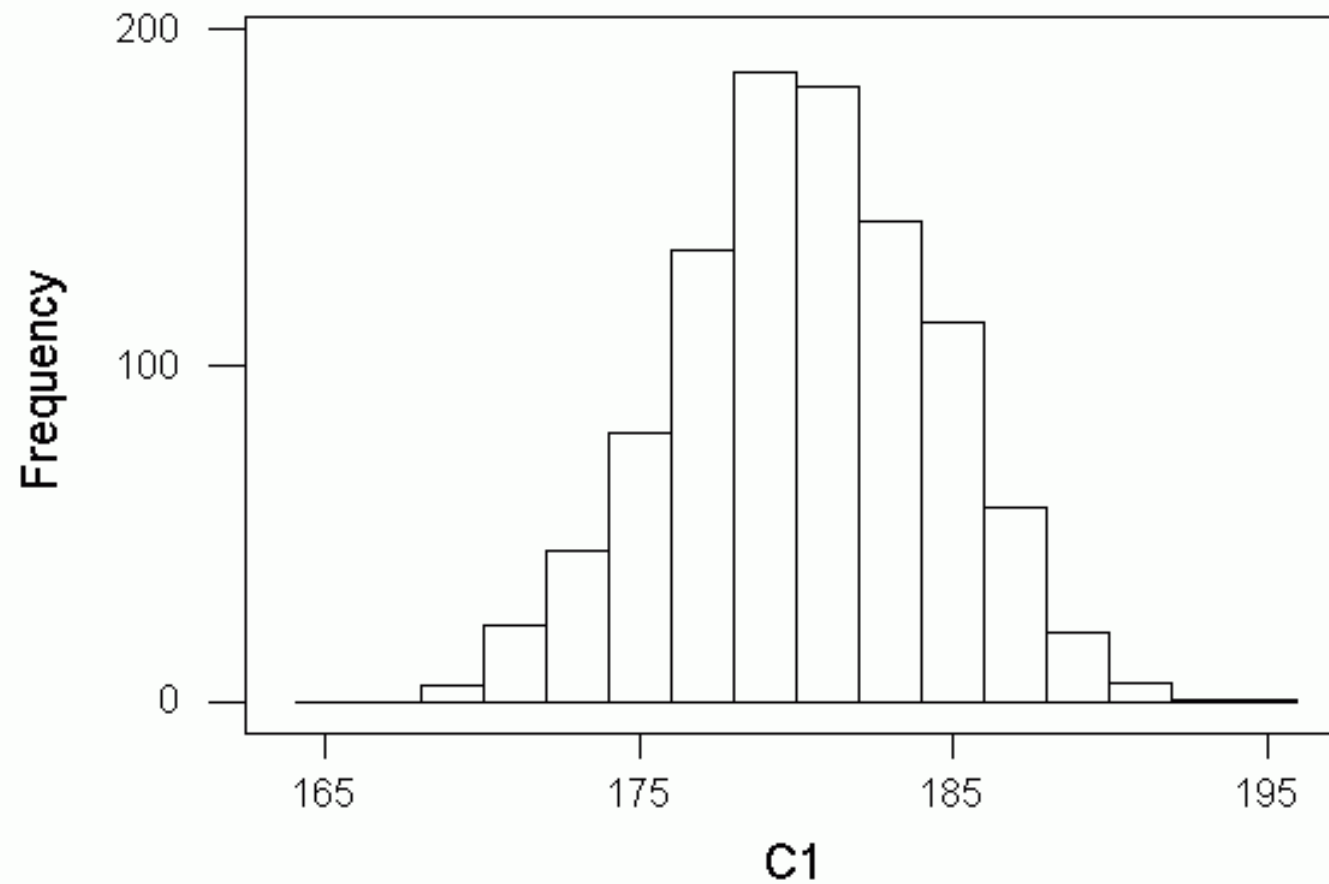


Shape skewed to left.

Centre around 18 (90% of 20).

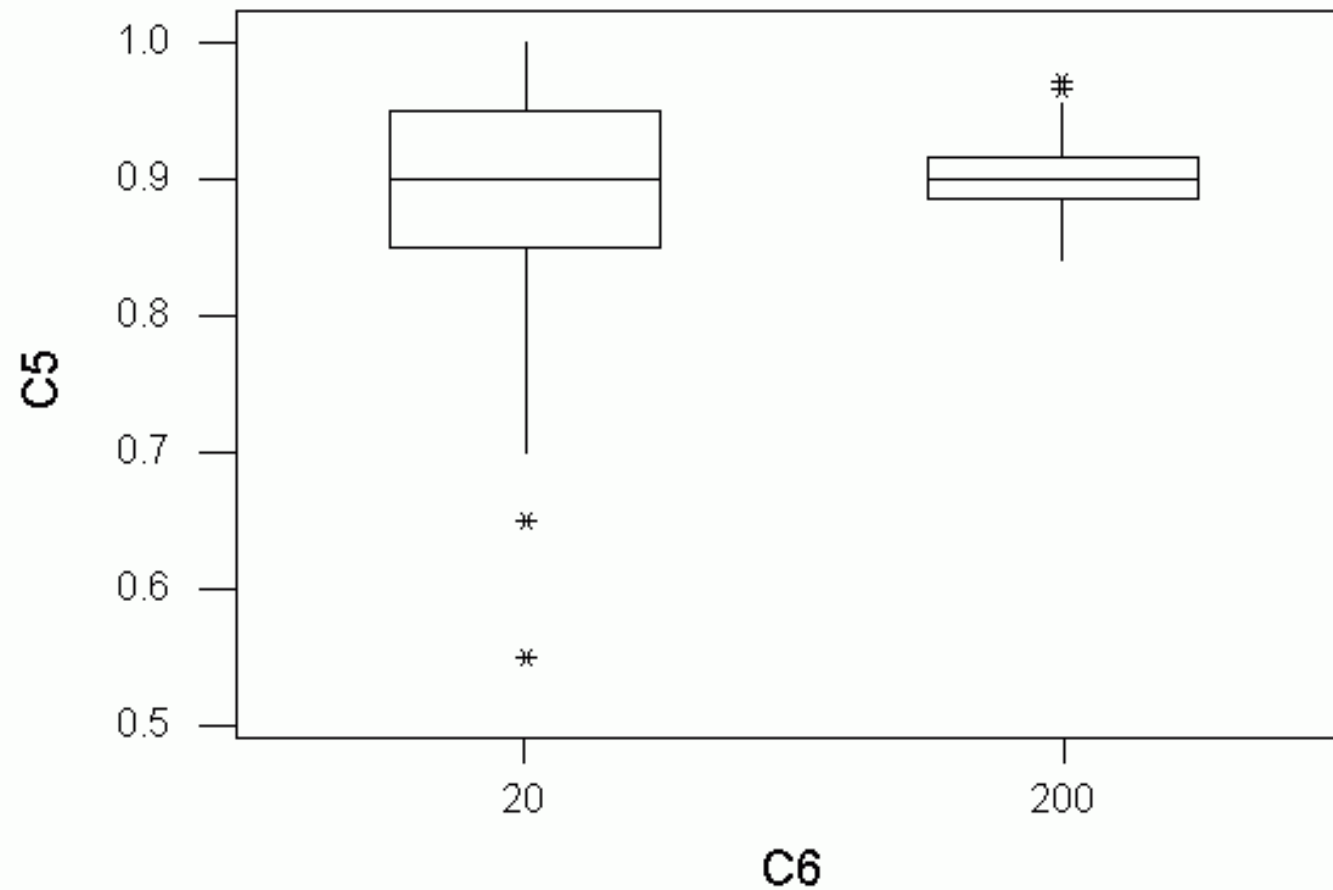
One sample had only 11 satisfactory TVs.

Now try with samples of 200 TVs ($n = 200, p = 0.90$):



Centre around 180 (90% of 200). Shape like *normal distribution*.

Counts difficult to compare (out of 20 vs. 200). Calculate *proportions* (divide by 20 or 200), show boxplots of proportions.



For both sample sizes, centre of sampling distribution close to 0.90.
But sampling distribution for $n = 200$ much less spread out, and more symmetrically shaped.

General fact for sample proportions based on large sample:

- sample proportion almost certainly very close to the population proportion. **Law of Large Numbers.**
- sampling distribution shape very close to *normal*.

The point of all this

Sampling distribution of count has binomial distribution; for large samples, sampling distribution of count or proportion close to normal.

That is, if we know population, we know what samples from it will look like.

In particular, know how close sample proportion likely to be to population proportion.

Later, provides key to reversing logic: going from sample to population. *This* is useful thing in practice – eg. tells factory manager whether enough of *all* TVs off production line work satisfactorily.

Summary

- May have a population divided into two parts: “success” and “failure” say.
- If:
 - sample size of n trials fixed
 - successes equally likely on each trial
 - trials independentthen number of successes described by binomial distribution.
- Often describes sampling from this population (unless sample is most of population)

- Can simulate from a population, to see what samples from it may look like.
- In a large sample:
 - sample proportion will be close to population proportion
 - distribution of possible sample proportions will look normal
- Later, reverse logic: from sample, draw conclusion about population.

Sampling distribution of sample mean (§5.2)

Counts and proportions describe *categorical* data. *Numerical* data: describe using sample mean, median, SD etc.

Here, concentrate on *sample mean* and how close it might be to population mean. Notation: \bar{x} .

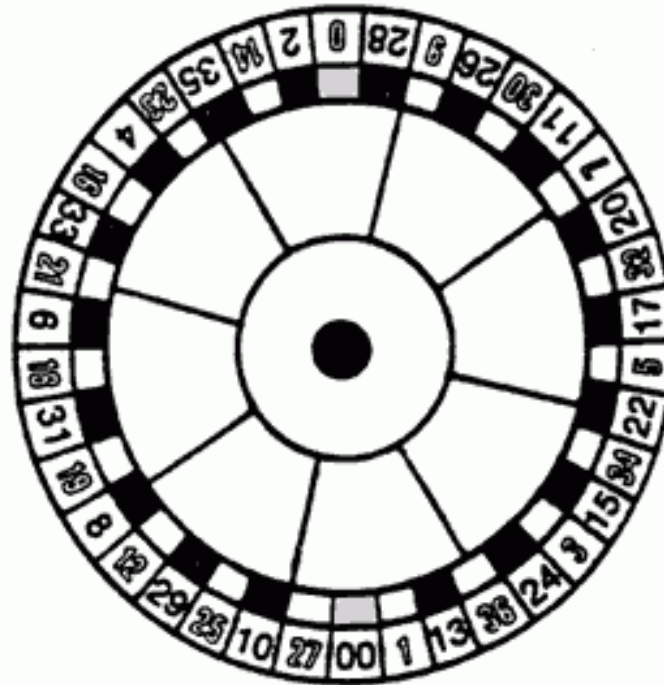
Same attitude as proportions: start from population, see what samples from it look like.

Roulette

Gambling game popular in casinos. Based on wheel with numbers
(from fattonys-roulette.com):

The American Roulette Wheel

The American roulette wheel has 36 numbers and the signs 0 and 00. The American wheels used all over the United States as well as in many locations in the Caribbean islands are, except for the double 00 option, very similar in construction to the French wheels.



Ball spun around wheel, eventually falls into “pocket” next to a number. Bet on single number or combination.

Wheel has numbers 1–36 plus 0 and 00. Bets paid as if 0 and 00 absent (gives casino small edge).

- “High”: win on 19–36, lose otherwise. Win \$1, lose \$1.
- “4 numbers”: win on 7, 8, 9 or 10 (say), lose otherwise. Win \$8, lose \$1.

Simulated roulette in Minitab

Column with numbers 1–36, 0, 00. For each bet, create column with winnings for each number (worksheet `roulette`).

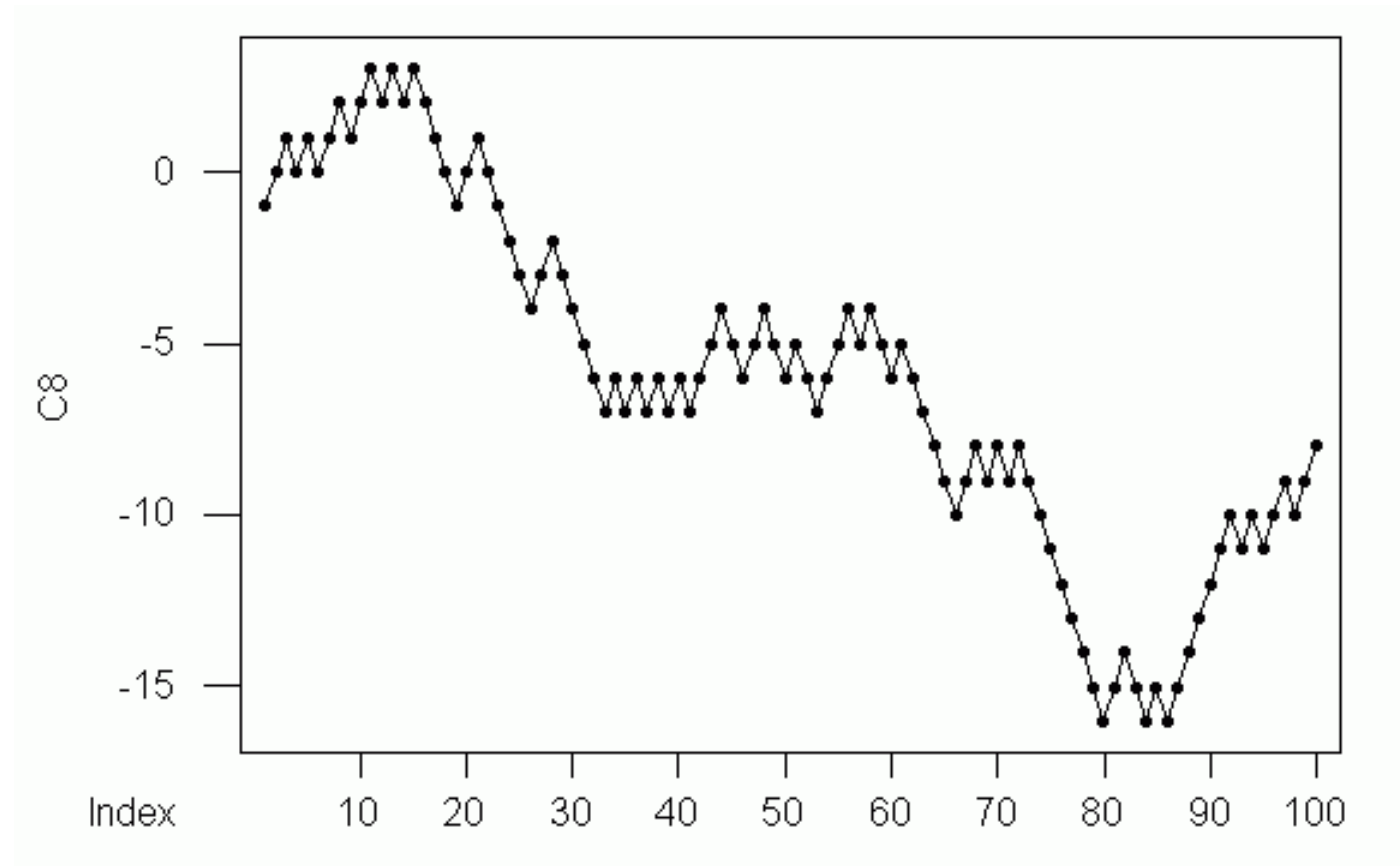
Session window: turn on command language (Editor, Enable Command Language).

Simulate 100 plays of high-low bet. Calc, Random Data, Sample from Columns. Sample from C1 and C2 (to get winning number and winnings). Put results in empty columns, say C6, C7. Check “sample with replacement” box.

See evolution of winnings over time: Calc, Calculator. Store Result in C8. Scan through list of functions to Partial Sums, double-click. In Formula box, see `PARS (number)` . Double-click C7, OK. C8 contains total winnings so far.

Plot this over time: Graph, Time Series Plot. Double-click C8, OK.

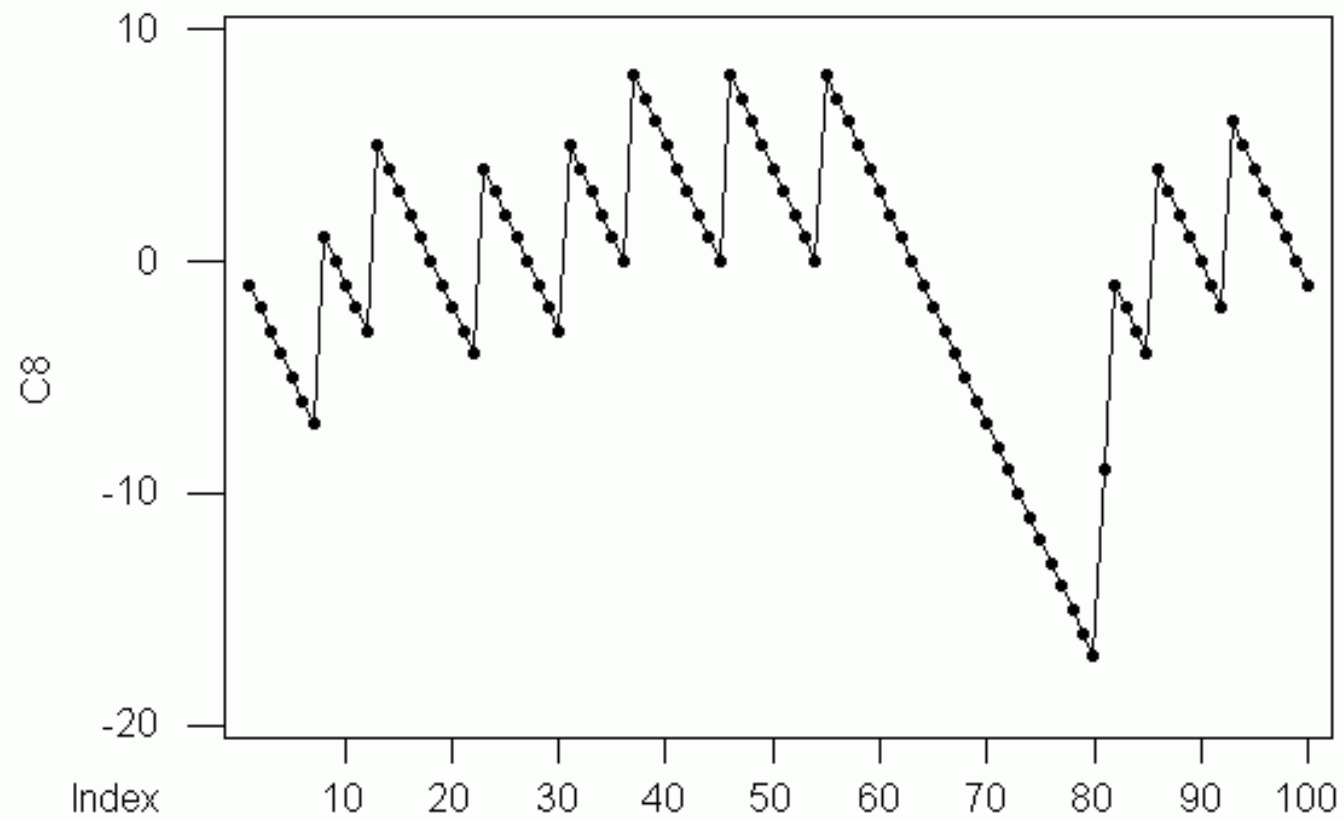
Example run:



More runs: select commands in Session window back to Sample, Edit-Copy. Go to end of Session window (last MTB >), Edit-Paste, Enter. Get another graph.

Try with 4-number bet. Calc, Random Data, Sample from Columns. Sample from C1 and C3, put results in C6, C7 (as before), check With Replacement. Total winnings, plot as before.

Typical picture more “jagged”: lose for some time, occasionally win big.



Here, down \$1 at end, but fluctuates wildly.

Repeated simulations suggest 4-number-bet results “wilder” than for high-low – more likely to come out ahead, but also more likely to lose big.

Mathematical calculations support this. For single plays, work out mean and SD of winnings per play for 2 bets:

Bet	Mean	SD
High-low	−0.0526	0.9986
4-number	−0.0526	2.7620

Sampling distribution of sample mean

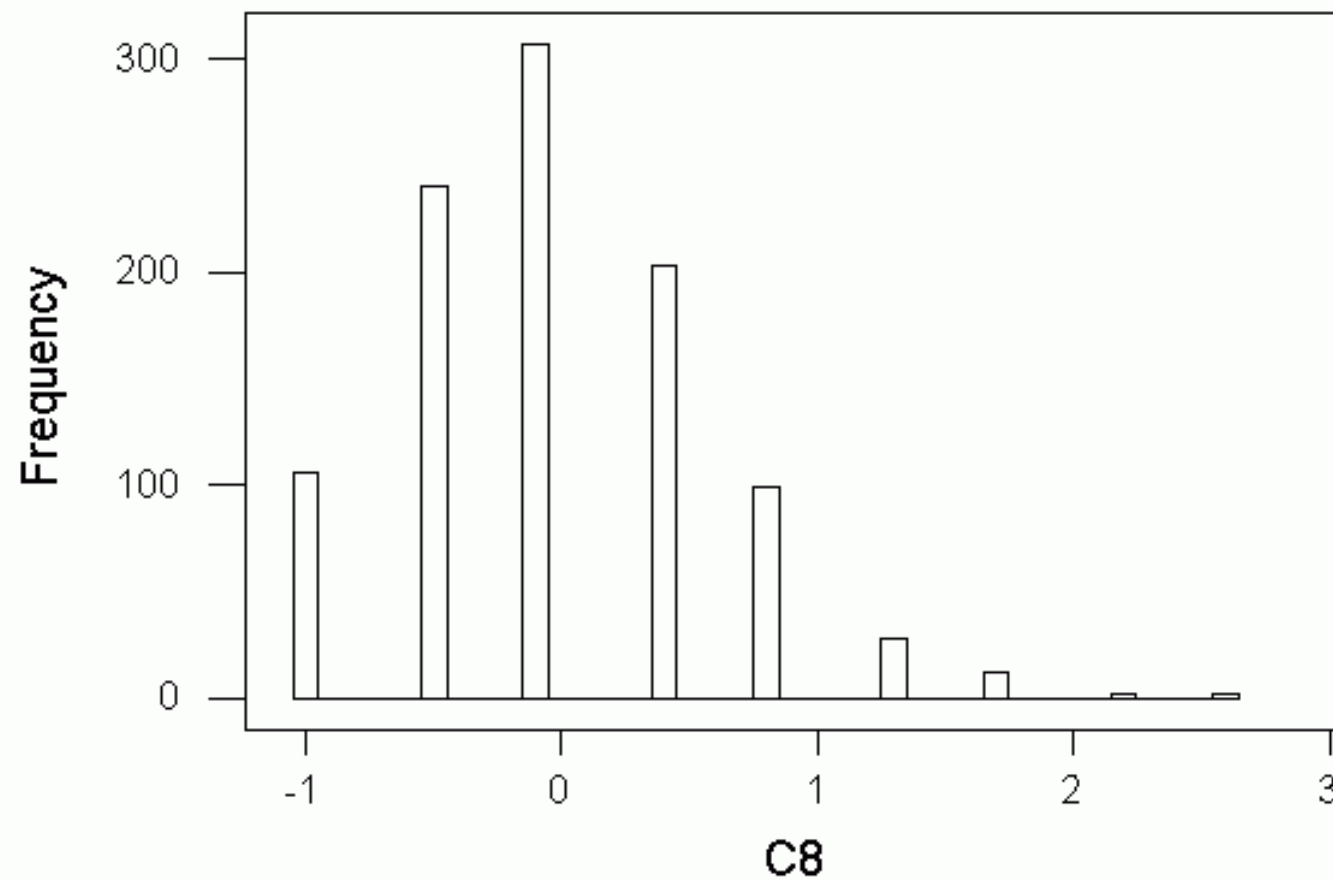
Can think of 100 plays of roulette as *sample* from *population* of “all possible plays”.

Mathematical results. Write population mean as μ (“mu”) and population SD as σ (“sigma”).

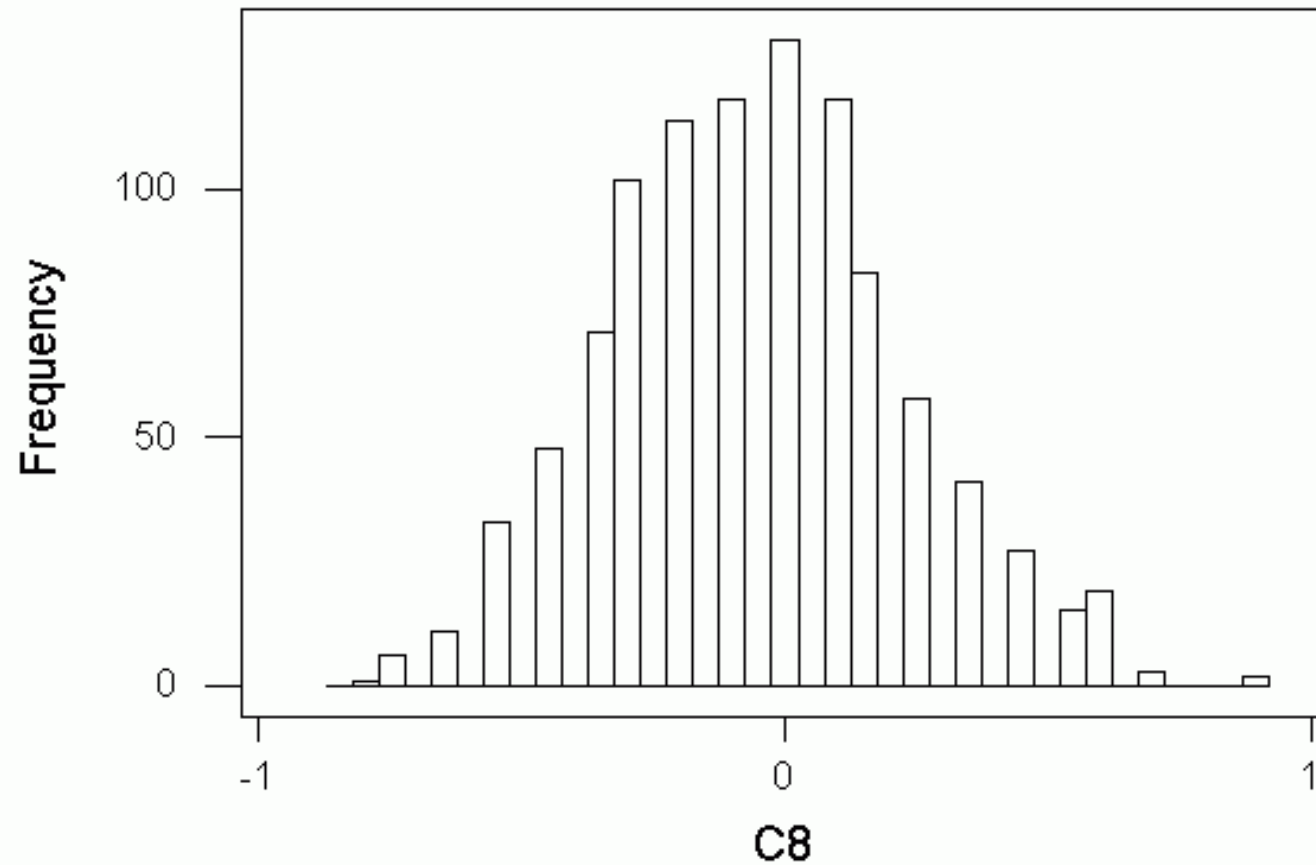
Then sampling distribution of sample mean \bar{x} has:

- mean μ (same as population mean)
- SD σ / \sqrt{n} (*smaller* than population SD).

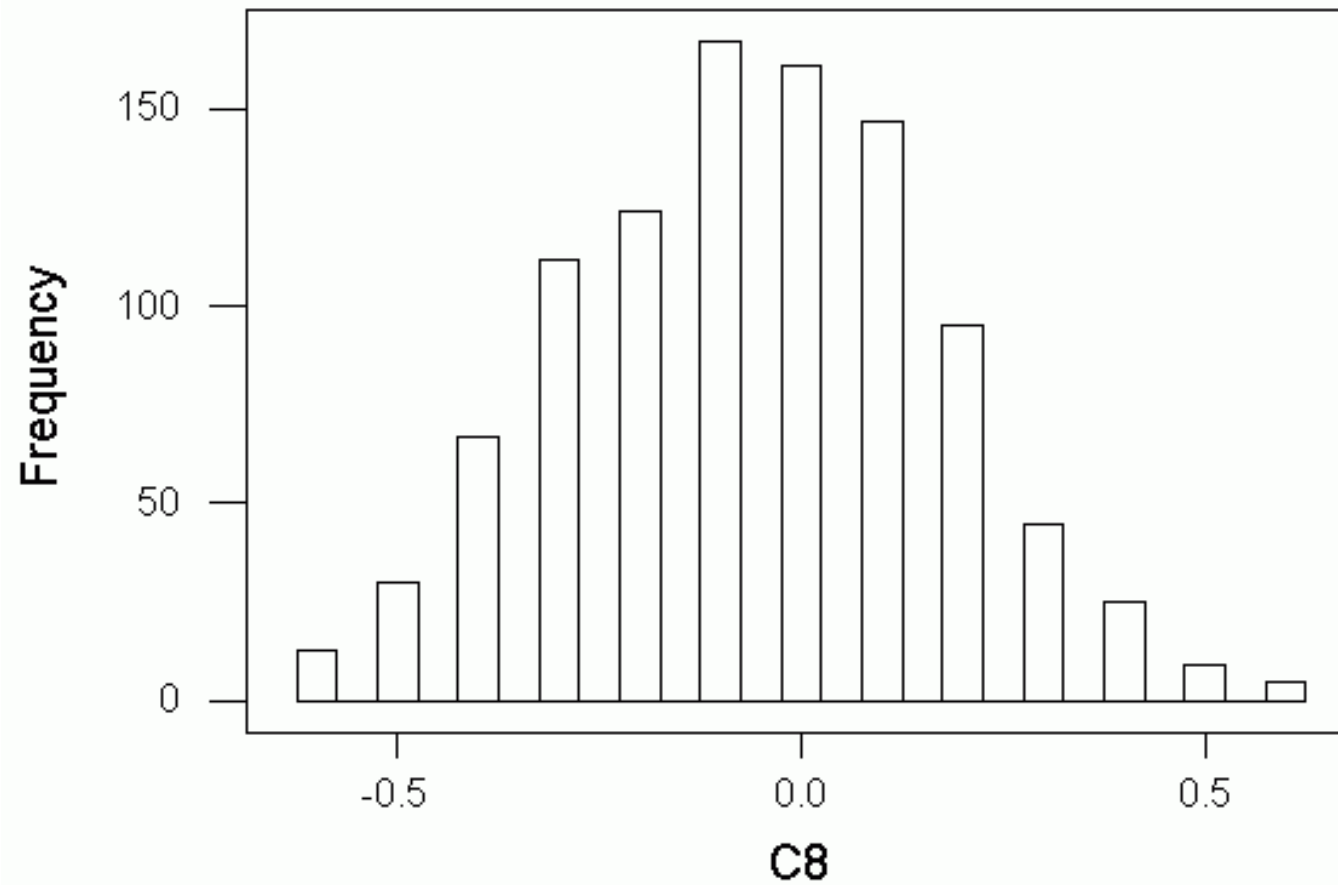
Investigate *shape* by simulation. Depends on sample size – for roulette, times each bet made. 4-number, 20 bets:



Skewed right. 100 bets: less spread, shape like normal:



High-low, 20 bets: also shape like normal:



Central Limit Theorem

So where does normal distribution come from?

Fact: draw simple random sample size n from *any* population. If n large, sampling distribution of \bar{x} has approx. **normal** shape.

Remarkable fact, called **central limit theorem**.

Required largeness of n depends on population – must be large enough to “iron out” skewness. Roulette: high-low population symmetric to start, so $n = 20$ large enough. But 4-number population skewed (small chance to win big), so needed $n = 100$.

Calculations for sample means

Central Limit Theorem allows use of normal distribution to calculate chances, provided sample size n big enough.

First: get mean μ , SD σ of population. (Also mean, SD of *single observation* randomly drawn from population.)

Next: calculate mean, SD of *sampling distribution*.

Finally: use Minitab to complete calculation.

Roulette: high-low bet, 100 plays.

Mean of population -0.0526 , SD of population 0.9986 . Sampling distribution has mean -0.0526 (same), SD $0.9986/\sqrt{100} = 0.09986$.

Chance of profit? Mean winnings > 0 . Calc, Probability Distributions, Normal. Put in mean, SD. Click Input Constant, put in 0.

Result 0.7008 is chance of mean winnings *less than 0* (loss).

Chance of profit $1 - 0.7008 = 0.2992$.

Chance of losing \$0.10 per play or worse? Replace 0 by -0.10 , repeat. Answer 0.3175.

Compare with 4-number bet, 100 plays. Population mean -0.0526 , SD 2.7620 . Sampling distribution mean -0.0526 , SD $2.7620/\sqrt{100} = 0.2762$.

Chance of profit? As before: put in mean, SD, 0. Result 0.5755 , so chance of profit $1 - 0.5755 = 0.4245$.

Chance of losing \$0.10 per play or worse? Change 0 to -0.10 again. Answer 0.4319 .

4-number bet offers bigger chance of profit (0.42 vs. 0.30), but bigger chance of large loss (0.43 vs. 0.32).

Summary

- Understand (sampling) distribution of sample mean from population with mean μ and SD σ .
- Can do by simulation or mathematics.
- Mathematics: sampling distribution of sample mean \bar{x} has mean μ and SD σ/\sqrt{n} .
- For large sample, sampling distribution has normal shape (central limit theorem).
- Enables (approx.) calculations about values of sample mean when population mean, SD known.

Chapter 6: Introduction to Inference

Estimating with confidence (§6.1)

Previous simulations, calculations say that

if I know about the population, I can say what kind of samples I might get from that population. In particular: sample mean and sample proportion have sampling distributions with mean, SD I can calculate, and a normal shape if n large.

But in practice, *don't* know about population; just have one sample from it.

Note logic of this example:

SAT-M scores have SD $\sigma = 100$ points. Take random sample of 500 California high-school seniors, give test to all. Sample mean is 461. What about mean score of *all* California seniors?

- Sample size large. So sampling distribution of sample mean approx. normal, mean μ , SD $100/\sqrt{500} = 4.5$.
- 68-95-99.7 rule: in about 95% of all samples, \bar{x} within 2 times right SD (ie. $2 \times 4.5 = 9$ points) of μ .
- same as saying μ within 9 points of \bar{x} .
- So interval $\bar{x} - 9$ to $\bar{x} + 9$ has μ in it somewhere for 95% of all possible samples.

Example: $\bar{x} = 461$, so interval from $461 - 9 = 452$ to $461 + 9 = 470$. Called **95% confidence interval for μ** .

Understand process:

- Randomness from “all possible samples”. So confidence in *procedure*, not any one answer.
- Cannot know whether our one interval from 95% “good” ones that contain μ , or 5% “bad” ones that don’t.

See figure 6.3 in text:

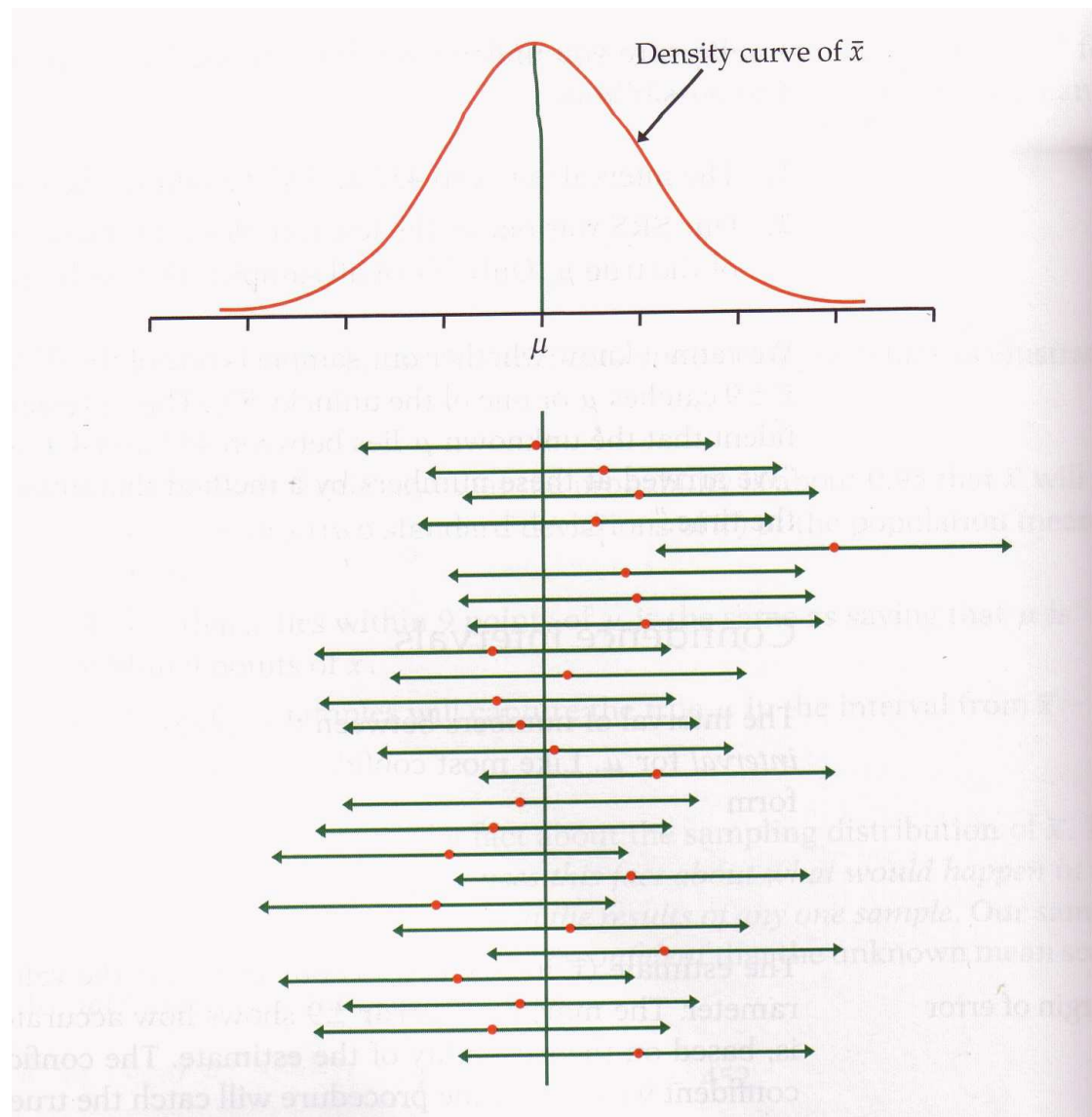


FIGURE 6.2 Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that covers μ .

Confidence intervals

Calculated 95% confidence interval above was $\bar{x} \pm 9$. Typical interval is

$$\text{estimate} \pm \text{margin of error.}$$

Estimate calculated from sample; margin of error expresses accuracy of estimate.

Many different kinds of confidence interval, depending on parameter being estimated, sampling methods etc., but all have:

- two numbers giving lower and upper limits for interval
- “confidence level” (95% above) giving chance that procedure gives interval containing parameter

Confidence interval for population mean

By going up and down 2 times the right SD, got 95% confidence interval (more or less) because of 68-95-99.7.

Can make confidence interval for any level of confidence. Lower confidence level (like 90%) gives shorter interval, but greater chance of interval *not* containing μ .

General formula for confidence interval:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where choose z^* to get confidence level right:

Level	90%	95%	99%
z^*	1.645	1.960	2.576

Example of confidence interval

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.

95% confidence interval uses $z^* = 1.96$:

$$4.63 \pm 1.96 \times 3.7 / \sqrt{90} = 4.63 \pm 0.76,$$

that is from 3.87 to 5.39 days.

99% confidence interval uses $z^* = 2.576$;

$$4.63 \pm 2.576 \times 3.7 / \sqrt{90} = 4.63 \pm 1.00,$$

from 3.63 to 5.63 days.

99% confidence interval bigger than 95%, because have to be more confident in answer (only wrong in 1% of all possible samples).

Confidence interval for mean in Minitab

In practice, use software for calculations. Data in worksheet `hospital`.

Stat, Basic Statistics, 1-sample Z. in Variables, select column (“stay”), ensure Confidence Interval selected, put 95 in Level, put in 3.7 for Sigma:

The assumed sigma = 3.70

Variable	N	Mean	StDev	SE Mean	95.0 % CI
stay	90	4.633	3.785	0.390	(3.869, 5.398)

basically as before.

Repeat for 99%: change Level to 99:

The assumed sigma = 3.70

Variable	N	Mean	StDev	SE Mean	99.0 % CI
stay	90	4.633	3.785	0.390	(3.629, 5.638)

Summary

- Know how far apart sample and population means might be, so can make guess at possible values for population mean based on sample.
- Meaning of eg. 90% confidence interval for pop. mean: in 90% of all possible samples, procedure will give interval containing population mean. (Confidence is in *procedure*).
- General formula for confidence interval for pop. mean, when σ known: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ where choose z^* to get confidence level right.
- 95% CI more likely to contain pop. mean than 90% interval, but interval will be bigger (less precise).

How confidence intervals behave

Size of confidence interval depends on *margin of error* $z^* \sigma / \sqrt{n}$.

Want *small* margin of error – then pinned down parameter precisely.

If margin of error too large:

- use lower confidence level
- reduce population SD σ
- increase sample size n .

Sometimes σ can be reduced by measuring more carefully, but usually only good way to reduce margin of error is larger sample.

Choosing sample size in advance

Previously, used available data and accepted margin of error. But can plan study more carefully: decide on required margin of error, then figure out sample size needed.

In formula, m is margin of error (\pm part of interval):

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

Hospital example: in 95% interval, margin of error 0.76 for $n = 90$, using $\sigma = 3.7$. How many patient records needed to reduce margin of error to 0.5?

Formula:

$$n = \left(\frac{1.96 \times 3.7}{0.5} \right)^2 = 210.4.$$

To be safe, round up: need to sample 211 patient records to get this small a confidence interval.

No free lunch! Sample is over twice as big as original 90, so will cost much more.

Cautions with confidence intervals

- Works for *simple random sample*, not for stratified/multistage samples.
- Sample mean can be affected by *outliers*.
- Sample size large enough to “iron out” skewness etc.
- *Must know population SD σ .*

Last point dealt with in Chapter 7.

Summary

- Margin of error can be made smaller by taking larger sample.
- Can choose sample size required for desired margin of error:
$$n = \left(\frac{z^* \sigma}{m} \right)^2 .$$
- This CI calculation only works for simple random sample with population SD σ known.

Tests of Significance (§6.2)

Confidence intervals answer questions like “What values of population parameter plausible?” Test of significance answers “Is this given value plausible?”

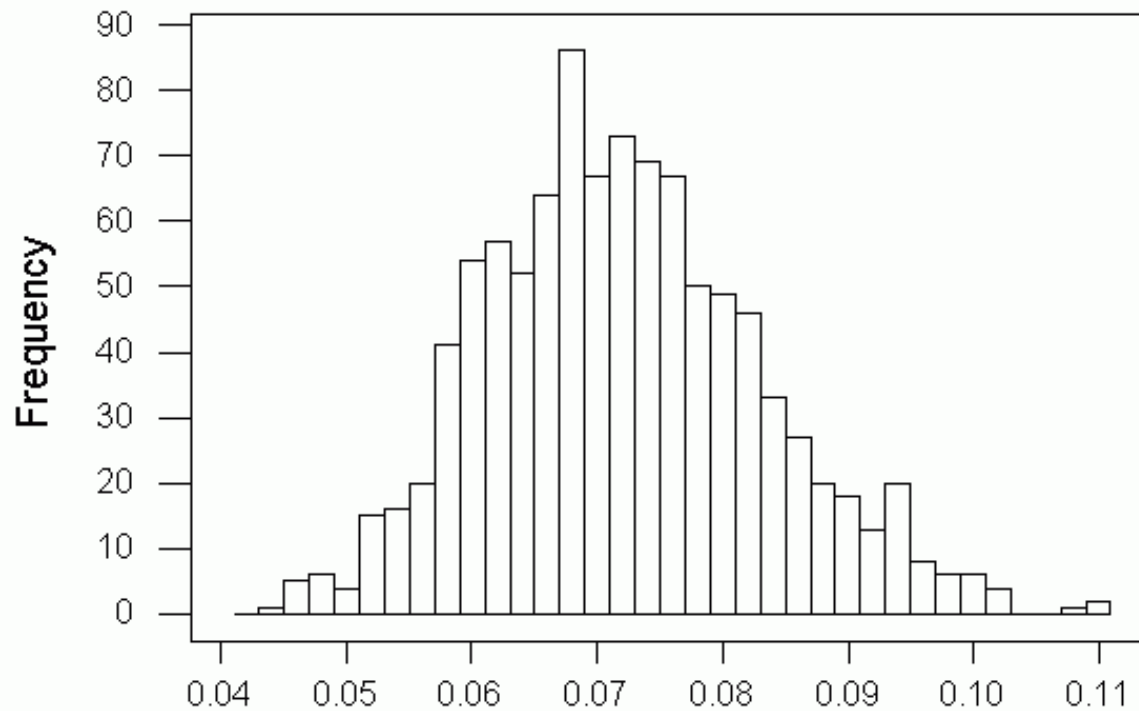
- suppose unemployment rate is 7.2% (as in 2003)
- take random sample of 500 people, get 23 unemployed (4.6%).
- examine kinds of random samples with $n = 500$ and rate 7.2%
- find that 4.6% unusually low in comparison
- conclude unemployment rate changed (such sample rare if population rate 7.2%)

Using sample to *draw conclusion* about population: *particular* population proportion, 7.2%, not plausible.

This logic called **test of significance**.

Might be wrong: is *possible* to get sample proportion 4.6% from population proportion 7.2%, but very unlikely. Always have to accept risk of being wrong.

Sampling distribution of sample proportion if rate 7.2% (simulation):



Chance of observing 4.6% *or more extreme* very low.

Test procedure for population mean μ

- assume that a given value for μ correct, eg. $\mu = 10$. Called **null hypothesis**. Opposite, $\mu \neq 10$, called **alternative hypothesis**.
- Take random sample from population, calculate \bar{x} .
- Find chance of statistic value as extreme or more extreme, *if null hypothesis true*. Called **P-value**.
- If P-value small, **reject** null hypothesis in favour of alternative. If not, **do not reject** null hypothesis.

Analogy: court of law

Accused innocent until proven guilty	Null hypothesis assumed true until proven wrong
Found guilty only with strong evidence	Reject only if P-value small
Innocent	Null hypothesis true
Guilty	Null hypothesis false
Found not guilty	Do not reject null hypothesis
Found guilty	Reject null hypothesis

Legal		Decision	
		Find not guilty	Find guilty
Truth	Innocent	Correct	Serious error
	Guilty	Error	Correct

Testing		Decision	
		Not reject null	Reject null
Truth	Null hyp. true	Correct	Type I error
	Null hyp. false	Type II error	Correct

Stating hypotheses

Example: Calcium levels in blood of healthy young adults vary with mean $\mu = 9.5$ mg/dl and SD $\sigma = 0.4$ mg/dl. Clinic in rural Guatemala measured blood Ca levels of 180 pregnant women; sample mean $\bar{x} = 9.58$.

Trying to prove, based on sample, that something different about these women. Thing “trying to prove” is *alternative* hypothesis: in symbols, $\mu \neq 9.5$ for these women.

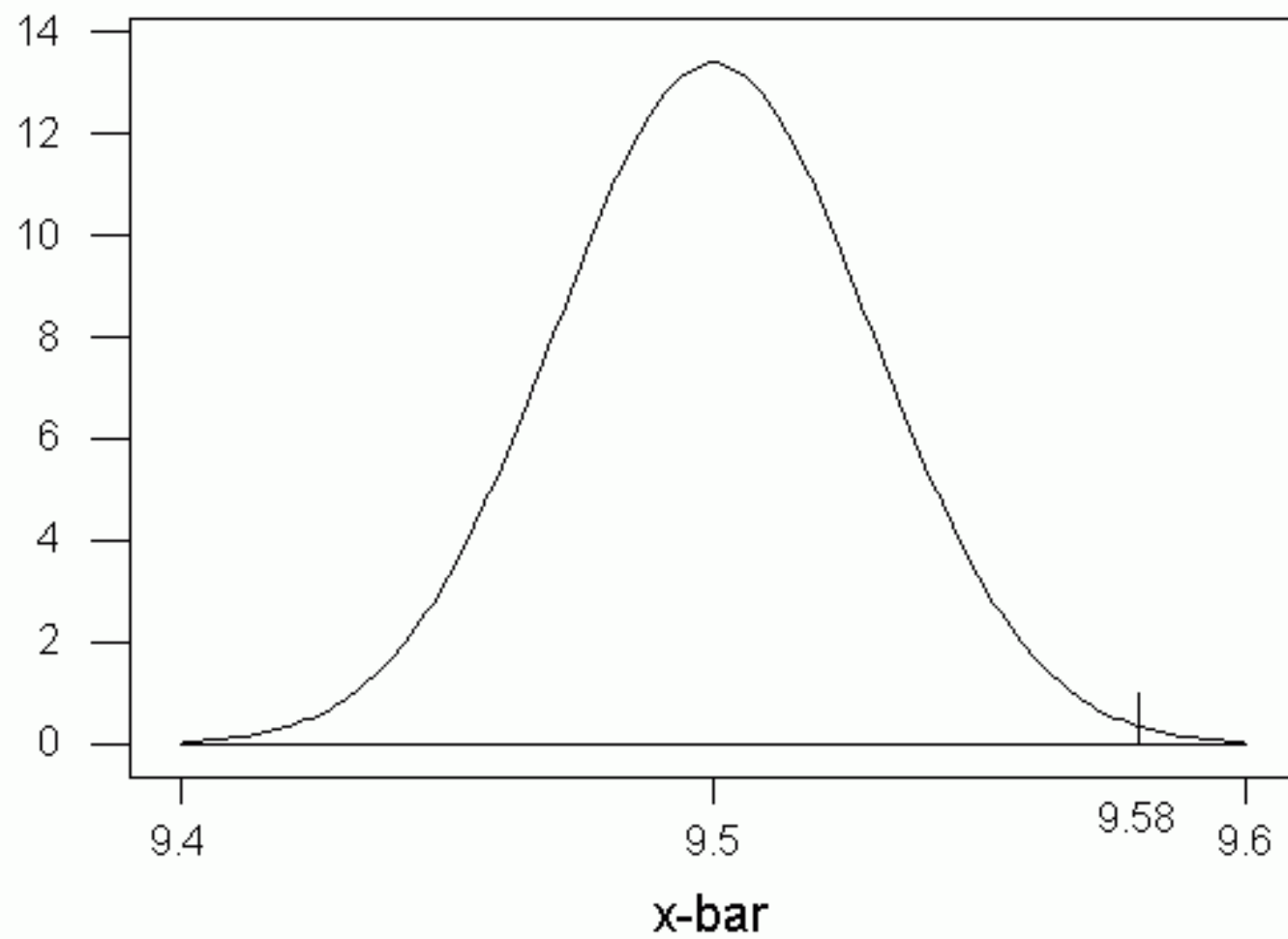
This alternative hypothesis **two-sided**. $\mu > 9.5$ would have been **one-sided**.

Other conclusion is that these women same as healthy young adults generally. “Same” is *null* hypothesis, $\mu = 9.5$.

Sampling distribution

Ask: if null hypothesis true, is sample mean of 9.58 typical or untypical?

To answer, look at *sampling distribution* of sample mean. If null hypothesis true, sampling distribution has mean $\mu = 9.5$, SD $\sigma/\sqrt{n} = 0.4/\sqrt{180} = 0.0298$ and approx. normal shape (because of large sample).



9.58 is a long way out.

P-values

How far out is “far out”?

Idea: find chance of being *as extreme or more extreme* in sampling distribution from null hypothesis.

“Extreme” means: higher if sample statistic is high (as in example: 9.58 vs. 9.5), lower if sample statistic low. (Drawing picture often helpful.)

In example: want chance of being higher than 9.58 in normal distribution, mean 9.5, SD $0.4/\sqrt{180} = 0.0298$.

Same as §1.3:

$$z = \frac{9.58 - 9.5}{(0.4/\sqrt{180})} = 2.68,$$

then Table A gives prob. 0.9963 below, 0.0037 above.

Test two-sided, so P-value $2 \times 0.0037 = 0.0074$.

z here called **test statistic**: stepping stone to getting P-value.

Statistical significance

Reject null hypothesis if P-value “small”. But how small?

Compare P-value to number α (alpha) chosen in advance.

$\alpha = 0.05$ common choice. If P-value smaller, result called **statistically significant at level α** .

“Significant” here means only “evidence against null hypothesis reaches stated level” – not “significant” in sense of “important”.

Statement of conclusion in example: Result statistically significant (P=0.0074).

Choice of α reflects strength of evidence against null hypothesis regarded as convincing. Case-by-case.

Small α like 0.01 requires strong evidence. Hard to reject null hypothesis when true (good), but also hard to reject when false (bad).

Large α like 0.10: Often reject null hypothesis when true (bad), but also often reject when false (good).

No free lunch!

Another example

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.

Researcher wants $\mu > 10$, so is alternative hypothesis. Null hypothesis $\mu = 10$ (always “=”).

Work as if null hyp. true – Sampling distribution of \bar{x} normal, mean 10, SD $5/\sqrt{49} = 0.714$.

Where does 10.94 come on this distribution? (Can draw picture.)

Want chance of being *greater than* because alt. hyp. has $>$.

Test statistic is

$$z = \frac{10.94 - 10}{(5/\sqrt{49})} = 1.32;$$

prob. of being below this (Table A) is 0.9066, so P-value is

$1 - 0.9066 = 0.0934$. (Don't double.)

Not statistically significant at 0.05 level; *do not reject* null hyp.

Sample mean this high in about 10% of samples if population mean really 10.

Have *not* proved that population mean GHQ is 10 – have only shown that it could be 10 (but could be many other values as well).

(Confidence interval for μ here: 9.54 to 12.34. Anything in here “plausible” value for μ .)

One and two-sided P-values

- If alternative has \neq (2-sided):
 - if $z > 0$, P-value is twice prob. of greater
 - if $z < 0$, P-value is twice prob. of less
- If alternative has $<$ (1-sided):
 - if $z > 0$, P-value is large: do not reject null. “Wrong side”.
 - if $z < 0$, P-value is prob. of less.
- If alternative has $>$ (1-sided):
 - if $z > 0$, P-value is prob. of greater
 - if $z < 0$, P-value is large: do not reject null. “Wrong side”.

Tests of significance in Minitab

When have data, can get Minitab to do all the calculations. In worksheet ghq, 49 scores in column C1.

Stat, Basic Statistics, 1-sample z (again). Select C1, click Test Mean. Change 0.0 to 10, change Alternative to Greater Than. Put in value 5 for sigma. Click OK:

```
Test of mu = 10.000 vs mu > 10.000
The assumed sigma = 5.00
```

Variable	N	Mean	StDev	SE Mean	Z	P
ghq	49	10.939	5.113	0.714	1.31	0.095

Minitab's P-value 0.095, conclusion as before.

Summary

- Is given population parameter value plausible?
- What we try to prove called alternative hypothesis (accused is guilty).
- Status quo called null hypothesis (always contains $=$). (accused innocent)
- Calculate P -value: chance of result as or more extreme, if null hypothesis true. Small P -value, reject null hypothesis, otherwise don't reject.
- To decide “small”, choose α before doing test. Reject null if P -value smaller than α . Often choose $\alpha = 0.05$.

Confidence intervals and tests of significance

Confidence interval: *what values are plausible?*

Significance test: *is this value plausible?*

Related ideas. Suggests: pick null hypothesis value for μ . If this null hypothesis not rejected (“plausible”), value inside confidence interval. If null hypothesis rejected, value outside.

True, provided: (a) doing 2-sided test, (b) match up α and confidence level. (Eg. for 95% interval, use $\alpha = 0.05$.)

Weights (kg) of 24 male runners. In worksheet
runner_weight. Use $\sigma = 4.5$ kg, $\alpha = 0.05$.

Suppose null hypothesis is $\mu = 59$ (against $\mu \neq 59$):

Test of mu = 59.000 vs mu not = 59.000

The assumed sigma = 4.50

Variable	N	Mean	StDev	SE Mean	Z	P
weight	24	61.792	4.808	0.919	3.04	0.0024

Null hypothesis rejected. 59 outside confidence interval.

Now try null hyp. $\mu = 62$:

Test of mu = 62.000 vs mu not = 62.000

The assumed sigma = 4.50

Variable	N	Mean	StDev	SE Mean	Z	P
weight	24	61.792	4.808	0.919	-0.23	0.82

Null hyp. not rejected. 62 inside confidence interval.

Finally, null hyp. $\mu = 64$:

Test of $\mu = 64.000$ vs $\mu \neq 64.000$

The assumed sigma = 4.50

Variable	N	Mean	StDev	SE Mean	Z	P
weight	24	61.792	4.808	0.919	-2.40	0.016

Null hyp. rejected, 64 outside interval.

To check this, calculate 95% confidence interval (to go with $\alpha = 0.05$):

The assumed sigma = 4.50

Variable	N	Mean	StDev	SE Mean	95.0 % CI
weight	24	61.792	4.808	0.919	(59.991, 63.592)

As predicted, 62 inside, 59 and 64 outside.

P-value for testing $\mu = 64$ vs. $\mu \neq 64$ was 0.016. If used $\alpha = 0.01$, would not reject. So 64 should be *inside* 99% confidence interval:

The assumed sigma = 4.50

Variable	N	Mean	StDev	SE Mean	99.0 % CI
weight	24	61.792	4.808	0.919	(59.425, 64.158)

as indeed it is.

Use and abuse of tests (§6.3)

Calculations for tests simple (with software), but wise use of tests difficult. Each test valid in certain circumstances with certain assumptions.

In testing a mean, as we have done:

- Works for *simple random sample*, not for stratified/multistage samples.
- Sample mean can be affected by *outliers*.
- Sample size large enough to “iron out” skewness etc.
- *Must know population SD σ .*

same as for confidence intervals.

Strength of evidence vs. decision-making

Best way to give result of significance test is to give *P-value*. Shows strength of evidence against null hyp. Enables reader to judge whether evidence “strong enough”.

To use test to make decision, must choose α *before looking at data*. Choice depends on consequences of wrong decision. If rejecting null in favour of alternative expensive, need strong evidence to reject null. Subjective, extra-statistical.

$\alpha = 0.05$ often used, reasons mainly historical. No real difference between P-values 0.049 and 0.051.

Statistical and practical significance

Example: test $\mu = 20$ vs. $\mu \neq 20$ using sample $n = 10,000$, $\sigma = 0.2$. $\bar{x} = 20.005$: P-value 0.012. If $\alpha = 0.05$, *reject null hyp.* even though sample mean very close to 20.

Difference between 20 and 20.005 very small, maybe of no practical relevance. P-value says only that with this big a sample, sample mean of 20.005 unusually high.

Statistical significance not same as practical significance.

Other way around too: with small sample, can get sample mean far from null hyp. but not statistically significant.

Don't ignore lack of significance

Tests with large P-values also important, even though not statistically significant.

Typical science: researcher has new theory about some effect (alternative hyp.) tested against existing theory (null). Tries to gather evidence against null.

Suppose now P-value not small. Could indicate theory wrong, or flaw in experiment. If theory plausible, result worth knowing about.

However, tendency in scientific work only to publish statistically significant results.

Statistical inference not always valid

Badly designed surveys/experiments give invalid results – cannot rescue by clever analysis.

Need to use proper experimental design and appropriate analysis, with right kind of randomization.

But also face data not from experiments. Eg. diameters of holes bored in engine blocks in car-making. Check whether reasonable to treat as independent observations from normal distribution using descriptives (graphs, numbers).

In general, *learn how data produced*, assess whether test/interval meaningful.

Searching for significance

Right way: decide on effect sought, design experiment to assess effect, significance test on results.

But tempting: collect a bunch of data, do a bunch of tests, see what is significant.

Example: industrial process with desired mean 50, SD 10. Take samples size 50 each day; test $\mu = 50$ vs. $\mu \neq 50$.

Results:

Test of $\mu = 50.00$ vs $\mu \neq 50.00$

The assumed sigma = 10.0

Variable	N	Mean	StDev	SE Mean	Z	P
C1	50	49.69	11.35	1.41	-0.22	0.83
C2	50	52.21	8.39	1.41	1.56	0.12
C3	50	51.39	8.71	1.41	0.99	0.32
C4	50	50.49	9.61	1.41	0.35	0.73
C5	50	49.43	9.50	1.41	-0.40	0.69
C6	50	49.02	10.91	1.41	-0.69	0.49
C7	50	52.55	8.64	1.41	1.80	0.072
C8	50	52.88	8.30	1.41	2.04	0.042
C9	50	48.25	10.20	1.41	-1.24	0.22
C10	50	49.55	7.81	1.41	-0.32	0.75

Result on 8th day (C8) significant at $\alpha = 0.05$. What happened?

Answer: nothing. Data randomly generated from normal distribution mean 50, SD 10. Result only of doing many tests.

Summary

- Value for parameter inside confidence interval — value plausible — null hyp. of that value wouldn't be rejected. True as long as test 2-sided and α , confidence level match up.
- Tests come with assumptions. Here: simple random sample, mean appropriate for “centre”, sample size large enough, must know σ .
- P-value measures strength of evidence (smaller = stronger). Use for decision-making by choosing α first and rejecting null if $\text{P-value} < \alpha$.
- Statistical significance not same as practical significance.

- Lack of (statistical) significance may mean plausible theory incorrect.
- Statistical inference depends on correct design, randomization, analysis. Can sometimes treat data “as if” random sample.
- Doing many tests will probably produce some significant results by chance.

Power (§6.4)

Two types of error in significance testing:

- Type I: rejecting null hyp. when true
- Type II: not rejecting null hyp. when false

Concentrated mostly on Type I (by choosing α), but type II also important. Want to design experiment that has some chance of detecting desired effect – a low chance of type II error, or high **power** — a high chance to *correctly* reject the null.

Can sometimes figure power by calculation, but usually easier to simulate. (Minitab can also do power exactly – later.)

Power example

Consider text example 6.28 (p. 431). Researchers testing whether exercise program increases total bone mineral content (TBBMC) in young women over 6-month period. Based on previous data, prepared to assume population SD is 2. Intend to use $n = 25$ subjects.

How likely are they to detect change of 1 unit in TBBMC? That is, if actually $\mu = 1$ and go through test procedure, how often would they reject null with P-value < 0.05 ?

Simulate data from *true* population (mean 1, SD 2). Carry out test for each simulated sample, see how many times null rejected.

Minitab: do 50 simulated tests: Calc, Random Data, Normal.

Generate 25 rows of data, store in C1–C50 (type into box). Mean 1, SD 2. Click OK. Then Stat, Basic Statistics, 1-sample Z. Variables C1–C50, test mean 0, alternative “greater than”, sigma 2.

Output like this:

Test of $\mu = 0.000$ vs $\mu > 0.000$

The assumed sigma = 2.00

Variable	N	Mean	StDev	SE Mean	Z	P
C1	25	1.236	1.671	0.400	3.09	0.0010
C2	25	0.675	2.057	0.400	1.69	0.046
C3	25	0.319	2.449	0.400	0.80	0.21
C4	25	0.965	2.469	0.400	2.41	0.0080
C5	25	1.123	1.982	0.400	2.81	0.0025
C6	25	0.743	1.749	0.400	1.86	0.032
C7	25	0.781	1.825	0.400	1.95	0.026
C8	25	0.311	1.747	0.400	0.78	0.22
...						

6 rejections in 8 tests shown; got 39 rejections in 50 tests, estimate power $39/50 = 0.78$.

Minitab does power directly: Stat, Power and Sample Size, 1-sample Z. Select “Calculate power for each sample size”, fill in sample size 25. “Difference” is difference between null hyp. and true means, here 1 (put in). Put in 2 for sigma (bottom). Click Options, select Greater Than for alternative. Click OK:

1-Sample Z Test

Testing mean = null (versus > null)

Calculating power for mean = null + 1

Alpha = 0.05 Sigma = 2

Sample

Size	Power
25	0.8038

Difference detectable in 80% of all possible samples. (Simulation close.) Researchers have good chance of success.

Determining sample size

At beginning of study, need to know how many subjects/experimental units to use. Common way of deciding: choose power for alternative of interest, then use software to find sample size making this work.

Suppose researchers of previous example actually wanted power 0.90 to detect an increase of 1 unit in TBBMC. Would need bigger sample than 25, but how big? Minitab: Stat, Power and Sample Size, 1-sample Z as before. Select “Sample size for each power”, fill in 0.90 for power; 1 for difference, 2 for sigma (as before). In Options, make sure Greater Than selected for alternative.

Results:

1-Sample Z Test

Testing mean = null (versus $>$ null)

Calculating power for mean = null + 1

Alpha = 0.05 Sigma = 2

Sample Size	Target Power	Actual Power
35	0.9000	0.9054

35 subjects are needed. Actually, 35 subjects gives power $>$ 0.90, but 34 subjects gives power $<$ 0.90, so 35 safe.

Summary

- Power is chance of rejecting null hypothesis when it is false.
Want power to be high.
- Study power by deciding what change is important to detect, and then finding chance of detecting that change.
- Can simulate: assume important change has happened (change in μ), simulate from that population, do original test, see how often original null rejected.
- Can calculate power directly with Minitab.
- Before study, can determine sample size needed to obtain a specified power against a specified alternative.

Chapter 7: Inference for Distributions

Inference for Population Mean (§7.1)

In Chapter 6, did tests and confidence intervals for population mean μ , but needed to know population SD σ – not realistic!

Can always calculate *sample* SD. Use symbol s .

Obvious remedy: use sample SD in place of population SD.

Investigate confidence intervals by simulation.

Minitab: generate 100 random samples of size $n = 5$ *in rows*. Calc, Random Data, Normal. Generate 100 rows, store in C1-C5, mean 10, SD 3.

Calculate sample means: Calc, Row Statistics. Click Mean under Statistic, select columns C1-C5, store results in C7. Do sample SDs same way, selecting Standard Deviation and storing in C8.

Now calculate confidence intervals “by hand”. First get “margin of error” for 95% interval as $1.96 \times s/\sqrt{5}$: Calc, Calculator. Store result in C9, in Expression box type $1.96 * C8 / \text{sqrt}(5)$ (or find “square root” in the list of functions). Click OK.

Ends of confidence interval for each sample are mean minus margin of error and mean plus margin of error, ie. $C7 - C9$ and $C7 + C9$.

Use Calculator again, store in C11, C12.

My results (selected):

Row	x-bar	s	+/-	C10	C11	C12
1	10.1868	2.23416	1.95833		8.2285	12.1451
2	10.1366	1.47570	1.29351		8.8430	11.4301
3	8.9175	1.49579	1.31112		7.6064	10.2286
12	7.1230	2.77318	2.43080		4.6922	9.5538
13	10.1301	2.28851	2.00597		8.1242	12.1361
31	12.7054	2.02251	1.77281		10.9326	14.4782
32	11.0099	2.66603	2.33688		8.6730	13.3468

Population mean 10, so confidence intervals should contain 10.

Those from rows 1,2,3,13,32 do, those from 12 and 31 don't.

In my simulation, only 89 of 100 simulated intervals contain 10.

Supposed to be 95% interval, but only about 89%.

The t distributions

Simulations showed that using Chapter 6 procedure with s not σ gave intervals that are *too short*.

Previously got margin of error using 1.96 for 95% interval, 1.645 for 90% interval. Idea: make these numbers bigger to account for lack of knowledge of σ .

- Large sample: sample SD s probably close to σ . Loss of information small, adjustment small.
- Small sample: sample SD s might be far from σ . Loss of information large, adjustment large.

So for 95% confidence interval, margin of error calculated with something other than 1.96, but exact number depends on sample size.

Correct number comes from **t-distribution**. Distribution different for each sample size; traditionally labelled by **degrees of freedom (df)**, which is $n - 1$.

Get number from *table*, inside back cover of book:

TABLE D *t* distribution critical values

df	Upper tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
<i>z</i> *	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

For sample size $n = 5$, df $5 - 1 = 4$. Find 95% along bottom of table, look up to 4 df, get 2.776. Much bigger than 1.96.

Now, try $n = 51$. 50 df row, 95% column, 2.009. Much nearer 1.96.

With a *very* large sample, t^* is very close to 1.96, eg. with 1000 df.

Example

When people buy bicycles, they often buy other accessories too (helmet, water bottle, etc.) A store took a random sample of 12 bike-buying customers, and found accessory purchases to be (in \$) 38, 65, 82, 114, 77, 19, 142, 93, 63, 107, 58, 76.

Sample mean $\bar{x} = 77.83$, sample SD $s = 33.51$. $n = 12$, so 11 df. For 95% confidence interval, number from table is 2.201. So interval is

$$77.83 \pm 2.201 \frac{33.51}{\sqrt{12}},$$

from \$56.50 to \$99.12.

t confidence intervals in Minitab

In practice, we have (or can get) the data in Minitab. Then get Minitab to do the whole calculation. I entered the 12 numbers into column C1. Then: Stat, Basic Statistics, 1-sample t. Select column C1, ensure Confidence Interval checked and 95.0 in box. Click OK. (Don't need to give "sigma".) Results:

T Confidence Intervals

Variable	N	Mean	StDev	SE Mean	95.0 % CI
C1	12	77.83	33.51	9.67	(56.54, 99.13)

a more accurate version of our calculation.

Summary

- When σ not known, using procedure of Chapter 6 with s in place of σ gives confidence intervals that are too short.
- Adjustment depends on sample size. Use t distribution with df $n - 1$.
- Confidence interval for μ now $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$, where t^* from table.

Significance tests based on t distribution

To continue bike-accessories example: at similar bike shop, mean accessory sales are \$90 per bike sold. Is there evidence that accessory sales at this shop less?

Significance test. Population mean sales μ . Alternative $\mu < 90$, null $\mu = 90$.

Cannot do using previous test procedure because *don't know* σ . So calculate test statistic t using sample SD, and get P-value from Table D.

Here, $\bar{x} = 77.83$, $s = 33.51$. So:

$$t = \frac{77.83 - 90}{(33.51/\sqrt{12})} = -1.26.$$

P-value is prob. of being below -1.26 . But table only gives *positive* values of t . t distribution is symmetric, like normal, so P-value also prob. of being *above* 1.26 .

Go to table with $12 - 1 = 11$ df. Look along 11 df row, and find where 1.26 is – between 1.088 and 1.363 . Look to very top row for those values: P-value between 0.10 and 0.15 .

Approximate, but accurate enough: not small, so do not reject null.
The population mean could be 90 .

1-sided and 2-sided P-values

- If alternative has \neq (2-sided):
 - if $t < 0$, ignore minus sign
 - then P-value is twice what you get from Table D.
- If alternative has $<$ (1-sided):
 - if $t > 0$, P-value is large: do not reject null. “Wrong side”.
 - if $t < 0$, ignore minus sign, get P-value from Table D.
- If alternative has $>$ (1-sided):
 - if $t > 0$, get P-value from Table D.
 - if $t < 0$, P-value is large: do not reject null. “Wrong side”.

Minitab: data in column C1. Stat, Basic Statistics, 1-sample t (as for interval). Click Test Mean, fill in 90. Change Alternative to “less than”. Click OK:

T-Test of the Mean

Test of $\mu = 90.00$ vs $\mu < 90.00$

Variable	N	Mean	StDev	SE Mean	T	P
acc_sold	12	77.83	33.51	9.67	-1.26	0.12

P-value 0.12 not small, cannot reject null hypothesis. No evidence that mean sales less than \$90. (Note that 0.12 consistent with “between 0.10 and 0.15” from before.)

Matched pairs

When designing experiment, *comparative* usually better (treatment vs. control etc.) – protects against other variables affecting result.

Two ways to compare:

- **Matching:** pair each subject with another similar subject. One subject gets treatment, other control. Alternative: get “before”, “after” measurements from each subject.
- **Randomizing:** divide subjects into 2 large groups at random. One group gets treatment, other control.

Think about matched pairs.

Measurements come in pairs: each “before” with particular “after”, each “treatment” with particular “control”.

Example: “Apnea” is brief stoppage of breathing during sleep. 13 premature infants given drug; measure “apneic episodes per hour” before and after drug given. Data like this:

Row	before	after
1	1.71	0.13
2	1.25	0.88
10	0.67	0.75
13	1.96	1.13

All infants except #10 had fewer episodes after.

Idea: *differences*, after minus before. Write μ for population mean difference. Trying to prove drug *reduces* apnea, so alternative hyp. $\mu < 0$, null $\mu = 0$.

Test just 1-sample t -test on differences. Get 90% confidence interval too.

Minitab: Stat, Basic Statistics, Paired t. Want after minus before, so `a f t e r` is 1st sample, `b e f o r e` 2nd (illogical!) Click Options, change Confidence Level to 90, Alternative to “less than”.

Paired T for after - before

	N	Mean	StDev	SE Mean
after	13	0.984	0.833	0.231
before	13	1.751	0.855	0.237
Difference	13	-0.767	0.524	0.145

90% CI for mean difference: (-1.026, -0.508)

T-Test of mean difference = 0 (vs < 0): T-Value = -5.28

P-Value = 0.000

P-value “0.000” very small, so reject null hypothesis; real decrease in apnea. Same infants before and after, so decrease due to drug.

With 90% confidence, drug reduces apnea by between 0.5 and 1 episodes per hour. (Gives idea of *size* of drug effect.)

Robustness of t procedures

Mathematics of t procedures assumes that population has normal-distribution shape (and therefore that mean, SD measure centre, spread).

But in practice, shape must be badly non-normal to cause problems, particularly if sample large. More important to have simple random sample. Guidelines:

- $n < 15$: don't use t if clearly non-normal shape or outliers.
- $15 \leq n < 40$: beware only of outliers or strong skewness.
- $n > 40$: unlikely to be any trouble.

Assess using pictures (histogram, normal quantile plot).

For matched pairs, only *differences* matter; original variables can have any shape.

If an inference procedure doesn't depend much on its assumptions (as here), called **robust**.

What to do if you can't use the t procedures?

Two main approaches to choose from:

- *transform* the data so that the distribution is more nearly normal.
In Example 7.10 in text (p. 466), analysis is based on logarithms of data values because data very right-skewed.
- *use a different test*, such as those in Chapter 15, which do not use the distribution shape at all (but not such good tests if the distribution shape is close to normal).

Power and sample size for t tests

How likely is t -test to reject null hyp. when it is false?

Mathematics complicated (needs to account for replacing σ by s), but can use simulation, or Minitab.

Return to apnea example. With sample of 13 infants, how likely is detection of decrease of 0.25 episodes/hour?

Minitab: Power and sample size, Calculate power for each sample size. Fill in sample size 13, difference -0.25 (lower better). Need value for σ ; use our sample SD 0.524. Click Options, select alternative “less than”.

Results:

1-Sample t Test

Testing mean = null (versus < null)

Calculating power for mean = null - 0.25

Alpha = 0.05 Sigma = 0.524

Sample

Size	Power
------	-------

13	0.4910
----	--------

Only about 50-50 chance to detect this size difference.

How big a sample size would be needed to make this power 0.8?

Power & sample size, 1-sample t as before. Click Calculate Sample Size. Fill in power 0.80, difference -0.25 . Check sigma correct, check alternative (via Options):

Testing mean = null (versus $< \text{null}$)

Calculating power for mean = null - 0.25

Alpha = 0.05 Sigma = 0.524

Sample Size	Target Power	Actual Power
29	0.8000	0.8055

Need 29 infants to have this chance of detecting a 0.25 decrease.

Summary

- Significance test when σ unknown: replace σ by s , use t distribution to get P-value. In practice, use software.
- With before-after measurements on same individual (measurements on paired individuals), matched-pair analysis: t procedures on differences (null: mean difference is 0).
- In practice, with reasonably large sample(s), can use t procedures with most data (even though mathematics assumes data normal). Otherwise: transformation, sign test.
- When null hypothesis known to be false, can find chance of (correctly) rejecting it: power. Can also find sample size needed to obtain desired power against given alternative.

Comparing two means (§7.2)

Earlier: best experiments *comparative*. Compare by matching or by randomizing.

Example: teaching reading. Compare new method to standard method. Can't teach a child by both methods, so: take random sample of 20 children, choose at random 8 children for new method (other 12 get standard method).

Children taught by qualified people for 6 months, given reading test at end. Record test score.

Two *separate* samples, no pairing. Needs different analysis.

Two populations. (Reading example: “all children taught by new method”, “all children taught by standard method”.)

Population 1: mean μ_1 , SD σ_1 . Population 2: mean μ_2 , SD σ_2 .

Don't know any of these. Don't care about actual values of μ_1, μ_2 – care about how they *compare*. So think about $\mu_1 - \mu_2$.

Reading example: 1 is “new”, 2 is “standard”. $\mu_1 = \mu_2$ or

$\mu_1 - \mu_2 = 0$ means two methods equally good, $\mu_1 > \mu_2$ or

$\mu_1 - \mu_2 > 0$ means new better.

Collect samples from each population. Population 1: sample size n_1 , sample mean \bar{x}_1 , sample SD s_1 . Population 2: sample size n_2 , sample mean \bar{x}_2 , sample SD s_2 .

Sample sizes can be same, but don't have to be (8 and 12 in reading example).

Don't know population SDs σ_1, σ_2 , so use sample SDs s_1, s_2 to estimate them, use t -distribution.

The details

To see why: p. 486–488 of the text. Actually t -distribution only approx. correct, but usually good enough.

Calculate

$$d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Base df on smaller sample size.

Test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{d}$$

and confidence interval is

$$\bar{x}_1 - \bar{x}_2 \pm t^* d.$$

Reading example revisited

Want to detect any differences between methods (new method might be worse).

Trying to prove $\mu_1 \neq \mu_2$. Requires test of significance with alternative hyp. $\mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$. Null says $\mu_1 = \mu_2$ or $\mu_1 - \mu_2$.

Data, labelling 1 as “new”: $n_1 = 8, \bar{x}_1 = 77.13, s_1 = 4.85$.
 $n_2 = 12, \bar{x}_2 = 72.33, s_2 = 6.34$.

Thus

$$d = \sqrt{\frac{4.85^2}{8} + \frac{6.34^2}{12}} = 2.508.$$

Test statistic

$$t = \frac{77.13 - 72.33}{2.508} = 1.91.$$

Using $8 - 1 = 7$ df, P-value between 0.05 and 0.10 (2-sided).

Cannot quite reject null: not quite evidence that new reading method makes difference.

95% confidence interval for difference in means, using $t^* = 2.365$:

$$77.13 - 72.33 \pm (2.365)(2.508),$$

from -1.13 to 10.73 .

Minitab: enter “new” scores in one column, “standard” in another (worksheet `reading`).

Stat, Basic Statistics, 2-sample t. Click “Samples in different columns”. Select “new” as 1st column, “standard” as 2nd. Ensure alternative “not equal”.

Two sample T for new vs standard

	N	Mean	StDev	SE Mean
new	8	77.13	4.85	1.7
standard	12	72.33	6.34	1.8

95% CI for mu new - mu standard: (-0.5, 10.1)

T-Test mu new = mu standard (vs not =):

T = 1.91 P = 0.073 DF = 17

Note different df, from formula page 498. Hence smaller P-value, shorter CI than hand calculation.

P-value 0.073, no evidence of difference as before.

Confidence interval from -0.5 to 10.1 , includes 0. “No difference” plausible, though suggests positive effect of new method.

(Minitab gives both test and interval; choose as necessary.)

Sample sizes small in example (8 and 12); might not have much *power* to detect useful difference. (Sample means 77.13 and 72.33 about 5 marks apart – raising reading scores on average by 5 marks might be important.)

Robustness of two-sample procedures

- Generally as for 1-sample t , only replacing “sample size” with sum $n_1 + n_2$.
- Better to use equal sample sizes when possible.
- Overall conclusion: non-normality/outliers not a problem in large sample sizes; worry mainly about outliers in moderate-sized samples.
- Difficult to assess normality anyway in small samples.

Power and sample size for two-sample t tests

Will we be able to prove that new reading method has effect? In general, how likely is rejection of null hyp. when it is false? This is *power*.

Minitab's power calculation assumes equal sample sizes and equal population SDs. To be safe, use *smaller* sample size, *larger* SD.

Reading example: sample sizes 8 and 12 (use 8), sample SDs 4.85, 6.34 (so use 6.34 as guess of population SD). Suppose that increase of 5 marks meaningful.

Stat, Power and Sample Size, 2-sample t. Select “Power for each sample size”, type in sample size 8. Enter difference 5 (marks). Enter 6.34 for “sigma”. Click Options, ensure Alternative is “not equal”, alpha 0.05.

2-Sample t Test

Testing mean 1 = mean 2 (versus not =)

Calculating power for mean 1 = mean 2 + 5

Alpha = 0.05 Sigma = 6.34

Sample

Size	Power
------	-------

8	0.3123
---	--------

Little chance of detecting this difference.

How many children should be in each sample to get this power up to 70% (0.70)?

Minitab: Stat, Power and Sample Size, 2-sample t. Select “Calculate sample size for each power value”, fill in 0.70 for power, 5 for difference. Ensure “sigma” still 6.34, Options correct:

2-Sample t Test

Testing mean 1 = mean 2 (versus not =)

Calculating power for mean 1 = mean 2 + 5

Alpha = 0.05 Sigma = 6.34

Sample Size	Target Power	Actual Power
21	0.7000	0.7033

Need 21 children *in each group* to get this much power.

Summary

- When data not paired, use two-sample t -procedures to compare two means.
- Hypotheses based on difference between population means.
- Sample sizes don't have to be the same.
- These procedures generally robust (provided $n_1 + n_2$ large enough).
- For power, use smaller n , larger s . For sample size, use larger s . Answer gives number of individuals in each group.

Chapter 8: Inference for Proportions

Inference for a single proportion (§8.1)

Suppose that a new snack food is being tried out. 500 students sampled at random, and 71 of them said they would definitely buy it.

Idea: “success” (definitely buy), “failure” (wouldn’t). Count successes in sample (71) or proportion ($71/500 = 0.142$).

What can we say about proportion of *all* students who would buy the new snack? (Confidence interval.)

Company estimates new snack profitable if more than 10% of all students buy it – evidence for this? (Significance test.)

Significance tests for proportions

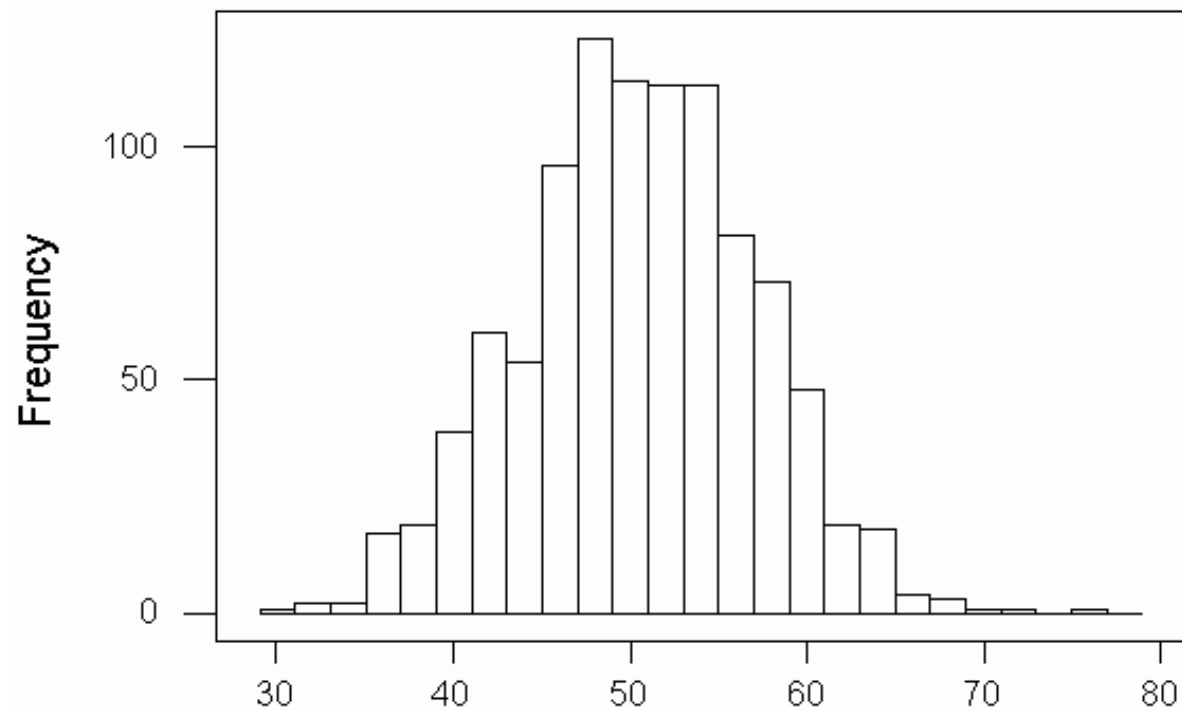
Back in chapter 5, studied sampling distribution of proportions: if we know the population proportion, can work out what kinds of sample proportions we might get.

Under usual assumptions for sampling, **binomial distribution** describes number of successes, with success prob. or population proportion p .

Example: is there evidence that more than 10% of all students would buy? Trying to prove this: alternative hypothesis $p > 0.10$.
Null hyp. $p = 0.10$.

Now, *if null hyp. true*, know population proportion p : 0.10. So number of people in sample who will buy has binomial distribution with $n = 500$ and $p = 0.10$. Is observed value 71 (of 500) consistent?

Simulate: Calc, Random Data, Binomial. Generate (say) 1000 rows of data, store in C1. Number of Trials 500, Probability of Success 0.10.



If p really 0.10, could observe 71 successes, but very unlikely. So conclude that we observed 71 successes because $p > 0.10$: reject null hypothesis in favour of alternative.

Mathematics: can calculate P-values exactly. Used by Minitab.

Stat, Basic Statistics, 1 proportion. Click Summarized Data, fill in number of trials (500), successes (71). Click Options. Test Proportion 0.10, Alternative “greater than”:

Test of $p = 0.1$ vs $p > 0.1$

Sample	X	N	Sample p	95.0 % CI	Exact P-Value
1	71	500	0.142000	(0.112597, 0.175711)	0.002

P-value is very small (0.002), so reject null hyp. without question.

Can be almost certain that more than 10% of students will buy.

Confidence interval for proportion

Continuing example: Minitab gave 95% confidence interval of 0.113 to 0.176. So, from information in sample, believe that between 11.3% and 17.6% of all students will buy this snack food.

Where did this come from?

Recall connection between confidence interval and 2-sided test.

Value *inside* interval “plausible” – would not reject that null hyp. But value *outside* not plausible, would reject.

Try this: test hypothesis $p = 0.113$ using our data, 2-sided:

Test of $p = 0.113$ vs $p \text{ not} = 0.113$

Sample	X	N	Sample p	95.0 % CI	Exact P-Value
1	71	500	0.142000	(0.112597, 0.175711)	0.056

Don't quite reject null hyp., so 0.113 just inside interval. Now try 0.176:

Test of $p = 0.176$ vs $p \text{ not } = 0.176$

Sample	X	N	Sample p	95.0 % CI	Exact P-Value
1	71	500	0.142000	(0.112597, 0.175711)	0.046

This time, would just reject, so 0.176 just outside.

Minitab gets 95% confidence interval by finding the two values of p giving P-value exactly 0.05 (can be done exactly by mathematics).

Test and CI without Minitab

Recall some stuff from §5.1:

- *number of successes* has binomial distribution.
- *sample proportion of successes* has mean p and SD $\sqrt{p(1 - p)/n}$.
- If sample large, number and proportion of successes has approx. normal distribution.

Test of significance

As before, testing null $p = 0.10$ against alternative $p > 0.10$ with data of 71 successes in 500 trials, so $\hat{p} = 71/500 = 0.142$.

Use *null hypothesis* value of p to calculate test statistic:

$$z = \frac{\hat{p} - p}{\sqrt{p(1 - p)/n}} = \frac{0.142 - 0.1}{\sqrt{(0.10)(0.90)/500}} = 3.13$$

with 1-sided P-value (Table A) of $1 - 0.9991 = 0.0009$. Compare approx. 0.0009 and exact 0.002 from Minitab.

Conclusion (either way): reject null and conclude that $p > 0.10$.

Confidence interval

Problem: in $\sqrt{p(1-p)/n}$, have no value for p . So use best value we have, namely \hat{p} . Hence formula:

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

In example, 95% CI is

$$0.142 \pm 1.96 \sqrt{(0.142)(0.858)/500} = 0.142 \pm 0.031,$$

or from 0.111 to 0.173. This is approx.; compare exact 0.113 to 0.176.

Usually, as here, approximation very good.

Power and sample size for proportions

As always with test of significance, want to know whether interesting alternative has some chance of being detected with current sample size (power), or what sample size needed to be able to detect alternative of interest.

To recycle snack food example: if in fact 12% of students would buy product, how likely are we to reject null hyp. $p = 0.10$ in favour of alternative $p > 0.10$ with sample of 500 students?

Minitab: Stat, Power and Sample Size, 1 proportion. Calculate Power for Each Sample Size, fill in sample size 500, alternative p 0.12 (this is the “true” 12%). Put in null hypothesis p , 0.10, at bottom. Click Options, select Greater Than.

Testing proportion = 0.1 (versus > 0.1)

Calculating power for proportion = 0.12

Alpha = 0.05 Difference = 0.02

Sample

Size	Power
500	0.4434

Only sometimes detect this difference.

How many students should we sample to have power 0.60 when in fact 12% of students will buy the product?

Minitab: Stat, Power and Sample Size, 1 proportion. Select Calculate Sample Size for Each Power Value. Fill in power 0.60, alternative value $p = 0.12$, ensure Hypothesized Value still 0.10.

Click Options, ensure still Greater Than:

Testing proportion = 0.1 (versus > 0.1)

Calculating power for proportion = 0.12

Alpha = 0.05 Difference = 0.02

Sample Size	Target Power	Actual Power
829	0.6000	0.6001

Need to sample 829 students to have 60% chance of rejecting null hyp. when $p = 0.12$.

Summary

- Do significance test for proportion by asking: if null hypothesis p correct, is observed #successes plausible or not?
- Get P-value using binomial distribution (or software).
- Confidence interval for p : “undo” test by asking what values of p would make you fail to reject null.
- Calculate power by giving null hypothesis value of p and correct value of p , along with sample size.
- Given null hypothesis p , correct p and desired power, can calculate sample size needed to achieve power.

Inference for two proportions (§8.2)

Often do comparative studies: have two groups (eg. treatment and control) and want to compare groups (eg. test null hypothesis that groups are same).

If data are measured: calculate means, SDs, use methods of §7.2.

If data are success/failure: have two binomial counts of successes, one for each group. What to do?

Example (8.47): In a study of (a randomly chosen) syntax text, are references to females more or less likely to refer to juveniles (“girl” vs. “woman”) than references to males (“boy” vs. “man”) in 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.

Thinking of “success” as a “juvenile” reference, number of successes has a binomial distribution. Sample proportions are $\hat{p}_1 = 48/60 = 0.80$ and $\hat{p}_2 = 52/132 = 0.394$.

Getting confidence interval for $p_1 - p_2$

Suppose X_1 has a binomial distribution with n_1 trials and success probability p_1 , and X_2 has also with n_2 trials and prob. p_2 .

Then: X_1 has mean $n_1 p_1$, variance $n_1 p_1 (1 - p_1)$ and SD $\sqrt{n_1 p_1 (1 - p_1)}$. Also $\hat{p}_1 = X_1 / n_1$ has mean p_1 , variance $p_1 (1 - p_1) / n_1$.

Same applies for X_2 and \hat{p}_2 .

Also, if n_1 and n_2 are large (only case we consider), \hat{p}_1 and \hat{p}_2 approx. normally distributed.

Can estimate population proportion difference $p_1 - p_2$ by sample proportion difference $D = \hat{p}_1 - \hat{p}_2$. D is difference of normals, so its variance is *sum* of individual variances:

$$\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}.$$

We don't know p_1 and p_2 , so replace by sample estimates to get SD (standard error) of D as

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

Then a confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z^* SE_D.$$

Numbers from example: $\hat{p}_1 = 0.80, n_1 = 60$;

$\hat{p}_2 = 0.394, n_2 = 132$. Hence

$$SE_D = \sqrt{\frac{(0.80)(0.20)}{60} + \frac{(0.394)(0.606)}{132}} = 0.067,$$

and for a 95% confidence interval, $z^* = 1.96$, giving

$$0.80 - 0.394 \pm (1.96)(0.067) = 0.406 \pm 0.131,$$

from 0.275 to 0.537.

We think the proportion of words for females that are juvenile is bigger than the proportion of words for males that are juvenile by between 0.275 and 0.537.

This confidence interval contains only positive values, so we believe that the proportion of juvenile words for females is bigger.

Confidence interval for two proportions using Minitab

In Minitab, select Stat, Basic Statistics, 2 Proportions. Select Summarized Data (the bottom one). Fill in the number of trials and the number of successes for each group. For our data, there are 60 trials and 48 successes, then 132 trials and 52 successes.

If you want a confidence level other than 95%, click Options and change it.

This command will also give you a test of significance, which you can ignore for now (we will learn it later).

My output looks like this:

Sample	X	N	Sample p
1	48	60	0.800000
2	52	132	0.393939

Estimate for $p(1) - p(2)$: 0.406061

95% CI for $p(1) - p(2)$: (0.274942, 0.537179)

This is the same as our calculation by hand.

Test of significance for two proportions

Confidence interval says “how far apart could proportions be?”

To see whether proportions believably the same, need a test of significance, with null hypothesis $p_1 = p_2$. Base test on

$$D = \hat{p}_1 - \hat{p}_2,$$

and reject null if too far from (above, below) zero.

Now,

$$SE_D = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

but if null is true, then $p_1 = p_2 = p$, say, so that SE_D becomes

$$SE_{Dp} = \sqrt{\frac{p(1 - p)}{n_1} + \frac{p(1 - p)}{n_2}} = \sqrt{p(1 - p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

Of course, don't know p either, but estimate it using overall proportion of successes:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$

So procedure for test is this:

1. Calculate \hat{p} from data.
2. Calculate SE_{Dp} using \hat{p} in place of p .
3. Calculate test statistic $z = (\hat{p}_1 - \hat{p}_2) / SE_{Dp}$
4. Get P-value from normal distribution.

To illustrate with example data

$X_1 = 48, n_1 = 60, X_2 = 52, n_2 = 132$ and alternative $p_1 \neq p_2$:

$$\hat{p} = (48 + 52)/(60 + 132) = 100/192 = 0.521;$$

$$SE_{Dp} = \sqrt{(0.521)(0.479) \left(\frac{1}{60} + \frac{1}{132} \right)} = 0.0778;$$

$$z = \frac{0.800 - 0.394}{0.0778} = 5.22.$$

This is off the end of the table, so the P-value is close to $2 \times 0 = 0$.

We can conclude that the two proportions are different.

In this example, the sample proportions are very different, so (not surprisingly) we could conclude that the population proportions are different. In general, proving that population proportions are different requires either

- sample proportions that are very different, or
- sample sizes that are very large.

Minitab produces the same answer for our data. Select Stat, Basic Statistics, 2 Proportions; click Summarized Data and fill in the sample values. Click Options; make sure that Alternative is Not Equal, and select Use Pooled Estimate. Click OK. Get the output from before plus this:

Test for $p(1) - p(2) = 0$ (vs not = 0): $Z = 5.22$ P-Value = 0.000

Summary of §8.2

- To compare population proportions p_1 and p_2 , use difference in sample proportions $D = \hat{p}_1 - \hat{p}_2$.
- Standard error $SE_D = \sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}$.
- For CI, substitute sample proportions in SE_D ; interval is $\hat{p}_1 - \hat{p}_2 \pm z^* SE_D$.
- For test, null is $p_1 = p_2$. Use SE_D with p replacing p_1 and p_2 .
- Estimate p using overall sample proportion of successes.
- Test statistic $z = (\hat{p}_1 - \hat{p}_2)/SE_{Dp}$, get P-value from normal table.