

MATH1905 Statistics (Advanced)

Michael Stewart

Semester 2, 2016

Outline

- 1 Welcome
 - Lecture 1
 - Overview
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models

What is statistics?

- Two levels:
 - ▶ analysis of data in and of itself;
 - ▶ interpretation of data as being somehow representative of the process creating it.
- The course is divided into 3 parts:
 - ▶ data analysis;
 - ▶ probability;
 - ▶ inference.

Data Analysis

- Summarising data.
- Presenting data.
- Extracting “information”.

Probability

- Provides mathematical models for the data-generation process.
- Our observed “sample” is just one of a large number of possible “samples”.
- A set of (mathematical) tools that allow us to develop methods of...

... Inference

- Using probability models and the data to make sensible statements ("inferences") about the data-generation process.

Computing: R and RStudio

- R is

- ▶ a free software environment for statistical computing and graphics.
- ▶ It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- ▶ Website: <https://www.r-project.org/> .

- RStudio is

- ▶ a cross-platform GUI (graphical user interface) for R which can function as a full-blown IDE (integrated development environment).
- ▶ However, it provides a user-friendly interface even for beginners.
- ▶ RStudio provide an open-source (free) version as well as professional (paid) versions (which are not directly relevant to us, although this “model” of software development often provides high-quality stuff for free).
- ▶ Website: <https://www.rstudio.com> .

Trying R

- Students are encouraged to install both R and RStudio on their own computers (RStudio will not work without R; install R first).
- A useful website to get started is **Try R**: <http://tryr.codeschool.com> (this series of tutorials can be completed *without registering*).

An example: Oral contraceptive: effect on blood pressure?

```
OC=scan("http://maths.usyd.edu.au/math1905/r/OC.txt")
OC
```

Read 24 items

```
[1] 115.7 125.9 122.9 125.2 139.3 141.4 141.3 123.0 135.7 124.0 111.8 120.1
[13] 117.7 119.5 131.0 132.8 132.7 117.7 128.1 118.6 138.7 94.7 127.2 126.5
```

```
u=OC[1:10]
u
```

```
[1] 115.7 125.9 122.9 125.2 139.3 141.4 141.3 123.0 135.7 124.0
```

```
nu=OC[-(1:10)]
nu
```

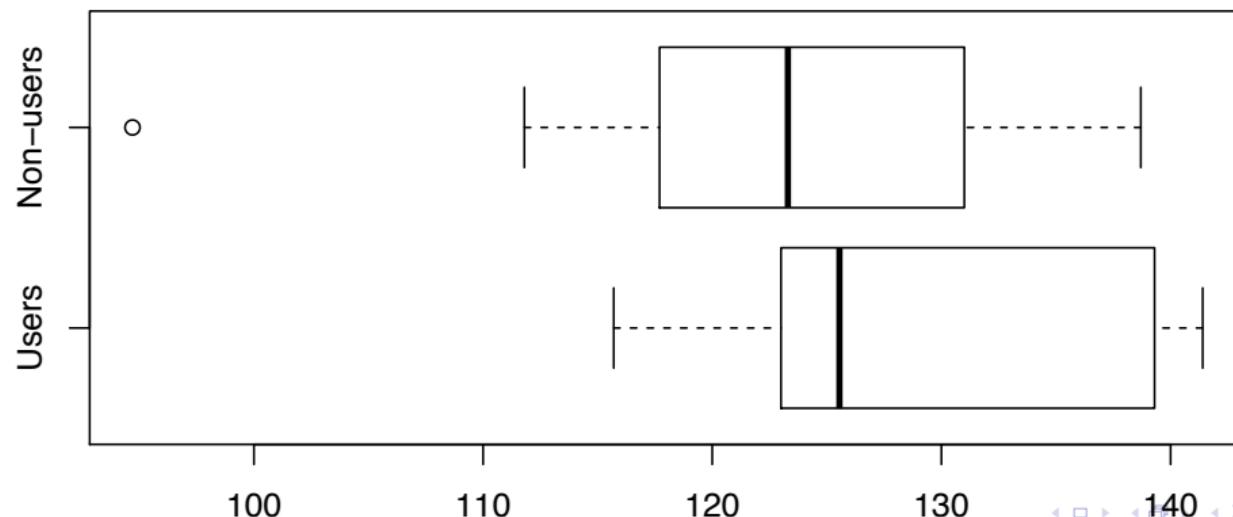
```
[1] 111.8 120.1 117.7 119.5 131.0 132.8 132.7 117.7 128.1 118.6 138.7 94.7
[13] 127.2 126.5
```

```
labels=c("Users", "Non-users")
```

```
labels
```

```
[1] "Users"      "Non-users"
```

```
boxplot(u,nu, horizontal=T, names=labels)
```



Outline

1 Welcome

2 Data Analysis
● Lecture 2

- Data types
- Discrete and Continuous data summaries
- Discrete Data Summaries in R
- Continuous Data Summaries in R

● Lecture 3
● Lecture 4
● Lecture 5

3 Probability

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

Overview

- Data types:
 - ▶ discrete;
 - ▶ continuous.
- Data Summaries:
 - ▶ discrete:
 - ★ frequency table;
 - ★ ordinate diagram;
 - ▶ continuous:
 - ★ stem and leaf plots;
 - ★ histograms;
 - ★ boxplots.
- Numerical Summaries:
 - ▶ median;
 - ▶ quartiles.

Discrete Data

- Can only take values in a set of isolated/discrete points;
- Classic example: *integers*.

Continuous Data

- In contrast to discrete data, continuous data can (in principle) take any value in a (given, known possibly infinite-lengthed) interval.
- In fact such data is generally rounded to a certain number of decimal places and so is in fact discrete in that sense.
- The discrete or continuous nature of the *quantities represented* by the data determine what kind of probability model we try to fit to that data.
- Discrete and Continuous models are fundamentally different.

Summarising discrete data

- An elementary way to summarise/reduce discrete data is to produce a frequency table which counts how many times each distinct value occurs.
- The R function `table()` gives a frequency table.
- A frequency table can be used to produce an *ordinate diagram*.

Summarising continuous data

There are various ways we shall (graphically) summarise continuous data:

- stem-and-leaf plot;
- histogram;
- boxplot.

Stem-and-leaf plots

- These are best explained by way of example (see example below, and on pages 9–12 of the Cartoon Guide).
- In brief:
 - ▶ each data value is “split” (e.g. at the decimal point of between 10’s and 100’s) into a “stem” and a “leaf”;
 - ▶ each *leaf* is reduced to 1 significant digit;
 - ▶ leaves are then represented on rows, paired with their stems.

Histograms

- A set of nonoverlapping intervals is chosen whose union covers the range of the data.
- A rectangle is drawn on top of each interval whose **area** reflects the frequency of values in that interval.

Boxplots

- Median and quartiles are obtained (more on this later).
- Box is drawn between quartiles (against a scale).
- Outliers are determined:
 - ▶ interquartile range (IQR) is determined (i.e. width of the box);
 - ▶ any value *more* than 1.5 IQRs away from the box (in either direction) is deemed an *outlier*.
- Whiskers are drawn to the extreme non-outliers.
- Outliers are marked separately.

Discrete Data Summaries in R

Dataset 1 from Phipps & Quine p131 is read in using scan():

```
PQ1=scan("http://maths.usyd.edu.au/math1905/r/PQdataset1.txt")
PQ1
```

Read 400 items

```
[1] 2 2 4 4 4 5 2 4 7 7 4 7 5 2 8 6 7 4 3 4 3 3 2 4 2
[26] 5 4 2 8 6 3 6 6 10 8 3 5 6 4 4 7 9 5 2 7 4 4 2 4 4
[51] 4 3 5 6 5 4 1 4 2 6 4 1 4 7 3 2 3 5 8 2 9 5 3 9 5
[76] 5 2 4 3 4 4 1 5 9 3 4 4 6 6 5 4 6 5 5 4 3 5 9 6 4
[101] 4 4 5 10 4 4 3 8 3 2 1 4 1 5 6 4 2 3 3 3 3 7 4 5 1
[126] 8 5 7 9 5 8 9 5 6 6 4 3 7 4 4 7 5 6 3 6 7 4 5 8 6
[151] 3 3 4 3 7 4 4 4 5 3 8 10 6 3 3 6 5 2 5 3 11 3 7 4 7
[176] 3 5 5 3 4 1 3 7 2 5 5 5 3 3 4 6 5 6 1 6 4 4 4 6 4
[201] 4 2 5 4 8 6 3 4 6 5 2 6 6 1 2 2 2 5 2 2 5 9 3 5 6
[226] 4 6 5 7 1 3 6 5 4 2 8 9 5 4 3 2 2 11 4 6 6 4 6 2 5
[251] 3 5 7 2 6 5 5 1 2 7 5 12 5 8 2 4 2 1 6 4 5 1 2 9 1
[276] 3 4 7 3 6 5 6 5 4 4 5 2 7 6 2 7 3 5 4 4 5 4 7 5 4
[301] 8 4 6 6 5 3 3 5 7 4 5 5 5 6 10 2 3 8 3 5 6 6 4 2 6
[326] 6 7 5 4 5 8 6 7 6 4 2 6 1 1 4 7 2 2 5 7 4 6 4 5 1 5
[351] 10 8 7 5 4 6 4 4 7 5 4 3 1 6 2 5 3 3 3 7 4 3 7 8 4
[376] 7 3 1 4 4 7 6 7 2 4 5 1 3 12 4 2 2 8 7 6 7 6 3 5 4
```

Frequency Table

The command `table()` produces a frequency table:

```
table(PQ1)
```

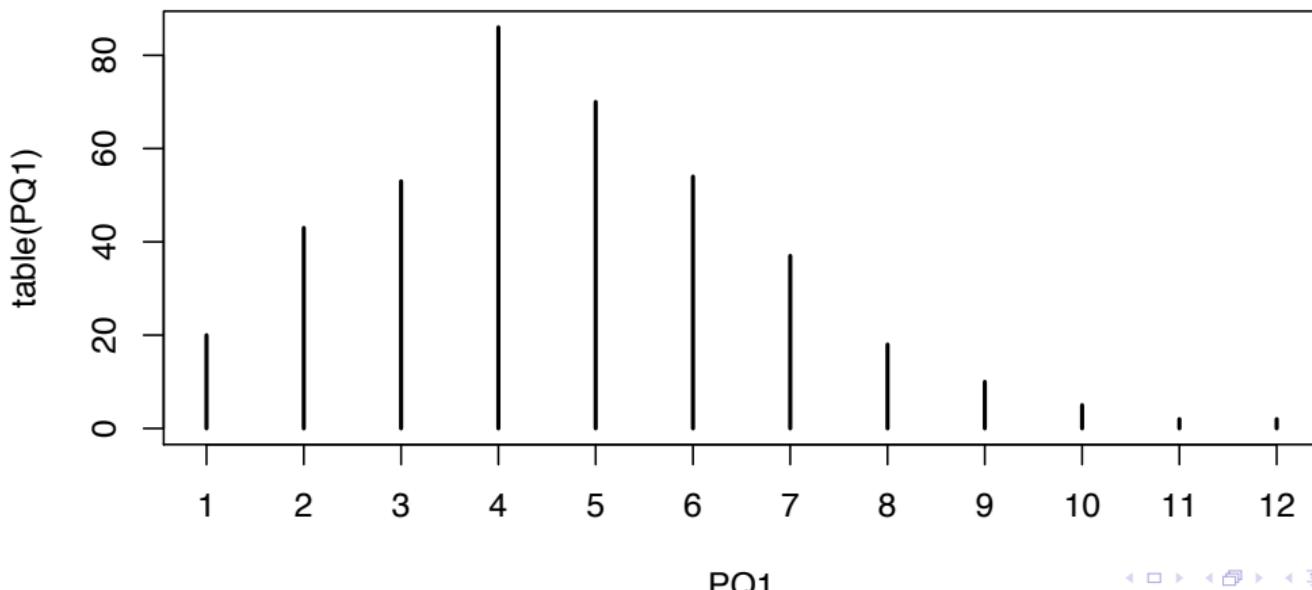
PQ1

1	2	3	4	5	6	7	8	9	10	11	12
20	43	53	86	70	54	37	18	10	5	2	2

Ordinate Diagram

When given as an argument to `plot()`, an ordinate diagram is produced:

```
plot(table(PQ1))
```



Continuous Data Summaries in R

We read in the male and female heights from the data on page 9 of Gonick & Smith and then combine them into a single vector:

```
GS9m=scan("http://maths.usyd.edu.au/math1905/r/GS9m.txt")
```

Read 57 items

```
GS9m
```

```
[1] 140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180  
[20] 135 170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160  
[39] 135 160 130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155
```

```
GS9f=scan("http://maths.usyd.edu.au/math1905/r/GS9f.txt")
```

Read 35 items

```
GS9f
```

```
[1] 140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135  
[20] 125 118 122 115 102 115 150 110 116 108 95 125 133 110 150 108
```

```
GS9=c(GS9m , GS9f )
GS9
```

```
[1] 140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180
[20] 135 170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160
[39] 135 160 130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155
[58] 140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135
[77] 125 118 122 115 102 115 150 110 116 108 95 125 133 110 150 108
```

Stem-and-leaf Plot

A stem-and-leaf plot is produced using `stem()`. The `scale=` parameter controls (roughly) how many rows appear, relative to the default of 1.

```
stem(GS9)
```

The decimal point is 1 digit(s) to the right of the |

8		5
10		288002556688
12		000123555550000013555688
14		00002555558000000000035555555557
16		000045000055
18		000500005
20		5

We see that here the default is to put the break between the tens and units and to put 20 leaves per stem. So all values 80 to 99 inclusive would be represented on the first row above; that 5 actually represents 95, not 85.

If we ask for approximately double the number of rows, it now keeps the break at the same place but puts only 10 leaves per stem:

```
stem(GS9, scale=2)
```

The decimal point is 1 digit(s) to the right of the |

9		5
10		288
11		002556688
12		00012355555
13		0000013555688
14		00002555558
15		00000000035555555557
16		000045
17		000055
18		0005
19		00005
20		
21		5

Now only values 90 to 99 inclusive would be represented on the first row above.

Asking for `scale=3` or `scale=4` results in the same plot, with again the break between the tens and the units but only putting 5 leaves per stem. So now the first row only represents 95 to 99 inclusive:

```
stem(GS9, scale=3)
```

The decimal point is 1 digit(s) to the right of the |

```
9 | 5
10 | 2
10 | 88
11 | 002
11 | 556688
12 | 000123
12 | 55555
13 | 0000013
13 | 555688
14 | 00002
14 | 555558
15 | 00000000003
15 | 5555555557
16 | 00004
16 | 5
17 | 0000
17 | 55
18 | 000
18 | 5
19 | 0000
19 | 5
20 |
20 |
21 |
21 | 5
```

```
stem(GS9, scale=4)
```

The decimal point is 1 digit(s) to the right of the |

```
9 | 5
10 | 2
10 | 88
11 | 002
11 | 556688
12 | 000123
12 | 55555
13 | 0000013
13 | 555688
14 | 00002
14 | 555558
15 | 00000000003
15 | 55555555557
16 | 00004
16 | 5
17 | 0000
17 | 55
18 | 000
18 | 5
19 | 0000
19 | 5
20 |
20 |
21 |
21 | 5
```

Putting scale=5 pushes the break to the decimal point:

```
stem(GS9, scale=5)
```

The decimal point is at the |

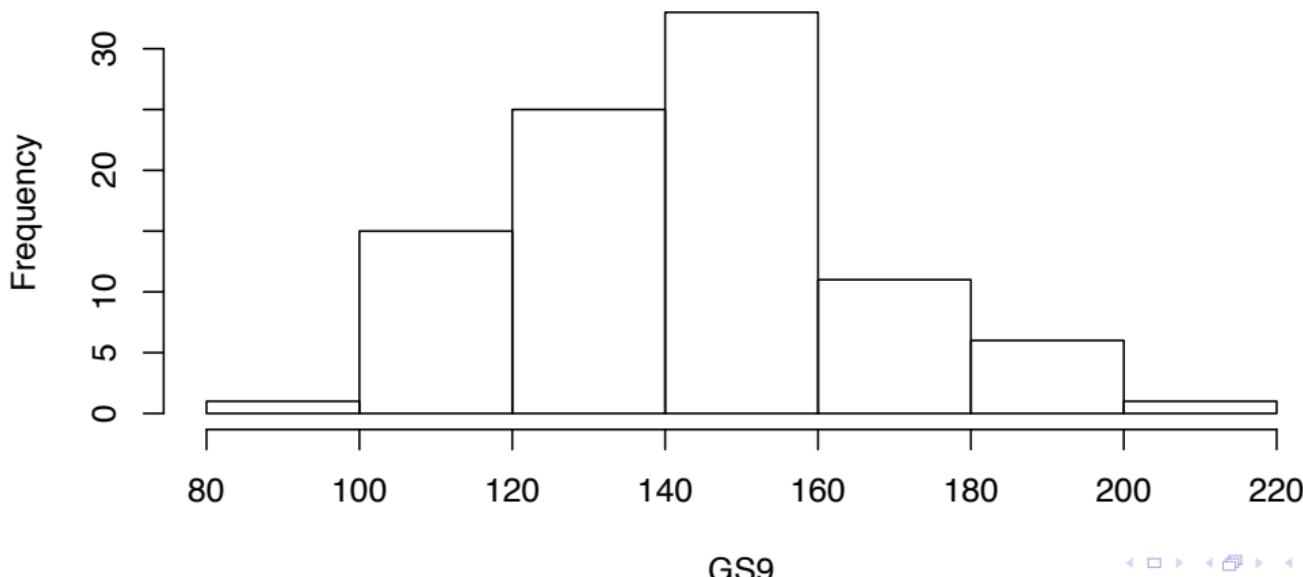
```
94 | 0  
96 |  
98 |  
100 |  
102 | 0  
104 |  
106 |  
108 | 00  
110 | 00  
112 | 0  
114 | 00  
116 | 00  
118 | 00  
120 | 0000  
122 | 00  
124 | 00000  
126 |  
128 |  
130 | 000000  
132 | 0  
134 | 000  
136 | 0  
138 | 00  
140 | 0000  
142 | 0  
144 | 00000  
146 |  
148 | 0  
150 | 0000000000  
152 | 0  
154 | 0000000000  
156 | 0
```

Histogram

Default histogram of GS9:

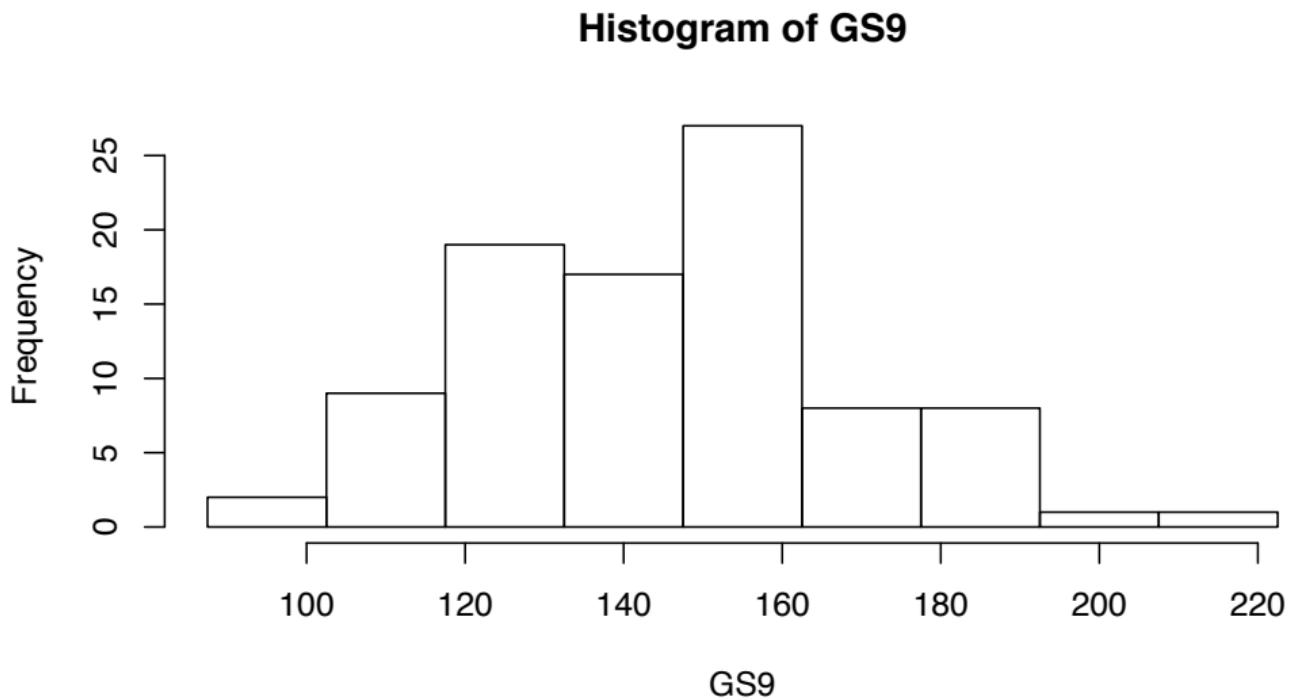
```
hist(GS9)
```

Histogram of GS9



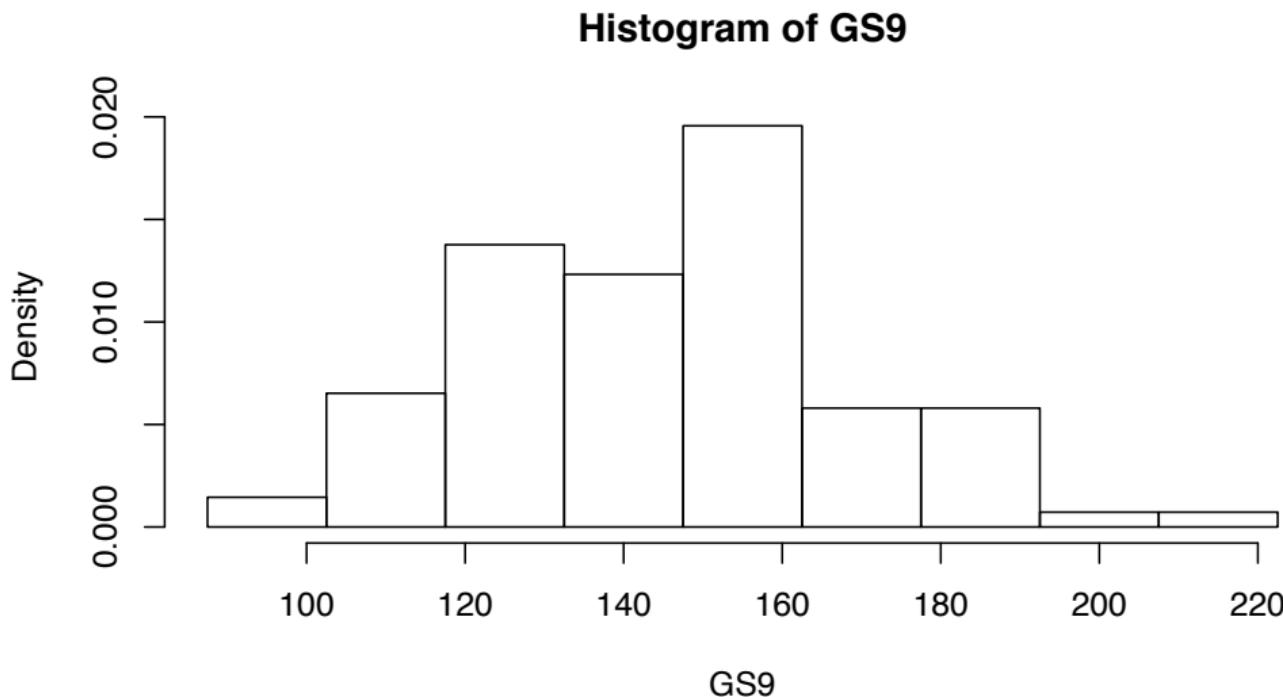
We can tell it what intervals to use using the `breaks=` argument:

```
br=seq(from=87.5,to=222.5,by=15)  
hist(GS9,breaks=br)
```



Setting `prob=T` gives a *probability* or *relative frequency histogram*, that is it changes the vertical scale so that the area under the histogram is 1 unit.

```
hist(GS9, breaks=br, prob=T)
```



How are endpoints handled?

- Histogram endpoint rules in R: by default each interval includes its *right* endpoint (not its left), unless we specify `right=FALSE`:

```
x=1:6  
x
```

```
[1] 1 2 3 4 5 6
```

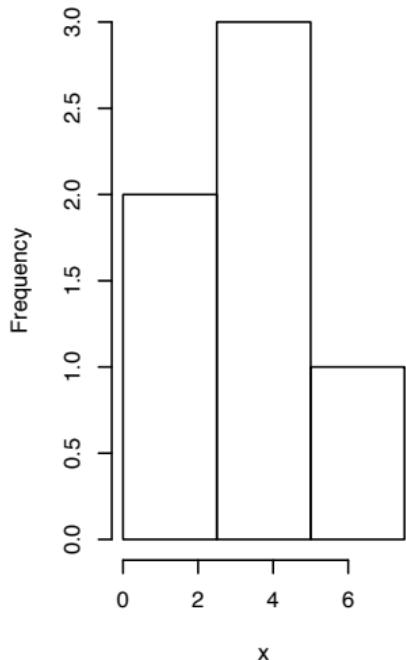
```
br=c(0,2.5,5,7.5)  
br
```

```
[1] 0.0 2.5 5.0 7.5
```

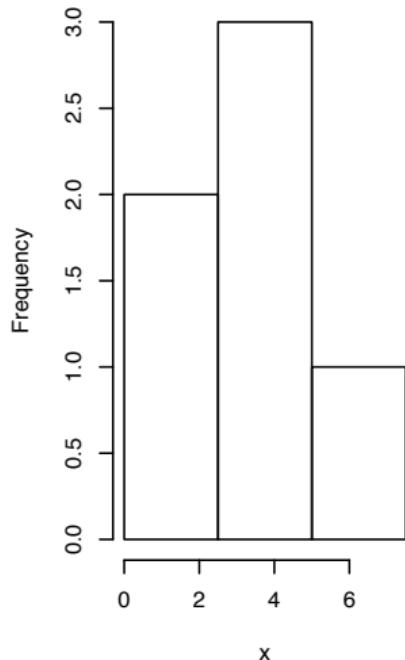
- Note we have a data value equal to an interval boundary here (the value 5).

```
par(mfrow=c(1,3)) # "Multiple Figures by ROW"  
hist(x, breaks=br)  
hist(x, breaks=br, right=T)  
hist(x, breaks=br, right=F, ylim=c(0,3))
```

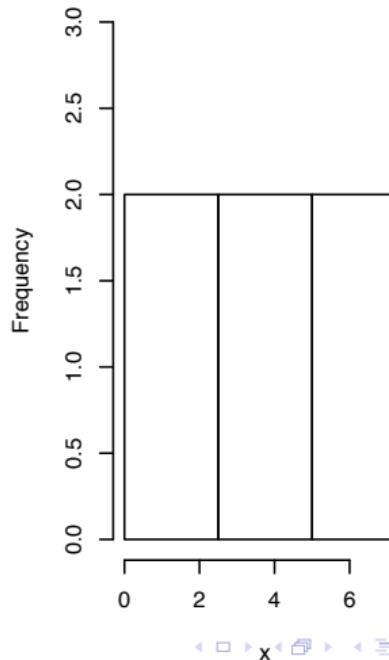
Histogram of x



Histogram of x



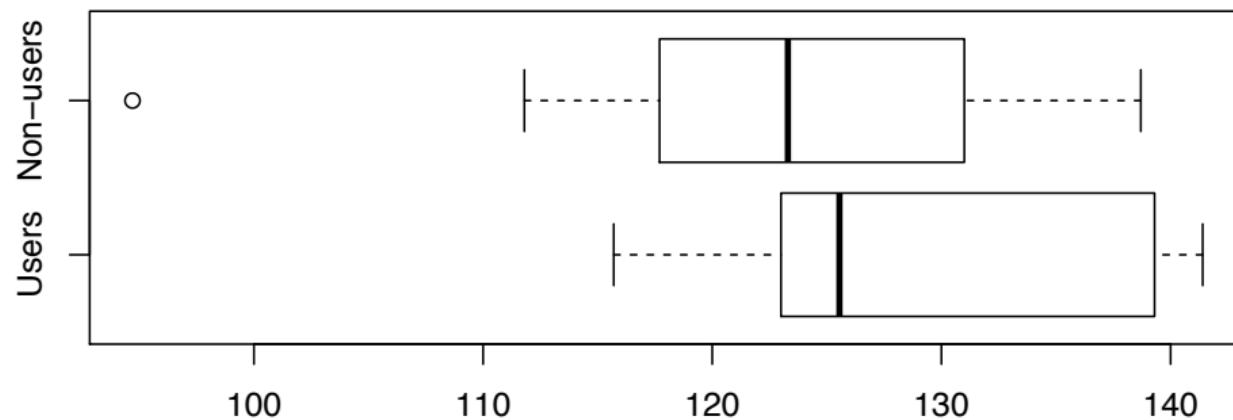
Histogram of x



Boxplots

- Boxplots are constructed using the *quartiles* which cut the ordered data into 4 (roughly) equal pieces. Recall this earlier example:

```
OC=scan("http://maths.usyd.edu.au/math1905/r/OC.txt")
u=OC[1:10]
nu=OC[-(1:10)]
labels=c("Users","Non-users")
boxplot(u,nu,horizontal=T,names=labels)
```



Outline

1 Welcome

2 Data Analysis
● Lecture 2
● Lecture 3

● Numerical Summaries

● Lecture 4
● Lecture 5

3 Probability

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

Median

- The “middle score”.
- Suppose x_1, \dots, x_n represents “the data” and let $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ represent the corresponding “order statistics” (so $x_{(1)} = \min_i x_i$, $x_{(n)} = \max_i x_i$, etc).
- Then the **median** \tilde{x} (Phipps and Quine notation) is given by
 - ▶ $\tilde{x} = x_{(\frac{n+1}{2})}$ if n is odd;
 - ▶ $\tilde{x} = \frac{1}{2} \left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right]$ if n is even.

Some properties of the median

- If $z_i = a + x_i$ (for some constant a) then $\tilde{z} = \tilde{x} + a$.
- If $z_i = bx_i$ (for some constant b) then $\tilde{z} = b\tilde{x}$.
- If $z_i = a + bx_i$ (for some constants a and b) then $\tilde{z} = a + b\tilde{x}$.

These are desirable properties of any “measure of location”.

An example using R

- Recall earlier example

```
OC=scan("http://maths.usyd.edu.au/math1905/r/OC.txt")
```

Read 24 items

```
u=OC[1:10]  
nu=OC[-(1:10)]  
u
```

```
[1] 115.7 125.9 122.9 125.2 139.3 141.4 141.3 123.0 135.7 124.0
```

```
nu
```

```
[1] 111.8 120.1 117.7 119.5 131.0 132.8 132.7 117.7 128.1 118.6 138.7 94.7  
[13] 127.2 126.5
```

```
median(u)
```

```
[1] 125.55
```

```
median(nu)
```

```
[1] 123.3
```

```
sort(u)
```

```
[1] 115.7 122.9 123.0 124.0 125.2 125.9 135.7 139.3 141.3 141.4
```

```
cbind(1:10, sort(u))
```

	[,1]	[,2]
[1,]	1	115.7
[2,]	2	122.9
[3,]	3	123.0
[4,]	4	124.0
[5,]	5	125.2
[6,]	6	125.9
[7,]	7	135.7
[8,]	8	139.3
[9,]	9	141.3
[10,]	10	141.4

```
sort(u)[5:6]
```

```
[1] 125.2 125.9
```

```
mean(sort(u)[5:6])
```

```
[1] 125.55
```

```
sort(nu)
```

```
[1] 94.7 111.8 117.7 117.7 118.6 119.5 120.1 126.5 127.2 128.1 131.0 132.7  
[13] 132.8 138.7
```

```
sort(nu)[7:8]
```

```
[1] 120.1 126.5
```

```
mean(sort(nu)[7:8])
```

```
[1] 123.3
```

Quartiles and Quantiles

- Given a “sample” x_1, x_2, \dots, x_n , a p -th sample quantile (for $0 \leq p \leq 1$) is *any* value which has
 - ▶ “roughly” $100p\%$ of the x_i ’s below it *and*
 - ▶ “roughly” $100(1 - p)\%$ of the x_i ’s above it.
- Examples are
 - ▶ deciles ($p = 0.1$ or multiples thereof);
 - ▶ quartiles:
 - ★ lower quartile (Q_1) corresponds to $p = 0.25$;
 - ★ upper (or third) quartile (Q_3) corresponds to $p = 0.75$;
- Unfortunately there is not universal agreement on how to make this precise (i.e. get rid of “roughly”).

Tukey's quartiles and five-number summary

- J.W. Tukey was a pioneer in the field of data analysis (inventor of
 - ▶ boxplot
 - ▶ the word “software”
 - ▶ many, many other things).
- His definition of quartiles was very straightforward:
 - ▶ Q_1 is the median of the lower half of the data which *includes* the median if n is odd.
 - ▶ So if the sample size $n = 4q + r$ for integers $q \geq 0$ (quotient), $0 \leq r \leq 3$ (remainder),
 - ★ if $r = 0$, $Q_1 = \frac{1}{2} [x_{(q)} + x_{(q+1)}]$;
 - ★ if $r = 1$ or 2 , $Q_1 = x_{(q+1)}$;
 - ★ if $r = 3$, $Q_1 = \frac{1}{2} [x_{(q+1)} + x_{(q+2)}]$.
 - ▶ Q_3 defined similarly for upper half of the data.
 - ▶ **Five number summary:** minimum, Q_1 , median, Q_3 , maximum.
- Agrees with Phipps & Quine except when $r = 3$ (P&Q say $Q_1 = x_{(q+1)}$ in that case).

Quartiles/Five-number summary in R

- Quartiles in R (and generally in fact) are a **nightmare**.
- The function `quantile()` (not `quartile`, `quantile`) gives 9 different methods for computing quartiles!
- Consider the elementary dataset 1,2,3,4,5,6:

```
1:6
```

```
[1] 1 2 3 4 5 6
```

```
quantile(1:6)
```

```
0% 25% 50% 75% 100%
1.00 2.25 3.50 4.75 6.00
```

```
quantile(1:6, type=1)
```

0% 25% 50% 75% 100%
1 2 3 5 6

```
quantile(1:6, type=2) # This is used in Phipps & Quine
```

0% 25% 50% 75% 100%
1.0 2.0 3.5 5.0 6.0

```
quantile(1:6, type=3)
```

0% 25% 50% 75% 100%
1 2 3 4 6

```
quantile(1:6, type=4)
```

0% 25% 50% 75% 100%
1.0 1.5 3.0 4.5 6.0

```
quantile(1:6, type=5)
```

0% 25% 50% 75% 100%
1.0 2.0 3.5 5.0 6.0

```
quantile(1:6, type=6)
```

```
0% 25% 50% 75% 100%
1.00 1.75 3.50 5.25 6.00
```

```
quantile(1:6, type=7) # this is the default in R
```

```
0% 25% 50% 75% 100%
1.00 2.25 3.50 4.75 6.00
```

```
quantile(1:6, type=8)
```

```
0%      25%      50%      75%      100%
1.000000 1.916667 3.500000 5.083333 6.000000
```

```
quantile(1:6, type=9)
```

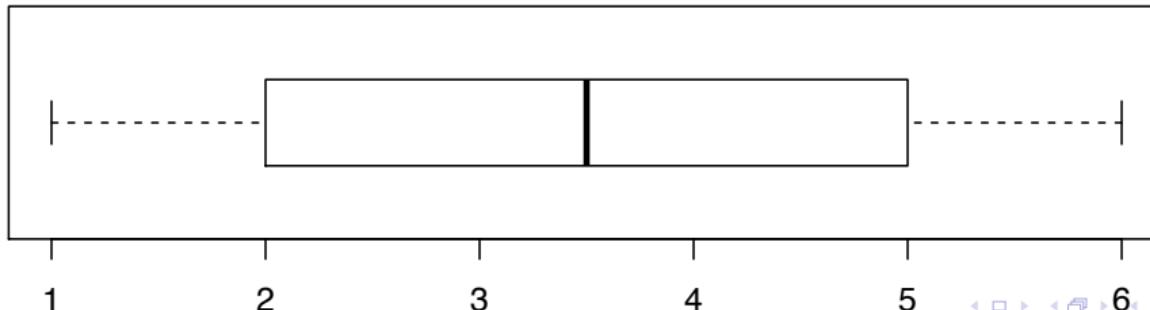
```
0%      25%      50%      75%      100%
1.0000 1.9375 3.5000 5.0625 6.0000
```

- To make matters worse, the function `boxplot()` uses *none* of these.
- Rather it uses Tukey's five-number summary, given by the R function `fivenum()`, which happens to agree with both `type=2` and `type=5` in this example.
- Note that Phipps and Quine's method of computing quartiles corresponds to `type=2`:

```
fivenum(1:6)
```

```
[1] 1.0 2.0 3.5 5.0 6.0
```

```
boxplot(1:6, horizontal=T)
```



In this example `fivenum()`/`boxplot()` agree with type=2 but *not* type=5:

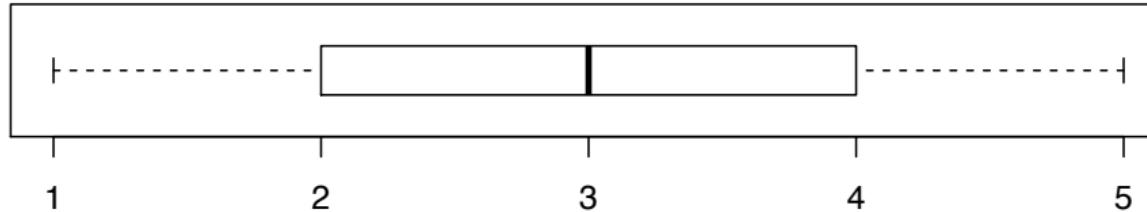
```
quantile(1:5, type=2)
```

```
0% 25% 50% 75% 100%
1    2    3    4    5
```

```
quantile(1:5, type=5)
```

```
0% 25% 50% 75% 100%
1.00 1.75 3.00 4.25 5.00
```

```
fivenum(1:5)
boxplot(1:5, horizontal=T)
```



while here they agree with neither:

```
quantile(1:7, type=2)
```

```
0% 25% 50% 75% 100%
1     2     4     6     7
```

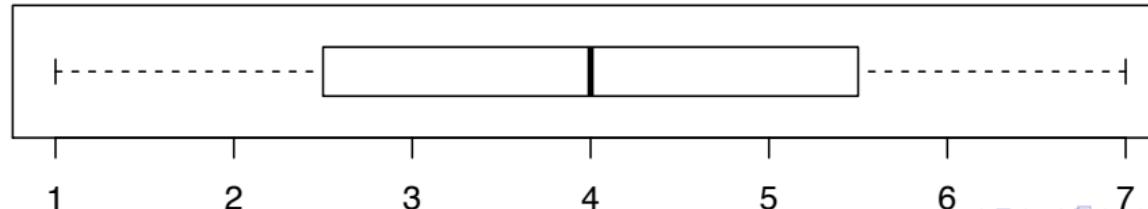
```
quantile(1:7, type=5)
```

```
0% 25% 50% 75% 100%
1.00 2.25 4.00 5.75 7.00
```

```
fivenum(1:7)
```

```
[1] 1.0 2.5 4.0 5.5 7.0
```

```
boxplot(1:7, horizontal=T)
```



- There is clearly no “natural” or “best” definition of sample quartile.
- For the sake of sanity we stick to Tukey’s 5-number summary, so our boxplots agree with our calculations; and we can avoid using `quantile()`!!
- Also, this happens to agree with the method used in the Cartoon Guide by Gonick & Smith.
- Note that the function `IQR()` uses `quantile()`. It is best to just use `fivenum()`, or something like

```
fivenum(u)[c(2,4)]
```

```
[1] 123.0 139.3
```

```
diff(fivenum(u)[c(2,4)])
```

```
[1] 16.3
```

or even

```
fivenum(u)[4] - fivenum(u)[2]
```

```
[1] 16.3
```

A property of the IQR

- The $\text{IQR} = Q_3 - Q_1$ is a common “measure of spread”. Note that if $z_i = a + bx_i$ then $\text{IQR}_z = |b| \text{IQR}_x$.
- This is an important property of a measure of spread (as we shall see).

The Mean

- A very important numerical summary is the (arithmetic) mean: if we have n values the mean is $\frac{\text{sum of values}}{n}$.
- Do not get confused with geometric or harmonic means.
- Sample means have very nice theoretical properties.

Σ (Sigma)-notation

- A more compact way to write the mean is via “Sigma-notation”.
- A shorthand for the sums of a given sequence x_1, x_2, \dots, x_n is
$$(x_1 + x_2 + \dots + x_n) = \sum_{i=1}^n x_i = \sum_{j=1}^n x_j = \sum_{k=1}^n x_k.$$
- We adopt the convention that the mean of x_1, x_2, \dots, x_n is denoted \bar{x} , the mean of y_1, y_2, \dots, y_n is denoted \bar{y} , etc.
- So $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.
- It is easily checked that if $z_i = a + bx_i$ then $\bar{z} = a + b\bar{x}$.

Computing the mean from a frequency table

- In general suppose **discrete** data x_1, x_2, \dots, x_n is condensed into a frequency table as follows:

Values	v_1	v_2	\dots	v_k	Total
Freq's	f_1	f_2	\dots	f_k	n

- Then note that we can write

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{j=1}^k f_j v_j}{\sum_{j=1}^k f_j}.$$

Mean in R

```
PQ1=scan("http://maths.usyd.edu.au/math1905/r/PQdataset1.txt")
PQ1
```

Read 400 items

```
[1] 2 2 4 4 4 5 2 4 7 7 4 7 5 2 8 6 7 4 3 4 3 3 2 4 2 2
[26] 5 4 2 8 6 3 6 6 10 8 3 5 6 4 4 7 9 5 2 7 4 4 2 4 4
[51] 4 3 5 6 5 4 1 4 2 6 4 1 4 7 3 2 3 5 8 2 9 5 3 9 5
[76] 5 2 4 3 4 4 1 5 9 3 4 4 6 6 5 4 6 5 5 4 3 5 9 6 4
[101] 4 4 5 10 4 4 3 8 3 2 1 4 1 5 6 4 2 3 3 3 7 4 5 1
[126] 8 5 7 9 5 8 9 5 6 6 4 3 7 4 4 7 5 6 3 6 7 4 5 8 6
[151] 3 3 4 3 7 4 4 4 5 3 8 10 6 3 3 6 5 2 5 3 11 3 7 4 7
[176] 3 5 5 3 4 1 3 7 2 5 5 5 3 3 4 6 5 6 1 6 4 4 4 6 4
[201] 4 2 5 4 8 6 3 4 6 5 2 6 6 1 2 2 2 5 2 2 5 9 3 5 6
[226] 4 6 5 7 1 3 6 5 4 2 8 9 5 4 3 2 2 11 4 6 6 4 6 2 5
[251] 3 5 7 2 6 5 5 1 2 7 5 12 5 8 2 4 2 1 6 4 5 1 2 9 1
[276] 3 4 7 3 6 5 6 5 4 4 5 2 7 6 2 7 3 5 4 4 5 4 7 5 4
[301] 8 4 6 6 5 3 3 5 7 4 5 5 5 6 10 2 3 8 3 5 6 6 4 2 6
[326] 6 7 5 4 5 8 6 7 6 4 2 6 1 4 7 2 5 7 4 6 4 5 1 5
[351] 10 8 7 5 4 6 4 4 7 5 4 3 1 6 2 5 3 3 3 7 4 3 7 8 4
[376] 7 3 1 4 4 7 6 7 2 4 5 1 3 12 4 2 2 8 7 6 7 6 3 5 4
```

```
length(PQ1)
```

```
[1] 400
```

```
sum(PQ1)
```

```
[1] 1872
```

```
mean(PQ1)
```

```
[1] 4.68
```

```
freqs=table(PQ1)      # the first row here are just labels!  
freqs
```

PQ1

1	2	3	4	5	6	7	8	9	10	11	12
20	43	53	86	70	54	37	18	10	5	2	2

```
sum(freqs)
```

```
[1] 400
```

```
values=1:12  
sum(values*freqs)/sum(freqs)
```

```
[1] 4.68
```

Comparing Mean and Median

```
prices=scan("http://www.maths.usyd.edu.au/math1905/r/prices.txt")
prices
```

Read 108 items

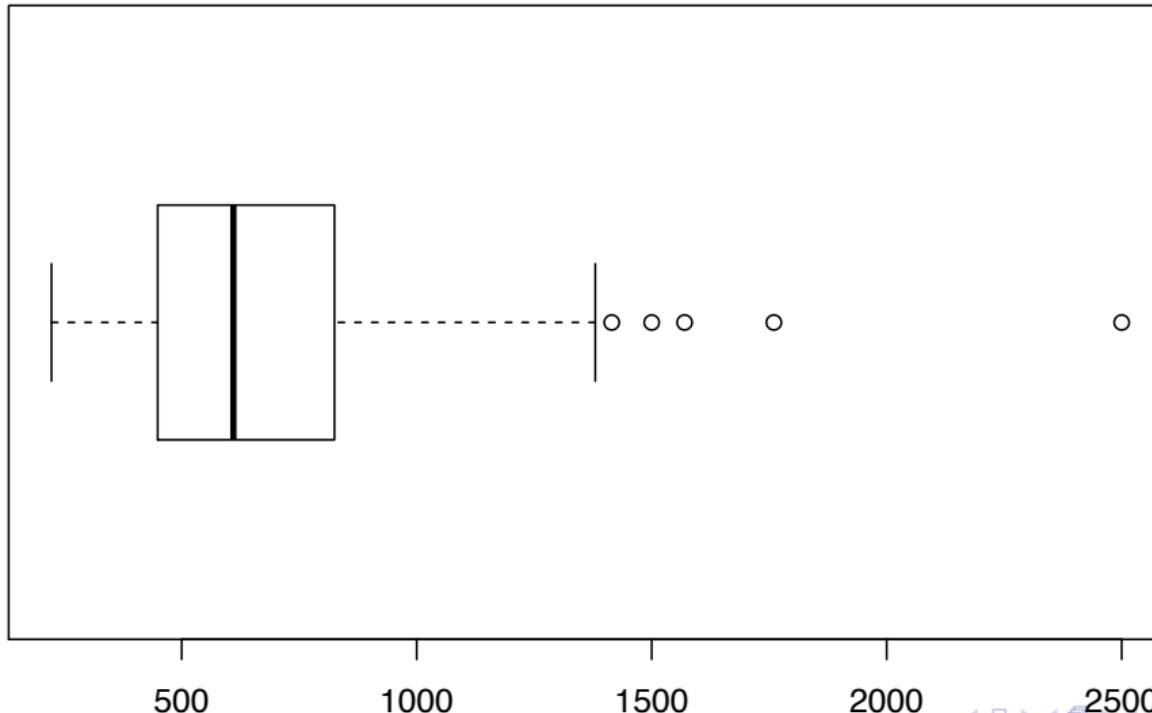
```
[1] 710.0 1070.0 692.0 237.5 1760.0 620.0 470.0 284.0 405.0 700.0
[11] 615.0 445.0 261.0 223.0 328.0 1210.0 1198.0 640.0 590.0 575.0
[21] 1500.0 480.0 495.0 730.0 738.0 850.0 1100.0 516.0 917.0 775.0
[31] 1350.0 653.0 1415.0 650.0 1280.0 715.0 580.0 635.0 2500.0 670.0
[41] 771.0 1225.0 631.0 582.0 500.0 575.0 1060.0 245.0 416.0 313.0
[51] 337.5 581.0 766.0 440.0 1125.0 388.0 897.0 875.0 690.0 712.0
[61] 400.0 326.0 491.5 605.0 367.0 390.0 1380.0 900.0 740.0 441.0
[71] 1351.0 620.0 343.0 605.0 960.0 527.0 500.0 500.0 572.0 555.0
[81] 379.0 690.0 448.0 390.0 571.0 862.0 600.0 675.0 384.0 485.0
[91] 410.0 370.0 1256.0 1570.0 450.0 1085.0 391.0 696.0 503.0 520.0
[101] 785.0 1210.0 922.0 496.0 566.0 800.0 358.0 637.0
```

```
stem(prices)
```

The decimal point is 2 digit(s) to the right of the |

2		24568
3		13344677889999
4		0112445557899
5		0000002236777888889
6		01122234445578999
7		001123447789
8		0568
9		00226
10		679
11		03
12		011368
13		558
14		2
15		07
16		
17		6
18		
19		
20		
21		
22		
23		
24		
25		0

```
boxplot(prices, horizontal=T)
```



```
sum(prices)
```

```
[1] 75794.5
```

```
length(prices)
```

```
[1] 108
```

```
sum(prices)/length(prices)
```

```
[1] 701.8009
```

```
mean(prices)
```

```
[1] 701.8009
```

```
median(prices)
```

```
[1] 610
```

Note the mean is greater than the median.

How do things change if we remove the largest observation?

```
pr2=sort(prices)[-108]  
mean(pr2)
```

```
[1] 684.9953
```

```
median(pr2)
```

```
[1] 605
```

```
(mean(pr2)-mean(prices))/mean(prices)
```

```
[1] -0.02394639
```

```
(median(pr2)-median(prices))/median(prices)
```

```
[1] -0.008196721
```

The relative change in the mean is about 3 times as much as for the median; it is “3 times as sensitive” to the effect of the outlier.

Outline

1 Welcome

2 Data Analysis

- Lecture 2
- Lecture 3
- Lecture 4

- Sample SD and Variance
- Bivariate Data
- The least-squares regression line

- Lecture 5

3 Probability

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

Variance

- Suppose x_1, x_2, \dots, x_n is a list of numbers.
- The **population variance** of these values is denoted and defined as

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the mean of the x_i 's.

- The **sample variance** of these values is denoted and defined as

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} \sigma_x^2.$$

- These measure the spread of the x_i 's in some sense: if $z_i = a + bx_i$ (for $i = 1, 2, \dots, n$) then

$$\sigma_z^2 = b^2 \sigma_x^2$$

and

$$s_z^2 = b^2 s_x^2.$$

Standard Deviation (SD)

- The **population standard deviation** of the x_i 's is

$$\sigma_x = \sqrt{\sigma_x^2}.$$

- The **sample standard deviation** of the x_i 's is

$$s_x = \sqrt{s_x^2}.$$

- Note that if $z_i = a + bx_i$ then

$$\sigma_z = |b|\sigma_x \text{ and } s_z = |b|s_x,$$

and so are “proper” measures of spread in this sense.

Sample SD in R

```
prices=scan("http://www.maths.usyd.edu.au/math1905/r/prices.txt")
sd(prices)
```

```
Read 108 items
[1] 367.3027
```

```
sum((prices-mean(prices))^2)/107
```

```
[1] 134911.3
```

```
sqrt(sum((prices-mean(prices))^2)/107)
```

```
[1] 367.3027
```

```
fivenum(prices)
```

```
[1] 223 449 610 825 2500
```

```
iqr=fivenum(prices)[4]-fivenum(prices)[2]
iqr
```

```
[1] 376
```

```
pr2=sort(prices)[-108]  
sd(pr2)
```

```
[1] 324.6445
```

```
fivenum(pr2)
```

```
[1] 223.0 449.0 605.0 792.5 1760.0
```

```
iqr2=fivenum(pr2)[4]-fivenum(pr2)[2]  
iqr2
```

```
[1] 343.5
```

```
(sd(prices)-sd(pr2))/sd(prices)
```

```
[1] 0.1161392
```

```
(iqr-iqr2)/iqr
```

```
[1] 0.08643617
```

The relative difference is not so extreme but the IQR is less sensitive to the effect of the outlier.



Why two versions?

- The “population” versions make more sense if the x_i ’s constitute a *whole population*.
- The “sample” versions make more sense if the x_i ’s are regarded as a sample from some other population *and* we are using the sample variance (of the sample) as an *estimate* of the population variance (of the population)!

Sample Variance

The little simulation exercise below shows why we divide by $n - 1$ in the sample variance. Consider the elementary population $\{1, 2, \dots, 10\}$.

```
pop=1:10  
mu=mean(pop)  
mu
```

```
[1] 5.5
```

```
sig.sq=mean((pop-mu)^2)  
sig.sq
```

```
[1] 8.25
```

It's "population variance" σ^2 is 8.25.

- Suppose now that the “population variance” σ^2 is in fact unknown but we wish to estimate it based on a random sample taken *with replacement* of size 2.
- One way to estimate σ^2 is to form the mean-squared difference from the mean, that is use $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ (the “population variance of the sample”) where $n = 2$, X_1, X_2 are the sample and \bar{X} is their average.
- The little simulation below does this 10000 times: the estimates are saved each time in the vector `p.var` (for “population variance”)

```
p.var=0
for(i in 1:10000){
  samp=sample(pop, size=2, replace=T)
  m=mean(samp)
  p.var[i]=mean((samp-m)^2)
}
p.var[1:10]
```

```
[1] 9.00 0.25 2.25 1.00 9.00 1.00 1.00 4.00 0.25 6.25
```

```
mean(p.var)
```

```
[1] 4.008175
```

```
sig.sq
```

```
[1] 8.25
```

Note that on average, the estimates are too small.

However if we instead compute the “sample variance” $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ then:

```
s.var=2*p.var  
s.var[1:10]
```

```
[1] 18.0 0.5 4.5 2.0 18.0 2.0 2.0 8.0 0.5 12.5
```

```
mean(s.var)
```

```
[1] 8.01635
```

```
sig.sq
```

```
[1] 8.25
```

The moral of the story is that if a dataset can be interpreted as a random sample from some population, then the *sample variance of the sample* is a better estimator of the *population variance of the population*.

Analysing bivariate data

- Suppose we have ordered pairs $(x_1, y_1), \dots, (x_n, y_n)$ and we wish to summarise/describe any apparent relationship that might exist between the 2 variables.
- It may be of some value to “model” each $y_i = f(x_i) + \varepsilon_i$ for
 - ▶ some “nice” function $f(\cdot)$ and
 - ▶ “random” errors ε_i .
- The if we are given a “fresh” x -value we may be able to use our model to give a better prediction of the corresponding y -value.

- How can we “pick” or “fit” a nice function to the points?
- We can separate this procedure into 2 “steps”:
 - ① find a function from a certain class which is “closest” to the data in some sense;
 - ② assessing whether the resultant *residuals* $\varepsilon_i = y_i - f(x_i)$ appear to be “random”.
- We firstly focus on 1.

The principle of least squares

- For given points $(x_1, y_1), \dots, (x_n, y_n)$, how do we fit a *line* to these points?
- For any candidate line $y = a + bx$ we obtain a corresponding set of *residuals* (see sketch)

$$\varepsilon_1 = y_1 - (a + bx_1),$$

⋮

$$\varepsilon_n = y_n - (a + bx_n).$$

- A “close fit” gives a set of residuals which are *small* in some sense.
- Can we find an ‘a’ and ‘b’ to minimise the **sum of squares of residuals**:

$$S(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2 ?$$

- The answer is “YES” so long as the x_i ’s are *not all the same*.

Two-step minimisation

- We can think of the minimisation as occurring in **two steps**:

$$\begin{aligned}\min_{a,b} S(a, b) &= \min_b \left[\min_a S(a, b) \right] \\ &= \min_b \left[\hat{S}(b) \right].\end{aligned}$$

Inner minimisation

As a function of a ,

$$\begin{aligned} S_b(a) = S(a, b) &= \sum_{i=1}^n (y_i - a - bx_i)^2 \\ &= \sum_{i=1}^n [(y_i - bx_i) - a]^2 \\ &= \sum_{i=1}^n [(y_i - bx_i)^2 - 2(y_i - bx_i)a + a^2] \\ &= \sum_{i=1}^n (y_i - bx_i)^2 - 2a \sum_{i=1}^n (y_i - bx_i) + na^2 \\ &= S_2 - 2aS_1 + na^2 \end{aligned}$$

is a **parabola**.

- It is minimised at that value a such that

$$0 = \frac{\partial S_b}{\partial a} = -2S_1 + 2na = -2 \sum_{i=1}^n (y_i - bx_i) + 2na = 0$$

that is

$$a = \frac{1}{n} \sum_{i=1}^n (y_i - bx_i) = \bar{y} - b\bar{x}.$$

- Replacing a with $\bar{y} - b\bar{x}$ gives a line with equation

$$y = a + bx = (\bar{y} - b\bar{x}) + bx = \bar{y} + b(x - \bar{x}).$$

- So for any candidate slope b , the “best” intercept is such that the line goes through the point (\bar{x}, \bar{y}) ,
 - ▶ to see this, substitute $x = \bar{x}$ and $y = \bar{y}$ into the equation above; these points satisfy the equation!

Outer minimisation

As a function of b , the function

$$\begin{aligned}\hat{S}(b) &= \sum_{i=1}^n [y_i - (\bar{y} - b\bar{x}) - bx_i]^2 \\ &= \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2b \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= S_{yy} - 2bS_{xy} + b^2 S_{xx}\end{aligned}$$

is also a **parabola**.

It is minimised at that value b such that

$$0 = \frac{\partial \hat{S}(b)}{\partial b} = -2S_{xy} + 2bS_{xx}$$

that is

$$b = \frac{S_{xy}}{S_{xx}}.$$

The least-squares regression line

In summary:

The least-squares regression line/ based on points $(x_1, y_1), \dots, (x_n, y_n)$ is given by $y = a + bx$ where

$$b = S_{xy}/S_{xx},$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

and

$$a = \bar{y} - b\bar{x}.$$

As usual, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Some comments

- Note that if we reflect the points in the “ $y = x$ line”, i.e. swap the roles of y and x , the new least-squares regression line is **not** obtained by reflecting the old one in the “ $y = x$ line”.
- The “new slope” would be S_{xy}/S_{yy} which is **not** $1/b$

Bivariate data in R: scatterplots

- The .csv (comma-separated-variables) files are obtained from the Bureau of Meteorology website and contain daily maximum and minimum temperatures for all of 2012:

```
max2012=read.csv("http://www.sydney.edu.au/science/maths/math1905/r>IDCJAC0010_066062_2012_Data.csv")
min2012=read.csv("http://www.sydney.edu.au/science/maths/math1905/r>IDCJAC0011_066062_2012_Data.csv")
```

- These commands define special R objects called *data frames*. These are sort of like matrices, but each column can be of a different type (numeric, logical, character, etc.).

The columns are named:

```
names(max2012)
```

```
[1] "Product.code"  
[2] "Bureau.of.Meteorology.station.number"  
[3] "Year"  
[4] "Month"  
[5] "Day"  
[6] "Maximum.temperature..Degree.C."  
[7] "Days.of.accumulation.of.maximum.temperature"  
[8] "Quality"
```

```
names(min2012)
```

```
[1] "Product.code"  
[2] "Bureau.of.Meteorology.station.number"  
[3] "Year"  
[4] "Month"  
[5] "Day"  
[6] "Minimum.temperature..Degree.C."  
[7] "Days.of.accumulation.of.minimum.temperature"  
[8] "Quality"
```

The following show the first 3 rows of each:

```
max2012[1:3,]
```

	Product.code	Bureau.of.Meteorology.station.number	Year	Month	Day
1	IDCJAC0010	66062	2012	1	1
2	IDCJAC0010	66062	2012	1	2
3	IDCJAC0010	66062	2012	1	3

	Maximum.temperature..Degree.C.	Days.of.accumulation.of.maximum.temperature
1	25.3	1
2	26.5	1
3	26.8	1

	Quality
1	Y
2	Y
3	Y

```
min2012[1:3 ,]
```

	Product.code	Bureau.of.Meteorology.station.number	Year	Month	Day
1	IDCJAC0011	66062	2012	1	1
2	IDCJAC0011	66062	2012	1	2
3	IDCJAC0011	66062	2012	1	3
	Minimum.temperature..Degree.C.	Days.of.accumulation.of.minimum.temperature			
1	16.0				1
2	18.1				1
3	20.6				1
	Quality				
1	Y				
2	Y				
3	Y				

We can extract columns by appending a dollar sign and an abbreviated name (enough to uniquely identify the column):

```
y=max2012$Max  
x=min2012$Min
```

- Computing the least-squares regression line is a simple special case of fitting a (more general) *linear model*, the fitting of which is the job of the R function `lm()`.

```
fit=lm(y~x)
```

- The value returned by `lm()` is a special kind of R object, which certain *generic* R functions know what to do with.

- One such generic function is `summary()`:

```
summary(fit)
```

Call:
`lm(formula = y ~ x)`

Residuals:

Min	1Q	Median	3Q	Max
-7.1230	-1.7323	-0.0515	1.4903	10.7179

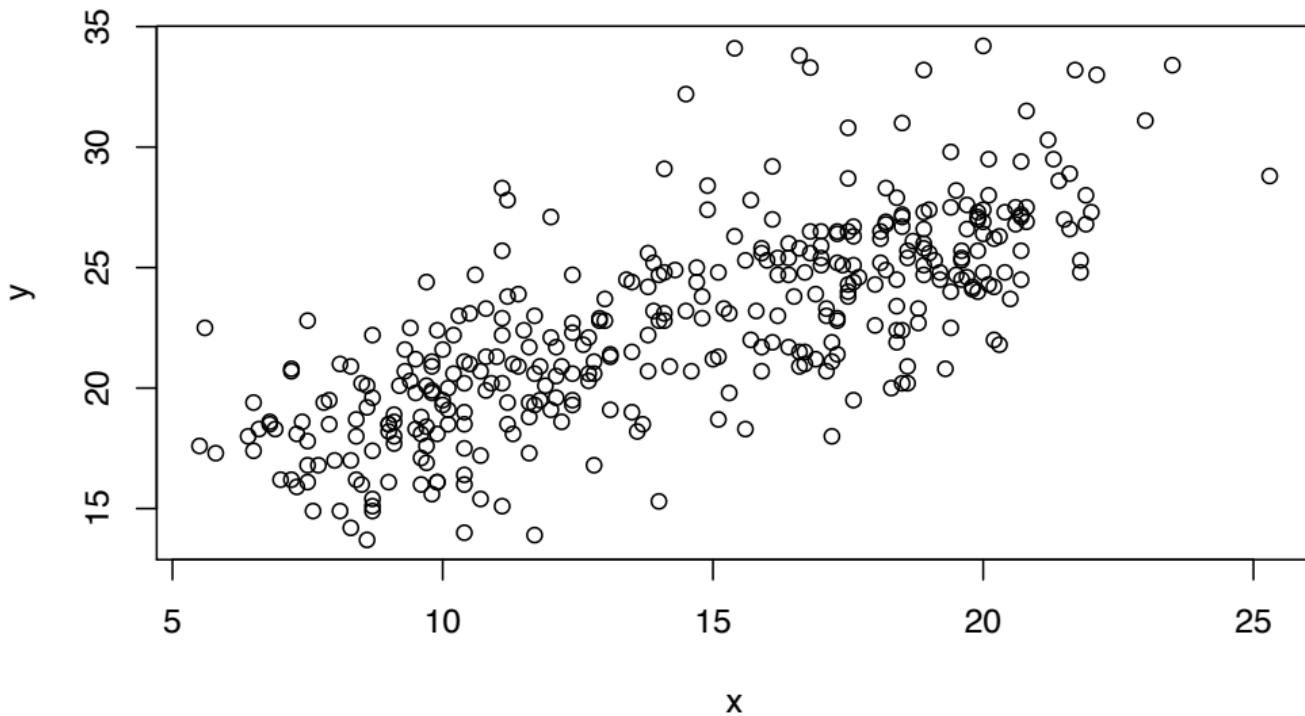
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	12.83121	0.48322	26.55	<2e-16 ***
x	0.68513	0.03214	21.32	<2e-16 ***

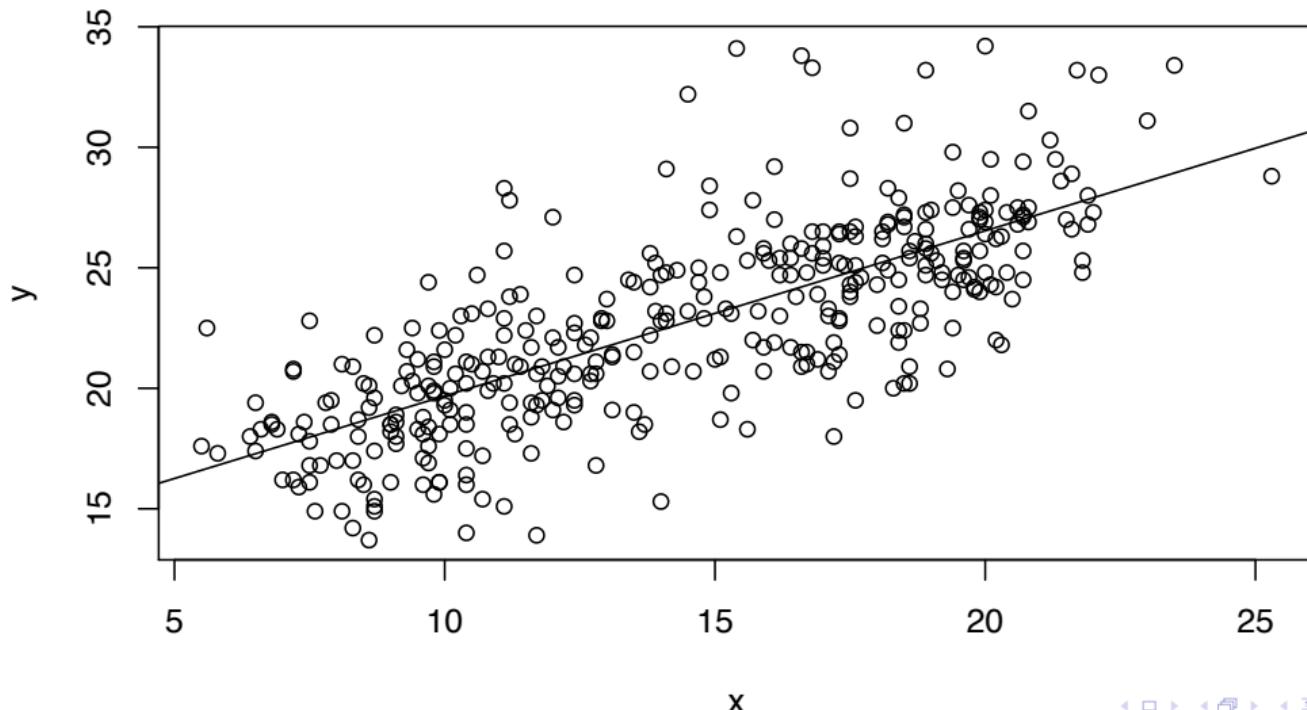
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1

Residual standard error: 2.743 on 364 degrees of freedom
Multiple R-squared: 0.5553, Adjusted R-squared: 0.5541
F-statistic: 454.5 on 1 and 364 DF, p-value: < 2.2e-16

```
plot(x, y)
```



```
plot(x,y)  
abline(fit)
```



- The function `summary()` is a generic function, in that it works on many different kinds of R objects, tailoring what it does (including how it prints) depending on what type of R object is passed to it.
- In this case, it prints an `lm.fit` summary, including the values of the intercept and slope. These appear in the object `coef(fit)` or `fit$coef`:

```
fit$coef
```

```
(Intercept)           x  
12.8312126   0.6851253
```

```
coef(fit)
```

```
(Intercept)           x  
12.8312126   0.6851253
```

In order to see what is going on “under the hood”, we can do the necessary calculations “by hand”:

```
Sxx = sum((x - mean(x))^2)
Sxy = sum((x - mean(x)) * (y - mean(y)))
Sxx
```

```
[1] 7286.243
```

```
Sxy
```

```
[1] 4991.99
```

```
b = Sxy / Sxx
b
```

```
[1] 0.6851253
```

```
a = mean(y) - b * mean(x)
a
```

```
[1] 12.83121
```

Outline

1 Welcome

2 Data Analysis

- Lecture 2
- Lecture 3
- Lecture 4
- Lecture 5

- Computing formulae
- Correlation
- Assessing the linear fit
- Residual plots

3 Probability

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

More on the mean and the median

- There is another way to compare the mean and the median: each can be viewed as the “minimiser” of a certain “distance”.
- Suppose we have a dataset x_1, \dots, x_n . Then one way to measure the “distance” of a constant c from these is the sum of squared deviations:

$$Q(c) = \sum_{i=1}^n (x_i - c)^2 = \sum_{i=1}^n x_i^2 - 2c \sum_{i=1}^n x_i + nc^2$$

which is a *quadratic* function of c .

- We can find the value of c that minimises $Q(\cdot)$ by differentiating and setting to 0:

$$Q'(c) = -2 \sum_{i=1}^n x_i + 2nc = 0 \Rightarrow c = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

```
x=c(1,2,3,3,4,5,7,9,10,15,20)
```

```
x
```

```
[1] 1 2 3 3 4 5 7 9 10 15 20
```

```
c=0:200/10
```

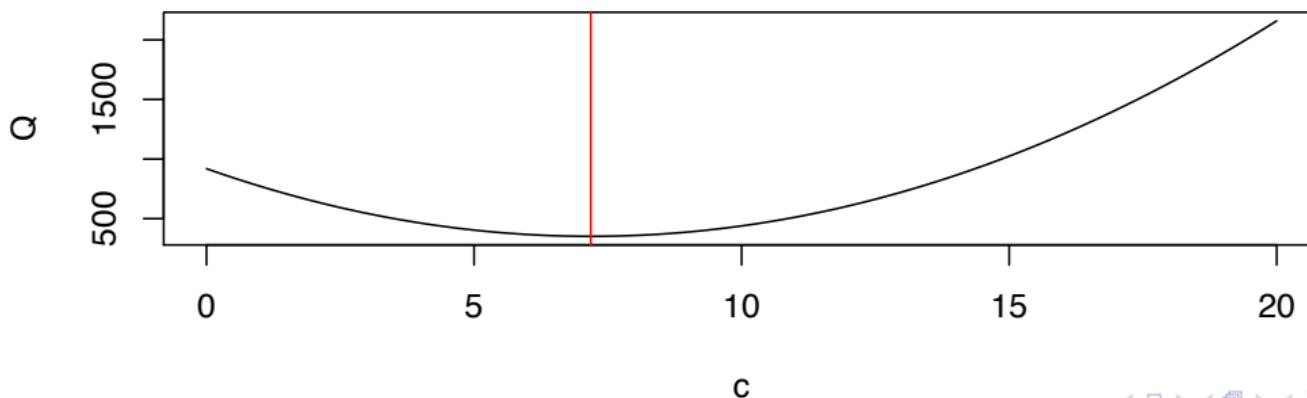
```
c
```

```
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4  
[16] 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9  
[31] 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4  
[46] 4.5 4.6 4.7 4.8 4.9 5.0 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9  
[61] 6.0 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8 6.9 7.0 7.1 7.2 7.3 7.4  
[76] 7.5 7.6 7.7 7.8 7.9 8.0 8.1 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9  
[91] 9.0 9.1 9.2 9.3 9.4 9.5 9.6 9.7 9.8 9.9 10.0 10.1 10.2 10.3 10.4  
[106] 10.5 10.6 10.7 10.8 10.9 11.0 11.1 11.2 11.3 11.4 11.5 11.6 11.7 11.8 11.9  
[121] 12.0 12.1 12.2 12.3 12.4 12.5 12.6 12.7 12.8 12.9 13.0 13.1 13.2 13.3 13.4  
[136] 13.5 13.6 13.7 13.8 13.9 14.0 14.1 14.2 14.3 14.4 14.5 14.6 14.7 14.8 14.9  
[151] 15.0 15.1 15.2 15.3 15.4 15.5 15.6 15.7 15.8 15.9 16.0 16.1 16.2 16.3 16.4  
[166] 16.5 16.6 16.7 16.8 16.9 17.0 17.1 17.2 17.3 17.4 17.5 17.6 17.7 17.8 17.9  
[181] 18.0 18.1 18.2 18.3 18.4 18.5 18.6 18.7 18.8 18.9 19.0 19.1 19.2 19.3 19.4  
[196] 19.5 19.6 19.7 19.8 19.9 20.0
```

```
mean(x)
```

```
[1] 7.181818
```

```
Q=0
for (i in 1:length(c)){
  Q[i]=sum((x-c[i])^2)      # This computes Q(c) for each value in the vector c
}
plot(c,Q,type="l")
abline(v=mean(x),col="red")
```



Sum of absolute values

- Consider the *piecewise linear* distance function

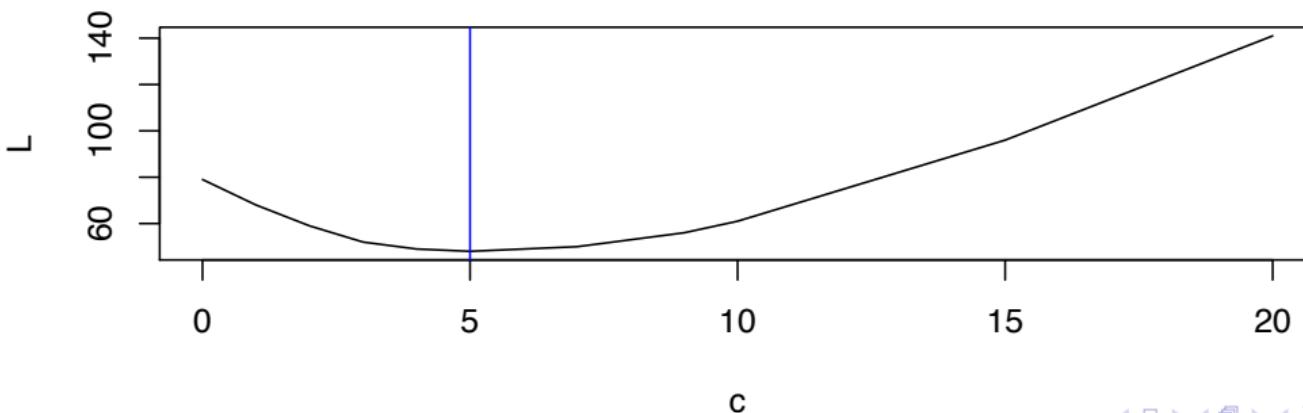
$$L(c) = \sum_{i=1}^n |x_i - c|$$

- It turns out the value of c that minimises $L(c)$ is the *median* of the x_i 's:

```
median(x)
```

```
[1] 5
```

```
L=0
for (i in 1:length(c)){
  L[i]=sum(abs(x-c[i])) # This computes  $L(c)$  for each value in the vector c
}
plot(c,L,type="l")
abline(v=median(x),col="blue")
```

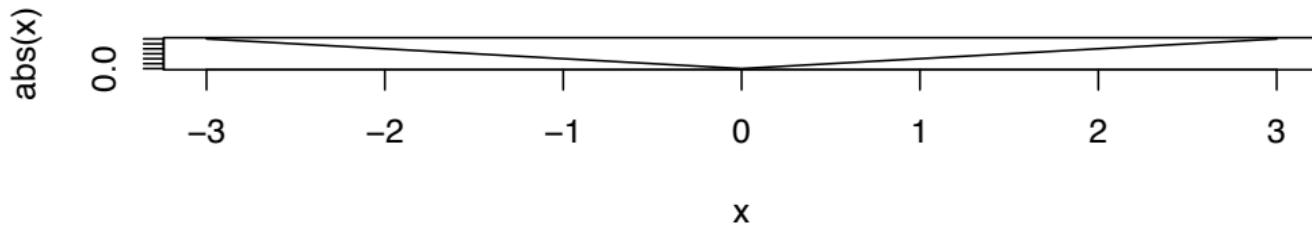


How does this work (not examinable)?

- The function $f(x) = |x|$ is
 - continuous** at all real x ;
 - differentiable** at all real x **except** $x = 0$, in particular

$$f'(x) = \begin{cases} -1 & \text{for } x < 0, \\ +1 & \text{for } x > 0. \end{cases}$$

```
x=-300:300/100
plot(x,abs(x),type="l")
```



- The function $L(c) = \sum_{i=1}^n |x_i - c|$, as a function of c is the *sum* of n functions $|x_1 - c|, |x_2 - c|, \dots, |x_n - c|$:
 - each of these are continuous for all c , so $L(c)$ is continuous for all c also;
 - $|x_i - c|$ is differentiable everywhere *except* at $c = x_i$:

$$\frac{d(|x_i - c|)}{dc} = \begin{cases} -1 & \text{for } c < x_i, \\ +1 & \text{for } c > x_i. \end{cases}$$

- Thus so long as c is not equal to any x_i , $L'(c)$ is well-defined and is given by

$$\begin{aligned} L'(c) &= \sum_{i=1}^n \frac{d(|x_i - c|)}{dc} \\ &= (\text{no. } x_i \text{'s less than } c) - (\text{no. } x_i \text{'s greater than } c) \end{aligned}$$

- Thus when $n = 2k + 1$ is odd (and all x_i 's distinct), the derivative is equal to
 - ▶ $-n$ for $c < x_{(1)}$;
 - ▶ $-n + 2$ for $x_{(1)} < c < x_{(2)}$;
 - ▶ \dots
 - ▶ $-n + 2k (< 0)$ for $x_{(k)} < c < x_{(k+1)}$;
 - ▶ $-n + 2k + 2 (> 0)$ for $x_{(k+1)} < c < x_{(k+2)}$;
 - ▶ \dots
 - ▶ $+n$ for $c > x_{(n)}$.
- So $L'(c)$ changes sign exactly at the median $x_{(k+1)}$.
- When $n = 2k$ is even, then we have that $L'(c) = 0$ for the whole interval $(x_{(k)}, x_{(k+1)})$, so *any value* in that interval attains the minimum.

Least absolute deviation regression (still not examinable)

- This can be used to derive an algorithm to compute a **least absolute deviation regression line**:
 - ▶ for each candidate slope b , choose a to minimise

$$L_b(a) = \sum_{i=1}^n |(y_i - bx_i) - a|$$

- According to the results above, this a is given by

$$a = \text{median}_{i=1,\dots,n} |y_i - bx_i|$$

- We would need a computer to then find the best slope b .

Computing formulae (examinable again)

- Both versions of variance and SD involve the sum of squared deviations from the mean also denoted S_{xx} in a least-squares regression context:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

- As shown in Tutorial 2 the identities

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = \sum_{i=1}^n x_i^2 - n(\bar{x})^2$$

always hold and indeed if this needs to be computed "by hand", the *middle* version is the most "efficient" way to do so (it minimises the effects of rounding errors).

- As also seen in Tutorial 2 5(c) the quantity

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}$$

and this far right-hand side is the preferred way to compute this “by hand”.

Correlation

- The slope of the least-squares regression line depends crucially on the scale of the x_i 's and y_i 's.
- E.g. if the y_i 's are $^{\circ}\text{C}$ and are changed to $^{\circ}\text{F}$ then the corresponding least-squares slope (and intercept) will change.
- The **correlation coefficient** can be viewed as a scale-free version of the least-squares slope.

- It can be expressed as r where

$$\text{l.s. slope } = b = r \frac{s_y}{s_x}.$$

- The ratio s_y/s_x accounts for all effects of scale.
- Whatever remains is “absorbed” into r .
- If r needs to be computed by hand, note that

$$r = b \frac{s_x}{s_y} = \frac{S_{xy}}{S_{xx}} \sqrt{\frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{S_{xx}} \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

Summary

The **correlation coefficient** associated with ordered pairs $(x_1, y_1), \dots, (x_n, y_n)$ is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

and

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2.$$

Properties/interpretation

- Recall that for any a and b , the “residual sum of squares” satisfies

$$0 \leq S(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2$$

with *equality* only if $y = a + bx_i$ for each i , i.e. the points are all on a line.

- Also recall that the *minimum* of this is (according to the “outer minimisation” last lecture, see slide 78) is attained at $b = S_{xy}/S_{xx}$:

$$\begin{aligned} 0 \leq \min_{a,b} S(a, b) &= \min_b \hat{S}(b) = \min_b [S_{yy} - 2bS_{xy} + b^2 S_{xx}] \\ &= S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} \\ &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} \end{aligned} \tag{*}$$

after substituting in the minimiser $b = S_{xy}/S_{xx}$.

- Note that from (*),
we get

$$\frac{S_{xy}^2}{S_{xx}} \leq S_{yy} \Rightarrow \frac{S_{xy}^2}{S_{xx}S_{yy}} \leq 1$$

that is

$$r^2 \leq 1$$

or $-1 \leq r \leq 1$, with equality only if the points lie on a straight line.

- More precisely
 - ▶ $r = 1$ if and only if the points lie on a line with *positive* slope;
 - ▶ $r = -1$ if and only if the points lie on a line with *negative* slope;

note the slope cannot be zero in either case!

Proportion of “variation” explained by the linear relationship

- The “spread” or “variation” of the y_i ’s can be characterised by the quantity

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = (n - 1)s_y^2.$$

This is sometimes called the *total sum of squares*.

- The quantity

$$S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

on the right-hand side of (*) above is in fact the *residual sum of squares*, the sum of squares of residuals about the least-squares line.

- The difference between these

$$S_{yy} - \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{S_{xy}^2}{S_{xx}}$$

is called the *regression sum of squares*.

Decomposition

- We thus have the following decomposition of the total sum of squares:

$$\begin{aligned}\text{Total sum of squares} &= \text{regression sum of squares} \\ &\quad + \text{residual sum of squares.}\end{aligned}$$

- If the residual sum of squares is zero, the points are on a straight line and **all** the variation is explained by the linear relationship.
- If the regression sum of squares is zero, the least-squares line has slope zero and indeed the linear relationship explains **none** of the variation.
- The **proportion of variation** explained by the linear relationship is

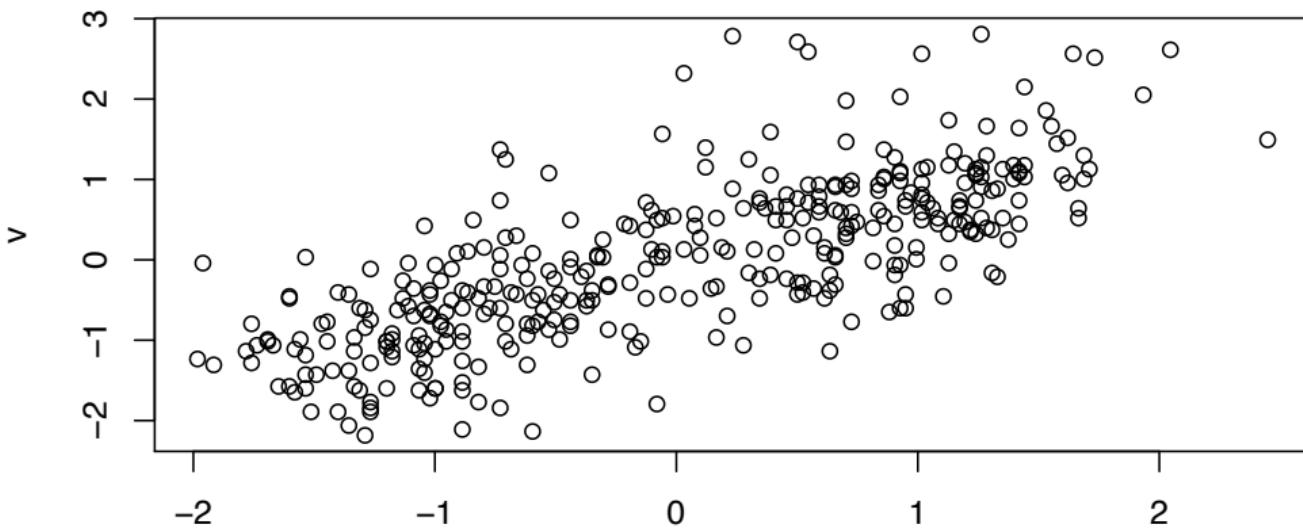
$$\frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = r^2$$

and this provides another interpretation of r^2 .

Correlation in R

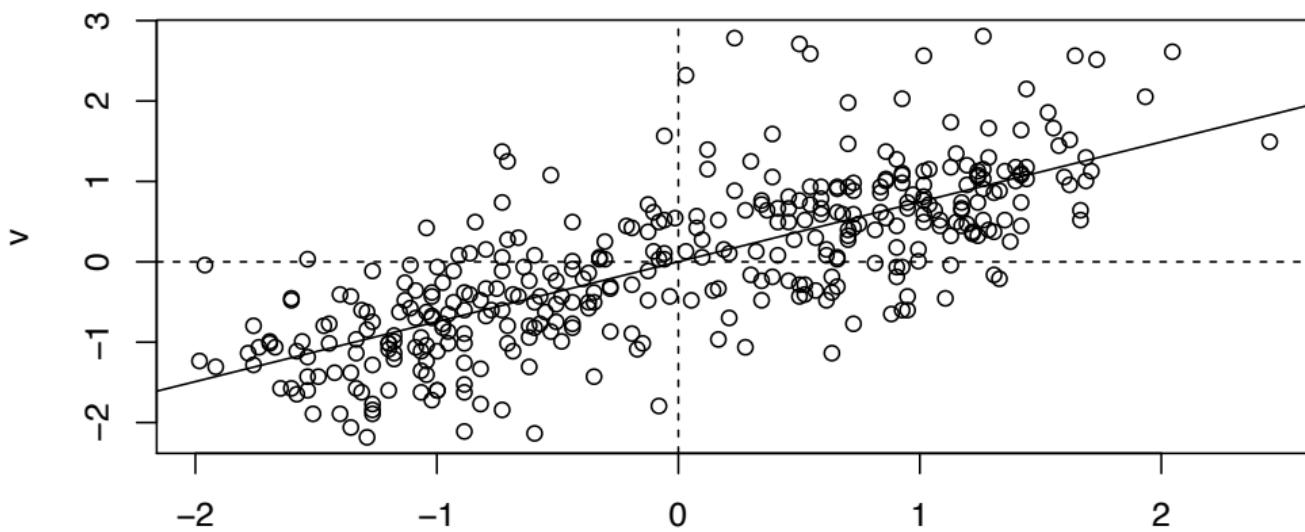
Recall x and y from lecture 4. We firstly standardise both variables, creating u and v :

```
u=(x-mean(x))/sd(x)
v=(y-mean(y))/sd(y)
plot(u,v)
```



We then compute the least-squares line for these:

```
plot(u,v)
abline(h=0,lty=2) # lty=2 makes the lines dashed
abline(v=0,lty=2)
fit.st=lm(v~u)
abline(fit.st)
```



```
coef(fit.st)
```

```
(Intercept)          u  
-7.266031e-17  7.451786e-01
```

- The scatterplot looks the same, except for the markings on the axes.
- Note the least-squares line goes through the origin on the plot, even though there is some rounding error in the lm.fit.

Also, note that the slope is exactly the correlation between the *original* variables:

```
cor(x, y)
```

```
[1] 0.7451786
```

```
Syy = sum((y - mean(y))^2)
Sxy / sqrt(Sxx * Syy)
```

```
[1] 0.7451786
```

Note also that given the sd's and correlation we can recover the least-squares line slope as follows:

```
r=cor(x,y)  
r
```

```
[1] 0.7451786
```

```
r*sd(y)/sd(x)
```

```
[1] 0.6851253
```

```
coef(fit)      # coefficients of the original fit
```

(Intercept)	x
12.8312126	0.6851253

Proportion of variation explained by the regression

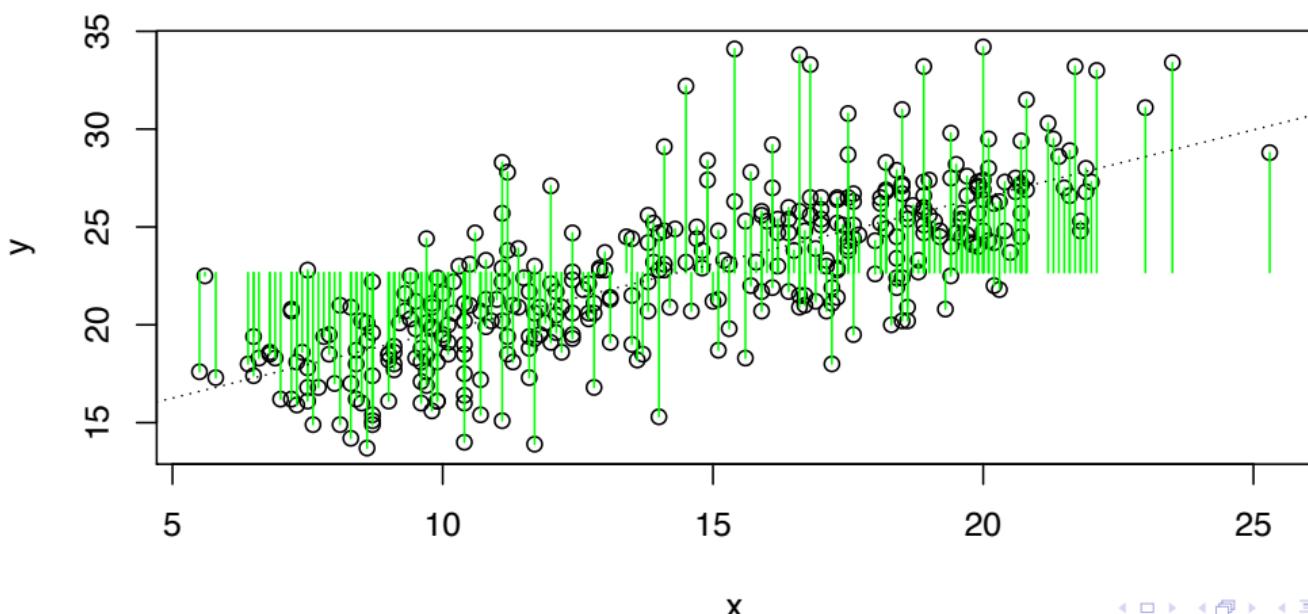
The average y-coordinate is

```
mean(y)
```

```
[1] 22.66913
```

The green lines show the vertical difference between each y-coordinate and the y-average:

```
plot(x,y)
for(i in 1:length(u)){
  lines(c(x[i],x[i]),c(mean(y),y[i]),col="green")
}
abline(fit,lty=3)
```



The sum of squares of these vertical differences is Syy , the “Total Sum of Squares”:

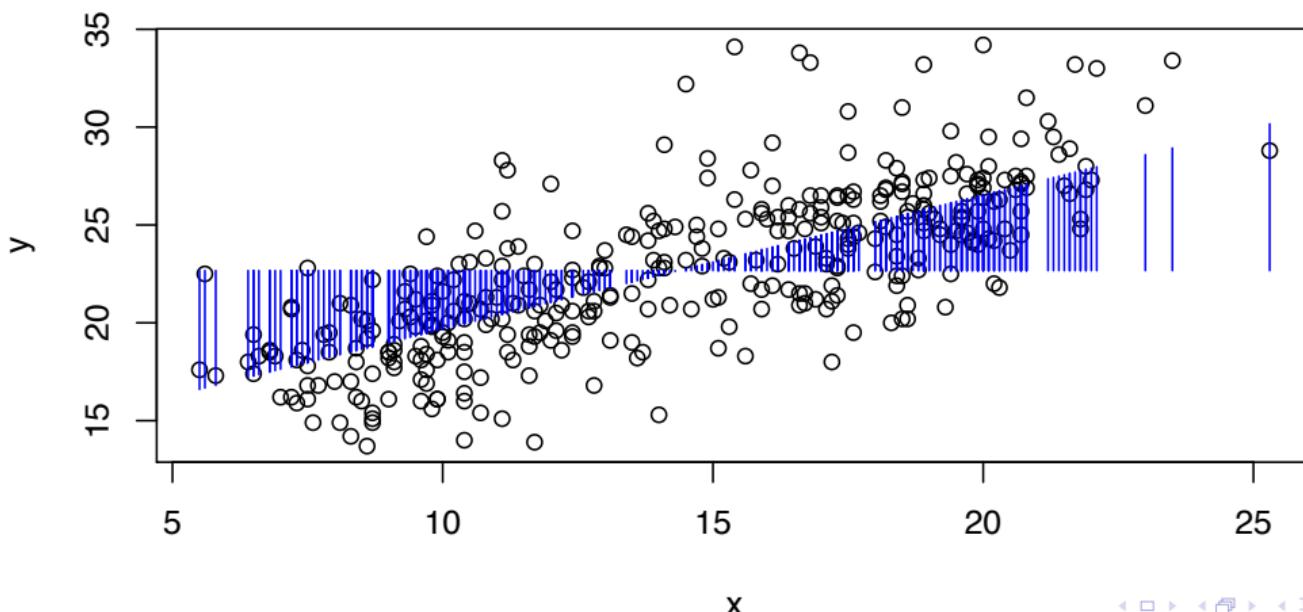
```
Syy = sum((y - mean(y))^2)  
Syy
```

```
[1] 6159.181
```

This measures the “variation” in the y-coordinates in some sense; if we divide it by $(n - 1)$ we get the sample variance of the y's.

- Writing $\hat{y}_i = a + bx_i$ for the i -th fitted value and $\hat{\varepsilon}_i = y_i - \hat{y}_i$ for the i -th residual, we decompose each green difference into two components: $y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{blue}} + \underbrace{\hat{\varepsilon}_i}_{\text{red}}$
- The first component is the part explained by the linear relationship. These are shown for each point in blue below:

```
plot(x,y)
yhat=fitted(fit)
for(i in 1:length(u)){
  lines(c(x[i],x[i]),c(mean(y),yhat[i]),col="blue")
}
```



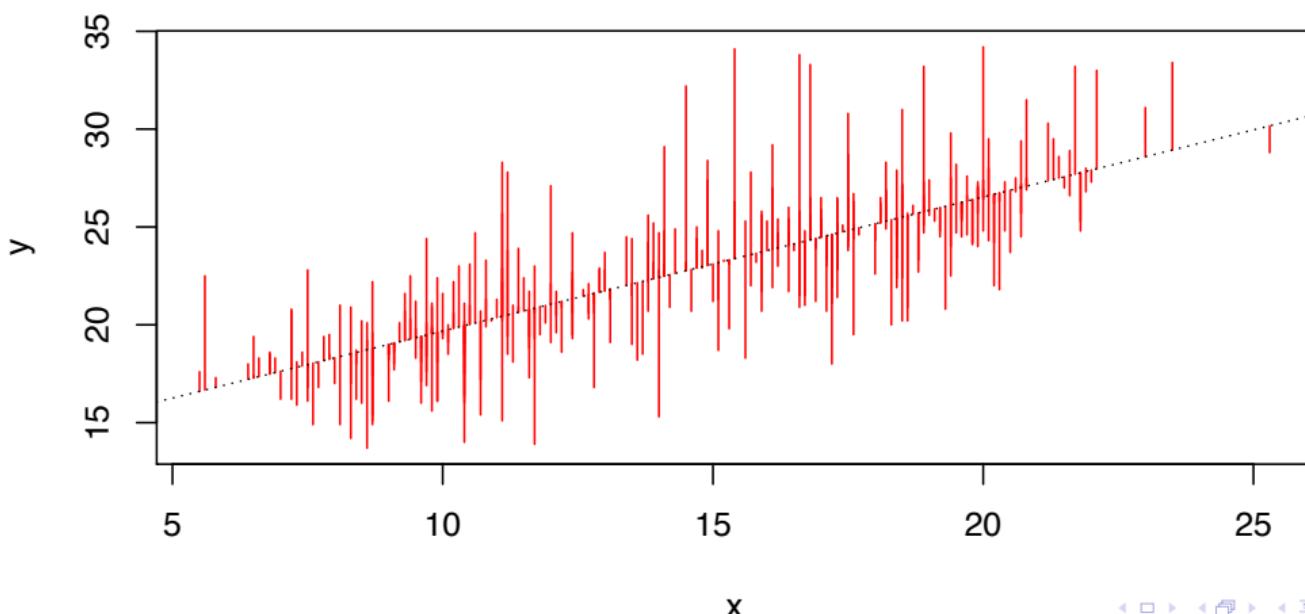
The sum of squares of the blue components is the *Regression sum of squares*:

```
RegSS = sum((yhat - mean(y))^2)  
RegSS
```

```
[1] 3420.139
```

The second component is the residual, the component of the variation *not* explained by the linear relationship and is shown for each point below in red:

```
plot(x,y,pch="")
for(i in 1:length(u)){
  lines(c(x[i],x[i]),c(yhat[i],y[i]),col="red")
}
abline(fit,lty=3)
```



The sum of squares of these is the *Residual sum of squares*:

```
ResSS = sum((y - yhat)^2)
```

```
ResSS
```

```
[1] 2739.042
```

```
RegSS
```

```
[1] 3420.139
```

```
RegSS + ResSS
```

```
[1] 6159.181
```

Note that the Regression and Residual sums of squares add up to the Total sum of squares S_{yy} :

```
Syy
```

```
[1] 6159.181
```

Moreover, the *proportion* of the Total “explained” by the Regression is simply r^2 :

```
RegSS/Syy
```

```
[1] 0.5552912
```

```
r^2
```

```
[1] 0.5552912
```

Interpreting the correlation coefficient

- We can interpret the numerical value of r^2 as the “proportion of variability explained by the linear relationship”. In other words if r^2 is close to 1, then the points are close to a straight line.
- However this does not mean that
 - ▶ if r^2 is close to 1 the linear relationship provides a good fit
 - ▶ if r^2 is near zero there is no “relationship” between the y_i 's and x_i 's (but it *does* mean there is no *linear* relationship!).

Assessing the linear fit

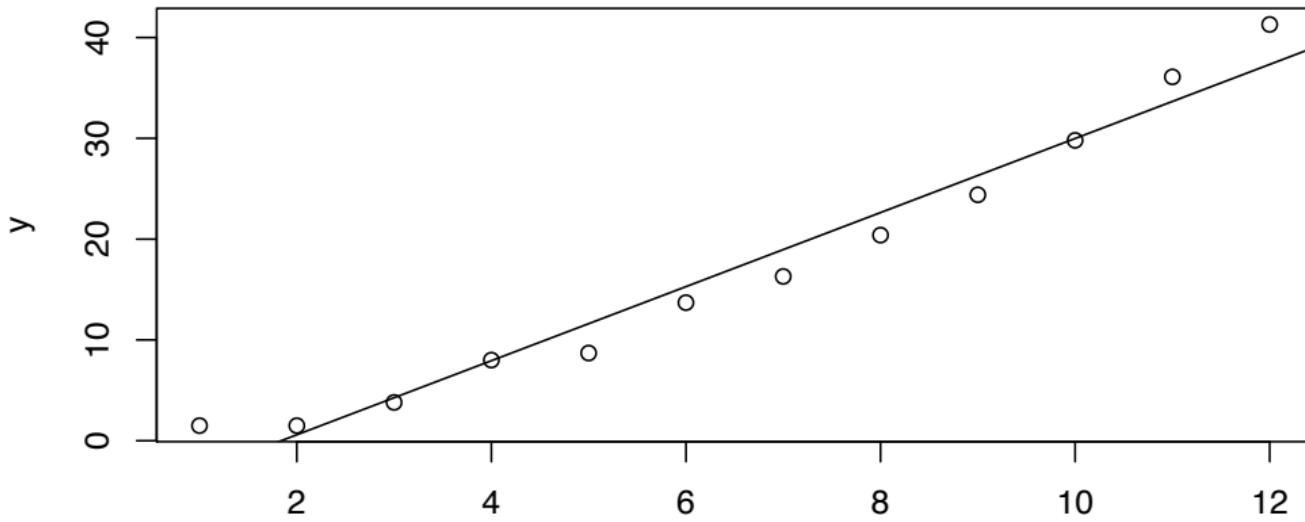
Recall our initial aim, to find an intercept and slope so that in fitting a line $y = a + bx$ to a set of points $(x_1, y_1), \dots, (x_n, y_n)$, the resultant residuals $y_i - (a + bx_i)$ look like “random errors”. Consider the following example:

```
x <- 1:12
y <- c(1.5, 1.5, 3.8, 8, 8.7, 13.7, 16.3, 20.4, 24.4, 29.8, 36.1, 41.3)
cor(x,y)
```

```
[1] 0.9822209
```

The correlation suggests there is a strong linear association between this x and y . However plotting we see that

```
plot(x,y)
fit=lm(y~x)
abline(fit)
```



To see the non-linear relationship more clearly, we can fit a **quadratic** regression function as follows:

```
z=x^2  
fit.quad=lm(y~x+z)  
summary(fit.quad)
```

Call:

```
lm(formula = y ~ x + z)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.92740	-0.56879	-0.09743	0.55711	1.31071

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	-0.07045	0.81779	-0.086	0.9332							
x	0.80792	0.28923	2.793	0.0209 *							
z	0.22050	0.02166	10.181	3.08e-06 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	,	1

Residual standard error: 0.7913 on 9 degrees of freedom

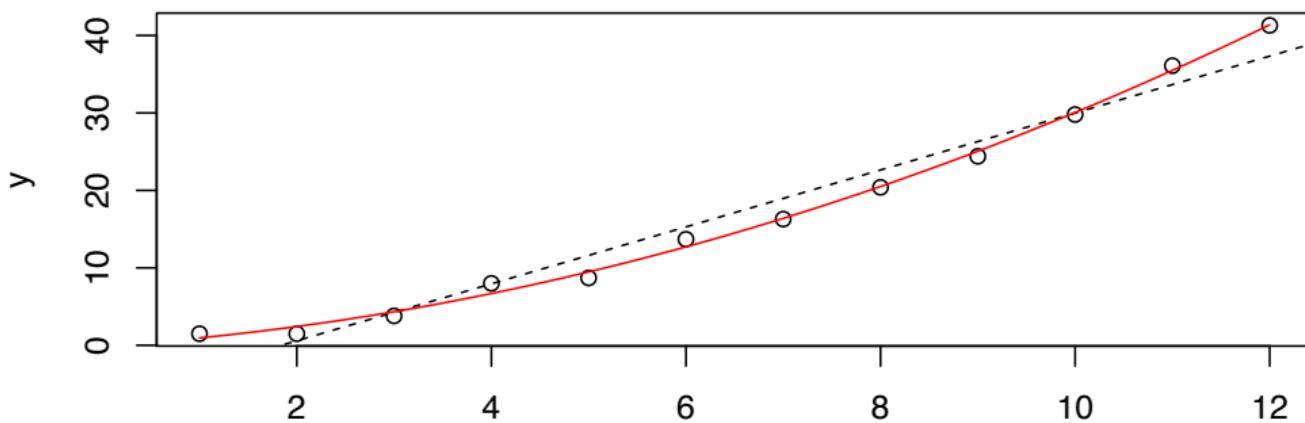
Multiple R-squared: 0.9972, Adjusted R-squared: 0.9966

F-statistic: 1594 on 2 and 9 DF, p-value: 3.335e-12

```
co=coef(fit.quad)  
co
```

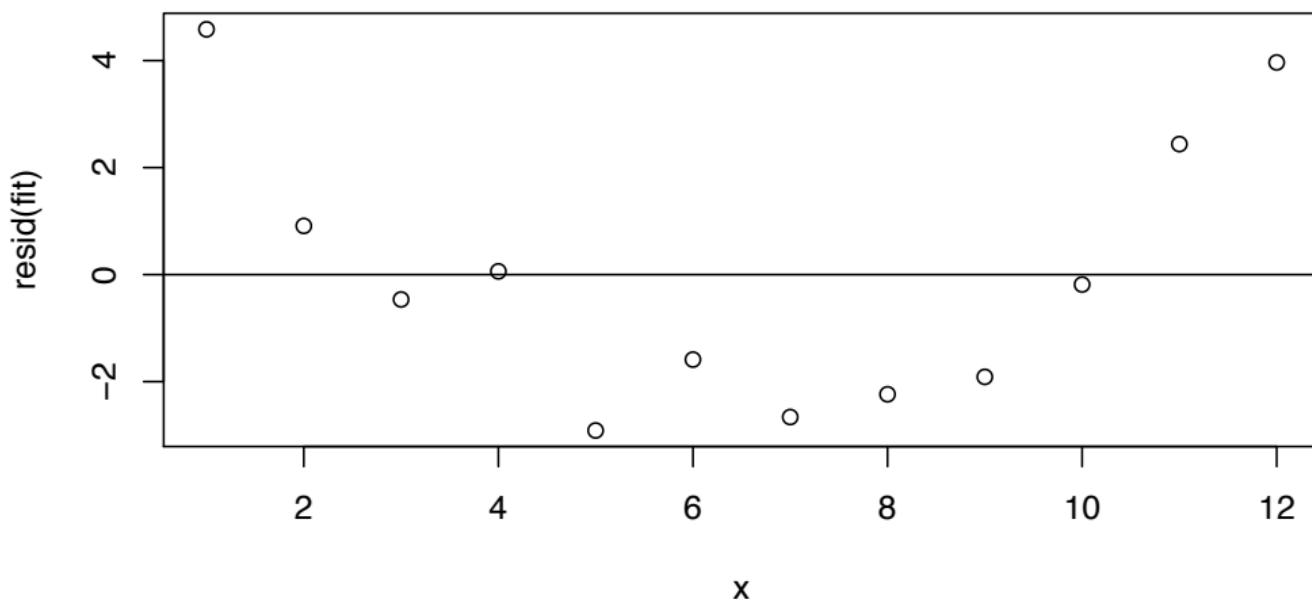
```
(Intercept)           x           z  
-0.07045455  0.80791708  0.22050450
```

```
plot(x,y)  
abline(fit,lty=2)  
curve(co[1]+co[2]*x+co[3]*x^2,add=TRUE,col="red")
```



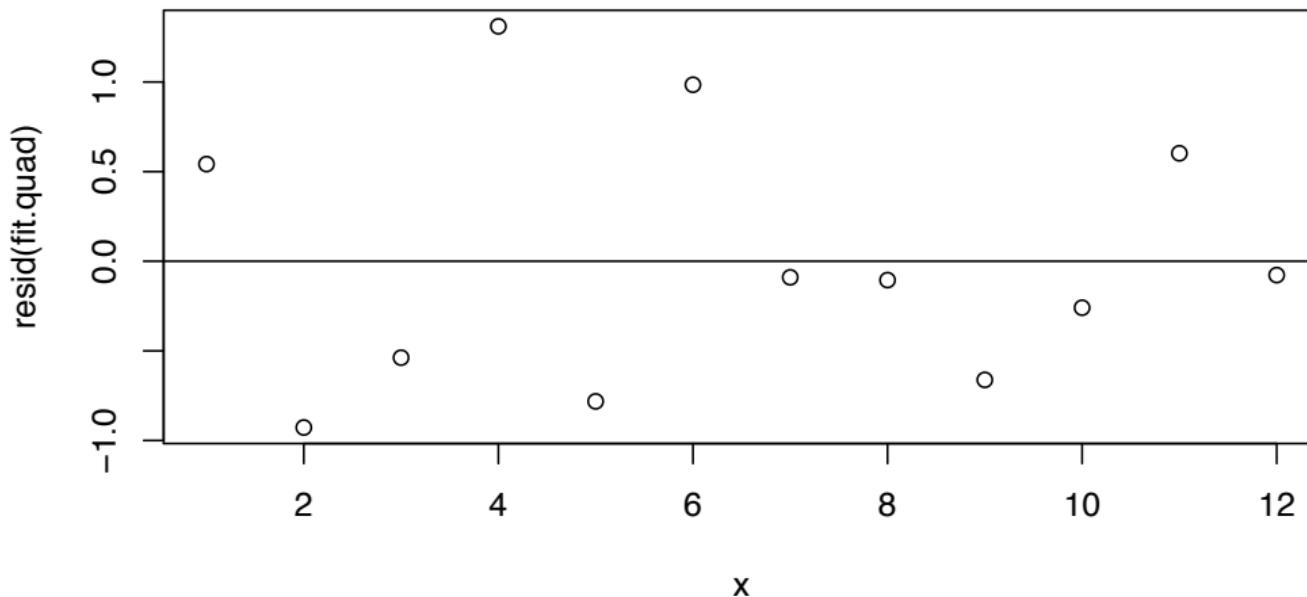
Firstly, the residual plot from the linear-only fit:

```
plot(x, resid(fit))  
abline(h=0)
```



Now the residuals from the quadratic fit:

```
plot(x,resid(fit.quad))  
abline(h=0)
```

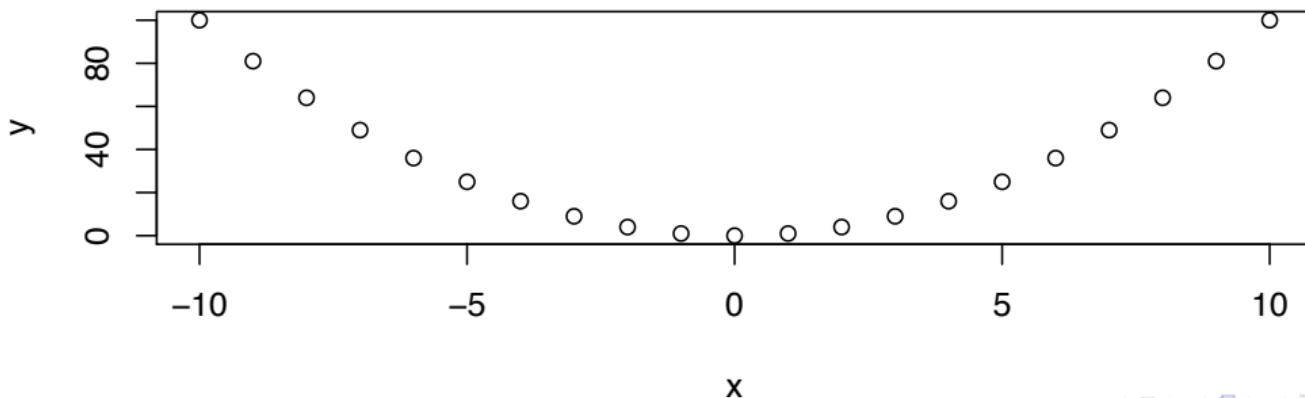


Note also that just because the correlation (squared) is close to zero, it doesn't mean there is not a strong relationship between the y_i 's and x_i 's:

```
x=-10:10  
y=x^2  
cor(x,y)
```

```
[1] 0
```

```
plot(x,y)
```



Anscombe Data

Consider finally the anscombe datasets:

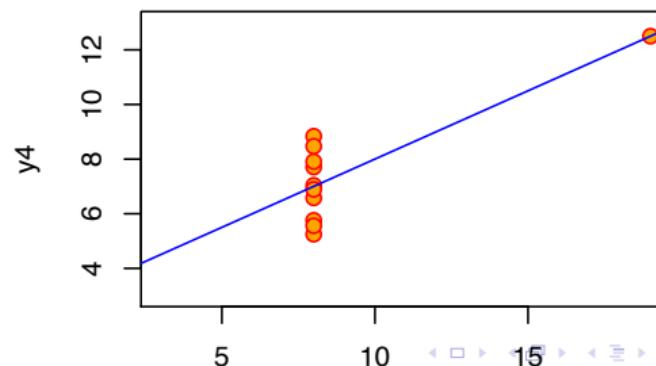
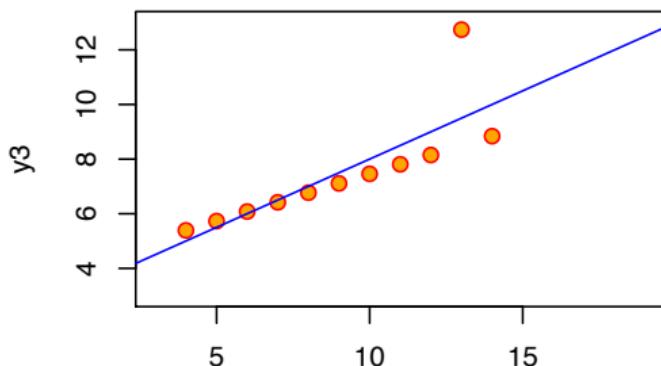
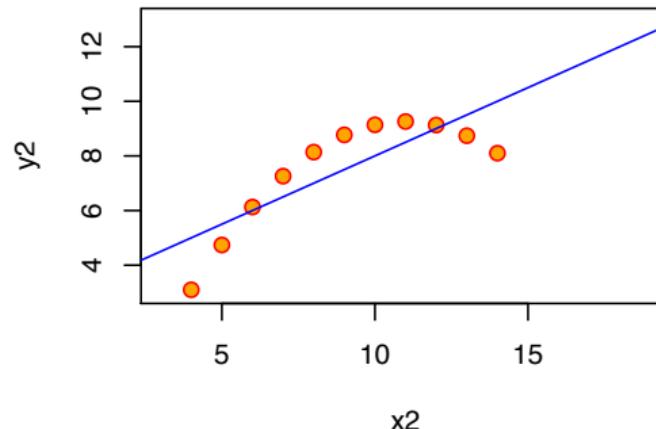
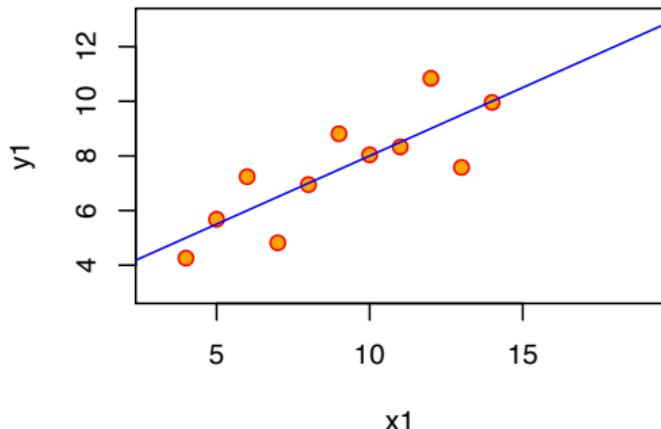
```
require(stats); require(graphics)  
anscombe
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

The summary statistics (means, sds, correlation) are all the same for these.

However plotting them reveals all:

Anscombe's 4 Regression data sets



Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
 - Lecture 6
 - Probability
 - Axioms of probability
 - Combinatorics
 - Some special sample spaces
 - Lecture 7
 - Lecture 8
 - Lecture 9
 - Lecture 10
 - Lecture 11
 - Lecture 12
 - Lecture 13
 - Lecture 14
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models

- **Probability theory** is a mathematical framework for modelling uncertainty.
- Originated in the analysis of games of chance in 17th century.
 - ▶ Started in context of “games” with a *finite number of equally likely outcomes*, the so-called “classical definition of probability”.
- Was “modernised” in the 1920s by Kolmogorov, part of measure theory.

Some definitions and notation

Term	Definition	Usual notation
Sample space	Set of all possible outcomes	Ω
Outcomes	Elements of the sample space	$\omega \in \Omega$
Events	Subsets of the sample space	$A \subset \Omega$
Union (of 2 events)	$\{\omega \in \Omega : \omega \in A \text{ OR } \omega \in B\}$	$A \cup B$
Intersection (ditto)	$\{\omega \in \Omega : \omega \in A \text{ AND } \omega \in B\}$	$A \cap B$
Complement (of A)	$\{\omega \in \Omega : \omega \notin A\}$	A^c
Empty set	Set with no elements	\emptyset
Collection(s) of events		\mathcal{A}, \mathcal{B}
Mutually exclusive	No outcomes in common	$A \cap B = \emptyset$
Number of outcomes in A		$\#A$

Some observations

- $\Omega^c = \emptyset$, $\emptyset^c = \Omega$.
- For any $A \subset \Omega$, $B \subset \Omega$ then

$$A \cup B = (A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)$$

(see sketch).

- Note that
 - ▶ \cup (union) is like addition and
 - ▶ \cap (intersection) is like multiplication:

compare

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

with

$$(a + b)c = ac + bc.$$

Axioms of probability

- Suppose Ω is a sample space and that \mathcal{A} is a “suitable collection” of events in Ω which includes both Ω itself and \emptyset (indeed if $A \in \mathcal{A}$ then so too is $A^c \in \mathcal{A}$).
- A **probability** $P(\cdot)$ on \mathcal{A} is any set of rules assigning a number $P(A)$ to each $A \in \mathcal{A}$ such that
 - ▶ **(A1)** $P(A) \geq 0$ for all $A \in \mathcal{A}$;
 - ▶ **(A2)** $P(\Omega) = 1$;
 - ▶ **(A3)** If A_1, A_2, \dots in \mathcal{A} are *mutually exclusive* then

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

- These three “conditions” are collectively known as the **axioms of probability**.
 - ▶ (A3) is referred to as “countable additivity”.

Example: coin tossing

- Suppose a coin is flipped three times in a row.
- Define the sample space as $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.
- Let the “suitable collection” of events \mathcal{A} be *all subsets of Ω* (*aside: how many such subsets?*).
- For any $A \in \mathcal{A}$ (i.e. any $A \subset \Omega$) define

$$P(A) = \frac{\#A}{8} = \frac{\text{number of outcomes in } A}{8} .$$

- It is easy to show that the axioms (A1), (A2) and (A3) are satisfied by this $P(\cdot)$.
- We show below that this $P(\cdot)$ is a (well-defined) **probability** on this \mathcal{A} .

Classical definition of probability

- The previous example is a special case of the **classical definition of probability**:
 - ▶ For any finite sample space Ω let \mathcal{A} be the set of all subsets.
 - ▶ For any subset $A \in \mathcal{A}$ (i.e. any $A \subset \Omega$) define

$$P(A) = \frac{\#A}{\#\Omega}.$$

- This defines a **probability** on \mathcal{A} .
- To see this, note that (A1) and (A2) follows easily, and (A3) follows because for any two events A and B that are mutually exclusive,

$$\#(A \cup B) = \#A + \#B.$$

A comment on (A3): countable additivity

- The “third axiom” (A3) concerns an infinite sequence of events.
- What if there are only a finite number of events in the collection \mathcal{A} ?
- It is still possible to find an *infinite sequence of mutually exclusive events* by appending $\emptyset, \emptyset, \emptyset, \dots$ to the end of any finite sequence of mutually exclusive events.

Infinite sample spaces

- Suppose $\Omega = \{\omega_1, \omega_2, \dots\}$ is a *countably infinite* set, e.g. the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$.
- Assign a (non-negative) weight p_i to each outcome ω_i in such a way that the *infinite sum*

$$p_1 + p_2 + \cdots = 1.$$

- Then consider the collection $\mathcal{A} = \text{all subsets of } \Omega$ (there are infinitely many subsets here).
- Define for any $A \in \mathcal{A}$,

$$P(A) = \sum_{i: \omega_i \in A} p_i.$$

- It is easily checked that this scheme defines a **probability** on \mathcal{A} .

Example: more coin tossing

- Suppose now that a coin is flipped repeatedly until the first head is obtained.
- Represent the sample space as $\mathbb{N} = \{1, 2, \dots\}$, where the (positive integer) outcome i means that the first head occurs on the i -th flip.
- Assign to each positive integer i the weight

$$p_i = \frac{1}{2^i} .$$

- Since

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1 ,$$

this defines a probability on this collection \mathcal{A} .

A more complicated example

- Suppose the sample space is $\Omega = [0, 1]$, the unit interval.
- Define for any interval $I \subset [0, 1]$,

$$P(I) = \text{length of } I.$$

Here I may include or exclude either endpoint.

- Consider the collection \mathcal{A} consisting of all events A of the form $A = I_1 \cup I_2 \cup I_3 \cup \dots$ where I_1, I_2, \dots are all mutually exclusive intervals.
- For any such event A define

$$P(A) = P(I_1) + P(I_2) + P(I_3) + \dots.$$

- This defines a probability on this collection \mathcal{A} (which is strictly smaller than *all subsets of* $[0, 1]$).
 - ▶ Note that (A3) is satisfied *by construction*.

Uniform distribution

- The previous example is known as the *uniform probability (or distribution) on $[0, 1]$* .
- The random-number generators on hand calculators are “imitations” of this probability.
- Thus under this probability
 - ▶ $P((0, x)) = x$ for any $0 < x \leq 1$;
 - ▶ $P((x, y)) = y - x$ for any $0 < x < y < 1$;
- Let $Q = \{q = a/b\}$ be all rational numbers in $[0, 1]$ where the numerator a and denominator b (both integers) have no common factor. Then it is possible to enumerate the elements in Q in a sequence:

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \dots$$

We may thus express $Q = \{q_1, q_2, q_3, \dots\}$ or equivalently as

$$Q = [q_1, q_1] \cup [q_2, q_2] \cup [q_3, q_3] \cup \dots$$

i.e. as a union of (countably) infinitely many (closed) intervals of length zero. Thus $P(Q) = 0$.

More general class of distributions

- The uniform probability can be generalised by simply changing the definition of $P(I)$ for intervals I .
- Let $f(\cdot)$ be a function with the following properties:
 - ▶ $f(x) \geq 0$ for all real x ;
 - ▶ $\int_{-\infty}^{\infty} f(x) dx = 1$.
- Define for any interval $I = (a, b)$ (or $(a, b]$ or $[a, b)$ or $[a, b]$),

$$P(I) = \int_a^b f(x) dx.$$

- Then, as before for any $A = I_1 \cup I_2 \cup \dots$ for mutually nonoverlapping intervals I_1, I_2, \dots ,

$$P(A) = P(I_1) + P(I_2) + \dots$$

- This also defines a probability on the collection of all events A of this form.

Consequences of the axioms

For any A in a collection \mathcal{A} we have $A \subset \Omega$, $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$ (A , A^c are mutually exclusive). So

- **(C1)** $1 \stackrel{(A2)}{=} P(\Omega) = P(A \cup A^c) \stackrel{(A3)}{=} P(A) + P(A^c)$ i.e. $P(A^c) = 1 - P(A)$.
- **(C2)** $1 - P(A) = P(A^c) \stackrel{(A1)}{\geq} 0 \Rightarrow P(A) \leq 1$.
- **(C3)** $P(A) = P((A \cap B) \cup (A \cap B^c)) \stackrel{(A3)}{=} P(A \cap B) + P(A \cap B^c)$ (for any other $B \in \mathcal{A}$, since then $(A \cap B)$ and $(A \cap B^c)$ are mutually exclusive).
- **(C4)**

$$\begin{aligned}P(A \cup B) &= P(A \cap B) + P(A \cap B^c) + P(A^c \cap B) \\&= P(A) + P(B) - P(A \cap B)\end{aligned}$$

(this is verified by applying (C3) above to $P(B)$ and rearranging).

- We don't need to verify these consequences "directly" in each example, although we possibly could.
- They follow simply by virtue that the axioms (A1), (A2) and (A3) hold.
- Thus in any example, whenever the axioms hold, so too do these consequences.

- We now focus on some special cases of **classical** probability.
- According to the classical defintion of probability, for any event A which is a subset of a sample space Ω ,

$$P(A) = \frac{\#A}{\#\Omega}$$

and so we just need to be able to count the number of elements in any such event A and sample space Ω .

- However, in some example this is not as easy as it sounds.

Counting selections

- Suppose we have a two-stage procedure:
 - ▶ at stage A, we have k possible choices
 - ▶ at stage B, the range of possible choices *depends* on the choice at the first stage:
 - ★ if we choose choice 1 at stage A, there are n_1 choices at stage B;
 - ★ if we choose choice 2 at stage A, there are n_2 choices at stage B;
 - ★ etc.
- Then the total number of ways we can choose how to perform the procedure is

$$n_1 + n_2 + \cdots + n_k$$

(see sketch).

- In the special case where $n_1 = n_2 = \cdots = n_k = n$ (say) this reduces to

$$kn.$$

Multiplication principle

- We can extend this **special case** to more than two stages:
 - ▶ Suppose we have s stages.
 - ▶ Suppose that there are m_1 choices for stage 1.
 - ▶ Suppose that at stage 2
 - ★ the *actual choices* depend on the choice at stage 1 **but** the *number* of choices is the same regardless, and is equal to m_2 .
 - ▶ Suppose that at stage 3
 - ★ the *actual choices* depend on the choices at stages 1 and 2 **but** the *number* of choices is the same regardless, and is equal to m_3 .
 - ▶ etc.
 - ▶ Then the *total* number of ways we can choose how to perform the whole s -stage procedure is

$$m_1 m_2 \cdots m_s .$$

Examples: drawing with replacement

- Suppose we have an urn with 3 balls, numbered 1, 2 and 3 and we are planning to draw out 2 balls *with replacement*:
 - ▶ There are 3 ways the first draw can be performed: 1, 2 or 3.
 - ▶ There are also 3 ways the second draw can be performed (since the first ball is replaced).
- There are thus $3 \times 3 = 9$ ways we can perform this procedure. (see sketch of tree diagram)

Examples: drawing *without* replacement

- Suppose we have an urn with 3 balls, numbered 1, 2 and 3 and we are planning to draw out 2 balls *without replacement*:
 - ▶ There are 3 ways the first draw can be performed: 1, 2 or 3.
 - ★ If the first draw is 1, there are two choice for the second draw: 2 or 3
 - ★ If the first draw is 2, there are two choice for the second draw: 1 or 3
 - ★ If the first draw is 3, there are two choice for the second draw: 1 or 2.
- Note that although the *actual choices* for the second draw depend on the first draw, the *number of choices* is the same regardless, i.e. is 2.
- Thus according to the multiplication principle, the total number of ways this procedure can be performed is

$$3 \times 2 = 6.$$

(see sketch of tree diagram).

Sets of sequences

- Suppose that for two positive integers n and N ,

$${}^N\mathcal{S}_n = \text{all sequences of length } n \text{ from } \{1, 2, \dots, N\}.$$

- E.g. for $n = 3$ and $N = 2$ we get

$$\begin{aligned} {}^2\mathcal{S}_3 &= \{(1, 1, 1), (1, 1, 2), (1, 2, 1), (1, 2, 2) \\ &(2, 1, 1), (2, 1, 2), (2, 2, 1), (2, 2, 2)\} \end{aligned}$$

- Then we have

$$\#{}^N\mathcal{S}_n = N^n.$$

- This is proved easily using the multiplication principle:
 - there are N choices for the first position;
 - ...
 - there are N choices for the n -th position.
- Sometimes ${}^N\mathcal{S}_n$ is written as $\{1, 2, \dots, N\}^n$ (product set notation).

Sets of n -permutations

- An n -permutation of a set of *distinct* objects is a sequence of n elements from it with **no repetitions**.
- Define

$${}^N\mathcal{P}_n = \text{all } n\text{-permutations from } \{1, 2, \dots, N\}.$$

- Note we must have $1 \leq n \leq N$.
- E.g. for $n = 2, N = 3$ we have

$${}^3\mathcal{P}_2 = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)\}.$$

- Then we have

$$\# {}^N\mathcal{P}_n = \underbrace{N(N-1) \cdots (N-n+1)}_{n \text{ factors}}.$$

- Again, this is verified using the multiplication principle:
 - ▶ There are N choices for the first position.
 - ▶ For each of these, there are $(N - 1)$ numbers left for second position.
 - ▶ Whatever the first two numbers, there are a further $(N - 2)$ numbers left to occupy the third position.
 - ▶ ...
 - ▶ After $(n - 1)$ positions are filled, there are $N - (n - 1) = N - n + 1$ possible choices for the last (n -th) position.

Outline

1 Welcome

2 Data Analysis

3 Probability

● Lecture 6

● Lecture 7

- Permutations and Combinations
- Conditional Probability

● Lecture 8

● Lecture 9

● Lecture 10

● Lecture 11

● Lecture 12

● Lecture 13

● Lecture 14

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

(Ordinary) permutations

- Suppose we have a set of N distinct objects (for positive integer N).
- Then an N -permutation is often just called a permutation of the elements of the set (i.e. the N - prefix is dropped).
- Thus using the above result we have

$$\#^N \mathcal{P}_N = \underbrace{N(N-1)\cdots 1}_{N \text{ factors}} = N! .$$

Factorial

- We call the quantity $N! = N(N - 1) \cdots 1$ “ N -factorial”.
- For $1 \leq n < N$ we may write

$$\#^N \mathcal{P}_n = \underbrace{N(N - 1) \cdots (N - n + 1)}_{n \text{ factors}} = \frac{N!}{(N - n)!} = {}^N P_n .$$

- For the case $n = N$ this notation seems to fail since then we get

$$\#^N \mathcal{P}_N = \frac{N!}{(N - N)!} = \frac{N!}{0!} .$$

- But we know that $\#^N \mathcal{P}_N = N!$.
- So, if we define $0! = 1$ then this notation can still make sense in the case $n = N$.
- Thus we define the factorial function for non-negative integers as

$$N! = \begin{cases} 1 & \text{for } N = 0, \\ N(N - 1) \cdots 1 & \text{for } N = 1, 2, \dots \end{cases}$$

Sets of n -combinations

- An n -combination of a set of *distinct* objects is simply a subset of n of its elements.
- Define

$${}^N\mathcal{C}_n = \text{all } n\text{-combinations from } \{1, 2, \dots, N\}.$$

- E.g. for $n = 2, N = 3$ we have

$${}^3\mathcal{C}_2 = \{(1, 2), (1, 3), (2, 3)\}.$$

- Comparing this with ${}^3\mathcal{P}_2$ we see that
 $\#{}^3\mathcal{P}_2$ is *bigger*:
 - ▶ ${}^3\mathcal{P}_2$ contains all possible *re-orderings* i.e. *permutations* of each element of ${}^3\mathcal{C}_2$.
- Each n -combination contains n distinct elements and so thus can be *reordered* i.e. *permuted* $n!$ ways.
- Thus each n -combination corresponds to $n!$ n -permutations.

"N-choose-n": binomial coefficients

- We thus have

$$\#{}^N\mathcal{C}_n = \frac{\#{}^N\mathcal{P}_n}{n!} = \frac{{}^N\mathcal{P}_n}{n!} = \frac{N!}{(N-n)!n!} = \binom{N}{n},$$

(also written as ${}^N\mathcal{C}_n$) and this holds for all integers $N \geq 1$ and $0 \leq n \leq N$ so long as we define $0! = 1$.

- The notation $\binom{N}{n}$ is usually read as “ N -choose- n ” reflecting the fact that it gives the number of ways of choosing (a subset of) n objects from a larger collection of size N .
- This quantity appears in the following form of *Newton's binomial formula*: for any real x and positive integer n ,

$$(1+x)^n = \sum_{i=0}^n \binom{n}{i} x^i$$

and as such is also referred to as a *binomial coefficient* (see also section 17, page 81 of the HSC syllabus).

Pascal's triangle

- (see sketch)
- Note the symmetry $\binom{N}{n} = \binom{N}{N-n}$.

Sampling interpretation

- We can interpret the sets ${}^N\mathcal{S}_n$, ${}^N\mathcal{P}_n$ and ${}^N\mathcal{C}_n$ as *sets of possible samples* in a literal sense:
 - ▶ ${}^N\mathcal{S}_n$ consists of all possible **ordered samples taken with replacement** of size n from the *population* $\{1, 2, \dots, N\}$;
 - ▶ ${}^N\mathcal{P}_n$ consists of all possible **ordered samples taken without replacement** of size n from the *population* $\{1, 2, \dots, N\}$;
 - ▶ ${}^N\mathcal{C}_n$ consists of all possible **unordered samples taken without replacement** of size n from the *population* $\{1, 2, \dots, N\}$

Applications

- Suppose a coin is flipped 7 times in such a way that all possible sequences of H 's and T 's are equally likely. What is the probability of (exactly) 3 H 's?
- **Solution:**

▶ *How many outcomes in the sample space?:*

- ★ There is a one-to-one correspondence between these and the sequences in 2S_7 : map $1 \leftrightarrow H$ and $2 \leftrightarrow T$.
- ★ So there are $\# {}^2S_7 = 2^7 = 128$ equally likely outcomes in the sample space.

▶ *How many of these have 3 H 's?:*

- ★ There is a one-to-one correspondence between these and the subsets in 7C_3 : values in such a subset give the positions in the sequence where the H 's occur.
- ★ So there are $\# {}^7C_3 = \binom{7}{3} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$ of these.

▶ So

$$P(3 \text{ } H\text{'s in 7 such flips}) = \frac{\binom{7}{3}}{2^7} = \frac{35}{128}.$$

More generally

- Suppose a coin is flipped n times in such a way that all possible sequences of H 's and T 's are equally likely. What is the probability of (exactly) x H 's (for some integer $0 \leq x \leq n$)?
- Using the same reasoning:

$$P(x \text{ } H\text{'s in } n \text{ such flips}) = \frac{\binom{n}{x}}{2^n}.$$

Sampling Without Replacement

- Suppose that
 - ▶ a habitat contains 30 animals of a certain species of which 5 are tagged;
 - ▶ 6 of these 30 animals are captured in such a way that all possible samples of size 6 are equally likely (a “random sample”).
- What is the probability that 2 of these 6 are tagged?
- The desired probability may be expressed as the ratio

$$\frac{\text{no. samples with 2 tagged (and thus 4 untagged)}}{\text{total no. samples size 6}}.$$

- The denominator is $\#{}^{30}C_6 = \binom{30}{6}$.

Multiplication principle

- We can evaluate the numerator using the multiplication principle.
- Suppose the animals are “numbered” 1 to 30 and that 1 to 5 are the *tagged* ones.
- Then each sample counted in the numerator contains
 - ① a subset of size 2 from $\{1, 2, 3, 4, 5\}$ (there are $\binom{5}{2}$ of these) *and*
 - ② a subset of size 4 from $\{6, 7, \dots, 30\}$ (**note**: this set has 25 elements, so there are $\binom{25}{4}$ of these).
- Thus by the multiplication principle there are $\binom{5}{2} \binom{25}{4}$ ways to
 - ① pick a subset of size 2 from $\{1, 2, 3, 4, 5\}$ *and then*
 - ② pick a subset of size 4 from $\{6, 7, \dots, 30\}$.
- The desired probability is thus

$$\frac{\binom{5}{2} \binom{25}{4}}{\binom{30}{6}}.$$

More generally

- Suppose that
 - ▶ a habitat contains N animals of a certain species of which M are tagged;
 - ▶ n of these N animals are captured in such a way that all possible samples of size n are equally likely (a “random sample”).
- What is the probability that x of these n are tagged?
- **Solution:**

$$P(x \text{ tagged (and thus } (n-x) \text{ untagged})) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}.$$

Conditional Probability

- Suppose we are rolling a 6-sided die (with sides numbered 1 to 6) once.
- Then we may represent the sample space as $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Let
 - ▶ $A = \text{"number showing } < 4" = \{1, 2, 3\}$ and
 - ▶ $B = \text{"number showing is even"} = \{2, 4, 6\}$.
- Then $A \cup B = \{1, 2, 3, 4, 6\}$ and $A \cap B = \{2\}$.
- If the die is rolled in such a way that all sides are equally likely then $P(A) = P(B) = 0.5$,

$$P(A \cup B) = \frac{5}{6}$$

and

$$P(A \cap B) = \frac{1}{6}.$$

- In the die-rolling example above suppose that we are *told* an even number is rolled but nothing else.
- What *then* is the probability of getting a number < 4 ?
- Effectively here we are *shrinking* the sample space down from $\Omega = \{1, 2, 3, 4, 5, 6\}$ to $B = \{2, 4, 6\}$.
- In that case, if the 3 remaining possible outcomes are all equally likely then

$$P(\text{no. } < 4 \text{ GIVEN it is even}) = \frac{\#\{2\}}{\#\{2, 4, 6\}} = \frac{1}{3}.$$

- This idea is generalised by the concept of *conditional probability*.

Definition

- Suppose we have a collection of events \mathcal{A} , a probability $P(\cdot)$ defined on it and 2 events $A, B \in \mathcal{A}$ such that $P(B) > 0$.
- The **conditional probability** of A given B is written and defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Interpretation

- Suppose Ω is a finite sample space and let $P(\cdot)$ denote the classical definition of probability on Ω .
- If we interpret Ω as a population of individuals then we may interpret $P(A)$ as the *proportion* of individuals in the population with a certain “attribute” A .
- In a similar way another event B can be used to define a **subpopulation** of Ω and moreover the conditional probability $P(A|B)$ can be interpreted as the proportion of individuals in the *subpopulation* with attribute A .

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
 - Lecture 6
 - Lecture 7
 - Lecture 8
 - Bayes rule
 - Independence
 - Non-equally-likely outcomes
 - Combining “experiments” independently
 - The binomial distribution
 - Lecture 9
 - Lecture 10
 - Lecture 11
 - Lecture 12
 - Lecture 13
 - Lecture 14
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models

Bayes rule

- Note that conditional probability gives us another way to express the probability of an intersection:

$$P(A \cap B) = P(A|B)P(B),$$

so long as $P(B) > 0$.

- Also recall that we can express

$$P(A) = P(A \cap B) + P(A \cap B^c).$$

- Combining these last two ideas we can write

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

- We can thus *reverse the conditioning* as follows:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

This is known as **Bayes' Rule**.

- Thus knowing $P(A|B)$, $P(A|B^c)$ and $P(B)$ is enough to compute $P(B|A)$ (note $P(B^c) = 1 - P(B)$).

Bayes' rule example

- Suppose that in a certain population 1% of individuals have a certain disease.
- A certain diagnostic test is available for the disease and it performs as follows:
 - ▶ for those *with* the disease, it detects the disease 98% of the time
 - ▶ for those *without* the disease, it gives a "false positive" 5% of the time.
- Suppose a person is tested (assume they are randomly picked from the population) and they have a positive test result.
- What is the (conditional) probability that they have the disease?

- First let us more clearly define things:
 - Suppose the whole population is Ω and that a single person is picked at random.
 - Define the following events:
 - D = “person picked has the disease”;
 - $+$ = “person picked tests positive for the disease”.
- The information above can be translated into
 - $P(D) = 0.01$ and so $P(D^c) = 1 - P(D) = 0.99$;
 - $P(+/D) = 0.98$;
 - $P(+/D^c) = 0.05$.
- The desired (conditional) probability is $P(D|+)$.
- Applying Bayes' rule directly gives

$$\begin{aligned}
 P(D|+) &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} \\
 &= \frac{0.98 \times 0.01}{(0.98 \times 0.01) + (0.05 \times 0.99)} \\
 &= \frac{0.0098}{0.0098 + 0.0495} \approx 0.1652
 \end{aligned}$$

which is close to $1/6$.

- This is to say, for every “true positive” there are (about) 5 “false positives”, simply because most people in the population don’t have the disease.
- To see this more clearly, if we let $-$ denote the complement of $+$ then note that
 - ▶ $P(-|D) = 1 - P(+|D) = 0.02$ and
 - ▶ $P(-|D^c) = 1 - P(+|D^c) = 0.95$

and so

$$P(D \cap +) = P(+|D)P(D) = 0.98 \times 0.01 = 0.0098,$$

$$P(D \cap -) = P(-|D)P(D) = 0.02 \times 0.01 = 0.0002,$$

$$P(D^c \cap +) = P(+|D^c)P(D^c) = 0.05 \times 0.99 = 0.0495 \text{ and}$$

$$P(D \cap -) = P(-|D^c)P(D^c) = 0.95 \times 0.99 = 0.9405.$$

- So if there are 10,000 in the population, we can cross-classify them as follows:

Disease Status	Test Positive	Test Negative	Totals
With Disease	98	2	100
Without Disease	495	9405	9900
Totals	593	9407	10000

- Thus the conditional probability $P(D|+)$ = 98/593 ≈ 0.1652 is the proportion of people who have the disease in the “Test Positive” subpopulation.
- Roughly speaking: *for rare diseases, most positive test results are false positives.*

Independence

- Recall the 6-sided die example: $\Omega = \{1, 2, 3, 4, 5, 6\}$,

$$A = \text{number showing is } < 4 = \{1, 2, 3\};$$

$$B = \text{number showing is even} = \{2, 4, 6\};$$

$$A \cap B = \{2\}.$$

- If each number is equally likely then

$$P(A) = \frac{1}{2} = P(B)$$

and

$$P(A \cap B) = \frac{1}{6}.$$

Thus

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

- In other words, knowing the number is even *reduces* the chance of a number less than 4 from its *unconditional* value of $P(A) = \frac{1}{2}$ to $\frac{1}{3}$.
- Also, since $B^c = \{1, 3, 5\}$, $A \cap B^c = \{1, 3\}$, $P(A \cap B^c) = \frac{1}{3}$ we have

$$P(A|B^c) = \frac{1/3}{1/2} = \frac{2}{3}.$$

- So knowing the number is odd *increases* the chance of a number less than 4.
- These differences occur because the numbers less than 4 are **not** evenly distributed between the odds and the evens.

- However, compare this to what happens if we replace A with $C = \text{number showing is } \leq 4 = \{1, 2, 3, 4\}$.
- Now $P(C) = \frac{2}{3}$, $C \cap B = \{2, 4\}$, $P(C \cap B) = \frac{1}{3}$ and

$$P(C|B) = \frac{P(C \cap B)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

- Also, $C \cap B^c = \{1, 3\}$, $P(C \cap B^c) = \frac{1}{3}$ so

$$P(C|B^c) = \frac{P(C \cap B^c)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

- So in this case we have

$$P(C|B) = P(C|B^c) = P(C).$$

Knowing we have an odd or even number gives us no "information" about whether the number is ≤ 4 or not;

- ▶ the numbers ≤ 4 are **equally distributed** among the odds and the evens.
- In this sense C does **not depend** on B in the same way A does.

Definition

- Suppose $P(\cdot)$ is a probability defined on a collection of events \mathcal{A} .
- Two events $A \in \mathcal{A}$ and $B \in \mathcal{A}$ are said to be **independent** if and only if

$$P(A \cap B) = P(A)P(B).$$

Some comments

- In the tutorial we show that if A and B are independent then so too are A and B^c .
- It follows then that the pairs (A^c, B) and (A^c, B^c) are also pairs of independent events.
- Note that a *consequence* of A and B being independent is that (assuming $P(B) > 0$),

$$P(A|B) \stackrel{\text{def}}{=} \frac{P(A \cap B)}{P(B)} \stackrel{\text{indep}}{=} \frac{P(A)P(B)}{P(B)} = P(A) \stackrel{\text{similarly}}{=} P(A|B^c)$$

using the first point.

More than 2 events

- Any collection of events are said to be independent if the probability of any intersection of any group of them is equal to the corresponding product of individual probabilities.
- In that case, any of them may be replaced by their complements to give another set of independent events, as in the case of 2 events.

Pairwise independence

- It is possible that three events A , B and C are such that each pair are independent but $P(A \cap B \cap C) \neq P(A)P(B)P(C)$.
- In such a case these events are said to be *pairwise independent*.
- For example suppose two fair dice are rolled in such a way that all 36 possible pairs are equally likely. Then
 - ▶ A = first roll is even,
 - ▶ B = second roll is odd,
 - ▶ C = both rolls are the same

are *pairwise independent*, but not (totally) independent.

Non-equally-likely outcomes

- Suppose Ω is a finite sample space and that for each $\omega \in \Omega$, $p(\omega) \geq 0$ and

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

- Then (as we have seen) the rule which assigns to each $A \subset \Omega$ the value

$$P(A) = \sum_{\omega \in A} p(\omega)$$

defines a probability on $\mathcal{A} = \text{all subsets of } \Omega$.

Example

- A 6-sided die is weighted in such a way that when thrown, the chance of each side landing face-upwards is **proportional** to the number showing on it. **Find** $P(6)$.
- **Solution:** For some value v , we have probabilities for each face as follows:

Face	1	2	3	4	5	6	Total
Prob.	v	$2v$	$3v$	$4v$	$5v$	$6v$	$21v = 1$

Therefore $v = 1/21$ and thus $P(6) = \frac{6}{21} = \frac{2}{7}$.

Combining “experiments” independently

- Suppose Ω_1 and Ω_2 are two sample spaces with respective probabilities defined on their respective collections of events \mathcal{A}_1 and \mathcal{A}_2 .
- We may define a **new** sample space by forming the *product set*

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

and define, for any $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ the probability

$$P(A_1 \cap A_2) = P_1(A_1)P_2(A_2).$$

- This corresponds to performing the two “experiments” in sequence in such a way that each event involving only the first stage is independent of any event only involving the second stage.

Example

- Roll the weighted die from the previous example twice independently. What is the probability of obtaining a total of 5?
- **Solution:**

$$\begin{aligned}P(\text{Total} = 5) &= P(1, 4) + P(2, 3) + P(3, 2) + P(4, 1) \\&= P_1(1)P_2(4) + P_1(2)P_2(3) + P_1(3)P_2(2) + P_1(4)P_2(1) \\&= \left[\left(\frac{1}{21} \times \frac{4}{21} \right) + \left(\frac{2}{21} \times \frac{3}{21} \right) \right] \times 2 = \frac{20}{21^2}.\end{aligned}$$

The binomial distribution

- The weighted die above is rolled 5 times independently.
- What is the probability of getting (exactly) 3 6's?
- **Solution:** Let $A_j = \text{"6 occurs on roll } j\text{"}$, for $j = 1, 2, 3, 4, 5$. Then
 - ▶ $P(A_1) = P(A_2) = \dots = P(A_5) = \frac{2}{7}$;
 - ▶ A_1, A_2, \dots, A_5 are all independent
- There are various sequences of A_j 's and A_j^c 's that give 3 6's: one such is

$$A_1 \cap A_2 \cap A_3 \cap A_4^c \cap A_5^c$$

which has probability

$$P(A_1)P(A_2)P(A_3)P(A_4^c)P(A_5^c) = \left(\frac{2}{7}\right)^3 \left(\frac{5}{7}\right)^2.$$

- How many such sequences are there?
 - ▶ By analogy with the coin-tossing example there are $\binom{5}{3}$ ($\#\mathcal{C}_3$; each subset gives the positions in the sequence where the 6's occur), each having the same probability.
- So we have $P(3 \text{ 6's}) = \binom{5}{3} \left(\frac{2}{7}\right)^3 \left(\frac{5}{7}\right)^2$.

- More generally if we have
 - n independent “trials”, each considered a “success” or a “failure”;
 - $P(\text{success}) = p$ is the same for each trial;

then

$$P(x \text{ successes}) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, \dots, n.$$

- This is known as the **binomial(n, p) distribution**.

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
 - Lecture 6
 - Lecture 7
 - Lecture 8
 - Lecture 9**
 - Infinite Sample Spaces
 - Uncountably Infinite Sample Spaces
 - Random Variables
 - Discrete Random Variables
 - Lecture 10
 - Lecture 11
 - Lecture 12
 - Lecture 13
 - Lecture 14
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models

Infinite Sample Spaces

- Consider the experiment where I flip a coin until the first head occurs.
- We can represent the sample space as the set of positive integers (natural numbers) i.e. $\Omega = \mathbb{N} = \{1, 2, 3, \dots\}$, where the integer gives the *position in the sequence of flips where the first head occurs*.
- There are infinitely many possible outcomes (at least in theory).
- From a practical point of view we might limit the number of attempts, but it is perfectly reasonable to consider a version without this restriction.

Defining a probability on (subsets of) \mathbb{N}

- As we have seen, the rule whereby we

- assign a non-negative weight p_i to each positive integer $i \in \mathbb{N}$ such that $\sum_{i=1}^{\infty} p_i = 1$ and
- assign to any subset $A \subset \mathbb{N}$ the number

$$P(A) = \sum_{i \in A} p_i$$

defines a (proper) probability on $\mathcal{A} = \text{all subsets of } \mathbb{N}$.

Example

- Assign to each $i \in \mathbb{N}$ the weight $p_i = 2^{-i}$, so that

Outcome i	Weight p_i
1	$1/2$
2	$1/4$
3	$1/8$
\vdots	\vdots

- It is straightforward to show that the geometric series $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1$.

- Thus we can assign probabilities to the following events:
 - $A = \text{"the first head occurs on the 5th flip"}$. Then $P(A) = 2^{-5} = \frac{1}{32}$.
 - $B = \text{"none of the first 4 flips are heads"}$. Then the first flip lands on the 5th flip or later, so then

$$P(B) = P(\{5, 6, 7, \dots\})$$

We can express this as the infinite sum

$$\frac{1}{2^5} + \frac{1}{2^6} + \dots = \frac{1}{2^4} \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots \right) = \frac{1}{2^4} = \frac{1}{16}$$

or as $1 - P(B^c)$, that is $1 - P(\{1, 2, 3, 4\})$ which is

$$1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} \right) = 1 - \frac{8 + 4 + 2 + 1}{16} = 1 - \frac{15}{16} = \frac{1}{16}.$$

Uncountably Infinite Sample Spaces

- The natural numbers is said to be *countably infinite* in that they can be placed in a sequence where each one has a well-defined position.
- The real line, or indeed any interval (with positive length) on the other hand is said to be *uncountably infinite* in that such an arrangement of points is not possible (contrast this with the rational numbers: they are countably infinite while the irrationals then are not).
- As we have seen, to define probabilities on the real line (or even just on an interval) we thus need a different idea.
- For any interval I we define

$$P(I) = \int_I f(x) dx$$

where $f(\cdot)$ is a *probability density function* (PDF) satisfying

- ▶ $f(x) \geq 0$ for all real x ;
- ▶ $\int_{-\infty}^{\infty} f(x) dx = 1$.

Comment on \mathcal{A}

- We then let the collection of events under consideration be all A which can be expressed in the form

$$A = I_1 \cup I_2 \cup \dots$$

for some (possibly countably infinite) collection of non-overlapping intervals I_1, I_2, \dots , in which case

$$P(A) = P(I_1) + P(I_2) + \dots$$

- Note that it is possible to come up with subsets I of the real line for which the integral $\int_I f(x) dx$ is not well defined.
- We shall not consider those, thus in such an example the “suitable collection of events” \mathcal{A} is *not* “all possible subsets”.

Radioactive decay example

- The waiting time (in hours) until a particle decays from some radioactive material is random and is described by the probability density function

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

- Find

- The probability the particle does not decay in the next 1 hour,
- The *conditional* probability it does not decay in the next 2 hours *given* it does not decay in the first hour.

- **Solution:** Let

- ▶ $A = \text{particle does not decay in the 1st hour}$ and
- ▶ $B = \text{particle does not decay in the 1st 2 hours}$

- Then the answer for 1. is

$$P(A) = \int_1^{\infty} e^{-x} dx = [-e^{-x}]_1^{\infty} = 0 - (-e^{-x}) = e^{-1} \approx 0.368.$$

- The answer for 2. is $P(B|A) = P(A \cap B)/P(A)$.

- ▶ Note that since $B \subset A$, $B \cap A = B$. Thus in this case $P(B|A) = P(B)/P(A)$.
- ▶ Since

$$P(B) = \int_2^{\infty} e^{-x} dx = \dots = e^{-2}$$

we have that the answer for 2. is also

$$P(B)/P(A) = e^{-2}/e^{-1} = e^{-1} \approx 0.368.$$

Random Variables

- A random variable is simply a rule that assigns a unique number to each outcome in a sample space; in other words a *real-valued function defined on the sample space*.
- We have seen many examples of random variables already:
 - ▶ number of heads in a fixed number of coin flips
 - ▶ total showing on two rolls of a die
 - ▶ number of tagged animals in the recapture sample
 - ▶ number of flips until the first head
 - ▶ value obtained under the “uniform-on-(0,1)” probability/distribution
 - ▶ time until radioactive particle decays

Let us look at an example in more detail.

Coin Tossing Example

- Suppose we flip a coin 3 times. As we have seen there are $2^3 = 8$ different possible outcomes.
- One way to represent the sample space is

$$\Omega = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}.$$

- There are various random variables we could define on this sample space:
 - ▶ X_1 = the number of heads on the first flip;
 - ▶ X_2 = the number of heads on the second flip;
 - ▶ X_3 = the number of heads on the third flip;
 - ▶ U = the total number of heads on the first two flips;
 - ▶ V = the total number of heads on the last two flips;
 - ▶ T = the total number of heads in all three flips.

- We can visualise these in the following table, giving the value of each random variable for each possible outcome ω in the sample space Ω :

ω	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$U(\omega)$	$V(\omega)$	$T(\omega)$
TTT	0	0	0	0	0	0
TTH	0	0	1	0	1	1
THT	0	1	0	1	1	1
THH	0	1	1	1	2	2
HTT	1	0	0	1	0	1
HTH	1	0	1	1	1	2
HHT	1	1	0	2	1	2
HHH	1	1	1	2	2	3

- Random variables can be used to describe events.
- In the example above, we have
 - ▶ $\{T = 2\} = \{THH, HTH, HHT\}$ = “total is 2” ;
 - ▶ $\{X_1 = 1\} = \{HTT, HTH, HHT, HHH\}$ = “first flip is a head”;
 - ▶ $\{V = 2\} = \{THH, HHH\}$ = “last two flips are both heads”;
- In fact here the notation $\{T = 2\}$ is really shorthand for $\{\omega \in \Omega | T(\omega) = 2\}$.

Discrete Random Variables

- A random variable that only takes values in a *separated (discrete) set of points* is called a **discrete random variable**.
- An important special case of a discrete random variable is an **integer-valued random variable**.
- All the examples in the coin flipping example above are of this type.

Probability Distributions of Discrete Random Variables

- Suppose we have a sample space Ω and a discrete random variable X defined on it and that \mathcal{A} is the “suitable collection of events” under consideration.
- Note: if Ω is countable then X must be discrete; also we may then take $\mathcal{A} = \text{“all subsets of } \Omega\text{”}$.
- Once a probability is defined on \mathcal{A} then X “inherits” a **probability distribution** from it.
- Specifically, suppose $\{x_1, x_2, \dots\}$ is the set of all possible values X can take, that is $\{X(\omega) | \omega \in \Omega\}$.
- Assuming each event of the form $\{X = x_j\}$ is in \mathcal{A} (if Ω is countable and \mathcal{A} “all subsets of Ω ” this will hold), the probability distribution of X is given by
 - ▶ the **set of all possible values** $\{x_1, x_2, \dots\}$ and
 - ▶ the **corresponding set of probabilities** $P(X = x_1), P(X = x_2), \dots$

Coin tossing example: heads and tails equally likely

- In the example above, suppose that all 8 outcomes are equally likely. Then the random variables X_1, X_2, X_3, T, U, V all inherit the following probability distributions:

x	0	1
$P(X_1 = x)$	$\frac{1}{2}$	$\frac{1}{2}$

x	0	1
$P(X_2 = x)$	$\frac{1}{2}$	$\frac{1}{2}$

x	0	1
$P(X_3 = x)$	$\frac{1}{2}$	$\frac{1}{2}$

Note that X_1, X_2, X_3 are all **identically distributed**.

u	0	1	2
$P(U = u)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

v	0	1	2
$P(V = v)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Note that U and V are also identically distributed.

t	0	1	2	3
$P(T = t)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Heads and tails possibly not equally likely

- Suppose now that we want to allow for the coin to *not be fair*.
- If we instead decide that $P(\text{head}) = p$ not necessarily equal to 0.5 *and that the flips are independent*, then we get the following set of weights for the outcomes (as seen in a previous lecture):

ω	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$U(\omega)$	$V(\omega)$	$T(\omega)$	Weight
TTT	0	0	0	0	0	0	$(1-p)^3$
TTH	0	0	1	0	1	1	$p(1-p)^2$
THT	0	1	0	1	1	1	$p(1-p)^2$
THH	0	1	1	1	2	2	$p^2(1-p)$
HTT	1	0	0	1	0	1	$p(1-p)^2$
HTH	1	0	1	1	1	2	$p^2(1-p)$
HHT	1	1	0	2	1	2	$p^2(1-p)$
HHH	1	1	1	2	2	3	p^3

- Note now that

$$\begin{aligned}
 P(X_1 = 1) &= P(\{\text{HTT}, \text{HTH}, \text{HHT}, \text{HHH}\}) \\
 &= p(1-p)^2 + 2p^2(1-p) + p^3 \\
 &= p[(1-p)^2 + 2p(1-p) + p^2] \\
 &= p[1 - 2p + p^2 + 2p - 2p^2 + p^2] \\
 &= p
 \end{aligned}$$

although note that we could have worked this out noting that $P(X_1 = 1) = P(\text{head on first}) = p$.

- The full probability distributions of X_1, X_2, X_3, T, U, V are now given by

x	0	1	x	0	1	x	0	1
$P(X_1 = x)$	$1 - p$	p	$P(X_2 = x)$	$1 - p$	p	$P(X_3 = x)$	$1 - p$	p

- ▶ Note again that X_1, X_2, X_3 are identically distributed.

u	0	1	2	v	0	1	2
$P(U = u)$	$(1 - p)^2$	$2p(1 - p)$	p^2	$P(V = v)$	$(1 - p)^2$	$2p(1 - p)$	p^2

- ▶ Note again that U and V are identically distributed.

t	0	1	2	3
$P(T = t)$	$(1 - p)^3$	$3p(1 - p)^2$	$3p^2(1 - p)$	p^3

The Binomial Distribution

- In fact, all these random variables have **binomial distributions** (see lecture 8) in that for some positive integer n and $0 \leq p \leq 1$,
 - ▶ the set of possible values x is $0, 1, \dots, n$ and
 - ▶ the probability of taking the value x is given by the formula

$$\binom{n}{x} p^x (1-p)^{n-x}.$$

- We use $B(n, p)$ as shorthand for “binomial distribution with n trials and success probability p ”. In the example above,
 - ▶ X_1, X_2, X_3 are all $B(1, p)$ (also called $\text{Bernoulli}(p)$) random variables;
 - ▶ U and V are both $B(2, p)$ random variables;
 - ▶ T is a $B(3, p)$ random variable.
- We also write $X_1 \sim B(1, p)$, $U \sim B(2, p)$, $T \sim B(3, p)$ etc.; that is “ \sim ” is short for “is distributed as”.

Using R

- R has a built-in function for the binomial probability distribution called `dbinom()`.
 - ▶ If $X \sim B(n, p)$ then $P(X = x)$ is given by `dbinom(x, n, p)`.
- For example for the weighted die example where $P(\text{rolling } 6) = 2/7$ we could compute $P(\text{rolling } 6 \text{ 3 times in 5 attempts}) = P(X = 3)$ where $X \sim B(5, \frac{2}{7})$ as follows:

```
dbinom(3, 5, 2/7)
```

```
[1] 0.118998
```

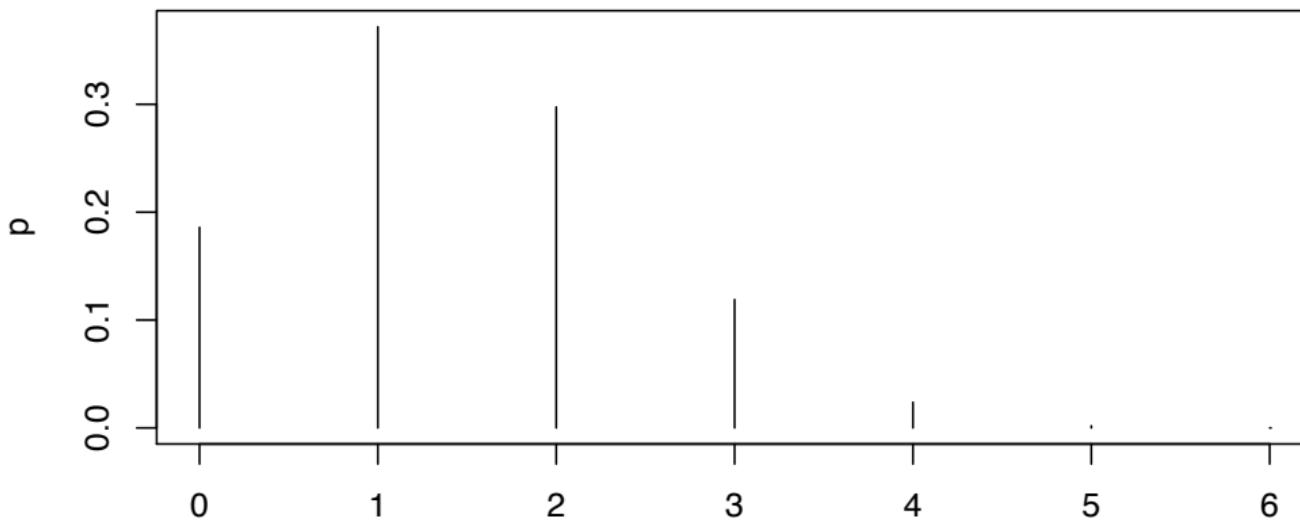
Note we can also compute this directly as $\binom{5}{3} \left(\frac{2}{7}\right)^3 \left(\frac{5}{7}\right)^2$:

```
choose(5, 3) * ((2/7)^3) * ((5/7)^2)
```

```
[1] 0.118998
```

We can visualise discrete probability distributions by creating ordinate diagrams:

```
x=0:6  
p=dbinom(x , 5 , 2/7)  
plot(x , p , type="h")
```



The Hypergeometric Distribution

- Suppose there are B black and W white balls in an urn and n are sampled *without replacement* in such a way that each possible sample is equally likely.
- Then if X denotes the number of black balls in the sample, X is said to have a **hypergeometric distribution**:

$$P(X = x) = \frac{\binom{B}{x} \binom{W}{n-x}}{\binom{B+W}{n}},$$

for each possible x .

- The possible values of x are those such that
 - ▶ $x \leq B$ (the no. black balls in the sample is no more than the no. in the urn) and
 - ▶ $(n - x) \leq W$ i.e. $x \geq n - W$ (the no. white balls in the sample is no more than the no. in the urn).
- These constraints are easy to remember since the two combinatorial coefficients in the numerator remind us:
 - ▶ $\binom{B}{x}$ reminds us $x \leq B$
 - ▶ $\binom{W}{n-x}$ reminds us $n - x \leq W$.

- The number of tagged animals in the sample for the capture-recapture example discussed earlier (see handwritten notes lecture 7 p5 and lecture 8) has a hypergeometric distribution;
 - ▶ indeed the general statement above is justified using identical reasoning employed in the analysis of that example.
- Using R, $P(X = x)$ is given by `dhyper(x,B,W,n)`.

Comparing the Binomial and Hypergeometric Distributions

- In fact, if we have the same urn as described above but instead change to sampling *with* replacement, then X has a binomial distribution with the same n and $p = B/(B + W)$, the *proportion* of black balls in the urn.
- Each draw is then independent of all others, and the probability of drawing a black ball at a single draw is $B/(B + W)$ (if each ball is equally likely).

The Geometric Distribution

- X is said to have a geometric distribution with success probability p if for all $x = 1, 2, 3, \dots$,

$$P(X = x) = (1 - p)^{x-1} p.$$

- This is the first example of a random variable taking an infinite number of possible values. The probabilities decrease “geometrically” however for each x , $P(X = x) > 0$.
- This distribution may be interpreted as that of the number of flips until the first head when flipping a coin independently which has $P(\text{head}) = p$.

Caution

- Note that if we let $Y =$ the number of tails *before* the first head then we can link X and Y by the equation $Y = X - 1$ and indeed then

$$P(Y = y) = P(X - 1 = y) = P(X = y + 1) = (1 - p)^y p$$

for $y = 0, 1, 2, \dots$

- This distribution is *also* called a geometric distribution and indeed the R function `dgeom(y, p)` returns probabilities for this second form of the geometric distribution.
- Be careful using R to compute geometric probabilities.
 - ▶ If in doubt check the R online help: `?dgeom`

Outline

1 Welcome

2 Data Analysis

3 Probability

- Lecture 6
- Lecture 7
- Lecture 8
- Lecture 9
- **Lecture 10**

- Expectation of X
- Expectation of a Sum
- Expectation of $g(X)$
- Linear Function of X
- Probability Generating Functions
- Variance of a Sum
- Joint Distribution of Discrete Random Variables
- Independent Random Variables.

- Lecture 11
- Lecture 12
- Lecture 13
- Lecture 14

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

Expectation of X

- An important property of a random variable X is its **expectation**, or expected value, or mean.
- For a *discrete* random variable X the expectation is defined as

$$E(X) \stackrel{\text{def}}{=} \sum_x xP(X = x),$$

the summation being over *all possible values x that X can take*. It is thus a *weighted* average of all possible values, the weight associated with each value being the probability of X taking that value.

- Note that it is a property of the probability distribution only. Different random variables on different sample spaces with the *same* distribution will have the same expectation.

- When the sample space is countable we can represent the expectation of a random variable X as

$$E(X) = \sum_{\omega \in \Omega} X(\omega)p(\omega)$$

where $p(\omega)$ is the "weight" associated with (the individual probability of) the outcome ω .

- To see this, split the sum into a double sum as follows:

$$\begin{aligned}\sum_{\omega \in \Omega} X(\omega)p(\omega) &= \sum_{\text{all poss. } x} \left(\sum_{\omega: X(\omega)=x} X(\omega)p(\omega) \right) \\ &= \sum_{\text{all poss. } x} \left(\sum_{\omega: X(\omega)=x} xp(\omega) \right) \\ &= \sum_{\text{all poss. } x} x \left(\sum_{\omega: X(\omega)=x} p(\omega) \right) \\ &= \sum_x xP(X=x).\end{aligned}$$

Interpretation of Expectation

- The expectation is a “measure of centre” or “measure of location” of the distribution of a random variable.
- It can literally be interpreted as the “long-run average” value we would see if we repeated the experiment a large number of times.

Some examples

- Suppose X has the following distribution:

x	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

Then

$$\begin{aligned}E(X) &= (0 \times 0.4) + (1 \times 0.3) + (2 \times 0.2) + (3 \times 0.1) \\&= 0 + 0.3 + 0.4 + 0.3 = 1.\end{aligned}$$

- If $X \sim B(2, p)$ then recall that X 's distribution is given by

x	0	1	2
$P(X = x)$	$(1 - p)^2$	$2p(1 - p)$	p^2

So

$$\begin{aligned}E(X) &= (0 \times (1 - p)^2) + (1 \times 2p(1 - p)) + (2 \times p^2) \\&= 2p - 2p^2 + 2p^2 = 2p.\end{aligned}$$

In fact we shall show later that if $X \sim B(n, p)$ then $E(X) = np$.

- Suppose X has the geometric distribution given by

$$P(X = x) = (1 - p)^{x-1} p$$

for $x = 1, 2, 3, \dots$ and some $0 < p < 1$. Then

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} x P(X = x) \\ &= \sum_{x=1}^{\infty} x (1 - p)^{x-1} p \\ &= p \{ 1 + 2(1 - p) + 3(1 - p)^2 + \dots \}. \end{aligned}$$

We show in the tutorial that the curly-bracketed factor equals $1/p^2$ and so

$$E(X) = 1/p.$$

Expectation of a Sum

- Suppose we have two random variables X and Y defined on the same *countable* sample space Ω .
- Then the sum of these, $S = X + Y$ is another random variable on the same sample space.
- Then using the second form of expectation given above, we can say that

$$\begin{aligned} E(S) &= \sum_{\omega \in \Omega} S(\omega)p(\omega) \\ &= \sum_{\omega \in \Omega} [X(\omega) + Y(\omega)]p(\omega) \\ &= \sum_{\omega \in \Omega} X(\omega)p(\omega) + \sum_{\omega \in \Omega} Y(\omega)p(\omega) \\ &= E(X) + E(Y). \end{aligned}$$

- So we always have (for *countable* Ω): **the expectation of a sum is the corresponding sum of expectations.**

Expectation of $g(X)$

- If X is a random variable defined on a sample space Ω and $g(\cdot)$ is some function then $Y = g(X)$ is just another random variable defined on Ω , given by

$$Y(\omega) = g(X(\omega)).$$

- Of course, if a probability $P(\cdot)$ is defined on (subsets of) Ω then Y , as well as X inherits a probability distribution from this.
- If X is discrete then Y is also discrete and so as usual the expectation of Y would be given by

$$E(Y) = \sum_y y P(Y = y).$$

- However, we also have

$$E(Y) = \sum_x g(x) P(X = x)$$

where the sum is over all possible values x that X can take; sometimes it is more convenient to calculate $E(Y)$ this way.

- To see this, note that the event $\{Y = y\}$ is the union

$$\cup_{x: g(x)=y} \{X = x\}$$

that is Y takes the value y if and only if X takes a value x such that $g(x) = y$ (e.g. if $g(x) = x^2$ so that $Y = X^2$ then $\{Y = y\} = \{X = -\sqrt{y}\} \cup \{X = +\sqrt{y}\}$).

- So then (by the third axiom),

$$P(Y = y) = \sum_{x: g(x)=y} P(X = x). \quad (*)$$

- Next, split the sum $\sum_x g(x)P(X = x)$ into a double sum:

$$\begin{aligned} \sum_x g(x)P(X = x) &= \sum_y \left(\sum_{x: g(x)=y} g(x)P(X = x) \right) \\ &= \sum_y \left(\sum_{x: g(x)=y} yP(X = x) \right) \\ &= \sum_y y \left(\sum_{x: g(x)=y} P(X = x) \right) = \sum_y yP(Y = y) \end{aligned}$$

according to $(*)$ above.

Important Examples

- **Moments** The *m-th moment* of X is $E(X^m)$.
- **Central Moments** The *m-th central moment* of X is $E[(X - \mu)^m]$ where $\mu = E(X)$ is the ordinary expectation; this is the (ordinary) *m-th moment* of $(X - \mu)$, the *centred version of X* .
- **Variance**

- ▶ The second central moment is called the *variance*:

$$\text{Var}(X) = E[(X - \mu)^2]$$

where $\mu = E(X)$.

- ▶ This is important for describing the *spread* or *dispersion* of X about the “centre” μ .
- ▶ In the tutorial we show that the following **computing formula** holds for the variance:

$$\text{Var}(X) = E(X^2) - [E(X)]^2.$$

Variance examples

- Suppose X has the following distribution:

x	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

Then as shown earlier $E(X) = 1$. We use the computing formula
 $\text{Var}(X) = E(X^2) - [E(X)]^2$ where

$$\begin{aligned}E(X^2) &= \sum_x x^2 P(X = x) \\&= (0^2 \times 0.4) + (1^2 \times 0.3) + (2^2 \times 0.2) + (3^2 \times 0.1) \\&= 0 + 0.3 + 0.8 + 0.9 = 2.\end{aligned}$$

And so $\text{Var}(X) = 2 - 1^2 = 1$.

- If $X \sim B(2, p)$ then recall that X 's distribution is given by

x	0	1	2
$P(X = x)$	$(1 - p)^2$	$2p(1 - p)$	p^2

and that $E(X) = 2p$. Again we use the formula $\text{Var}(X) = E(X^2) - [E(X)]^2$ where

$$\begin{aligned}
 E(X^2) &= \sum_x x^2 P(X = x) \\
 &= (0^2 \times (1 - p)^2) + (1^2 \times 2p(1 - p)) + (2^2 \times p^2) \\
 &= 2p - 2p^2 + 4p^2 = 2p + 2p^2.
 \end{aligned}$$

So $\text{Var}(X) = 2p + 2p^2 - (2p)^2 = 2p - 2p^2 = 2p(1 - p)$.

- In fact we shall show later that if $X \sim B(n, p)$ then $\text{Var}(X) = np(1 - p)$.

More examples of $E[g(X)]$

• Factorial Moments

- ▶ The m -th factorial moment of a (usually integer-valued) random variable X is given by

$$E \underbrace{[X(X - 1) \cdots (X - m + 1)]}_{m \text{ factors}}$$

so there are m factors inside the product.

- ▶ We can establish various relationships between the factorial moments and the ordinary and central moments.
 - ★ For example, the second factorial moment is (since the expectation of a sum is the sum of the expectations):

$$\begin{aligned} E[X(X - 1)] &= E(X^2 - X) = E(X^2) - E(X) \\ &= \{Var(X) + [E(X)]^2\} - E(X), \end{aligned}$$

the last equality following after using the computing formula for the variance given on slide 228 above. Thus

$$Var(X) = E[X(X - 1)] + E(X) - [E(X)]^2.$$

• Probability Generating Function

- ▶ If X is an *integer-valued* random variable then the *probability generating function* or PGF is the function of s given by $E(s^X)$.
- ▶ We study the properties of PGFs below.

Linear Function of X

An important special case is where we start with a random variable X and then define a new random variable

$$Y = g(X) = a + bX$$

for some constants a and b .

Expectation

- It is straightforward to show that

$$\begin{aligned} E(Y) &= E(a + bX) \\ &= \sum_x (a + bx) P(X = x) \\ &= a \sum_x P(X = x) + b \sum_x x P(X = x) \\ &= a + bE(X). \end{aligned}$$

- As special cases we have

- ▶ $E(X + a) = E(X) + a;$
- ▶ $E(bX) = bE(X).$

Variance

- We may compute $\text{Var}(Y) = \text{Var}(a + bX)$ in various ways.
- By definition, with $\mu_Y = E(Y) = a + bE(X) = a + b\mu_X$ we have

$$\begin{aligned}\text{Var}(Y) &= E[(Y - \mu_Y)^2] \\ &= E\left\{[(a + bX) - (a + b\mu_X)]^2\right\} \\ &= E[b^2(X - \mu_X)^2] = b^2E[(X - \mu_X)^2] = b^2\text{Var}(X).\end{aligned}$$

- The second-last equality follows because for any random variable Z , $E(b^2Z) = b^2E(Z)$ using an application of the result concerning expectations immediately above.

Standardised Version of X

- Suppose $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.
- Then the *standardised version of X* is that linear function of X which has expectation 0 and variance 1, and is given by

$$Z = \frac{X - \mu}{\sigma}.$$

- **Example:** Suppose $X \sim B(2, 0.7)$. Write down the standardised version of X .
- **Solution:** Since $E(X) = 2 \times 0.7 = 1.4$ and $\text{Var}(X) = 2 \times 0.7 \times 0.3 = 0.42$ the standardised version of this X is

$$Z = \frac{X - 1.4}{\sqrt{0.42}}.$$

Probability Generating Functions

- Suppose X is a random variable only taking values $0, 1, 2, \dots$ (some of these may have probability zero, but *no other* real numbers have positive probability).
- Then the PGF of X is the function

$$\pi_X(s) = E(s^X) = \sum_{x=0}^{\infty} s^x p_x = p_0 + p_1 s + p_2 s^2 + \dots$$

where here we have written $p_x = P(X = x)$ for short.

- If X only takes finitely many different values then $\pi_X(s)$ is a *polynomial*.
- In general $\pi_X(s)$ is a *power series* and notice that $\pi_X(1) = 1$.

Differentiating power series term-by-term

- Recall the following facts concerning power series:
 - ① the partial sums converge (i.e. the limit $\lim_{n \rightarrow \infty} \sum_{x=0}^n p_x s^x$ exists and is finite) for all $|s| < r$ where the *radius of convergence* $r \leq \infty$ depends on the coefficients p_0, p_1, \dots
 - ② for every s such that $|s| < r$, the power series can be differentiated term-by-term.
- Note that since the PGF equals 1 for $s = 1$, the radius of convergence for must be *at least* 1, and furthermore we can differentiate term-by-term for any $0 \leq s \leq 1$.

Examples: binomial

We use binomial probabilities to derive an expression for $(1 + t)^n$ (for $t > 0$) and then derive the binomial PGF:

- Suppose $X \sim B(n, p)$. Note then that we have

$$\begin{aligned} 1 = \sum_{x=0}^n P(X = x) &= \sum_{x=0}^n \binom{n}{x} p^x (1 - p)^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1 - p)^n. \end{aligned}$$

In other words

$$\sum_{x=0}^n \binom{n}{x} \left(\frac{p}{1-p}\right)^x = (1 - p)^{-n}.$$

- Putting $t = \frac{p}{1-p}$ we have $\frac{1}{1-p} = 1 + t$ and so we can write this as

$$\sum_{x=0}^n \binom{n}{x} t^x = \sum_{x=0}^n \binom{n}{x} \left(\frac{p}{1-p}\right)^x = (1 - p)^{-n} = (1 + t)^n.$$

- Then the PGF of X is given by

$$\begin{aligned}
 \pi_X(s) &= E(s^X) = \sum_{x=0}^n s^x P(X=x) \\
 &= \sum_{x=0}^n s^x \binom{n}{x} p^x (1-p)^{n-x} \\
 &= (1-p)^n \sum_{x=0}^n \binom{n}{x} \left[\frac{sp}{1-p} \right]^x \\
 &= (1-p)^n \left[1 + \frac{sp}{1-p} \right]^n \\
 &= (1 - p + sp)^n
 \end{aligned}$$

and so is a polynomial in s of degree n .

Geometric

- Suppose X has the geometric distribution given by $P(X = x) = (1 - p)^{x-1}p$ for $x = 1, 2, 3, \dots$ and some $0 < p < 1$.
- Then the PGF of X is given by

$$\begin{aligned}\pi_X(s) &= E(s^X) = \sum_{x=1}^{\infty} s^x P(X = x) \\&= \sum_{x=1}^{\infty} s^x (1 - p)^{x-1} p \\&= ps \{1 + s(1 - p) + [s(1 - p)]^2 + [s(1 - p)]^3 + \dots\} \\&= \frac{ps}{1 - s(1 - p)}\end{aligned}$$

for $|s(1 - p)| < 1$ (which includes $0 \leq s \leq 1$ for $0 < p < 1$).

Derivatives at 0 and 1

- We may differentiate the power series term-by-term for any $0 \leq s \leq 1$. The first few derivatives are

$$\pi'_X(s) = \sum_{x=0}^{\infty} p_x x s^{x-1} = p_1 + 2p_2 s + 3p_3 s^2 + \dots = \sum_{x=1}^{\infty} p_x x s^{x-1}$$

$$\pi''_X(s) = \sum_{x=0}^{\infty} p_x x(x-1) s^{x-2} = 2p_2 + 3 \times 2p_3 s + 4 \times 3p_4 s^2 + \dots$$

$$= \sum_{x=2}^{\infty} p_x x(x-1) s^{x-2}$$

$$\pi'''_X(s) = \sum_{x=0}^{\infty} x(x-1)(x-2) p_x s^{x-3} = 3 \times 2p_3 + 4 \times 3 \times 2p_4 s + 5 \times 4 \times 3p_5 s^2 + \dots$$

$$= \sum_{x=3}^{\infty} x(x-1)(x-2) p_x s^{x-3}$$

- In general the m -th derivative is

$$\begin{aligned}\pi_X^{(m)}(s) &= \sum_{x=0}^{\infty} x(x-1)\cdots(x-m+1)p_xs^{x-m} \\ &= \sum_{x=m}^{\infty} x(x-1)\cdots(x-m+1)p_xs^{x-m}.\end{aligned}$$

- If we evaluate $\pi_X^{(m)}(s)$ at $s = 0$, we see from the *far right-hand-side* that

$$\pi_X^{(m)}(0) = m!p_m = m!P(X = m)$$

so the derivatives at $s = 0$ give the *actual distribution* of X .

- If we evaluate $\pi_X^{(m)}(s)$ at $s = 1$ we see from the *middle* expression that

$$\begin{aligned}\pi_X^{(m)}(1) &= \sum_{x=0}^{\infty} x(x-1)\cdots(x-m+1)p_x \\ &= \sum_{x=0}^{\infty} x(x-1)\cdots(x-m+1)P(X=x) \\ &= E[X(X-1)\cdots(X-m+1)] ,\end{aligned}$$

the m -th factorial moment of X .

- Recall that the first two factorial moments can be used to derive the expectation and variance of X .

Examples

- If X is $B(n, p)$ then the PGF is

$$\pi_X(s) = (1 - p + sp)^n.$$

- The first two derivatives are

$$\pi'_X(s) = n(1 - p + sp)^{n-1}p$$

and

$$\pi''_X(s) = n(n-1)(1 - p + sp)^{n-2}p^2.$$

Thus setting $s = 1$ we get that

$$E(X) = \pi'_X(1) = np$$

and

$$E[X(X - 1)] = n(n-1)p^2.$$

Recall then that

$$\begin{aligned}Var(X) &= E[X(X - 1)] + E(X) - [E(X)^2] = n(n-1)p^2 + np - n^2p^2 \\&= n^2p^2 - np^2 + np - n^2p^2 = np(1 - p).\end{aligned}$$

- If X is geometric with $P(X = x) = (1 - p)^{x-1}p$ for $x = 1, 2, \dots$, then the PGF of X is

$$\pi_X(s) = \frac{ps}{1 - s(1 - p)}.$$

It is straightforward to check that the first two derivatives are

$$\pi'_X(s) = \frac{p}{[1 - s(1 - p)]^2} \quad \text{and} \quad \pi''_X(s) = \frac{2p(1 - p)}{[1 - s(1 - p)]^3}.$$

Thus (as we have already seen) the expectation is

$$E(X) = \pi'_X(1) = \frac{1}{p}$$

while the second factorial moment is

$$E[X(X - 1)] = \pi''_X(1) = \frac{2(1 - p)}{p^2}. \quad \text{Finally,}$$

$$\begin{aligned} \text{Var}(X) &= E[X(X - 1)] + E(X) - [E(X)]^2 = \frac{2(1 - p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} \\ &= \frac{2 - 2p + p - 1}{p^2} = \frac{1 - p}{p^2}. \end{aligned}$$

Variance of a Sum

- Suppose X and Y are both defined on the same countable sample space (and a probability is defined on it too).
- Then if $E(X) = \mu_X$ and $E(Y) = \mu_Y$ we have already seen that

$$E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y.$$

- What can we say about $\text{Var}(X + Y)$?

- By definition it is

$$\begin{aligned}Var(X + Y) &= E\{(X + Y) - (\mu_X + \mu_Y)\}^2 \\&= E\{(X - \mu_X) + (Y - \mu_Y)\}^2 \\&= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \\&= E[(X - \mu_X)^2] + E[(Y - \mu_Y)^2] + 2E[(X - \mu_X)(Y - \mu_Y)] \\&= Var(X) + Var(Y) + 2Cov(X, Y)\end{aligned}$$

only depending on how X and Y vary *together* through the covariance

$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$; the variances only depend on the individual distributions of the X and Y .

- So in general the variance of a sum is **not** the sum of the variances, at least if the covariance is not zero.
- However we shall see that in a special case that the covariance vanishes.

Joint Distribution of Discrete Random Variables

- Suppose X and Y are *discrete* random variables both defined on the same sample space.
- Then the **joint distribution** of X and Y is the function of two real variables given by

$$\begin{aligned} p(x, y) &= P(X = x, Y = y) \\ &= P\{(X = x) \cap (Y = y)\} \\ &= \sum_{\substack{\omega: X(\omega)=x \\ \text{AND } Y(\omega)=y}} p(\omega). \end{aligned}$$

Example

- Recall the coin-tossing example from last lecture:

ω	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$U(\omega)$	$V(\omega)$	$T(\omega)$	Weight
TTT	0	0	0	0	0	0	$(1-p)^3$
TTH	0	0	1	0	1	1	$p(1-p)^2$
THT	0	1	0	1	1	1	$p(1-p)^2$
THH	0	1	1	1	2	2	$p^2(1-p)$
HTT	1	0	0	1	0	1	$p(1-p)^2$
HTH	1	0	1	1	1	2	$p^2(1-p)$
HHT	1	1	0	2	1	2	$p^2(1-p)$
HHH	1	1	1	2	2	3	p^3

- Consider firstly the joint distribution of X_1 and X_2 :

$$\begin{aligned}
 P(X_1 = 0, X_2 = 0) &= P(\{\text{TTT}, \text{TTH}\}) = (1-p)^3 + p(1-p)^2 \\
 &= (1-p)^2 [(1-p) + p] \\
 &= (1-p)^2
 \end{aligned}$$

$$\begin{aligned}
 P(X_1 = 0, X_2 = 1) &= P(\{\text{THT}, \text{THH}\}) = p(1-p)^2 + p^2(1-p) \\
 &= p(1-p) [(1-p) + p] \\
 &= p(1-p)
 \end{aligned}$$

$$\begin{aligned}
 P(X_1 = 1, X_2 = 0) &= P(\{\text{HTT}, \text{HTH}\}) = p(1-p)^2 + p^2(1-p) \\
 &= p(1-p) [(1-p) + p] \\
 &= p(1-p)
 \end{aligned}$$

$$\begin{aligned}
 P(X_1 = 1, X_2 = 1) &= P(\{\text{HHT}, \text{HHH}\}) = p^2(1-p) + p^3 \\
 &= p^2 [(1-p) + p] \\
 &= p^2.
 \end{aligned}$$

- Consider next the joint distribution of X_1 and U :

$$\begin{aligned} P(X_1 = 0, U = 0) &= P(\{\text{TTT}, \text{TTH}\}) = (1-p)^3 + p(1-p)^2 \\ &= (1-p)^2 [(1-p) + p] \\ &= (1-p)^2 \end{aligned}$$

$$\begin{aligned} P(X_1 = 0, U = 1) &= P(\{\text{THT}, \text{THH}\}) = p(1-p)^2 + p^2(1-p) \\ &= p(1-p) [(1-p) + p] \\ &= p(1-p) \end{aligned}$$

$$P(X_1 = 0, U = 2) = P(\emptyset) = 0$$

$$P(X_1 = 1, U = 0) = P(\emptyset) = 0$$

$$\begin{aligned} P(X_1 = 1, U = 1) &= P(\{\text{HTT}, \text{HTH}\}) = p(1-p)^2 + p^2(1-p) \\ &= p(1-p) [(1-p) + p] \\ &= p(1-p) \end{aligned}$$

$$\begin{aligned} P(X_1 = 1, U = 2) &= P(\{\text{HHT}, \text{HHH}\}) = p^2(1-p) + p^3 \\ &= p^2 [(1-p) + p] \\ &= p^2. \end{aligned}$$

Expectation of $g(X, Y)$

- For any such pair of random variables, for any other function of two variables $g(x, y)$, $Z = g(X, Y)$ is just another random variable.
- The expectation of this new random variable can be expressed as

$$\sum_{\omega \in \Omega} Z(\omega)p(\omega) = \sum_{\omega \in \Omega} g(X(\omega), Y(\omega))p(\omega)$$

- If we decompose this into a *triple* sum according to the possible values of X and Y we get

$$\begin{aligned}& \sum_x \sum_y \left(\sum_{\omega: X(\omega)=x \text{ AND } Y(\omega)=y} g(X(\omega), Y(\omega))p(\omega) \right) \\&= \sum_x \sum_y \left(\sum_{\omega: X(\omega)=x \text{ AND } Y(\omega)=y} g(x, y)p(\omega) \right) \\&= \sum_x \sum_y g(x, y) \left(\sum_{\omega: X(\omega)=x \text{ AND } Y(\omega)=y} p(\omega) \right) \\&= \sum_x \sum_y g(x, y)P(X=x, Y=y)\end{aligned}$$

- We shall return to this after we introduce independent random variables.

Independent Random Variables.

- We have seen how *events* A and B on a sample space can be independent: this means $P(A \cap B) = P(A)P(B)$.
- Random variables in turn can be used to define events e.g. $\{X = 0\}$, $\{Y \geq 2\}$ $\{0 \leq T \leq 2\}$ etc.
- Suppose X and Y are *discrete* random variables. We say that they are **independent** if for all possible values x for X and y for Y ,

$$P(X = x, Y = y) = P\{(X = x) \cap (Y = y)\} = P(X = x)P(Y = y).$$

- That is to say, every event “involving X ” is independent of every event “involving Y ”.

Example (cont'd)

- In the coin-tossing example above, note that

$$P(X_1 = 0, X_2 = 0) = (1 - p)^2 = P(X_1 = 0)P(X_2 = 0)$$

$$P(X_1 = 0, X_2 = 1) = (1 - p)p = P(X_1 = 0)P(X_2 = 1)$$

$$P(X_1 = 1, X_2 = 0) = p(1 - p) = P(X_1 = 1)P(X_2 = 0)$$

$$P(X_1 = 1, X_2 = 1) = p^2 = P(X_1 = 1)P(X_2 = 1)$$

and so X_1 and X_2 are indeed independent (by design in fact!).

- However note that

$$P(X_1 = 1, U = 0) = 0 \neq p(1 - p)^2 = P(X_1 = 1)P(U = 0)$$

so X_1 and U are *not* independent.

- This makes sense, since $U = X_1 + X_2$ clearly *depends* on X_1 .

An Interesting Consequence

- If X and Y (both discrete) are independent then for any two functions $g(\cdot)$ and $h(\cdot)$ we have that

$$\begin{aligned} & E[g(X)h(Y)] \\ &= \sum_x \sum_y g(x)h(y)P(X = x, Y = y) \\ &= \sum_x \sum_y g(x)h(y)P(X = x)P(Y = y) \text{ by independence} \\ &= \sum_x g(x)P(X = x) \sum_y h(y)P(Y = y) \\ &= E[g(X)]E[h(Y)], \end{aligned}$$

that is the say the expectation of the product of **any functions** of X and Y is the corresponding product of expectations.

Applied to Covariance

- Note, in examining the variance of a sum of random variables we found that

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

where $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$, $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

- If X and Y are independent, we have from the “interesting consequence” above that

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E(X - \mu_X)E(Y - \mu_Y) \\ &= [E(X) - \mu_X][E(Y) - \mu_Y] = 0\end{aligned}$$

- We have thus shown that **if X and Y are independent then**

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Outline

1 Welcome

2 Data Analysis

3 Probability

- Lecture 6
- Lecture 7
- Lecture 8
- Lecture 9
- Lecture 10
- **Lecture 11**

- PGF of Sum of Independent Random Variables
- Generalisation of the binomial distribution: multinomial
- A Limiting Case of the binomial: Poisson distribution

- Lecture 12
- Lecture 13
- Lecture 14

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

PGF of Sum of Independent Random Variables

- Suppose X and Y are integer-valued and independent.
- Then the PGF of their sum is $T = X + Y$ is

$$\begin{aligned}\pi_T(s) &= E(s^T) \\ &= E(s^{X+Y}) \\ &= E(s^X s^Y) \\ &= E(s^X)E(s^Y) \text{ by independence} \\ &= \pi_X(s)\pi_Y(s).\end{aligned}$$

- That is, the PGF of the sum is the *product* of the PGFs.

Examples: sum of binomials

- Suppose $X \sim B(m, p)$ and $Y \sim B(n, p)$ are independent. Then we have already seen that
 - ▶ $\pi_X(s) = (1 + p - sp)^m$
 - ▶ $\pi_Y(s) = (1 + p - sp)^n.$

So the sum $T = X + Y$ has PGF

$$\pi_T(s) = \pi_X(s)\pi_Y(s) = (1 + p - sp)^{m+n}.$$

- But this is precisely the $B(m + n, p)$ PGF.
- Thus $T \sim B(m + n, p).$

- This can be extended to the sum of **any number** of independent binomial random variables with the **same** p .
- In particular we can say that if X_1, X_2, \dots, X_n are independent $B(1, p)$ then their sum

$$T = X_1 + X_2 + \cdots + X_n \sim B(n, p).$$

- This is precisely what is going on with the coin-tossing example introduced in lecture 9, which is the $n = 3$ case.
 - ▶ As is easily checked, X_1, X_2, X_3 are independent $B(1, p)$ and $T = X_1 + X_2 + X_3 \sim B(3, p)$ there.

Generalisation of the binomial distribution: multinomial

- The binomial distribution applies when we have a sequence of **independent** trials where each trial has only one of **two** possible outcomes, conveniently called “Success” and “Failure”.
- What if we have more than two?

- Suppose we have 3 possibilities, say, there are balls of 3 different colours in an urn.
 - Let us label the colours R , B and G .
 - Suppose that
 - ▶ the proportion of red balls in the urn is r ,
 - ▶ the proportion of blue balls in the urn is b ,
 - ▶ the proportion of green balls in the urn is g ,
- so that $r + b + g = 1$.
- Suppose also we have 9 trials where we randomly draw from the urn **with replacement**.
 - What is the probability of getting 4 R , 3 B and 2 G ?

- **Solution:** We interpret “randomly” and “with replacement” together to mean the draws are **independent**.
- Then, if the sequence of colours is $RRRRBBBGGG$ this has probability

$$r^4 b^3 g^2,$$

by independence.

- ▶ To see how this construction works more explicitly, generate all outcomes by starting with a collection of events and taking certain intersections of them:

- ★ R_1, R_2, \dots, R_9 all have probability r ;
- ★ B_1, B_2, \dots, B_9 all have probability b ;
- ★ G_1, G_2, \dots, G_9 all have probability g ;
- ★ $R_i \cup B_i \cup G_i = \Omega$ for each $i = 1, 2, \dots, 9$;
- ★ any collection of 9 of these events with all different subscripts are independent,
- ★ any event corresponding to a given sequence of colours can be expressed as the intersection of such a sequence of 9 events e.g. the sequence

$$RRRRBBBGGG = R_1 \cap R_2 \cap R_3 \cap R_4 \cap B_5 \cap B_6 \cap B_7 \cap G_8 \cap G_9$$

- ★ So by independence

$$P(RRRRBBBGGG) = P(R_1)P(R_2)P(R_3)P(R_4)P(B_5)P(B_6)P(B_7)P(G_8)P(G_9) = r^4 b^3 g^2.$$

- Similarly the sequence $BBBRRRRGG$ also has probability $r^4b^3g^2$.
- In fact any possible sequence with 4 R 's, 3 B 's and 2 G 's has probability $r^4b^3g^2$.
- The *real* question is: **How many different possible words (i.e. sequences of length 9) are there with 4 R 's, 3 B 's and 2 G 's?**
- We can construct such a word/sequence in a fixed sequence of steps
 - ① Pick 4 positions for the R 's: there are $\binom{9}{4}$ ways to do this.
 - ② Pick 3 of the remaining 5 positions for the B 's: there are $\binom{5}{3}$ ways to do this.

Then the G 's are placed in the 2 remaining empty slots.

- There are thus $\binom{9}{4} \binom{5}{3}$ such sequences.

- Note also that we can write this product in a slightly different way:

$$\binom{9}{4} \binom{5}{3} = \frac{9!}{4!5!} \frac{5!}{2!3!} = \frac{9!}{4!3!2!}.$$

- Written in this way we call it a **multinomial coefficient**. The binomial coefficient then is the special case where there are only two “colours”.
- Thus the *probability* of getting 4 *R*'s, 3 *B*'s and 2 *G*'s is thus

$$\frac{9!}{4!3!2!} r^4 b^3 g^2.$$

More generally:

- If we have
 - ▶ x_1 of the first letter,
 - ▶ x_2 of the second letter,
 - ▶ \dots
 - ▶ x_k of the k -th letter where $x_1 + x_2 + \dots + x_k = n$

then there are

$$\frac{n!}{x_1!x_2!\cdots x_k!}$$

different possible words we can make.

- This number is known as a **multinomial coefficient** and generalises the binomial coefficient.

- If we are “randomly/independently picking letters” where
 - ▶ the probability of getting the first letter is p_1 ,
 - ▶ the probability of getting the first letter is p_2 ,
 - ▶ ...
 - ▶ the probability of getting the first letter is p_k ,

where $p_1 + p_2 + \cdots + p_k = 1$ then the probability of getting a particular **given** word(sequence) with

- ▶ x_1 of the first letter,
- ▶ x_2 of the second letter,
- ▶ ...
- ▶ x_k of the k -th letter is

$$p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

- The probability of getting **any** of the $\frac{n!}{x_1!x_2!\cdots x_k!}$ possible such words is then

$$\frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

- This is the **multinomial distribution**. The case $k = 2$ gives the binomial distribution but in that case we usually write $x_1 = x$, $x_2 = n - x$, $p_1 = p$, $p_2 = 1 - p$.

A Limiting Case of the binomial: Poisson distribution

- Some situations can be well-modelled by a binomial distribution with a relatively “large n ” but a relatively small “mean np ” e.g.
 - ▶ number of accidents at an intersection in a week,
 - ▶ number of emails arriving in an hour
 - ▶ number of radioactive particles emitted in an hour
 - ▶ number of cells in a microscope slide
 - ▶ etc
- For each of these examples, we can perhaps imagine that a measurement scale (often time and/or space) is divided up into smaller “intervals” and that the number of “occurrences” in each interval is either 0 or 1.
- Then the total number of “occurrences” over the entire observation window could be modelled as a sum of $B(1, p)$ random variables, assuming independence and that the intervals are chosen so that an “occurrence” occurs in each with equal probability.

- In all such cases, how one might choose the intervals is not so important as the overall **expected number** λ of occurrences.
- We can then imagine a binomial $B(n, p)$ random variable with “large n ” and expectation $np = \lambda$, i.e. $p = \lambda/n$.
- Does the exact choice of n matter? It turns out not really, so long as it is large enough.
- How so? Consider the following animation showing ordinate diagrams of binomial $B(n, p)$ distributions with expectation $np = 5$ and increasing n : Poisson-animation.gif
- It appears that so long as $np = 5$, as n increases the probability distribution “settles down” and indeed we seem to have identified the “limiting distribution” in red.
- Indeed we can derive this limiting distribution. Suppose $X \sim B(n, \lambda/n)$ for some fixed $\lambda > 0$. Then note that

$$P(X = 0) = \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}$$

as $n \rightarrow \infty$ (we verify this in an upcoming tutorial exercise).

- For $1 \leq x \leq n$,

$$\begin{aligned}
 P(X = x) &= \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
 &= \frac{n(n-1)\cdots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(\frac{n-\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^n \\
 &= \underbrace{\frac{n}{n-\lambda} \frac{n-1}{n-\lambda} \cdots \frac{n-x+1}{n-\lambda}}_{x \text{ factors}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n \frac{\lambda^x}{x!}}_{\rightarrow e^{-\lambda}}
 \end{aligned}$$

- Note that the x factors in the leading product are all of the form, for non-negative b and (integer) a ,

$$\frac{n-a}{n-b} = \frac{1 - \frac{a}{n}}{1 - \frac{b}{n}} \rightarrow 1$$

as $n \rightarrow \infty$.

- Thus we have shown that for $x = 1, 2, \dots$ as $n \rightarrow \infty$, $P(X = x) \rightarrow \frac{e^{-\lambda} \lambda^x}{x!}$ and indeed this formula also holds for $x = 0$ if we take $0! = 1$.

Definition

- We say that a random variable X has a **Poisson distribution** with mean/expectation λ if

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

for $x = 0, 1, 2, \dots$ (here we understand $0! = 1$). We then write $X \sim \text{Pois}(\lambda)$ for short.

- We “set it up” so that the mean is λ but it would be nice to be able to verify that this distribution gives an expectation of λ directly.
- One way to do this (since X only takes non-negative integer values) is to compute the probability generating function.

Probability Generating Function

- Note firstly that if $X \sim \text{Pois}(\lambda)$ then

$$1 = \sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!}$$

and so we may recover the well-known power series for the exponential function by multiplying both sides by e^λ :

$$e^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{6} + \dots$$

(note $2! = 2$, $3! = 6$, etc.)

- Then the PGF of X is given by

$$\begin{aligned}\pi_X(s) &= E(s^X) = \sum_{x=0}^{\infty} s^x P(X = x) \\ &= \sum_{x=0}^{\infty} s^x \underbrace{\frac{e^{-\lambda} \lambda^x}{x!}}_{e^{\lambda s}} = e^{-\lambda} \underbrace{\sum_{x=0}^{\infty} \frac{(\lambda s)^x}{x!}}_{e^{\lambda s}} = e^{-\lambda} e^{\lambda s} = e^{-\lambda(1-s)}.\end{aligned}$$

- Taking a few derivatives with respect to s we see that

$$\frac{d^m \pi_X(s)}{ds^m} = \lambda^m e^{-\lambda(1-s)}$$

and so the m -th factorial moment

$$E[X(X-1)\cdots(X-m+1)] = \left. \frac{d^m \pi_X(s)}{ds^m} \right|_{s=1} = \lambda^m e^{-\lambda(1-s)} = \lambda^m.$$

- So taking $m = 1$ we see that

$$E(X) = \lambda$$

as we would hope!

- Taking $m = 2$ shows us that $E[X(X-1)] = \lambda^2$ and so

$$\text{Var}(X) = E[X(X-1)] + E(X) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

- This is exactly what we would expect, since $B(n, \lambda/n)$ has variance

$$n \frac{\lambda}{n} \left(1 - \frac{\lambda}{n}\right) = \lambda \left(1 - \frac{\lambda}{n}\right) \rightarrow \lambda$$

as $n \rightarrow \infty$.

Applications

- Suppose the number of insects on a fruit tree is modelled as a $\text{Pois}(3)$ random variable. Find the probability that on a given tree there are 4 such insects.

Solution: $e^{-3}3^4/(4!)$. We can compute this either by using R as a “big calculator”:

```
exp(-3)*(3^4)/factorial(4)
```

```
[1] 0.1680314
```

or using the function `dpois()` as follows:

```
dpois(4,3)
```

```
[1] 0.1680314
```

- In any given time period of length t (in minutes) the number of emails received at a certain address is modelled as a Poisson($2t$) random variable. Find

- ▶ the probability of 10 emails in the next 2 minutes

Solution: The number in the next 2 minutes is Poisson(4) and so

```
dpois(10, 4)
```

```
[1] 0.005292477
```

- ▶ the probability of at most 2 emails in the next 5 minutes.

Solution: The number in the next 5 minutes is Poisson(10) and so we want

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$ which we can compute various ways, perhaps the most instructive being

```
x=0:2
```

```
x
```

```
[1] 0 1 2
```

```
dpois(x, 10)
```

```
[1] 4.539993e-05 4.539993e-04 2.269996e-03
```

```
sum(dpois(x, 10))
```

```
[1] 0.002769396
```

Sum of Independent Poissons

- Suppose that $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ are independent.
- Then the PGF of $T = X + Y$ is

$$\begin{aligned}\pi_T(s) &= E(s^T) = E(s^{X+Y}) \\ &= E(s^X) E(s^Y) \text{ by independence,} \\ &= e^{-\lambda(1-s)} e^{-\mu(1-s)} \\ &= e^{-(\lambda+\mu)(1-s)}.\end{aligned}$$

- But this is just the PGF of the $\text{Pois}(\lambda + \mu)$ distribution.
- Thus the sum of independent Poisson random variables also has a Poisson distribution.

Outline

1 Welcome

2 Data Analysis

3 Probability

- Lecture 6
- Lecture 7
- Lecture 8
- Lecture 9
- Lecture 10
- Lecture 11
- Lecture 12

- Continuous Random Variables
- Examples: the general uniform distribution
- The general exponential distribution
- Changes of Location and/or Scale
- The general normal distribution

- Lecture 13
- Lecture 14

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

Continuous Random Variables

- We have already seen some examples of continuous random variables: e.g. the time until a radioactive particle decays.
- Such a random variable X is completely described by its **probability density function** (PDF) $f_X(x)$, so that for any x ,

$$P(X \leq x) = \int_{-\infty}^x f_X(t) dt .$$

- Note that we can use any dummy variable inside the integral **except** x , since it is appearing in the limits of integration.
- The function

$$F_X(x) = P(X \leq x)$$

is called the *cumulative distribution function* (CDF) of X and completely determines the distribution of X .

- For instance, if $-\infty < b < a < \infty$ we can say that if $X \leq b$ then it is also $\leq a$ or not, that is

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}.$$

- Moreover the two events on the RHS are mutually exclusive and so by the third axiom,

$$P(X \leq b) = P(X \leq a) + P(a < X \leq b)$$

which means we can write

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) \\ &= \int_a^b f_X(x) dx. \end{aligned}$$

- Note also that for any single point x ,

$$P(X = x) \leq P(x - \varepsilon < X \leq x) = F_X(x) - F_X(x - \varepsilon)$$

for all $\varepsilon > 0$.

- Thus if $F_X(x)$ is *continuous* at x then we see that letting $\varepsilon \rightarrow 0$, $F_X(x - \varepsilon) \rightarrow F_X(x)$ and so

$$P(X = x) = 0.$$

- **Note:** By the fundamental theorem of calculus, if the PDF is continuous then it is the derivative of the CDF:

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

- **Note:** $P(X \leq x)$ is also well-defined *for all real x* if X is discrete; **all random variables (whether discrete or continuous) have a CDF.**

Examples: the general uniform distribution

- **Uniform distribution** For any $a < b$, if X has the uniform distribution over (a, b) then its PDF is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

- The corresponding CDF is

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x \geq b \end{cases}$$

and so is piecewise linear and continuous.

- We then write $X \sim U(a, b)$ for short.

The general exponential distribution

- A random variable X is said to have an *exponential distribution with mean μ* if its PDF is

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The standard normal distribution

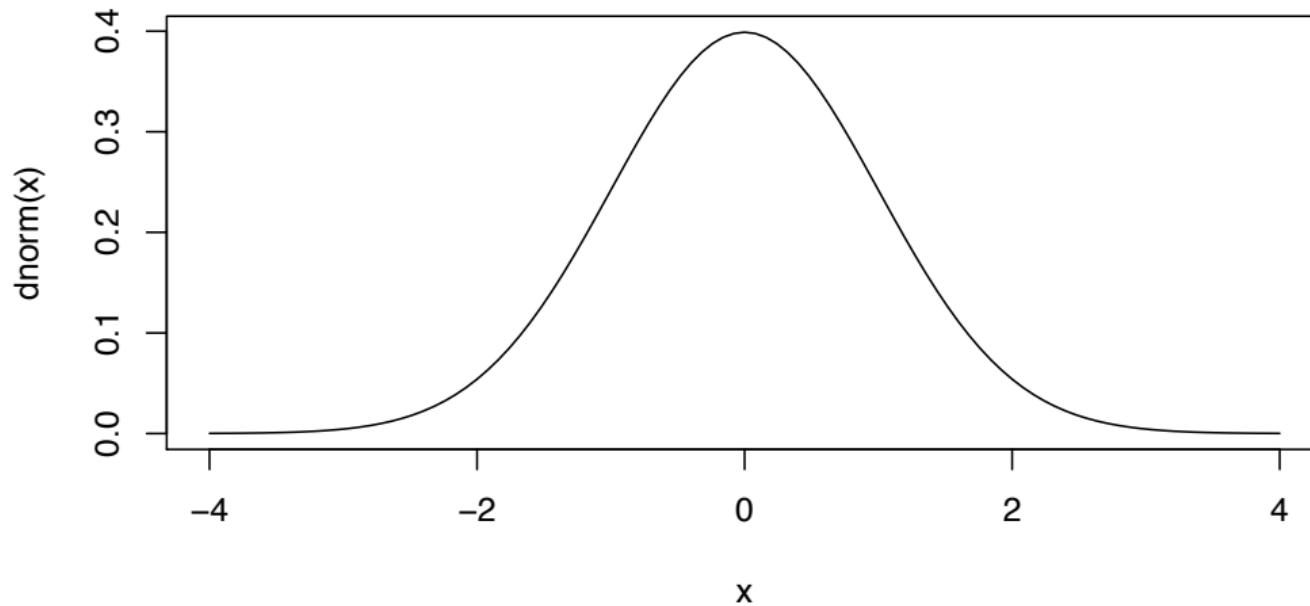
- The **standard normal distribution** is that with PDF given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

for all real x .

- The R function `dnorm(x)` computes $\phi(x)$:

```
curve(dnorm(x), from=-4, to=4)
```



- If X has this distribution we write $X \sim N(0, 1)$ for short (we see the significance of the 0 and 1 shortly).
- The constant $\frac{1}{\sqrt{2\pi}}$ is (again) a normalisation constant so the PDF integrates to 1 (*this can be proved in second year by changing a two-dimensional integral to polar coordinates*).

Expectations: $E(X)$

- By analogy with the discrete case, the expectation of a continuous random variable X with PDF $f_X(x)$ is given by

$$E(X) = \int_{-\infty}^{\infty} xf_X(x) dx,$$

so long as the integral is well-defined.

- One case where it is *not* is the Cauchy distribution where the PDF is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \text{ for } -\infty < x < \infty.$$

- To see this, note that although $\int_{-\infty}^{\infty} f_X(x) dx = 1$,

$$\int_0^{\infty} xf_X(x) dx = +\infty \text{ whereas } \int_{-\infty}^0 xf_X(x) dx = -\infty$$

(these two are verified in the tutorial) and so

$$E(X) = \int_{-\infty}^{\infty} xf_X(x) dx = \int_0^{\infty} xf_X(x) dx + \int_{-\infty}^0 xf_X(x) dx = \infty - \infty$$

which is undefined.

Other examples

- If $X \sim U(a, b)$ then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf_X(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{(b+a)(b-a)}{2(b-a)} \\ &= \frac{a+b}{2}, \end{aligned}$$

the midpoint of the interval which makes sense since the distribution is *symmetric* about that point.

- If $X \sim N(0, 1)$ then since $x\phi(x)$ is an **odd** function of x ,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\phi(x) dx = \int_{-\infty}^0 x\phi(x) dx + \int_0^{\infty} x\phi(x) dx \\ &= - \int_0^{\infty} x\phi(x) dx + \int_0^{\infty} x\phi(x) dx \end{aligned}$$

which is zero, so long as the integral from 0 to ∞ is finite.

- But note that

$$\begin{aligned} \int_0^{\infty} x\phi(x) dx &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} xe^{-\frac{1}{2}x^2} dx = \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{1}{2}x^2} \right]_0^{\infty} \\ &= \frac{1}{\sqrt{2\pi}} [0 - (-1)] \\ &= \frac{1}{\sqrt{2\pi}} < \infty \end{aligned}$$

and so $E(X) = 0$.

- If X is exponential then $E(X)$ is given by the integral

$$E(X) = \int_{-\infty}^{\infty} xf_X(x) dx = \int_0^{\infty} x \frac{1}{\mu} e^{-\frac{x}{\mu}} dx.$$

- Changing variables inside the integral via $y = x/\mu$, we get $x = y\mu$, $dx = \mu dy$ and so

$$E(X) = \int_0^{\infty} ye^{-y} \mu dy = \mu \int_0^{\infty} ye^{-y} dy = \mu,$$

(unsurprisingly!).

Expectations: $E[g(X)]$

- More generally, for any function $g(\cdot)$,

$$E [g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- In particular as in the discrete case with $\mu_X = E(X)$, the variance of X is given by

$$\text{Var}(X) = E [(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

and indeed the computing formula

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

also holds in the continuous case.

Linear Functions

- As in the discrete case, if $Y = a + bX$ for constants $b > 0$ and a then

$$E(Y) = E(a + bX) = a + bE(X)$$

and

$$\text{Var}(Y) = \text{Var}(a + bX) = b^2 \text{Var}(X).$$

- Examples

- Suppose $X \sim U(0, 1)$. Then the m -th moment is

$$E(X^m) = \int_{-\infty}^{\infty} x^m f_X(x) dx = \int_0^1 x^m dx = \left[\frac{x^{m+1}}{m+1} \right]_0^1 = \frac{1}{m+1}.$$

So then in particular

- $E(X^2) = \frac{1}{3}$;
- $\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$.

- Suppose $X \sim N(0, 1)$. Then since $E(X) = 0$, we can write

$$\text{Var}(X) = E(X^2) = \int_{-\infty}^{\infty} x^2 \phi(x) dx = \int_{-\infty}^{\infty} (-x)(-x\phi(x)) dx .$$

- Then, since $\frac{d\phi(x)}{dx} = -x\phi(x)$, integrating by parts gives

$$\text{Var}(X) = [-x\phi(x)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(x) dx = 0 - 0 + 1 = 1$$

(this uses the fact that $x\phi(x) \rightarrow 0$ as $x \rightarrow \pm\infty$).

- Suppose X is exponential with mean μ . Then

$$E(X^2) = \int_0^\infty x^2 \frac{1}{\mu} e^{-\frac{x}{\mu}} dx = \mu^2 \int_0^\infty y^2 e^{-y} dy$$

again after changing variables inside the integral.

- It can be shown via induction (and integration by parts) that for positive integer m ,
 $\int_0^\infty y^m e^{-y} dy = m!$, thus

$$E(X^2) = 2\mu^2$$

and thus

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = 2\mu^2 - \mu^2 = \mu^2.$$

Changes of Location and/or Scale

- In the case of a linear function, $Y = a + bX$ we can say more than just what $E(Y)$ and $\text{Var}(Y)$ are: *we can write out the whole distribution of Y in terms of that of X .*
- Suppose X is a continuous random variable with CDF $F_X(x) = P(X \leq x)$ and PDF $f_X(x) = \frac{dF_X(x)}{dx}$.
- What is the distribution of $Y = X + a$, for any constant a ?
- We can easily write down the CDF of Y :

$$F_Y(y) = P(Y \leq y) = P(X + a \leq y) = P(X \leq y - a) = F_X(y - a).$$

Differentiating we get the PDF of Y :

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(y - a)}{dy} = f_X(y - a).$$

- Suppose now that $Y = bX$ for some constant $b > 0$. In a similar way we can write down the CDF of Y :

$$F_Y(y) = P(Y \leq y) = P(bX \leq y) = P(X \leq y/b) = F_X(y/b).$$

(since $b > 0$, direction of inequality doesn't change).

- Differentiating (paying careful heed to the chain rule), we get the PDF of Y :

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= \frac{dF_X(y/b)}{dy} \stackrel{\text{chain rule}}{=} \frac{dF_X(x)}{dx} \Big|_{x=y/b} \frac{d}{dy} \left(\frac{y}{b} \right) = \frac{1}{b} f_X \left(\frac{y}{b} \right). \end{aligned}$$

- Let's combine these two steps. Suppose X is a continuous random variable with CDF $F_X(x) = P(X \leq x)$ and PDF $f_X(x) = \frac{dF_X(x)}{dx}$.
- For some constants $b > 0$ and a define $Y = a + bX$. Then Y has CDF

$$F_Y(y) = F_X\left(\frac{y-a}{b}\right)$$

and differentiating gives that Y has PDF

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy} F_X\left(\frac{y-a}{b}\right) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right).$$

- The previous two results are then special cases of this last one.

Examples

- **Uniform** Suppose $X \sim U(0, 1)$. Then X has PDF

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

- If for some $b > 0$ we define $Y = bX$ then Y has PDF

$$\begin{aligned} f_Y(y) &= \frac{1}{b} f_X\left(\frac{y}{b}\right) = \begin{cases} \frac{1}{b} & \text{if } 0 < \frac{y}{b} < 1 \\ 0 & \text{otherwise;} \end{cases} \\ &= \begin{cases} \frac{1}{b} & \text{if } 0 < y < b \text{ (since } b > 0\text{)} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

which is of course the $U(0, b)$ PDF.

The general normal distribution

- **Normal** Suppose $Z \sim N(0, 1)$ with PDF $f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ and for some $\sigma > 0$ and μ we define $X = \mu + \sigma Z$.
- Then we immediately have that
 - ▶ $E(X) = \mu + \sigma E(Z) = \mu$ and
 - ▶ $Var(X) = \sigma^2 Var(Z) = \sigma^2$(recall that $E(Z) = 0$ and $Var(Z) = 1$).
- But of course we can say more. The PDF of X is given by

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

- We then say that X has a **normal distribution with mean μ and variance σ^2** and we write $X \sim N(\mu, \sigma^2)$ for short (this explains the $N(0, 1)$ notation: “with mean 0 and variance 1”).

- Note also that the exponential distribution with mean μ can be obtained via a scale change: if
 - ▶ X is exponential with mean 1 and
 - ▶ we define $Y = \mu X$,then Y is exponential with mean μ .
- To see this note that the PDFs of X and Y are related via

$$f_Y(y) = \frac{1}{\mu} f_X\left(\frac{y}{\mu}\right).$$

Outline

1 Welcome

2 Data Analysis

3 Probability

- Lecture 6
- Lecture 7
- Lecture 8
- Lecture 9
- Lecture 10
- Lecture 11
- Lecture 12
- **Lecture 13**

- Independent continuous random variables
- Random Sample Sums and Means
- Sum of Independent Normal Random Variables
- Normal approximation to Binomial probabilities
- Continuity Correction

- Lecture 14

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

Comment on marginal distributions

- For two discrete random variables X and Y we have already seen that for any function $g(x, y)$, the new random variable $g(X, Y)$ has expectation given by the following double sum: $E[g(X, Y)] = \sum_x \sum_y g(x, y)P(X = x, Y = y)$ where the sum extends over all possible pairs (x, y) . However if $g(x, y) = g(x)$ does not depend on y we see that

$$\begin{aligned} E[g(X)] &= \sum_x \sum_y g(x)P(X = x, Y = y) \\ &= \sum_x g(x) \left[\sum_y P(X = x, Y = y) \right] \\ &= \sum_x g(x)P(X = x) \end{aligned}$$

- This shows us that to recover the distribution of X only (ignoring Y) that is its *marginal distribution* we simply add the joint distribution over y :
 $P(X = x) = \sum_y P(X = x, Y = y).$

Brief recap: independent *discrete* random variables

- Discrete random variables X_1, \dots, X_n are independent if and only if for all possible x_1, \dots, x_n ,

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n).$$

- A consequence of this is that a similar statement holds if we replace each “=” sign (inside the probabilities) with a “ \leq ”:

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n). \quad (*)$$

- The left-hand side is called the *joint* CDF of X_1, \dots, X_n ; independence means the joint CDF is the product of the *marginal* CDFs.
- We have seen explicit examples of samples spaces (and probabilities defined on its subsets) where independent *discrete* RVs exist.

Independent *continuous* random variables

- How do we characterise independence of continuous random variables?
- Recall that for any n events A_1, \dots, A_n , $A_1 \cap \dots \cap A_n \subset A_1$ and so

$$P(A_1 \cap \dots \cap A_n) \leq P(A_1).$$

- Recall also that $P(X_1 = x_1) = 0$ for a continuous RV X_1 and all real x_1 , and so for continuous RVs X_1, \dots, X_n

$$P(X_1 = x_1, \dots, X_n = x_n) \leq P(X_1 = x_1) = 0$$

and thus we trivially have

$$P(X_1 = x_1, \dots, X_n = x_n) = 0 = P(X_1 = x_1) \cdots P(X_n = x_n)$$

for **any** collection of continuous random variables.

- We thus need a *different* way to characterise independence for continuous random variables

Use the CDF form

- Note however that the *second* characterisation of independence (*) on slide (303) also makes sense for continuous random variables.
- Thus **any** random variables X_1, \dots, X_n (whether discrete or continuous) are said to be independent if and only if

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n).$$

- To properly construct examples of independent continuous random variables we need the idea of multiple integrals which are not formally met until second year mathematics.
- We shall just note that it is possible to construct a sample space (and probabilities on its subsets) on which exist random variables X_1, \dots, X_n with any desired distribution (discrete or continuous) such that they are all *independent*.

Random Sample Sums and Means

- In many situations we might model data as values taken by independent random variables X_1, X_2, \dots, X_n , all with the same distribution (characterised by a CDF $F(\cdot)$).
- In such a situation we often say that these form a “random sample” from a “population” described by $F(\cdot)$.
- We might then identify the common expectation and variance of the X_i 's as the “population mean” μ and “population variance” σ^2 .

- Then it is straightforward to show that with
 - ▶ $T = X_1 + \cdots + X_n$ the sample total (sum) and
 - ▶ $\bar{X} = T/n = \frac{1}{n}(X_1 + \cdots + X_n)$ the sample mean,
- we have

$$E(T) = E(X_1) + \cdots + E(X_n) = n\mu$$

$$\text{Var}(T) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = n\sigma^2 \text{ by independence,}$$

$$E(\bar{X}) = \frac{1}{n}E(T) = \frac{1}{n}n\mu = \mu \text{ and}$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\text{Var}(T) = \frac{\sigma^2}{n}.$$

Sum of Independent Normal Random Variables

- An important special case is when we have independent normal random variables.
- Suppose $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent.
- Then we *already* have that with $T = X + Y$,
 - ▶ $E(T) = \mu_X + \mu_Y$
 - ▶ $\text{Var}(T) = \sigma_X^2 + \sigma_Y^2$
- The *special* extra result is that **T is also normally distributed.**
 - ▶ This needs multiple integrals to prove; we thus defer proving this until second year.
- This is of course easily extended to sums of more than two independent normal random variables.

- An important special case is where X_1, X_2, \dots, X_n are independent $N(\mu, \sigma^2)$ random variables (i.e. all have the same distribution). Then
 - ▶ $T = X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$;
 - ▶ $\bar{X} = T/n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

We can also say that the **standardised sum/average**

$$\frac{T - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Normal as (another) limiting version of a Binomial

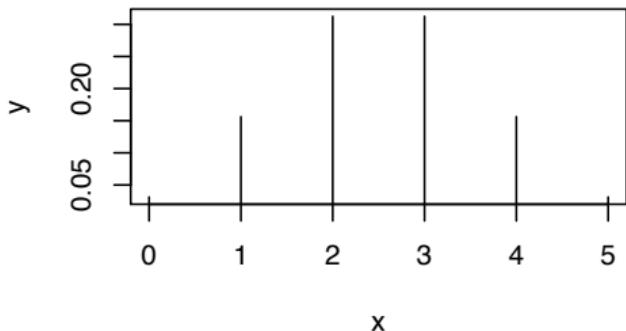
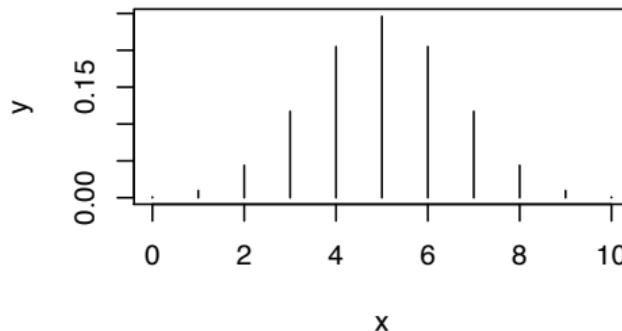
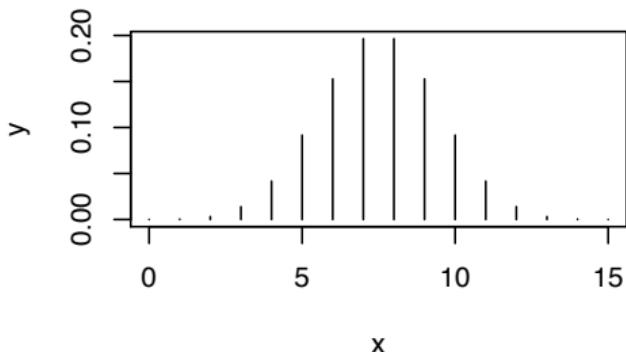
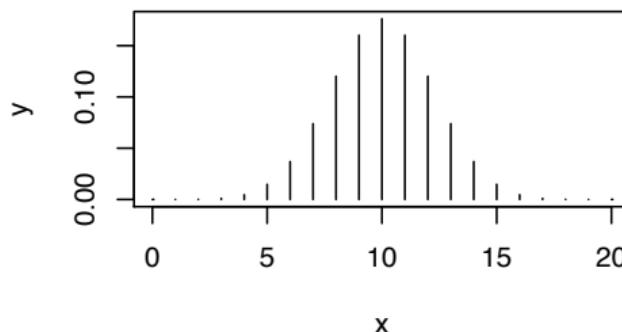
- We have already seen that if $X \sim B(n, p)$ and “ n is large”, we can perhaps use the Poisson distribution as an approximation to the distribution of X .
- However this only holds when p is small, specifically if the product np is “not too large”, preferably a known number.
- A complementary scenario is when
 - ▶ $X \sim B(n, p)$
 - ▶ n is “large”
 - ▶ p is “not necessarily small”.

Examples

- X is the number of boys (not identical twins!) in the next 1000 births at a hospital
- X is the number of “dwarf” pea plants in a plot of 300 such plants, $\frac{1}{4}$ of which are predicted to be “dwarf”.
- etc.

Consider the following graphical output, showing ordinate diagrams of $B(n, 0.5)$ distributions for progressively larger values of n :

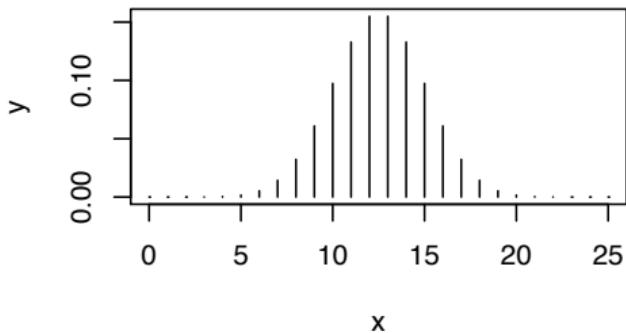
```
par(mfrow=c(2,2))
n=c(5,10,15,20)
for(i in 1:4){
  x=0:n[i]
  y=dbinom(x,n[i],0.5)
  lab=paste("B(",n[i],",0.5)",sep="")
  plot(x,y,type="h",main=lab)
}
```

$B(5,0.5)$  $B(10,0.5)$  $B(15,0.5)$  $B(20,0.5)$ 

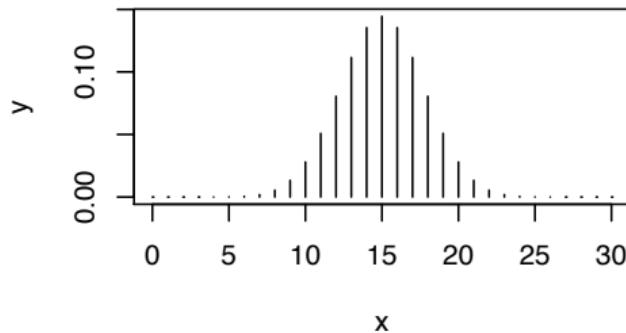
```
par(mfrow=c(2,2))
n=c(25,30,35,40)
for(i in 1:4){
  x=0:n[i]
  y=dbinom(x,n[i],0.5)
  lab=paste("B(",n[i],",0.5)",sep="")
  plot(x,y,type="h",main=lab)
}
```

It is clear that the distributions are becoming “bell-shaped”:

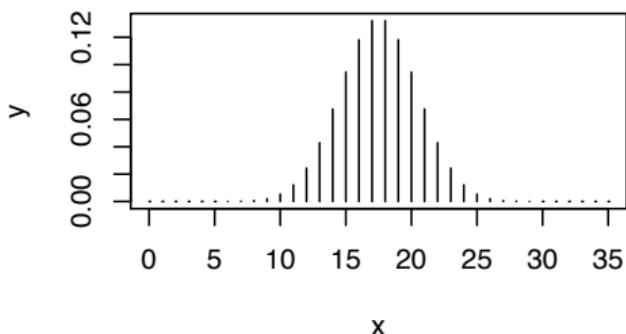
B(25,0.5)



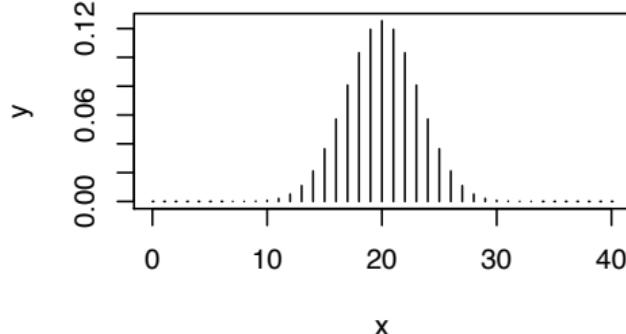
B(30,0.5)



B(35,0.5)

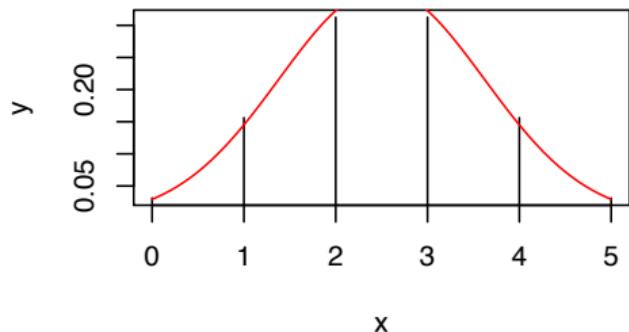
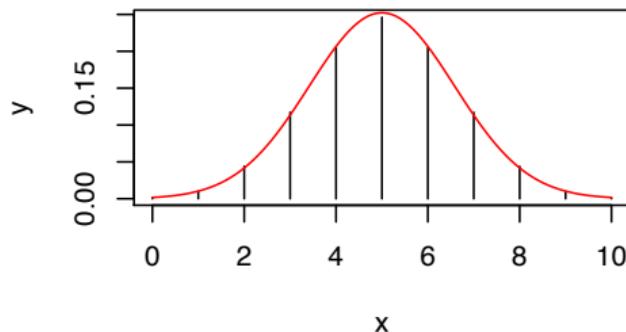
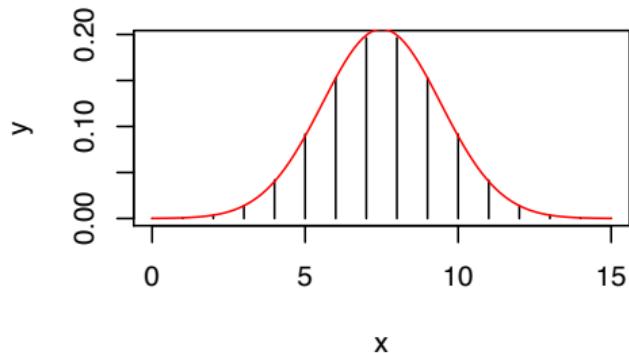
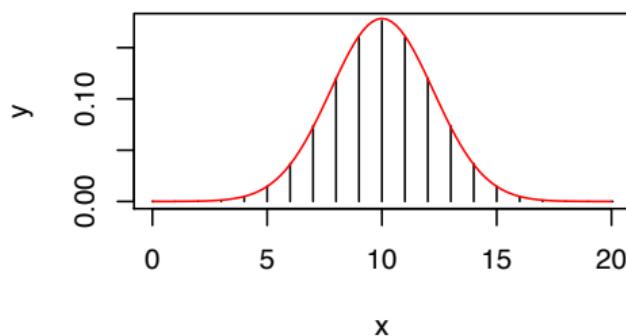


B(40,0.5)



Look what happens when we superimpose the *normal* PDF with the same mean and variance:

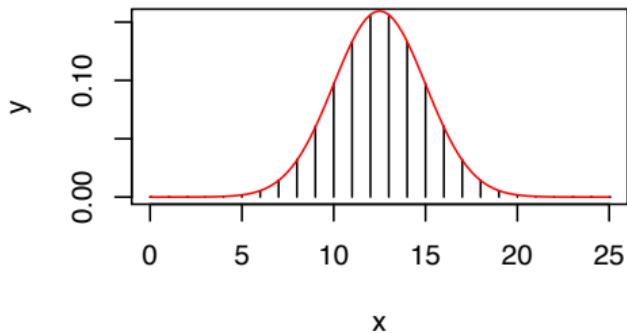
```
par(mfrow=c(2,2))
n=c(5,10,15,20)
for(i in 1:4){
  x=0:n[i]
  y=dbinom(x,n[i],0.5)
  lab=paste("B(",n[i],",0.5)",sep="")
  plot(x,y,type="h",main=lab)
  curve(dnorm(x,m=n[i]*.5,s=sqrt(n[i])/2),add=T,col="red")
}
```

$B(5,0.5)$  $B(10,0.5)$  $B(15,0.5)$  $B(20,0.5)$ 

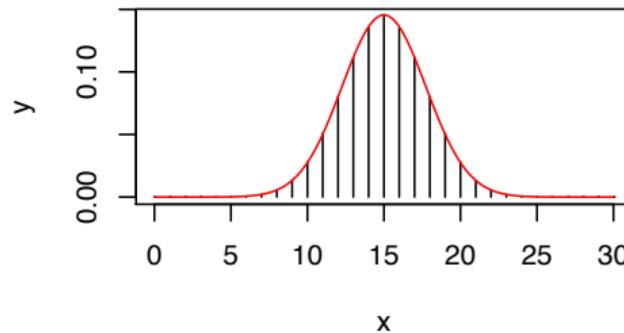
```
par(mfrow=c(2,2))
n=c(25,30,35,40)
for(i in 1:4{
  x=0:n[i]
  y=dbinom(x,n[i],0.5)
  lab=paste("B(",n[i],",0.5)",sep="")
  plot(x,y,type="h",main=lab)
  curve(dnorm(x,m=n[i]*.5,s=sqrt(n[i])/2),add=T,col="red")
}
```

The red normal curve is uncannily close to the tops of the ordinates:

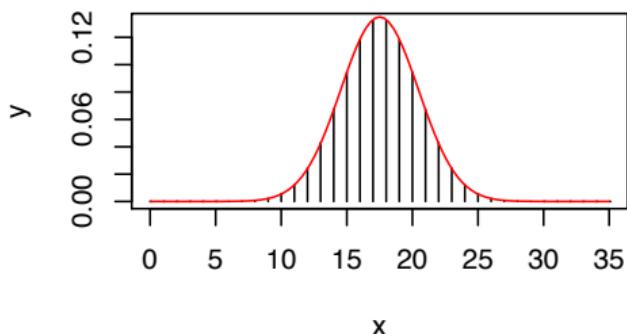
B(25,0.5)



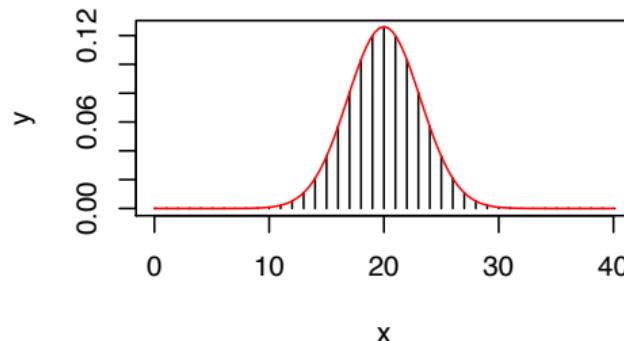
B(30,0.5)



B(35,0.5)

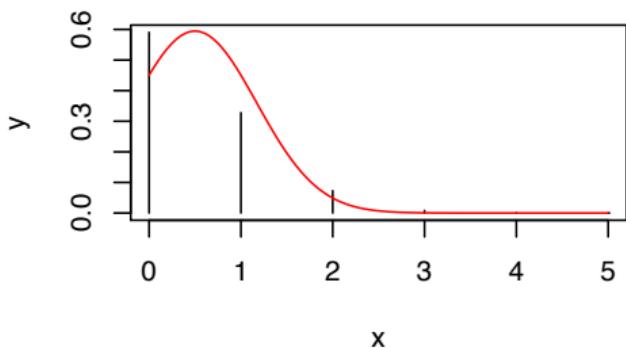
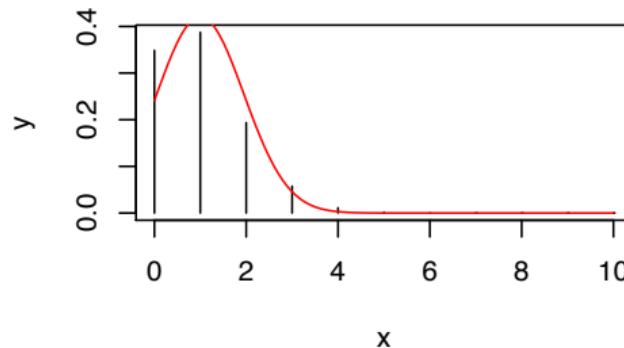
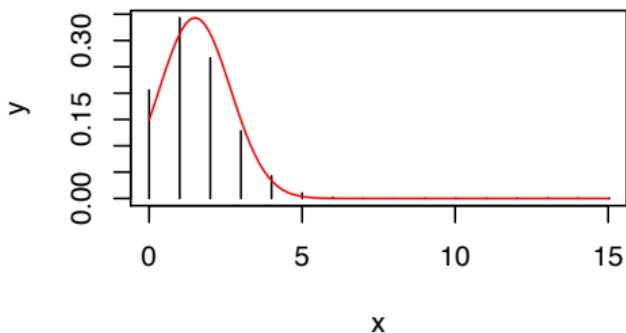
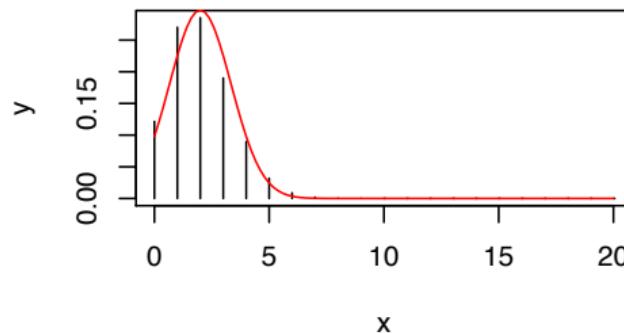


B(40,0.5)



Let us try with a different $p \neq \frac{1}{2}$ (so the resulting binomial distribution is **not** symmetric):

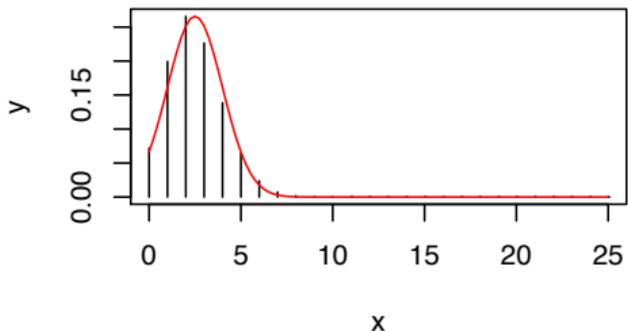
```
p=0.1
par(mfrow=c(2,2))
n=c(5,10,15,20)
for(i in 1:4){
  x=0:n[i]
  y=dbinom(x,n[i],p)
  lab=paste("B(",n[i],",",p,",")",sep="")
  plot(x,y,type="h",main=lab)
  curve(dnorm(x,m=n[i]*p,s=sqrt(n[i]*p*(1-p))),add=T,col="red")
}
```

$B(5,0.1)$  $B(10,0.1)$  $B(15,0.1)$  $B(20,0.1)$ 

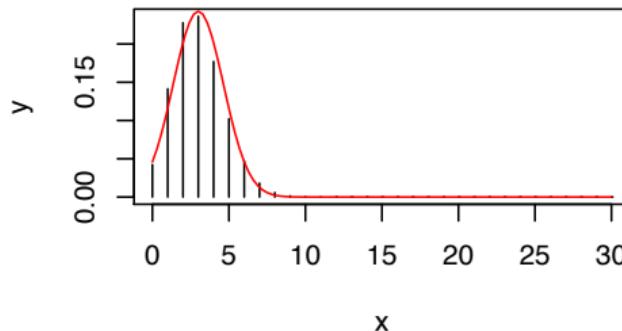
```
par(mfrow=c(2,2))
n=c(25,30,35,40)
for(i in 1:4){
  x=0:n[i]
  y=dbinom(x,n[i],p)
  lab=paste("B(",n[i],",",",",p,",")",sep="")
  plot(x,y,type="h",main=lab)
  curve(dnorm(x,m=n[i]*p,s=sqrt(n[i]*p*(1-p))),add=T,col="red")
}
```

For $p = 0.1$ the agreement isn't quite so good except for the larger n , but it is clearly improving as n increases:

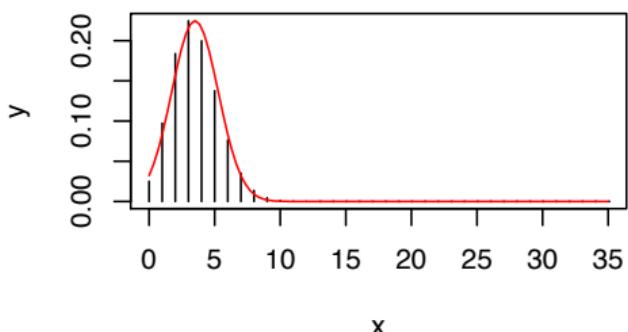
B(25,0.1)



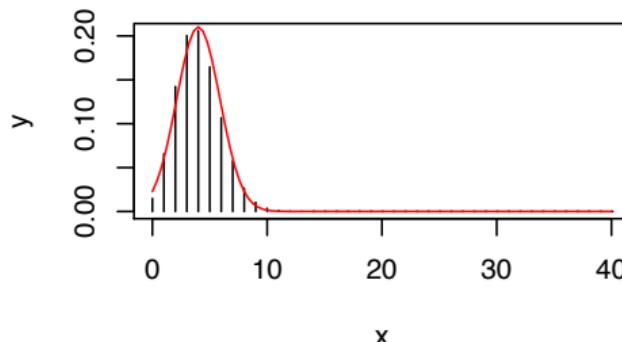
B(30,0.1)



B(35,0.1)



B(40,0.1)



We can express this observed “closeness” informally as follows:

If $X \sim B(n, p)$ **for “large n”, then** $P(X = x) \approx \frac{1}{\sqrt{np(1-p)}} \phi \left(\frac{x-np}{\sqrt{np(1-p)}} \right).$

Normal approximation to Binomial probabilities

- Since a binomial random variable takes integer values, the ordinates on these plots are all 1 unit apart.
- Thus if we were to add together the heights of these up to some point, say x , then the corresponding area under the approximating normal PDF would be “close” to the sum of all the “heights”.
- That is to say, as a naive first approximation, we might approximate the binomial probability

$$P(X \leq x) \approx P(Y \leq x)$$

where $Y \sim N(np, np(1 - p))$, that is Y is normal with the same expectation (mean) and variance as X .

Continuity Correction

- However, if we think about this a little we perceive a slight problem.
- Since X is integer-valued,

$$P(X \leq x) = P(X < x + 1).$$

- However, since Y is continuous,

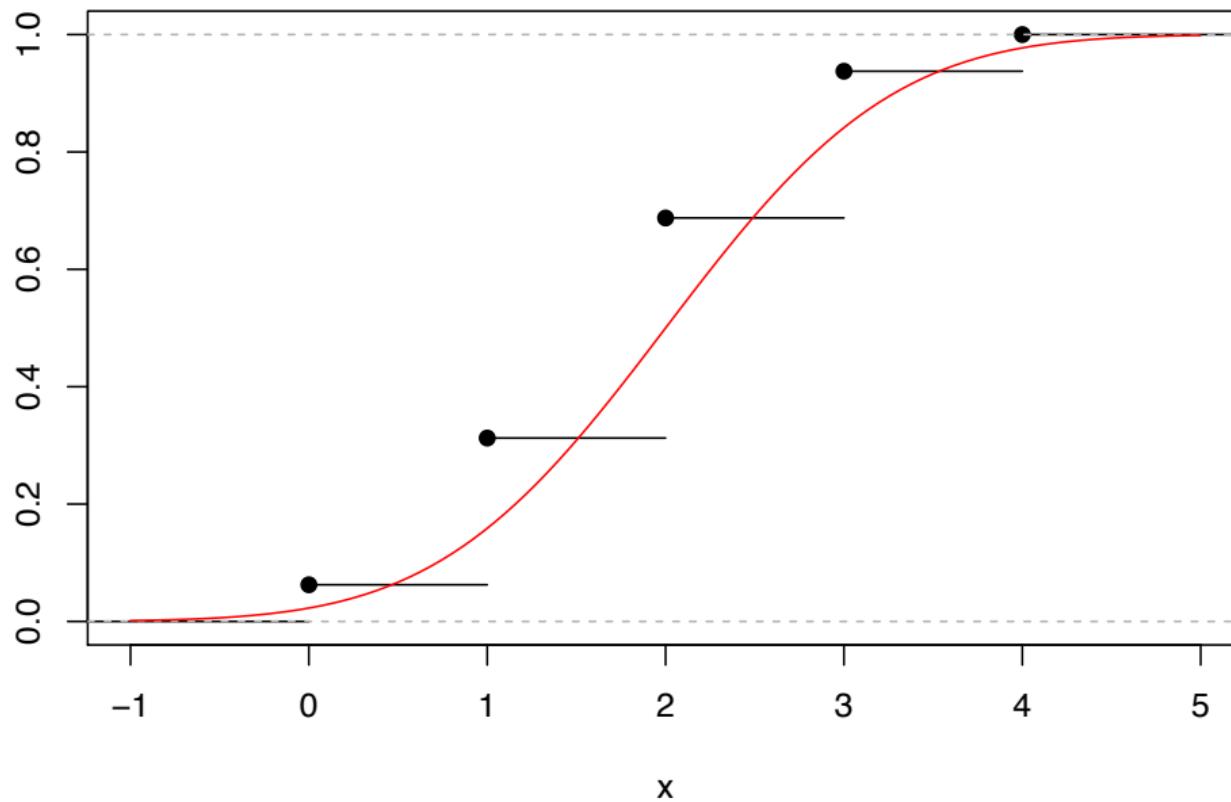
$$P(Y \leq x) = P(Y < x) \neq P(Y < x + 1).$$

- So we have two different normal approximations to the same binomial probability. We can get a slightly better insight into what is going on here by directly comparing the CDFs.

- Below appears the CDF $P(X \leq x)$ of $X \sim B(4, 0.5)$ (in black, a step function) and the CDF $P(Y \leq y)$ of the approximating normal random variable $Y \sim N(2, 1)$ (in red):

```
p=dbinom(0:4,4,.5)
x=rep(0:4,16*p)
plot(ecdf(x),main="B(4,0.5) CDF (black) N(2,1) CDF (red)",ylab="")
curve(pnorm(x,2,1),add=T,col="red")
```

$B(4,0.5)$ CDF (black) $N(2,1)$ CDF (red)



- Note that while the (continuous) red curve “follows” the black step function quite closely, it is *impossible* for it to provide a good approximation at the jump points (the integers).
- However **half-way** between the integers the two are very close.
- This suggests using the approximation for **integer** x :

$$P(X \leq x) = P\left(X \leq x + \frac{1}{2}\right) \approx P\left(Y \leq x + \frac{1}{2}\right).$$

- This “ $+\frac{1}{2}$ ” is known as a **correction for continuity** and improves normal approximations to binomial probabilities.

- How about we try $P(Y > y)$ for a value y half-way in between, that is

$$\begin{aligned}P(X > 7.5) \approx P(Y > 7.5) &= P\left(Z > \frac{7.5 - 8}{\sqrt{4.8}}\right) \\&= 1 - \Phi\left(-\frac{1}{2\sqrt{4.8}}\right) \\&= \Phi\left(+\frac{1}{2\sqrt{4.8}} \approx 0.23\right) \approx 0.5910.\end{aligned}$$

- This is *much closer* to the true value, giving an error of only about 0.007!

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
 - Lecture 6
 - Lecture 7
 - Lecture 8
 - Lecture 9
 - Lecture 10
 - Lecture 11
 - Lecture 12
 - Lecture 13
 - **Lecture 14**
 - The Sample Variance
 - The χ^2 (Chi-Squared) Distribution
 - The χ_d^2 distribution
 - Joint Distribution of a Normal Sample Mean and Sample Variance
 - Pivots; the Studentised mean
 - The Central Limit Theorem
 - Probability Wrap-up
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models

The Sample Variance

- If X_1, X_2, \dots, X_n are independent random variables, each with expectation μ and variance σ^2 then we have seen that the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ has expectation and variance given by

$$E(\bar{X}) = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

- We can also consider the two forms of variance of the X_i 's: the population variance

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

and the *sample variance*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

- We are now in a position to understand why this last one is used, by comparing the expectations of the two random variables V and S^2 .
- To facilitate this, use the computing formula for the sum of squared deviations from the mean:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} = \sum_{i=1}^n X_i^2 - n(\bar{X})^2.$$

- Taking expectations gives

$$E \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] = \sum_{i=1}^n E(X_i^2) - nE(\bar{X})^2$$

giving an expression involving the *second moments* of each X_i and \bar{X} .

- Then note that for any random variable Y , the computing formula for the $\text{Var}(Y) = E(Y^2) - [E(Y)]^2$ means we can write the second moment in terms of the expectation and variance, i.e.

$$E(Y^2) = \text{Var}(Y) + [E(Y)]^2.$$

- Since for each i , $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ we get

$$E(X_i^2) = \sigma^2 + \mu^2$$

while for \bar{X} we get

$$E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2.$$

- Substituting these in gives

$$\begin{aligned} E \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] &= \sum_{i=1}^n (\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) \\ &= n(\sigma^2 + \mu^2) - (\sigma^2 + n\mu^2) \\ &= (n-1)\sigma^2, \end{aligned}$$

the terms involving μ cancelling out.

- Hence we see that

$$E(V) = E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \frac{1}{n}(n-1)\sigma^2 = \left(1 - \frac{1}{n}\right) \sigma^2$$

while

$$E(S^2) = E \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

- Thus if we wanted to use either of these to **estimate** σ^2 , we would find that “on average”
 - ▶ V would slightly “underestimate” σ , while
 - ▶ S^2 would not;
- S^2 is “on target” in the sense of expectation.
- This is the main reason the $\frac{1}{n-1}$ -version sample variance is used: its expectation is the “population”“ variance σ^2 .

The χ^2 (Chi-Squared) Distribution

- Suppose that $Z \sim N(0, 1)$. What is the distribution of $X = Z^2$?
- The CDF of X is given by, for any $x > 0$,

$$\begin{aligned}F_X(x) &= P(X \leq x) = P(Z^2 \leq x) \\&= P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\&= P(Z \leq \sqrt{x}) - P(Z < -\sqrt{x}) \\&= \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) \\&= \Phi(\sqrt{x}) - [1 - \Phi(\sqrt{x})] \\&= 2\Phi(\sqrt{x}) - 1\end{aligned}\tag{\dagger}$$

where the equality (\dagger) follows from the symmetry of the $N(0, 1)$ distribution.

- Also, $F_X(x) = 0$ for $x \leq 0$ since $X = Z^2$ can only take non-negative values, and $P(Z = 0) = 0$.

- For $x > 0$, the PDF is obtained by differentiating the CDF:

$$f_X(x) = \frac{dF_X(x)}{dx} = 2\phi(\sqrt{x}) \frac{1}{2}x^{-1/2} = \frac{1}{\sqrt{2\pi}}x^{\frac{1}{2}-1}e^{-\frac{1}{2}x}$$

using the chain rule (we write it in this form for reasons that will become apparent).

- This is the PDF of the **chi-squared distribution with 1 degree of freedom**.
- If X has this distribution we write $X \sim \chi_1^2$.

The χ_d^2 distribution

- Suppose X_1, \dots, X_d are independent χ_1^2 random variables.
- Then their sum $Y = X_1 + \dots + X_d$ is said to have the **chi-squared distribution with d degrees of freedom** and we write $Y \sim \chi_d^2$.
- It turns out (this is not proved until second year) that the PDF of such a Y is of the form

$$f_Y(y) = C_d y^{\frac{d}{2}-1} e^{-\frac{y}{2}}$$

for $y > 0$ and 0 for $y \leq 0$ (the constant C_d is just there so the PDF integrates to 1).

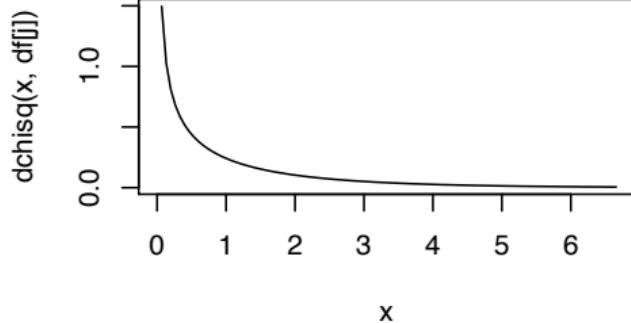
- The exponential distribution with mean 2 is also χ_2^2 , so when $d = 2$, $C_d = \frac{1}{2}$. Thus
 - ▶ if Z_1, Z_2 are indep $N(0, 1)$ the *mean square* $\frac{Z_1^2 + Z_2^2}{2}$ is exponential with mean 1;
 - ▶ a sum of n independent exponential random variables (all with the same mean) have a (possibly rescaled) χ_{2n}^2 distribution.

Properties of χ_d^2

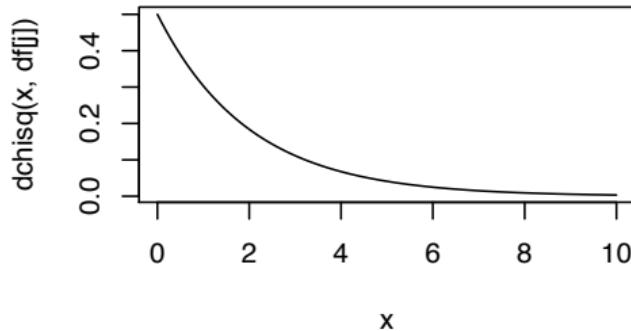
- If $X \sim \chi_d^2$ then $E(X) = d$ and $Var(X) = 2d$.
- Differentiating the χ_d^2 PDF shows us that the PDF increases up to $\frac{d}{2}$ and then decreases.
- In fact the shape is **right-skewed** (longer right tail than left) but it becomes *more symmetric* as d increases:

```
par(mfrow=c(2,2))
df=c(1,2,10,20)
up=df+4*sqrt(2*df)
for(j in 1:4){
  lab=paste("Chi-squared with",df[j],"df")
  curve(dchisq(x,df[j]),from=0,to=up[j],main=lab)
}
```

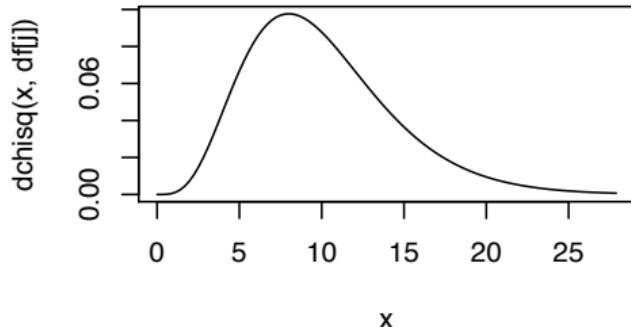
Chi-squared with 1 df



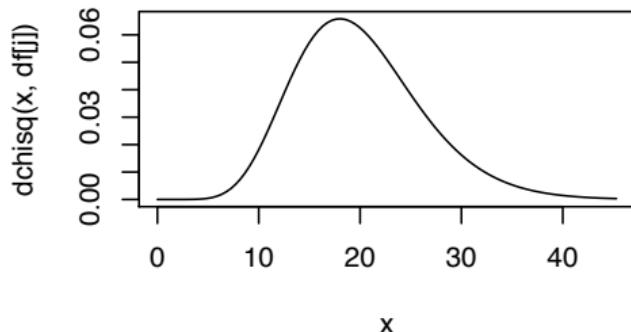
Chi-squared with 2 df



Chi-squared with 10 df



Chi-squared with 20 df



Joint Distribution of a Normal Sample Mean and Sample Variance

- Suppose X_1, X_2, \dots, X_n are independent $N(\mu, \sigma^2)$ random variables. Another way to say this is that they form a “random sample from a $N(\mu, \sigma^2)$ population”.
- We know already that
 - ▶ the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;
 - ▶ the sample variance $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ has $E(S^2) = \sigma^2$.
- Can we say more about S_X^2 ? The answer is yes.
- In fact it is known what
 - ▶ the full distribution of S_X^2 is, and more than this,
 - ▶ the **joint** distribution of S_X^2 and \bar{X} , in particular **they are independent**.
- The fact that they are independent is proved in third year, we shall just take it as given.

The $N(0, 1)$ case

- Let us start with Z_1, Z_2, \dots, Z_n independent $N(0, 1)$ (we will extend to general μ and σ^2 later).
- Write $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$ for their mean and $S_Z^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$ for their sample variance.
- Now we already know that
 - ▶ since $\bar{Z} \sim N(0, \frac{1}{n})$, $\sqrt{n}\bar{Z} \sim N(0, 1)$ so $n\bar{Z}^2 \sim \chi_1^2$;
 - ▶ $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$.

- The computing formula for the sum of squared deviations from the mean gives us that

$$\sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - n(\bar{Z})^2$$

that is

$$\underbrace{\sum_{i=1}^n Z_i^2}_{\sim \chi_n^2} = \underbrace{\sum_{i=1}^n (Z_i - \bar{Z})^2}_{\sim ???} + \underbrace{n(\bar{Z})^2}_{\sim \chi_1^2}.$$

- However, since a sum of independent χ^2 random variables is also χ^2 , if $\sum_{i=1}^n (Z_i - \bar{Z})^2 \sim \chi_{n-1}^2$ then the RHS would be χ_n^2 (by the independence of \bar{Z} and $\sum_{i=1}^n (Z_i - \bar{Z})^2 = (n-1)S_Z^2$).
- It turns out this is in fact *if and only if*.
- So then $S_Z^2 \sim \frac{1}{n-1} \chi_{n-1}^2$ *independently of* $\bar{Z} \sim N(0, \frac{1}{n})$.

The general $N(\mu, \sigma^2)$ case

- Suppose now that X_1, X_2, \dots, X_n are independent $N(\mu, \sigma^2)$ random variables.
- Note firstly that if we define $Z_i = (X_i - \mu)/\sigma$ as the standardised version of X_i then Z_1, Z_2, \dots, Z_n are independent $N(0, 1)$.
- Then $X_i = \mu + \sigma Z_i$ and $\bar{X} = \mu + \sigma \bar{Z}$ where $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$ so $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- Thus we can write the sample variance as

$$\begin{aligned} S_X^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n [\mu + \sigma Z_i - (\mu + \sigma \bar{Z})]^2 \\ &= \frac{\sigma^2}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2. \end{aligned}$$

- So $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ independently of $S_X^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$.

Pivots; the Studentised mean

- Suppose we are modelling data as a normal random sample that is as values taken by independent $N(\mu, \sigma^2)$ random variables X_1, X_2, \dots, X_n .
- Then as noted above, with $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ the “sample mean” and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ the “sample variance”, as noted above
 - ① $\bar{X} \sim N(\mu, \sigma^2)$ independently of
 - ② $S^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$.

From 1. we have already seen that the **standardised mean**

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).$$

- This is a **pivot**, that is a *function of the sample and the parameters whose distribution does not depend on the parameters.*

Student's *t*-ratio

- The statistician W.S. Gosset (who wrote under the pseudonym "Student") studied the similar quantity obtained by replacing the "population standard deviation" σ with its sample version $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

- What is the distribution of this T ?
- Note it can be written as

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} / \frac{S}{\sigma} = \frac{Z}{Y}$$

where $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$ independently of $Y = S/\sigma \sim \sqrt{\frac{\chi_{n-1}^2}{n-1}}$.

- Note that T is also a **pivot**, the distributions of Z and Y here are both free of μ and σ .

- “Student” derived the PDF of T . A double integral can be evaluated (2nd or 3rd year problem) to show that this PDF is of the form

$$f_T(t) = C_n \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

where the constant C_n is chosen so that the PDF integrates to 1.

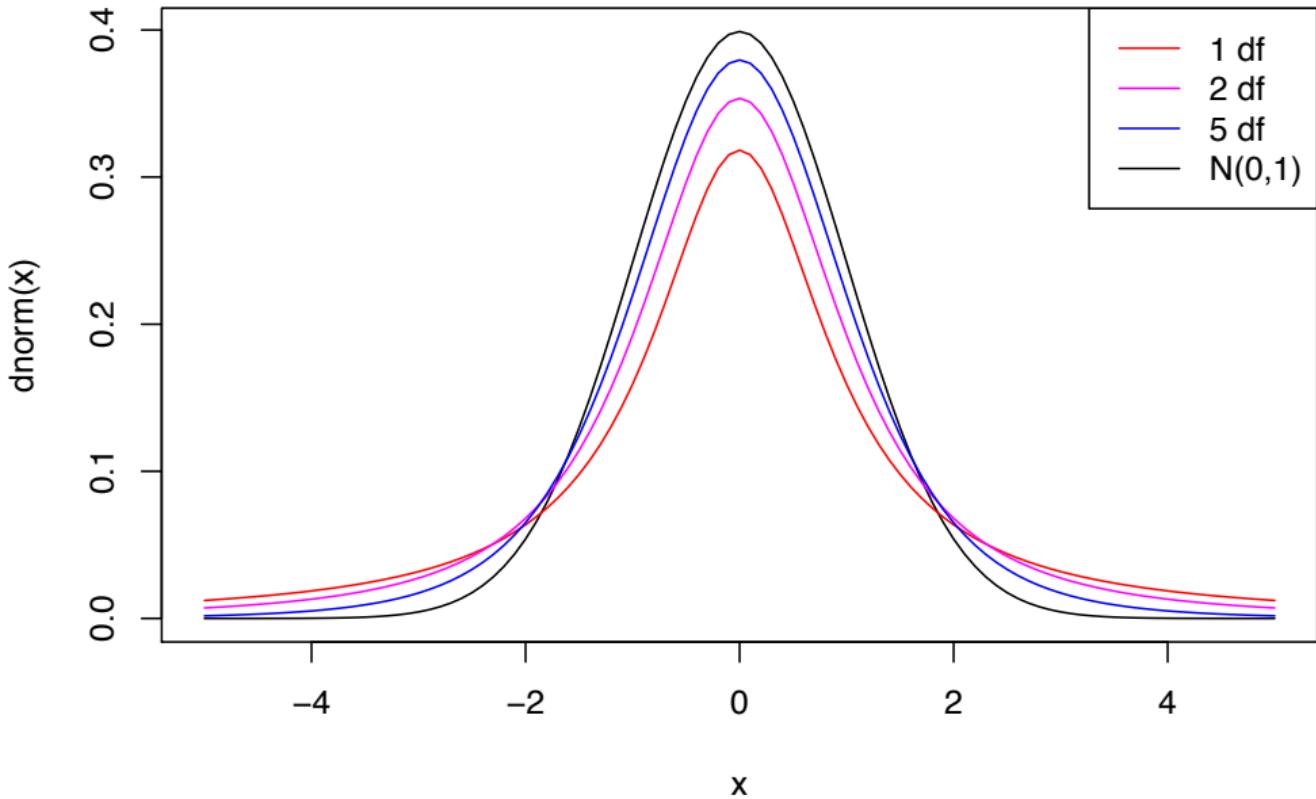
- This is called **Student's t -distribution with $(n - 1)$ degrees of freedom (df)**.
- We write $T \sim t_{n-1}$ for short.
- Also, the pivot T is referred to as the **Studentised mean** (as opposed to the *standardised mean*).
- The shape of Student's t distribution is similar to normal:
 - ▶ symmetric
 - ▶ bell-shaped

however it has *longer tails* than the normal, which makes sense.

- Replacing the fixed σ with a random version S adds “a little extra randomness” which thus increases the spread.

- Some examples are given below:

```
curve(dnorm(x),from=-5,to=5)
curve(dt(x,df=1),add=T,col="red")
curve(dt(x,df=2),add=T,col="magenta")
curve(dt(x,df=5),add=T,col="blue")
legend("topright",legend=c(paste(c(1,2,5),"df"),"N(0,1)" ),
      lty=1,col=c("red","magenta","blue","black"))
```



- Note that as the sample size/degrees of freedom increase,
 - ▶ the variance of S decreases;
 - ▶ the distribution of the “extra randomness” in the denominator gets more and more concentrated about 1;
 - ▶ the increase in spread between t_{n-1} and $N(0, 1)$ reduces.

Indeed it can be shown that as $n \rightarrow \infty$,

$$f_T(t) = \underbrace{C_n}_{\rightarrow e^{-\frac{1}{2}t^2}} \left(1 + \frac{t^2}{n-1}\right)^{-n/2} \xrightarrow{\substack{\rightarrow \frac{1}{\sqrt{2\pi}} \\ \text{ }} } \phi(t).$$

The Central Limit Theorem

- We noted above that for “large n ”, the $B(n, p)$ distribution is “approximately normal”.
- In fact this is a special case of a much more general and important result in statistical theory.
- The key observation in generalising this is that if $S_n \sim B(n, p)$ we can *represent* S_n as a *sum of independent $B(1, p)$ random variables* X_1, X_2, \dots, X_n :

$$S_n = X_1 + X_2 + \cdots + X_n .$$

- More generally if X_1, X_2, \dots, X_n are **independent and identically distributed** (iid) random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$ (remember it is possible for a random variable to have an infinite variance!), then if
 - $S_n = \sum_{i=1}^n X_i$ we have $E(S_n) = n\mu$ and $\text{Var}(S_n) = n\sigma^2$;
 - $\bar{X}_n = \frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^n X_i$ we have $E(\bar{X}) = \mu$ and $\text{Var}\left(\frac{\sigma^2}{n}\right)$ (as we already knew).
- However, the standardised version (of both),

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

has (by design)

- $E(Z_n) = 0$ and
- $\text{Var}(Z_n) = 1$.

Standardised sums approx $N(0, 1)$ for large n

- But most importantly, the *CDF* of Z_n has a special limiting property: for all real z ,

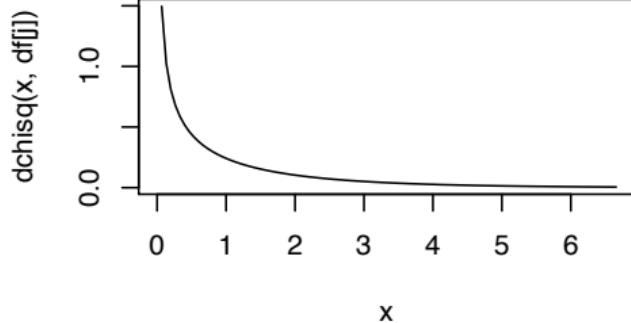
$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z) = P(Z \leq z) \text{ where } Z \sim N(0, 1).$$

- This is to say, as the sample size n increases, the distribution of the *standardised sum/mean* gets closer and closer to the $N(0, 1)$ distribution.
- This holds **even if the original X_i 's are not normal**. Indeed as we have (empirically) already seen, it even holds if the X_i 's are $B(1, p)$ (i.e. only taking values 0 and 1!).

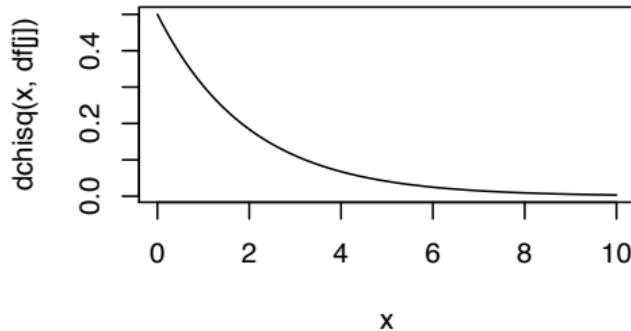
- We have also had a graphical glimpse of this effect when looking at χ^2 PDFs.
- Recall that the sum of n independent χ_1^2 's (with the same scale parameter!) is χ_n^2 . Hence if Z_1, Z_2, \dots, Z_n are independent $N(0, 1)$ then their sum $S_n = Z_1^2 + Z_2^2 + \dots + Z_n^2$ has a χ_n^2 distribution.
An extended version of the gamma PDF pictures from last lecture is given below:

```
par(mfrow=c(2,2))
df=c(1,2,10,20)
up=df+4*sqrt(2*df)
for(j in 1:4){
  lab=paste("Chi-squared with", df[j], "df")
  curve(dchisq(x, df[j]), from=0, to=up[j], main=lab)
}
```

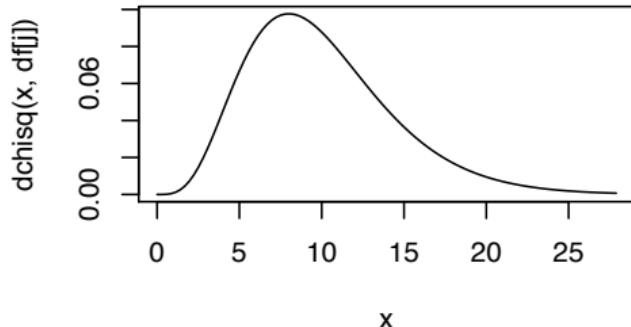
Chi-squared with 1 df



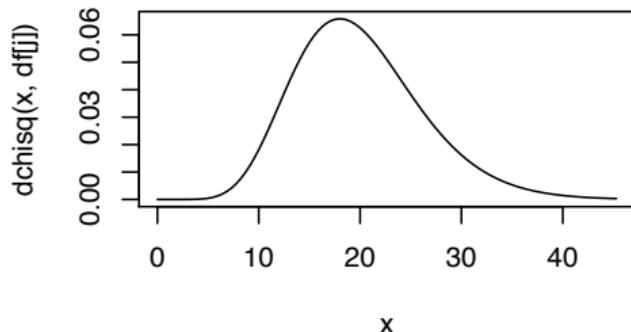
Chi-squared with 2 df



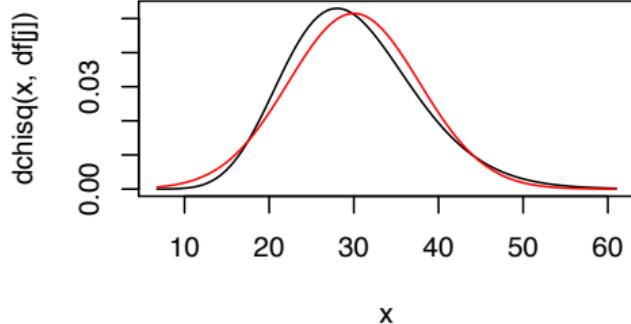
Chi-squared with 10 df



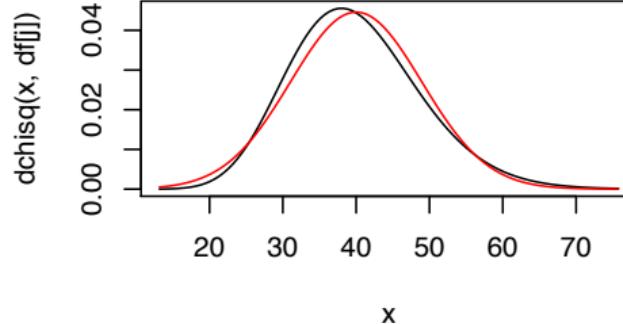
Chi-squared with 20 df



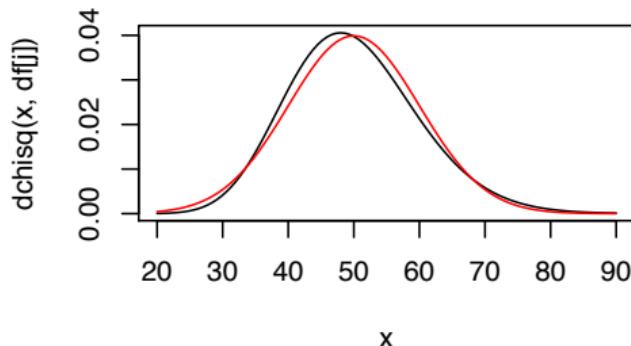
Chi-squared with 30 df



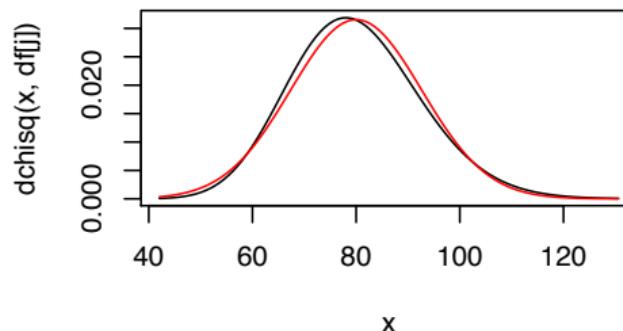
Chi-squared with 40 df



Chi-squared with 50 df

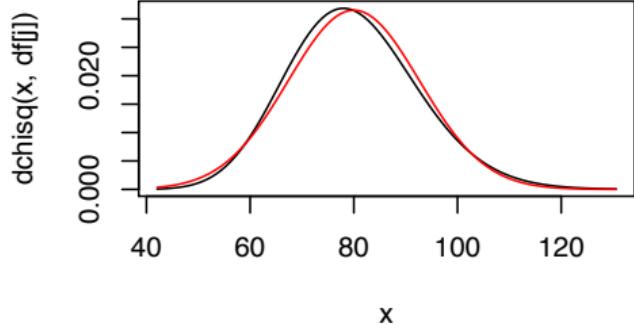


Chi-squared with 80 df

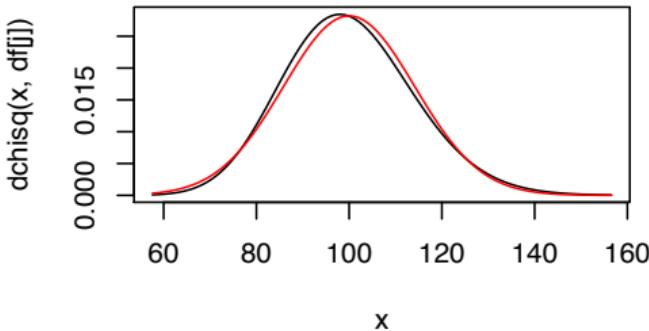


```
par(mfrow=c(2,2))
df=c(80,100,200,400)
up=df+4*sqrt(2*df)
lo=df-3*sqrt(2*df)
for(j in 1:4){
  lab=paste("Chi-squared with ", df[j], " df")
  curve(dchisq(x,df[j]), from=lo[j], to=up[j], main=lab)
  curve(dnorm(x,m=df[j],s=sqrt(2*df[j])), col="red", add=T)
}
```

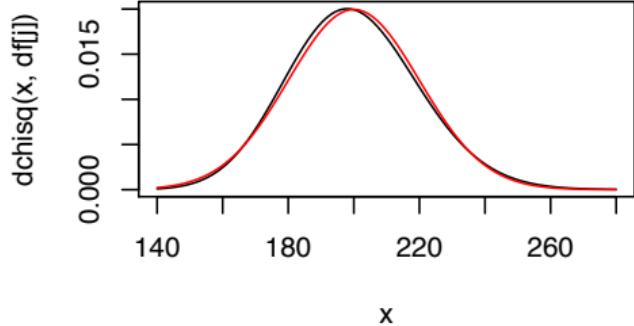
Chi-squared with 80 df



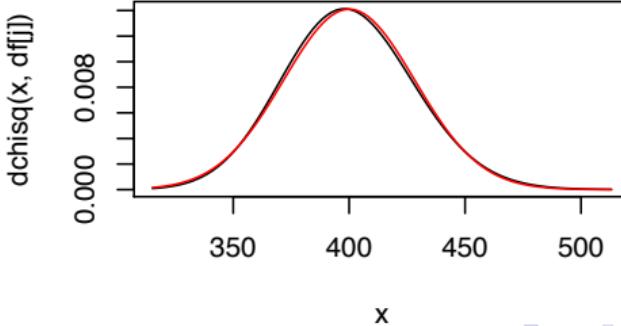
Chi-squared with 100 df



Chi-squared with 200 df



Chi-squared with 400 df



Practical Implications

- What is the use of this Central (i.e. “most important”) Limit Theorem?
- The practical use of *any* limiting result as $n \rightarrow \infty$ is that the limit on the “right-hand-side” may serve as a reasonably accurate approximation to the “left-hand-side” for moderate-to-large n , so that

$$P(Z_n \leq z) \approx \Phi(z)$$

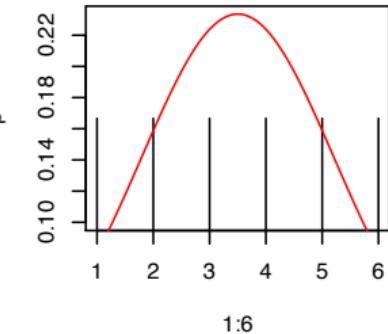
for n “moderate-to-large”.

- This comes down to “how fast is the convergence”?

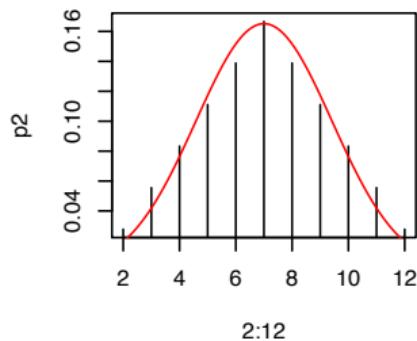
Example: sums of die rolls

- Let us first consider X_1, X_2, \dots, X_n distributed as the roll of a fair 6-sided die, so taking values 1, 2, ..., 6 with equal probabilities.
- The following graphs have ordinate diagrams of the (discrete) distributions of the sum of n (independent) 6-sided (fair) die rolls, for $n = 1, 2, \dots, 9$.
- Superimposed in red is the PDF of the corresponding normal distribution with the same expectation ($3.5n$) and variance ($35n/12$).

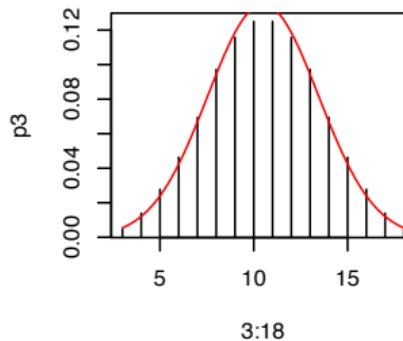
n=1



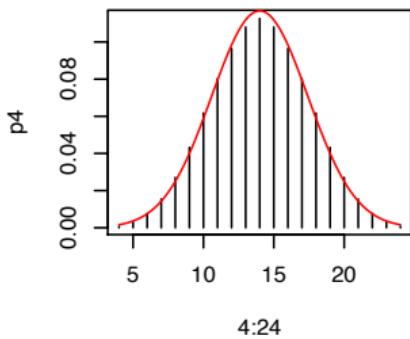
n=2



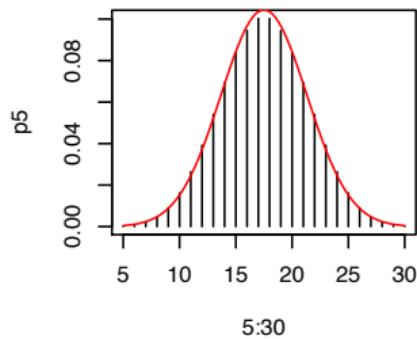
n=3



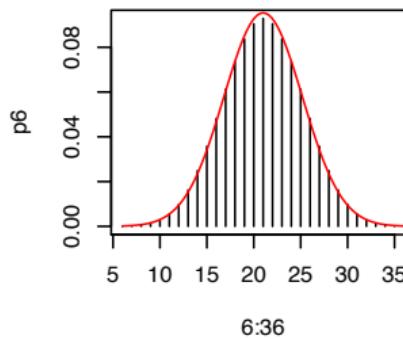
n=4



n=5



n=6



- We see that after n as small as 9 the distribution of the sum of n die rolls is very close to the normal distribution, at least in this "graphical" sense.
- To compare them more closely, consider the table below. We consider the case $n = 9$ and only compare up to $x = 30$ since the distribution is symmetric and $P(X \leq 31) = P(Y \leq 31) = 0.5$ ($x = 31.5$ is the middle of the distribution).
- The columns are
 - ▶ x ;
 - ▶ $\text{cdf.9: } P(X \leq x)$;
 - ▶ na.cc: the normal approximation to $P(X \leq x)$ with continuity correction, given by $P(Y \leq x + \frac{1}{2})$ where $Y \sim N(9 \times 3.5, 9 \times \frac{35}{12})$;
 - ▶ err: the absolute error $|P(X \leq x) - P(Y \leq x + \frac{1}{2})|$;
- The errors in the last column are pretty small, and are for the most part small even compared to the quantity they are approximating (all except in the extreme tails).

	x	cdf .9	na.cc	err
[1,]	9	0.00000009922903	0.000008775693	0.000008676464
[2,]	10	0.00000099229030	0.000020766651	0.000019774361
[3,]	11	0.00000545759666	0.000047386126	0.000041928529
[4,]	12	0.00002183038663	0.000104278150	0.000082447764
[5,]	13	0.00007094875654	0.000221338493	0.000150389737
[6,]	14	0.00019865651832	0.000453224652	0.000254568133
[7,]	15	0.00049574823452	0.000895453042	0.000399704807
[8,]	16	0.00112624949195	0.001707395589	0.000581146097
[9,]	17	0.00236313935249	0.003142590822	0.000779451469
[10,]	18	0.00462804196515	0.005584917655	0.000956875690
[11,]	19	0.00852804053625	0.009586242378	0.001058201841
[12,]	20	0.01487859923538	0.015897430552	0.001018831317
[13,]	21	0.02470018940837	0.025480968484	0.000780779076
[14,]	22	0.03917760567495	0.039491289632	0.000313683957
[15,]	23	0.05957512510796	0.059209971354	0.000365153754
[16,]	24	0.08710760872326	0.085928669531	0.001178939192
[17,]	25	0.12277736895417	0.120783293484	0.001994075470
[18,]	26	0.16719912964233	0.164556992989	0.002642136653
[19,]	27	0.22044235110882	0.217483580307	0.002958770802
[20,]	28	0.28192158207590	0.279092324711	0.002829257365
[21,]	29	0.35036133259031	0.348135170057	0.002226162533
[22,]	30	0.42385223765432	0.422626012110	0.001226225544

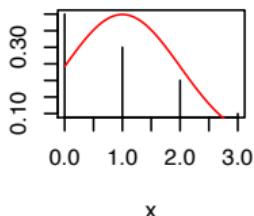
Example: sums of asymmetric discrete random variables

- Let us consider one more discrete example, this time where the distribution of the X_i 's is *not* symmetric:

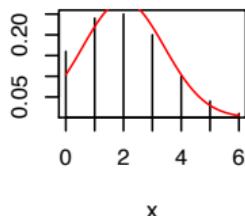
x	0	1	2	3
$P(X_i = x)$	0.4	0.3	0.2	0.1

- We see that a larger n is required before the sampling distribution of the sum appears (roughly) normal:

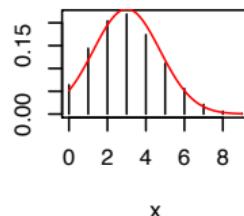
$n=1$



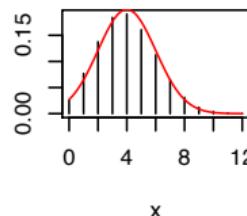
$n=2$



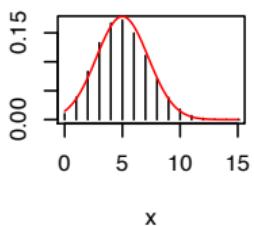
$n=3$



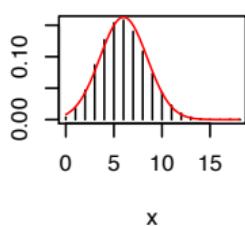
$n=4$



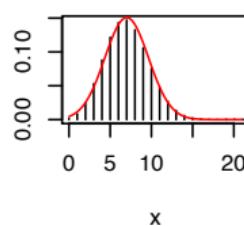
$n=5$



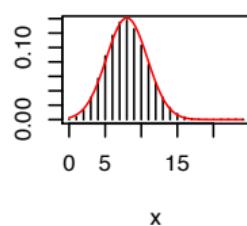
$n=6$



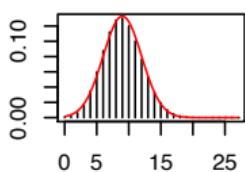
$n=7$



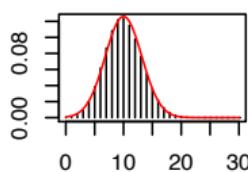
$n=8$



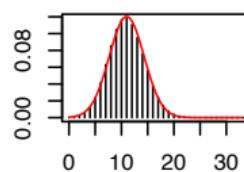
$n=9$



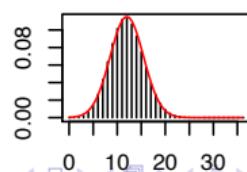
$n=10$



$n=11$



$n=12$



How large does n have to be?

- These examples perhaps beg the question: *How large does n have to be for the normal approximation to be “good enough”?*
- These examples perhaps indicate that there is **no simple answer to this question** (the Phipps & Quine book uses the rule of thumb “so long as $n \geq 25$ ” but this is overly simplistic).
- The real answer is “**it depends**”.
- It depends on how **non-normal** the distribution of the X_i ’s being added is.
 - ▶ For very skewed distributions like the chi-squared, $n > 100$ is required before the sampling distribution of the sum is well approximated by a normal distribution.
 - ▶ For skewed distributions which have shorter tails (outliers) $n = 15$ may be enough (as in the last example above)
 - ▶ For symmetric, short tailed distributions $n = 5$ or 6 may be enough (like die rolls).
 - ▶ Finally, if the X_i ’s actually have a normal distribution, $n = 1$ is enough!!

Probability Wrap-up

- We are now ready to proceed to the “business end” of the course, **statistical inference**.
- Before we do, we recap some key results which will be used heavily in what follows.
- Suppose X_1, X_2, \dots, X_n constitute a random sample from a population with mean μ and variance σ^2 . This is to say, these are **independent and identically distributed (iid)** random variables with
 - ▶ common distribution (CDF) $F(\cdot)$,
 - ▶ $E(X_1) = \mu$ and
 - ▶ $Var(X_1) = \sigma^2$

(since the X_i 's all have the same distribution, whatever applies to X_1 applies to all of them).

- Then with $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ their average, we can say

- ① $E(\bar{X}) = \mu$.
- ② $Var(\bar{X}) = \frac{\sigma^2}{n}$.
- ③ If the X_i 's are **normal** then \bar{X} is also normal, equivalently the standardised version

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Also the Studentised version

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Here $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ denotes the (random) sample variance and t_{n-1} denotes a random variable with Student's t -distribution with $n - 1$ degrees of freedom.

- If the X_i 's are **not normal** but n is "large" then \bar{X} is *approximately* normal; equivalently the standardised version

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{\text{approx}}{\sim} N(0, 1).$$

We cannot say anything about the Studentised mean T_n in this case.

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
 - Lecture 15
 - Statistical Inference
 - Inference Concerning a Population Mean when the Population Variance is Known
 - Interpretation of p-values
 - An alternative formulation
 - Lecture 16
 - Lecture 17
 - Lecture 18
- 5 Inference Part 2: Discrete models

Probability Wrap-up

- We are now ready to proceed to the “business end” of the course, **statistical inference**.
- Before we do, we recap some key results which will be used heavily in what follows.
- Suppose X_1, X_2, \dots, X_n constitute a random sample from a population with mean μ and variance σ^2 .
- This is to say, these are **independent and identically distributed** (iid) random variables with common distribution/CDF $F(\cdot)$, $E(X_1) = \mu$ and $\text{Var}(X_1) = \sigma^2$ (since the X_i 's all have the same distribution, whatever applies to X_1 applies to all of them).

- Then with $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ their average, we can say

- ① $E(\bar{X}) = \mu$.
- ② $Var(\bar{X}) = \frac{\sigma^2}{n}$.
- ③ If the X_i 's are **normal** then \bar{X} is also normal, equivalently the standardised version

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Also the Studentised version

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Here $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ denotes the (random) sample variance and t_{n-1} denotes a random variable with Student's t -distribution with $n - 1$ degrees of freedom.

- If the X_i 's are **not normal** but n is "large" then \bar{X} is *approximately* normal; equivalently the standardised version

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{\text{approx}}{\sim} N(0, 1).$$

We cannot say anything about the Studentised mean T_n in this case.

- The general problem of statistical inference can be summarised as follows:
 - ▶ we wish to “learn something” about a certain *population*
 - ▶ it is not possible/feasible to observe the “whole population”
 - ▶ therefore, we obtain a *representative sample* from the population and then try to learn something of the population *based on what we see in the sample*
- Of course this last step is **imperfect** in that whatever statements we might make about the population based on what we see in the sample will involve making some kind of error; without seeing the whole population we can never be completely certain of anything we might say about it.
- However if we employ appropriate probabilistic and statistical reasoning, we can **quantify in a very specific way** this inherent uncertainty. In this way we can guard against making “overly confident” statements; we can hopefully convey in a clear way exactly how certain or uncertain we might be.

Inference Concerning a Population Mean when the Population Variance is Known

- Suppose we have data x_1, x_2, \dots, x_n which we are *modelling* as values taken by independent and identically distributed (iid) random variables whose common expectation is μ which is *unknown* and whose common variance is σ_0^2 which is *known*.
- The aim is to make some inference about μ . Exactly what sort of inference depends on the context.

Motivating Examples

- We shall consider two motivating examples
 - ① birthweights of babies whose mothers smoke
 - ② marks on a draft standardised test

We shall see that each of these is treated differently in light of the natural “scientific question” that arises.

Birthweights of babies of mothers who smoke

- It is assumed that the birthweights (in kg) of full-term babies follow a $N(3.4, 0.5^2)$ distribution.
- It is believed that smoking during pregnancy reduces birthweight on average.
- Suppose that the birthweights of 14 babies whose mothers smoked heavily during pregnancy are given below:

```
x
```

```
[1] 2.5 2.9 2.8 3.1 3.4 3.4 3.5 2.8 3.3 2.8 3.7  
[12] 2.3 3.4 3.0
```

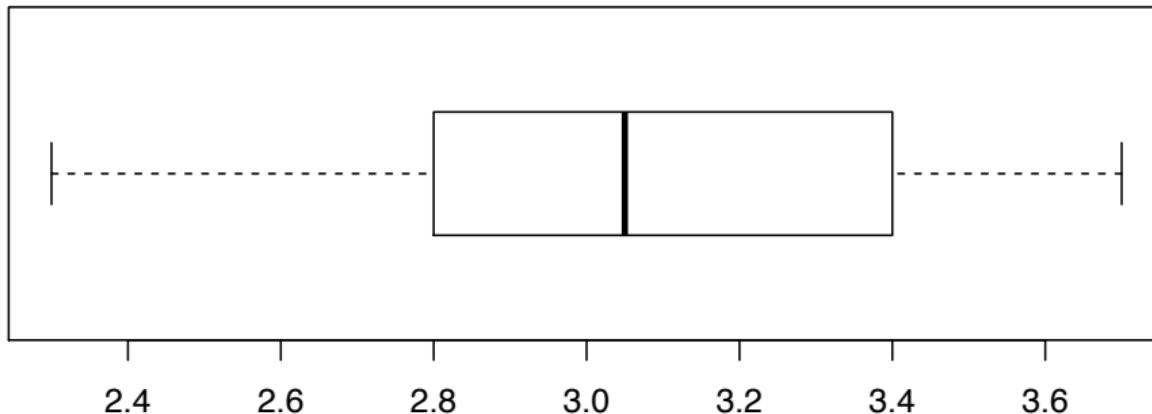
```
mean(x)
```

```
[1] 3.064286
```

```
sd(x)
```

```
[1] 0.4049827
```

```
boxplot(x, horizontal=T)
```



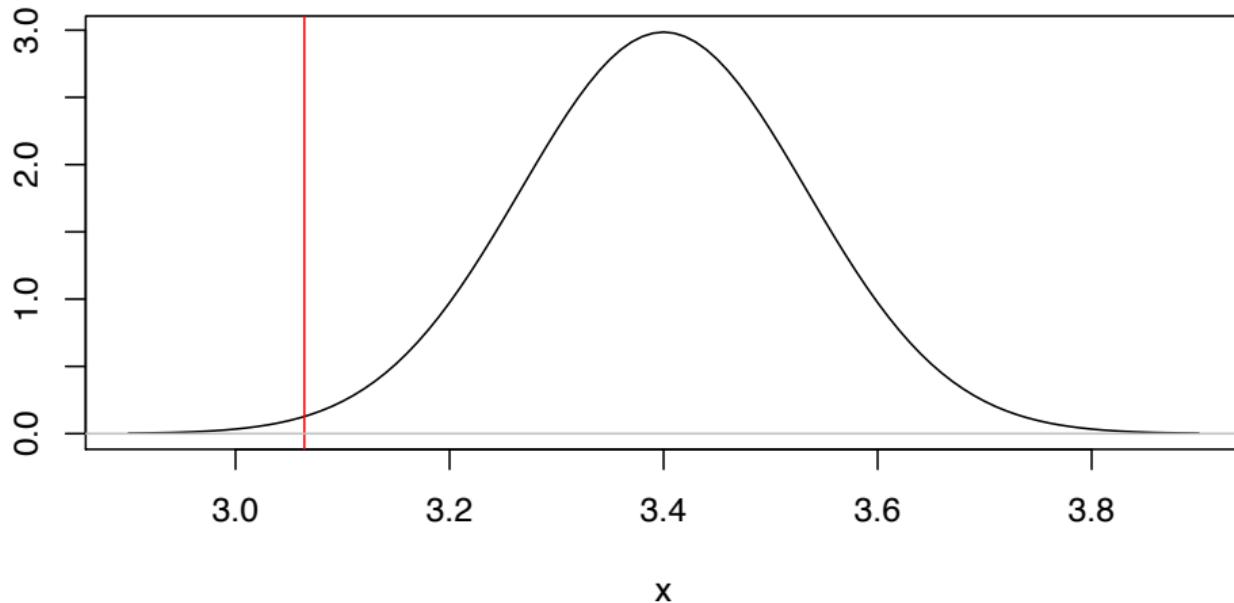
The average of these weights is *less* than 3.4 but the *real* question is: is this **significantly less** than 3.4?

Assuming No Difference

- To address this, let us first deduce how the sample mean would behave as a *random variable* if the data were indeed values taken by iid $N(3.4, 0.5^2)$ random variables X_1, X_2, \dots, X_{14} , that is if there was **no difference** between birthweights for babies of mothers who smoked and the general population (which have been assumed to be $N(3.4, 0.5^2)$).
- In that case $\bar{X} = \frac{1}{14} \sum_{i=1}^n X_i \sim N\left(3.4, \frac{\sigma_0^2}{n} = \frac{0.5^2}{14} \approx 0.134^2\right)$.
- The plot below shows the PDF the random sample mean would have if the birthweights for babies of mothers who smoked were indeed $N(3.4, 0.5^2)$; it is the $N\left(3.4, \frac{\sigma_0^2}{n} = \frac{0.5^2}{14} \approx 0.134^2\right)$ PDF.

```
curve(dnorm(x,m=3.4,s=.5/sqrt(14)),from=2.9,to=3.9,
      main='Distribution of sample mean under assumption of "no change"\n(observed value in red)',
      ylab="")
abline(v=mean(x),col="red")
abline(h=0,col="grey")
```

**Distribution of sample mean under assumption of "no change"
(observed value in red)**



- The observed value is rather extreme for the “no difference” assumption; it appears to be a bit “unusually small” for this assumption.
- This perhaps casts doubt on the assumption of no difference, which in turn perhaps constitutes some evidence that indeed the mean birthweight of babies of mothers who smoke is less than the assumed value of 3.4.
- The methods we shall develop permit us to convert this comment into a clearer quantitative statement, in various ways.

Marks on draft standardised test

- A standard test is prepared each year and is carefully adjusted so that the distribution of marks has mean 60 and standard deviation 10.
- The draft version of the test is tried out on a smaller number of candidates to see if the distribution is about right.
- In the past, the standard deviation of the marks on the draft test has been around 10, although the mean of the marks can vary from 60 if the draft test is too easy or too hard.
- Suppose that 50 candidates sit the draft test and the marks are as follows:

```
matrix(y, 5, 10)
```

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 62   54   54   41   65   71   63   72   59   49
[2,] 59   62   68   84   47   62   69   72   56   74
[3,] 78   49   72   60   68   46   56   62   74   61
[4,] 47   65   67   68   66   44   70   67   67   71
[5,] 58   66   89   70   55   62   43   63   51   57
```

```
mean(y)
```

```
[1] 62.3
```

```
sd(y)
```

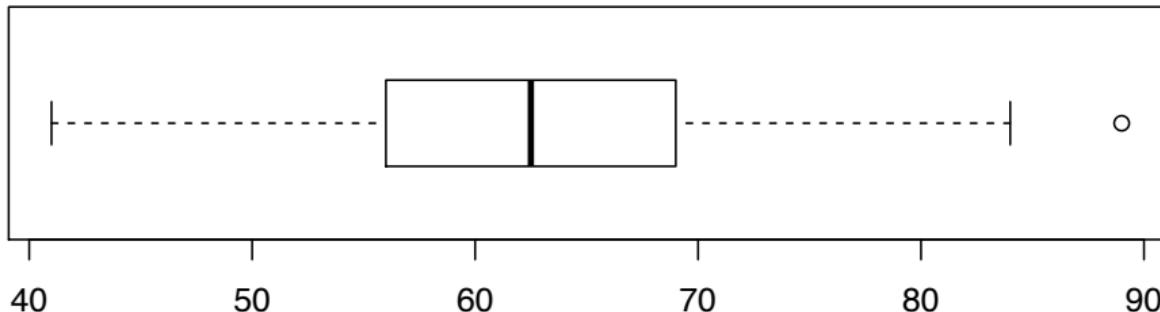
```
[1] 10.33174
```

```
stem(y)
```

The decimal point is 1 digit(s) to the right of the |

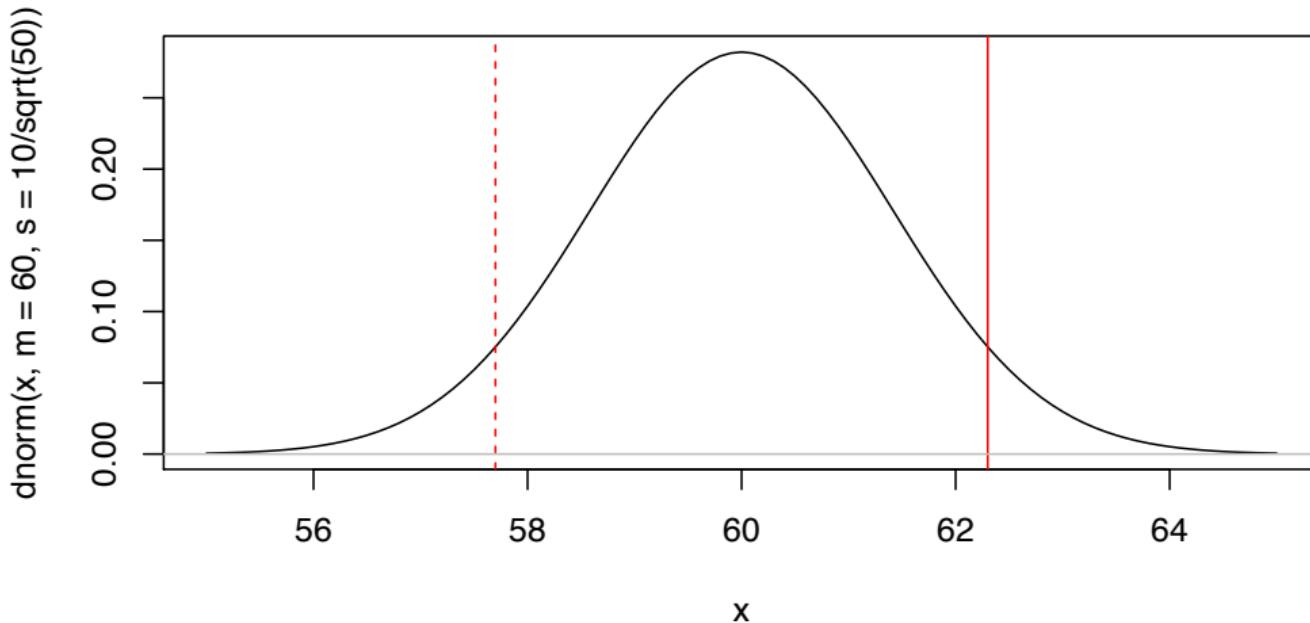
4 134
4 67799
5 144
5 5667899
6 012222233
6 55667778889
7 001122244
7 8
8 4
8 9

```
boxplot(y, horizontal=T)
```



- Now the average mark is different from 60, but the real question is: “Is it **significantly different** from 60?”
- On the following plot we have marked the observed average mark with a solid red vertical line, but we have also added a red dashed line equidistant from 60 on the other side, for reasons which will become apparent later.

Approximate PDF of average mark under mu=60, sigma=10



Treating these as Detection problems

- In both cases we can interpret the data analyses as *detection problems*:
 - ▶ for the birthweight data we are trying to *detect* if there is any **reduction** in birthweight from the “normal” mean of 3.4
 - ▶ for the test scores data we are trying to *detect* if the long-run average mark is **different** to 60.

Birthweights

- Imagine designing, **before the data were collected**, a procedure that would “detect” if a drop in birthweights had occurred.
- The design process might look like the following:
 - ① pick a “significance level” or “false alarm rate” $0 < \alpha < 1$, e.g. $\alpha = 0.05$.
 - ② choose a corresponding “critical value” $c(\alpha) < 3.4$ such that *if the observed sample mean is less than $c(\alpha)$ we shall declare that a reduction in average birthweights has been “detected”*.
- How is $c(\alpha)$ chosen exactly?
 - ▶ We know that if we choose $c(\alpha)$ too close to 3.4 we could easily get a “low” \bar{x} even if the true birthweights are indeed $N(3.4, 0.5^2)$, just by chance alone.
 - ▶ But if we choose it too far below 3.4, we may never get a detection unless the actual reduction is “huge”.

- We choose $c(\alpha)$ so that the *probability of a false alarm* is α .
- That is,

$$P(\bar{X} \leq c(\alpha)) = \alpha$$

with the probability computed under the assumption **there is no real reduction**.

- According to the $N(0, 1)$ table, the value that cuts off 5% in the lower tail is

```
qnorm(0.05)
```

```
[1] -1.644854
```

- More generally, we need to go 1.645 standard deviations below the mean.

- In this example, if there is no reduction then $\bar{X} \sim N(3.4, 0.134^2)$ and so the critical value for $\alpha = 0.05$ is

$$c(0.05) = 3.4 - (1.645 \times 0.134) \approx 3.180.$$

This value is 1.645 standard deviations below the mean.

- If we want a smaller false alarm rate we would choose, say, $\alpha = 0.01$

```
qnorm(0.01)
```

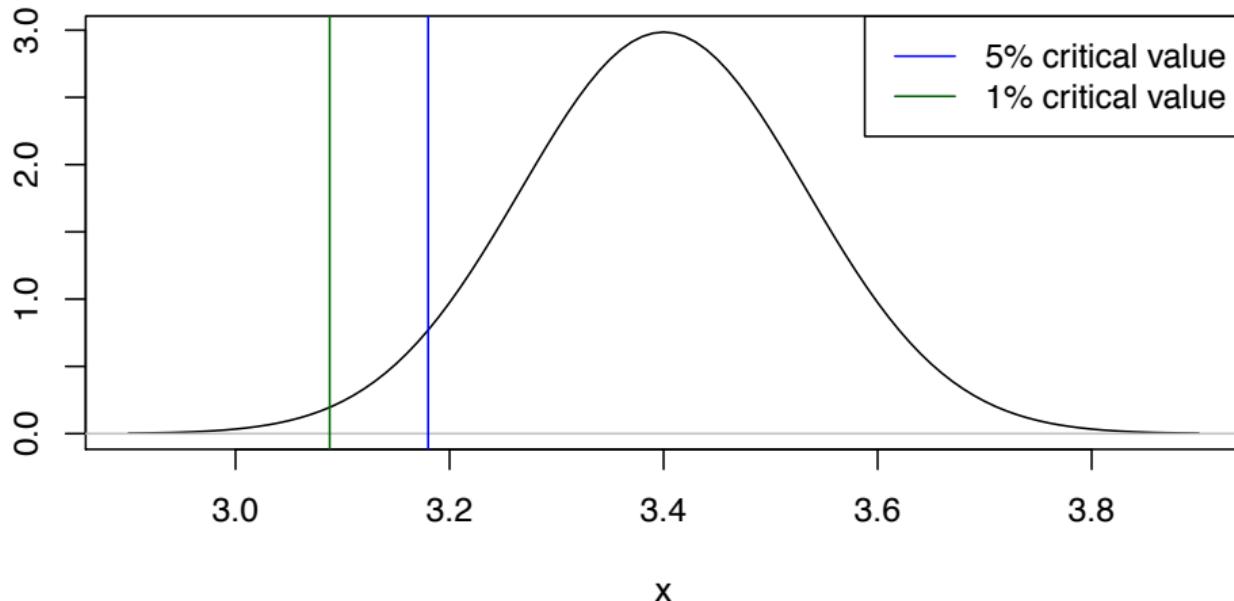
```
[1] -2.326348
```

in which case since the lower 1% point of any normal distribution is 2.326 standard deviations below the mean,

$$c(0.01) = 3.4 - (2.326 \times 0.134) \approx 3.088.$$

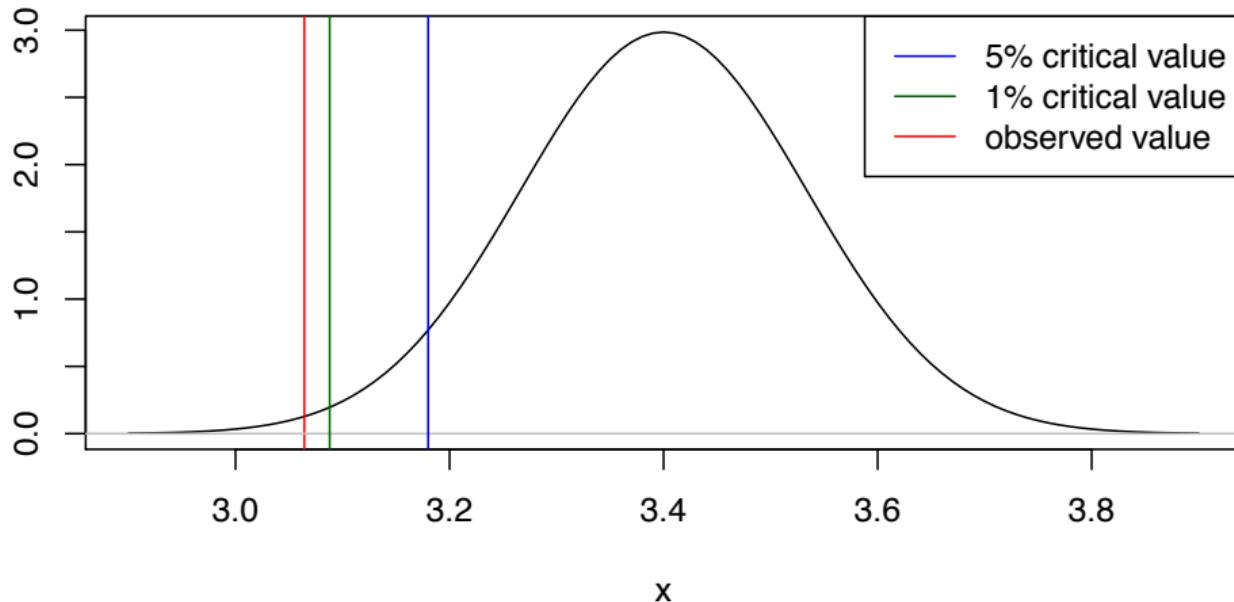
- We visualise these on the plot below:

Distribution of sample mean under assumption of "no change" (with 5% and 1% critical values)



- Now we are ready to look at the data (again, although imagine we hadn't seen it yet!).
- The observed sample mean is in fact 3.064286 which would be declared significant at both the 5% level and the 1% level since it is smaller than both the 5% and 1% critical values.

Distribution of sample mean under assumption of "no change" (with 5% and 1% critical values, observed value in red)



- We can describe *precisely* how significant it is by quoting the corresponding **observed significance level** which is that value of α so that the observation is *exactly* equal to the corresponding critical value $c(\alpha)$.
- In this case, we determine this simply by computing

$$P(\bar{X} \leq 3.064) \text{ when } \bar{X} \sim N(3.4, 0.134^2)$$

which is

$$P\left(\frac{\bar{X} - 3.4}{0.134} \leq \frac{3.064 - 3.4}{0.134}\right) = P(Z \leq -2.51)$$

where $Z = \frac{\bar{X}-3.4}{0.134} \sim N(0, 1)$.

- We can interpret this number -2.51 as meaning that the observed value was 2.51 standard deviations below the mean, assuming *no reduction*.
- This probability is

```
pnorm(-2.51)
```

[1] 0.006036558

which is (as we would expect) slightly smaller than 1%.

- This is the significance level α that would have put our observation exactly equal to the critical value $c(\alpha)$.
- The observed significance level is also called the **p-value**.
- This is also equal to the lower tail area below the PDF to the left of the red line on the plots above.

Marks on draft standardised test

Now imagine designing a procedure for detecting if the average mark of the draft test is *different* to 60:

- ① Pick a significance level/false alarm rate $0 < \alpha < 1$ e.g. $\alpha = 0.05$;
- ② Choose a critical value $c(\alpha)$ such that if the observed mean mark \bar{x} is such that $|\bar{x} - 60| > c(\alpha)$, i.e. if
 - ▶ $\bar{x} > 60 + c(\alpha)$ or
 - ▶ $\bar{x} < 60 - c(\alpha)$

then we shall declare that a difference in the (population) average mark from 60 has been “detected”.

- Note the difference: a discrepancy above or below 60 is potentially of interest.
- Furthermore, a discrepancy in either direction *of the same size* is regarded as equally significant.
- This is due to the symmetry of the distribution under the assumption of “no change”.
- How do we choose $c(\alpha)$ in this example?
- Well we declare a difference detected if the event $\{|\bar{X} - 60| > c(\alpha)\}$ occurs.
- We want to make sure this is equal to α when there is no real difference, that is we need $c(\alpha)$ to satisfy

$$P(|\bar{X} - 60| > c(\alpha)) = \alpha$$

when $E(\bar{X}) = 60$.

- Note we have made no assumption here about the precise distribution of the X_i 's.
- However by the Central Limit Theorem we can assume that \bar{X} is at least approximately normal.
- Indeed by looking at the boxplot we can see the population is nice and symmetric without long tails/outliers so we can confident that $n = 50$ is “large enough” for the normal approximation to the distribution of \bar{X} to be good.
- So let us assume that *when there is no difference*,

$$\bar{X} \stackrel{\text{approx}}{\sim} N\left(60, \frac{10^2}{50} = 2\right).$$

- Then since $(\bar{X} - 60)/\sqrt{2} \stackrel{\text{approx}}{\sim} N(0, 1)$ the critical value needs to satisfy

$$\alpha = P\left(\frac{|\bar{X} - 60|}{\sqrt{2}} > \frac{c(\alpha)}{\sqrt{2}}\right) \approx P(|Z| > c(\alpha)/\sqrt{2})$$

where $Z \stackrel{\text{approx}}{\sim} N(0, 1)$.

- If $\alpha = 0.05$ then note that

$$P(|Z| \leq 1.96) = P(-1.96 \leq Z \leq 1.96) = 0.95$$

(to see this, note the following:)

```
pnorm(1.96) - pnorm(-1.96)
```

[1] 0.9500042

and so

$$P(|Z| > 1.96) = 0.05.$$

- More generally, we need the critical region to be ± 1.96 standard deviations away from the mean. Thus we must have

$$\frac{c(0.05)}{\sqrt{2}} = 1.96$$

i.e.

$$c(0.05) = 1.96 \times \sqrt{2} \approx 2.77.$$

- Thus with a significance level/false-alarm rate of 5% in this problem, we declare “a difference detected” if the observed value \bar{x} lies *outside* the interval

$$60 \pm 2.77 = [57.23, 62.77].$$

- For a significance level/false-alarm rate of 1%, since

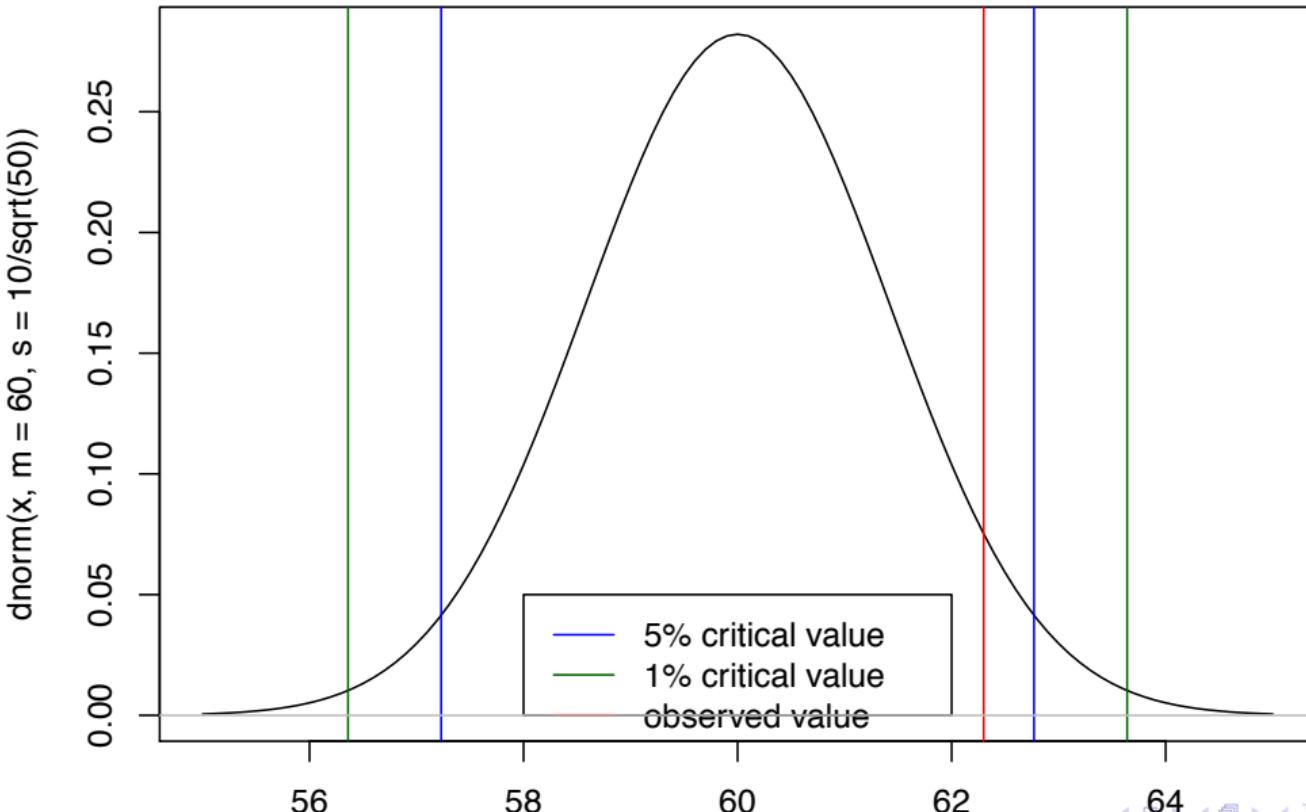
```
pnorm(2.576) - pnorm(-2.576)
```

[1] 0.9900049

the critical value for 1% is $2.576 \times \sqrt{2} \approx 3.64$ and thus we declare “a difference is detected at the 1% level of significance” if the observed value \bar{x} is *outside* the interval

$$60 \pm 3.64 = [56.36, 63.64].$$

Approximate PDF of average mark under mu=60, sigma=10 (5% and 1% critical regions indicated)



- Indeed we compute the *observed significance level* by finding that value of α so that the observed value 62.3 is right on the boundary of the critical region, that is we evaluate the probability

$$\begin{aligned} P(|\bar{X} - 60| > |62.3 - 60|) \\ &= P\left(\frac{|\bar{X} - 60|}{\sqrt{2}} > \frac{|62.3 - 60|}{\sqrt{2}} \approx 1.6263\right) \\ &\approx P(|Z| > 1.6263) \end{aligned}$$

where $Z \sim N(0, 1)$.

- To evaluate this last probability, note that it can be written as

$$1 - P(|Z| \leq 1.6263) = 1 - P(-1.6263 \leq Z \leq 1.6263)$$

which is given by

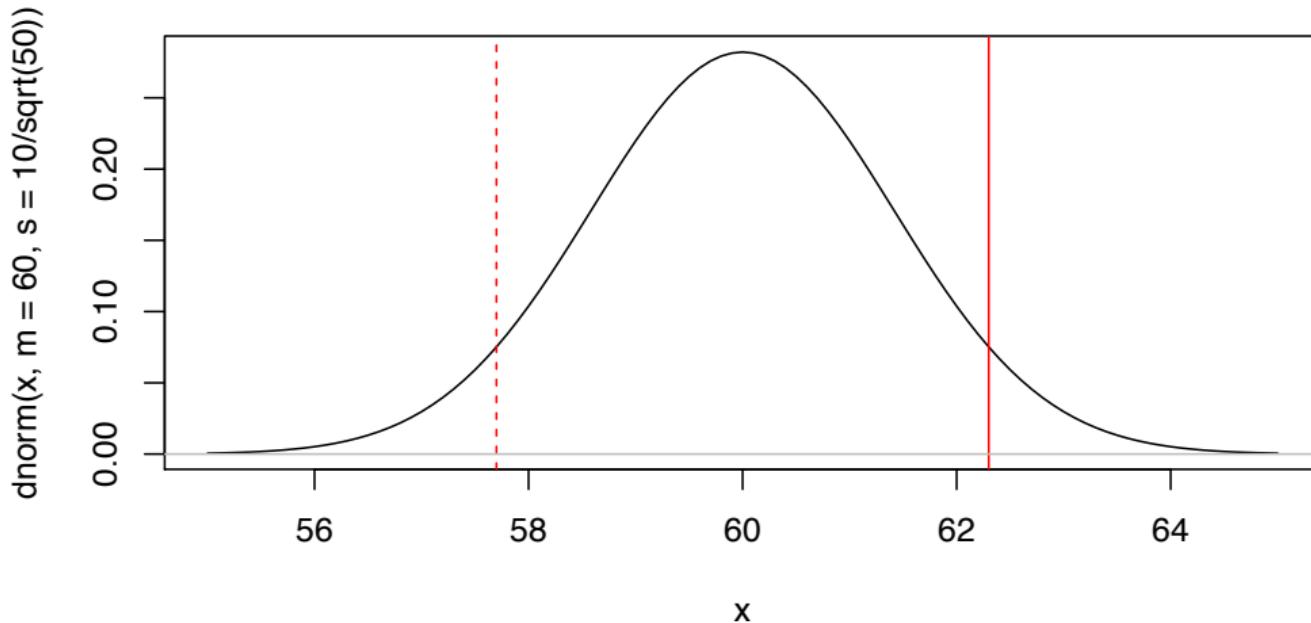
```
1 - (pnorm(1.6263) - pnorm(-1.6263))
```

[1] 0.1038859

a little over 10%.

- We knew it would be bigger than 5% because the result was not significant at the 5% level.
- This observed significance level or p-value is also exactly the sum of the two tail areas on either side of 60 beyond the red lines in the original plot above, which we reproduce below:

Approximate PDF of average mark under mu=60, sigma=10



Interpretation of p-values

- In both of the examples above we had a scientific question asking if some apparent effect was significantly above/below/different to some reference value.
- The observed significance level or p-value is a measure of this significance. It gives **the probability of at least as significant a result purely by chance**.
- The **smaller** it is, the stronger the evidence against the **assumed hypothesis**.

Null Hypothesis

- In each case we have an underlying **null hypothesis**. Here “null” means zero, or more loosely “no effect”.
 - ▶ In the birthweights example, the null hypothesis is H_0 : “birthweights of babies whose mothers smoke $\sim N(3.4, 0.5^2)$ ”
 - ▶ In the marks on the draft test example the null hypothesis is H_0 : long-run average mark is 60 (and sd is 10)

Test Statistic

- In each case we could identify a *test statistic* whose distribution is known if the corresponding null hypothesis is true:
 - ▶ In the birthweights example, if H_0 is true then the sample mean $\bar{X} \sim N\left(3.4, \frac{0.5^2}{14} \approx 0.134^2\right)$. Equivalently the Z -statistic

$$Z = \frac{\bar{X} - 3.4}{0.134} \sim N(0, 1)$$

if H_0 is true, in which case this is the standardised version of \bar{X} .

- ▶ In the draft test marks, if H_0 is true then (by the Central Limit Theorem) $\bar{X} \stackrel{\text{approx}}{\sim} N(60, 2)$ and so the Z -statistic

$$Z = \frac{\bar{X} - 60}{\sqrt{2}} \sim N(0, 1)$$

if H_0 true.

Converting observed value of test statistic into a p-value

- The p-value can be phrased as

*The probability of at least as much evidence against H_0 as was observed **under the assumption H_0 is true**.*

- In the birthweights example, because only a reduction was anticipated, *smaller* the observed value of the Z-statistic, the *more* evidence of H_0 this constitutes. The observed value of the Z-statistic is -2.51. Thus the event “at least as much evidence as observed” is

$$\{Z \leq -2.51\}.$$

The probability of this is the p-value.

- In the draft test marks example, because a difference above 60 or below 60 is equally significant, we have a **two-sided** test. The observed value of the Z-statistic is 1.6263. The event “at least as much evidence as was observed” is the event

$$\{|Z| > 1.6263\} = \{Z > 1.6263\} \cup \{Z < -1.6263\}$$

and the probability of this is the p-value.

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
 - Lecture 15
 - Lecture 16
 - Broader Inferences
 - Inference based on Student's t -distribution
 - Lecture 17
 - Lecture 18
- 5 Inference Part 2: Discrete models

A Statistical Model

- In the above setup, we are simply trying to measure the strength of evidence against a single hypothesis.
- The only assumptions we make are about what happens if the hypothesis is true.
- We make **no** assumptions about what might be happening if the hypothesis is false, other than vague statements about whether a test statistic might tend to take larger and/or smaller values.
- However, we may wish to make a more directed statement about the mechanism generating the data, other than “it is like H_0 or not like H_0 ”.
- To do so we need a broader model that not only describes the “population” when H_0 holds, but also when H_0 doesn’t hold.

Birthweights

- Let us model the data as values taken by independent $N(\mu, 0.5^2)$ random variables.
- So we are allowing this population of interest “birthweights of babies whose mothers smoke” to possibly have a different mean μ to 3.4 (that of the general population), but we are also assuming **there is no change in the variance**, and that the population is also normal.
 - This may or may not be reasonable; we shall discuss this further later on, let us stick with it for now.
- We thus interpret μ as the “population mean” birthweight, in the population of birthweights of babies whose mothers smoke.
- Any **inference** about the population then reduces to inference concerning the unknown μ .

Estimating μ

- Whatever the value of μ , we do know that *as a random variable* the sample mean

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} = \frac{0.5^2}{14} \approx 0.134^2\right).$$

- So we know that whatever value \bar{X} might take, it should not be “too far” from the true μ , in some sense.
- Thus the observed value \bar{x} of the sample mean can be used as an *estimate* of μ , in this case 3.064286.
- However we know that it is not *exactly* equal to μ , but that we have made some error in using the observed value \bar{x} as an estimate of μ .
- *An estimate on its own is useless without some idea of the size of the error we have made in using it.*
- One way to do this to quote a **standard error**.

Standard Error

- The estimate $\bar{x} = 3.064286$ is the observed value of the random variable or **estimator** $\bar{X} \sim N(\mu, 0.134^2)$.
- One measure of the uncertainty inherent in using this random variable as an estimator is a *measure of spread of its distribution*.
- One such measure of spread is its **standard deviation**, in this case

$$SD(\bar{X}) = \frac{0.5}{\sqrt{14}} \approx 0.134.$$

- This can be interpreted as the “likely size of the error” or a little more precisely, its *square* is the “expected squared error” we would make using \bar{X} as an estimator of μ .
- **Summary** Under our statistical model, our estimate of μ is 3.064 with a standard error of 0.134.

More Precise Inference

- Providing an estimate and a standard error is a first important step in describing the population under our assumed statistical model.
- However, we have not fully utilised everything we know under this model. We have only used $SD(\bar{X})$ rather than its whole distribution; we have not yet exploited the fact that it is *normally distributed*.
- Recall that the **pivot**

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} = \frac{\bar{X} - \mu}{0.5/\sqrt{14}} = \frac{\text{estimate} - \text{true value}}{\text{standard error}} \sim N(0, 1)$$

whatever be the true value μ .

Confidence Interval

- It is straightforward to see that if $Z \sim N(0, 1)$ then the special value 1.96 satisfies

$$P(-1.96 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z < -1.96) = 0.95.$$

```
pnorm(1.96)
```

```
[1] 0.9750021
```

```
pnorm(1.96) - pnorm(-1.96)
```

```
[1] 0.9500042
```

- Applying this to our $N(0, 1)$ pivot we have that

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}} \leq 1.96\right) = 0.95$$

$$P\left(-1.96 \frac{\sigma_0}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

$$P\left(-1.96 \frac{\sigma_0}{\sqrt{n}} \leq \mu - \bar{X} \leq 1.96 \frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - 1.96 \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

- This is to say that the *random interval* given by $\bar{X} \pm 1.96 \frac{\sigma_0}{\sqrt{n}}$ has a particular probabilistic property:
 - it *covers* the true, unknown value of μ with probability 0.95.

Interpretation

- Now this “coverage probability” is a property of the *random interval* $\bar{X} \pm 1.96 \frac{\sigma_0}{\sqrt{n}}$.
- If we actually compute the corresponding interval according to $\bar{x} \pm 1.96 \frac{\sigma_0}{\sqrt{n}}$ we get an ordinary interval.
- We don't know if the “true value” μ is in the interval or not, and we shall never know.
- All we do know is that the *procedure* we have used (assuming the model is true) is such that in the “long run” it successfully covers the target (approximately) 95% of the time.
- Consider the following simulation: this simulates a $N(3.4, 0.5)$ random sample of size 14 and computes the estimate and corresponding 95% confidence interval 50 times. At the bottom we see which of the first 20 intervals covered 3.4 and which ones didn't:

```

est=0
intervals=matrix(0,50,2)
for (i in 1:50){
  samp=rnorm(14,m=3.4,s=0.5)
  est[i]=mean(samp)
  intervals[i,]=est[i]+c(-1,1)*1.96*0.5/sqrt(14)
}
covers=(intervals[,1]<3.4)&(intervals[,2]>3.4)
cbind(est,intervals,covers)[1:20,]

```

	est	covers
[1,]	3.564209	3.302293 3.826126
[2,]	3.153839	2.891923 3.415755
[3,]	3.588444	3.326528 3.850360
[4,]	3.395480	3.133564 3.657396
[5,]	3.579536	3.317620 3.841452
[6,]	3.847145	3.585229 4.109061
[7,]	3.422898	3.160982 3.684814
[8,]	3.429898	3.167982 3.691814
[9,]	3.311825	3.049909 3.573741
[10,]	3.233344	2.971428 3.495260
[11,]	3.400109	3.138193 3.662025
[12,]	3.667724	3.405808 3.929640
[13,]	3.402096	3.140180 3.664012
[14,]	3.613274	3.351358 3.875190
[15,]	3.625907	3.363991 3.887823
[16,]	3.374861	3.112945 3.636777
[17,]	3.684285	3.422369 3.946201
[18,]	3.206120	2.944204 3.468036
[19,]	3.545167	3.283251 3.807083
[20,]	3.277416	3.015499 3.539332

- In this case

```
sum(covers)
```

```
[1] 47
```

so 47 of the intervals covered 3.4.

- If we re-ran the simulation, it might be a different number (in fact the number of intervals covering is random, with a $B(50, 0.95)$ distribution!)
- Let us run it a few more times:

```
est=0
intervals=matrix(0,50,2)
for (i in 1:50){
  samp=rnorm(14,m=3.4,s=0.5)
  est[i]=mean(samp)
  intervals[i,]=est[i]+c(-1,1)*1.96*0.5/sqrt(14)
}
covers=(intervals[,1]<3.4)&(intervals[,2]>3.4)
sum(covers)
```

[1] 45

In this batch of 50, 45 of the intervals covered 3.4. Again...

```
est=0
intervals=matrix(0,50,2)
for (i in 1:50){
  samp=rnorm(14,m=3.4,s=0.5)
  est[i]=mean(samp)
  intervals[i,]=est[i]+c(-1,1)*1.96*0.5/sqrt(14)
}
covers=(intervals[,1]<3.4)&(intervals[,2]>3.4)
sum(covers)
```

[1] 47

In this batch of 50, 47 of the intervals covered 3.4. Again...

```
est=0
intervals=matrix(0,50,2)
for (i in 1:50){
  samp=rnorm(14,m=3.4,s=0.5)
  est[i]=mean(samp)
  intervals[i,]=est[i]+c(-1,1)*1.96*0.5/sqrt(14)
}
covers=(intervals[,1]<3.4)&(intervals[,2]>3.4)
sum(covers)
```

[1] 47

In this final batch of 50, 47 of the intervals covered 3.4.

Applied to Example

- So for our example we have
 - ▶ observed value $\bar{x} = 3.064286$;
 - ▶ standard error $\sigma_0/\sqrt{n} = 0.5/\sqrt{14} \approx 0.134$;
 - ▶ 95% confidence interval is this given by
$$3.064 \pm 1.96 * 0.134 \Rightarrow [2.930, 3.197].$$
- We can interpret this interval as a “range of plausible values” for the unknown μ (mean birthweight of babies of mothers who smoke) in a certain sense.

Different “confidence levels”

- The multiplier 1.96 was chosen so that the overall “confidence level” was 95%.
- However higher (or lower) confidence levels can be obtained by simply choosing a different multiplier.
- For instance, for 99% we would need the value c such that

$$P(-c \leq Z \leq c) = 0.99 .$$

- ▶ We can find this value using the R `qnorm()` (the exact inverse of `pnorm()`), or by reading normal tables “backwards”.

- Since $P(Z < -c) = P(Z > c)$ by symmetry and $P(Z < -c) + P(-c \leq Z \leq c) + P(Z > c) = 1$ we have that

$$P(Z < -c) = P(Z > c) = 0.005$$

and so

$$P(Z \leq c) = P(Z < -c) + P(-c \leq Z \leq c) = 0.995.$$

- Equivalently `pnorm(c)=0.995`. Thus the desired multiplier c is given by

```
qnorm(0.995)
```

```
[1] 2.575829
```

- Let us perform a slightly more involved simulation where we compute both the 95% and 99% confidence intervals, and repeat 200 times.

```
est=0
intervals95=matrix(0,200,2)
intervals99=matrix(0,200,2)
for (i in 1:200){
  samp=rnorm(14,m=3.4,s=0.5)
  est[i]=mean(samp)
  intervals95[i,]=est[i]+c(-1,1)*1.96*0.5/sqrt(14)
  intervals99[i,]=est[i]+c(-1,1)*2.576*0.5/sqrt(14)
}
covers95=(intervals95[,1]<3.4)&(intervals95[,2]>3.4)
covers99=(intervals99[,1]<3.4)&(intervals99[,2]>3.4)
sum(covers95)
```

```
[1] 188
```

```
sum(covers99)
```

```
[1] 198
```

Marks on standardised test example

- The estimate here is 62.3 and the standard error is $\sqrt{2}$. Thus
 - ▶ a 95% confidence interval for μ is $62.3 \pm 1.96\sqrt{2}$ which gives [59.53, 65.07].
 - ▶ a 99% confidence interval for μ is $62.3 \pm 2.576\sqrt{2}$ which gives [58.66, 65.94].

One-sided Confidence Interval

- The type of confidence interval above is a natural way of expressing the uncertainty of the estimate when deviations in either direction are equally significant.
- The interval provides a “set of plausible values” of μ , more precisely a set of values distributed in *either direction* about the estimate that are plausible for μ in a certain sense.
- However given the original scientific question about these birthweights, namely that smoking perhaps lowers birthweights, we may wish a different approach.

- Given that the estimate 3.064 is less than the “special” value of 3.4, we may instead ask “what is the **largest** value that is consistent with the data in this sense”?
- We thus reason as follows: with $Z = \frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}} \sim N(0, 1)$ the special value 1.645 satisfies

$$P(-1.645 \leq Z) = 0.95$$

$$P\left(-1.645 \leq \frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}}\right) = 0.95$$

$$P\left(-1.645 \frac{\sigma_0}{\sqrt{n}} \leq \bar{X} - \mu\right) = 0.95$$

$$P\left(\mu \leq \bar{X} + 1.645 \frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

- So the random interval $(-\infty, \bar{X} + 1.645 \frac{\sigma_0}{\sqrt{n}})$ contains the unknown true value μ with probability 0.95.
- In our example we thus get $3.064 + (1.645 * 0.134) = 3.28443$.
- We can then quote the *one-sided 95% confidence interval* $(-\infty, 3.28443)$ or the *upper 95% confidence limit* 3.28443.

- Since

```
qnorm(0.01)
```

[1] -2.326348

so that

$$P(-2.326 \leq Z) = 0.99,$$

the same reasoning gives that an upper 99% confidence limit for μ in this example is given by $3.064 + (2.326 * 0.134) = 3.375684$.

- Note that this is still *less* than the “special” (hypothesised) value of 3.4, so 3.4 is not plausible/consistent with the data at the 99% confidence level.
- This bears an eerie similarity to the original hypothesis test above, where we concluded that since the observed \bar{x} was less than the 1% critical value that the data were “significantly different” to the hypothesis at the 1% level.
- We shall see that this similarity indicates a deep connection between the two procedures.

Alternative hypothesis; Z -statistic

- We can develop an extended version of hypothesis testing in this broader statistical model framework.
- In particular we can model the data generating mechanism both when the null hypothesis holds and when it doesn't.
- This is formalised by the concept of an **alternative hypothesis**.

- Suppose we are testing the null hypothesis $H_0: \mu = \mu_0$ for some known value μ_0 (e.g. 3.4 as in the birthweights example).
- We again start with the fact that under our statistical model that $\bar{X} \sim N\left(\mu, \frac{\sigma_0^2}{n}\right)$, the **pivot**

$$\frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}} \sim N(0, 1)$$

whatever be the true value of μ .

- Note again that the denominator is precisely that *standard error* associated with the estimate.

- If we replace μ in the pivot above by its **hypothesised value** μ_0 we get the Z -statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}.$$

This now only depends on known values and the data and so once the data is obtained its value can be computed.

- Under the model, as a random variable the Z -statistic

$$Z = \underbrace{\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}}_{\text{pivot} \sim N(0,1)} + \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \sim N\left(\frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}, 1\right)$$

and so in general its distribution is unknown.

- However **when H_0 is true** $Z \sim N(0, 1)$.

Birthweights example

- Having obtained our estimate $\bar{x} = 3.064$ and standard error 0.134 we note that z , the observed value of the Z statistic is $z = \frac{3.064 - 3.4}{0.134} \approx -2.50533$.
- This is to say that the sample mean is 2.50533 standard errors below the hypothesised value 3.4.

Alternative hypothesis

There are 3 cases:

“Less”

We are only interested in cases where the true μ is **less** than μ_0 . In this case we say that the alternative hypothesis is $H_1: \mu < \mu_0$.

- For a given significance level $0 < \alpha < 1$ we “reject” H_0 in favour of H_1 if $z < c(\alpha)$ where $c(\alpha)$ is the *lower α* point of the $N(0, 1)$ distribution, e.g.
 - ▶ for $\alpha = 0.05$, $c(\alpha) = -1.645$;
 - ▶ for $\alpha = 0.01$, $c(\alpha) = -2.326$.
- The *observed level of significance* or *p-value* is $P(Z \leq z)$ where $Z \sim N(0, 1)$ i.e. if H_0 is true.

“Greater”

We are only interested in cases where the true μ is **greater** than μ_0 . In this case we say that the alternative hypothesis is $H_1: \mu > \mu_0$.

- For a given significance level $0 < \alpha < 1$ we “reject” H_0 in favour of H_1 if $z > c(\alpha)$ where $c(\alpha)$ is the *upper α* point of the $N(0, 1)$ distribution, e.g.
 - ▶ for $\alpha = 0.05$, $c(\alpha) = +1.645$;
 - ▶ for $\alpha = 0.01$, $c(\alpha) = +2.326$.
- The *observed level of significance* or *p-value* is $P(Z \geq z)$ where $Z \sim N(0, 1)$ i.e. if H_0 is true.

“Not equal”

We are interested in either of the above two cases. In this case we say that the alternative hypothesis is $H_1: \mu \neq \mu_0$.

- For a given significance level $0 < \alpha < 1$ we “reject” H_0 in favour of H_1 if $|z| > c(\alpha)$ (i.e. $z > c(\alpha)$ or $z < -c(\alpha)$) where $c(\alpha)$ is the *upper* $\alpha/2$ point of the $N(0, 1)$ distribution, i.e. so that if $Z \sim N(0, 1)$ (H_0 true) then
$$P(|Z| \geq c(\alpha)) = P(Z < -c(\alpha)) + P(Z > c(\alpha)) = \alpha$$
e.g.
 - for $\alpha = 0.05$, $c(\alpha) = 1.96$;
 - for $\alpha = 0.01$, $c(\alpha) = 2.576$.
- The *observed level of significance* or *p-value* is $P(|Z| \geq |z|) = 2P(Z \geq |z|)$ where $Z \sim N(0, 1)$ i.e. if H_0 is true.

- Cases “Less” and “Greater” are called **one-sided tests**.
- Case “Not equal” is a **two-sided test**.
- Careful examination of the examples above shows the birthweights example here to be precisely case “Less”, while the marks on the standardised test example is case “Not equal”.

Inference based on Student's t -distribution

- The inferential procedures in the previous section all have one particular feature: **the variance of the population is assumed known.**
- In some problems this is not realistic. However, we can make up for this in a kind of "trade-off": **If we are willing to assume the population is normal, we can relax the assumption that we know the variance.**
- This is because when the population is normal, we know the distribution of the pivot

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

where as usual \bar{X} and S denote the sample mean and sample sd, respectively and μ is the *true* population mean.

- Let us revisit the birthweights example. There are various points of view we can take:
 - a “pure hypothesis test” of H_0 : “birthweights normal with mean 3.4”, using small values of the T -statistic
$$T = \frac{\bar{X} - 3.4}{S/\sqrt{14}}$$
as evidence against H_0 .
 - This does not require making any assumptions about what might be going on if H_0 is false.
 - assume a model where the birthweights are normal with mean μ unknown; then we can
 - construct a confidence interval for μ (one-sided makes more sense here)
 - test the same H_0 above i.e. $\mu = 3.4$ against the alternative $\mu < 3.4$
- This is a model that makes more assumptions but allows us to say more in the end (subject to the assumptions being reasonable).

Pure Hypothesis Test

- This only measures evidence against one particular assumption, that the birthweights are normal with mean 3.4, where it is understood (due to the context, where mothers who smoke are suspected of having lower birthweight babies) that we would only “reject” this hypothesis for a “small enough sample mean”.

```
x=c(2.5, 2.9, 2.8, 3.1, 3.4, 3.4, 3.5, 2.8, 3.3, 2.8, 3.7, 2.3, 3.4, 3.0)  
x
```

```
[1] 2.5 2.9 2.8 3.1 3.4 3.4 3.5 2.8 3.3 2.8 3.7 2.3 3.4  
[14] 3.0
```

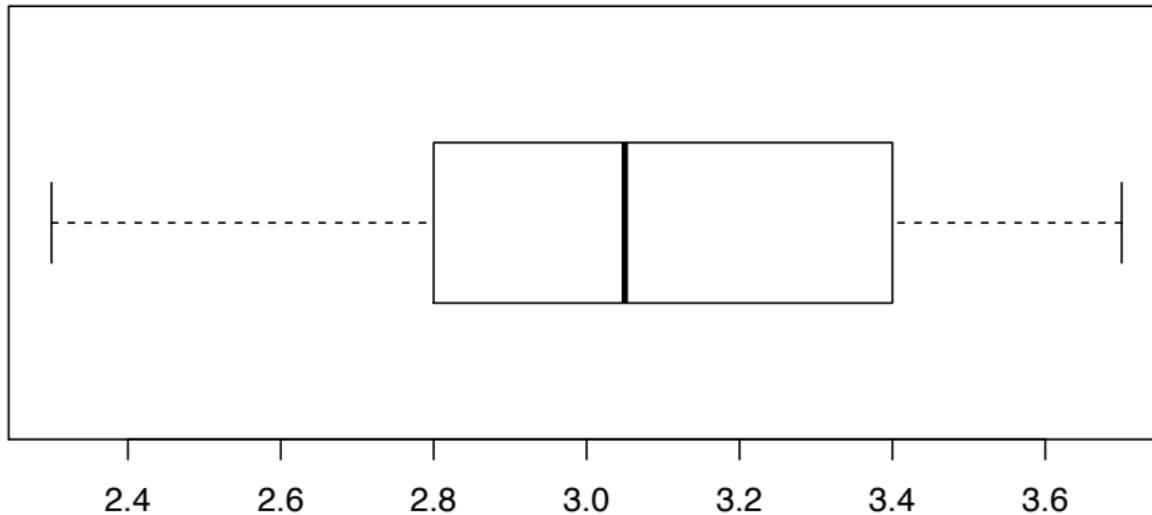
```
mean(x)
```

```
[1] 3.064286
```

```
sd(x)
```

```
[1] 0.4049827
```

```
boxplot(x, horizontal=T)
```



- Before where we assumed we knew the population variance we could directly describe the anticipated distribution of the sample mean \bar{X} when H_0 is true.
- Now we cannot, however we can make a similar assessment: is \bar{X} much smaller than 3.4 *in light of the value of the sample sd?*
- If H_0 is true the T -statistic

$$T = \frac{\bar{X} - 3.4}{S/\sqrt{14}} \sim t_{13}.$$

- If we are to think of this as a “detection” problem, then we need a critical value $c(\alpha)$ so that

$$P(T \leq c(\alpha)) = \alpha$$

when H_0 is true i.e. if $T \sim t_{13}$.

- This value can be found using the R function `qt()` (the Student's t analog to `qnorm()`, the inverse of the cdf `pt()`): for $\alpha = 0.05$ the critical value is

```
qt (.05 , df=13)
```

[1] -1.770933

while for $\alpha = 0.01$ the critical value is

```
qt (.01 , df=13)
```

[1] -2.650309

- Compare these to the corresponding one-sided critical values for the $N(0, 1)$ distribution:

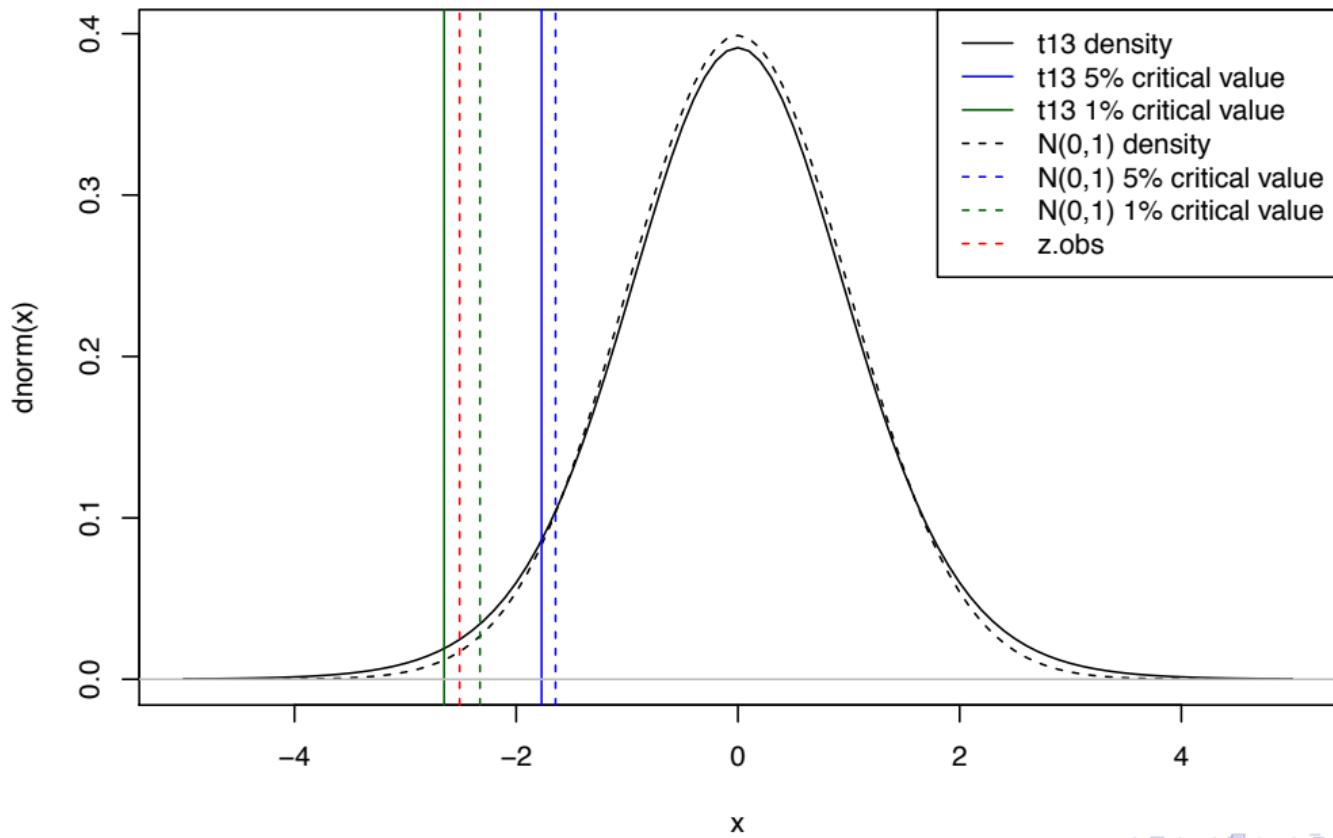
```
qnorm(.05)
```

```
[1] -1.644854
```

```
qnorm(.01)
```

```
[1] -2.326348
```

- They are bigger (i.e. larger in size and “more negative”), since for such a small sample size (14) there is extra variability on the denominator of T compared to the Z -statistic.
- Recall that in that case (using the Z -statistic, with the variance assumed known), the result was significant at both the 5% and 1% levels.
- In particular note that the 1% critical value is *bigger* than the observed value of the Z -statistic from last lecture.



- The observed value is $t = (\bar{x} - 3.4)/(s/\sqrt{14})$ i.e.

```
t.obs=(mean(x)-3.4)/(sd(x)/sqrt(14))  
t.obs
```

[1] -3.101683

- This is bigger than the observed value of the z-statistic from last lecture because the "assumed known" population sd of 0.5 has been replaced with the *smaller* sample sd 0.4049827.
- Indeed it is larger than the 1% critical value and so is significant at both the 1% and 5% levels.
- The observed level of significance is given by

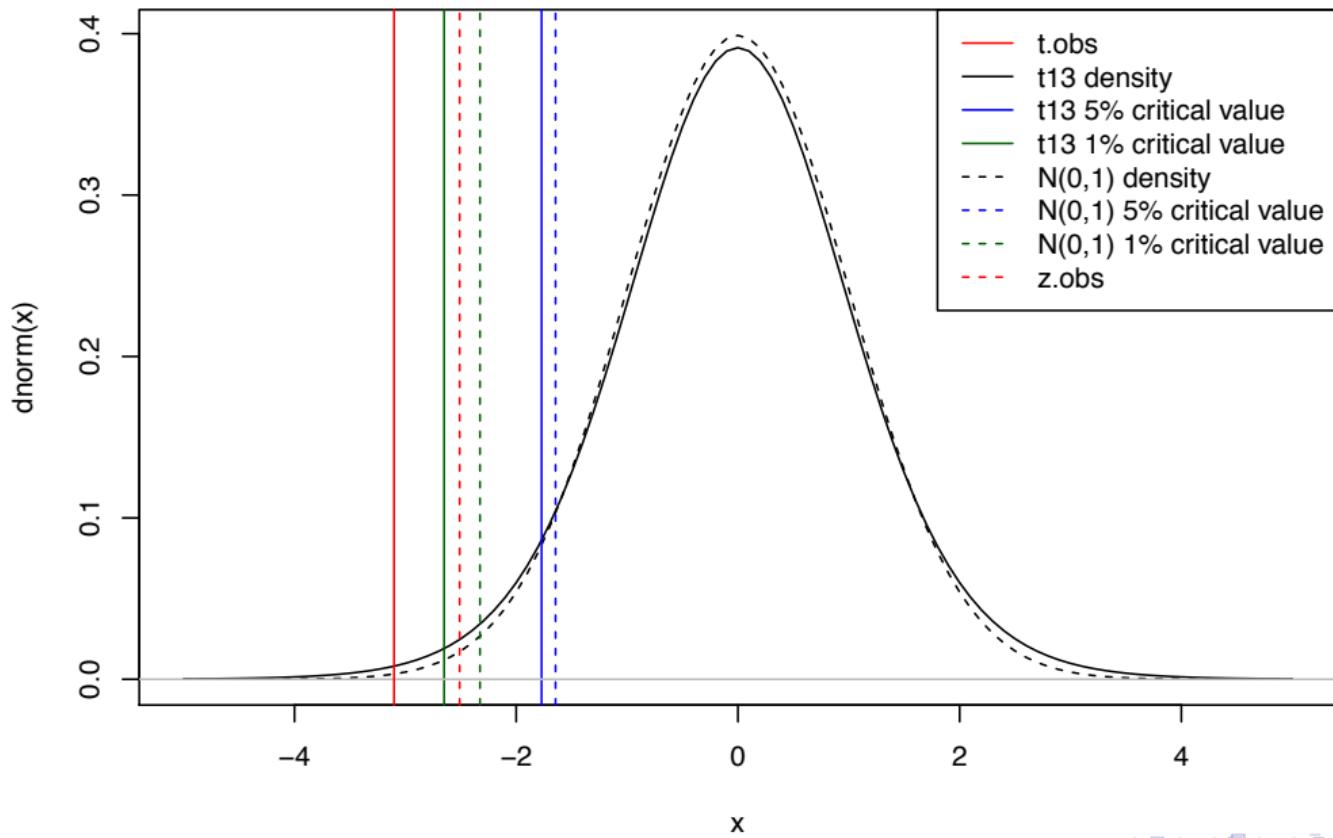
$$P(t_{13} \leq -3.102)$$

which is given by

```
pt(t.obs, df=13)
```

[1] 0.004209727

... which is interestingly slightly below 1%, as it was for the z-test.



- While this approach (“pure hypothesis test”) makes very few assumptions, it doesn’t let us say much, only that there is some evidence against the claim that the birthweights of babies of mothers who smoke have the same distribution as “normal” babies.
- To say more, we make more assumptions.

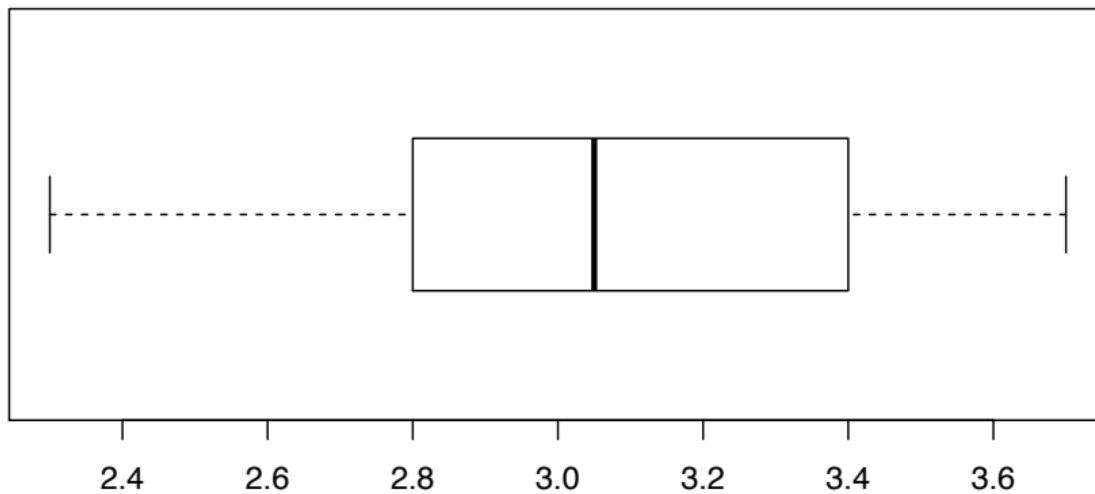
Broader statistical model

- Let us now model the birthweights of babies of mothers who smoke as (independent) $N(\mu, \sigma^2)$ random variables, for μ and σ *both unknown*.
- The population mean μ is the “parameter of interest” while σ^2 here is regarded as a “nuisance parameter”; we don’t really care what value it is, although not knowing it does make our work *slightly* harder.
- We need to use a t -distribution instead of a $N(0, 1)$ distribution, and we may lose some **power** that is, the probability of detecting a difference is slightly lower under this model, we would need a slightly larger effect to get the same “significance” in some sense.

Checking assumptions

- Are these “slightly stronger” assumptions reasonable? One way to get a feel for the answer to this is to look at a boxplot:

```
boxplot(x, horizontal=T)
```



Upper confidence limits

- We derive the form of the upper confidence limits based on the t_{13} -distribution.
- We simply use the 5% and 1% critical values from the one-sided t -tests above.

95%

- Since

```
qt(0.05, df=13)
```

[1] -1.770933

we also have that

```
1-pt(-1.771, df=13)
```

[1] 0.9500057

so $P(t_{13} \geq -1.771) = 0.95$.

- So then whatever the true value of μ ,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{14}} \sim t_{13}$$

and so

$$P\left(\frac{\bar{X} - \mu}{S/\sqrt{14}} \geq -1.771\right) = 0.95$$

$$P\left(\bar{X} - \mu \geq -1.771 \frac{S}{\sqrt{14}}\right) = 0.95$$

$$P\left(\bar{X} + 1.771 \frac{S}{\sqrt{14}} \geq \mu\right) = 0.95.$$

- Thus the one-sided 95% confidence interval for μ here is of the form

$$\left(-\infty, \bar{X} + 1.771 \frac{S}{\sqrt{14}}\right).$$

- That is to say, the *largest* value of μ consistent with the data in this sense is $\bar{X} + 1.771 \frac{S}{\sqrt{14}}$.
- The observed value of this upper limit is then

```
mean(x) + 1.771 * sd(x) / sqrt(14)
```

[1] 3.255972

- The interval $(-\infty, 3.256)$ does **not** include the special value 3.4 so at this “confidence level”, 3.4 is **not** a plausible value for μ .
- This is in perfect agreement with the finding from our one-sided t -test above; we would reject the null hypothesis $H_0: \mu = 3.4$ at the 5% level.

- In the same way as above, since

```
qt(0.01, df=13)
```

```
[1] -2.650309
```

we have that $P(t_{13} \geq -2.65) = 0.99$:

```
1-pt(-2.65, df=13)
```

```
[1] 0.9899941
```

- Thus the upper 99% confidence limit for μ is given by

```
mean(x)+2.65*sd(x)/sqrt(14)
```

```
[1] 3.351112
```

that is to say the one-sided 99% confidence interval for μ is given by $(-\infty, 3.351)$.

- This does not include the special value 3.4 either and so it is **not** a plausible value for μ at the 99% confidence level.
- This is in perfect agreement with our t -test findings above.
- The test result was significant at the 1% level so in both senses 3.4 is **not** a plausible value for μ at this level.

Specifying an Alternative Hypothesis

- Finally, we conduct a t -test under the stronger assumptions of our statistical model.
 - Model:** $X_1, \dots, X_{14} \sim N(\mu, \sigma^2)$ for μ and σ^2 both unknown
 - Hypotheses:** $H_0: \mu = 3.4$ against $H_1: \mu < 3.4$
 - Test Statistic:** $T = \frac{\bar{X} - 3.4}{S/\sqrt{14}} \sim t_{13}$ if H_0 true.
 - How is evidence measured?**: Since T would tend to take smaller values under H_1 , the p-value is given by $P(t_{13} \leq t_{\text{obs}})$ where t_{obs} is the observed value of the statistic T .

- **P-value** Since \bar{X} takes the value 3.064286 and S takes the value 0.4049827 the statistic takes the value -3.101683 and so the p-value is given by

```
t.obs=(mean(x)-3.4)/(sd(x)/sqrt(14))  
t.obs
```

```
[1] -3.101683
```

```
pt(t.obs,df=13)
```

```
[1] 0.004209727
```

- This is significant at the 1% level, and so provides strong evidence against the null hypothesis that the average birthweight of babies of mothers who smoke is 3.4 kg (against the alternative that it is less).

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
 - Lecture 15
 - Lecture 16
 - Lecture 17
 - Two-sample problems
 - Unequal variances; the Welch test
 - Using the R function `t.test()`
 - Lecture 18
- 5 Inference Part 2: Discrete models

Two-sample problems

- We shall study two different scenarios where we may perform t -tests or construct a corresponding confidence interval based on *two samples*, not one. These are
 - ▶ two *paired* samples;
 - ▶ two *independent* samples.

Two Paired Samples

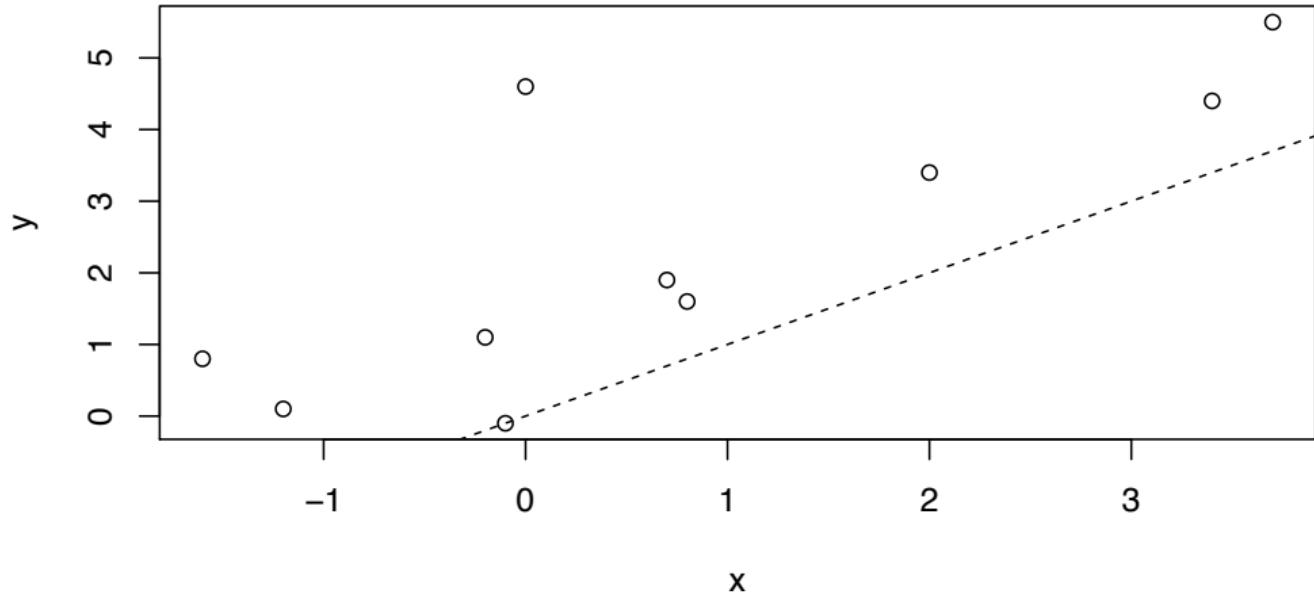
- The original paper by Student (1908) that introduced the “*t*-test” studied the (now famous) “sleep data”, which is present in R:

```
sleep
```

	extra	group	ID
1	0.7	1	1
2	-1.6	1	2
3	-0.2	1	3
4	-1.2	1	4
5	-0.1	1	5
6	3.4	1	6
7	3.7	1	7
8	0.8	1	8
9	0.0	1	9
10	2.0	1	10
11	1.9	2	1
12	0.8	2	2
13	1.1	2	3
14	0.1	2	4
15	-0.1	2	5
16	4.4	2	6
17	5.5	2	7
18	1.6	2	8
19	4.6	2	9
20	3.4	2	10

- This is a **data frame**, a special kind of R matrix where each column is possibly a different kind of “variable” and each row gives an observation.
- Ten individuals each tried two different drugs designed to improve sleep. Each observation is the “average gain in sleep” for each individual under each of the two drugs.
- It is perhaps more natural to plot these as ordered pairs, one point for each person.

```
x=sleep$extra[sleep$group==1]
y=sleep$extra[sleep$group==2]
plot(x,y)
abline(0,1,lty=2)
```



```
mean(x)
```

```
[1] 0.75
```

```
mean(y)
```

```
[1] 2.33
```

```
sd(x)
```

```
[1] 1.78901
```

```
sd(y)
```

```
[1] 2.002249
```

- All of the y -values are positive here, but some x -values are negative.
- All but one point is above the line $y = x$, the remaining point being **on** the line.
- Thus for all points $y \geq x$. This suggests the y drug might be better. But how do we test this exactly?

- The trick is to turn this into a **one-sample problem**, but obtaining a single sample of $y - x$ differences:

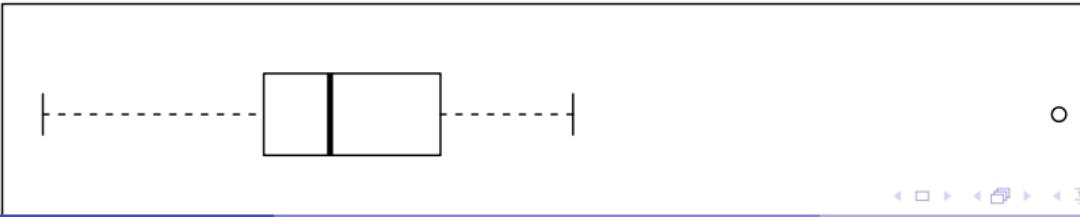
```
d=y-x  
mean(d)
```

```
[1] 1.58
```

```
sd(d)
```

```
[1] 1.229995
```

```
boxplot(d, horizontal=T)
```



- Aside from the outlier, the boxplot is reasonably symmetric.
- We might note that the outlier is not ideal, but perhaps one outlier is not critical.
- If we could use an alternative test that doesn't need an assumption of normality, that *may* be preferable (more on this later).
- However this is the famous sleep data! So let's analyse it with a *t*-test anyway.

Pure Hypothesis Test

- **Null Hypothesis:** H_0 : "differences are normal with mean zero"
- **Test Statistic:** $T = \bar{D}/(S/\sqrt{10})$ where \bar{D} is the mean difference and S is the sample sd of the differences. If H_0 true, $T \sim t_9$.
- **How is evidence measured?** Unless one drug was *anticipated* to be superior before the data was collected, we must treat this as a *two-sided* test. Thus if t_{obs} is the observed value of the statistic T , the p-value will be

$$P(|T| \geq |t_{\text{obs}}|) = 2P(T \geq |t_{\text{obs}}|) \text{ where } T \sim t_9.$$

- **P-value:** The observed value is

```
t.obs=mean(d)/(sd(d)/sqrt(10))  
t.obs
```

```
[1] 4.062128
```

- Therefore the (two-sided) p-value is

```
2*(1-pt(t.obs,df=9))
```

```
[1] 0.00283289
```

- This is significant at the 1% level, formally providing evidence against the hypothesis of "no difference".
- This thus *indirectly* suggests there is a difference.

Confidence Interval

- We assume the differences are normal with unknown mean μ (and unknown variance σ^2 too).
- The boxplot indicates this isn't *too* bad an assumption, although the outlier makes it less than ideal.
- Since, whatever the value of μ , the random variable

$$\frac{\bar{D} - \mu}{S/\sqrt{10}} \sim t_9$$

and since

```
qt(.025, df=9)
```

[1] -2.262157

we have that

$$P(-2.262 \leq t_9 \leq 2.262) = 0.95.$$

- Thus using the now (hopefully) familiar steps,

$$P\left(-2.262 \leq \frac{\bar{D} - \mu}{S/\sqrt{10}} \leq 2.262\right) = 0.95$$

$$P\left(-2.262 \frac{S}{\sqrt{10}} \leq \bar{D} - \mu \leq 2.262 \frac{S}{\sqrt{10}}\right) = 0.95$$

$$P\left(-2.262 \frac{S}{\sqrt{10}} \leq \mu - \bar{D} \leq 2.262 \frac{S}{\sqrt{10}}\right) = 0.95$$

$$P\left(\bar{D} - 2.262 \frac{S}{\sqrt{10}} \leq \mu \leq \bar{D} + 2.262 \frac{S}{\sqrt{10}}\right) = 0.95$$

- So the *random interval* given by $\bar{D} \pm 2.262S/\sqrt{10}$ contains μ with probability 0.95.
- If we determine the value(s) of these interval endpoints we get

```
mean(d)+c(-1,1)*2.262*sd(d)/sqrt(10)
```

[1] 0.7001754 2.4598246

- Note that this interval does *not* include zero.
- Thus at this 95% level 0 is **not** a plausible value for μ , hence suggesting a difference.

- The 99% interval is obtained in a similar way, this time using the multiplier

```
qt (.005 , df=9)
```

```
[1] -3.249836
```

so the 99% interval is

```
mean (d)+c (-1 , 1)*3.25*sd(d) / sqrt(10)
```

```
[1] 0.3158841 2.8441159
```

- This does not contain zero either, which is in perfect agreement with our earlier *t*-test.

Alternative Hypothesis

- The extended version of the hypothesis test which also models an alternative hypothesis is similar to above:
 - ▶ **Model:** D_1, \dots, D_{10} ind $N(\mu, \sigma^2)$, μ and σ^2 both unknown
 - ▶ **Hypotheses:** $H_0: \mu = 0$ against $H_1: \mu \neq 0$.
 - ▶ **the rest is the same as above, really:**
 - ▶ **Test Statistic:** $T = \bar{D}/(S/\sqrt{10})$ where \bar{D} is the mean difference and S is the sample sd of the differences. If H_0 true, $T \sim t_9$.
 - ▶ **How is evidence measured?** This is a *two-sided* test. Thus if t_{obs} is the observed value of the statistic T , the p-value will be $P(|T| \geq |t_{\text{obs}}|) = 2P(T \geq |t_{\text{obs}}|)$ where $T \sim t_9$.
 - ▶ **P-value:** The observed value is

```
t.obs=mean(d)/(sd(d)/sqrt(10))  
t.obs
```

```
[1] 4.062128
```

- Therefore the (two-sided) p-value is

```
2*(1-pt(t.obs,df=9))
```

```
[1] 0.00283289
```

- This is significant at the 1% level, formally providing evidence against the hypothesis of "no difference".
- This thus *indirectly* suggests there is a difference.

Two Independent Samples

- Recall how we get Student's t -distribution in the first place:
 - ▶ If \bar{X} and S^2 are the sample mean and variance from a normal random sample then $\bar{X} \sim N(\mu, \sigma^2)$ **independently of**

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

- ▶ Then the Studentised mean given by

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} / \frac{S}{\sigma} \sim \frac{Z}{\sqrt{\chi_{n-1}^2/(n-1)}}$$

with the $Z \sim N(0, 1)$ and χ_{n-1}^2 independent.

- Now consider what would happen if we had two normal random samples
 - ▶ X_1, \dots, X_m indep $N(\mu_X, \sigma^2)$ *independently of*
 - ▶ Y_1, \dots, Y_n indep $N(\mu_Y, \sigma^2)$

that is where the **two population variances are the same** but all 3 parameters μ_X , μ_Y and σ^2 were unknown.

- In particular the mean difference

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right),$$

since $\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y})$.

- Furthermore, the **standardised version**

$$\frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0, 1).$$

- If the σ were known we could use this as a test statistic to test $H_0: \mu_X = \mu_Y$.
- However if σ is unknown what do we do?

- Well, if we could find an estimator $\hat{\sigma}^2$ such that

$$\hat{\sigma}^2 \sim \sigma^2 \chi_d^2 / d$$

for some d (and which is independent of $\bar{X} - \bar{Y}$!) then the **Studentised** version

$$\frac{\bar{X} - \bar{Y}}{\hat{\sigma} \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \sim \frac{\bar{X} - \bar{Y}}{\sqrt{\chi_d^2 / d}} \sim t_d$$

under the null hypothesis $H_0: \mu_X = \mu_Y$.

- Can we find such an estimator $\hat{\sigma}^2$?

- Yes we can.
- Consider the two *sample variances* S_X^2 and S_Y^2 , which satisfy
 - ▶ $(m - 1)S_X^2 \sim \sigma^2 \chi_{m-1}^2$;
 - ▶ $(n - 1)S_Y^2 \sim \sigma^2 \chi_{n-1}^2$;
 - ▶ they are both independent (and independent of the sample means \bar{X} and \bar{Y} !).
- Then the sum

$$(m - 1)S_X^2 + (n - 1)S_Y^2 = \sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2 \sim \sigma^2 \chi_{m+n-2}^2$$

since the sum of independent χ^2 's is also χ^2 .

- So our “Pooled Sample Variance”

$$S_p^2 = \frac{(m - 1)S_X^2 + (n - 1)S_Y^2}{m + n - 2} = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{m + n - 2}$$

exactly fulfils the above requirement, with the degrees of freedom $d = m + n - 2$.

Example

- The following random samples are measurements of the heat-producing capacity (in millions of calories per ton) of specimens of coal from two mines.

```
mine1=c(8400,8230,8380,7860,7930)
mine2=c(7510,7690,7720,8070,7660)
meandiff=mean(mine1)-mean(mine2)
meandiff
```

```
[1] 430
```

```
sp=sqrt( (4*var(mine1)+4*var(mine2))/8 )
sp
```

```
[1] 230.3259
```

```
t.stat=meandiff/(sp*sqrt((1/5)+(1/5)))
t.stat
```

```
[1] 2.95186
```

- According to our reasoning above, this is the value taken by a statistic whose distribution is t_8 under the hypothesis that the two mine means are equal.
- This is a two-sided test so a p-value is

```
2*(1-pt(t.stat,df=8))
```

```
[1] 0.01837337
```

- So this result is significant at the 2% level but not the 1% level.
- It nonetheless constitutes some evidence against the hypothesis of no difference between the mines.

Two-sided confidence interval

- The denominator of the statistic is the standard error of the estimate of the mean difference in the numerator.
- Since

```
qt (.025 , df=8)
```

```
[1] -2.306004
```

a 95% (two-sided) confidence interval for the population mean difference is given by

```
meandiff + c(-1,1)*2.306*sp*sqrt((1/5)+(1/5))
```

```
[1] 94.08299 765.91701
```

- To see why, note that **whatever the values of μ_X and μ_Y** , the Studentised mean difference

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_8 .$$

- Then, since

$$P\left(-2.306 \leq \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \leq 2.306\right) = 0.95,$$

applying the (by now I hope) familiar manipulations we can write

$$P\left(\bar{X} - \bar{Y} - 2.306 S_p \sqrt{\frac{1}{m} + \frac{1}{n}} \leq \mu_X - \mu_Y \leq \bar{X} - \bar{Y} + 2.306 S_p \sqrt{\frac{1}{m} + \frac{1}{n}}\right) = 0.95.$$

- The values above are the observed values of these interval endpoints.
- Note that zero is not in the interval, it is not a “plausible value” at the 95% confidence level.
- Put another way, the observed mean difference is significantly different from 0 at the 5% level.

- However, a two-sided 99% confidence interval uses the multiplier

```
qt(0.005, df=8)
```

```
[1] -3.355387
```

since then $P(-3.355 \leq t_8 \leq 3.355) = 0.99$.

- The resultant interval is

```
meandiff + c(-1, 1)*3.355*sp*sqrt((1/5)+(1/5))
```

```
[1] -58.72574 918.72574
```

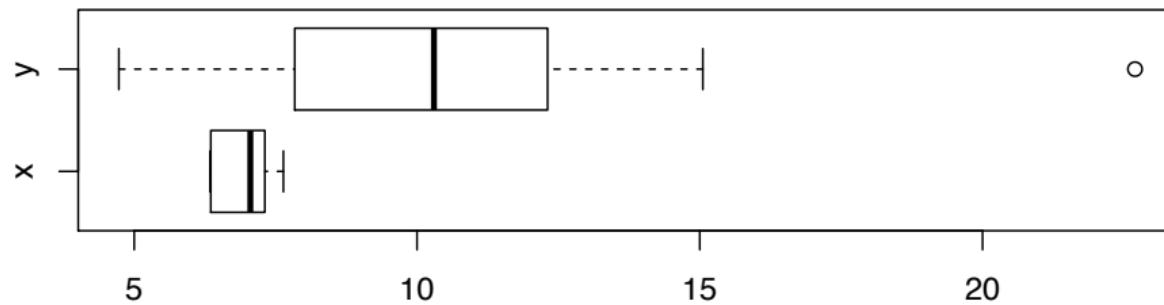
which *includes* zero.

- So at the 99% confidence level, zero is a plausible value for the true population mean difference.

- These two confidence intervals are in perfect agreement with the t-tests above:
 - ▶ the sample mean difference is *significantly* different at the 5% level \Leftrightarrow the 95% confidence interval does *not* include zero \Leftrightarrow “zero not plausible” (at this level);
 - ▶ the sample mean difference is *not significantly* different at the 1% level \Leftrightarrow the 99% confidence interval *does* include zero \Leftrightarrow “zero plausible” (at this level);

Unequal variances; the Welch test

- The two-sample t -test developed above uses a model where the two unknown normal populations are assumed to have the *same variance*.
- However in some situations this may not be a reasonable assumption.
- For instance if two samples of sizes 10 (x) and 13 (y) have boxplots that look like



it would not be reasonable to assume these were samples from normal populations *with the same variance*:

- However it might be of interest to test if the apparent differences in mean/location are *significantly* different (in either a one-sided or two-sided sense).
- A natural model is that the two samples are (respectively) values taken by X_1, \dots, X_m which are independent $N(\mu_X, \sigma_X^2)$ independently of Y_1, \dots, Y_n which are independent $N(\mu_Y, \sigma_Y^2)$.
- We must ask ourselves two questions:
 - ① Can we identify a sensible/natural test statistic?
 - ② If so, what is its distribution if the two population means are equal, i.e. under the null hypothesis $H_0: \mu_X = \mu_Y$?
- The answer to the first question is not too difficult: the sample mean difference

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)$$

(this follows since the sample means are normal and independent and so their difference is also normal).

- So under this model an estimator of $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y}$; its standard deviation is

$$SD(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}.$$

- If the two population variances were known then we could use as test statistic

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim N(0, 1) \text{ if } H_0 \text{ is true.}$$

- However if the population variances are **unknown**, all of the *t*-tests we have seen suggest replacing the unknown σ 's here with their sample versions, giving the statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}} \sim ??? \text{ if } H_0 \text{ is true.}$$

- Note that this statistic has the same basic form as all others considered so far:

$$\frac{\text{estimate} - \text{hypothesised value}}{\text{(estimated) standard error}}$$

The only remaining difficulty is to determine its distribution when H_0 is true.

- It turns out **this does not have an (exact) Student's-t distribution when H_0 is true.**
- However this distribution is *well approximated* by the Student's-t distribution with a special degrees-of-freedom that depends on the data.
- It has a complicated form but can be computed easily enough using R.

Using the R function `t.test()`

- Recall the *t*-test we performed on the birthweights data:

```
x=c(2.5, 2.9, 2.8, 3.1, 3.4, 3.4, 3.5, 2.8, 3.3, 2.8, 3.7, 2.3, 3.4, 3.0)  
x
```

```
[1] 2.5 2.9 2.8 3.1 3.4 3.4 3.5 2.8 3.3 2.8 3.7 2.3 3.4  
[14] 3.0
```

```
mean(x)
```

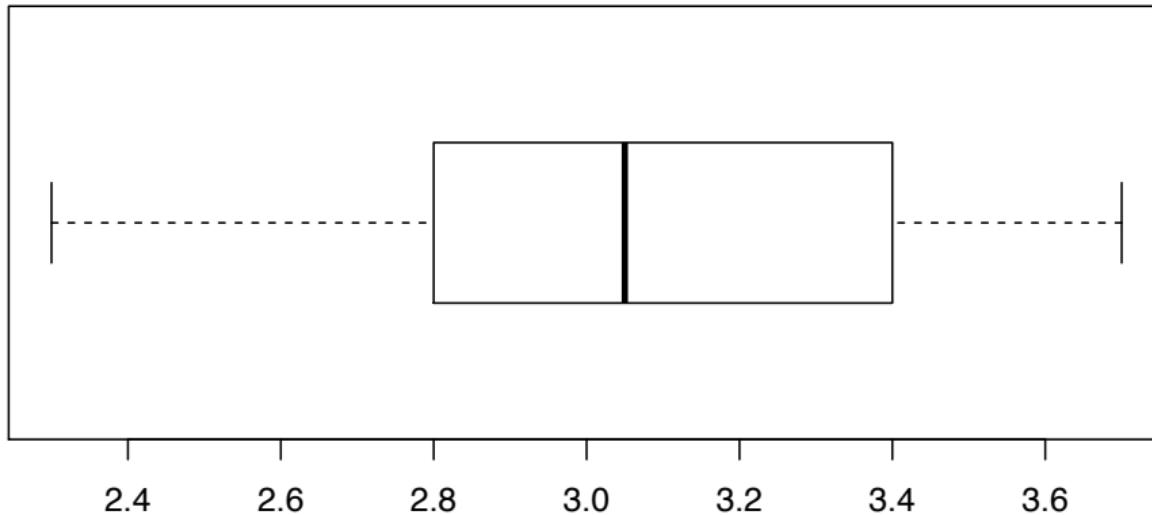
```
[1] 3.064286
```

```
sd(x)
```

```
[1] 0.4049827
```

```
n=length(x)
```

```
boxplot(x, horizontal=T)
```



```
stat=(mean(x)-3.4)/(sd(x)/sqrt(n))
stat
```

```
[1] -3.101683
```

```
pt(stat,df=n-1)
```

```
[1] 0.004209727
```

- The corresponding **one-sided** 95% and 99% upper confidence limits were obtained according to (for the 95% level):

```
mult.95=qt(.05,df=n-1)
mult.95
```

```
[1] -1.770933
```

```
mean(x)-mult.95*sd(x)/sqrt(n)
```

```
[1] 3.255965
```

(and for the 99% level):

```
mult.99=qt(.01,df=n-1)
mult.99
```

```
[1] -2.650309
```

```
mean(x)-mult.99*sd(x)/sqrt(n)
```

```
[1] 3.351145
```

- This can all be performed in a single line as follows:

```
t.test(x, mu=3.4, alternative="less")
```

One Sample t-test

```
data: x
t = -3.1017, df = 13, p-value = 0.00421
alternative hypothesis: true mean is less than 3.4
95 percent confidence interval:
-Inf 3.255965
sample estimates:
mean of x
3.064286
```

- Notice that the output also returns a **one-sided** confidence interval (since the test is specified as one-sided) at the 95% level (the default).

- Different levels can be used e.g.:

```
t.test(x, mu=3.4, alternative="less", conf.level=.99)
```

One Sample t-test

```
data: x
t = -3.1017, df = 13, p-value = 0.00421
alternative hypothesis: true mean is less than 3.4
99 percent confidence interval:
 -Inf 3.351145
sample estimates:
mean of x
3.064286
```

- For Student's sleep data, recall that we did a two-sided test on the *single* sample of differences:

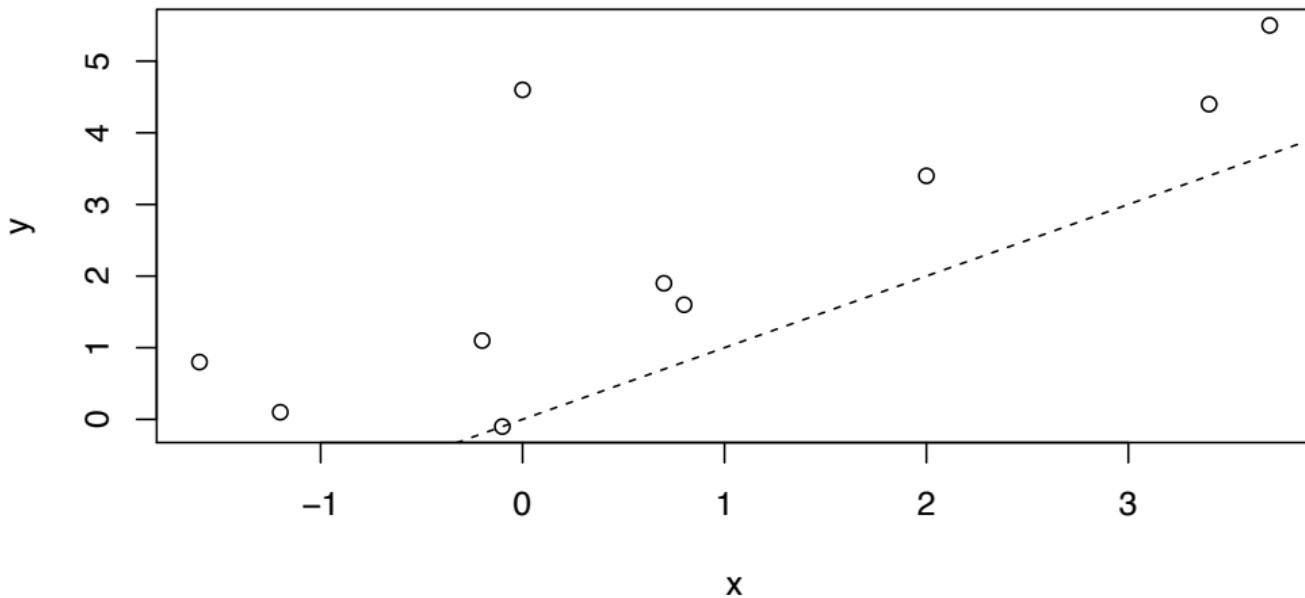
```
x=sleep$extra[sleep$group==1]  
y=sleep$extra[sleep$group==2]  
x
```

```
[1] 0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
```

```
y
```

```
[1] 1.9  0.8  1.1  0.1 -0.1  4.4  5.5  1.6  4.6  3.4
```

```
plot(x,y)
abline(0,1,lty=2)
```



```
mean(x)
```

```
[1] 0.75
```

```
mean(y)
```

```
[1] 2.33
```

```
sd(x)
```

```
[1] 1.78901
```

```
sd(y)
```

```
[1] 2.002249
```

```
d=x-y
```

```
d
```

```
[1] -1.2 -2.4 -1.3 -1.3  0.0 -1.0 -1.8 -0.8 -4.6 -1.4
```

```
mean(d)
```

```
[1] -1.58
```

```
sd(d)
```

```
[1] 1.229995
```

```
n=length(d)
stat=mean(d)/(sd(d)/sqrt(n))
stat
```

```
[1] -4.062128
```

```
2*(1-pt(abs(stat),df=n-1))
```

```
[1] 0.00283289
```

- We can use `t.test()` on the single sample of differences:

```
t.test(d)
```

One Sample t-test

```
data: d
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of x
-1.58
```

- Note that the default

- ▶ null value for mu is 0,
- ▶ alternative is two-sided.

To be sure we could supply these

```
t.test(d, mu=0, alt="two.sided")
```

One Sample t-test

```
data: d
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of x
-1.58
```

- Note also that we can give the original two samples as arguments as long as we indicate they are paired:

```
t.test(x,y,paired=TRUE)
```

Paired t-test

```
data: x and y
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

- Note what happens if we forget to indicate paired=TRUE:

```
t.test(x,y)
```

Welch Two Sample t-test

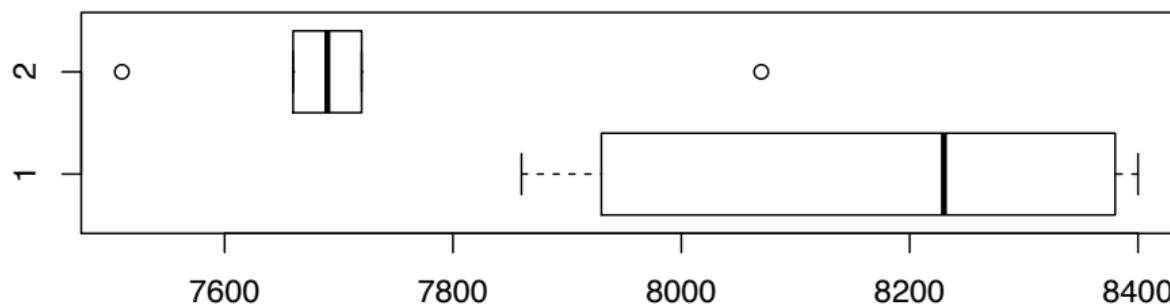
```
data: x and y
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.3654832 0.2054832
sample estimates:
mean of x mean of y
0.75      2.33
```

- It performs a Welch test! And look at the large p-value!
- This shows how important it is to use a paired test when it is appropriate.

Two Independent Samples

- Recall the mines example, where we performed a two *independent* sample *t*-test with the assumption of **equal variances**:

```
mine1=c(8400,8230,8380,7860,7930)
mine2=c(7510,7690,7720,8070,7660)
boxplot(mine1,mine2,horizontal=T)
```



- Note that these are both samples of size 5.

- The boxplots are dramatically different mainly because for mine1, the middle 3 values are close together; the ranges (max-min) are similar for both though:

```
max(mine1)-min(mine1)
```

```
[1] 540
```

```
max(mine2)-min(mine2)
```

```
[1] 560
```

```
meandiff=mean(mine1)-mean(mine2)  
meandiff
```

```
[1] 430
```

```
sp=sqrt( (4*var(mine1)+4*var(mine2))/8 )  
sp
```

```
[1] 230.3259
```

```
t.stat=meandiff/(sp*sqrt((1/5)+(1/5)))  
t.stat
```

```
[1] 2.95186
```

```
2*(1-pt(t.stat,df=8))
```

```
[1] 0.01837337
```

- To perform this using `t.test()` we need to indicate `var.equal=TRUE`:

```
t.test(mine1,mine2,var.equal=T)
```

Two Sample t-test

```
data: mine1 and mine2
t = 2.9519, df = 8, p-value = 0.01837
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 94.08239 765.91761
sample estimates:
mean of x mean of y
 8160      7730
```

- Changing the confidence level to 99% gives an interval including 0:

```
t.test(mine1,mine2,var.equal=T,conf.level=0.99)
```

Two Sample t-test

```
data: mine1 and mine2
t = 2.9519, df = 8, p-value = 0.01837
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
-58.78216 918.78216
sample estimates:
mean of x mean of y
8160      7730
```

- Notice what happens if we accidentally forget to indicate `var.equal=TRUE`:

```
t.test(mine1,mine2)
```

Welch Two Sample t-test

```
data: mine1 and mine2
t = 2.9519, df = 7.7039, p-value = 0.01916
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 91.82189 768.17811
sample estimates:
mean of x mean of y
 8160      7730
```

```
t.test(mine1,mine2,conf.level=0.99)
```

Welch Two Sample t-test

```
data: mine1 and mine2
t = 2.9519, df = 7.7039, p-value = 0.01916
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -64.29063 924.29063
sample estimates:
mean of x mean of y
 8160      7730
```

- The p-values are much the same.

- This is usually the case for these tests and this leads many to recommend that one should always use the Welch test, since it is best when the true population variances are different, and about the same as the classical two-sample t -test when they are equal.
- However note one other thing: the confidence intervals are wider for the Welch test. In some sense they are needlessly wide so there is still perhaps a good reason to use the classical test when it is ok to do so (i.e. if the boxplots indicate similar spread), although these differences become negligible for larger sample sizes.
- For this reason the default two-sample test in R is the Welch test.

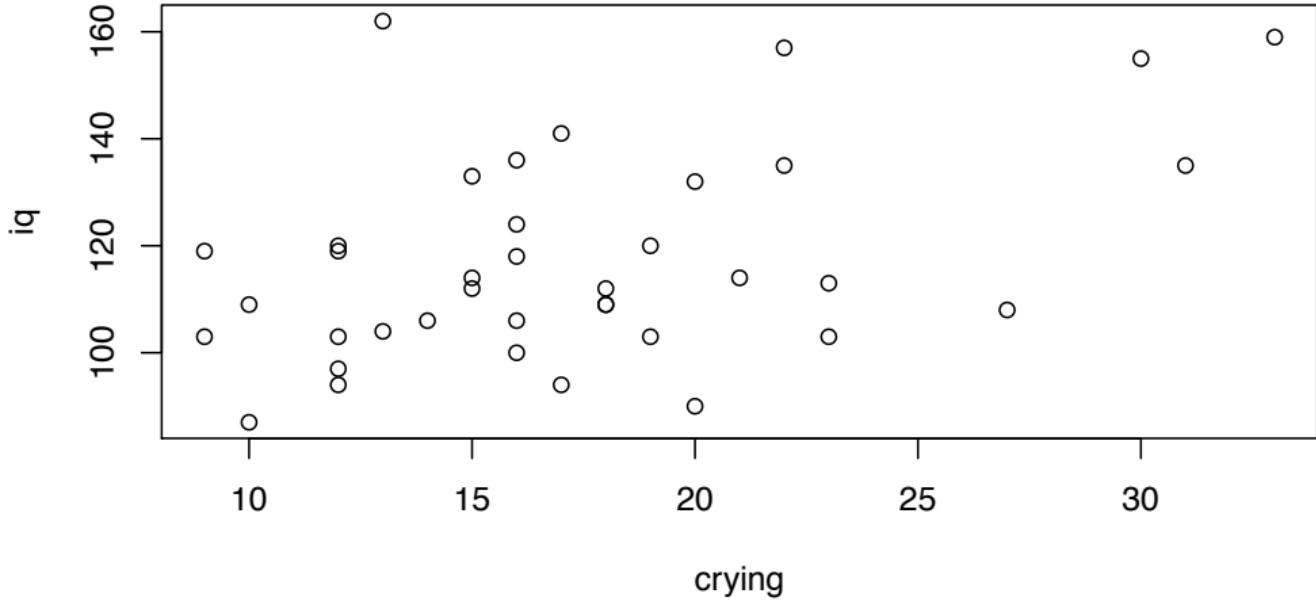
Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
 - Lecture 15
 - Lecture 16
 - Lecture 17
 - **Lecture 18**
 - Inference for linear regression models with normal errors
 - Derivations of distributions
 - Summary of inference for normal error linear regression
- 5 Inference Part 2: Discrete models

Inference for linear regression models with normal errors

- Babies who cry a lot immediately after birth may be more easily stimulated and hence have a higher IQ.
- To test this theory a study in the 1960's provoked 38 babies at around 4 days old to cry and measured the intensity of their crying (in some sense).
- Then their IQs were determined at age 3.

```
 crying <-  
 c(10, 20, 17, 12, 12, 16, 19, 12, 9, 23, 13, 14, 16, 27, 18,  
 10, 18, 15, 18, 23, 15, 21, 16, 9, 12, 12, 19, 16, 20, 15, 22,  
 31, 16, 17, 30, 22, 33, 13)  
 iq <-  
 c(87, 90, 94, 94, 97, 100, 103, 103, 103, 104, 106, 106,  
 108, 109, 109, 109, 112, 112, 113, 114, 114, 114, 118, 119, 119, 119,  
 120, 120, 124, 132, 133, 135, 135, 136, 141, 155, 157, 159, 162)  
 plot(crying,iq)
```



- The correlation here is

```
cor(crying, iq)
```

```
[1] 0.4549725
```

which the authors of the study claim is “significant”. Let us add the least-squares line to this plot.

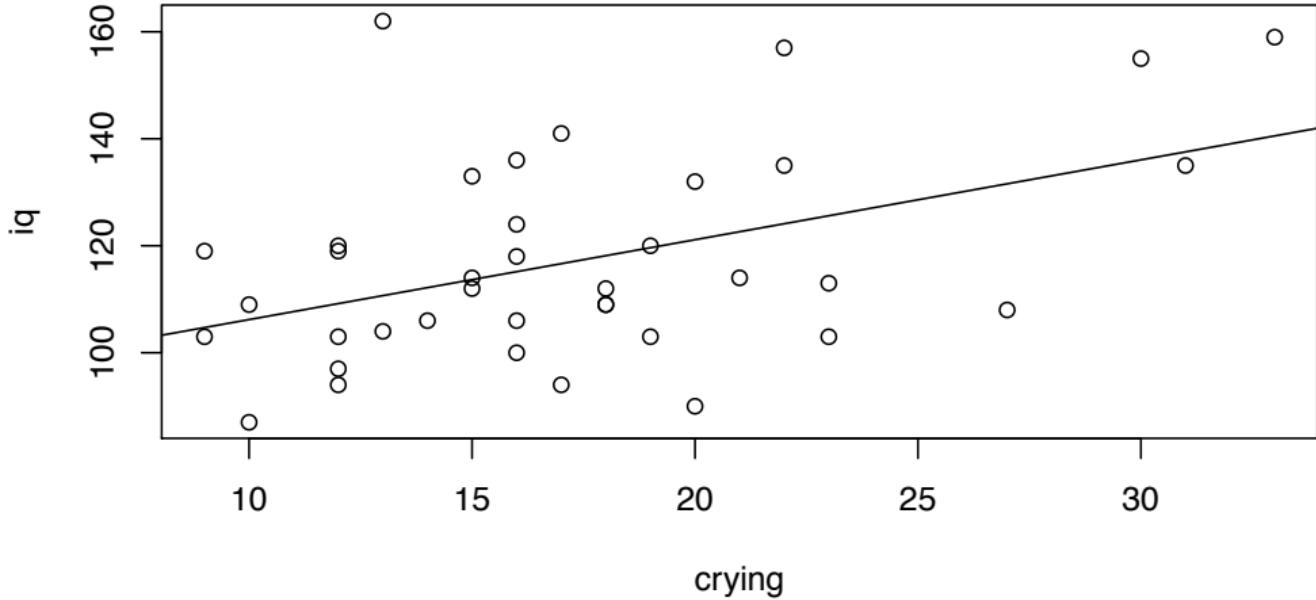
```
x=crying  
y=iq  
Sxx=sum((x-mean(x))^2)  
Sxy=sum((x-mean(x))*(y-mean(y)))  
b=Sxy/Sxx  
b
```

```
[1] 1.492897
```

```
a=mean(y)-b*mean(x)  
a
```

```
[1] 91.2683
```

```
plot(crying, iq)
```



- One way to interpret this as significant is to claim that the fitted slope 1.492897 is *significantly different from/greater than* zero (depending on if it is a one-sided or two-sided test).
- To make such a statement we need to model *something* as random variables here.
- The simplest such model is the following: the points are $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ where
 - ▶ x_1, \dots, x_n are **given constants** and
 - ▶ the Y_i 's are normal random variables with a **common variance** and whose **expectations** are linearly related to the x_i 's.
- Specifically, for some **unknown** constants α and β (and σ^2 !), for $i = 1, 2, \dots, n$,

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2).$$

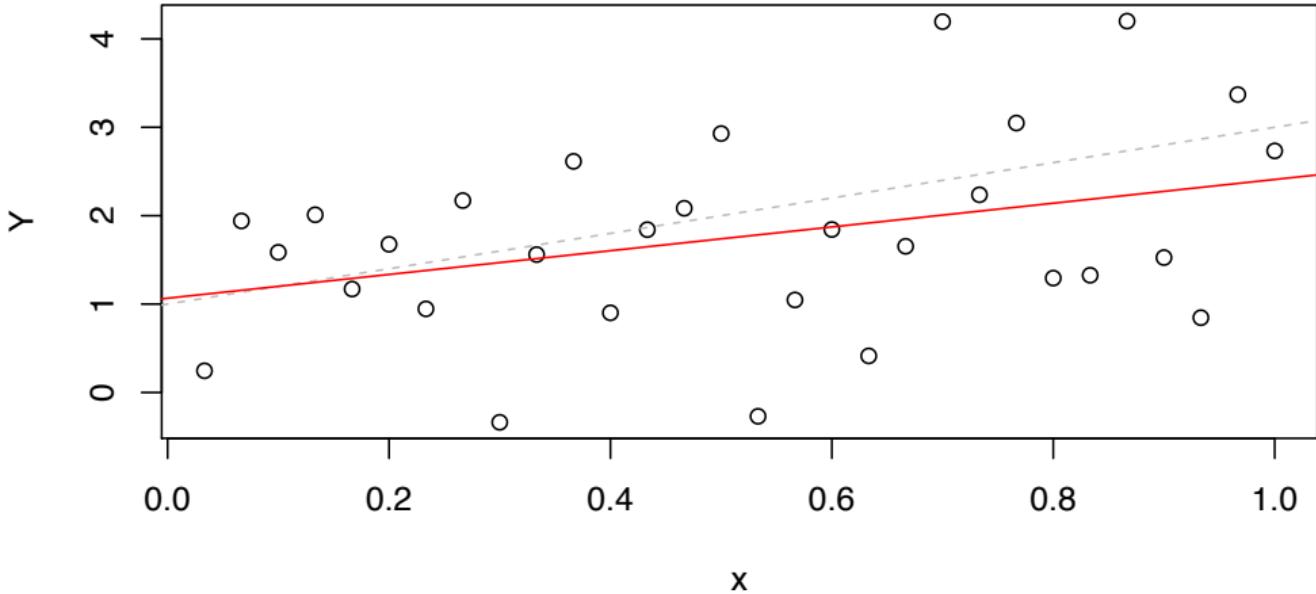
- Alternatively, if we define “errors” $\varepsilon_i = Y_i - (\alpha + \beta x_i)$ then $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent $N(0, \sigma^2)$ random variables and we can express each Y_i then as

$$Y_i = \alpha + \beta x_i + \varepsilon_i.$$

Simulations

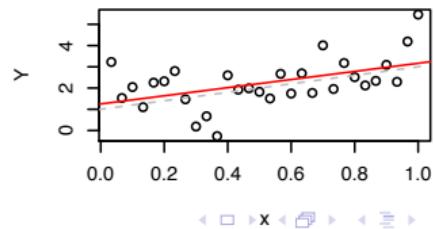
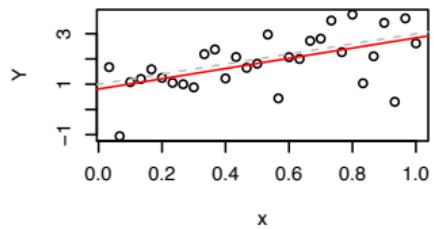
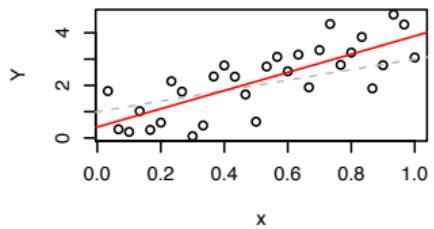
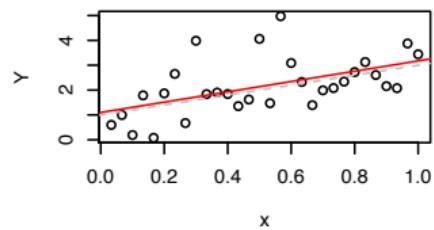
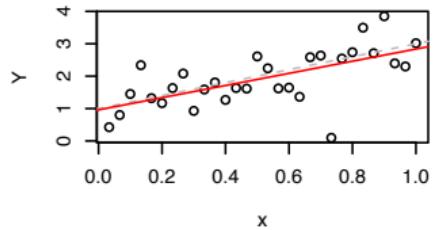
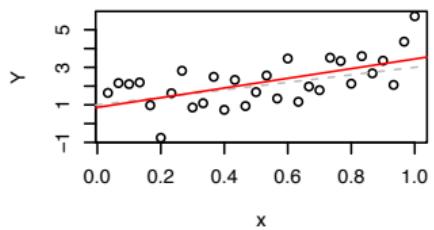
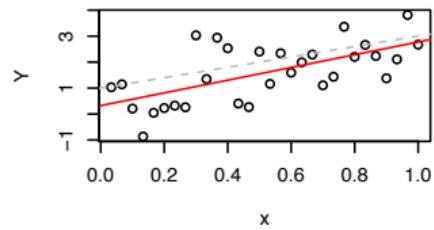
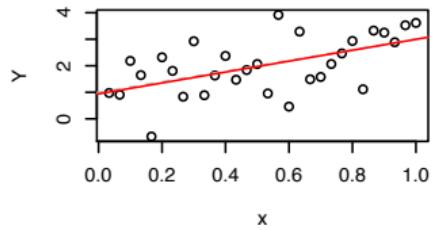
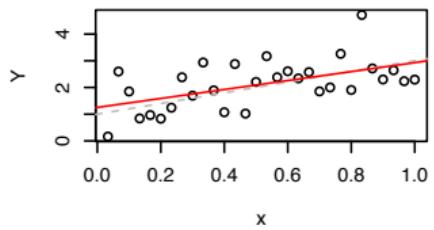
- Let us do some simulations under this model with normal errors and a convenient set of x_i 's

```
x=(1:30)/30
eps=rnorm(30)
Y=1+2*x+eps
plot(x,Y)
abline(1,2,lty=2,col="grey")
SxY=sum((x-mean(x))*(Y-mean(Y)))
Sxx=sum((x-mean(x))^2)
beta.hat=SxY/Sxx
alpha.hat=mean(Y)-beta.hat*mean(x)
abline(alpha.hat,beta.hat,col="red")
```



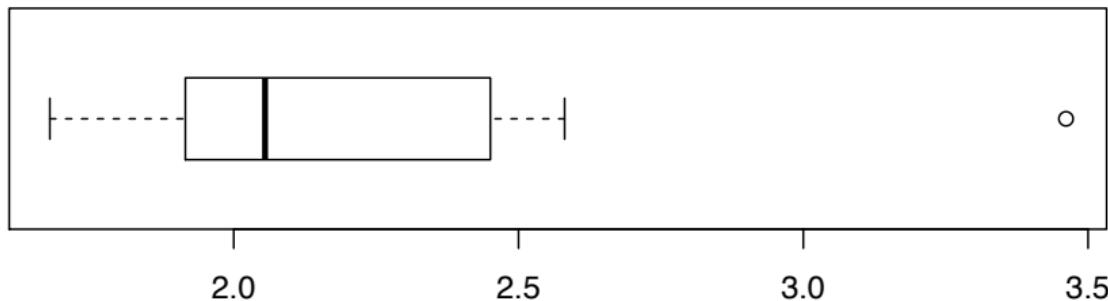
- Note here that the *true* regression line (grey dashed) is $y = 1 + 2x$ while the *fitted* or *estimated* regression line (red solid) is $y = 1.066481 + 1.343009 x$.
- Let's do 9 more like this

```
par(mfrow=c(3,3))
alpha.hat.vec=0
beta.hat.vec=0
for (i in 1:9){
  eps=rnorm(30)
  Y=1+2*x+eps
  SxY=sum((x-mean(x))*(Y-mean(Y)))
  Sxx=sum((x-mean(x))^2)
  beta.hat=SxY/Sxx
  alpha.hat=mean(Y)-beta.hat*mean(x)
  plot(x,Y)
  abline(1,2,col="grey",lty=2)
  abline(alpha.hat,beta.hat,col="red")
  alpha.hat.vec[i]=alpha.hat
  beta.hat.vec[i]=beta.hat
}
```

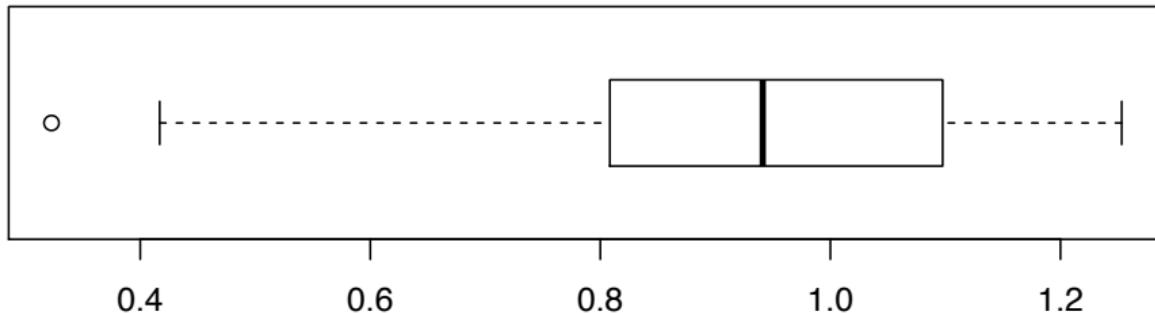


- So you can see that even when the *true* regression line is the same, for different sets of observations we get different estimates of the intercept and slope.

```
boxplot(beta.hat.vec, horizontal=T)
```



```
boxplot(alpha.hat.vec, horizontal=T)
```



- Note that the estimates are (roughly) centred on the corresponding true values.

- In fact in this model we can derive the **exact** distributions of the estimators of intercept and slope. We do so in the following section.
- Also there is another parameter which needs to be estimated, the *error variance* σ^2 .
- This is estimated using

$$\hat{\sigma}^2 = \frac{\text{Residual sum of squares}}{n - 2} = \frac{1}{n - 2} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

where $\hat{\varepsilon}_i = Y_i - (\hat{\alpha} + \hat{\beta}x_i)$ is the *i-th residual* (or estimated error).

Slope estimator $\hat{\beta}$

- The first step in deriving the distribution of the estimator of the slope is to realise we can write $\hat{\beta}$ as a **linear combination of the Y_i 's**.
- To see this write

$$\begin{aligned} S_{XY} &= \sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}) \\ &= \left[\sum_{i=1}^n (x_i - \bar{x}) Y_i \right] - \bar{Y} \sum_{i=1}^n (x_i - \bar{x}) \\ &= \sum_{i=1}^n (x_i - \bar{x}) Y_i \end{aligned}$$

since $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

- Then

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}} = \sum_{i=1}^n c_i Y_i$$

where $c_i = (x_i - \bar{x})/S_{xx}$.

- As a linear combination of independent normals, $\hat{\beta}$ is also normal. Its mean and variance are given by

$$E(\hat{\beta}) = E\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i E(Y_i) = \sum_{i=1}^n c_i (\alpha + \beta x_i) = \beta \sum_{i=1}^n c_i x_i$$

(since $\sum_{i=1}^n c_i = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) = 0$) and

$$Var(\hat{\beta}) = Var\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i^2 Var(Y_i) = \sigma^2 \sum_{i=1}^n c_i^2.$$

- All that remains is to determine the two sums $\sum_i c_i x_i$ and $\sum_i c_i^2$.
- Note though that (similarly to above with S_{XY}) we can rewrite

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^n x_i(x_i - \bar{x}) - \bar{x} \sum_{i=1}^n (x_i - \bar{x}) \\ &= \sum_{i=1}^n x_i(x_i - \bar{x}) \end{aligned}$$

again since $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

- So then

$$\sum_{i=1}^n x_i c_i = \sum_{i=1}^n x_i \frac{(x_i - \bar{x})}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^n x_i(x_i - \bar{x}) = \frac{S_{xx}}{S_{xx}} = 1.$$

- Also,

$$\sum_{i=1}^n c_i^2 = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_{xx}} \right)^2 = \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{S_{xx}}{S_{xx}^2} = \frac{1}{S_{xx}}.$$

Summary

- In summary $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \sigma^2/S_{xx}$, that is

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right).$$

Intercept estimator $\hat{\alpha}$

- Similar steps can be employed to derive the distribution of the intercept estimator $\hat{\alpha}$, by writing it as a linear combination of the Y_i 's:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = \frac{1}{n} \sum_{i=1}^n Y_i - \bar{x} \sum_{i=1}^n c_i Y_i = \sum_{i=1}^n \left[\frac{1}{n} - \bar{x}c_i \right] Y_i = \sum_{i=1}^n d_i Y_i$$

- Then using similar steps as for $\hat{\beta}$,

$$E(\hat{\alpha}) = \sum_{i=1}^n d_i E(Y_i) = \dots = \alpha$$

and

$$Var(\hat{\alpha}) = \sum_{i=1}^n d_i^2 Var(Y_i) = \dots = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

so that

$$\hat{\alpha} \sim N \left(\alpha, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] \right).$$

Error variance estimator $\hat{\sigma}^2$

- In this model we have *actual* (normally distributed) errors

$$\varepsilon_i = Y_i - (\alpha + \beta x_i) \sim N(0, \sigma^2)$$

and *residuals* (or estimated errors)

$$\hat{\varepsilon}_i = Y_i - (\hat{\alpha} + \hat{\beta} x_i).$$

- The sums of squares of the actual errors then has a χ_n^2 distribution (times σ^2):

$$\begin{aligned}\sum_{i=1}^n \varepsilon_i^2 &= \sum_{i=1}^n [Y_i - (\alpha + \beta x_i)]^2 = \sigma^2 \sum_{i=1}^n \left[\frac{Y_i - (\alpha + \beta x_i)}{\sigma} \right]^2 \\ &= \sigma^2 \sum_{i=1}^n Z_i^2 \sim \sigma^2 \chi_n^2\end{aligned}$$

since each $Z_i = [Y_i - (\alpha + \beta x_i)]/\sigma \sim N(0, 1)$.

- If we replace two unknown parameters with their estimators, giving the *residual sum of squares*, we **lose two degrees of freedom**:

$$\sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2 \sim \sigma^2 \chi_{n-2}^2.$$

- Compare this to the one-sample case: if X_1, \dots, X_n are independent $N(\mu, \sigma^2)$ then we have

$$\sum_{i=1}^n (X_i - \mu)^2 = \sigma^2 \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \sigma^2 \chi_n^2$$

since each $Z_i = (X_i - \mu)/\sigma \sim N(0, 1)$ while if we replace μ with its estimator \bar{X} we lose one degree of freedom:

$$\sum_{i=1}^n (X_i - \bar{X})^2 \sim \sigma^2 \chi_{n-1}^2.$$

- Thus we divide by $n - 2$ so $E(\hat{\sigma}^2) = \sigma^2$, its target:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2.$$

A t -test for the slope

- Thus we have

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

independently of

$$\hat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n-2}}{n-2}.$$

- Thus whatever the true value of β the random variable

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{S_{xx}}} = \underbrace{\frac{\hat{\beta} - \beta}{\sigma/\sqrt{S_{xx}}}}_{\sim N(0,1)} \Bigg/ \underbrace{\frac{\hat{\sigma}}{\sigma}}_{\sim \sqrt{\frac{\chi^2_{n-2}}{n-2}}} \sim t_{n-2}$$

- This can be used to construct a test statistic for testing the null hypothesis $H_0: \beta = \beta_0$ for any given null value β_0 .

- In particular for testing $H_0: \beta = 0$ (i.e. there is no linear relationship between the x_i 's and the $E(Y_i)$'s), the test statistic would be simply

$$T = \frac{\hat{\beta}}{\hat{\sigma}/\sqrt{S_{xx}}}.$$

- The denominator here is the (estimated) *standard error* of the estimate in the numerator.
- In the crying/IQ example we have the estimate

```

x=crying
y=iq
Sxx=sum((x-mean(x))^2)
Sxy=sum((x-mean(x))*(y-mean(y)))
b=Sxy/Sxx
b

```

[1] 1.492897

- The corresponding standard error is $\hat{\sigma}/\sqrt{S_{xx}}$. We can manually compute the residual sum of squares (and hence the estimated se) as follows:

```
a=mean(y)-b*mean(x)  
a
```

```
[1] 91.2683
```

```
resid=y-(a+b*x)  
sum(resid^2)
```

```
[1] 11023.39
```

```
sigma.hat=sqrt(sum(resid^2)/36)  
sigma.hat
```

```
[1] 17.49872
```

```
se=sigma.hat/sqrt(Sxx)  
se
```

```
[1] 0.4870011
```

- So the t -statistic for testing zero slope is

b/se

[1] 3.065489

- In a tutorial exercise we show that the observed value of this statistic is equal to

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is observed value of the sample correlation coefficient.

- So in fact we may perform a t -test of the null hypothesis of “zero slope” knowing only the value of the sample correlation coefficient r (and the number of points n).
- This makes sense since the t -statistic is independent of scale and the sample correlation coefficient is a “scale-free” version of the least-squares slope.

- We verify this in our example below, and compute a (one-sided) p-value:

```
r=cor(crying,iq)  
r
```

```
[1] 0.4549725
```

```
n=length(iq)  
n
```

```
[1] 38
```

```
stat=r*sqrt(n-2)/sqrt(1-r^2)  
stat
```

```
[1] 3.065489
```

```
1-pt(stat,df=n-2)
```

```
[1] 0.00205265
```

- More formally:
 - ▶ **Model:** $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ for $i = 1, 2, \dots, n$ are independent
 - ▶ **Hypotheses:** $H_0: \beta = 0$ vs $H_1: \beta > 0$ (the test is one-sided, since it presupposes IQ will increase with crying intensity)
 - ▶ **Test statistic:** $T = \hat{\beta}/(\hat{\sigma}/\sqrt{S_{xx}}) \sim t_{n-2}$ if H_0 true
 - ▶ **P-value:** Observed value of statistic is 3.065489 giving a p-value of $P(t_{36} \geq 3.065489) = 0.00205265$.
- This is a very small p-value (significant at the 1% level) indicating strong evidence against the null hypothesis of no linear relationship.
- This indirectly lends support to the idea there is some (linear) relationship between crying intensity at 4 days and IQ at 3 years.

Comment

- I hope that this was **not** a case of firstly identifying the children with high IQ and then going back to the crying data and choosing a funny way of measuring intensity in order to give a significant result; the original paper is not clear on this.

Using R

- All of this can be performed using the R function `lm()` (which stands for “linear model”).
- It uses R’s “formula” syntax:

```
fit=lm(iq~crying)
summary(fit)
```

```
Call:
lm(formula = iq ~ crying)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.126 -11.426 -2.126  10.860  51.324 

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)    
(Intercept)  91.268     8.934 10.216 0.0000000000035
crying       1.493     0.487   3.065    0.00411  
                                                        
(Intercept) ***
crying      ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.5 on 36 degrees of freedom
Multiple R-squared:  0.207, Adjusted R-squared:  0.185 
F-statistic: 9.397 on 1 and 36 DF,  p-value: 0.004105
```

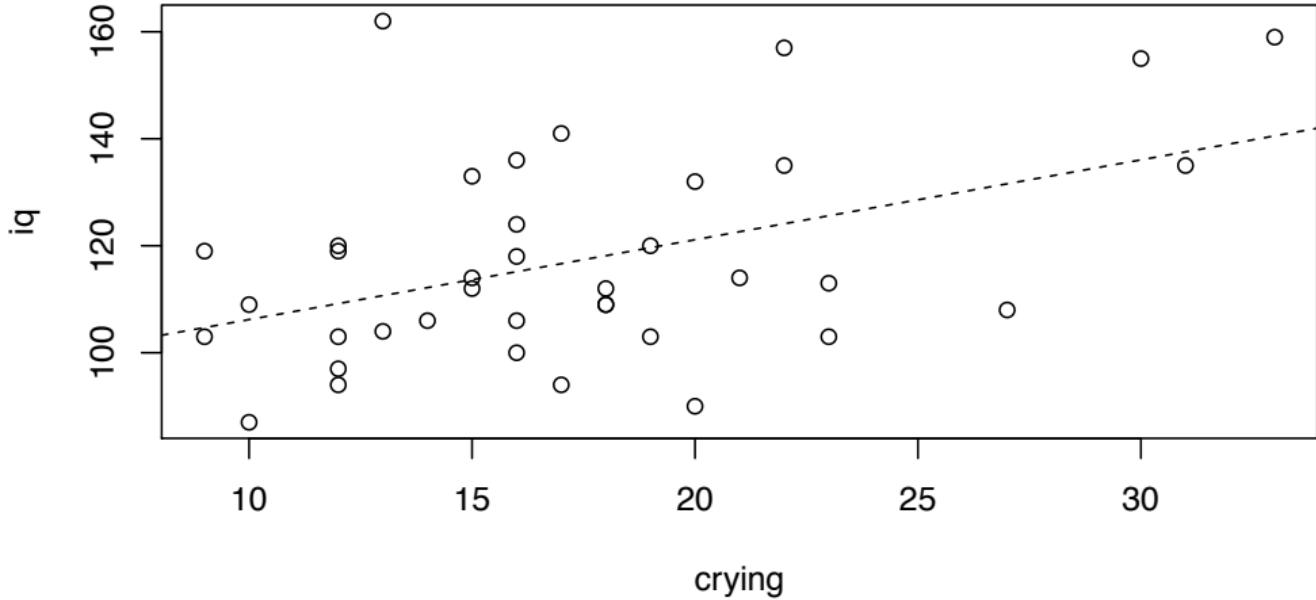
- Note in the output, the line titled “crying”. It has
 - ▶ an estimate (1.493), and
 - ▶ standard error (0.487),
 - ▶ t -statistic (3.07) and it also has a corresponding *two-sided* p-value (which is twice the one-sided p-value we computed directly).
- There is no facility for indicating a one- or two-sided test here.
- Note also the entry “Multiple R-squared”, this is precisely the square of the sample correlation coefficient and gives the proportion of the variance in the Y_i 's explained by the linear regression:

```
cor(crying,iq)^2
```

[1] 0.207

- Note also that using `abline(fit)` adds the least-squares regression line to the current plot:

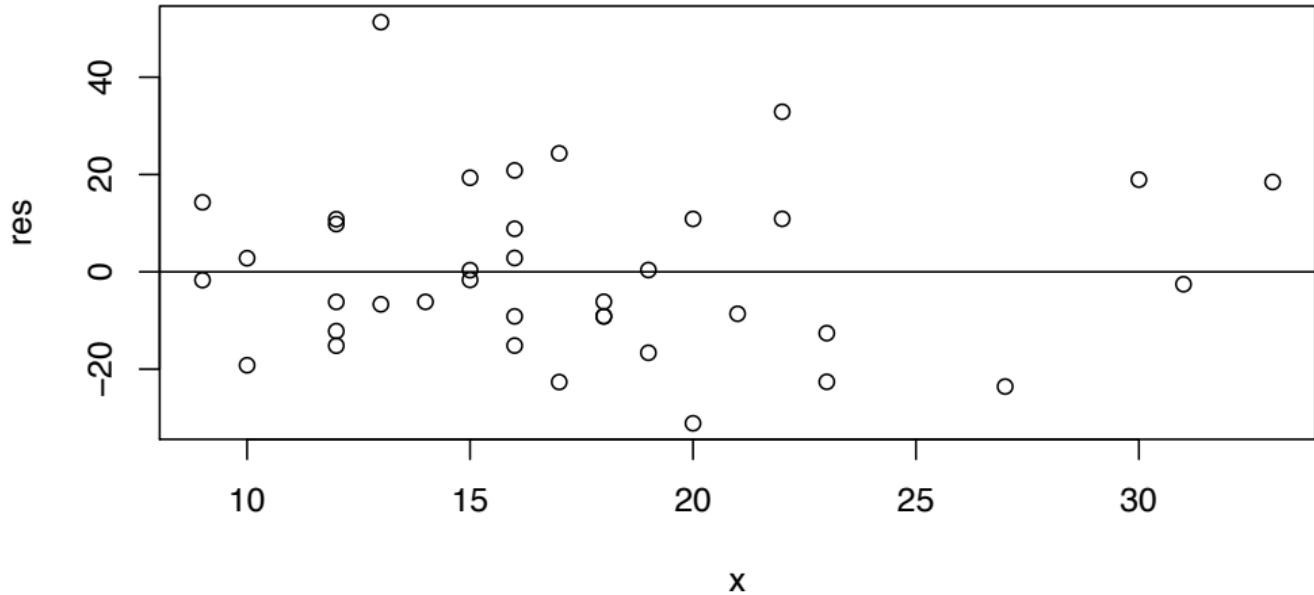
```
plot(crying,iq)
abline(fit,lty=2)
```



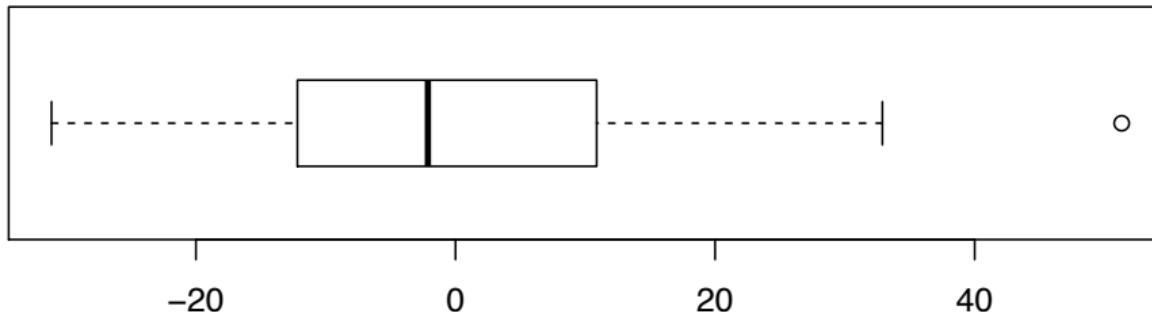
Model checking

- We could use these procedures on any set of points and get a p-value, confidence interval, etc.
- However the validity of these will depend on if the model assumptions are reasonable.
- Common qualitative checks include
 - ▶ plotting the residuals against x to see if there is any other *systematic* variability not captured by the linear regression
 - ▶ looking at a boxplot of residuals to assess (approximate) normality of these residuals.
- For the crying/IQ data we have

```
res=resid(fit)
plot(x,res)
abline(0,0)
```



```
boxplot(res, horizontal=T)
```



- There is one outlier (corresponding to the baby with the highest IQ, who had quite a low crying intensity).
- Aside from this there is no strong systematic pattern (e.g. curvature) in the residual vs x plot and the boxplot looks reasonably symmetric.
- There is no strong indication that the assumption of normality is not reasonable.

Summary of inference for normal error linear regression

- Points are modelled as $(x_1, Y_1), \dots, (x_n, Y_n)$ for
 - ▶ known constants x_1, \dots, x_n ;
 - ▶ each $Y_i = \alpha + \beta x_i + \varepsilon_i$ for
 - ★ $\varepsilon_1, \dots, \varepsilon_n$ independent $N(0, \sigma^2)$ "errors";
 - ★ unknown constants $\sigma^2 > 0$ (the error variance), α (the intercept) and β (the slope);

With

- ▶ $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$,
- ▶ $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ as usual,
- ▶ $S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$ (now a random variable) and
- ▶ $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ (also a random variable)

the estimates of the parameters are

- ▶ $\hat{\beta} = S_{xY}/S_{xx}$ for the slope;
- ▶ $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$ for the intercept;
- ▶ $\hat{\sigma}^2 = \frac{\text{Residual Sum of Squares}}{n-2} = \frac{\text{Total Sum of Squares} - \text{Regression Sum of Squares}}{n-2} = \frac{1}{n-2} \left(S_{YY} - \frac{S_{xY}^2}{S_{xx}} \right)$ for the error variance.

- The estimates are distributed as

- ▶ $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$;
- ▶ $\hat{\alpha} \sim N\left(\alpha, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$);
- ▶ $\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-2}/(n-2)$, that is $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.

- Standard Errors:

- ▶ $se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$;
- ▶ $se(\hat{\alpha}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$;

- Inference:

- ▶ for the slope is based on the pivot $\frac{\hat{\beta} - \beta}{\text{se}(\hat{\beta})} \sim t_{n-2}$ whatever the true value of β , that is
 - ★ to test $H_0: \beta = \beta_0$ use test statistic $\frac{\hat{\beta} - \beta_0}{\text{se}(\hat{\beta})} \sim t_{n-2}$ if H_0 true;
 - ★ for the special value $\beta_0 = 0$ the observed value of the test statistic equals $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ where r is the observed value of the sample correlation coefficient;
 - ★ for confidence intervals use $\hat{\beta} \pm c \text{se}(\hat{\beta})$ for appropriate "cut-off" c from the t_{n-2} distribution.
- ▶ for the intercept is based on the pivot $\frac{\hat{\alpha} - \alpha}{\text{se}(\hat{\alpha})} \sim t_{n-2}$ whatever the true value of α .

- Using R:

- ▶ `plot(x,y)`: produces scatterplot
- ▶ `fit=lm(y~x)`: computes “linear model” fit using y as the dependent variable
- ▶ `abline(fit)`: adds least-squares line to scatterplot
- ▶ `summary(fit)`: prints estimates, t-statistics, two-sided p-values etc.
- ▶ `res=resid(fit)`: extracts residuals
- ▶ `plot(x,res)`: plots residuals against x
- ▶ `boxplot(res)`: checks for approximate normality of residuals

- Model Checking:

- ▶ plot residuals against x ;
- ▶ look at boxplot of residuals.

Final comments on (inference for) regression

- We “interpret” the slope of the regression line as the amount of change we expect in the response variable (Y) for a change of 1 unit in the explanatory variable (x).
- A high correlation does not necessarily indicate *causation*.
- Two variables can be highly correlated without there being a causal relationship between them.
- An excellent example is between *shoe size* and *reading ability* in children; these tend to be highly correlated, however,
 - ▶ having big feet does not *make* you a good reader and conversely,
 - ▶ reading a lot does not make your feet grow faster.
- Of course the hidden/unobserved causal factor here is *age*; both shoe size and reading ability increase with age.

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models
 - Lecture 19
 - Inference concerning multinomial probabilities
 - Estimating parameters
 - Testing independence
 - Comparing multinomial vectors
 - Lecture 20
 - Lecture 21
 - Lecture 22

Motivating example

- When we roll a 6-sided die, if we think it is *fair* then we would expect (roughly) equal numbers of each possible result.
- If we roll the die 600 times we would expect to get 100 of each.
- Suppose that after 600 rolls the following *observed frequencies* are obtained:

Result	1	2	3	4	5	6
Obs Freq	111	78	101	100	113	97

- These are different from “all equal”.
- But are they **significantly** different from “all equal”?

Probability model

- Let us introduce a probability model for these data.
- Suppose that the probabilities of $1, 2, \dots, 6$ are given by p_1, p_2, \dots, p_6 with each $p_i \geq 0$ and $\sum_{i=1}^6 p_i = 1$.
- Note that although there are 6 parameters here, since they must obey the constraint $\sum_{i=1}^6 p_i = 1$ there are really only **5 free parameters**; once we know any 5 of them, we can deduce the 6th by subtraction.
- The observed frequencies (O_1, O_2, \dots, O_6) are then a **random vector** with a **multinomial** $(600; p_1, p_2, \dots, p_6)$ distribution, and we write $(O_1, O_2, \dots, O_6) \sim \text{Mult}(600; p_1, p_2, \dots, p_6)$ for short.

Multinomial distribution

- The multinomial is a generalisation of the binomial; if $X \sim B(n, p)$ then the vector $(X, n - X) \sim \text{Mult}(n; p, 1 - p)$ distribution or equivalently $(n - X, X) \sim \text{Mult}(n; 1 - p, p)$.
- Also, the *marginal* distribution of each O_i is $B(600, p_i)$; if we focus on, say, O_1 , the observed frequency of 1's and lump all the others into a second category of "not 1's", then we get

Result	Probability
1	p_1
not 1	$p_2 + p_3 + \cdots + p_6 = 1 - p_1$

- Equivalently, the random vector $(O_1, 600 - O_1) \sim \text{Mult}(600; p_1, 1 - p_1)$.

Testing a hypothesis

- What we really want to do is test the null hypothesis $H_0: p_1 = p_2 = \dots = p_6 = \frac{1}{6}$ against the alternative $H_1: \text{not } H_0$.
- A totally “unrestricted” alternative like this is indicative of a **goodness of fit** test, i.e. testing a certain null hypothesis H_0 against **any** alternative not covered by H_0 .

Pearson's test statistic

- A common test statistic for this test is **Pearson's Chi-Squared Statistic** (sometimes called "Pearson's X^2 "), given by

$$\sum_{i=1}^g \frac{(O_i - E_i)^2}{E_i},$$

observed frequency expected fq

where E_i is the "expected frequency" of group i under H_0 .

- If
 - ▶ there are g groups and
 - ▶ under H_0 the g probabilities are *completely specified*
- then when H_0 is true the statistic is (approximately, for large n) distributed as χ_{g-1}^2 .
- In our motivating example we have $g = 6$ groups, each with expected frequency 100. Thus here the statistic takes the value

$$\frac{11^2 + 22^2 + 1^2 + 0^2 + 13^2 + 3^2}{100} = \frac{121 + 484 + 1 + 0 + 169 + 9}{100} = \frac{784}{100} = 7.84.$$

- As with all hypothesis tests, the p-value is given as “the probability of at least as much evidence against H_0 , as was observed, assuming H_0 is in fact true”.
- Clearly, the **larger** the statistic, the more the observed frequencies O_i are disagreeing with the expected frequencies E_i .
- Thus an observation constituting *at least as much evidence against H_0* as this would be one with a statistic **at least as large as 7.84**.
- Since the statistic is (approximately) χ^2_5 when H_0 is true, the (approximate) p-value is given by

$$P(\chi^2_5 \geq 7.84).$$

Using R

- The (approximate) p-value is then given as follows:

```
1 - pchisq(7.84, df=5)
```

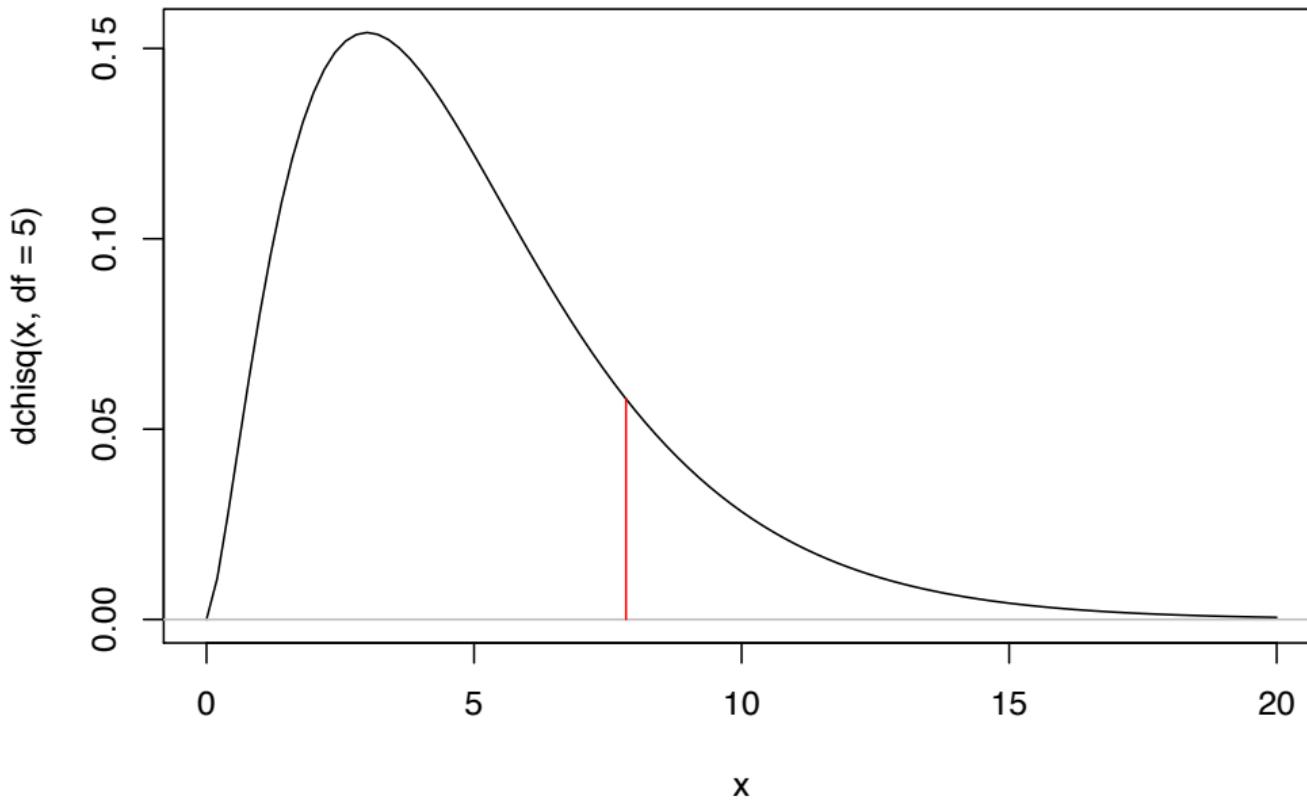
```
[1] 0.1652768
```

- We can visualise this as follows:

```
curve(dchisq(x, df=5), from=0, to=20,
       main="Chi-squared PDF with observed value in red")
abline(h=0, col="grey")
lines(c(7.84, 7.84), c(0, dchisq(7.84, df=5)), col="red")
```

- The area under the PDF to the right of the red line at 7.84 is the p-value.

Chi-squared 5 PDF with observed value in red



Using tables

- When a computer is not available, bounds can be obtained for the p-value by consulting certain “critical values” presented in chi-squared tables.
- On page 4 of the formula sheet for quizzes a limited set of such critical values is given.
- Note that the two values 6.626 and 8.115 satisfy
 - $P(\chi_5^2 \geq 6.626) \approx 0.25$ and
 - $P(\chi_5^2 \geq 8.115) \approx 0.15$.
- Since 7.84 is between these, we could conclude that the

$$0.15 \leq (\text{approx.}) \text{ p-value} = P(\chi_5^2 \geq 7.84) \leq 0.25 .$$

Conclusion

- This p-value is not small at all, about $1/6$; something that happens one-sixth of the time is not at all uncommon.
- There is thus no evidence against the claim that the die is fair.

Estimating parameters

- A generalisation of the above test is where
 - ▶ we have g groups,
 - ▶ under the alternative hypothesis the p_i 's are completely unspecified but
 - ▶ under the null hypothesis the p_i 's are **not completely specified** but depend on a certain number k of **free** parameters
- Then so long as the parameters are estimated “appropriately” (more on this later) then if the null hypothesis is true Pearson's statistic has an (approximate, for large n) χ^2_{g-k-1} distribution.

Motivating example

- There are various blood-typing systems.
- A certain gene on chromosome 4 has two (co-dominant) forms or *alleles*, labelled M and N.
- Each person has one copy of the gene from each parent.
- Thus a person could have MM, (both parents passed on M), NN (both parents passed on N) or MN (each parent passed on a different allele).
- “Co-dominant” means that there is a physical manifestation of *both* forms, i.e. a blood test can detect an M antigen and an N antigen.

Hardy-Weinberg equilibrium

- If a certain population is in (Hardy-Weinberg) equilibrium, then this process of passing on 0, 1 or 2 of a given allele (form N, say) is well-modelled by a binomial random variable.
- Loosely speaking, the “gene pool” is randomly mixed in such a way that a certain proportion p of the genes in the population are of form N, and each parent contributes their copy of the gene independently and $P(\text{passes on form N}) = p$ for both.
- It is known that this proportion, and the corresponding allele frequencies, differ greatly between different human populations, see for instance
http://www.mun.ca/biology/scarr/MN_bloodgroup.html.

- In 1937 a group of 1029 residents of Hong Kong had their MN blood type determined.

Type	Freq
MM	342
MN	500
NN	187

- Are these frequencies consistent with the population being in Hardy-Weinberg equilibrium?

- That is to say, are these consistent with probabilities of the form

Type	Prob
MM	$(1 - p)^2$
MN	$2p(1 - p)$
NN	p^2

for some $0 < p < 1$?

- One way to interpret this is to associate with each person a random variable X giving the number of N alleles, so this can be 0, 1 or 2, and model this as a $B(2, p)$ random variable;
- then
 - $P(MM) = P(X = 0) = (1 - p)^2$,
 - $P(MN) = P(X = 1) = 2p(1 - p)$ and
 - $P(NN) = P(X = 2) = p^2$.

Estimating the parameter p

- Before we can compute Pearson's statistic we need to determine expected frequencies under H_0 .
- Thus we need to find the value p so the resultant class probabilities $(1 - p)^2$, $2p(1 - p)$ and p^2 are “closest to the data” in some sense.
- We can treat this as a binomial estimation problem: each observation is recording the number of successes in two trials (i.e. how many of the two genes each person received from their parents is “N”).
- There are thus 2058 trials all together giving a total of 2058 copies of the gene, yielding $500+2*187=874$ in the form “N”.
- The overall proportion of form “N” is thus $874/2058 \approx 0.425$.

- Replacing p in the above table with the estimate $\hat{p} = 0.425$ yields the “fitted probabilities”

Type	Prob	Fitted Prob
MM	$(1 - p)^2$	0.3306
MN	$2p(1 - p)$	0.4888
NN	p^2	0.1806

- Multiplying this last column by the total frequency 1029 gives expected frequencies as follows:

Type	Prob	Fitted Prob	Exp Freq	Obs Freq
MM	$(1 - p)^2$	0.3306	340.21	342
MN	$2p(1 - p)$	0.4888	502.92	500
NN	p^2	0.1806	185.86	187

- Now these look remarkably close to the original observed frequencies.

- Indeed the value of Pearson's statistic is given in the R code below:

```
ObsF=c(342,500,187)  
ObsF
```

```
[1] 342 500 187
```

```
n=sum(ObsF)  
n
```

```
[1] 1029
```

```
p.hat=(500+2*187)/(2*n)  
p.hat
```

```
[1] 0.4246842
```

```
Fitted.Probs=dbinom(0:2,2,p.hat)  
Fitted.Probs
```

```
[1] 0.3309883 0.4886550 0.1803566
```

```
ExpF=Fitted.Probs*n  
ExpF
```

```
[1] 340.587 502.826 185.587
```

```
cbind(ObsF, ExpF)
```

	ObsF	ExpF
[1,]	342	340.587
[2,]	500	502.826
[3,]	187	185.587

```
stat = sum(((ObsF - ExpF)^2) / ExpF)  
stat
```

```
[1] 0.03250408
```

- This is (ridiculously) small, however the formal (approximate) p-value is given by

```
1 - pchisq(stat, df = 3 - 1 - 1)
```

```
[1] 0.8569258
```

- The p-value is then of course very large, indicating **no evidence at all** against the Hardy-Weinberg Equilibrium/Binomial null hypothesis.

What does “estimated appropriately” mean

- The distribution theory requires that any parameters estimated under H_0 are either
 - ▶ the values of the parameters that make the resultant Pearson statistic as *small as possible* (so-called “minimum chi-squared” estimators) or
 - ▶ are “close enough” to these in a particular sense.
- We shall not concern ourselves too much with this issue, only noting that the parameter estimates should always be functions of the multinomial observations themselves.

Testing goodness of fit of continuous distributions

- One pitfall in this area concerns using Pearson's test to test *goodness of fit of continuous distributions*.
- This is a somewhat outdated type of procedure but is still discussed in some books (an example involving the normal distribution appeared in the third edition of Phipps and Quine, although wisely this was removed from the fourth edition).
- The idea is to see whether a given sample of continuous data is well explained by a distribution from a parametric family, e.g. $N(\mu, \sigma^2)$ for some μ and σ .
- The basic approach is to
 - ▶ **somehow** divide the range of the data into intervals and count how many observations land in each interval; these are the *observed frequencies*;
 - ▶ estimate the parameters using the data;
 - ▶ compute probabilities and thus expected frequencies for each interval using the distribution corresponding to the estimated parameter values;
 - ▶ compute Pearson's statistic.

Various issues

- There are various issues with this general procedure: how exactly
 - ▶ are the intervals chosen (there is no real “natural” or “obvious” way to do this)?
 - ▶ are the parameters estimated?

Choosing intervals

- The χ^2 -theory really only applies if the intervals are in some sense “determined independently of the data”.
- This is rarely the case in practice, although the effect of choosing the intervals using the data is usually ignored.

Estimating parameters

- It is tempting to use **all of the data** to estimate the parameters, rather than just the counts in each interval.
- However doing so gives a **bigger** statistic than if we used the minimum chi-squared method.
- This then leads to smaller p-values, and thus “false significance” which is a big no-no.

Use with extreme caution

- While this procedure is still in use we only recommend it's use if
 - ▶ the intervals may be (essentially) chosen without reference to the data
 - ▶ *only the counts in the intervals* are used to estimate the parameters.
- We shall not discuss this topic any further.

How many degrees of freedom?

- There is a nice way to interpret the *degrees of freedom* in these chi-squared tests.
- In all cases, when the null hypothesis H_0 is true the statistic (approximately) has a chi-squared distribution with degrees of freedom equal to

(no. free parameters estimated under H_1) –
(no. free parameters estimated under H_0).

- Of course, when there are g groups and H_1 is “totally unconstrained probabilities” then there are $g - 1$ free parameters which are all estimated using the corresponding observed *proportions*, i.e. the observed frequencies divided by the total frequency.
- When the p_i 's are *totally determined* under H_0 (as in the fair dice example) there are zero free parameters estimated under H_0 and so the difference above is $g - 1$.
- If there are k free parameters estimated under H_0 then the difference is $(g - 1) - k$.

Motivating example

- Consider the following table of counts, which classify a sample of 229 individuals from a certain population according to two factors,
 - Blood Group at four levels: O , A , AB , B ;
 - Severity of Tuberculosis at 3 levels: not present, minimal, moderate-to-advanced.

	Blood Group	O	A	AB	B	Totals
Severity	Mod-Adv	7	5	3	13	28
	Minimal	27	32	8	18	85
	Not Present	55	50	7	4	116
	Totals	89	87	18	35	229

Full model

- We can model the 12 counts as a multinomial random vector (assuming this is an appropriately random sample).
- More specifically, the 12 categories each have their own probability:

	Blood Group	O	A	AB	B	Totals
Severity	Mod-Adv	p_{11}	p_{12}	p_{13}	p_{14}	r_1
	Minimal	p_{21}	p_{22}	p_{23}	p_{24}	r_2
	Not Present	p_{31}	p_{32}	p_{33}	p_{34}	r_3
	Totals	c_1	c_2	c_3	c_4	1

- Note that
 - ▶ each p_{ij} is precisely the probability that a **single** person drawn from this population will “land” in row i and column j ;
 - ▶ the row sum r_i gives the probability that a randomly drawn person will “land” in row i ;
 - ▶ the column sum c_j gives the probability that a randomly drawn person will “land” in column j ;
- This is the “full model”. Note that it has **11 free parameters**;
 - ▶ given any 11 of the p_{ij} ’s we can deduce the 12th, also the r_i ’s and c_j ’s are functions of the p_{ij} ’s.

Null hypothesis

- The real scientific question here is if there is any relationship between Blood Group and Tuberculosis Severity.
- We could make some kind of determination on this question by testing the null hypothesis that there is **no** relationship between them.
- In probability terms, “no relationship” could be formulated as

$$P(\text{in row } i \cap \text{ in column } j) = P(\text{in row } i)P(\text{in column } j),$$

that is

$$p_{ij} = r_i c_j$$

for each possible (i,j) combination.

- That is the row factors and column factors (as events) are *independent*.

- This gives a reduced model as follows:

	Blood Group	O	A	AB	B	Totals
Severity	Mod-Adv	$r_1 c_1$	$r_1 c_2$	$r_1 c_3$	$r_1 c_4$	r_1
	Minimal	$r_2 c_1$	$r_2 c_2$	$r_2 c_3$	$r_2 c_4$	r_2
	Not Present	$r_3 c_1$	$r_3 c_2$	$r_3 c_3$	$r_3 c_4$	r_3
	Totals	c_1	c_2	c_3	c_4	1

- This now all depends only on the r_i 's and c_j 's. But note also that these are not all free:
 - the r_i 's add to 1 so there are only 2 free parameters among the r_i 's;
 - the c_j 's also add to 1 so there are only 3 free parameters among the c_j 's;
- There are thus **5 free parameters** under this reduced model, i.e. this null hypothesis of independence.

Computing the Pearson statistic

- To compute the Pearson statistic we need estimates of the group probabilities under H_0 , which we then multiply by the total frequency to get expected frequencies.
- Thus we need estimates of the row and column probabilities:
 - ▶ the row probabilities (the r_i 's) are estimated by dividing each *row* total by the overall total;
 - ▶ the column probabilities (the c_j 's) are estimated by dividing each *column* total by the overall total
- The estimated cell *probabilities* are then the products of the corresponding estimated row and column probabilities.
- Thus an expression for the expected frequency of a given cell is

$$\begin{aligned} & \text{est. of cell prob} \times \text{total freq.} \\ &= \text{est. of row prob} \times \text{est. of col prob} \times \text{total freq.} \\ &= \frac{\text{row total}}{\text{total freq.}} \times \frac{\text{col total}}{\text{total freq.}} \times \text{total freq.} \\ &= \frac{\text{row total} \times \text{col total}}{\text{total freq.}}, \end{aligned}$$

- Suppose the counts are entered into a matrix in R:

```
r1=c(7,5,3,13)
r2=c(27,32,8,18)
r3=c(55,50,7,4)
OF=rbind(r1,r2,r3)
OF
```

```
[,1] [,2] [,3] [,4]
r1    7      5      3     13
r2   27     32      8     18
r3   55     50      7      4
```

```
dimnames(OF)[[2]]=c("c1","c2","c3","c4")
OF
```

```
c1 c2 c3 c4
r1 7 5 3 13
r2 27 32 8 18
r3 55 50 7 4
```

- Then the row and column sums can be found using `apply()`:

```
row.sums=apply(OF,1,sum)  
row.sums
```

```
r1  r2  r3  
28  85  116
```

```
col.sums=apply(OF,2,sum)  
col.sums
```

```
c1  c2  c3  c4  
89  87  18  35
```

```
n=sum(OF)  
n
```

```
[1] 229
```

- The statistic can then be computed using `outer()`:

```
outer(row.sums,col.sums)
```

```
c1   c2   c3   c4  
r1 2492 2436 504 980  
r2 7565 7395 1530 2975  
r3 10324 10092 2088 4060
```

```
EF=outer(row.sums,col.sums)/n  
EF
```

```
c1      c2      c3      c4  
r1 10.88210 10.63755 2.200873 4.279476  
r2 33.03493 32.29258 6.681223 12.991266  
r3 45.08297 44.06987 9.117904 17.729258
```

```
((OF-EF)^2)/EF
```

```
c1      c2      c3      c4  
r1 1.384905 2.987718789 0.2901591 17.770292  
r2 1.102482 0.002650794 0.2603077 1.931098  
r3 2.181478 0.797970462 0.4919461 10.631721
```

```
stat=sum(((OF-EF)^2)/EF)  
stat
```

```
[1] 39.83273
```

Degrees of freedom and p-value

- According to the general rule developed above, we have
 - ▶ 11 free parameters under the full model and
 - ▶ 5 free parameters under the null hypothesis
- Thus the test statistic should have (approximately) a chi-squared distribution with $11-5=6$ degrees of freedom if the null hypothesis is true.
- In this case then the (approximate) p-value is given by

```
1 - pchisq(stat, df = 6)
```

```
[1] 0.0000004913225
```

which is very very small indeed, thus there is very strong evidence **against** the null hypothesis of independence between Blood Group and Tuberculosis Severity.

- This *indirectly* suggests a connection between them.

A general formula for the degrees of freedom

- It is not hard to generalise this to a general two-way table with r rows and c columns.
 - ▶ Under the full model there are rc cells in the table and thus $rc - 1$ free parameters.
 - ▶ Under the null hypothesis there are $(r - 1) + (c - 1)$ free parameters: $r - 1$ for rows and $c - 1$ for columns.
 - ▶ Thus in general the difference is

$$rc - 1 - (r - 1) - (c - 1) = rc - r - c + 1 = (r - 1)(c - 1).$$

- Thus in a general r -by- c test of independence, the test statistic (approximately) has a χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom under the null hypothesis.

Test of Homogeneity

The use of Pearson's statistic can be extended beyond a test concerning a single multinomial vector.

Motivating example

- A study comparing the Canadian and US heart attack patients involved a random sample of 400 Canadian patients and 2600 US patients from a certain register.
- Various observations/measurements were made/taken on each, and follow-up questions were asked up to a year after their heart attack.
- One such question was the patient's own assessment of their quality of life one year later.

- The Canadian responses are below:

Quality of Life	Frequency
Much better	75
Somewhat better	71
About the same	96
Somewhat worse	50
Much worse	19
<hr/>	<hr/>
Total	311

- Here are the US responses:

Quality of Life	Frequency
Much better	541
Somewhat better	498
About the same	779
Somewhat worse	282
Much worse	65
<hr/>	<hr/>
Total	2165

- Note that not all patients have a response included here.

- These two sets of responses can each be modelled as multinomial random vectors.
- If we convert these into proportions, we get

Quality of Life	Canada	US
Much better	0.241	0.250
Somewhat better	0.228	0.230
About the same	0.309	0.360
Somewhat worse	0.161	0.130
Much worse	0.061	0.030
<hr/>	<hr/>	<hr/>
Total	1	1

- The scientific question is whether the apparent differences between these two vectors of estimated probabilities are *significantly different*.
- To do so we test the null hypothesis that there are **no differences** between the two. This is called a **test of homogeneity**.

Full Model

- The full model is that the two multinomial vectors have possibly different probabilities:

Quality of Life	Canada	US
Much better	p_1	q_1
Somewhat better	p_2	q_2
About the same	p_3	q_3
Somewhat worse	p_4	q_4
Much worse	p_5	q_5
—	—	—
Total	1	1

- Note that there are **8 free parameters here**: 4 free parameters among the p_i 's and 4 among the q_i 's.

The null hypothesis

The null hypothesis is that the two vectors of probabilities are the same:

Quality of Life	Canada	US
Much better	p_{01}	p_{01}
Somewhat better	p_{02}	p_{02}
About the same	p_{03}	p_{03}
Somewhat worse	p_{04}	p_{04}
Much worse	p_{05}	p_{05}
Total	1	1

There are only **4 free parameters here.**

8-4

Computing the Pearson statistic

- As usual, we need expected frequencies (and thus estimated probabilities) under H_0 before we can compute the Pearson statistic.
- In this case we just need to estimate the common vector of probabilities, p_{01}, \dots, p_{05} .
- A natural way to do this is to pool all the observations and then convert the (pooled) frequencies into proportions:

Pooled Data

Quality of Life	Frequency	Proportion
Much better	616	0.249
Somewhat better	569	0.230
About the same	875	0.353
Somewhat worse	332	0.134
Much worse	84	0.034
<hr/>		
Total	2476	1

- We then multiply this vector of estimated probabilities (i.e. proportions) by the total frequency for both Canada and then the US to get expected frequencies:

Expected Frequencies under Homogeneity

Canada

Quality of Life	Obs Freq	Exp Freq
Much better	75	77.37
Somewhat better	71	71.47
About the same	96	109.91
Somewhat worse	50	41.70
Much worse	19	10.55
<hr/>	<hr/>	<hr/>
Total	311	311.00

US

Quality of Life	Obs Freq	Exp Freq
Much better	541	538.63
Somewhat better	498	497.53
About the same	779	765.09
Somewhat worse	282	290.30
Much worse	65	73.45
<hr/>	<hr/>	<hr/>
Total	2165	2165.00

Two-way table format

- Such comparisons between multinomial vectors are often also represented as a two-way table (for good reason, as we shall see):

Quality of Life	Canada	US	Totals
	Obs (Exp)	Obs (Exp)	
Much better	75 (77.37)	541 (538.63)	616
Somewhat better	71 (71.47)	498 (497.53)	569
About the same	96 (109.91)	779 (765.09)	875
Somewhat worse	50 (41.70)	282 (290.30)	332
Much worse	19 (10.55)	65 (73.45)	84
Totals	311	2165	2476

- Note that again, the expected frequency in each cell is given by

$$\frac{(\text{row sum}) \times (\text{column sum})}{\text{total freq.}},$$

just as it was for the test of independence.

- We may thus use the same methods for computing the statistic as we did for the test of independence.
- The code below computes the statistic:

```
can=c(75,71,96,50,19)
us=c(541,498,779,282,65)

OF=cbind(can,us)
OF
```

```
can us
[1,] 75 541
[2,] 71 498
[3,] 96 779
[4,] 50 282
[5,] 19 65
```

```
row.sums=can+us
col.sums=c(sum(can),sum(us))
row.sums
```

```
[1] 616 569 875 332 84
```

```
col.sums
```

```
[1] 311 2165
```

```
EF=outer(row.sums,col.sums)/sum(OF)
round(EF,dig=3)
```

```
 [,1]    [,2]
[1,] 77.373 538.627
[2,] 71.470 497.530
[3,] 109.905 765.095
[4,] 41.701 290.299
[5,] 10.551  73.449
```

```
contributions=((OF-EF)^2)/EF
round(contributions,dig=3)
```

```
can    us
[1,] 0.073 0.010
[2,] 0.003 0.000
[3,] 1.759 0.253
[4,] 1.652 0.237
[5,] 6.766 0.972
```

```
stat=sum(contributions)
round(stat,dig=3)
```

- We have noted that there are 8 free parameters under the full model and 4 under the null hypothesis, giving thus $8-4=4$ degrees of freedom for the test.
- The (approximate) p-value is thus $P(\chi_4^2 \geq 11.725)$

which is

```
round(1-pchisq(stat ,df=4) ,dig=5)
```

```
[1] 0.01951
```

General formula for the degrees of freedom

- But the similarities between the test of homogeneity and the test of independence do not end with the formula for expected frequencies.
- Suppose we have c different multinomial random vectors, each of which has r categories (above we had $c = 2$ different multinomials, each with $r = 5$ categories).
- Then the full model has c different vectors of r probabilities; each vector has $r - 1$ free parameters, and so there are $c(r - 1)$ free parameters in total.
- The null hypothesis, on the other hand, only has a *single* vector of r probabilities, giving thus only $r - 1$ free parameters. r-row c-column
- Thus the difference is $c(r - 1) - (r - 1) = (r - 1)(c - 1)$ degrees of freedom for the test, exactly the same as the test of independence!

Conceptually very different; operationally identical

- Thus the two different kinds of chi-squared tests
 - ▶ test of independence
 - ▶ test of homogeneity

are conceptually very different:

- ▶ a test of independence applies when we have a random sample from a **single** population, and each individual is classified according to two factors, one with r levels, another with c levels;
 - ▶ a test of homogeneity applies when we have a random sample from each of c populations, and each individual is classified only according to a single factor with r levels.
- However it turns out that to perform each test identical steps are involved in both cases.

The R function `chisq.test()`

- We have thus considered four different kinds of chi-squared tests:
 - ① a one-way layout with probabilities completely determined under H_0 (i.e. goodness-of-fit, no parameters estimated);
 - ② a one-way layout with probabilities *not* completely determined under H_0 , but depending on k free parameters (i.e. goodness-of-fit with parameters estimated);
 - ③ a two-way layout test of independence;
 - ④ a two-way layout test of homogeneity.
- It turns out that 3 of these 4 types of chi-squared test can be performed using the R function `chisq.test()`.

Goodness of fit, no parameters estimated

- For type 1. above, the syntax is of the form `chisq.test(x,p...)`= where
 - ▶ `x` is a vector of counts
 - ▶ `p` is an (optional) vector of probabilities.
- For the dice example above,

```
x=c(111,78,101,100,113,97)
chisq.test(x)
```

Chi-squared test for given probabilities

```
data: x
X-squared = 7.84, df = 5, p-value = 0.1653
```

```
probs=rep(1/6,6)  
probs
```

```
[1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667  
[6] 0.1666667
```

```
chisq.test(x,p=probs)
```

Chi-squared test for given probabilities

```
data: x  
X-squared = 7.84, df = 5, p-value = 0.1653
```

- For types 3 and 4 (since these are operationally identical), the syntax is of the form `chisq.test(x)` where `x` is a matrix.

Test of independence

- For the test of independence example above:

```
r1=c(7,5,3,13)
r2=c(27,32,8,18)
r3=c(55,50,7,4)
OF=rbind(r1,r2,r3)
OF
```

	[,1]	[,2]	[,3]	[,4]
r1	7	5	3	13
r2	27	32	8	18
r3	55	50	7	4

```
print(chisq.test(OF))
```

Pearson's Chi-squared test

data: OF

X-squared = 39.833, df = 6, p-value = 0.0000004913

Warning message:

In chisq.test(OF) : Chi-squared approximation may be incorrect

- The warning is issued here because when there are some small expected frequencies, which can be seen explicitly as follows, the chi-squared approximation is possibly not accurate:

```
chisq.test(OF)$expected
```

	[,1]	[,2]	[,3]	[,4]
r1	10.88210	10.63755	2.200873	4.279476
r2	33.03493	32.29258	6.681223	12.991266
r3	45.08297	44.06987	9.117904	17.729258

Warning message:

In chisq.test(OF) : Chi-squared approximation may be incorrect

Two of these are less than 5 which flags the warning.

- In such a case an alternative method of approximating a p-value can be used, based on simulation:

```
chisq.test(OF, simulate.p.value=TRUE)
```

```
Pearson's Chi-squared test with simulated p-value  
(based on 2000 replicates)
```

```
data: OF  
X-squared = 39.833, df = NA, p-value = 0.0004998
```

- This returns a higher p-value, in fact since the simulation uses 2000 repetitions by default, this indicates that only 1 of the 2000 gave a higher statistic.

- Let's try with a higher number of repetitions:

```
chisq.test(OF, simulate.p.value=TRUE, B=20000)
```

Pearson's Chi-squared test with simulated p-value
(based on 20000 replicates)

```
data: OF
X-squared = 39.833, df = NA, p-value = 0.00005
```

or even

```
chisq.test(OF, simulate.p.value=TRUE, B=1000000)
```

Pearson's Chi-squared test with simulated p-value
(based on 1000000 replicates)

```
data: OF
X-squared = 39.833, df = NA, p-value = 0.000003
```

- In any case the p-value is very small, and indicates strong evidence against the hypothesis of independence between blood group and tuberculosis severity.

Test of homogeneity

For the test of homogeneity example above:

```
can=c(75,71,96,50,19)
us=c(541,498,779,282,65)
OF=cbind(can,us)
OF
```

```
can  us
[1,] 75 541
[2,] 71 498
[3,] 96 779
[4,] 50 282
[5,] 19 65
```

```
chisq.test(OF)
```

Pearson's Chi-squared test

```
data: OF
X-squared = 11.725, df = 4, p-value = 0.01951
```

Goodness of fit, parameters estimated

- For type 2 we need to perform the test manually, or at least compute the estimated cell probabilities manually;
 - `chisq.test()` will then compute the test statistic properly, but the degrees of freedom, and thus the returned p-value, will be wrong.
- Recall our earlier computations:

```
ObsF=c(342,500,187)  
ObsF
```

```
[1] 342 500 187
```

```
n=sum(ObsF)  
n
```

```
[1] 1029
```

```
p.hat=(500+2*187)/(2*n)  
p.hat
```

```
[1] 0.4246842
```

```
Fitted.Probs=dbinom(0:2,2,p.hat)
Fitted.Probs
```

```
[1] 0.3309883 0.4886550 0.1803566
```

```
ExpF=Fitted.Probs*n
ExpF
```

```
[1] 340.587 502.826 185.587
```

```
cbind(ObsF,ExpF)
```

```
ObsF      ExpF
[1,] 342 340.587
[2,] 500 502.826
[3,] 187 185.587
```

```
stat=sum(((ObsF-ExpF)^2)/ExpF)
stat
```

```
[1] 0.03250408
```

```
1-pchisq(stat,df=3-1-1)
```

```
[1] 0.8569258
```

- Note what happens if we try to use `chisq.test()` with our estimated `Fitted.Probs`:

```
chisq.test(ObsF, p=Fitted.Probs)
```

Chi-squared test for given probabilities

```
data: ObsF  
X-squared = 0.032504, df = 2, p-value = 0.9839
```

- We get the correct value of the statistic, but the wrong p-value.
- In such a case, one can extract the value of the statistic:

```
st=chisq.test(ObsF, p=Fitted.Probs)$stat  
st
```

X-squared
0.03250408

and then use this to compute the p-value manually, if this is easier than a direct manual calculation.

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models
 - Lecture 19
 - Lecture 20
 - Inference concerning a binomial success probability
 - Motivating examples
 - Hypothesis tests
 - One-sided tests
 - Two-sided tests
 - Using Pearson's statistic
 - Exact p-values using Pearson's statistic
 - The likelihood ratio test
 - Sign test
 - Lecture 21
 - Lecture 22

Inference concerning a binomial success probability

- In various applications we can model a single observed count x as the observed value of a binomial random variable $X \sim B(n, q)$ ¹ for some known integer n and unknown “success probability” parameter $0 \leq q \leq 1$.
- In such cases the parameter q has an interpretation and some inference concerning it is desired.
- Consider the following examples.

¹We use q instead of p for the binomial success probability parameter here to avoid confusion with p -values

1,2 are one sided

- ① A coin is suspected of favouring heads. It is to be flipped **20** times.
- ② It was known that 10% of teenagers in a certain population smoked. After an aggressive public awareness campaign concerning the dangers of smoking designed to reduce the proportion of teenagers smoking a random sample of **256** such teenagers was taken and it was determined what proportion of them smoked.
- ③ A coin is to be *tested for fairness* by flipping it **20** times.
- ④ A new political party scored 10% of the vote at the last election. *Has the level of support changed?* A random sample of **256** is to be taken from a large population of voters (an “opinion poll”).

3,4 not a clear direction

- We have deliberately **not** stated the outcomes here, since we should be able to specify a procedure **before we see the data**.
- This approach guarantees that, in particular, we will not mistakenly use a one-sided test where a two-sided test should be used.
- The *phrases emphasised thusly* in the last two examples are the key indicators that a two-sided approach is required.
- However for these examples to be of use we need some tangible outcomes to work with. Thus, suppose that respectively:
 - ① 15 (i.e. 75%) heads and 5 (i.e. 25%) tails are obtained.
 - ② 16 of the 256 (i.e. 6.25%) are defective.
 - ③ 15 (i.e. 75%) heads and 5 (i.e. 25%) tails are obtained.
 - ④ 16 of the 256 (i.e. 6.25%) support the new party.

Hypothesis tests

- The general setup is that $H_0: q = q_0$ for some known $0 < q_0 < 1$.
- In the first two examples above only deviations in one particular direction are of interest, or are anticipated.
- For the last two, deviations in either direction are potentially of interest.

- Thus the null and alternative hypotheses for the four examples are
 - ① $H_0: q = \frac{1}{2}$ against $H_1: q > \frac{1}{2}$.
 - ② $H_0: q = 0.1$ against $H_1: q < 0.1$.
 - ③ $H_0: q = \frac{1}{2}$ against $H_1: q \neq \frac{1}{2}$.
 - ④ $H_0: q = 0.1$ against $H_1: q \neq 0.1$.
- Thus examples 1 and 2 are *one-sided* while 3. and 4. are *two-sided*.

\$P\\$-values

- As with all hypothesis tests, the p-value is the probability (under H_0) of at least as much evidence against H_0 as was observed.
- Thus the process of assigning a *p*-value to an observation x has two steps:
 - ▶ identify a set S_x consisting of all possible observations constituting *at least as much evidence against H_0 as x* , and then
 - ▶ determine $P(S_x)$ when H_0 is true.

One-sided tests

There are two cases:

$$H_1: q > q_0$$

- If the alternative is $q > q_0$, then the *larger* the observation the more evidence against H_0 it constitutes.
- Thus the set S_x referred to above is $\{x, x + 1, \dots, n\}$ and so the p -value is

$$P(S_x) = P(X \geq x) \text{ assuming } X \sim B(n, q_0).$$

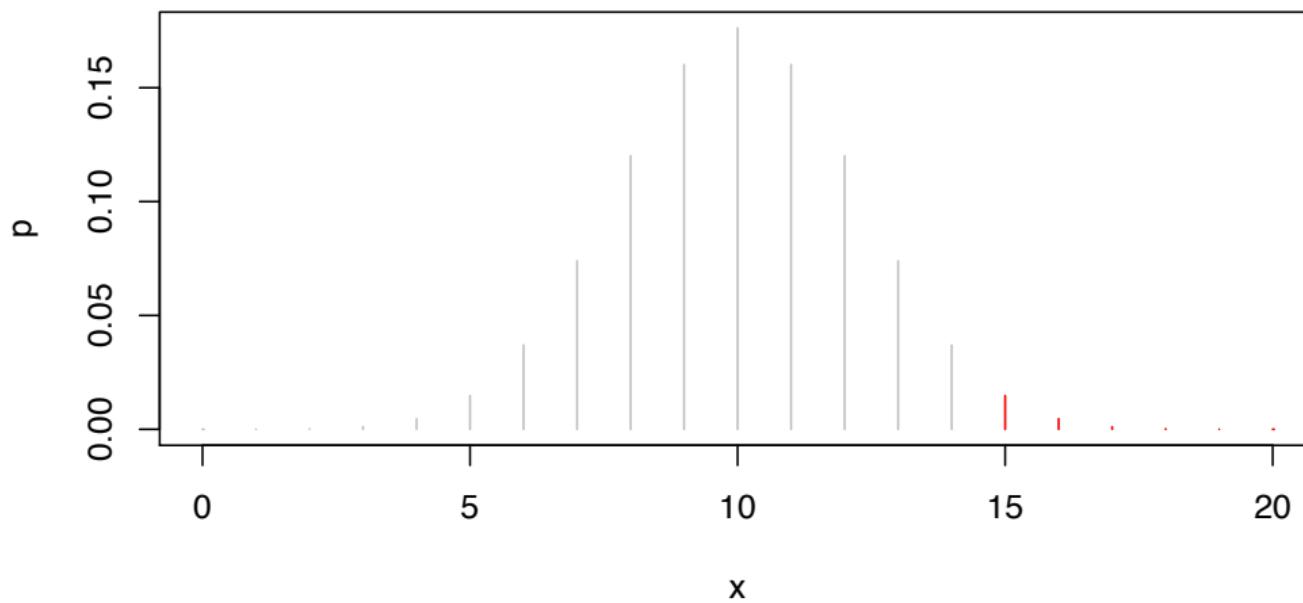
- In the example above where 15 heads are observed, the observations constituting *at least* as much evidence are 15 heads, 16 heads, 17 heads, ..., 20 heads.
- Thus there the p -value is $P(X \geq 15)$ when $X \sim B(20, 0.5)$ which is $1 - P(X \leq 14)$ and is obtained in R as follows:

```
1-pbinom(14,20,.5)
```

```
[1] 0.02069473
```

This is presented graphically below:

```
x=0:20  
p=dbinom(x,20,0.5)  
plot(x,p,type="h",col="grey")  
lines(x[16:21],p[16:21],col="red",type="h")
```



$$H_1: q < q_0$$

- If the alternative is $q < q_0$, then the *smaller* the observation the more evidence against H_0 it constitutes.
- Thus the set S_x referred to above is $\{0, 1, \dots, x\}$ and so the p-value is

$$P(S_x) = P(X \leq x) \text{ assuming } X \sim B(n, q_0).$$

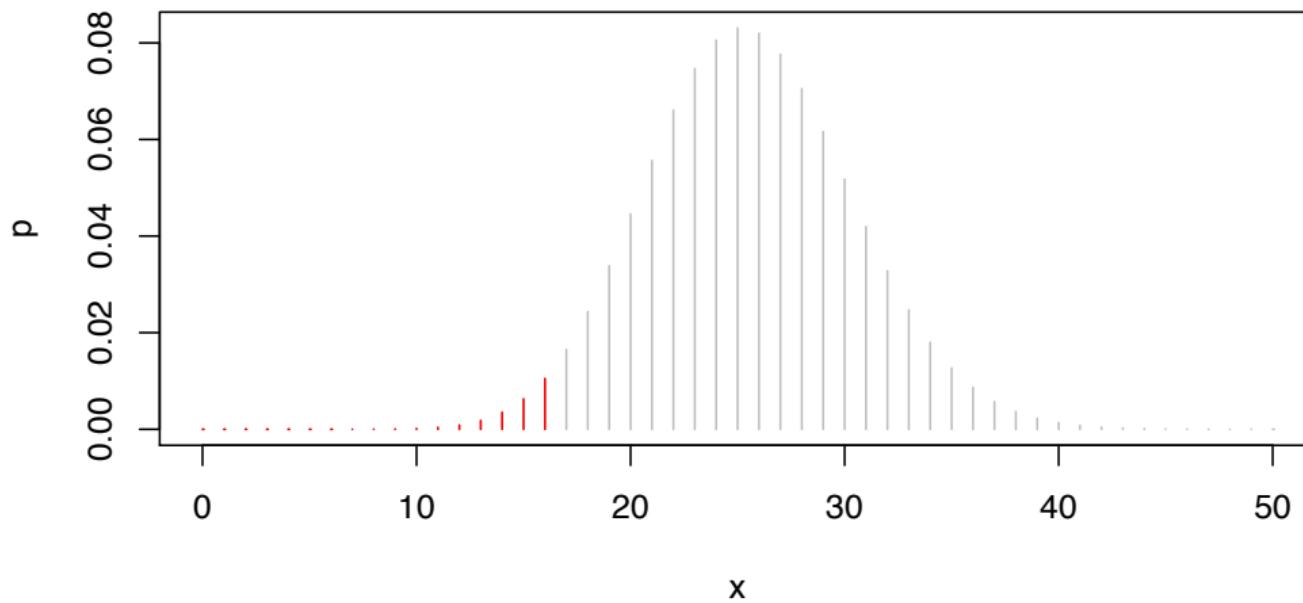
- In the example above where 16 of the 256 teenagers smoked, the observations constituting *at least* as much evidence are 0 smokers, 1 smoker, ..., 16 smokers.
- Thus there the p-value is $P(X \leq 16)$ when $X \sim B(256, 0.1)$ which is obtained in R as follows:

```
pbinary(16, 256, .1)
```

```
[1] 0.02360106
```

This is presented graphically below:

```
x=0:50  
p=dbinom(x, 256 ,0.1)  
plot(x,p,type="h",col="grey")  
lines(x[1:17] ,p[1:17] ,col="red",type="h")
```



Two-sided tests

- When the alternative hypothesis is $H_1: q \neq q_0$ there are (again) two cases:

The symmetric case: $q_0 = \frac{1}{2}$

- When the hypothesised value $q_0 = \frac{1}{2}$, the distribution of X under H_0 is *symmetric about $n/2$* .
- This symmetry means that we can equally well choose X or $n - X$ as test statistic and indeed these will each have the same distribution under H_0 i.e. $B(n, \frac{1}{2})$.
- In the coin tossing example where we are *testing for fairness* i.e. it is not known or suspected beforehand in which direction the coin might be biased, getting 15 heads (and 5 tails) is equally significant as getting 5 heads (and 15 tails);
 - we may use the number of heads **or** the number of tails as a test statistic.

The set S_x

- So if an observation is x , what is the set S_x of values constituting at least as much evidence against H_0 ?
- **Answer:** all observations *at least as far away from $n/2$ as x* (in either direction!).
- In the coin-tossing example, suppose for definiteness that X represents the number of heads and so the observed value $x = 15$.
- Then $S_x = \{0, 1, \dots, 5\} \cup \{15, 16, \dots, 20\}$.
- Thus the p-value would be

$$\begin{aligned} & P(X \leq 5) + P(X \geq 15) \text{ where } X \sim B(20, 0.5) \\ &= 2P(X \leq 5) \text{ by symmetry} \end{aligned}$$

- This is then

```
2*pbinom(5, 20, .5)
```

```
[1] 0.04138947
```

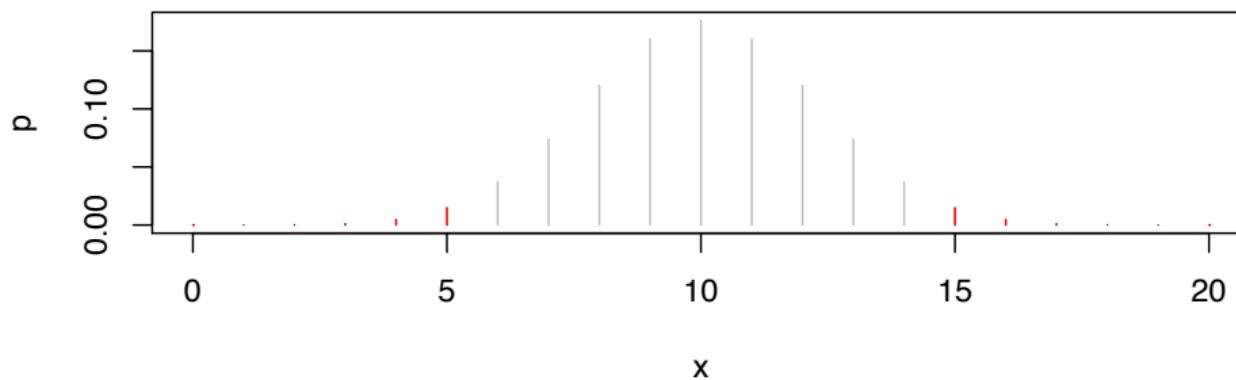
- Note that this is also twice the one-sided p-value obtained for example 1 above:

```
2*(1-pbinom(14,20,.5))
```

```
[1] 0.04138947
```

This is graphically presented below:

```
x=0:20  
p=dbinom(x,20,0.5)  
plot(x,p,type="h",col="grey")  
lines(x[16:21],p[16:21],col="red",type="h")  
lines(x[1:6],p[1:6],col="red",type="h")
```



- Note that if we instead chose to let the number of *tails* be the observation we would get $x = 5$ instead of $x = 15$ however we would get the same S_x , the same null distribution and thus the same *p*-value.
- Thus **in the symmetric case the two-sided p-value is twice the one-sided p-value** (assuming the latter is ≤ 0.5), just as it is for *Z*-tests and *t*-tests (where the test statistics also have symmetric distributions under the null hypothesis).

The asymmetric case: $q_0 \neq \frac{1}{2}$

- In this case, under the null hypothesis X does not have a symmetric distribution.
- There is no obvious, accepted natural way to convert an observation x into a two-sided p -value for such a case.
- The expected value of X under H_0 is nq_0 .
- Suppose for definiteness that the observed value x is *less* than nq_0 . Then how do we get a two-sided p -value from this?

Various proposals

- Various proposals have been put forward including:
 - ▶ double the one-sided p -value
 - ▶ define the set S_x as all values whose probabilities are $\leq P(X = x)$ under H_0 .
- We shall focus on two other approaches:
 - ▶ using Pearson's statistic
 - ▶ the likelihood ratio test

Using Pearson's statistic

- We have in fact already met a method of performing a two-sided binomial test: Pearson's chi-squared test, or more precisely a test using Pearson's statistic, using a chi-squared approximation to compute a p -value.
- Recall the simplest version of Pearson's test: model a vector of observed frequencies O_1, \dots, O_g as a $\text{Mult}(n; q_1, \dots, q_g)$ random vector, with $q_1 + \dots + q_g = 1$.
- To test the null hypothesis $H_0: q_1 = q_{01}, \dots, q_g = q_{0g}$, use Pearson's statistic

$$S = \sum_{i=1}^g \frac{(O_i - nq_{0i})^2}{nq_{0i}}$$

which has an approximate χ_{g-1}^2 distribution under H_0 .

- If S takes the value s , the approximate p -value is then

$$P(\chi_{g-1}^2 \geq s).$$

The alternative hypothesis here is simply H_1 : "not H_0 ".

- The simplest case is where $g = 2$ in which case we can instead write

- $O_1 = X$;
- $O_2 = n - X$;
- $q_1 = q$;
- $q_2 = 1 - q$;
- $q_{01} = q_0$.

and then we have that $X \sim B(n, q)$ and the test reduces to $H_0: q = q_0$ against $H_1: q \neq q_0$.

- Pearson's statistic is then

$$\begin{aligned}\frac{(X - nq_0)^2}{nq_0} + \frac{[(n - X) - n(1 - q_0)]^2}{n(1 - q_0)} &= (X - np_0)^2 \left[\frac{1}{nq_0} + \frac{1}{n(1 - q_0)} \right] \\ &= \frac{(X - nq_0)^2}{nq_0(1 - q_0)}\end{aligned}$$

since $(n - X) - n(1 - q_0) = nq_0 - X$.

- This shows us clearly why Pearson's statistic is approximately χ_1^2 in this case because under H_0 (as we have already seen), $X \sim B(n, q_0) \stackrel{\text{approx}}{\sim} N(nq_0, nq_0(1 - q_0))$ for large n .
- Thus the standardised version

$$\frac{X - nq_0}{\sqrt{nq_0(1 - q_0)}} \stackrel{\text{approx}}{\sim} N(0, 1)$$

and thus its square is approximately χ_1^2 (the distribution of Z^2 if $Z \sim N(0, 1)$).

- However we *need not rely on an approximation when the exact p-value is quite easy to compute.*

Potential inaccuracy of the χ^2 approximation

- Consider firstly the *symmetric* two-sided example we have already dealt with above. We can think of it as a multinomial/Pearson's chi-squared-type test with

	H	T
Obs	15	5
Exp	10	10

- As shown above, the observed value of Pearson's statistic is

$$\frac{[15 - (20 \times \frac{1}{2})]^2}{20 \times \frac{1}{2} \times \frac{1}{2}} = 5.$$

- An approximate p-value is obtained by supposing this has a χ_1^2 distribution, giving an approximate *p*-value of

```
1-pchisq(5, df=1)
```

```
[1] 0.02534732
```

- This is **much smaller** than the “exact” *p*-value of 0.04138947 obtained above. The χ^2 approximation is not very accurate in this case.
- We might expect similar inaccuracies in the asymmetric case.

Exact p-values using Pearson's statistic

not symmetric anymore, more common example

- In this simplest case of Pearson's testing procedure we can compute exact p -values quite easily rather than relying on the χ^2 approximation.
- If we are modelling $X \sim B(n, q)$ and testing $H_0: q = q_0$ against $H_1: q \neq q_0$, if the observed value is x then the p -value is given by

$$\begin{aligned} P \left\{ \frac{(X - nq_0)^2}{nq_0(1 - q_0)} \geq \frac{(x - nq_0)^2}{nq_0(1 - q_0)} \right\} &= P \{ (X - nq_0)^2 \geq (x - nq_0)^2 \} \\ &= P \{ |X - nq_0| \geq |x - nq_0| \} \end{aligned}$$

which is the probability that the observation X takes a value *at least as far away from the hypothesised mean nq_0 as was observed*.

- How this simplifies further depends on whether x is greater than or less than np_0 (the mean of X under H_0).

$x > np_0$

- If $x > np_0$ then $|x - np_0| = x - np_0$ and so the p-value reduces to

$$P(X \geq x) + P(X \leq np_0 - (x - np_0)) = P(X \geq x) + P(X \leq 2np_0 - x).$$

- ▶ So the value $2np_0 - x < np_0$ is the point equidistant from np_0 on the *low* side as x is on the *high* side.
- ▶ Note that this may or may not be an integer.

$x < np_0$

- If $x < np_0$ then $|x - np_0| = np_0 - x$ and so the p-value reduces to

$$P(X \leq x) + P(X \geq np_0 + (np_0 - x)) = P(X \leq x) + P(X \geq 2np_0 - x).$$

- ▶ Here $2np_0 - x > np_0$ is the point equidistant from np_0 on the *high* side as x is on the *low* side.
 - ▶ Again $2np_0 - x$ may or may not be an integer.
- Note that if $x = np_0$ then Pearson's statistic is zero and the p-value is thus 1.

Second example

- The exact p -value for the test using Pearson's statistic can be written as

$$\begin{aligned} P(|X - 25.6| \geq |16 - 25.6|) &= P(|X - 25.6| \geq 9.6) \\ &= P(X \leq 16) + P(X \geq 35.2) \\ &= P(X \leq 16) + P(X \geq 36). \end{aligned}$$

- Using R this is

```
pbinary(16, 256, .1)+1-pbinary(35, 256, .1)
```

[1] 0.04708377

The likelihood ratio test

- The *likelihood ratio* test is a general procedure which can be applied in almost *any* statistical testing situation.
- It is based on the **likelihood function** which is simply *the probability of the observed data as a function of the unknown parameter*.
- When we observe a single value x modelled as the value taken by $X \sim B(n, q)$ ² for some unknown q , the likelihood function is

$$L(q) = P_q(X = x) = \binom{n}{x} q^x (1 - q)^{n-x}.$$

- This is a function of 3 real variables: x , n and q however in a practical statistical context we would have
 - ▶ x and n **known** but
 - ▶ q **unknown**.

²We use q instead of p for the binomial success probability parameter here to avoid confusion with p -values

How “likely” is the hypothesised value?

- The likelihood function is used to indicate how consistent a certain value q might be with the data: the higher the likelihood function (the “more likely” it is), the better the agreement between parameter and data.
- In the tutorial exercises we show that the value of q that is “most likely” i.e. gives the *highest possible value of the likelihood function* is simply the observed proportion x/n . ³
- We can compare the relative likelihood of the hypothesised value q_0 to the “best possible” value x/n by evaluating the *ratio* of the likelihood function at these two values:

$$LR = \frac{L(x/n)}{L(q_0)}.$$

Note this *likelihood ratio* is always ≥ 1 .

- If q_0 is close to the observed proportion x/n this should be close to 1, so q_0 is “almost as likely” as the best possible value.
- The larger this is though, the *less likely* q_0 is compared to the best possible value.

³In other words x/n is the “maximum likelihood estimate” of q

Larger values of the likelihood ratio mean more evidence against H_0 .

- So larger values of the $\text{LR} = L(x/n)/L(q_0)$ indicate poorer agreement between data and hypothesised parameter value q_0 .
- We thus regard larger values of the LR statistic to constitute *more evidence against H_0* .
- Thus if the LR statistic takes the value y , the p -value would be

$$P(\text{LR} \geq y) \text{ when } H_0 \text{ is true}$$

- How is this computed? It is perhaps easiest to just use R directly.

Second example

- Recall we model $X \sim B(256, q)$ and are testing $H_0: q = 0.1$. The observed value was $x = 16$, giving an observed proportion of $16/256 = 0.0625$.
- Let us compute two vectors:
 - ① $P(X = x)$ for $x = 10, 11, \dots, 40$ when $q = 0.1$;
 - ② $P(X = x)$ for $x = 10, 11, \dots, 40$ when $q = x/256$.

```
x=10:40  
probs.q0=dbinom(x,256,0.1)  
max.probs=dbinom(x,256,x/256)
```

- The vector of ratios given by `max.probs/probs.q0` gives the value of the LR statistic for each of these potential observed values x .

```
data.frame(x, ratio=max.probs/probs.q0)
```

```
x      ratio
1 10 825.938871 #
2 11 317.807596 #
3 12 134.489741 #
4 13 62.119385 # All values x < 16 give LR larger than 9.75
5 14 31.116638 #
6 15 16.811259 #
7 16 9.749545 # <-- This is the observed value
8 17 6.044204
9 18 3.990915
10 19 2.797503
11 20 2.075722
12 21 1.626043
13 22 1.341625
14 23 1.163413
15 24 1.058252
16 25 1.007898
17 26 1.003462
18 27 1.042753
19 28 1.129397
20 29 1.273300
21 30 1.492474
22 31 1.816701
23 32 2.294046
24 33 3.002159
25 34 4.067947
26 35 5.702277
27 36 8.262217
28 37 12.364719 # All values x >= 37 give LR > 9.75 too.
29 38 19.098273 #
30 39 30.424696 #
31 40 49.957045 #
```

LR test p -value calculation

- The observed value of the LR statistic is ≈ 9.75 (since observed $x = 16$).
- The set of all observations x giving an LR statistic *at least as large* is

$$\{0, 1, 2, \dots, 16\} \cup \{37, 38, \dots, 256\}, .$$

- Therefore the p -value is

$$P(X \leq 16) + P(X \geq 37) = P(X \leq 16) + [1 - P(X \leq 36)] \quad \text{where } X \sim B(256, 0.1).$$

- This is given by

```
pbinary(16, 256, 0.1) + 1 - pbinary(36, 256, 0.1)
```

```
[1] 0.03842328
```

- The one-sided binomial tests we considered earlier are also LR tests.
 - ▶ The maximisation in the numerator is restricted to $q \geq q_0$ or $q \leq q_0$ depending on whether H_1 is $q > q_0$ or $q < q_0$ respectively.
 - ▶ Thus if the observed proportion x/n is not “in H_1 ” then the “most likely” estimate is simply q_0 and the LR is 1.
- A two-sided test for a Poisson parameter may be derived using the LR statistic; in that case the “most likely” parameter value is simply the sample mean (we expand on this in a tutorial exercise).
- A likelihood function may also be defined for continuous models. We shall not explore this any further except to point out that the Z-test and t-tests we have studied (except for the Welch test) are also LR tests.

Sign test

- An important application of *symmetric* binomial tests (i.e. when the null hypothesis is $H_0: p = \frac{1}{2}$) is the **sign test**.
- The model is that observations are modelled as independent random variables X_1, \dots, X_n , all with the same distribution, and these are then converted into their *signs*:
+ and - are equally and likely

$$S_i = \begin{cases} + & \text{if } X_i > 0 \\ 0 & \text{if } X_i = 0 \text{ and} \\ - & \text{if } X_i < 0. \end{cases}$$

- So the observations are reduced to a collection of random signs, with

$$P(S_i = +) = p_+ = P(X_i > 0),$$

$$P(S_i = 0) = p_0 = P(X_i = 0) \text{ and}$$

$$P(S_i = -) = p_- = P(X_i < 0).$$

- The null hypothesis to be tested is $H_0: p_+ = p_-$, so that positive signs and negative signs are equally likely.
- The test statistic is N_+ . the **number of + signs**.
- The sign test can be one-sided or two sided.
- So there are three possibilities for the alternative:
 - ① $H_1: p_+ > p_-$;
 - ② $H_1: p_+ < p_-$;
 - ③ $H_1: p_+ \neq p_-$.

A *conditional* test

- The key feature of the sign test is that it is a *conditional* test.
- That is to say, the p-value that is quoted is in fact a *conditional* probability.
- More precisely let N_0 be the (random) number of zeroes and write n_0 for the *observed* number of zeroes.
- If the observed number of plus signs is n_+ , the p-value is the *conditional probability*

$$\begin{aligned} P(N_+ \geq n_+ | N_0 = n_0) &\text{ when } p_+ = p_- \\ &= P(X \geq n_+) \text{ when } X \sim B\left(n - n_0, \frac{1}{2}\right). \end{aligned}$$

- This last result follows simply because conditional on there being $n - n_0$ non-zero signs, each one is either positive with (conditional) probability $\frac{p_+}{p_+ + p_-}$ and this is $\frac{1}{2}$ under H_0 .
- The full derivation of this result is beyond the scope of the course however it is very intuitive.

- Since the zeroes tell us nothing about the relative abundance of positive and negative signs, we simply discard them and regard the **effective** sample size as $n - n_0$ and treat the remaining signs as binomial trials.
- Thus to perform as sign test we
 - ① discard any zeroes, leaving $n - n_0$ non-zero signs;
 - ② regard the number of positive signs as a binomial random variable $X \sim B(n - n_0, p)$;
 - ③ test the null hypothesis $H_0: p = \frac{1}{2}$ (against an appropriate alternative).

Applications using the sign test.

- “taste tests” or other market research type exercises where subjects are asked to indicate preference for one (or neither) of two competing products
 - paired tests when the assumptions of a t -test are not reasonable and/or data may be rounded to perhaps a single decimal place
 - various examples are presented in the tutorial exercises.

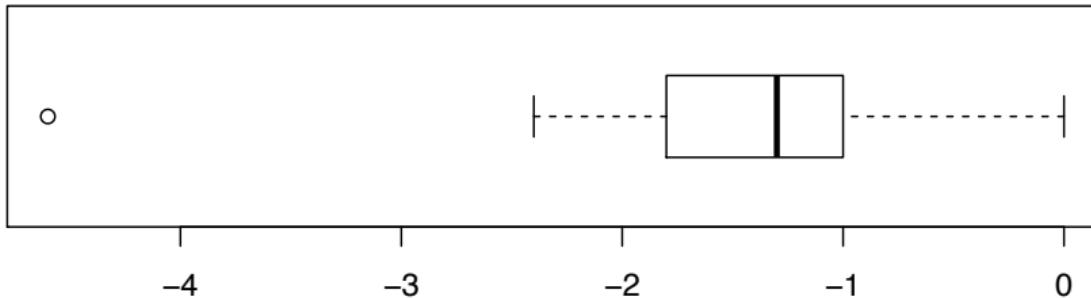
Example

- Consider Student's famous sleep data (recall from lecture 16)

```
x=sleep$extra[sleep$group==1]
y=sleep$extra[sleep$group==2]
d=x-y
d
```

```
[1] -1.2 -2.4 -1.3 -1.3  0.0 -1.0 -1.8 -0.8 -4.6 -1.4
```

```
boxplot(d, horizontal=T)
```



- We can however perform a sign test.

```
sort(d)
```

```
[1] -4.6 -2.4 -1.8 -1.4 -1.3 -1.3 -1.2 -1.0 -0.8 0.0
```

```
sum(d<0)
```

```
[1] 9
```

```
sum(d==0)
```

```
[1] 1
```

```
sum(d>0)
```

```
[1] 0
```

- Out of 10 differences, 1 is zero and 9 are negative.

- We thus perform a two-sided binomial test with $n = 9$, which gives a *two-sided* p-value of

```
2*pbinom(0,9,0.5)
```

```
[1] 0.00390625
```

that is

```
2^(-8)
```

```
[1] 0.00390625
```

- This is very strong evidence that there is a difference between the drugs, or rather that the probability of the y drug giving a greater improvement than the x drug is *not equal to* 0.5.
- Note that this is almost the same as the two-sided *t*-test p-value, which is reassuring.

Outline

- 1 Welcome
- 2 Data Analysis
- 3 Probability
- 4 Inference Part 1: Continuous models
- 5 Inference Part 2: Discrete models
 - Lecture 19
 - Lecture 20
 - **Lecture 21**
 - Binomial Confidence Interval
 - Constructing confidence intervals for a binomial success probability.
 - Lecture 22

Review of confidence intervals covered so far

- We have seen confidence intervals in the context of inference concerning a population mean (or a difference between two population means, or even a regression line slope parameter).
- In all cases they have taken one of three forms:

- ▶ a two-sided interval:

$$(est - c(se), est + c(se)); \quad \begin{matrix} \text{standard error} \\ \text{estimate} \end{matrix}$$

- ▶ a one-sided interval of the form

$$(-\infty, est + c(se)),$$

the right endpoint also being called an “upper confidence limit”;

- ▶ a one-sided interval of the form

$$(est - c(se), \infty),$$

the left endpoint also being called a “lower confidence limit”.

- In all cases the constant c is chosen in such a way that the *random interval* contains the true but unknown parameter value is equal to the confidence level.

Relationship between tests and confidence intervals

parameter value

- In all our earlier cases, we noted that there is a connection between confidence intervals and (families of) hypothesis tests.
- To make this clear, there are two ways that data can be consistent or not consistent with a possible parameter value at a certain “level”:
 - ▶ Given a significance level e.g. 0.05, data is said to be consistent with a certain parameter value if the corresponding p-value is bigger than 0.05; conversely it is said to **not** be consistent with a certain parameter value if the corresponding p-value is less than 0.05.
 - ▶ Given a confidence level e.g. 95%, data is said to be consistent with a certain parameter value if that value is included in the 95% confidence interval; it is said to **not** be consistent with a certain parameter value if that value is **not** included in the 95% confidence interval.
- It turns out these two senses are exactly equivalent, with the correspondence:

$$\alpha \text{ significance level} \Leftrightarrow 100(1 - \alpha)\% \text{ confidence level.}$$

- For example, the value c above for e.g. a 95% confidence interval for a population mean μ also equals a 5% critical value for a test based on the test statistic

$$\frac{\bar{x} - \mu_0}{\text{se}}.$$

- Thus e.g. in the two-sided case

$$p\text{-value} \geq 0.05 \Leftrightarrow \frac{|\bar{x} - \mu_0|}{\text{se}} \leq c \Leftrightarrow \mu_0 \text{ in the interval } \bar{x} \pm c \text{ se}$$

and conversely

$$p\text{-value} < 0.05 \Leftrightarrow \frac{|\bar{x} - \mu_0|}{\text{se}} > c \Leftrightarrow \mu_0 \text{ NOT in the interval } \bar{x} \pm c \text{ se}$$

- We may use this equivalence to construct confidence intervals for binomial p parameters, i.e. by *inverting tests*.

Exact, approximate, valid and conservative confidence intervals

- All the confidence interval procedures mentioned above for continuous models are **exact** procedures in that the **coverage probability** (the probability that the true parameter value is contained in the confidence interval) is exactly equal to the stated or **nominal** confidence level, *regardless what the true parameter value is*.
- In other situations, particularly **discrete** models, this is not always possible for all true parameter values.
- A **valid** procedure has coverage probability **at least** equal to the nominal level for all true parameter values.
- A **conservative** procedure has the coverage probability **greater than** the nominal level; such procedures usually achieve this property at the expense of having wider intervals than necessary.
- In most contexts “valid” and “conservative” are used interchangably although “valid” has positive connotations while “conservative” has negative connotations.
- An **approximate** procedure has the coverage probability approximately equal to the nominal confidence level. Usually such a procedure is not valid, in that for some true parameter values the coverage probability is below the nominal confidence level.

Historically popular but discredited binomial confidence intervals

- Historically, there have been two commonly used procedures for constructing a (two-sided) confidence interval for a binomial success probability parameter:
 - ① An *approximate* $100(1 - \alpha)\%$ confidence interval of the form
$$\text{estimate} \pm c \text{ s.e.}$$
where the "table value" c is the upper $\alpha/2$ point from the $N(0, 1)$ distribution. We call this the **Wald** interval.
 - ② A *valid* but *conservative* procedure due to **Clopper and Pearson** which is used by the R function `binom.test()`.
- We **do not recommend using either of these procedures.**
- Why?
 - ▶ The Wald interval, while easy to compute, has recently been shown to have terrible coverage probability properties.
 - ▶ The Clopper-Pearson interval is overly conservative, it gives needlessly wide intervals; there exist other valid methods which give narrower intervals.
- Instead we recommend alternatives obtained by *inverting certain tests*⁴.

⁴The Wald and Clopper-Pearson procedures can each also be interpreted as inverting a certain test; we discuss this later.

Constructing confidence intervals for a binomial success probability.

- According to the above equivalence, we can define e.g. a 95% confidence interval for a binomial p parameter by inverting a test as follows:

All values q_0 such that a test of $H_0: q = q_0$ has p-value at least 0.05.

- One-(respectively two-)sided level tests give one-(respectively two-)sided confidence intervals.
- In the binomial context one-sided tests are clear/unambiguous, so we start with one-sided confidence intervals.

One-sided examples

- Let us return to the motivating examples from the last few lectures.

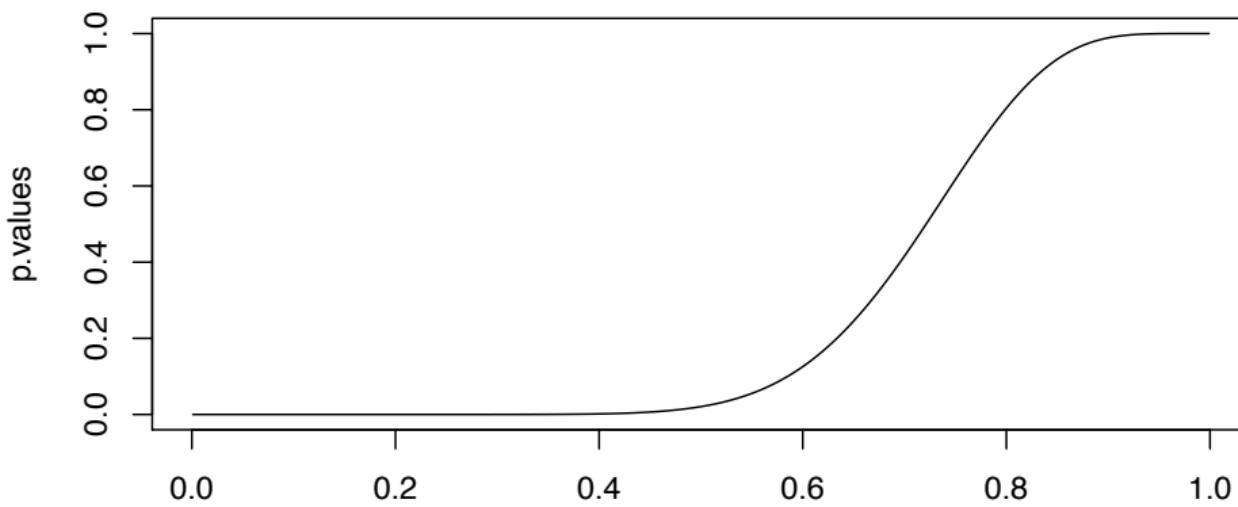
15 heads from 20 flips

- Suppose we suspect a coin may be biased in favour of heads; 20 independent flips turn up 15 heads.
- What is the lowest p consistent with this at the
 - 95% confidence level;
 - 99% confidence level?
- The corresponding one-sided tests are of the form $H_0: q = q_0$ against $H_1: q > q_0$; each such test would have (one-sided) p -value given by

$$P(X \geq 15) = 1 - P(X \leq 14) \text{ where } X \sim B(20, q_0).$$

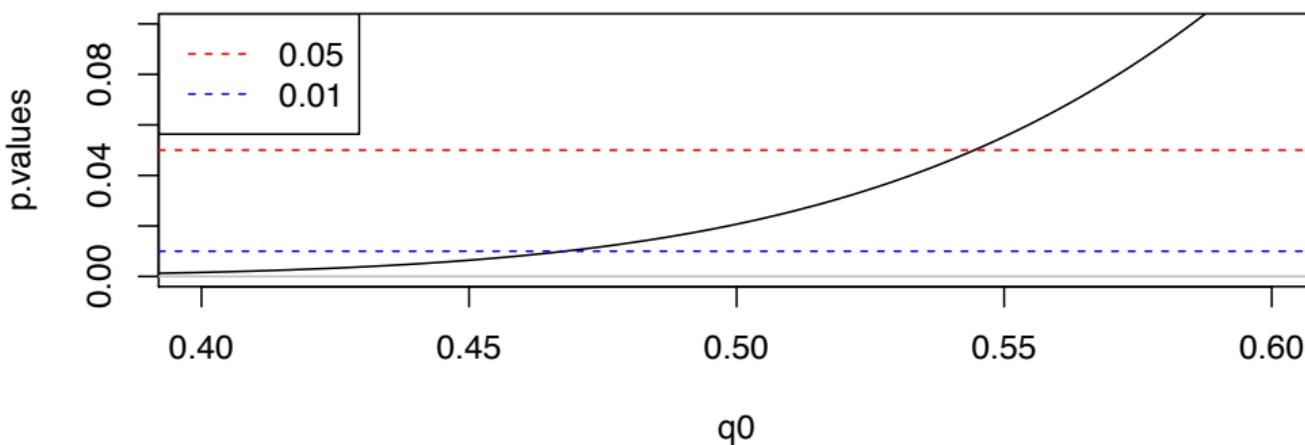
- The R code below computes this p-value for a fine grid of q_0 values, and plots these against q_0 :

```
q0=1:999/1000  
p.values=1-pbinom(14,20,q0)  
plot(q0,p.values,type="l")
```



The same plot is produced below, but we have zoomed in and added horizontal lines at 0.05 and 0.01:

```
plot(q0,p.values,type="l",xlim=c(0.4,0.6),ylim=c(0,.1))
abline(h=0,col="grey")
abline(h=0.05,col="red",lty=2)
abline(h=0.01,col="blue",lty=2)
legend("topleft",leg=c("0.05","0.01"),col=c("red","blue"),lty=c(2,2))
```



We can see the horizontal 0.05 line cuts at around 0.54, while the 0.01 line at about 0.47.

The R function `binom.test()`

- The easiest way to obtain these points explicitly in R is to use the `binom.test()` function which performs a test and produces a confidence interval.
- For one-sided tests this does just what we want **but it does the wrong thing for two-sided tests** (as we shall see later):

```
binom.test(15, 20, alt="greater")
```

Exact binomial test

```
data: 15 and 20
number of successes = 15, number of trials = 20,
p-value = 0.02069
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
0.5444176 1.0000000
sample estimates:
probability of success
0.75
```

- We see here that the 95% (the default level) one-sided interval is (0.5444,1).
- Note also that the default hypothesised q_0 is 0.5 and this p-value agrees with our earlier analysis.

- For a 99% interval,

```
binom.test(15,20,alt="greater",conf.level=.99)
```

Exact binomial test

```
data: 15 and 20
number of successes = 15, number of trials = 20,
p-value = 0.02069
alternative hypothesis: true probability of success is greater than 0.5
99 percent confidence interval:
0.4678936 1.0000000
sample estimates:
probability of success
0.75
```

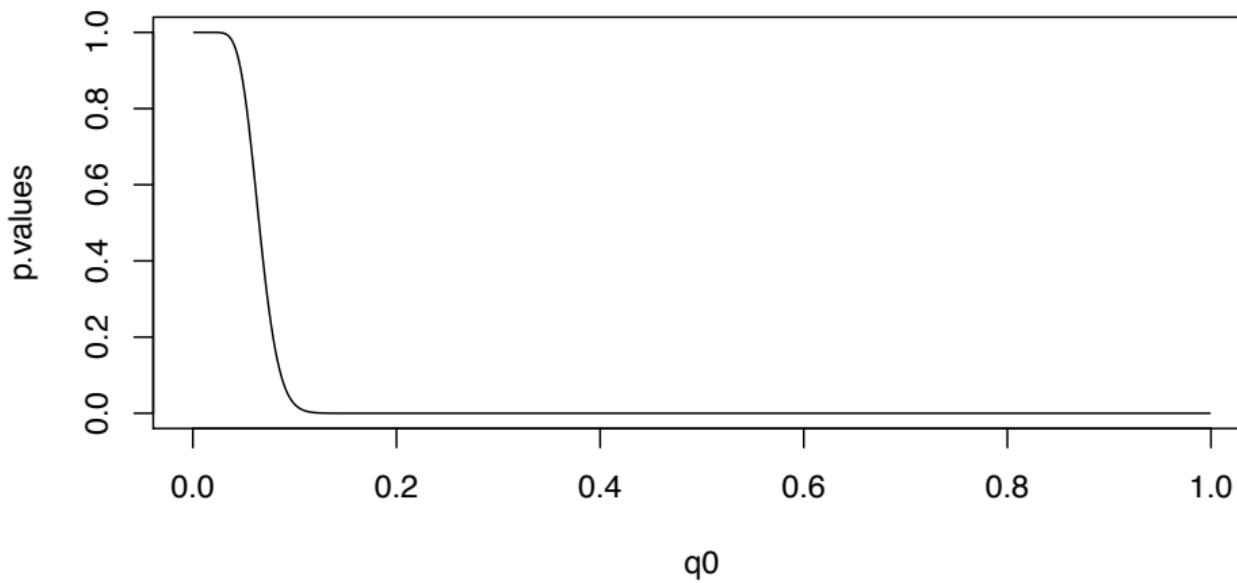
16 smokers from 256 teenagers

- It is put forward that an awareness campaign has reduced the smoking rate among teenagers from 10%.
- In a random sample of 256 such teenagers 16 smoked.
- What is the largest value of q (the true current proportion of teenage smokers) consistent with this at the
 - ▶ 95% confidence level;
 - ▶ 99% confidence level?
- The corresponding one-sided tests are of the form $H_0: q = q_0$ against $H_1: q < q_0$; each such test would have a one-sided p -value given by

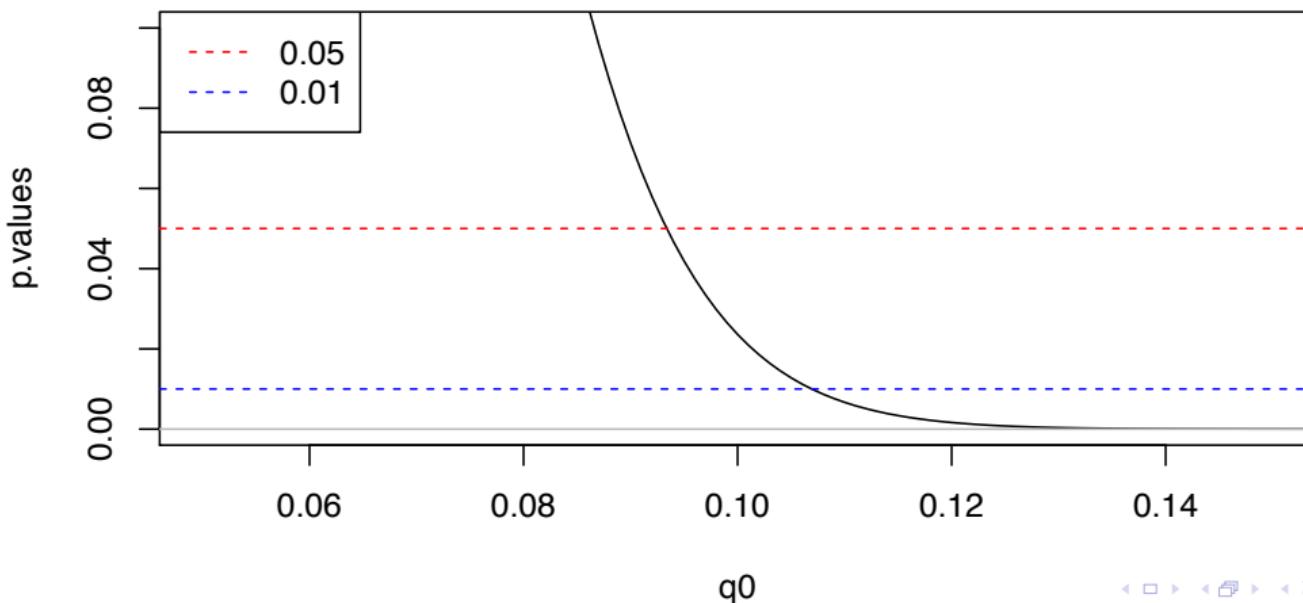
$$P(X \leq 16) \text{ where } X \sim B(256, p_0).$$

- Analogous plots to those above for the previous example appear below:

```
q0=1:999/1000  
p.values=pbinom(16,256,q0)  
plot(q0,p.values,type="l")
```



```
plot(q0,p.values,type="l",xlim=c(0.05,0.15),ylim=c(0,.1))
abline(h=0,col="grey")
abline(h=0.05,col="red",lty=2)
abline(h=0.01,col="blue",lty=2)
legend("topleft",leg=c("0.05","0.01"),col=c("red","blue"),lty=c(2,2))
```



- To determine the upper confidence limits, again use `binom.test()`:

- first for 95%:

```
binom.test(16,256,alt="less")
```

```
Exact binomial test

data: 16 and 256
number of successes = 16, number of trials = 256,
p-value < 2.2e-16
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
0.00000000 0.09338157
sample estimates:
probability of success
0.0625
```

- next for 99%:

```
binom.test(16,256,alt="less",conf.level=.99)
```

```
Exact binomial test

data: 16 and 256
number of successes = 16, number of trials = 256,
p-value < 2.2e-16
alternative hypothesis: true probability of success is less than 0.5
99 percent confidence interval:
0.000000 0.106923
sample estimates:
probability of success
0.0625
```

- Note, to get sensible p-values here in the testing context, add the `p=0.1` argument (although this is not needed for confidence intervals/limits):

- ▶ first for 95%:

```
binom.test(16,256,p=.1,alt="less")
```

```
Exact binomial test

data: 16 and 256
number of successes = 16, number of trials = 256,
p-value = 0.0236
alternative hypothesis: true probability of success is less than 0.1
95 percent confidence interval:
 0.00000000 0.09338157
sample estimates:
probability of success
0.0625
```

- ▶ next for 99%:

```
binom.test(16,256,p=.1,alt="less",conf.level=.99)
```

```
Exact binomial test

data: 16 and 256
number of successes = 16, number of trials = 256,
p-value = 0.0236
alternative hypothesis: true probability of success is less than 0.1
99 percent confidence interval:
 0.000000 0.106923
sample estimates:
probability of success
0.0625
```

Two-sided examples

- In the one-sided case, it is clear how everything should be done.
- In the two-sided case, however, it is not.
- Two historically popular methods (Wald and Clopper-Pearson) are **not recommended**.
- We instead recommend two other procedures:
 - ① an **approximate** procedure obtained by inverting the test based on Pearson's statistic (which uses a χ^2 approximation to obtain a *p*-value) called the **Wilson interval**;
 - ② a **valid/conservative** procedure obtained by inverting the (exact *p*-value) **likelihood ratio test**.

The binom R package

- The R computing language has a large number of “built-in” functions for performing all sorts of statistical computations.
- Being a full-blown computing language it is possible for users to extend R by writing their own functions.
- There are over 9,000 add-on packages (written and made available by ordinary R users) on CRAN, and more made available via other channels.
- One of these is the `binom` package which has functions for computing various additional *approximate* binomial confidence intervals.
- It contains functions `binom.wilson()` and `binom.lrt()` for computing Wilson and LR-test-based confidence intervals although note that `binom.lrt()` **uses a normal approximation**, it is not “exact” (we shall use our own code for the LR test-based intervals).
- You will need to install the `binom` library first (quite easy in RStudio by selecting the menu item Tools->Install Packages... and entering `binom` in the dialog box).
- Once the library is installed type

```
library(binom)
```

at the R console. Then the function `binom.wilson()` will be available.

15 heads in 20 flips: Wilson interval

- The 95% Wilson interval is given by the command

```
binom.wilson(15,20)
```

	method	x	n	mean	lower	upper
1	wilson	15	20	0.75	0.5312991	0.8881383

- The 99% Wilson interval is given by the command

```
binom.wilson(15,20,conf.level=.99)
```

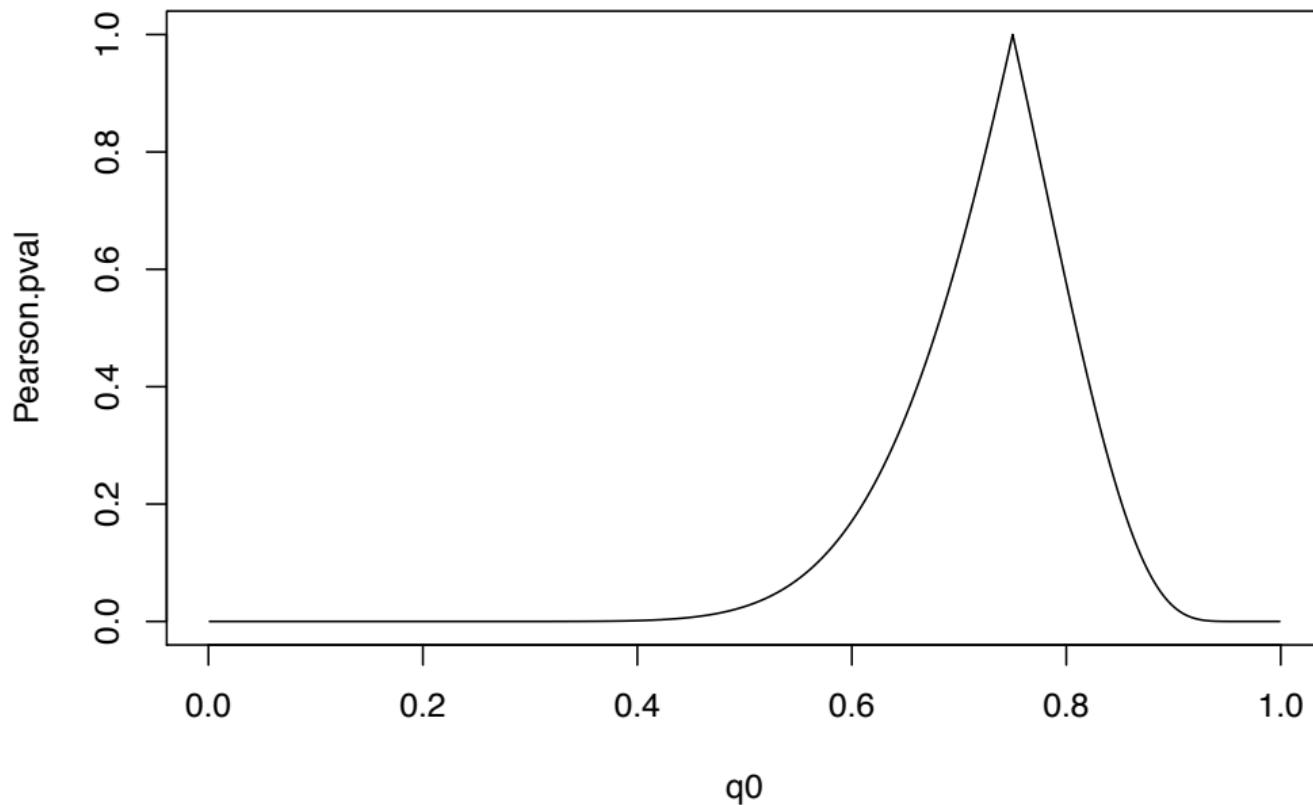
	method	x	n	mean	lower	upper
1	wilson	15	20	0.75	0.4628112	0.9126361

P -value plots: Pearson statistic (χ^2 approx. p -value)

- To see how these interval endpoints are obtained, we can plot the p -value of the test $H_0: q = q_0$ against q_0 .
- We can obtain the (χ^2 approximations for) the Pearson test p -values using the R function `chisq.test()`:

```
q0=1:999/1000
Pearson.pval=0
for (i in 1:999){
  Pearson.pval[i]=
    chisq.test(c(15,5),p=c(q0[i],1-q0[i]))$p.val
}
plot(q0,Pearson.pval,type="l",
      main="Approximate p-value for Pearson test of q=q0")
```

Approx p-value for Pearson test of $q=q_0$

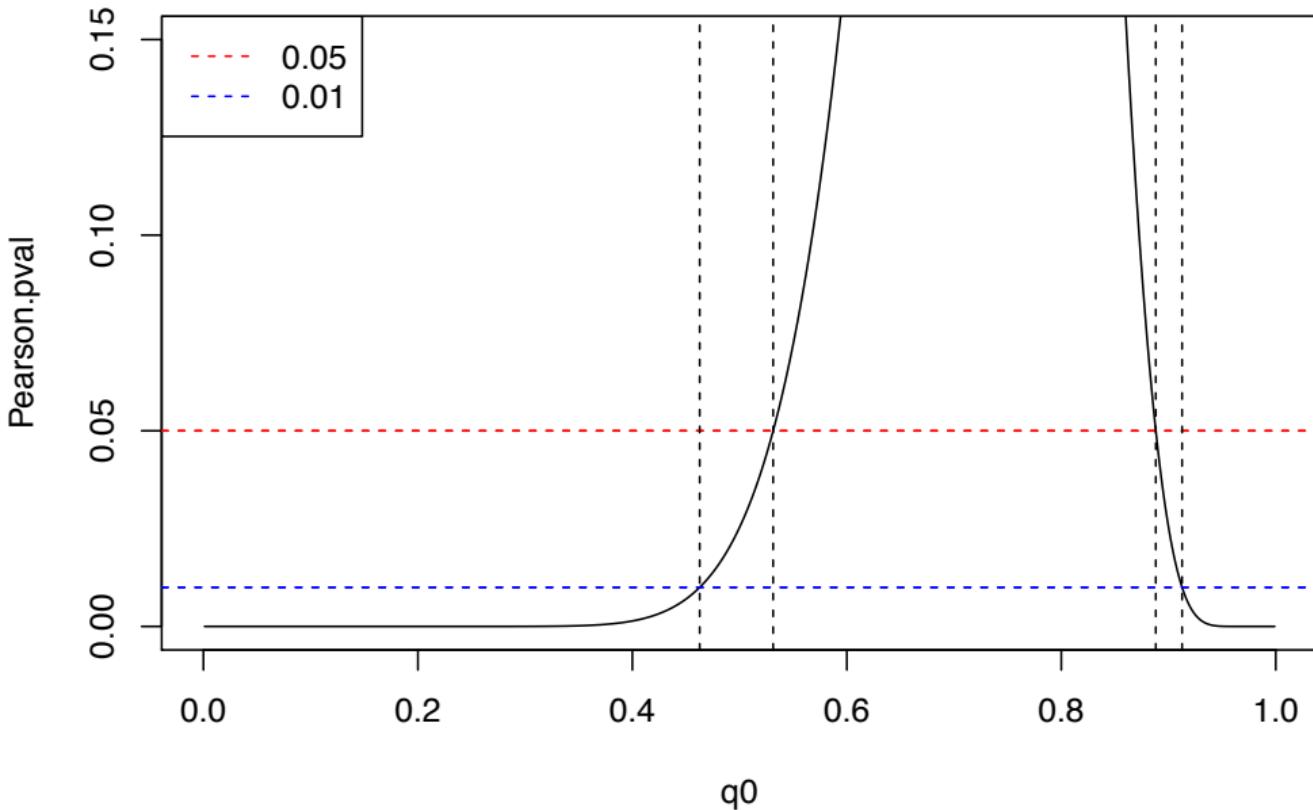


Zoomed-in version

- We focus on the lower part of the graph and add horizontal lines at 0.05, 0.01 and vertical lines at the endpoints returned by `binom.wilson()`

```
plot(q0, Pearson.pval, type="l", ylim=c(0,0.15))
abline(h=c(0.05,0.01), lty=2, col=c("red","blue"))
w95=binom.wilson(15,20)
w99=binom.wilson(15,20, conf.level=.99)
abline(h=c(0.05,0.01), lty=2, col=c("red","blue"))
abline(v=c(w95$lower,w95$upper,w99$lower,w99$upper), lty=2)
legend("topleft", legend=c("0.05","0.01"), lty=2, col=c("red","blue"))
```

- The values returned by `binom.wilson()` agree perfectly with the plot.



A note on the Wilson interval (**not examinable**)

- The calculations performed by `binom.wilson()` are not that hard to obtain directly.
- We need to find the values of q_0 such that

$$\text{Pearson statistic} = \frac{(x - nq_0)^2}{nq_0(1 - q_0)} \leq c,$$

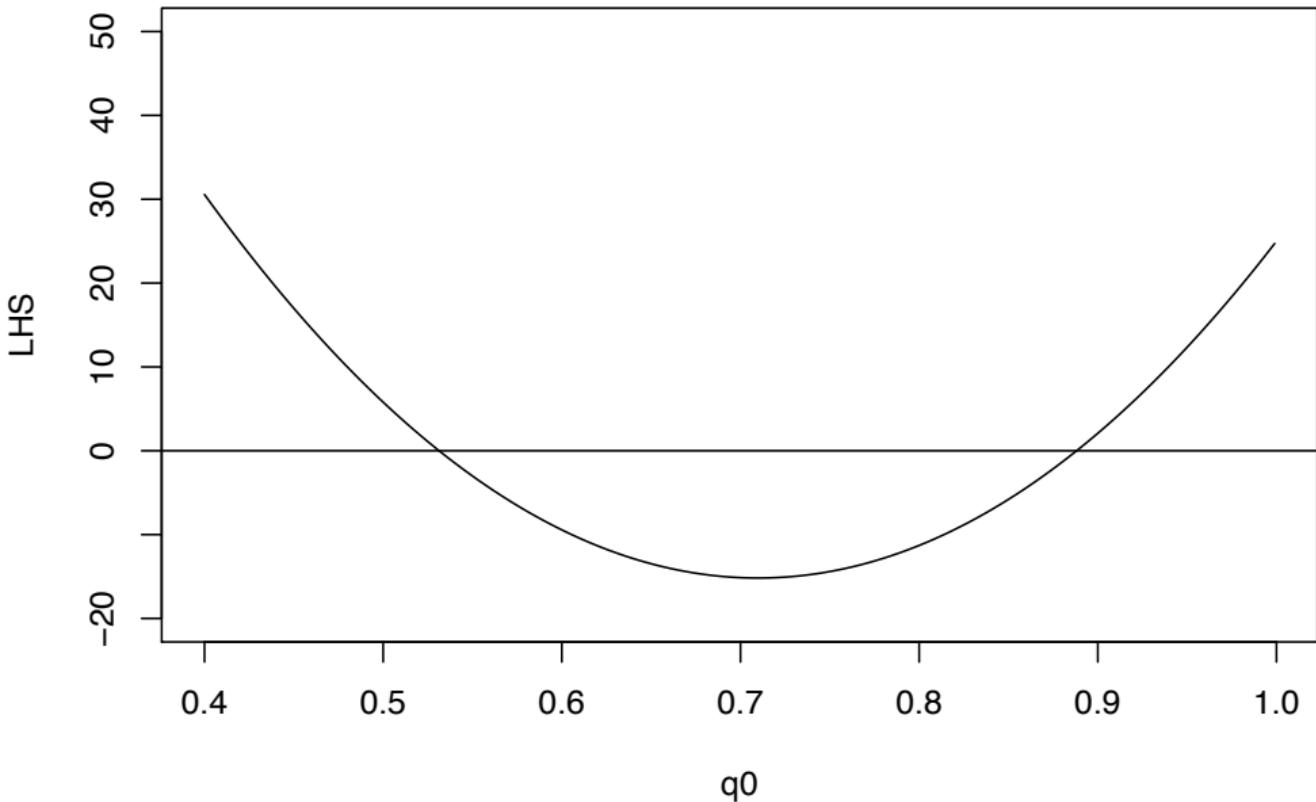
where c is a value from the χ^2_1 table satisfying $P(\chi^2_1 > c) = \alpha$; that is

$$x^2 - 2xnq_0 + n^2q_0^2 \leq (nq_0 - nq_0^2)c,$$

$$\text{equivalently } q_0^2(n^2 + nc) - q_0(2xn + n) + x^2 \leq 0.$$

- The function of q_0 on the left-hand side is plotted below for our example.

```
x=15; n=20; c=qchisq(.95, 1)
q0=400:999/1000
LHS=(n^2+n*c)*q0^2 - q0*(2*x*n+n*c) + x^2
plot(q0, LHS, type="l", ylim=c(-20,50))
abline(h=0)
```



- The roots of the equation are exactly the interval endpoints.
- Note also the midpoint of the interval (the $-\frac{b}{2a}$ term in the general formula $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$) is **weighted average** of $\hat{q} = \frac{x}{n}$ and $\frac{1}{2}$:

$$\frac{2xn + nc}{2n^2 + 2nc} = \left(\frac{n}{n+c} \right) \frac{x}{n} + \left(\frac{c}{c+n} \right) \frac{1}{2}$$

and so is shifted from \hat{q} in the direction of $\frac{1}{2}$. This reflects the *asymmetry* of $B(n, q)$ distributions when $q \neq \frac{1}{2}$.

P-value plots: likelihood ratio test (exact *p*-value)

- The R code in the file `binomLRT.R` creates the functions `binomLRtest()` and `binomLR.CI()`.
- The former computes the exact *p*-value of the two-sided test of $q = q_0$ based on the LR statistic:

```
source("binomLRT.R")
binomLRtest(15,20)      # default hypothesised q0=0.5
```

```
$x.obs
[1] 15

$p.value
[1] 0.04138947

$table
   x    probs0 max.probs      ratio p.values
 0 0.000001 1.000000 1048576.000000 0.000002
 1 0.000019 0.377354 19784.197000 0.000040
 2 0.000181 0.285180 1573.856300 0.000402
 3 0.001087 0.242829 223.354870 0.002577
 4 0.004621 0.218199 47.223665 0.011818
 5 0.014786 0.202331 13.684184 0.041389
 6 0.036964 0.191639 5.184418 0.115318
 7 0.073929 0.184401 2.494307 0.263176
 8 0.120134 0.179706 1.495873 0.503445
 9 0.160179 0.177055 1.105356 0.823803
10 0.176197 0.176197 1.000000 1.000000
11 0.160179 0.177055 1.105356 0.823803
12 0.120134 0.179706 1.495873 0.503445
13 0.073929 0.184401 2.494307 0.263176
14 0.036964 0.191639 5.184418 0.115318
15 0.014786 0.202331 13.684184 0.041389
16 0.004621 0.218199 47.223665 0.011818
17 0.001087 0.242829 223.354870 0.002577
18 0.000181 0.285180 1573.856300 0.000402
19 0.000019 0.377354 19784.197000 0.000040
20 0.000001 1.000000 1048576.000000 0.000002
```

- The latter uses the former to numerically solve equations to find the 2 values of q_0 above and below \hat{q} such that the p -value of the two-sided test of $q = q_0$ is exactly α ; this determines the $100(1 - \alpha)\%$ confidence interval.

```
binomLR.CI(15, 20, conf.level=c(0.95, 0.99))
```

```
$lower
```

```
[1] 0.5269067 0.4420938
```

```
$upper
```

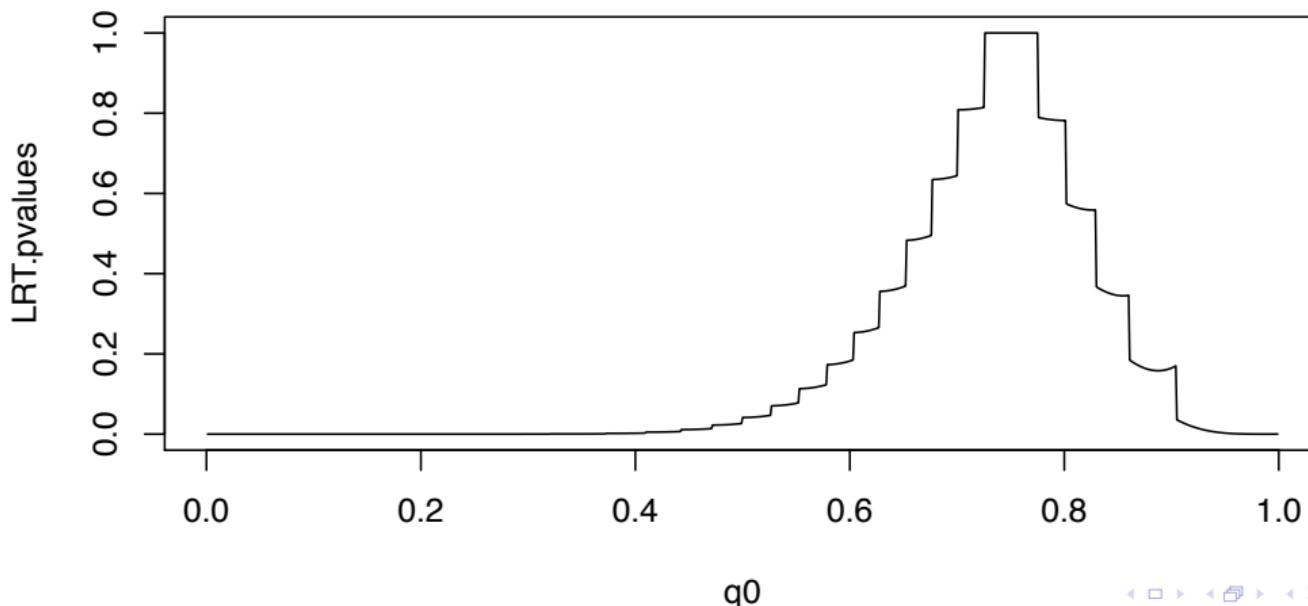
```
[1] 0.9046099 0.9311812
```

```
$conf.level
```

```
[1] 0.95 0.99
```

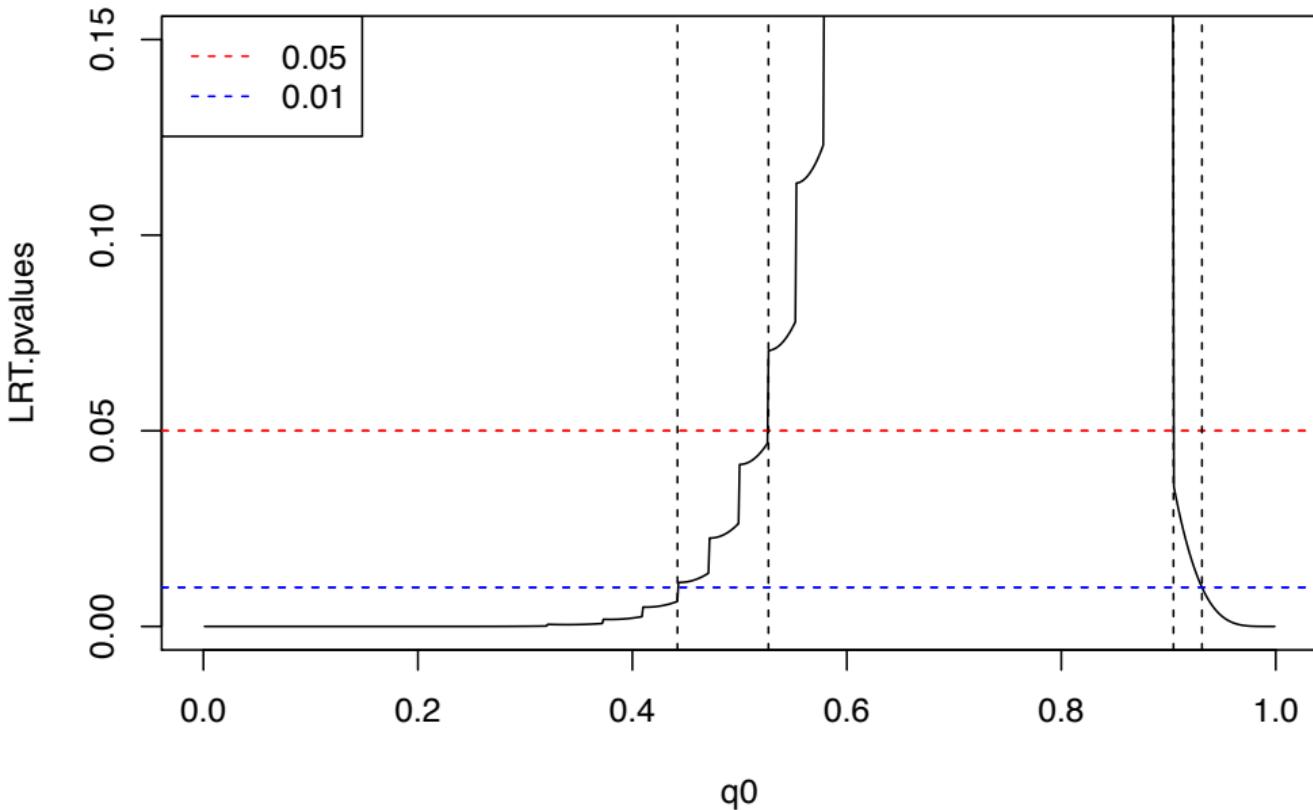
- To illustrate how these work we firstly use `binomLRtest()` to create p -value plots.

```
source("binomLRT.R")
q0=(1:999)/1000
LRT.pvalues=0
for (i in 1:999) LRT.pvalues[i]=binomLRtest(15,20,q0[i])$p.value
plot(q0,LRT.pvalues,type="l")
```



Zoomed-in version

```
plot(q0, LRT.pvalues, type="l", ylim=c(0, 0.15))
abline(h=c(0.05, 0.01), lty=2, col=c("red", "blue"))
CIs=binomLR.CI(15, 20, conf=c(0.95, 0.99))
abline(h=c(0.05, 0.01), lty=2, col=c("red", "blue"))
abline(v=c(CIs$lower[1], CIs$lower[2], CIs$upper[1], CIs$upper[2]))
legend("topleft", legend=c("0.05", "0.01"), lty=2,
       col=c("red", "blue"))
```



Outline

1 Welcome

2 Data Analysis

3 Probability

4 Inference Part 1: Continuous models

5 Inference Part 2: Discrete models

Lecture 19

Lecture 20

Lecture 21

Lecture 22

- Comparing procedures
- Exact/valid procedures: Clopper-Pearson vs LRT-based
- Approximate procedures: Wald vs Wilson

Comparing procedures

- You may be asking yourselves various questions like:
 - ▶ Why are there all these different intervals?
 - ▶ Isn't there just a "best" one?
 - ▶ Which one should I use?
- These are all reasonable questions. Unfortunately there are not always easy answers.
- For the moment we briefly discuss the following points:
 - ▶ What is better, approximate or exact/valid intervals?
 - ▶ Of each type, is there a "best" one?

Four intervals

- We shall compare two *approximate* intervals:
 - ▶ the Wilson interval;
 - ▶ the Wald interval.
- We shall also compare two *valid/exact* intervals:
 - ▶ the (exact) likelihood-ratio based interval;
 - ▶ the Clopper-Pearson interval.
- We shall compare these on the basis of
 - ▶ coverage probability;
 - ▶ “average” length.

Approximate or exact/valid?

- Some practitioners **insist** that all confidence intervals should be valid wherever possible, and that certainly *valid* procedures should always be preferred over *approximate* ones.
 - ▶ This is rather hard to argue with but ultimately comes down to a philosophical issue: what exactly is a confidence interval?
- Some practitioners are not too worried if a confidence interval is “approximate”, but if so then there might be other nice properties it should have:
 - ▶ one such property is that the “average” (in some sense) coverage probability is close to (or preferably exceeds) the nominal level.

Coverage probabilities

- If a $100(1 - \alpha)$ confidence interval is obtained by inverting a test, then remember:

q_0 is in the interval \Leftrightarrow p -value of test of $q = q_0$ is $> \alpha$.

- This *duality* can be used to compute the coverage probability of the procedure:

$$P_{q_0}(q_0 \text{ in the interval}) = P_{q_0}(p\text{-value} > \alpha).$$

Using functions in the `binom` package.

- Recall the `binom` R package has various functions for computing (mainly) *approximate* confidence intervals.
- In fact it has the following which we will find useful:
 - ▶ `binom.wilson()` (as we have already seen)
 - ▶ `binom.asymp()`: the Wald interval
 - ▶ `binom.exact()`: the Clopper-Pearson interval (same as `binom.test()` but more convenient)
- We also have our homemade `binomLR.CI()`

Strategy

- The strategy is to fix an n , say $n = 12$ and then
 - ① compute all possible intervals for $x = 0, 1, 2, \dots, n$
 - ② define a vector of q_0 values
 - ③ for each such value
 - ① compute the vector of probabilities for each x
 - ② determine the probability of coverage directly
 - ③ work out the average (weighted by probabilities) length

```
ClopperPearson=binom.exact(x=0:12, n=12)
```

```
ClopperPearson
```

	method	x	n	mean	lower	upper
1	exact	0	12	0.00000000	0.00000000	0.2646485
2	exact	1	12	0.08333333	0.002107593	0.3847962
3	exact	2	12	0.16666667	0.020862525	0.4841377
4	exact	3	12	0.25000000	0.054860645	0.5718585
5	exact	4	12	0.33333333	0.099246091	0.6511245
6	exact	5	12	0.41666667	0.151652230	0.7233303
7	exact	6	12	0.50000000	0.210944638	0.7890554
8	exact	7	12	0.58333333	0.276669686	0.8483478
9	exact	8	12	0.66666667	0.348875506	0.9007539
10	exact	9	12	0.75000000	0.428141538	0.9451394
11	exact	10	12	0.83333333	0.515862251	0.9791375
12	exact	11	12	0.91666667	0.615203835	0.9978924
13	exact	12	12	1.00000000	0.735351531	1.0000000

- The last columns are extracted as elements “lower” and “upper”

```
cbind(ClopperPearson$lower , ClopperPearson$upper)
```

```
          [,1]      [,2]
[1,] 0.000000000 0.2646485
[2,] 0.002107593 0.3847962
[3,] 0.020862525 0.4841377
[4,] 0.054860645 0.5718585
[5,] 0.099246091 0.6511245
[6,] 0.151652230 0.7233303
[7,] 0.210944638 0.7890554
[8,] 0.276669686 0.8483478
[9,] 0.348875506 0.9007539
[10,] 0.428141538 0.9451394
[11,] 0.515862251 0.9791375
[12,] 0.615203835 0.9978924
[13,] 0.735351531 1.0000000
```

What is “the Clopper-Pearson” anyway?

- This is an old procedure which is (fortunately) easy to describe.
- Recall the *one-sided (and exact)* confidence intervals, i.e. the lower and upper confidence limits.
- The 95% Clopper Pearson interval is simply:

(97.5% lower confidence limit, 97.5% upper confidence limit)

- There is thus
 - ▶ (at most) a 0.025 probability that the lower endpoint $> q_0$ and
 - ▶ (at most) a 0.025 probability that the upper endpoint $< q_0$.
- Thus the *non-coverage probability* is (at most) $0.025 + 0.025 = 0.05$.
- The “at most”’s here are due to the discreteness, and lead to the procedure being (overly) conservative.
- This is an inversion of the test which *doubles the smallest one-sided p-value*, i.e. $p\text{-value} = 2 \min(P(X \leq x), P(X \geq x))$ where $X \sim B(n, q_0)$.

binomLR.CI()

```
LRTlower=LRTupper=0
for (i in 1:13) {
  b=binomLR.CI(i-1,12)
  LRTlower[i]=b$lower
  LRTupper[i]=b$upper
}
cbind(LRTlower ,LRTupper)
```

	LRTlower	LRTupper
[1,]	0.000000000	0.2355110
[2,]	0.004262902	0.3545012
[3,]	0.030439998	0.4522526
[4,]	0.071872059	0.5478344
[5,]	0.122866629	0.6455674
[6,]	0.163738971	0.7060636
[7,]	0.199891521	0.8001085
[8,]	0.293936352	0.8362610
[9,]	0.354432603	0.8771334
[10,]	0.452165599	0.9281279
[11,]	0.547747393	0.9695600
[12,]	0.645498758	0.9957371
[13,]	0.764489020	1.0000000

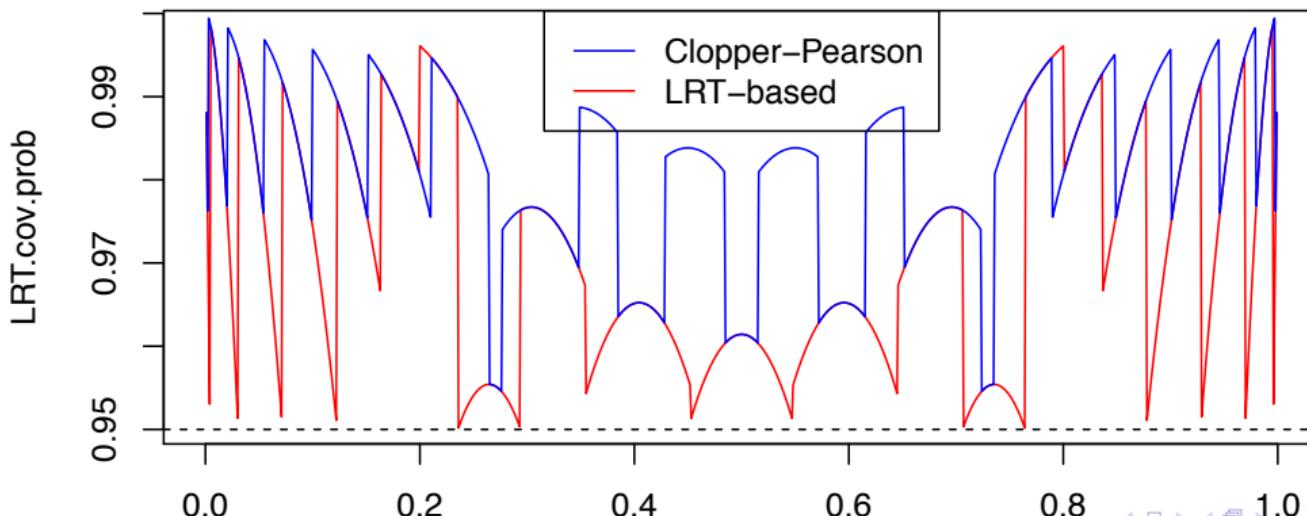
Computations

```
q0=(1:999)/1000
Clop.cov.prob=0
LRT.cov.prob=0
Clop.av.len=0
LRT.av.len=0
LRT.lens=LRTupper-LRTlower
Clop.lens=ClopperPearson$upper-ClopperPearson$lower
for (i in 1:999){
  probs=dbinom(0:12,12,q0[i])
  Clop.covers=(ClopperPearson$lower<=q0[i])&(ClopperPearson$upper>=q0[i])
  Clop.cov.prob[i]=sum(probs[Clop.covers])
  Clop.av.len[i]=sum(probs*Clop.lens)
  LRT.covers=(LRTlower<=q0[i])&(LRTupper>=q0[i])
  LRT.cov.prob[i]=sum(probs[LRT.covers])
  LRT.av.len[i]=sum(probs*LRT.lens)
}
```

Plots

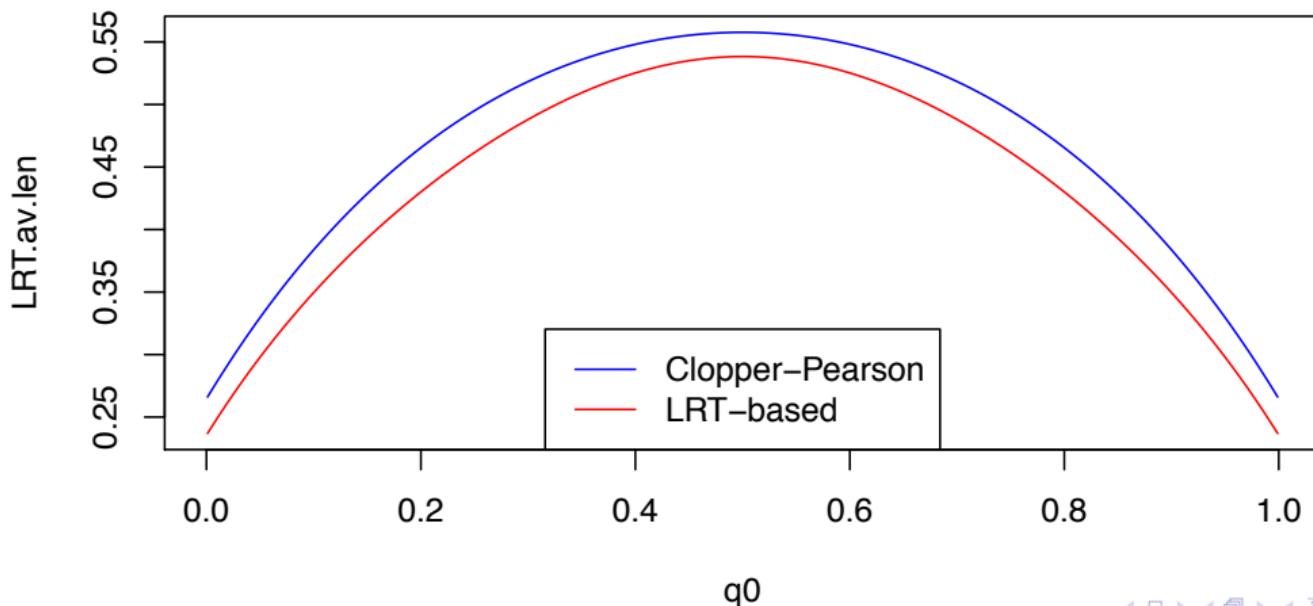
```
plot(q0,LRT.cov.prob,type="l",col="red",main="Coverage\u2225Probability\u222595%\u2225n=12")
abline(h=0.95,lty=2)
lines(q0,Clop.cov.prob,type="l",col="blue")
legend("top",col=c("blue","red"),lty=c(1,1),leg=c("Clopper-Pearson","LRT-based"))
```

Coverage Probability 95% n=12



```
plot(q0,LRT.av.len,type="l",col="red",main="Average Length 95% n=12",ylim=range  
abline(h=0.95,lty=2)  
lines(q0,Clop.av.len,type="l",col="blue")  
legend("bottom",col=c("blue","red"),lty=c(1,1),leg=c("Clopper-Pearson","LRT-based"))
```

Average Length 95% n=12



Summary

- Both procedures are “valid/conservative”.
- Clopper-Pearson for (almost) all q_0 has higher coverage probability (too high really) than the LRT-based interval.
- The average length of the LRT-based interval is uniformly less than the Clopper-Pearson.
- All-in-all,
 - ▶ the LRT-based interval is better (according to these measures at least) and is **recommended**;
 - ▶ the Clopper-Pearson is **not recommended**.

LRT-based interval

- There is a neat alternate way we can determine the coverage probability for the LRT-based interval.
- Suppose we want the coverage probability of the 95% LRT-based confidence interval when $n = 12$ and $q_0 = 0.3$.
- Recall that our home-made function `binomLRtest()` returns a table of all possible values and corresponding p -values.
- There are 13 possible observations, thus (at most) 13 different values of the LR statistic and thus (at most) 13 different p -values.

```
binomLRtest(3,12,0.4)$table
```

x	probs0	max.probs	ratio	p.values
0	0.002177	1.000000	459.393660	0.002496
1	0.017414	0.383995	22.050622	0.022401
2	0.063852	0.296094	4.637165	0.140753
3	0.141894	0.258104	1.818989	0.383550
4	0.212841	0.238446	1.120302	0.772970
5	0.227030	0.228605	1.006938	1.000000
6	0.176579	0.225586	1.277534	0.560129
7	0.100902	0.228605	2.265610	0.241656
8	0.042043	0.238446	5.671527	0.076901
9	0.012457	0.258104	20.719426	0.034858
10	0.002491	0.296094	118.845470	0.004987
11	0.000302	0.383995	1271.550000	0.000319
12	0.000017	1.000000	59604.645000	0.000017

- The p -value > 0.05 if and only if the LR statistic ≤ 5.67 .
- Conversely, p -value ≤ 0.05 if and only if the LR statistic ≥ 20.719 .
- From the table,

$$P(LR \geq 20.719) = 0.0349.$$

- Thus the coverage probability is 1 minus this, i.e.

$$\begin{aligned} P\{\text{interval contains } 0.3\} &= P\{p\text{-value} > 0.05\} \\ &= 1 - P\{p\text{-value} \leq 0.05\} \\ &= 1 - 0.0349 \\ &= 0.9651 \end{aligned}$$

- So in fact, the coverage probability of the (nominal) 95% LRT-based confidence interval is actually 96.51% when the true value is 0.3.

```
Wald=binom.asymp(x=0:12,n=12)
```

```
Wald
```

	method	x	n	mean	lower	upper
1	asymptotic	0	12	0.00000000	0.00000000	0.00000000
2	asymptotic	1	12	0.08333333	-0.073043554	0.2397102
3	asymptotic	2	12	0.16666667	-0.044191885	0.3775252
4	asymptotic	3	12	0.25000000	0.005004502	0.4949955
5	asymptotic	4	12	0.33333333	0.066616018	0.6000506
6	asymptotic	5	12	0.41666667	0.137727021	0.6956063
7	asymptotic	6	12	0.50000000	0.217103566	0.7828964
8	asymptotic	7	12	0.58333333	0.304393688	0.8622730
9	asymptotic	8	12	0.66666667	0.399949351	0.9333840
10	asymptotic	9	12	0.75000000	0.505004502	0.9949955
11	asymptotic	10	12	0.83333333	0.622474781	1.0441919
12	asymptotic	11	12	0.91666667	0.760289779	1.0730436
13	asymptotic	12	12	1.00000000	1.00000000	1.00000000

- Again, the last columns are extracted as elements “lower” and “upper”

```
cbind(Wald$lower , Wald$upper)
```

```
 [,1]      [,2]
[1,] 0.000000000 0.0000000
[2,] -0.073043554 0.2397102
[3,] -0.044191885 0.3775252
[4,] 0.005004502 0.4949955
[5,] 0.066616018 0.6000506
[6,] 0.137727021 0.6956063
[7,] 0.217103566 0.7828964
[8,] 0.304393688 0.8622730
[9,] 0.399949351 0.9333840
[10,] 0.505004502 0.9949955
[11,] 0.622474781 1.0441919
[12,] 0.760289779 1.0730436
[13,] 1.000000000 1.0000000
```

What is “the Wald interval” anyway?

- The $100(1 - \alpha)\%$ Wald interval, also called “the approximate interval” (e.g. in Phipps and Quine) or “the asymptotic interval” (e.g. in the manual for the `binom` R package) is given by

$$\hat{q} \pm c \sqrt{\frac{\hat{q}(1 - \hat{q})}{n}}$$

where $P(Z > c) = \alpha/2$.

- The name “Wald” arises because it is the inversion of the Wald test of $q = q_0$ which uses as test statistic

$$\frac{\hat{q} - q_0}{\sqrt{\frac{\hat{q}(1 - \hat{q})}{n}}}$$

and computes an (approximate) p -value assuming the test statistic is approximately $N(0, 1)$ under the null hypothesis.

Wilson interval

```
Wilson=binom.wilson(x=0:12, n=12)
```

```
Wilson
```

	method	x	n	mean	lower	upper
1	wilson	0	12	0.00000000	0.00000000	0.2424940
2	wilson	1	12	0.08333333	0.01486509	0.3538799
3	wilson	2	12	0.16666667	0.04696514	0.4480309
4	wilson	3	12	0.25000000	0.08894167	0.5323053
5	wilson	4	12	0.33333333	0.13812009	0.6093779
6	wilson	5	12	0.41666667	0.19326031	0.6804887
7	wilson	6	12	0.50000000	0.25378160	0.7462184
8	wilson	7	12	0.58333333	0.31951131	0.8067397
9	wilson	8	12	0.66666667	0.39062209	0.8618799
10	wilson	9	12	0.75000000	0.46769467	0.9110583
11	wilson	10	12	0.83333333	0.55196914	0.9530349
12	wilson	11	12	0.91666667	0.64612009	0.9851349
13	wilson	12	12	1.00000000	0.75750599	1.0000000

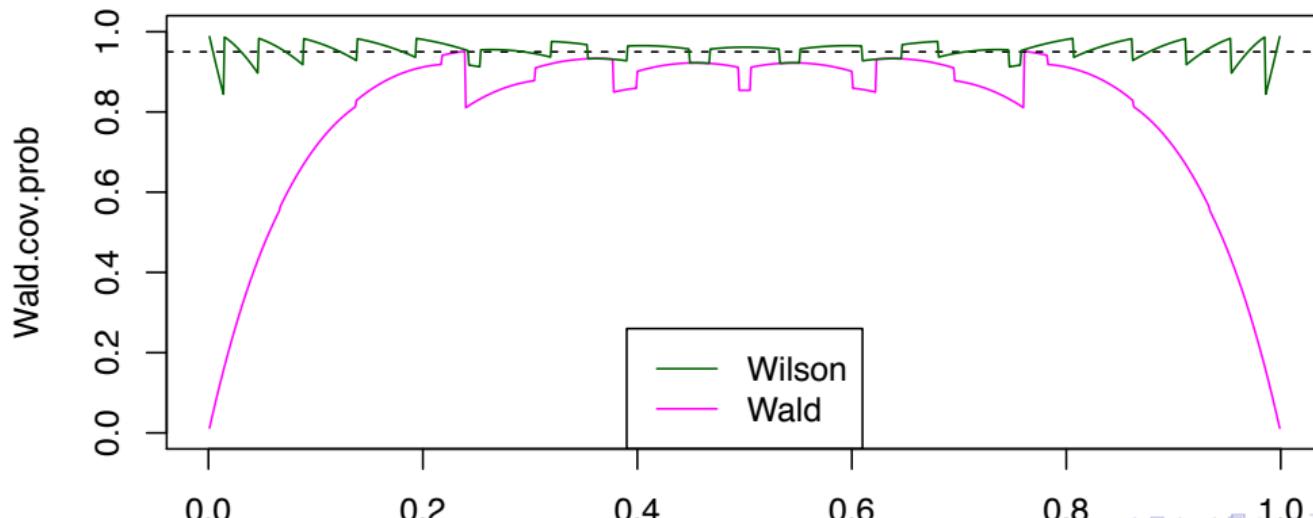
Computations

```
q0=(1:999)/1000
Wald.cov.prob=0
Wilson.cov.prob=0
Wald.av.len=0
Wilson.av.len=0
Wald.lens=Wald$upper-Wald$lower
Wilson.lens=Wilson$upper-Wilson$lower
for (i in 1:999){
  probs=dbinom(0:12,12,q0[i])
  Wald.covers=(Wald$lower<=q0[i])&(Wald$upper>=q0[i])
  Wald.cov.prob[i]=sum(probs[Wald.covers])
  Wald.av.len[i]=sum(probs*Wald.lens)
  Wilson.covers=(Wilson$lower<=q0[i])&(Wilson$upper>=q0[i])
  Wilson.cov.prob[i]=sum(probs[Wilson.covers])
  Wilson.av.len[i]=sum(probs*Wilson.lens)
}
```

Plots

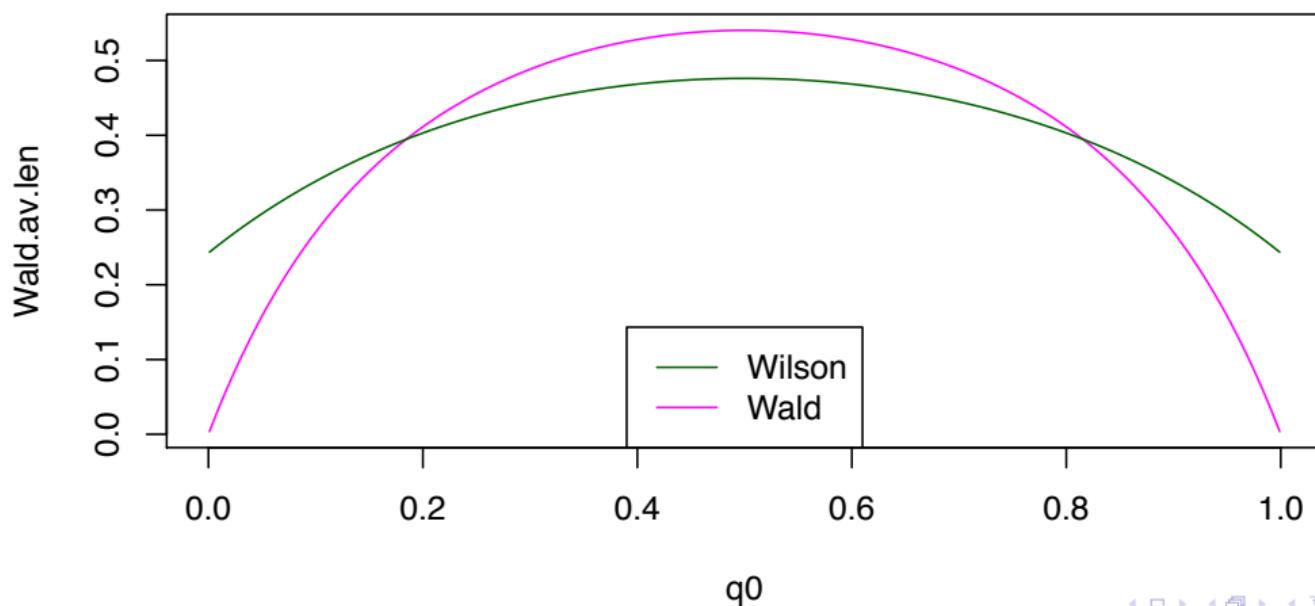
```
plot(q0,Wald.cov.prob,type="l",col="magenta",main="Coverage Probability 95% n=12")
abline(h=0.95,lty=2)
lines(q0,Wilson.cov.prob,type="l",col="DarkGreen")
legend("bottom",col=c("DarkGreen","magenta"),lty=c(1,1),leg=c("Wilson","Wald"))
```

Coverage Probability 95% n=12



```
plot(q0,Wald.av.len,type="l",col="magenta",main="Average Length 95% n=12")
abline(h=0.95,lty=2)
lines(q0,Wilson.av.len,type="l",col="DarkGreen")
legend("bottom",col=c("DarkGreen","magenta"),lty=c(1,1),leg=c("Wilson","Wald"))
```

Average Length 95% n=12



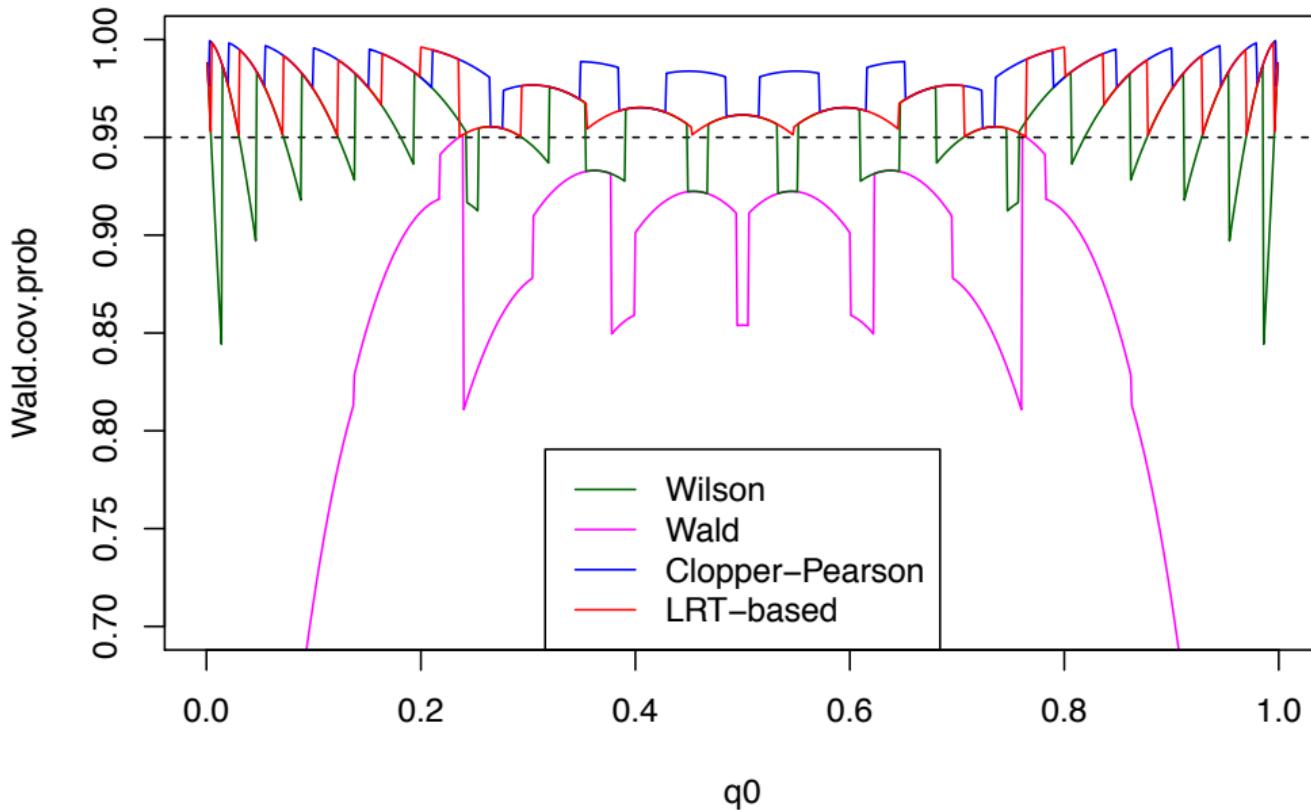
Summary

- Both procedures are “approximate”.
- The Wald interval clearly has **terrible** coverage probability properties, particularly near 0 and 1. Also, it almost never exceeds the nominal level.
- The Wilson interval on the other hand, seems to “oscillate” roughly about the nominal level.
- The Wilson interval also manages to have much shorter intervals, at least near 0.5.
- The Wald intervals shorter average length near 0 and 1 come at a price: terrible coverage probability there.
- In all,
 - ▶ the Wilson interval is better (according to these measures at least) and is **recommended**;
 - ▶ the Wald interval is **not recommended**.

All four

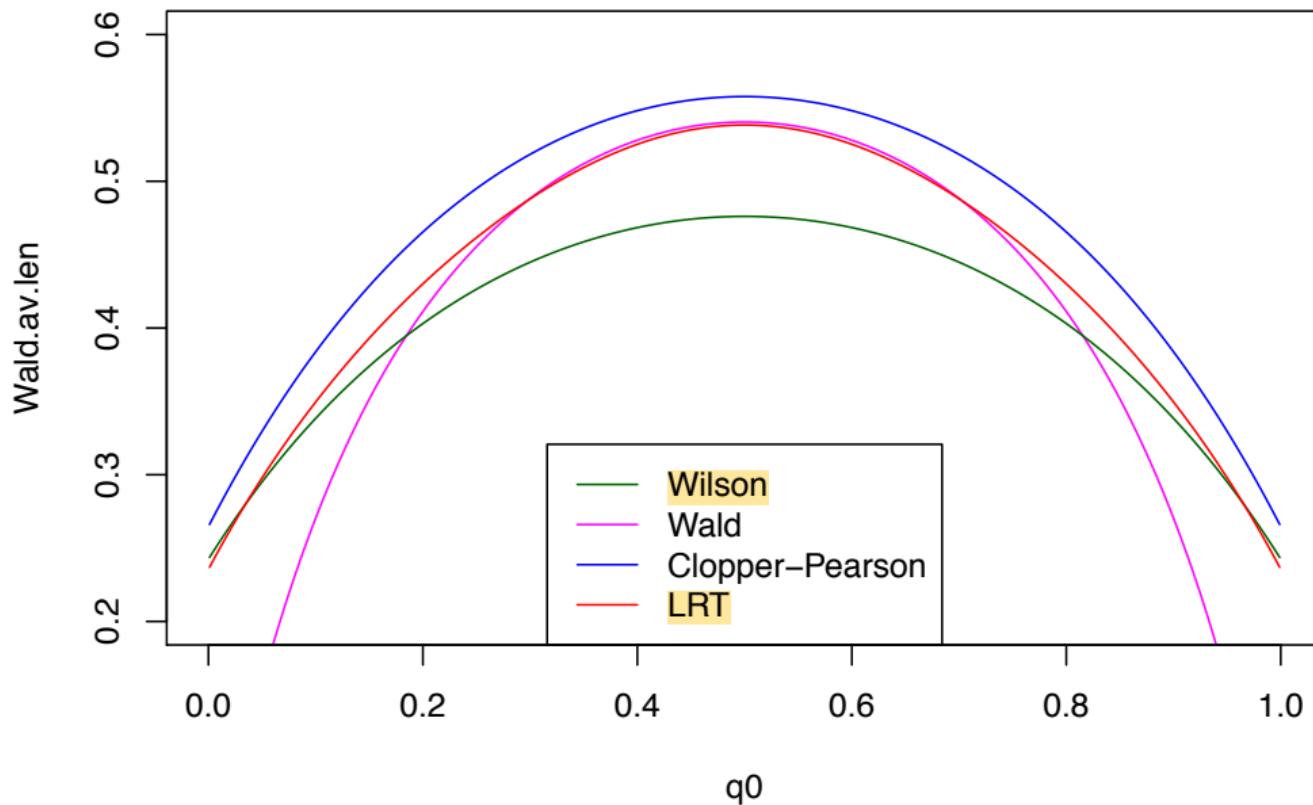
```
plot(q0,Wald.cov.prob,type="l",col="magenta",main="Coverage Probability Curves")
abline(h=0.95,lty=2)
lines(q0,Wilson.cov.prob,type="l",col="DarkGreen")
lines(q0,Clop.cov.prob,type="l",col="blue")
lines(q0,LRT.cov.prob,type="l",col="red")
legend("bottom",col=c("DarkGreen","magenta","blue","red"),lty=c(1,2,1,1))
```

Coverage Probability 95% n=12



```
plot(q0,Wald.av.len,type="l",col="magenta",main="Average Length")
abline(h=0.95,lty=2)
lines(q0,Wilson.av.len,type="l",col="DarkGreen")
lines(q0,Clop.av.len,type="l",col="blue")
lines(q0,LRT.av.len,type="l",col="red")
legend("bottom",col=c("DarkGreen","magenta","blue","red"),lty=c(1,
```

Average Length 95% n=12



Summary

- The Wald interval is so wide near 0.5 it is almost as wide as the LRT-based (exact) interval, but has coverage probability there below the nominal level. **It clearly performs poorly.**
- It seems that we recommend **both** the Wilson intervals and LRT-based intervals. Which of these is “better”?
 - ▶ It depends: some practitioners are happy with “average” coverage probability near the nominal level, and so would prefer the Wilson interval since it gives much shorter intervals on average than the LRT-based interval.
 - ▶ “Purists” would prefer the LRT-based interval since it **guarantees** coverage probability at least as large as the nominal level, but manages to give (relatively) short intervals. Also the LR test has other nice properties so there is some benefit from inverting it to get a confidence interval.
 - ▶ One slight factor in favour of the Wilson interval: it can be computed “by hand”, although most people would use a computer for this “calculation” anyway these days:

$$\frac{\hat{q} + \frac{c}{2n} \pm \sqrt{\frac{c\hat{q}(1-\hat{q})}{n} + \frac{c^2}{4n^2}}}{1 + \frac{c}{n}}$$

As usual $\hat{q} = x/n$, $P(\chi_1^2 \geq c) = \alpha$. Note this full formula is **not examinable**.