Semester 2, 2012 (Last adjustments: August 16, 2012)

Lecture Notes

MATH1905 Statistics (Advanced)

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod (at) sydney.edu.au

Semester 1, 2012 (Last adjustments: August 16, 2012)

Monday, 30th July 2012

Lecture 1 - Content

- ☐ **Outline of MATH1905**
- ☐ **First definitions**
- ☐ **Types of variables**
- ☐ **A very short introduction to R**
- ☐ **Visualizing data**

See Phipps & Quine Chapter 1, Sections 1.1 and 1.2.

Outline of MATH1905

"Alea jacta est - The die has been cast."

Julius Caesar, 10 January 49 BC

- Mathematical problems on games of chance date back to 1494 (Pacioli from Italy): What is the distribution of revenue?

  Definition 1. Statistics is the science of collecting, organizing, interpreting and reporting data.

- Probability (theory) is the appropriate language of statistics.

Knowledge based on evidence

- The scientific method is about getting knowledge based on (hard) evidence.
- This involves the following steps:
  1. Formulate question
  2. Collect relevant data
  3. Do statistical analysis of data
  4. Draw conclusions

MATH1905 – Statistics (Advanced) will help to get you started.

Unit information sheet:

- Web sites: `www.maths.usyd.edu.au/...`
  - `/u/UG/JM/MATH1905/` (for School of Mathemetics and Statistics material) or
  - `/u/jormerod/math1905/loc` (for John Ormerod's material).
- Lectures
- Tutorials
- Assessment
  - Exam
  - Quizzes
  - Assignments
- No textbook

Week-by-week lecture summary

1. Data analysis

- **Week 1:** Stem and leaf plots; relative frequencies and probability; histograms; 5-figure summaries; boxplots; R introduction.
- **Week 2:** Σ notation; sample mean; sample variance; bivariate data; correlation. $\Rightarrow$ tutorial classes start
- **Week 3:** Linear regression; residual plots; data analysis using R.

2. Probability

- **Week 4:** Axioms of probability; Venn diagrams; de Morgan's laws; inclusion-exclusion principle; counting principles; sampling; Bayes rule; independence.
- **Week 5:** Integer valued random variables; unordered selections; discrete random variables; mean and variance; probability generating functions.
- **Week 6:** Continuous rv's; mean and variance; standardized rv's; normal rv's.
- **Week 7:** Independent rv's; sums of independent normal rv's; sampling distributions; central limit theorem; normal approximation to binomial.
  $\Rightarrow$ QUIZ 1 in tutorial classes!

3. Statistical inference

- **Week 8:** Hypothesis testing; 1-sided and 2-sided test for a proportion; sign test.
- **Week 9:** Two sample binomial test; one sample Z-test; one sample t-test. $\Rightarrow$ ASSIGNMENT DUE!
- **Week 10:** Review of Z-test and t-test.
- **Week 11:** Two sample t-test; confidence intervals; confidence bounds.
- **Week 12:** χ^2 goodness of fit test. $\Rightarrow$ QUIZ 2 in tutorial classes!

4. Review

- **Week 13:** Review of data analysis, probability and statistical inference; past exam papers.

Introduction

Definition 2. A population is the set of all possible measurements of interest.

Definition 3. A sample is a subset of measurements from the population.

Definition 4. Data is the collection of measured variables.

Example (Length of words)**.** The first three stanzas of Waltzing Matilda are:

> Once a jolly swagman camped by a billabong Under the shade of a coolibah tree, And he sang as he watched and waited till his billy boiled: "You'll come a-waltzing Matilda, with me."
>
> Waltzing Matilda, waltzing Matilda You'll come a-waltzing Matilda, with me And he sang as he watched and waited till his billy boiled: "You'll come a-waltzing Matilda, with me."
>
> Down came a jumbuck to drink at that billabong. Up jumped the swagman and grabbed him with glee. And he sang as he shoved that jumbuck in his tucker bag: "You'll come a-waltzing Matilda, with me."

(Source: http://en.wikipedia.org/wiki/Waltzing_Matilda)

Having collecting the above data we now consider organizing the above words in terms of the number of letters in each word.

Example (Length of words – continued)**.** Variable character count of the $n = 15$ words in the first two linesof Waltzing Matilda, i.e.,

> Once a jolly swagman camped by a billabong Under the shade of a coolibah tree

is:

$$x_1 = 4,\ x_2 = 1,\ x_3 = 5,\ x_4 = 7,\ x_5 = 6,\ x_6 = 2,\ x_7 = 1,\ x_8 = 9,$$
$$x_9 = 5,\ x_{10} = 3,\ x_{11} = 5,\ x_{12} = 2,\ x_{13} = 1,\ x_{14} = 8,\ x_{15} = 4.$$

Population: all $N = 251$ possible word length counts in the entire poem; sample: the $n = 15$ words.

Types of variables

- **Nominal**: information given is a name, e.g. gender;
- **Ordinal**: the measurements can be naturally ordered, e.g. good, average, bad - or storm of category 1,2,3,4 or 5;
- **Quantitative**, which can be measured and is interpretable on either scale:
  - discrete, i.e. $\in \mathbb{N}$; from counting e.g. character count in previous example;
  - continuous ($\in \mathbb{R}$; e.g. length measurement)

Small and large data sets

- In general observations are denoted by $x_1, x_2, x_3, \ldots, x_n$.
- The sample size $= n$.
- If $n < 30 \Rightarrow n$ is small; if $n \geq 30 \Rightarrow n$ is large.
- Other rules of thumb exist, e.g. $n = 25, 50,$ or 100, to decide whether or not a data set is small or large.

Ordering observations

It's natural to order values. The ordered list of observations is

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}.$$

Definition 5. The ith smallest observation in a sample $x_1, x_2, x_3, \ldots, x_n$ is denoted by $x_{(i)}$ and is called the ith order statistic, $i = 1, \ldots, n$.

Example (Length of words).

The ordered values for our Waltzing Matilda data are

$$x_{(1)} = 1 \leq x_{(2)} = 1 \leq x_{(3)} = 1 \leq x_{(4)} = 2 \leq \ldots$$
$$\ldots \leq x_{(14)} = 8 \leq x_{(15)} = 9$$

This can be quickly obtained with the software R by executing

```
> x = c(4,1,5,7,6,2,1,9,5,3,5,2,1,8,4)
> sort(x)
 [1] 1 1 1 2 2 3 4 4 5 5 5 6 7 8 9
```

A very short introduction to R

What is R?

- R is a freeware 'clone' of the commercial package S-Plus based on the programming language S; (technically a 'function language'.)
- R can be downloaded (for free) from the R web site: http://cran.r-project.org/
- R has many 'inbuilt' mathematical & statistical commands.
- There are versions of R for all common operating systems.
- Reference PDF can be found on the course website.
- Many code examples in lecture/tutorial material.

A first dip into R

Elementary commands are either expressions or assignments.

- An expression is a command to simply display the result of a calculation, which is not retained in the computer's memory
- An assignment passes the result of a calculation to a variable name which is stored (but the result will not necessarily be printed out on the screen).

A simple R session

```
> 1*2 + sqrt(4) - 1/2
[1] 3.5
> x = 3.5
> (x+0.5)^2
[1] 16
```

Stored objects

- All assigned variables (or any other R objects) are stored by the computer until overwritten or explicitly deleted by the command `rm()` (for remove).
- To see what variables are stored, type `ls()` (for list) or `objects()`.

```
> x = 8; x;
[1] 8
> y = 3.1415
> ls()
[1] ".Last.value" "x"           "y"
> rm(x)
> objects()
[1] ".Last.value" "y"
```


Creating vectors in R

The command `c()` (for concatenate) creates R vectors.

```
> x = c(4,1,5,7,6,2,1,9,5,3,5,2,1,8,4)
> x
 [1] 4 1 5 6 2 1 9 5 3 5 2 1 8 4
```

The command `sort(x)` sorts the R vector `x`.

```
> sort(x)  # for sorting vectors
[1] 1 1 1 2 2 3 4 4 5 5 5 6 7 8 9
```

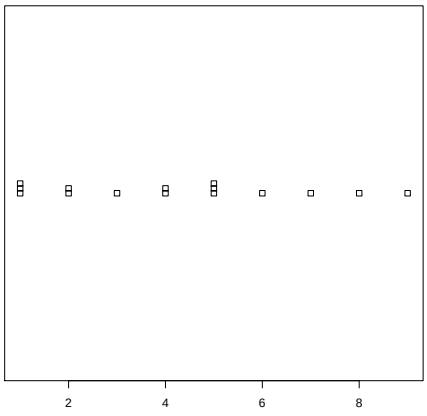
Exit R

```
> q()    # quits R
```

Visualizing data

Strip chart

- For small data sets;
- with `stripchart(x, method="stack")`.



Stem-and-leaf displays

- For small and not too large data sets;
- ordered or unordered; single, double or five stem version;
- with `stem(x, scale=1)`; if you change the `scale` parameter you get more/fewer stems - try `scale=`2^k, for $k = -2, -1, 0, 1, 2, 3$.

```
> stem(x, scale=2)

  The decimal point is at the |

  1 | 000
  2 | 00
  3 | 0
  4 | 00
  5 | 000
  6 | 0
  7 | 0
  8 | 0
  9 | 0
```

```
> stem(x, scale=1)

  The decimal point is at the |

  0 | 000
  2 | 000
  4 | 00000
  6 | 00
  8 | 00

> stem(x, scale=0.5)

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

  0 | 11122344
  0 | 5556789
```


Additional material for Lecture 1

General comment

At the end of each lecture I will provide some additional material and background information if appropriate.

More on stem-and-leaf displays

In a stem-and-leaf display all numbers are broken into two components: the stem = the leading digits; the leaf = the remaining digits. The R function `stem(x)` by default produces a stem-and-leaf display with default parameter `scale=1`. This does not mean that a single stem-and-leaf display is produced but rather what some underlying algorithm determines as the most appropriate. In practice you simply start with the default parameter. If you don't like the display either change the scale parameter to `scale=2` or `scale=0.5` or any other power of 2.

```
single stem version                double stem version         five stem version

stem | leafs                       stem | leafs               stem | leafs
   0 | 0 - 9                          0 | 0 - 4                  0 | 0 - 1
   1 | 1 1 2 3 9  (ordered)            0 | 5 - 9                  0 | 2 - 3
   2 | ...                            1 |                        0 | .
   3 | 4 2 3      (unordered)         1 |                        0 | .
   4 |                                .                          0 | 8 - 9
   .                                  .                          1 | 0 - 1
   .                                  .                           .
   .
```

Tuesday, 31st July 2012

Lecture 2 - Content

- ☐ **Absolute and relative frequencies**
- ☐ **Ordinate diagrams and histograms**
- ☐ **Cumulative frequencies and empirical distribution**
- ☐ **Five number summary**
- ☐ **Boxplot**

See Phipps & Quine Chapter 1, Sections 1.1, 1.2 and 2.

Absolute and relative frequencies

Example (Gold medals). A total of $n = 55$ countries had at least one olympic gold medal in Beijing 2008. In R absolute frequencies are obtained with `table(x)`, * = AU:

```
x
 1  2  3  4  5  6  7  8  9 13 *14* 16 19 23 36 51
19  9  9  3  2  1  3  1  1  1 * 1*  1  1  1  1  1
```

Let $x_j \in \mathbb{R}$ denote possible measurements, here $j = 1, \ldots, 16$.

Definition 6. The (absolute) frequency with which the value x_j occurred is denoted by f_j.

Definition 7. The relative frequency $:= \widehat{p}_j = \frac{f_j}{n}$.

E.g. $x_{10} = 13 \in \mathbb{R}$ has $f_{10} = 1$ and $\widehat{p}_{10} = \frac{1}{55}$ or $x_1 = 1$, $f_1 = 19$, $\widehat{p}_1 = \frac{19}{55}$.

Ordinate diagrams and histograms

- If all values are discrete (i.e. $\in \mathbb{N}$) draw an

  ordinate diagram $=$ plot of j against f_j;

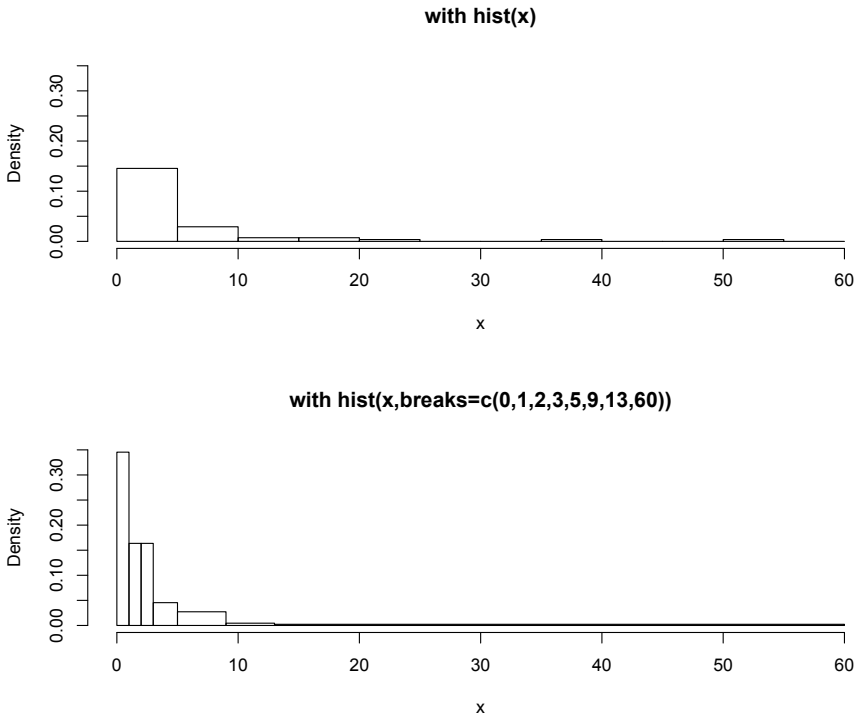
  - with `barplot(table(x))` but this omits empty x-values,

  - `barplot(tabulate(x, nbins = max(1, x)))` has a scale preserving x-axis.

- Condensing information can be very useful $\Rightarrow$ slicing $\mathbb{R}$!

- Choose intervals (or midpoints) of the form $[\, l_j\, ,\, u_j\,)$.
  E.g. $\mathbb{R} = (-\infty, 1) \cup [1, 2) \cup [2, 3) \cup [3, 5) \cup [5, 9) \cup [9, 13) \cup [13, \infty)$.

- Produce a histogram with `hist(x)`. Remember: frequencies are represented by the area and not hight.



Optimal Binwidths for Histograms (Not examinable)

There are many ways to draw a boxplot. However, the choices in how a boxplot are represented does matter!

Scott (1992) proved that the asymptotically optimal binwidth (based on various assumptions such as differentiability of the underlying density) is

$$\left(\frac{24\sqrt{\pi}}{n}\right)^{1/3}.$$

This can be used as a reasonable rule of thumb for constructing histograms. This is automated by the R command `hist(x,breaks="Scott")`.

Cumulative frequencies

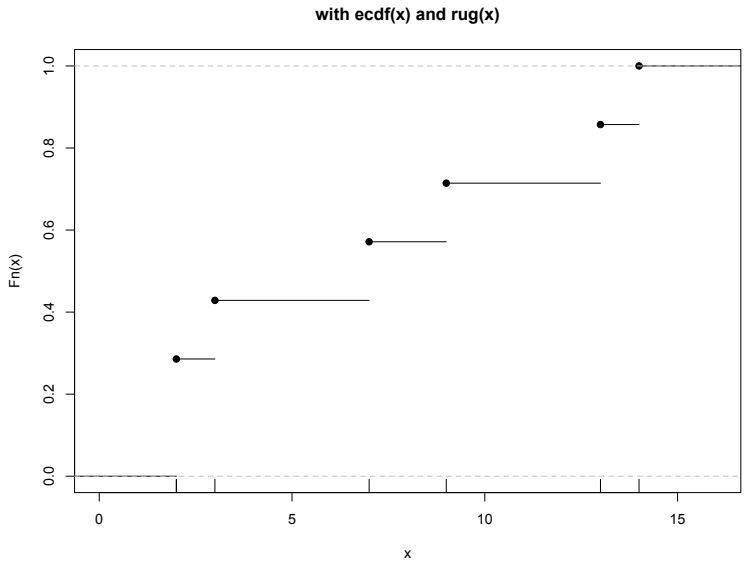
- Consider the number of gold medals for au, fr, jp, kr, nz, ch, and th:
  $x_1 = 14,\ x_2 = 7,\ x_3 = 9,\ x_4 = 13,\ x_5 = 3,\ x_6 = 2,\ x_7 = 2.$
- Ordering observations preserves the information!
  $x_{(1)} = 2 \leq x_{(2)} = 2 \leq \ldots \leq x_{(6)} = 13 \leq x_{(7)} = 14.$
- There are 6 different measurement values:

j	1	2	3	4	5	6
x_j	2	3	7	9	13	14
freq. f_j	2	1	1	1	1	1
cum. freq. F_j	2	3	4	5	6	7

  Hence, $F_j = f_1 + f_2 + \ldots + f_j$.
- Knowing frequencies or cumulative frequencies preserves the information!

Empirical distribution function (EDF)

EDFs are mathematically very useful, have many properties (monotone, continuous from the right) and can be drawn by plotting a step-function using x_j and F_j/n:



Five number summary

Definition 8. The minimum $= x_{(1)}$ and the maximum $= x_{(n)}$.

Definition 9. The range $= x_{(n)} - x_{(1)}$.

Definition 10. The median, $\widetilde{x}$, is a value such that at least half the observations (obs) are less than or equal to $\widetilde{x}$ and at least half the obs are greater or equal to $\widetilde{x}$.

Quartiles are medians of lower and upper half respectively:

Definition 11. The lower quartile, Q_1, is a value such that at least 25% of the obs are $\leq Q_1$ and at least 75% of the obs are $\geq Q_1$.

Definition 12. The upper quartile, Q_3, is a value such that at least 75% of the obs are $\leq Q_3$ and at least 25% of the obs are $\geq Q_3$.

Definition 13. The interquartile range, IQR $= Q_3 - Q_1$.

Five number summary (cont)

Definition 14. The five number summary is

$$(\text{min}, Q_1, \widetilde{x}, Q_3, \text{max}) = (Q_0, Q_1, Q_2, Q_3, Q_4)$$

and is visualized by the boxplot.

Example. Number of gold medals for au, fr, jp, kr, nz, ch, and th:

```
> x = c(14,7,9,13,3,2,2)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.500   7.000   7.143  11.000  14.000
```

or with

```
> quantile(x,c(0.00,0.25,0.50,0.75,1.00))
  0%  25%  50%  75% 100%
 2.0  2.5  7.0 11.0 14.0
```


Boxplot

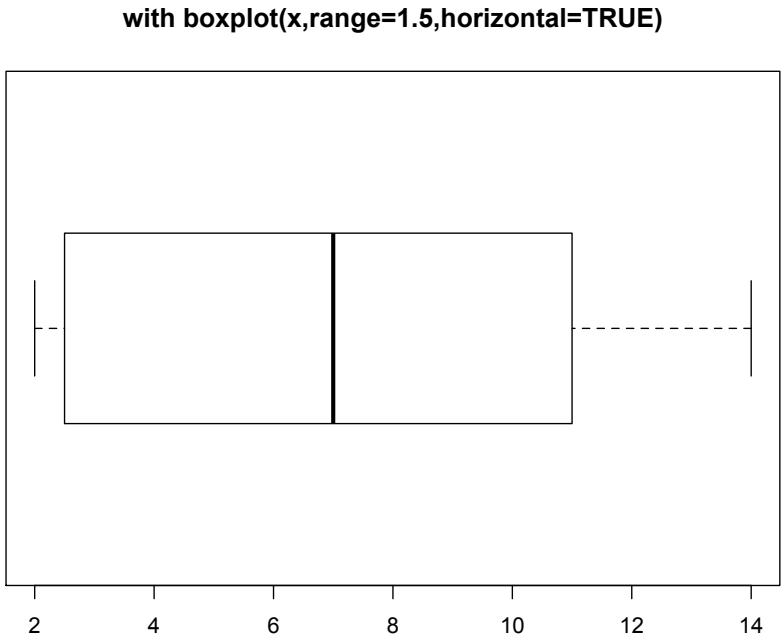
- Draw a box between Q_1 and Q_3;
- add midline at Q_2;
- draw whiskers to min and max if there are no outliers, otherwise to first point larger than LT and first point smaller than UT;
- draw all outlier candidates as points.

Definition 15. Potential outliers are points more than $r = c\times$ IQR beyond the ends of $[Q_1, Q_3]$, $c = 1.5$ is the default choice. Hence,

$$\begin{aligned}\text{Lower Threshold} &= \text{LT} = Q_1 - 1.5 \times \text{IQR},\\ \text{Upper Threshold} &= \text{UT} = Q_3 + 1.5 \times \text{IQR}.\end{aligned}$$

Other choices for c are $1, 1.5, 2, 2.5, 3, \ldots$. The larger c the fewer potential outliers are drawn as single points.

Boxplot (cont)



Boxplot (cont)

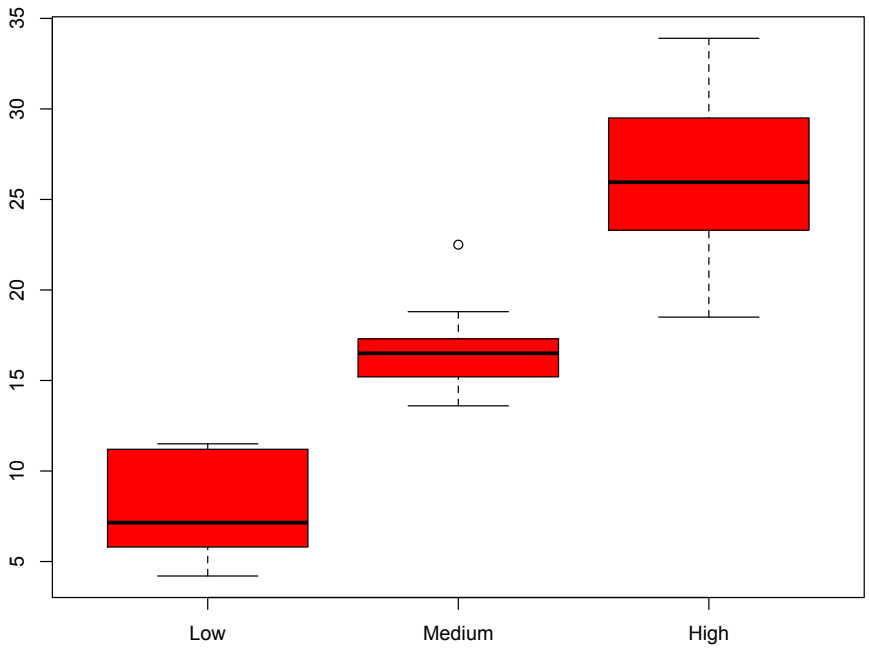
- A single boxplot is boring!
- Boxplots are powerful to compare a continuous variable (e.g. length, weight etc) with a nominal variable (e.g. treatment).
- Length of whisker in R is by default chosen to be $1.5 \times$ IQR, i.e. you don't need to specify `range = 1.5`.
- Boxplots give an easy impression of the shape of the data set:
  - Symmetrical: yes, no?
  - Skewed: left, right?
  - Right skewed = if boxplot is stretched to the right.

Example (Vitamin C and Tooth Growth).

- Data from an (old) experiment into the effects of vitamin C on tooth growth.
- 30 guinea pigs were divided (at random) into three groups of ten and treated with vitamin C (administered in orange juice).

- Group 1 dose was low, group 2 dose was medium and group 3 dose was high.
- Length of odontoblasts (teeth) measured as response variable.
- Reference: C. I. Bliss (1952) *The Statistics of Bioassay*. Academic Press.

Example (cont)



Example (cont)

```
> tapply(Length,Dose.fac,summary)
$Low
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.20    5.95    7.15    7.98   10.90   11.50

$Medium
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.60   15.27   16.50   16.77   17.30   22.50

$High
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.50   23.38   25.95   26.14   28.80   33.90
```


Comments for the five number summary

- A median can be calculated for

$$n \text{ odd:} \quad \widetilde{x} = x_{(\frac{n+1}{2})}$$
$$n \text{ even:} \quad \widetilde{x} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}).$$

- If $n/4 \in \mathbb{N}$ then $k = \frac{n}{4}$ and

$$Q_1 = \frac{1}{2}(x_{(k)} + x_{(k+1)}), \quad Q_3 = \frac{1}{2}(x_{(n-k)} + x_{(n-k+1)});$$

  otherwise $k = \lceil \frac{n}{4} \rceil$ and $Q_1 = x_{(k)}, \quad Q_3 = x_{(n-k+1)}$.

- The range covers 100% of the obs

$$x_i \in [\, x_{(1)} \,,\, x_{(n)} \,] \quad \text{for all } i = 1, \ldots, n,$$

  the IQR covers approximately 50% of the obs.

Density plots (Not examinable)

An alternative to histograms are (kernel) density plots. These are special smoothed positive functions which integrate to 1.

Example. Old Faithful is a cone geyser located in Wyoming, in Yellowstone National Park in the United States. It is also called the most predictable geographical feature on Earth erupting almost every 91 minutes. The data for length between consecutive eruptions can be obtained from the R code

```
> faithful$eruptions
  [1] 3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950 4.350 1.833 3.917 4.200 1.750 4.700 2.167 1.750 4.800
 [19] 1.600 4.250 1.800 1.750 3.450 3.067 4.533 3.600 1.967 4.083 3.850 4.433 4.300 4.467 3.367 4.033 3.833 2.017
 [37] 1.867 4.833 1.833 4.783 4.350 1.883 4.567 1.750 4.533 3.317 3.833 2.100 4.633 2.000 4.800 4.716 1.833 4.833
 [55] 1.733 4.883 3.717 1.667 4.567 4.317 2.233 4.500 1.750 4.800 1.817 4.400 4.167 4.700 2.067 4.700 4.033 1.967
 [73] 4.500 4.000 1.983 5.067 2.017 4.567 3.883 3.600 4.133 4.333 4.100 2.633 4.067 4.933 3.950 4.517 2.167 4.000
 [91] 2.200 4.333 1.867 4.817 1.833 4.300 4.667 3.750 1.867 4.900 2.483 4.367 2.100 4.500 4.050 1.867 4.700 1.783
[109] 4.850 3.683 4.733 2.300 4.900 4.417 1.700 4.633 2.317 4.600 1.817 4.417 2.617 4.067 4.250 1.967 4.600 3.767
[127] 1.917 4.500 2.267 4.650 1.867 4.167 2.800 4.333 1.833 4.383 1.883 4.933 2.033 3.733 4.233 2.233 4.533 4.817
[145] 4.333 1.983 4.633 2.017 5.100 1.800 5.033 4.000 2.400 4.600 3.567 4.000 4.500 4.083 1.800 3.967 2.200 4.150
[163] 2.000 3.833 3.500 4.583 2.367 5.000 1.933 4.617 1.917 2.083 4.583 3.333 4.167 4.333 4.500 2.417 4.000 4.167
[181] 1.883 4.583 4.250 3.767 2.033 4.433 4.083 1.833 4.417 2.183 4.800 1.833 4.800 4.100 3.966 4.233 3.500 4.366
[199] 2.250 4.667 2.100 4.350 4.133 1.867 4.600 1.783 4.367 3.850 1.933 4.500 2.383 4.700 1.867 3.833 3.417 4.233
[217] 2.400 4.800 2.000 4.150 1.867 4.267 1.750 4.483 4.000 4.117 4.083 4.267 3.917 4.550 4.083 2.417 4.183 2.217
```

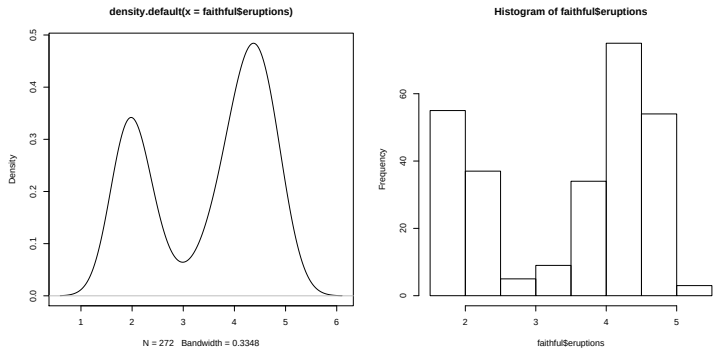


```
[235] 4.450 1.883 1.850 4.283 3.950 2.333 4.150 2.350 4.933 2.900 4.583 3.833 2.083 4.367 2.133 4.350 2.200 4.450
[253] 3.567 4.500 4.150 3.817 3.917 4.450 2.000 4.283 4.767 4.533 1.850 4.250 1.983 2.250 4.750 4.117 2.150 4.417
[271] 1.817 4.467
```

The density plot of this data can be obtained from the R code

```
> plot(density(faithful$eruptions))
```

Density plots are aesthetically pleasing to the eye when compared to histograms:



Some of the fundamental theory around density plots was developed by Australia Statisticians Matt Wand (my PhD supervisor) and Peter Hall (my grandsupervisor).

Additional material for Lecture 2

More on histograms

The best/nicest way to draw histograms is a matter of taste. The following rules serve as a guideline:

- Choose an appropriate number of intervals, e.g. $5 \leq k \leq 20$ or automated by $k = \lfloor \sqrt{n} \rfloor$, where $y = \lfloor x \rfloor \in \mathbb{N}$ is the function that returns the largest integer smaller or equal than x;
- choose appropriate interval boundaries of the form $[\ l_j\ ,\ u_j\ )$, $j = 1, \ldots, k$, e.g. equally spaced;
- determine the absolute/relative frequencies, i.e. the number of observations falling into each of the k intervals;
- draw the histogram such that the x-axis shows the *sliced* real numbers and draw rectangles on top of the histogram with **area proportional to the absolute/relative frequency**;
- don't forget to label both axes.

More on quartiles

Depending on the sample size and the sample itself it can occur that an entire interval satisfies the definition of the lower and upper quartile, respectively. To get a unique solution there exist multiple ways. The suggested unique solution on the previous slide is only one option and can be obtained in R by typing `quantile(x,type=2)`. Reading `help(quantile)` shows that this is the second unique solution out of nine implemented in R. The default option is `type=7`, i.e. if you just type `quantile(x)` this is what is done. What all definitions have in common is that a unique solution is produced by a particular weighted average of the two observations (order statistics) at either end of the interval $[x_{(k)}, x_{(k+1)}]$ and $[x_{(n-k-1)}, x_{(n-k)}]$, respectively, where $k = \lceil n/4 \rceil$.

Monday, 6th August 2012

Lecture 3 - Content

- $\sum$ **notation**
- **Sample mean**
- **Sample variance**
- **Transformation of data to symmetry**

See Phipps & Quine Chapter 1, Sections 3 and 4.

Review Σ notation

For the values $x_1 = 2$, $x_2 = 1$, $x_3 = 5$, $x_4 = 3$, i.e $n = 4$ we have:

$$\begin{aligned}
\sum_{i=1}^{4} x_i &= x_1 + x_2 + x_3 + x_4 = 2 + 1 + 5 + 3 = 11; \\
\sum_{i=1}^{n} x_i^2 &= 2^2 + 1^2 + 5^2 + 3^2 = 39; \\
\sum_{i=2}^{n-1}(3x_i + 2) &= 3\sum_{i=2}^{3} x_i + 2(\underbrace{n - 1 - 2 + 1}_{2}) = 18 + 4 = 22; \\
\sum_{i=1}^{n} cx_i &= c\sum_{i=1}^{n} x_i; \\
\sum_{i=1}^{n} c1 &= c\sum_{i=1}^{n} 1 = cn.
\end{aligned}$$

Review Σ notation

For the values $x_1 = 2$, $x_2 = 1$, $x_3 = 5$, $x_4 = 3$, $x_5 = 0$ i.e $n = 5$ we have:

$$\begin{aligned}
\sum_{i=1}^{5}\left(\frac{x_i - 3}{\sqrt{5}}\right)^2 &= \frac{1}{5}\sum_{i=1}^{5}(x_i - 3)^2 \\
&= \frac{1}{5}\sum_{i=1}^{5} x_i^2 - \frac{6}{5}\sum_{i=1}^{5} x_i + \frac{5}{5}9 \\
&= \frac{1}{5}(39 - 66 + 45) \\
&= \frac{18}{5}.
\end{aligned}$$

Sample mean, standard deviation and variance

Definition 16. The sample mean is the simple average of the observations. For observations $x_1, x_2, \ldots, x_n$

$$\overline{x}_n = \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Theorem 1. Given constants $a \in \mathbb{R}$ and $b \in \mathbb{R}$ and obs $x_1, x_2, \ldots, x_n$. Then the mean of the transformed observations $y_i = a \times x_i + b$, $i = 1, 2, \ldots, n$, is

$$\overline{y} = a \times \overline{x} + b.$$

Proof. Write down the left hand side of the equation and begin with the definition:

$$\begin{aligned}
\overline{y} &\overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} y_i \overset{\text{expand}}{=} \frac{1}{n}(y_1 + y_2 + \ldots + y_n) \\
&\overset{\text{replace}}{=} \frac{1}{n}\Big((ax_1 + b) + (ax_2 + b) + \ldots + (ax_n + b)\Big) \\
&\overset{\text{group}}{=} \frac{1}{n}\Big(a(x_1 + x_2 \ldots + x_n) + nb\Big) \\
&\overset{\text{simpl.}}{=} \frac{a}{n}\sum_{i=1}^{n} x_i + \frac{n}{n}b \\
&= a \times \overline{x} + b.
\end{aligned}$$

□

Change of working origin and unit

The theorem helps to transform data to a new working origin, a, and a new working unit, h:

$$d_i = \frac{x_i - a}{h} = \frac{1}{h}x_i - \frac{a}{h}, \quad i = 1, \ldots, n.$$

Thus,

$$\overline{d} = \frac{1}{h}\overline{x} - \frac{a}{h}$$

and solving for $\overline{x}$ yields

$$\overline{x} = h\overline{d} + a.$$

Example. Find the mean of x_i: 9.80, 9.81, 9.82, 9.84.

$a = 9.80$ and $h = 0.01$

$\Rightarrow d_i : 0, 1, 2, 4$. Thus, $\overline{d} = \frac{7}{4} = 1.75$ and $\overline{x} = 0.01 \times 1.75 + 9.80 = 9.8175$

Mean vs median

- Mean, $\overline{x}$, easier to calculate and to handle than the median, $\widetilde{x}$.
- If the data are approximately symmetric then $\overline{x} \approx \widetilde{x}$.
- If the data are skewed then the mean is pulled toward the long tail.
- $\widetilde{x}$ is robust against outliers and incorrect readings whereas $\overline{x}$ is not.

Example. Assume in the previous example 9.80 is misread as 3.80.

```
> x = c(3.80,9.81,9.82,9.84)
> mean(x)
[1] 8.3175
> median(x)
[1] 9.815
```

The mean is a Least Squares (LS) estimate!

Definition of Least Squares:

$$S(a) := \sum_{i=1}^{n} (x_i - a)^2; \quad \text{minimise} \;\; S(a).$$

Hence,

$$\begin{aligned} S(a) &= \sum (x_i^2 - 2ax_i + a^2) = \sum x_i^2 - 2a\left(\sum x_i\right) + n \times a^2 \\ &= \sum x_i^2 - 2an\overline{x} + na^2 \\ \Rightarrow \frac{\partial S(a)}{\partial a} &= S'(a) = -2n\overline{x} + 2na \end{aligned}$$

$S'(a)$ equals 0 if $a = \overline{x}$.

The median is a Least Absolute Deviation (LAD) estimate!

Definition of Least Absolute Deviation:

$$D(a) := \sum_{i=1}^{n} |x_i - a|; \quad \text{minimise} \;\; D(a).$$

For simplicity assume that $x_i \neq a$ for all x_i. Then

$$\frac{\partial D(a)}{\partial a} = -\sum_{i=1}^{n} \text{sign}(x_i - a)$$

where

$$\text{sign}(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$$

Thus a solution to $D'(a) = 0$ is the value a such that

$$\{ \text{ The number of } x_i\text{s greater than a } \} = \{ \text{ The number of } x_i\text{s less than a} \}.$$

In other words the sample median!

Sample variance and standard deviation

Definition 17. For data $x_1, x_2, \ldots, x_n$ the sample standard deviation s_x is

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

and the sample variance is

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

Alternative formula for s_x^2

($\sum$ index omitted)

$$\begin{aligned} S_{xx} &= \sum (x_i - \overline{x})^2 = \sum (x_i^2 - 2\overline{x} x_i + \overline{x}^2) \\ &= \sum x_i^2 - 2\overline{x} \sum x_i + n \times \overline{x}^2; \quad \text{with } \sum x_i = n \times \overline{x}, \\ &= \sum x_i^2 - n\overline{x}^2 \\ &= \sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2. \end{aligned}$$

Hence,

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right).$$

Change of working origin and unit (cont)

Exercise. Show that for working origin a and working unit h the variance of data $x_1, x_2, \ldots, x_n$ equals $h^2\times$ the variance of the transformed data

$$d_i = \frac{x_i - a}{h} \quad \text{i.e.} \quad x_i = h \cdot d_i + a \Rightarrow s_x^2 = h^2 \times s_d^2,$$

and therefore $s_x = h \times s_d$.

Example. Data: 340, 350, 360, 370, 380.

For $a = 360$, $h = 10$ we get

$d_i : -2, -1, 0, 1, 2$ and $\sum d_i = 0$, $\sum d_i^2 = 10$:

and

$$s_d^2 = \frac{1}{5-1}\left(10 - \frac{1}{5} \times 0^2\right) = 2.5.$$

So $s_x^2 = h^2 \times s_d^2 = 100 \times 2.5 = 250$.

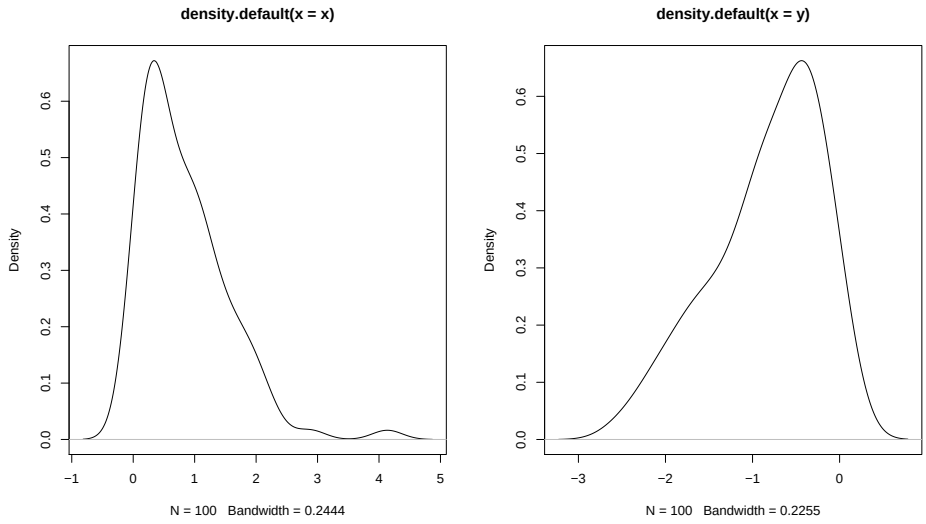
```
> x = c(340,350,360,370,380)
> var(x)
[1] 250
> sd(x)
[1] 15.81139
```

Skewed data

Definition 18. Data are said to be left skewed if the left tail of the density is longer than the right.

Definition 19. Data are said to be right skewed if the right tail of the density is longer than the left.

The data in the left hand side plot are right skewed whereas the data on the right hand side are left skewed.



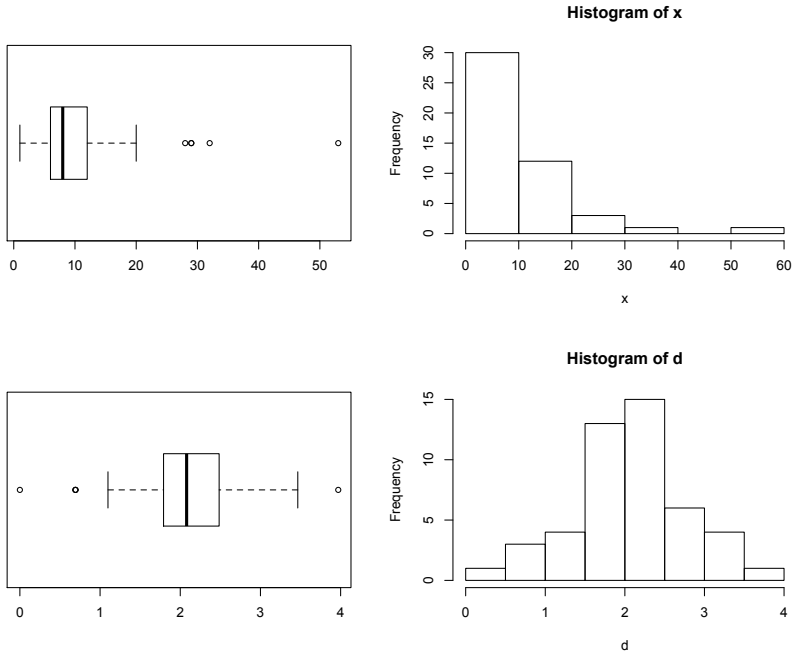
Transformations of data

- To have symmetric data can be a desirable property for some statistical methods.
- Data obtained as differences (e.g. from before/after studies) are often approximately symmetric.
- For right skewed data $\{x_i\}$
  - $d_i = x_i^a$, for various values of $a \in (0, 1)$.
  - $d_i = \log x_i$
  - $d_i = -x_i^{-a}$, for $a > 0$.
- Left skewed data $\{x_i\}$: transform into right skewed data by $d_i = -x_i$.
- For data $\{x_i\}$ recorded as proportions, i.e. $x_i \in (0, 1)$, the logit transform can be used: $d_i = \log \frac{x_i}{1-x_i}$.

Example (Swiss fertility data). Execute the following code in R:

```
> data(swiss)
> help(swiss)
> names(swiss)
> x = swiss$Education
> d = log(x)
> par(mfrow=c(2,2))
> boxplot(x,horizontal=TRUE)
> hist(x)
> boxplot(d,horizontal=TRUE)
> hist(d)
> summary(x)
> summary(d)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.00    8.00   10.98   12.00   53.00
> summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.792   2.079   2.099   2.485   3.970
```

Example (cont)



Additional material for Lecture 3

More on the sample variance and sample standard deviation

The sample variance is almost the average of squared distances to the sample mean. But instead of using n in the denominator a $(n-1)$ term is used. This will make perfect sense after STAT2011/2911 but probably not too much sense at this stage. Note that it is almost the average, particularly when n is large since $\lim_{n\to\infty} \frac{n}{n-1} = 1$ and

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

To fully appreciate to correct by $\frac{n}{(n-1)}$ the following concepts have to be understood first: random variables, function of random variables, expected value, covariance of random variables, unbiasedness.

Both the range of the sample and the standard deviation of the sample measure aspects of spread or scale. They are to a certain extent depend of each other. An equality is given in the theorem below.

Theorem (Thomson, 1955) Let $w = x_{(n)} - x_{(1)}$ be the range of the observations $x_1, \ldots, x_n$ then the sample standard deviation satisfies

$$\sqrt{\frac{1}{2(n-1)}} \leq \frac{s_x}{w} \leq \begin{cases} \frac{1}{2}\sqrt{\frac{n}{n-1}}; & n \text{ even}, \\ \frac{1}{2}\sqrt{\frac{n+1}{n}}; & n \text{ odd}. \end{cases}$$

Tuesday, 7th August 2012

Lecture 4 - Content

- **Bivariate data**
- **Scatterplot**
- **Correlation coefficient**

See Phipps & Quine Chapter 1, Section 5.

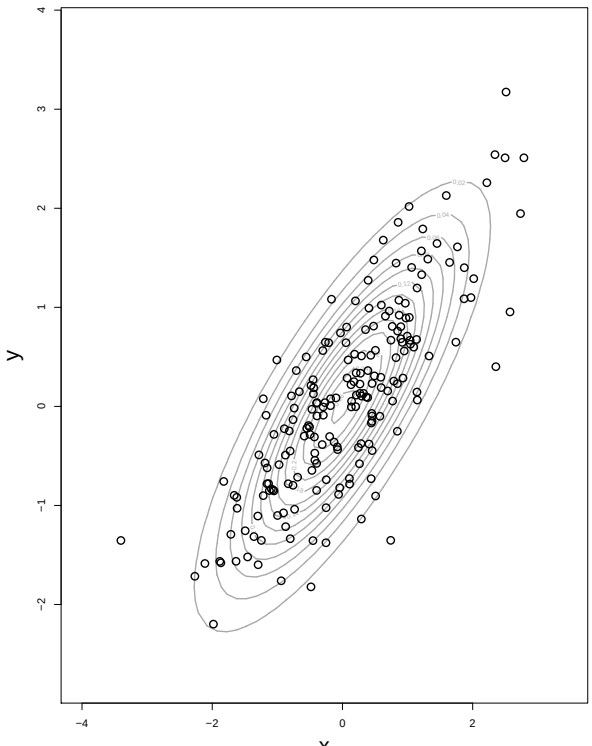
Bivariate data

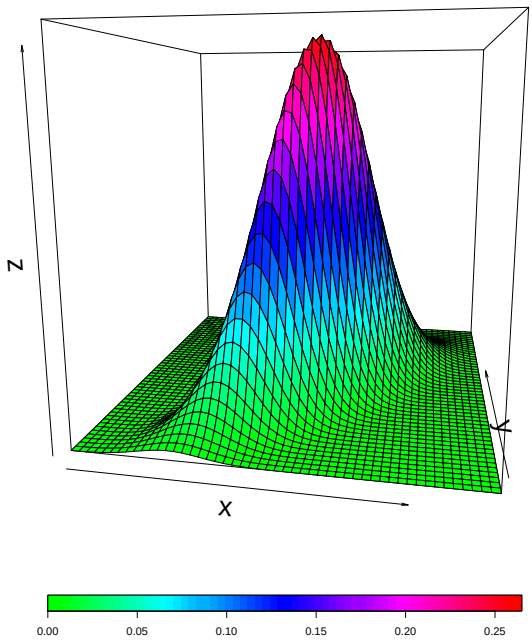
- So far univariate data only, i.e. observations on a single feature.
- In general multivariate data, e.g. bivariate data

$$x = \text{patient's age}$$
$$y = \text{patient's reaction time}$$

- The first step in the analysis of multivariate data is visualisation!

Visualisation!





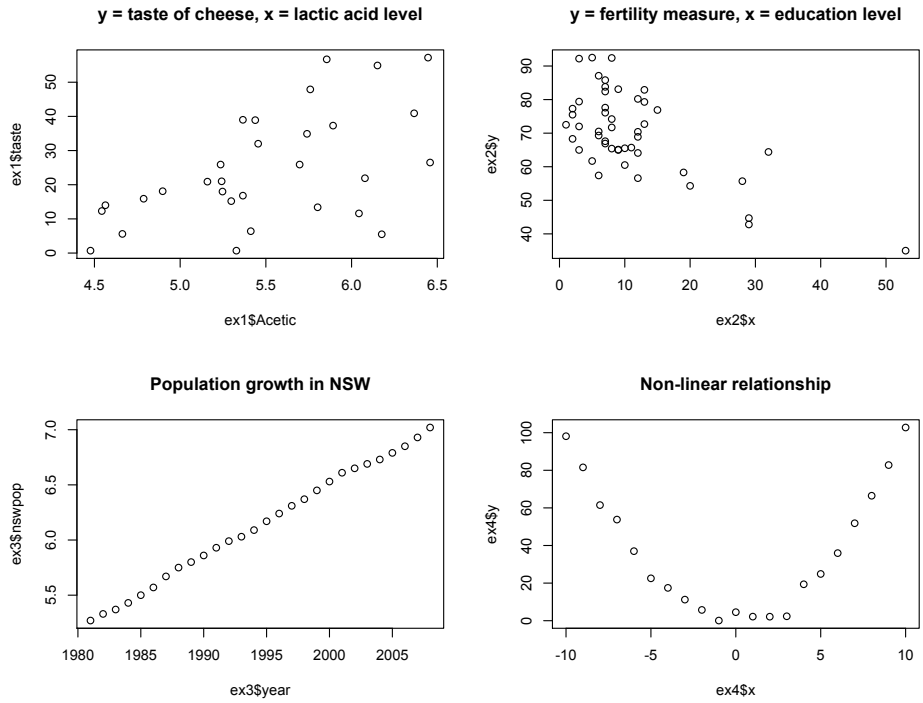
Scatterplot

For bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$ simply plot the points.

Four examples:

- Taste of matured cheese and lactic acid level. $(r = 0.55)$.
- Education level (x_i) and fertility level (y_i) of Swiss provinces (French speaking part) in $n = 47$. $(r = -0.66)$.
- Population growth in NSW between 1981 - 2008. $(r = 0.9992)$.
- Noisy non-linear functional relationship. $(r = 0.015)$.

Four examples (cont)



Correlation coefficient

Definition 20. The correlation coefficient is a numerical index that measures the degree of linear association between x and y,

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

Note that,

$$\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\
S_{yy} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \left(\sum_{i=1}^{n} y_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\sum_{i=1}^{n} x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)
\end{aligned}$$

In R: calculate r with `cor(x,y)`.

Example

Dose (in grams)	x	30	40	50	60	70	80	90	100
Breathing rate	y	16	14	13	13	11	12	9	9

To calculate r we need $n = 8$, $\sum_{i=1}^n x_i y_i = 5910$, $\sum_{i=1}^n x_i = 520$, $\sum_{i=1}^n y_i = 97$, $\sum_{i=1}^n x_i^2 = 38000$ and $\sum_{i=1}^n y_i^2 = 1217$. So

- $S_{xy} = \left(\sum_{i=1}^n x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) = 5910 - \frac{1}{8}520 \times 97 = -395$
- $S_{xx} = 38000 - \frac{1}{8}520^2 = 4200$
- $S_{yy} = 1217 - \frac{1}{8}97^2 = 40.875$
- $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-395}{\sqrt{4200 \times 40.875}} = -0.9533$ (to 4 d.p)

Properties of the correlation coefficient

i) The correlation coefficient is always between -1 and 1: $r \in [-1, 1]$.

ii) If $r = 1$ then all obs. (x_i, y_i) lie on a straight line with positive slope.

iii) If $r = -1$ then all obs. (x_i, y_i) lie on a straight line with negative slope.

iv) If $r = 0$ it does not follow that there is no relationship between x and y!

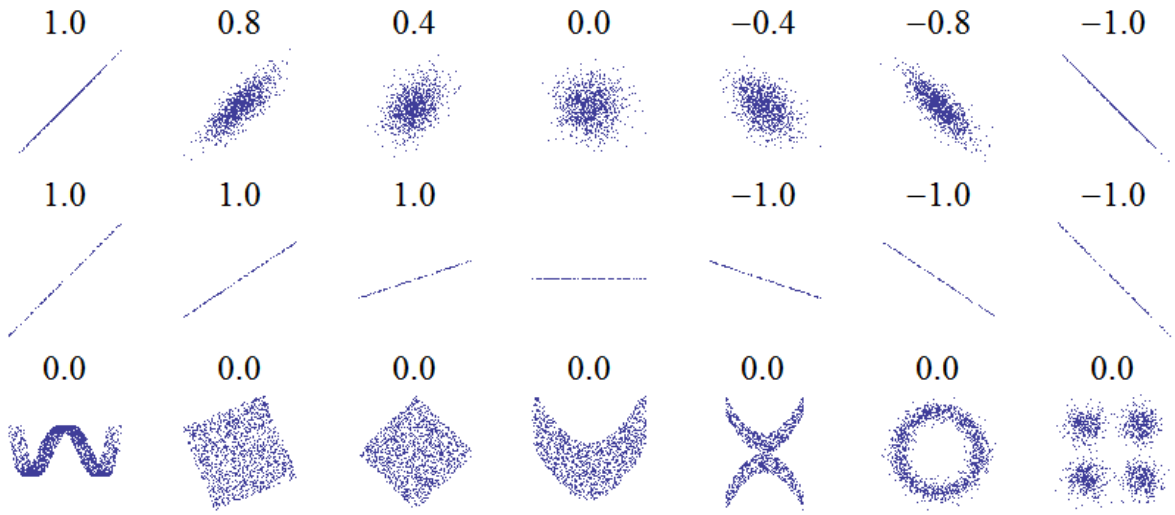
v) For high r (close to 1 or -1) it does not follow that there must be a relationship between x and y!

Proof. i) Is true because,

$$0 \leq \sum_{i=1}^n \left(\frac{x_i - \overline{x}}{\sqrt{S_{xx}}} - \frac{y_i - \overline{y}}{\sqrt{S_{yy}}}\right)^2 = \underbrace{\frac{\sum(x_i - \overline{x})^2}{S_{xx}}}_{=1} + \underbrace{\frac{\sum(y_i - \overline{y})^2}{S_{yy}}}_{=1} - 2\underbrace{\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{S_{xx}S_{yy}}}}_{=r} = 2 - 2r.$$

Hence it follows that $r \leq 1$. Similarly for $r \geq -1$ but with $\sum_{i=1}^n \left(\frac{x_i - \overline{x}}{\sqrt{S_{xx}}} + \frac{y_i - \overline{y}}{\sqrt{S_{yy}}}\right)^2$. □

Correlation Examples



Alternative formula for S_{xy}

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right).$$

Expanding and simplifying yields

$$\begin{aligned} S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i y_i - \overline{x} y_i - \overline{y} x_i + \overline{x}\,\overline{y}) \\ &= \sum x_i y_i - \underbrace{\overline{x}}_{\frac{1}{n}\sum x_i}\left(\sum y_i\right) - \overline{y}\underbrace{\sum x_i}_{n\overline{x}} + n\overline{x}\,\overline{y} \\ &= \sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) \underbrace{-n\overline{x}\,\overline{y} + n\overline{x}\,\overline{y}}_{=0} \end{aligned}$$

Theorem 2. Linear rescaling and translating of x or y values does not change the correlation coefficient r.

Proof: Let $u_i = a + bx_i$ and $v_i = c + dy_i$ where $b > 0$ and $d > 0$ then

$$S_{uv} = \sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v}).$$

Then

$$\begin{aligned} u_i - \overline{u} &= a + bx_i - \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(a + bx_i) \\ &= a + bx_i - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} a - \tfrac{b}{n}\textstyle\sum_{i=1}^{n} x_i \\ &= b(x_i - \overline{x}) \end{aligned}$$

Hence, $S_{uv} = bd\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = bdS_{xy}$. Similarly, $S_{uu} = b^2 S_{xx}$ and $S_{vv} = d^2 S_{yy}$. Then

$$r_{uv} = \frac{S_{uv}}{\sqrt{S_{uu}S_{vv}}} = \frac{bdS_{xy}}{\sqrt{b^2d^2S_{xx}S_{yy}}} = r_{xy}.$$

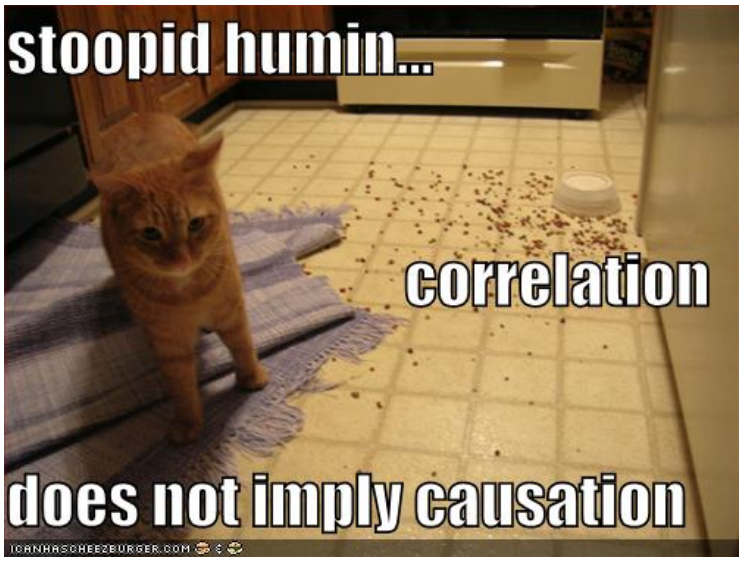
Example (Swiss fertility data). With R:

```
> x = swiss$Education
> y = swiss$Fertility
> length(x)
[1] 47
> c(sum(x),sum(x**2))
[1]  516 9918
> c(sum(y),sum(y**2))
[1]   3296.7 238416.9
> sum(x*y)
[1] 32526
> cor(x,y)
[1] -0.6637889
```

Thus, $S_{xx} = 9918 - \frac{(516)^2}{47} = 4252.979$, $S_{yy} = 7177.955$, $S_{xy} = 32526 - \frac{516\times 3296.7}{47}$. Hence,

$$r = \frac{-3667.557}{\sqrt{4252.979 \times 7177.955}} = -0.6637888.$$

Common misconception – Correlation is not cause!



Common misconception – Correlation is not cause!

Causation between two events implies a dependence between the two events.

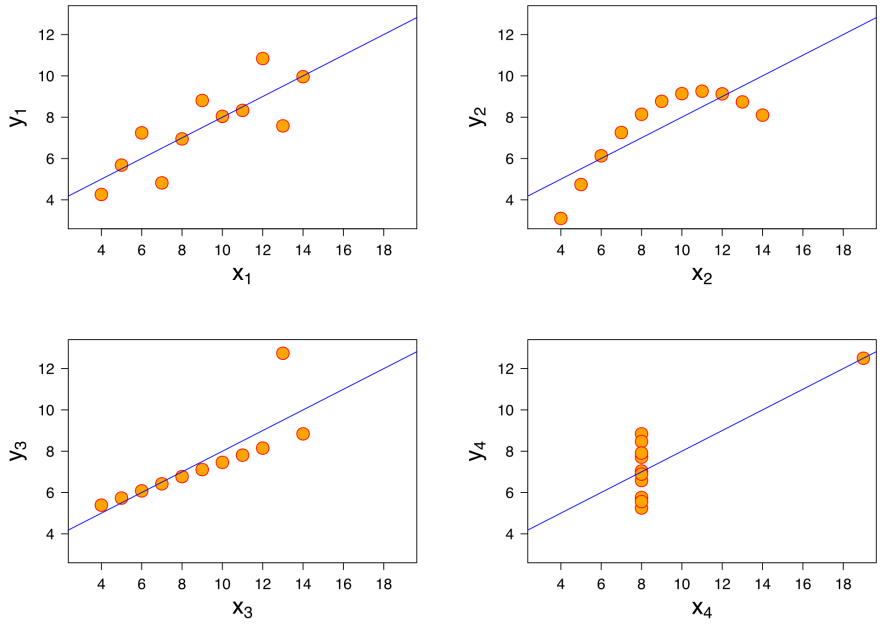
However, correlation cannot be used to infer a causal relationship between the variables because the cause of the underlying the correlation may be indirect and unknown, and high correlations can occur where no causal process exists.

For example, one may observe a correlation between the lecturer John Ormerod waking up and daybreak, though there is no direct causal relationship between these events, i.e. John Ormerod does not cause the sun to rise.

A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health, or does good health lead to good mood, or both?

Common misconception – Correlation does not mean linearity!

All of the examples below have a correlation of 0.816.

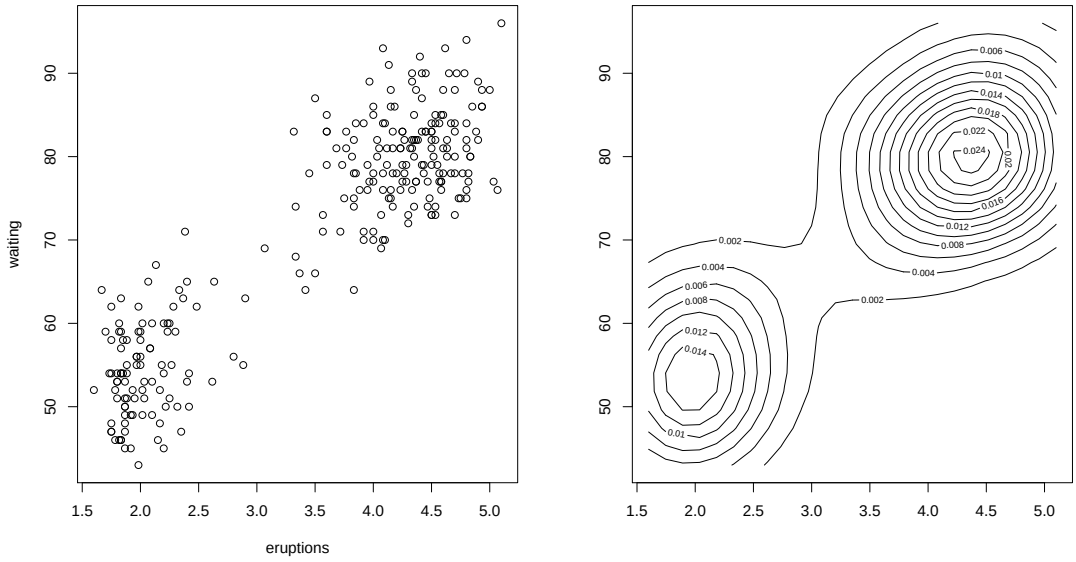


Contour plots (Not examinable)

Contour plots (based on density methods) are another useful way of looking at data.

On the next page the left hand side plot below is a scatterplot of the "Old faithful" dataset we saw earlier whereas the right hand side plot corresponds to the R command:

```
> library(MASS) # load an R library into memory
> contour(kde2d(faithful$eruptions,faithful$waiting))
```



More on measuring correlation (Not examinable)

The correlation coefficient r, often called Pearson correlation, is just one quantity that measures if two set of observations are correlated, i.e. are related. There are many more. Probably the second most famous is the Spearman rank correlation. Instead of using the original observations $(x_1, y_1), \ldots, (x_n, y_n)$ the corresponding ranks are analysed:

$$x_i \mapsto u_i = \mathsf{rank}(x_i)$$

$$y_i \mapsto v_i = \mathsf{rank}(y_i)$$

The Spearman rank correlation coefficient is simply the Pearson correlation coefficient for the u's and v's,

$$\rho = \frac{\sum \left(u_i - \overline{u}\right)\left(v_i - \frac{n+1}{2}\right)}{\sqrt{\sum \left(u_i - \frac{n+1}{2}\right)^2 \sum \left(v_i - \overline{v}\right)^2}}.$$

A toy example in R:

```
> x = c(5,4,2,1.5,3)
> y = c(5.2,4.7,2.8,1.9,4.1)
> u = rank(x)
> v = rank(y)
> cor(x,y)
[1] 0.9696742
> cor(u,v)
[1] 1
> plot(x,y)
```

The Spearman correlation coefficient measures to what extent the relationship of x and y is monotone. In the toy example the scatterplot of x and y shows a perfectly monotone relationship. Therefore, $\rho = 1$ whereas $r = 0.97$, i.e. is not exactly one because there could well be some quadratic relationship.

Monday, 13th August 2012

Lecture 5 - Content

□ Simple linear regression

See Phipps & Quine Chapter 1, Section 5.

Quotes about regression

Yale Law Professor Ian Ayres on regression:

> William Grove, completed a meta-analysis of 136 human versus machine studies. In only 8 out of 136 studies was expert opinion found to be appreciably more accurate than statistical prediction... Indeed, regression equations are so much better than humans... that even very crude regressions with just a few variables have been found to outpredict humans.

Cognitive psychologists Richard Nisbett and Lee Ross on regression:

> Human judges are not merely worse than optimal regression equations; they are worse than almost any regression equation.

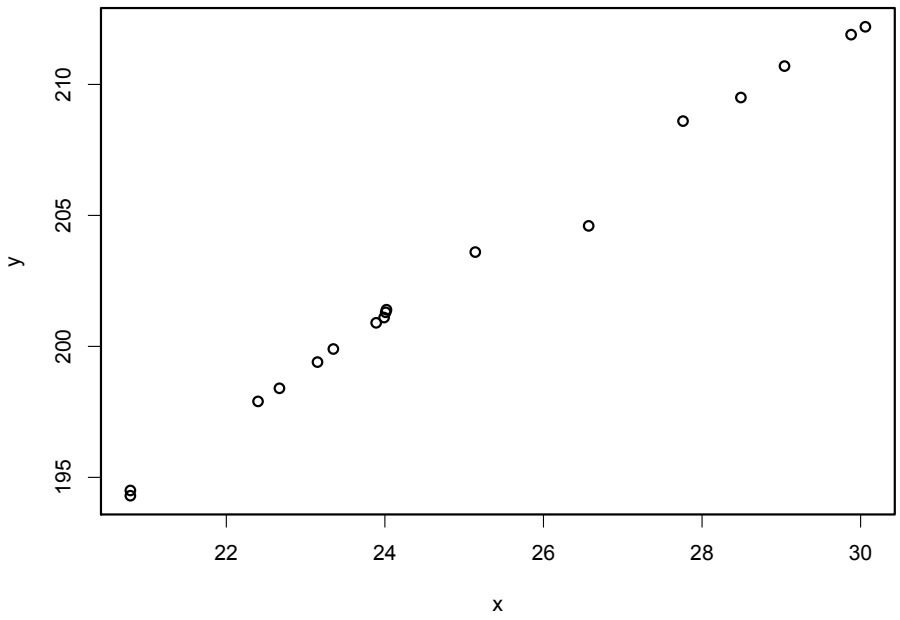
Simple linear regression

Linear regression seeks to model the relationship between the mean of a response variable, y, and a single explanatory variable x.

Example (Boiling point data)**.**

- Data on boiling point in degrees Fahrenheit (y) and pressure in inches of mercury (x), collected during an expedition in the Alps.
- Reference: Hand et al. (1994). *A Handbook of Small Data Sets*, London: C. & Hall.

The scatterplot shows that there is a clear relationship between y (temperature) and x (pressure).



Regression lines

For data $(x_1, y_1), \ldots, (x_n, y_n)$ we want to find a regression line that "fits" the data points. A simple linear regression model is

$$y_i = \underbrace{a + bx_i}_{\widehat{y}_i} + e_i = \widehat{y}_i + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where

- a is the intercept of the regression line,
- b is the slope of the regression line,
- $e_i = y_i - \widehat{y}_i$ is called the residual (error) of observation i.

The "best" regression line

Suppose we want to fit the "best" line $y = a + bx$ to the data.

There are a number of ways to define "best". We could choose a and b such that the sum of squared residuals is minimised:

$$M(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

or where the sum of absolute residuals is minimised:

$$D(a,b) = \sum_{i=1}^{n}|y_i - a - bx_i| = \sum_{i=1}^{n}|y_i - \widehat{y}_i| = \sum_{i=1}^{n}|e_i|$$

or where the maximum absolute residual is minimised:

$$H(a,b) = \max_i|y_i - a - bx_i| = \max_i|y_i - \widehat{y}_i| = \max_i|e_i|$$

The first problem corresponds to the "least squares" (LS) method, which chooses values of a and b which minimise the sum of the squares of these residuals. The other criteria are **much** harder to minimise.

The least squares regression line

Theorem 3. The least squares regression line, i.e. with a and b such that $M(a,b)$ is minimal, has intercept

$$a = a_{\mathsf{LS}} = \overline{y} - b_{\mathsf{LS}}\overline{x}$$

and slope

$$b = b_{\mathsf{LS}} = \frac{S_{xy}}{S_{xx}}.$$

Recall

- $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)$;
- $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2$.

Proof. Let $M = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$.
First minimise over a:

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

Hence,

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \Leftrightarrow n\overline{y} - na - nb\overline{x} = 0 \Rightarrow a = \overline{y} - b\overline{x}.$$

Then, substitute for a in the expression for M to get

$$M = \sum_{i=1}^{n}(y_i - \overline{y} - b(x_i - \overline{x}))^2 = S_{yy} - 2bS_{xy} + b^2 S_{xx}.$$

Minimise over b:

$$\frac{\partial M}{\partial b} = -2S_{xy} + 2bS_{xx} = 0 \Leftrightarrow b = \frac{S_{xy}}{S_{xx}}.$$

□

Example

In a study on the absorption of a drug, the dose x (in grams) and concentration in the urine y (in mg/g) were recorded as:

Dose (in grams)	x	46	53	37	42	34	29	60	44	41	48	33	40
Urine Concentration	y	12	14	11	13	10	8	17	12	10	15	9	13

$$\sum_{i=1}^{n} x_i = 507, \quad \sum_{i=1}^{n} y_i = 144, \quad n = 12$$
$$\sum_{i=1}^{n} x_i^2 = 22265, \quad \sum_{i=1}^{n} y_i^2 = 1802, \quad \sum_{i=1}^{n} x_i y_i = 6314$$

- $S_{xy} = \left(\sum_{i=1}^{n} x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 6314 - \frac{1}{12}(507)(144) = 230$
- $S_{xx} = \left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 22265 - \frac{1}{12}(507)^2 = 844.25$
- $b = \frac{S_{xy}}{S_{xx}} = \frac{230}{844.25} = 0.272431$ (to 6 d.p)
- $a = \overline{y} - b\overline{x} = \frac{144}{12} - 0.272431 \times \frac{507}{12} = 0.489782$ (to 6 d.p)

Fitted regression line

Because $a = \overline{y} - b\overline{x}$ we can write

$$y = a + bx = \overline{y} - b\overline{x} + bx = \overline{y} + b(x - \overline{x}),$$

so the regression line passes through the component-wise mean $(\overline{x}, \overline{y})$.

Correlation coefficient and regression slope

Recall that the correlation coefficient between vectors x and y is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \in [-1, 1].$$

Because,

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

Therefore, b and r have the same sign, both positive or both negative.

Example (Boiling point, cont). The $n = 17$ observations are:

```
> x
 [1] 20.79 20.79 22.40 22.67 23.15 23.35 23.89 23.99 24.02
[10] 24.01 25.14 26.57 28.49 27.76 29.04 29.88 30.06
> y
 [1] 194.5 194.3 197.9 198.4 199.4 199.9 200.9 201.1 201.4
[10] 201.3 203.6 204.6 209.5 208.6 210.7 211.9 212.2
```

To obtain a and b we first calculate the following auxiliary numbers:

$\sum x_i = 426$ $\quad$ $\sum y_i = 3,450.2$

$\sum x_i^2 = 10,821$ $\quad$ $\sum y_i^2 = 700,759$

$\sum x_i y_i = 86,735.5$

Thus,

$S_{xx} = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = 10,821 - \frac{426^2}{17} = 145.9412$

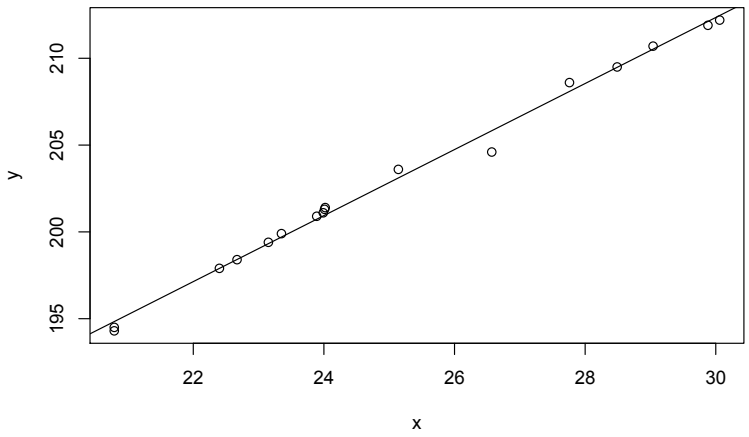
$S_{yy} = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2 = 530.7824$

$S_{xy} = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i) = 277.5421 \Rightarrow r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{277.5421}{278.3185} = 0.9972$

and we have $b = \frac{S_{xy}}{S_{xx}} = \frac{277.5421}{145.9412} = 1.90$ and $a = \frac{3450.2}{17} - 1.90 \times \frac{426}{17} = 155.3$.

In R: Execute,

```
> plot(x,y)        # as before, produces scatterplot of x against y
> abline(lm(y~x)) # lm(y~x) : lm() = linear model function;
                   # y~x means model y by x
> lm(y~x)
(Intercept)            x
    155.296        1.902
```



Additional material for Lecture 5

More on simple linear regression

A function $f(\beta) : \mathbb{R}^2 \mapsto \mathbb{R}$ is linear (a linear map) if and only if it preserves addition and scalar multiplication, i.e.

1. for all $\beta, \gamma \in \mathbb{R}^2$ we have $f(\beta + \gamma) = f(\beta) + f(\gamma)$,
2. for all $c \in \mathbb{R}$ we have $f(c\beta) = cf(\beta)$.

A simple linear regression is considered to be a function of the intercept a and slope b, given the information of the data, i.e. $(x_1, y_1), \ldots, (x_n, y_n)$. Therefore one can do much more with simple linear regression than just fitting a straight line. For example consider a simple transformation of the explanatory variable x such as $z = \log(x)$. Then,

$$y_i = a + b\log(x_i) + e_i = f(a, b|x_i) + e_i = a + bz_i + e_i = f(a, b|z_i) + e_i, \quad i = 1, \ldots, n,$$

is clearly a simple linear regression model since

$$y = f(a, b|x) = a + bz \Rightarrow f(ca, cb|z) = ca + cbz = cf(a, b|z) \quad \text{and} \quad f(a_0 + a, b_0 + b|z) = a_0 + b_0 z + a + bz = f(a_0, b_0|z) + f(a, b|z).$$

Linear regression models will turn out to be a very powerful instrument in the analysis of higher dimensional data and are in statistics as powerful as are Taylor or Fourier series in calculus. You can learn more on linear models in STAT2912 and much more in STAT3912.

Tuesday, 14th August 2012

Lecture 6 - Content

- ☐ **Semi-log transformation**
- ☐ **Residual plots**
- ☐ **Explaining variability**

See Phipps & Quine Chapter 1, Section 5.

Semi-log Transformations

Suppose an exponential trend of the type $y = A \times B^x$ is expected.
Take (natural) logs of both sides to obtain

$$\begin{aligned} \log(y) &= \log(A \times B^x) \\ &= \log(A) + \log(B) \times x \end{aligned}$$

and so if we put $Y = \log(y)$, $X = x$, $a = \log(A)$ and $b = \log(B)$ the line we now want to estimate is $Y = a + bX$.

Procedure: Perform a semi-log transform, i.e. $X_i = x_i$ and $Y_i = \log(y_i)$, then find a LSR line for the points (X_i, Y_i) for the line $Y = a + bX$ in the usual way. Lastly, transform back to obtain the fitted curve $y = A \times B^x$ (using $A = e^a$ and $B = e^b$).

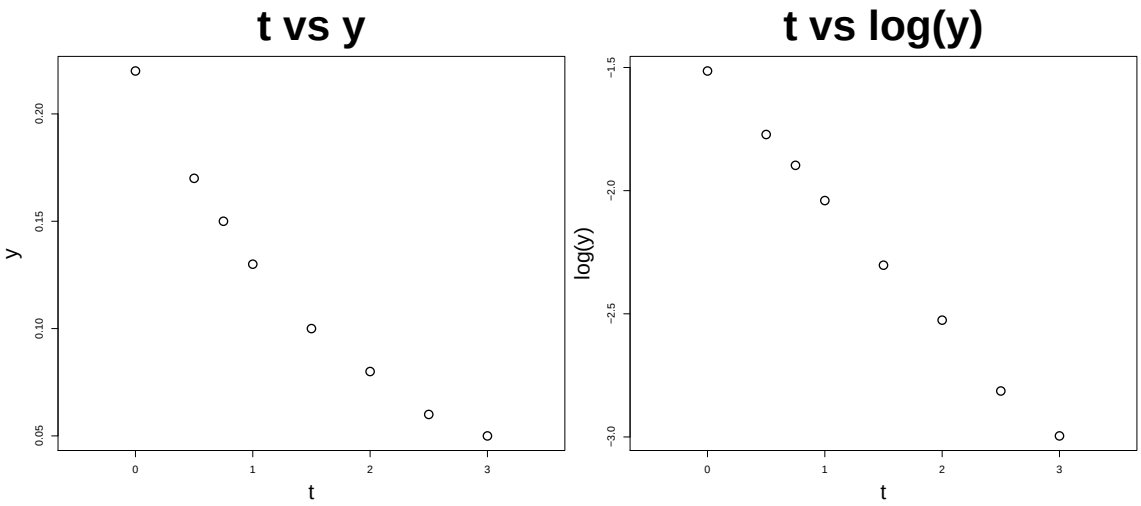
The Semi-log Transformation – Example

The alcoholic content, y (mg/ml) of a person's blood, t hours after drinking whisky, is displayed in the table below:

Time (h)	t	0.00	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Alcohol (mg/ml)	y	0.22	0.17	0.15	0.13	0.10	0.08	0.06	0.05

Why do you think that an exponential relationship, $y = A \times B^t$, might be an appropriate relationship?

The Semi-log Transformation – Plots



The Semi-log Transformation – Working Out

The original data is

t	0.00	0.50	0.75	1.00	1.50	2.00	2.50	3.00
y	0.22	0.17	0.15	0.13	0.10	0.08	0.06	0.05

Using a semi-log transformation we have

$X = t$	0.00	0.50	0.75	1.00	1.50	2.00	2.50	3.00
$Y = \log(y)$	-1.51	-1.77	-1.90	-2.04	-2.30	-2.53	-2.81	-3.00

Using these values:

$$\sum_{i=1}^{n} X_i = 11.25, \quad \sum_{i=1}^{n} Y_i = -17.86, \quad n = 8$$
$$\sum_{i=1}^{n} X_i^2 = 23.31, \quad \sum_{i=1}^{n} Y_i^2 = 41.77, \quad \sum_{i=1}^{n} X_i Y_i = -28.89$$

The Semi-log Transformation – Working Out

Using the values:

$$\sum_{i=1}^{n} X_i = 11.25, \quad \sum_{i=1}^{n} Y_i = -17.86, \quad n = 8$$
$$\sum_{i=1}^{n} X_i^2 = 23.31, \quad \sum_{i=1}^{n} Y_i^2 = 41.77, \quad \sum_{i=1}^{n} X_i Y_i = -28.89$$

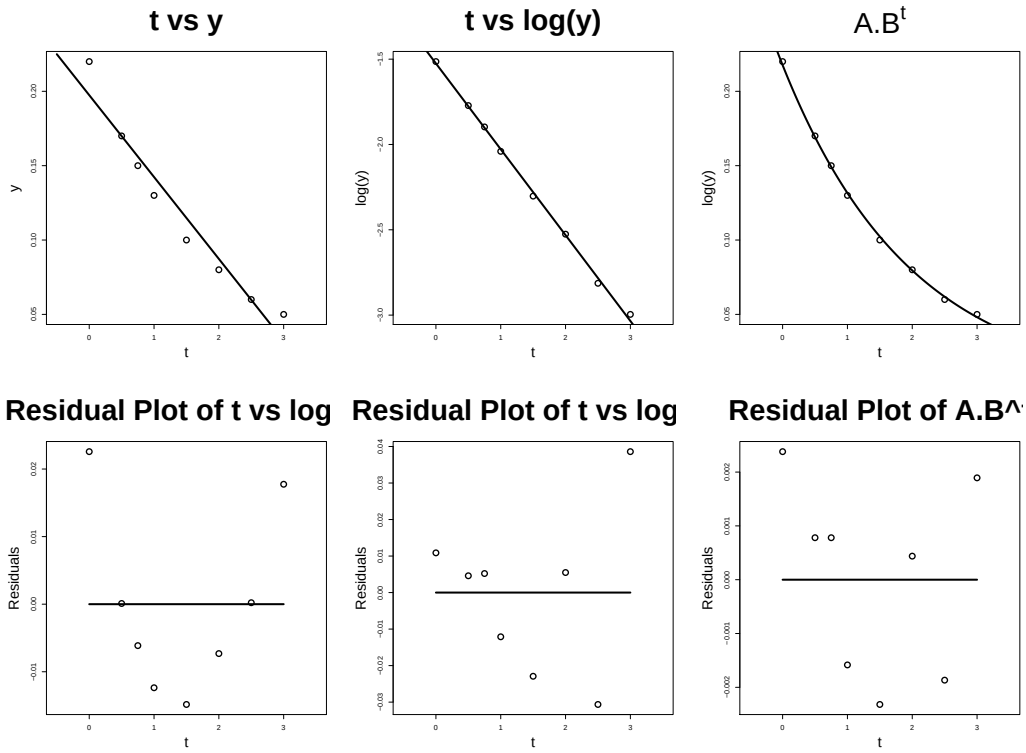
we have

- $S_{xy} = -3.76$
- $S_{xx} = 7.49$

Hence,

- $b = S_{xy}/S_{xx} = -0.5031074$
- $a = \overline{y} - b\overline{x} = -1.525005$
- $A = e^a = 0.2176199$
- $B = e^b = 0.6046488$

The Semi-log Transformation – Plots



Residual plots

- The scatterplot of y and x already indicates whether or not a straight line is a good model.
- A scatterplot of the residuals e against the explanatory variable x gives further insight and is called residual plot:
  - Is there any curvature left?
  - Are there any non horizontal patterns left?

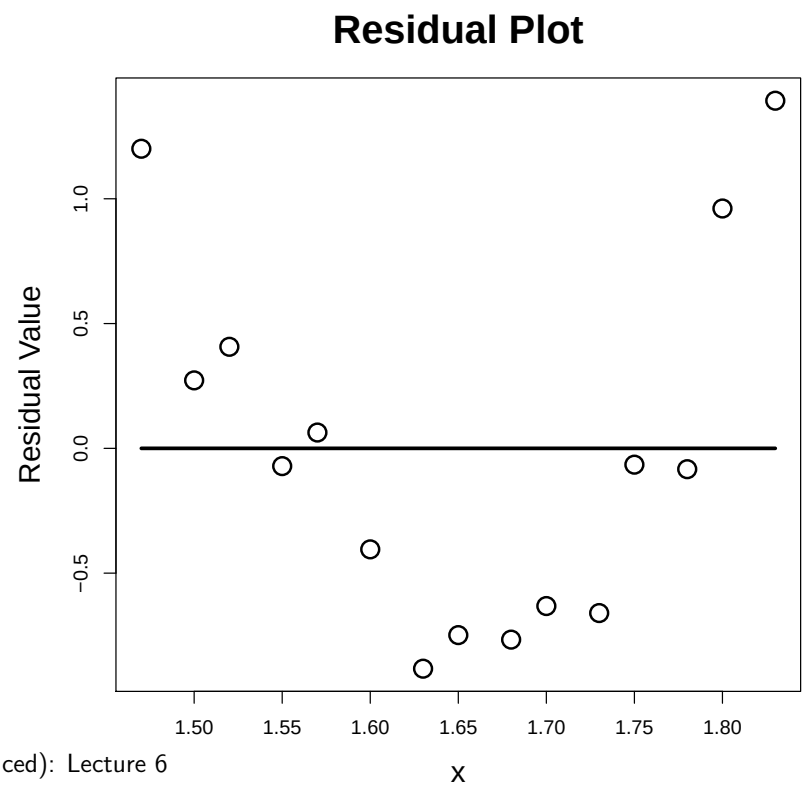
Remarks

- It is a property of the least squares method that

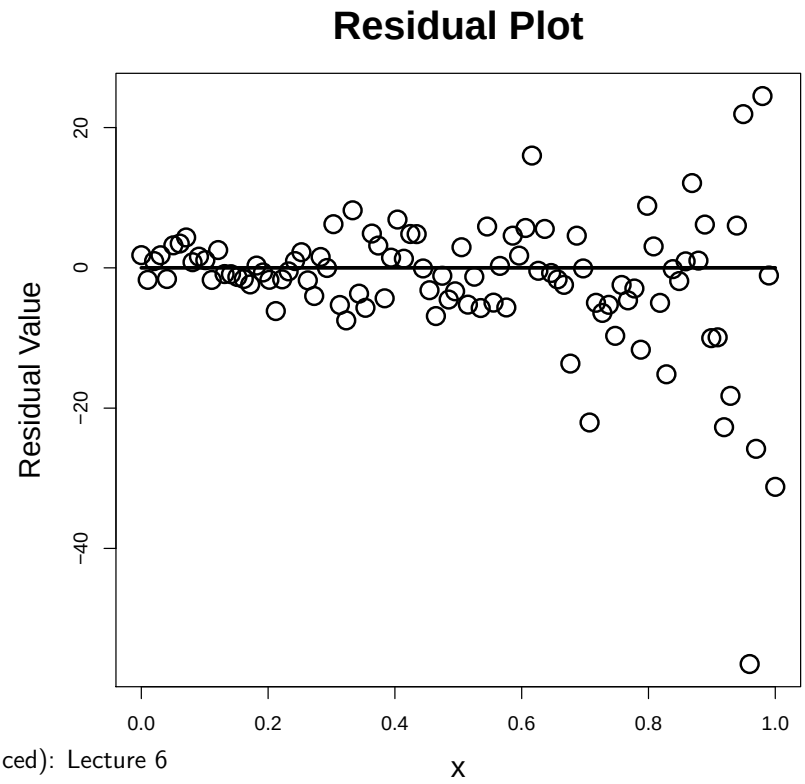
$$\sum_{i=1}^{n} e_i = 0 \;\Rightarrow\; \overline{e}_i = 0$$

  'local' failures, i.e. regions where there is curvature indicate that 'locally' a straight line is not an appropriate model.
- A boxplot and histogram of the residuals e can be drawn to assess symmetry and other aspects of the residuals.
- Overall, residuals should appear randomly scattered about zero.
- Long sequences of positive residuals followed by sequences of negative residuals in e_i vs x_i plot suggests that the error terms are not independent.
- Outliers can severely effect the quality of the fit.

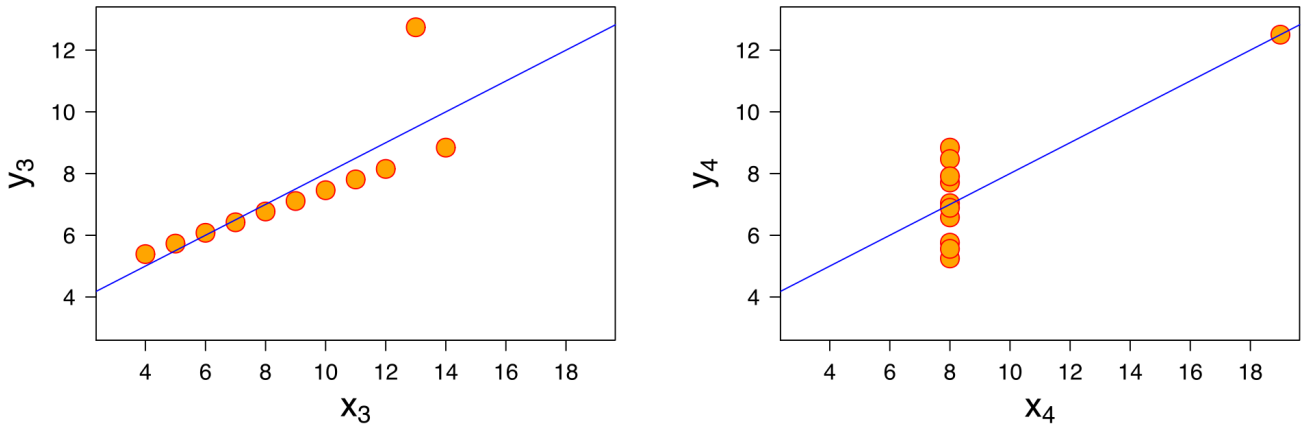
Residual plots - Nonlinear Example



Residual plots - Heteroscedasticity

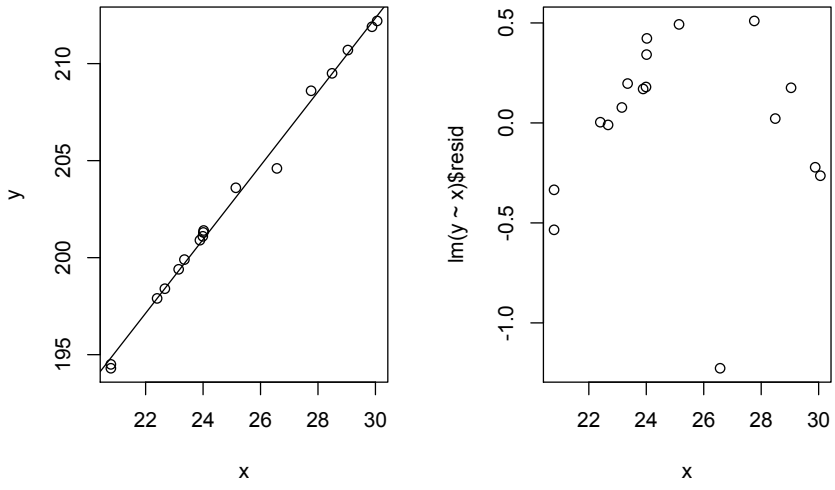


Residual plots - Outlier Example

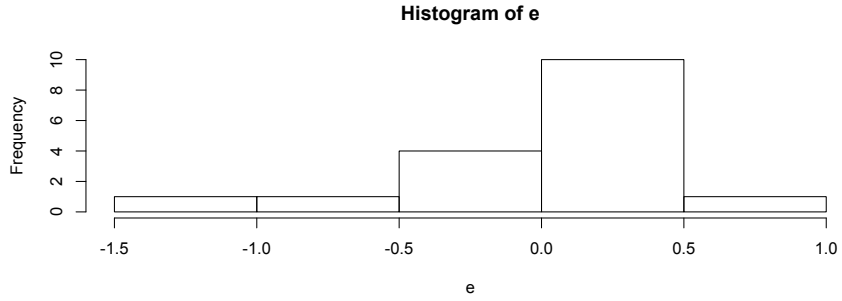


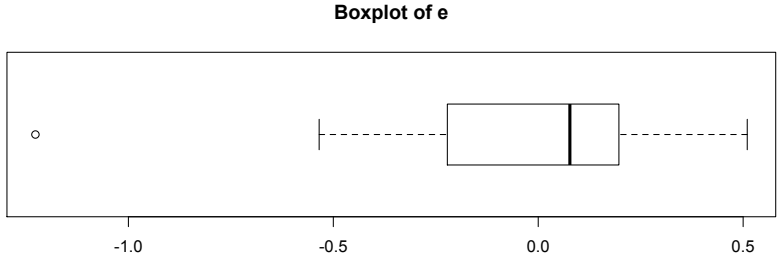
Example (Boiling point, cont). After executing the following lines in R:

```
> par(mfrow=c(1,2));
> plot(x,y)
> abline(coef(lm(y~x)))
> plot(x, lm(y~x)$resid)
```



```
> e = lm(y~x)$resid
> hist(e)
> boxplot(e, main="Boxplot of e")
```





Explaining variability

- The proportion of variability of y's explained by the regression on x is r^2, i.e.

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

- The variance of the y's is $s_y^2 = S_{yy}/(n-1)$.
- Recall, $\widehat{y}_i = a + bx_i = \overline{y} + \frac{S_{xy}}{S_{xx}}(x_i - \overline{x})$:

$$\begin{aligned}\text{RSS} = \sum e_i^2 &= \sum (y_i - \widehat{y}_i)^2 = \sum \left(y_i - \overline{y} - \frac{S_{xy}}{S_{xx}}(x_i - \overline{x}) \right)^2 \\ &= S_{yy} + \frac{S_{xy}^2}{S_{xx}^2} \sum (x_i - \overline{x})^2 - 2\frac{S_{xy}}{S_{xx}} \sum (y_i - \overline{y})(x_i - \overline{x}) \\ &= S_{yy} + \frac{S_{xy}^2}{S_{xx}} - 2\frac{S_{xy}^2}{S_{xx}} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.\end{aligned}$$

- Hence, $\frac{S_{yy} - \text{RSS}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$.

Semester 2, 2012 (Last adjustments: August 13, 2012)

Lecture Notes

MATH1905 Statistics (Advanced)

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod (at) sydney.edu.au

Semester 1, 2012 (Last adjustments: August 13, 2012)

Tuesday, 14th August 2012

Lecture 6 - Content

- ☐ **Sets**
- ☐ **Probability and counting**
- ☐ **Conditional probability**

Sets

Before we look at probability it is necessary to understand sets because probabilities are typically described in terms of sets where an 'event' occurs.

Definition 1. The set of all possible outcomes of an experiment is called a sample space, denoted by Ω. Any subset A of the sample space Ω, denoted by $A \subset \Omega$ is called an event.

Definition 2. The counting operator $N(A)$ is a set function that counts how many elements belong to the set (event) A.

Example (Sample spaces)**.**
Coin: $\Omega = \{H, T\} \Rightarrow N(\Omega) = 2$.
Dice: $\Omega = \{1, 2, 3, 4, 5, 6\} \Rightarrow N(\{1, 2, 5\}) = 3$;
Weight: $\Omega = \mathbb{R}^+ \Rightarrow N(\mathbb{R}^+) = \infty$.

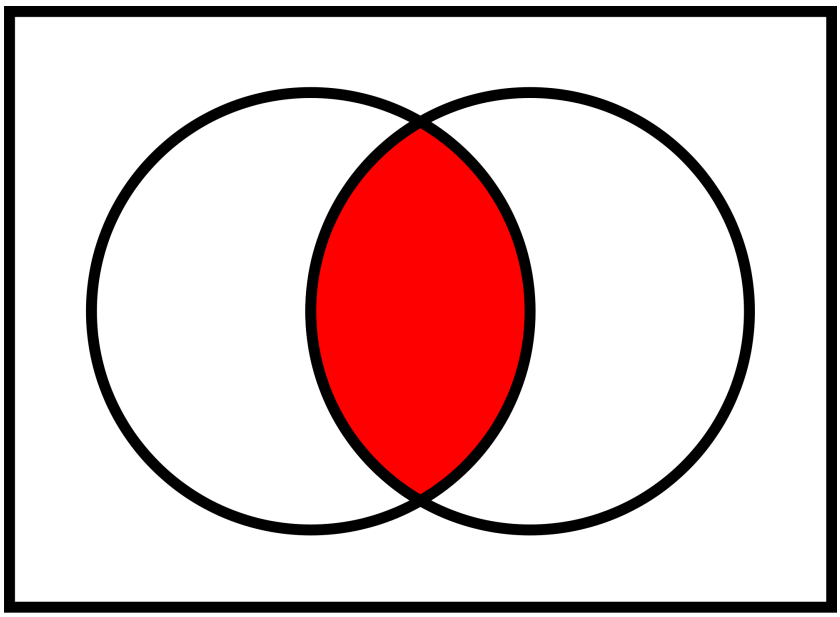
Set Notation

Before we introduce probability we need to introduce some notation. Let $A, B \subset \Omega$.

symbol	set theory	probability
Ω	largest set	certain event
$\emptyset$	empty set	impossible event
$A \cup B$	union of A and B	event A or event B
$A \cap B$	intersection of A and B	event A and event B
$A^C = \Omega \backslash A$	complement of A	not event A

Intersection Operator

The set $A \cap B$ denotes the set such that if $C \in A \cap B$ then $C \in A$ and $C \in B$ ($\cap$ is called the intersection operator).



Intersection Operator

Examples:

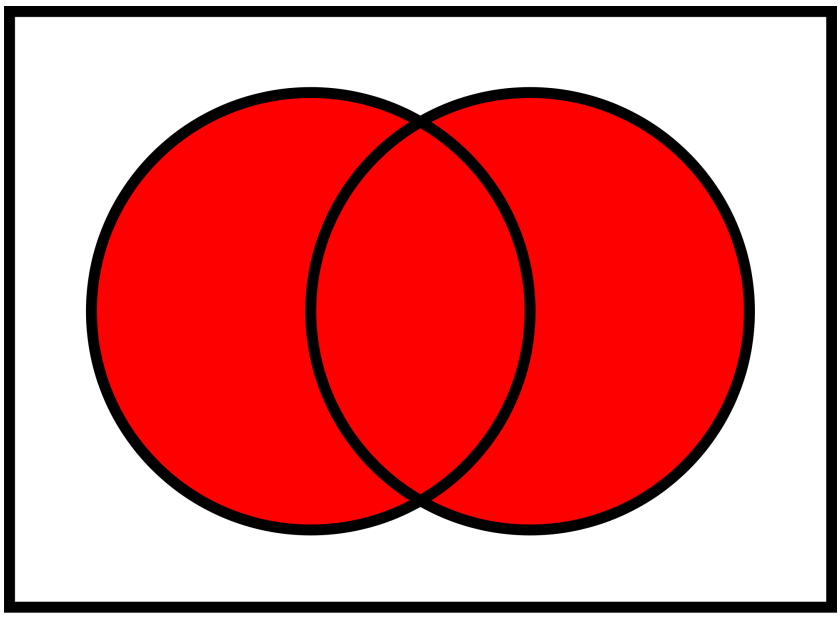
- $\{1, 2\} \cap \{\text{red}, \text{white}\} = \emptyset$.
- $\{1, 2, \text{green}\} \cap \{\text{red}, \text{white}, \text{green}\} = \{\text{green}\}$.
- $\{1, 2\} \cap \{1, 2\} = \{1, 2\}$.

Some basic properties of intersections:

- $A \cap B = B \cap A$.
- $A \cap (B \cap C) = (A \cap B) \cap C$.
- $A \cap B \subseteq A$.
- $A \cap A = A$.
- $A \cap \emptyset = \emptyset$.
- $A \subseteq B$ if and only if $A \cap B = A$.

Union Operator

The set $A \cup B$ denotes the set such that if $C \in A \cup B$ then $C \in A$ and/or $C \in B$ ($\cup$ is called the union operator).



Union Operator

Examples:

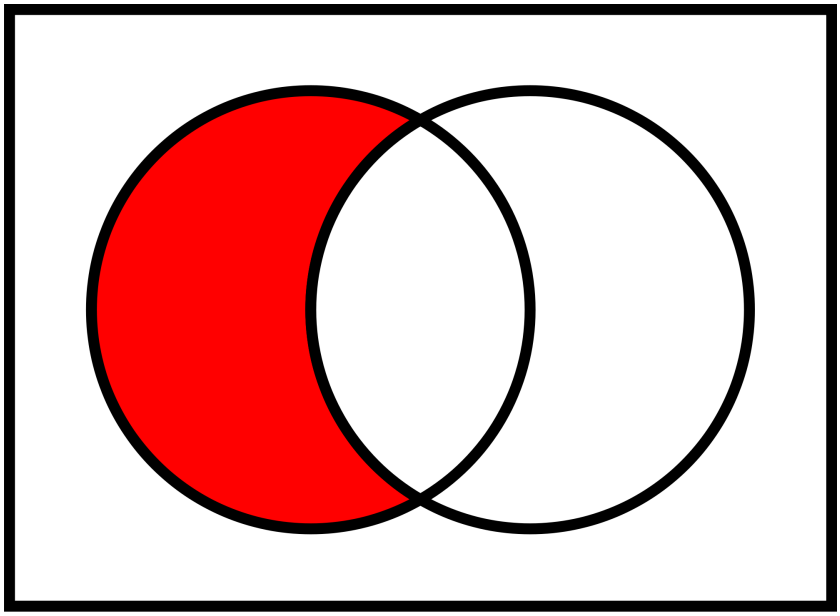
- $\{1, 2\} \cup \{\text{red, white}\} = \{1, 2, \text{red, white}\}$.
- $\{1, 2, \text{green}\} \cup \{\text{red, white, green}\} = \{1, 2, \text{red, white, green}\}$.
- $\{1, 2\} \cup \{1, 2\} = \{1, 2\}$.

Some basic properties of unions:

- $A \cup B = B \cup A$.
- $A \cup (B \cup C) = (A \cup B) \cup C$.
- $A \subseteq (A \cup B)$.
- $A \cup A = A$.
- $A \cup \emptyset = A$.
- $A \subseteq B$ if and only if $A \cup B = B$.

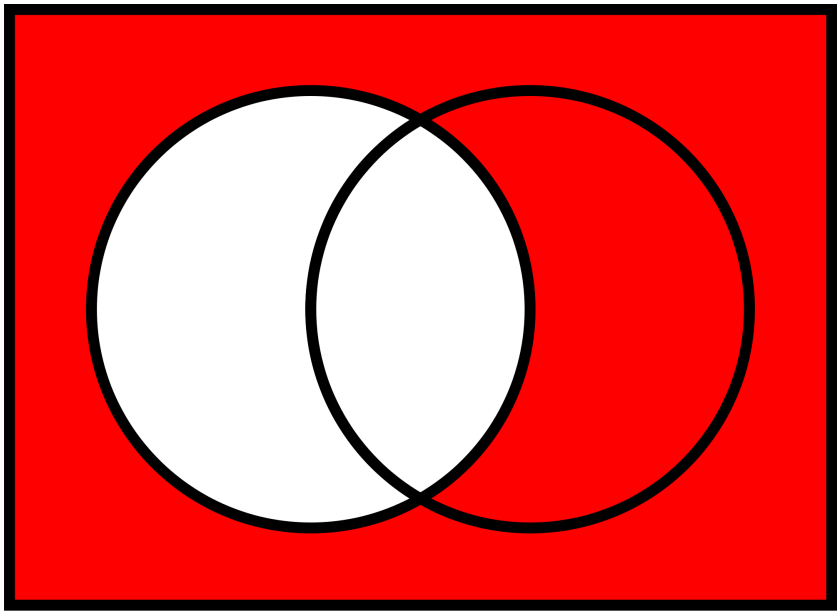
Set Minus

The set $A \setminus B$ denotes the set such that if $C \in A \setminus B$ then $C \in A$ and $C \notin B$.



Set Complement

The set $A^c = \Omega \setminus A$ denotes the set such that if $C \in A^c$ then $C \notin A$.



Set Minus

Examples:

- $\{1, 2\} \setminus \{\text{red, white}\} = \{1, 2\}$.
- $\{1, 2, \text{green}\} \setminus \{\text{red, white, green}\} = \{1, 2\}$.
- $\{1, 2\} \setminus \{1, 2\} = \emptyset$.
- $\{1, 2, 3, 4\} \setminus \{1, 3\} = \{2, 4\}$.

Some basic properties of complements:

- $A \setminus B \neq B \setminus A$.
- $A \cup A^c = \Omega$.
- $A \cap A^c = \emptyset$.
- $(A^c)^c = A$.
- $A \setminus A = \emptyset$.

Theorem 1. The complement of the union of A and B equals the intersection of the complements

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

Proof. Use Venn diagrams for LHS and RHS and colour areas. □

Theorem 2. de Morgan's Laws.

$$\left(\bigcap_{i=1}^{n} A_i \right)^c = \bigcup_{i=1}^{n} A_i^c$$

and

$$\left(\bigcup_{i=1}^{n} A_i \right)^c = \bigcap_{i=1}^{n} A_i^c$$

Counting – Ordered Sampling without replacement

Example (Ordered samples without replacement)**.** The number of ordered samples of size r we can draw without replacement from n objects is,

$$n \times (n-1) \times \ldots \times (n-r+1) = \frac{n!}{(n-r)!}$$

Recall: $0! = 1$.

Counting – Unordered Sampling without replacement

Example (Unordered samples without replacement)**.**

$${}^{n}C_{r} = \binom{n}{r} = \frac{n!}{r!(n-r)!} = n \text{ Choose } r.$$

Recall,

$${}^{n}C_{r} = {}^{n}C_{n-r}$$

since

$$\binom{n}{n-r} = \frac{n!}{(n-r)!((n-(n-r))!} = n \text{ Choose } r.$$

and so

$$\binom{n}{0} = \binom{n}{n} = 1$$

Sampling in R

```
# Creating ordered lists
n = 158;
x = 1:n;
set.seed(6)       # set random seed to 6 to reproduce results
sample(x)         # random permutation of nos 1,2,...,158: n! possibilities
sample(x,10)      # choose 10 numbers without replacement
sample(x,10,TRUE) # choose 10 numbers with replacement = bootstrap sampling
```


What is Probability?

1. Subjective probability expresses the strength of one's belief (the basis of Bayesian Statistics – a bit on that later).
2. Classical probability concept, mathematical answer for equally likely outcomes.

   Theorem 3. If there are n equally likely possibilities of which one must occur and s are regarded as favourable ($=$ successes), then the probability P of a success is given by s/n.

What is Probability?

3. The frequency interpretation of probability:

   Theorem 4. The probability of an event (or outcome) is the proportion of times the event occur in a long run of repeated experiments.

   or in words:

   > If an experiment is repeated n times under **identical conditions**, and if the event A occurs m times, then as n becomes large (i.e. in the long-run) the probability of A occurring is the ratio m/n.

What is Probability?

- The constancy of the gender ratio at birth. In Australia, the proportion of male births is fairly stable at 0.51. This long run relative frequency is used to estimate the probability that a randomly chosen birth is male.
- Cancer council records show the age standardised mortality rate from breast cancer in NSW was close to 20 per 100,000 over the years 1972-2000. For a randomly chosen woman, we use 0.0002 as the probability of breast cancer.

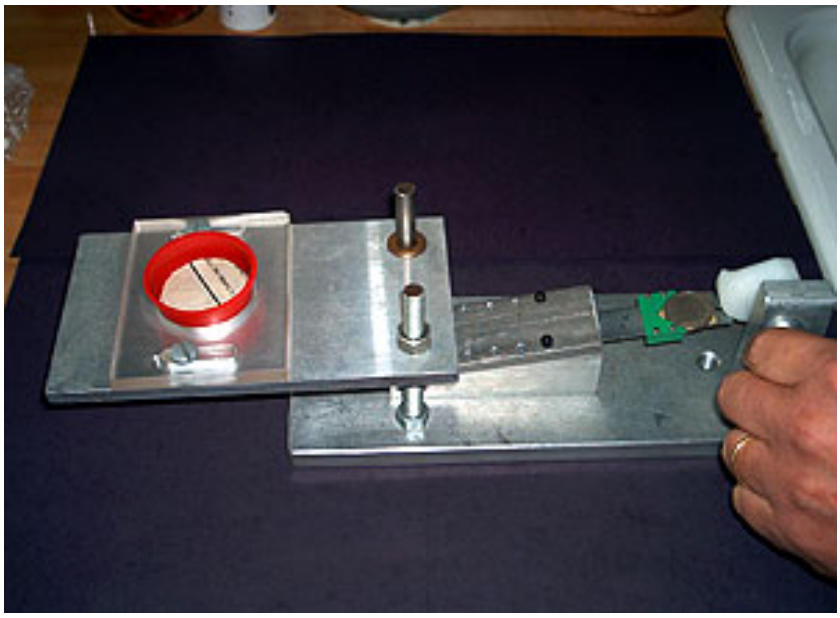
Example (Coin tossing).
Buffon (1707-1788):
Pearson (1857-1936):

Coin Tossing in R

```
table(sample(c("H","T"),4040,T))/4040
table(sample(c("H","T"),24000,T))/24000
```

Coin Tossing 2010's

In the 2010's Stanford Professor Persi Diaconis developed the "Coin Tosser 3000".



However, the machine is designed to flip a coin with the same result **every time**!

What is Probability?

4. Mathematical formulation of probability

Definition 3 (due to Andrey Kolmogorov, 1933)**.** Given a sample space Ω $A \subset \Omega$, we define $\mathrm{P}(A)$, the probability of A, to be a value of a non-negative additive set function that satisfies the following three axioms:
A1: For any event A, $0 \leq \mathrm{P}(A)$,
A2: $\mathrm{P}(\Omega) = 1$,
A3: If A and B are mutually exclusive events $(A \cap B = \emptyset)$, then

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).$$

A3': If $A_1, A_2, A_3, \ldots$ is a finite or infinite sequence of mutually exclusive events in Ω, then

$$\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_2) + \ldots.$$

$0 \leq \mathrm{P}(A) \leq 1$

Theorem 5. Assume the following 3 axioms:
A1: For any event $A \subset \Omega$, $0 \leq \mathrm{P}(A)$,
A2: $\mathrm{P}(\Omega) = 1$,
A3': If $A_1, A_2, A_3, \ldots$ is a finite or infinite sequence of mutually exclusive events in Ω, then

$$\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_2) + \ldots.$$

Then $0 \leq \mathrm{P}(A) \leq 1$.

Proof.

□

□

□

□

Example (Lotto). A lotto type barrel contains 10 balls numbered $1, 2, \ldots, 10$. Three balls are drawn.

i. How many distinct samples can be drawn?

ii. Event $A = \{1, 2, \ldots, 7\}$ (all numbers less than seven).

$$\mathrm{P}(A) = \qquad .$$

iii. $B =$ all drawn numbers are even: $\mathrm{P}(B) = \frac{1}{120} \times \binom{5}{3} = \frac{10}{120} = \frac{1}{12}$.

$$\mathrm{P}(A \cap B) =$$

iv. $\mathrm{P}(A \cup B)$? To answer this we need our next theorem.

Addition Theorem

Theorem 6 (Addition Theorem)**.** If A and B are any events in Ω, then

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

Proof. Use Venn diagrams, i.e. draw pictures  and colour regions.

use axioms only

□

Example (Lotto)**.** A lotto type barrel contains 10 balls numbered $1, 2, \ldots, 10$. Three balls are drawn.

i. How many distinct samples can be drawn? 120.

ii. Event $A = \{1, 2, \ldots, 7\}$ (all numbers less than seven). $\mathrm{P}(A) = \frac{7}{24}$.

iii. $B =$ all drawn numbers are even: $\mathrm{P}(B) = \frac{1}{12}$.
Also $\mathrm{P}(A \cap B) = 1/120$.

iv. $\mathrm{P}(A \cup B)$?

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = \frac{7}{24} + \frac{1}{12} - \frac{1}{120} = \frac{44}{120} = \frac{11}{30}.$$

Poincarés' Theorem

Theorem 7 (Poincarés' formula, not part of M1905)**.** Let $A_1, A_2, \ldots, A_n$ be any events in Ω. Then,

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i) - \sum_{i<j} \mathrm{P}(A_i \cap A_j) + \sum_{i<j<k} \mathrm{P}(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n-1} \mathrm{P}\left(\bigcap_{i=1}^{n} A_i\right).$$

(Unconditional) probability

- Recall 3 Axioms of probability.
- $\mathrm{P}(A^C) = 1 - \mathrm{P}(A)$ since $A \cap A^C = \emptyset$ hence, $1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^C) = \mathrm{P}(A) + \mathrm{P}(A^C)$.
- $\mathrm{P}(\emptyset) = 0$ because $\emptyset = \Omega^C$, hence $\mathrm{P}(\emptyset) = 1 - \mathrm{P}(\Omega)$.
- etc.

Conditional Probability – Another Motivating Example

What is the probability of the important event

$$A = (\text{starting salary after uni } \geq 60\text{k})?$$

What is the sample space Ω?

Possibilities are:

$$\begin{aligned}\Omega_1 &= \{\text{all students}\},\\ \Omega_2 &= \{\text{all male students}\},\\ \Omega_3 &= \{\text{all students with a maths degree}\}.\end{aligned}$$

Conclusion

- Probability depends on the underlying sample space Ω!
- Hence, if it is unclear to what sample space A refers to then make it clear by writing

  $$\mathrm{P}(A|\Omega) \quad \text{instead of} \quad \mathrm{P}(A)$$

  which we read as the conditional probability of A relative to Ω or given Ω, respectively.

Definition 4. If A and B are any events in Ω and $\mathrm{P}(B) \neq 0$ then, the conditional probability of A given B is

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

Additional material for Lecture 6

A combinatorial proof of the binomial theorem

The binomial theorem says

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Consider the more complicated product

$$(x_1+y_1)(x_2+y_2)\cdots(x_n+y_n)$$

Its expansion consists of the sum of 2^n terms, each term being the product of n factors. Each term consists either x_k or y_k, for each $k = 1, \ldots, n$. For example,

$$(x_1+y_1)(x_2+y_2) = x_1x_2 + x_1y_2 + y_1x_2 + y_1y_2$$

Now, there is $1 = \binom{n}{0}$ term with y terms only, $n = \binom{n}{1}$ with one x term and $(n-1)$ y terms etc. In general, there are $\binom{n}{k}$ terms with exactly k x's and $(n-k)$ y's. The theorem follows by letting $x_k = x$ and $y_k = y$.

More on set theory

The operation of forming unions, intersections and complements of events obey rules similar to the rules of algebra. Following some examples for events A, B and C:
Commutative law: $A \cup B = B \cup A$ and $A \cap B = B \cap A$
Associative law: $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$
Distributive law: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ and $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

Monday, 20th August 2012

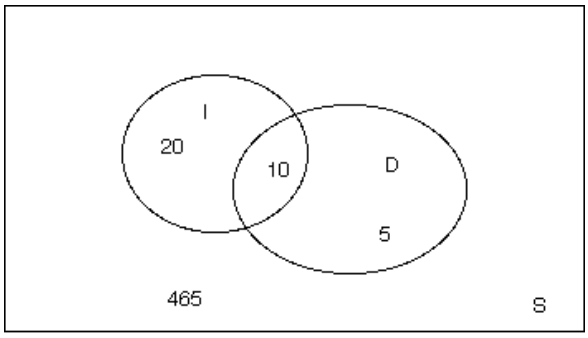
Lecture 7 - Content

- ☐ **Conditional probability**
- ☐ **Bayes rule**
- ☐ **Integer valued random variables**

Conditional probability (cont)

Example (Defect machine parts)**.** Suppose that 500 machine parts are inspected before they are shipped.

- $I =$ (a machine part is improperly assembled)
- $D =$ (a machine part contains one or more defective components)



Example (cont)

Assumption: equal probabilities in the selection of one of the machine parts.
$\Rightarrow$ Using the classical concept of probability we get:

General multiplication rule of probability

Theorem 8 (General multiplication rule of probability)**.** If A and B are any events in Ω, then

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(B) \times \mathrm{P}(A|B), \text{ if } \mathrm{P}(B) \neq 0, \text{ changing } A \text{ and } B \text{ yields} \\ &= \mathrm{P}(A) \times \mathrm{P}(B|A), \text{ if } \mathrm{P}(A) \neq 0. \end{aligned}$$

Proof. This holds because,

$$\mathrm{P}(A|B) := \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \text{ etc.}$$

□

What happens if $\mathrm{P}(A|B) = \mathrm{P}(A)$?
$\Rightarrow$ additional information of B is of no use $\Rightarrow$ special multiplication rule!

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B).$$

Definition of independence of events

Definition 5. If A and B are any two events in a sample space Ω, we say that A is independent of B if and only if

From the general multiplication rule it follows that if $\mathrm{P}(A|B) = \mathrm{P}(A)$ then $\mathrm{P}(B|A) = \mathrm{P}(B)$ and we say simply that A and B are independent.

Alternative View of Independence

Alternatively, if A and B are independent then $P(A \cap B) = P(A) \times P(B)$ and hence,

$$P(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \quad \text{(using Baye's rule)}$$

$$= \frac{\mathrm{P}(A) \times \mathrm{P}(B)}{\mathrm{P}(A)} \quad \text{(using independence)}$$

$$= \mathrm{P}(B).$$

which can also be interpreted as saying that knowing A does not effecting the probability of B.

Independence

In other words the events A and B are independent if the chance that one happens **remains the same** regardless of how the other turns out.

Example. Suppose that we toss a fair coin twice. Let
$A = \{\text{heads of the first toss}\}$
and
$B = \{\text{heads of the second toss}\}$.
Now suppose A occurred. Then

$$P(\{B \text{ knowning } A \text{ has happened}\}) = \tfrac{1}{2}.$$

Independence – Example 2

Example. Consider the following 6 boxes

1	2	3	1	2	3

Suppose we select a box at random, as it is drawn you see that it is **green**. Then

$$P(A = \{\text{getting a "2"}\}) = \frac{2}{6} = \frac{1}{3}$$

$$P(B = \{\text{getting a "2" if I know it is } \textbf{green}\}) = \frac{1}{3}$$

Knowing the selected box is **green** has not changed our knowledge about which numbers might be drawn.

Hence, the events A and B are independent.

Independence – Example 3

Example. Consider the following 6 boxes

1	1	2	1	2	2

Suppose we select a box at random, as it is drawn you see that it is **green**. Then

$$P(A = \{\text{getting a "2"}\}) = \frac{3}{6} = \frac{1}{2}$$

$$P(B = \{\text{getting a "2" if I know it is } \textbf{green}\}) = \frac{1}{3}$$

Knowing the selected box is **green** **HAS CHANGED** our knowledge about which numbers might be drawn.

Hence, the events A and B are **NOT** independent.

Independence – Example 4

Example. Two cards are drawn at random from an ordinary deck of 52 playing cards. What is the probability of getting two aces if

(a) the first card is replaced before the second is drawn?

(b) The first card is not replaced before the second card is drawn?

$\Rightarrow$ Independence is violated when the sampling is without replacement.

Independence – Example 5

Medical records indicate that the proportion of children who have had measles by the age of 8 is 0.4. The corresponding proportion for chicken pox is 0.5. The proportion who have had both diseases by the age of 8 is 0.3. An infant is randomly selected. Let A represent the event that he contracts measles, and B that he contracts chicken pox, by the age of 8 years.

- Estimate $P(A)$, $P(B)$ and $P(A \cap B)$.
  $P(A) = 0.4$, $P(B) = 0.5$ and $P(A \cap B) = 0.3$.
- Are A and B independent?
  $P(A) \times P(B) = 0.2 \neq P(A \cap B) = 0.3$, so NO, A and B are not independent.

Bayes rule

Example (The burgers are better...)**.** Assume you get your burgers

- 60% from supplier B_1
- 30% from supplier B_2
- 10% from supplier B_3

$\Rightarrow \mathrm{P}(B_1) = 0.6$, $\mathrm{P}(B_2) = 0.3$, and $\mathrm{P}(B_3) = 0.1$.

Interested in the event $A =$(good burger).

Example (cont)

It follows that,

$$A \;=\; A \cap (B_1 \cup B_2 \cup B_3) = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3).$$

Note that $(A \cap B_1)$, $(A \cap B_2)$ and $(A \cap B_3)$ are mutually exclusive.
By Axiom 3 we get

$$\begin{aligned} \mathrm{P}(A) \;&=\; \mathrm{P}((A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)) \\ &=\; \mathrm{P}(A \cap B_1) + \mathrm{P}(A \cap B_2) \;+ \mathrm{P}(A \cap B_3). \end{aligned}$$

Remember the general multiplication rule:
We already know that

$$\begin{aligned} \mathrm{P}(A \cap B) \;&=\; \mathrm{P}(B) \times \mathrm{P}(A|B), \text{ if } \mathrm{P}(B) \neq 0, \\ &=\; \mathrm{P}(A) \times \mathrm{P}(B|A), \text{ if } \mathrm{P}(A) \neq 0. \end{aligned}$$

Example (cont)

So we can write

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(B_1)\cdot \mathrm{P}(A|B_1) + \mathrm{P}(B_2)\cdot \mathrm{P}(A|B_2) + \mathrm{P}(B_3)\cdot \mathrm{P}(A|B_3) \\ &= 0.6\cdot \underbrace{\mathrm{P}(A|B_1)}_{0.95,\text{ very good}} + 0.3\cdot \underbrace{\mathrm{P}(A|B_2)}_{0.80,\text{ sufficient}} + 0.1\cdot \underbrace{\mathrm{P}(A|B_3)}_{0.65,\text{ insufficient}} \\ &= 0.875. \end{aligned}$$

What did the example teach us?

Strategy: decompose complicated events into mutually exclusive simple(r) events!

Total probability rule

Theorem 9 (Total probability rule)**.** If $B_1, B_2, \ldots, B_n$ are mutually exclusive events such that $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$ then for any event $A \subset \Omega$,

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(B_i) \times \mathrm{P}(A|B_i).$$

Example (Burger, cont)**.** We know already that supplier B_3 is bad. So what is $P(B_3|A)$ (if a burger is good is it from B_3)? By definition of the conditional probability, since $P(A) > 0$,

$$\begin{aligned} \mathrm{P}(B_3|A) &= \frac{\mathrm{P}(A \cap B_3)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B_3 \cap A)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B_3) \times \mathrm{P}(A|B_3)}{\sum_{i=1}^{3} \mathrm{P}(B_i) \times \mathrm{P}(A|B_i)} \\ &= \frac{0.1 \times 0.65}{0.875} = 0.074. \end{aligned}$$

After we know that a burger is good the probability that it comes from B_3 decreases from 0.1 to 0.074.

Bayes' rule or Theorem

What we just derived is the famous formula, called Bayes' rule or theorem.

Theorem 10 (Bayes Rule)**.** If $B_1, B_2, \ldots, B_n$ are mutually exclusive events such that $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$ then for any event $A \subset \Omega$,

$$\mathrm{P}(B_j|A) = \frac{\mathrm{P}(A|B_j) \times \mathrm{P}(B_j)}{\sum_{i=1}^{n} \mathrm{P}(A|B_i) \times \mathrm{P}(B_i)}.$$

The probabilities $\mathrm{P}(B_i)$ are called the priori probabilities and the probabilities $\mathrm{P}(B_i|A)$ the posteriori probabilities, $i = 1, \ldots, n$.

Reverend Thomas Bayes (1701 - 1761)

- Born in Hertfordshire (London, England),
- was a Presbyterian minister,
- studied: theology and mathematics,
- best known for Essay Towards Solving a Problem in the Doctrine of Chances ,
- where Bayes' Theorem was first proposed.
- Words: Bayes' rule, Bayes' Theorem, Bayesian Statistics.

Example of Bayes Rule – Screening test for Tuberculosis

	TB (D^+)	No TB (D^-)	
X-ray Positive (S^+)	22	51	73
X-ray Negative (S^-)	8	1739	1747
	30	1790	1820

What is the probability that a randomly selected individual has tuberculosis given that his or her X-ray is positive given that $P(D^+) = 0.000093$?

- $P(D^+) = 0.000093$ which implies that $P(D^-) = 0.999907$.
- $P(S^+|D^+) = 22/30 = 0.7333$
- $P(S^+|D^-) = 51/1790 = 0.0285$

$$\begin{aligned} P(D^+|S^+) &= \frac{P(S^+|D^+)P(D^+)}{P(S^+|D^+)P(D^+) + P(S^+|D^-)P(D^-)} \\ &= \frac{0.7333 \times 0.000093}{0.7333 \times 0.000093 + 0.0285 \times 0.999907} = 0.00239 \end{aligned}$$

Integer valued random variables

Many observed numbers are the random result of many possible numbers.

Definition 6. A random variable X is a real-valued function of the elements of a sample space Ω.

Note that such functions are denoted with capital letters and their images (outcomes) with lower case letters, e.g. x.

Examples.

- How many times (X) will you be caught speeding?
- What will your final mark (Y) for MATH1905 be?
- How old (Z, in years) do you think your stats lecturer is?

Random Variable Example – 3 Coins

Consider tossing three coins. The number of heads showing when the coins land is a random variable: it assigns the number 0 to the outcome {T, T, T}, the number 1 to the outcome {T, T, H}, the number 2 to the outcome {T, H, H}, and the number 3 to the outcome {H, H, H }.

Random Variable Example – 3 Coins

Events	Random Variable	Probability
TTT		
TTH		
THT		$P(X=0)=\frac{1}{8}$
THH	X = { Number of Heads }	$P(X=1)=\frac{3}{8}$
HTT		$P(X=2)=\frac{3}{8}$
HTH		$P(X=3)=\frac{1}{8}$
HHT		
HHH		

Random Variable Notation – 3 Coins

We use upper case letters to denote "unobserved" random variables, say X, and lower case letters to their observed values, in this case x.

For example, in the above example before the three coins land we denote the number of heads X, after the coins have landed we denote the number of coins x so that we can write $P(X = x)$.

The mother of all examples: Bernoulli trials!

Definition 7. Bernoulli trials satisfy the following assumptions:

(i) there are only two possible outcomes for each trial,

(ii) the probability of success is the same for each trial,

(iii) the outcomes from different trials are independent,

(iv) there are a fixed number n of Bernoulli trials conducted.

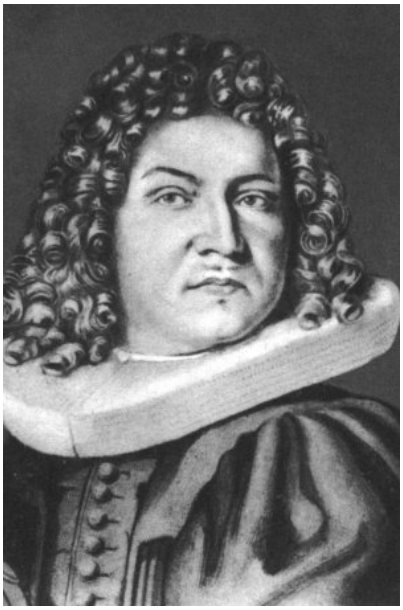
Example ($n = 1$, coin)**.** Ω: Head or Tail. We can describe the trial (before flipping the coin) in full detail. Consider a function

$$X : \{H, T\} \mapsto \{0, 1\} \quad \text{s.t.} \quad X(H) = x_H = 1 \quad \text{and} \quad X(T) = x_T = 0.$$

What is the probability that $X = x_H = 1$?

$$\mathrm{P}(X = 1) = \mathrm{P}(X = x_H) = \mathrm{P}(H) = p = 1/2 \Rightarrow \mathrm{P}(X = 0) = 1/2.$$

Jacob Bernoulli (1654–1705)



- Born in Basel (Switzerland),
- 1 of 8 mathematicians in his family,
- studied: theology $\rightarrow$ maths & astro,
- best known for Ars Conjectandi (The Art of Conjecture),
- application of probability theory to games of chance, introduction of the law of large numbers.
- Words: Bernoulli trial, Bernoulli numbers.

Tuesday, 21st August 2011

Lecture 8 - Content

- ☐ **Distribution of a random variable**
- ☐ **Binomial distribution**
- ☐ **Mean of a distribution**

Random Variables Reminder

Distribution of a random variable

Definition 8. The probability distribution of a integer-valued random variable X is a list of the possible values of X together with their probabilities

$$p_i = \mathrm{P}(X = i) \geq 0 \quad \text{and} \quad \sum_i p_i = 1.$$

There is nothing special with the subscript i; we could and will equally well use j, k, x etc.

Definition 9. The probability that the value of a random variable X is less than or equal to x, that is

$$F(x) = \mathrm{P}(X \leq x),$$

is called the cumulative distribution function or just the distribution function.

Also, note that for integer valued random variables that

$$P(X = x) = F(x) - F(x - 1).$$

Example ($n = 3$, IT problems)**.** A network is fragile. By experience: $\mathrm{P}(\mathsf{F}) = 0.1 = 1 - p$ that in any given week ≥ 1 major problem; $\mathrm{P}(\mathsf{S}) = 0.9 = p$ that there is none, respectively. Out of 3 weeks, how many weeks, X, had ≥ 1 problem and with what probability?

(a) All possible outcomes:

FFF SFF FSF FFS
SSF FSS SFS SSS

(b) What is the probability of each outcome? Use special multiplication rule of probability because sessions are independent!?

(c) What is the probability distribution of the number of successes, X, among the 3 sessions.

Example (cont)

$$\begin{aligned}
\mathrm{P}(X=0) &= \mathrm{P}(\mathsf{FFF}) = \mathrm{P}(\mathsf{F})\cdot\mathrm{P}(\mathsf{F})\cdot\mathrm{P}(\mathsf{F}) = (1-p)^3 \\
\mathrm{P}(X=1) &= \underbrace{\mathrm{P}(\mathsf{SFF}\cup\mathsf{FSF}\cup\mathsf{FFS})}_{\text{mutually exclusive events}} = \mathrm{P}(\mathsf{SFF})+\mathrm{P}(\mathsf{FSF})+\mathrm{P}(\mathsf{FFS}) \\
&= 3\times(1-p)^2p = \binom{3}{1}(1-p)^2p, \text{ select one } S \text{ out of } 3 \text{ trials.}
\end{aligned}$$

similarly we get for $X=2$ and $X=3$

$$\begin{aligned}
\mathrm{P}(X=2) &= \binom{3}{2}(1-p)p^2, \text{ select two S out of } 3 \text{ trials,} \\
\mathrm{P}(X=3) &= \binom{3}{3}p^3, \text{ select three S out of } 3 \text{ trials.}
\end{aligned}$$

Binomial distribution

We can generalise this result for any $n \geq 1$ and success probability $p \in [0,1]$.

Definition 10. The probability distribution of the number of successes $X=i$ in $n \in \mathbb{N}$ independent Bernoulli trials is called the binomial distribution,

$$p_i = \mathrm{P}(X=i) = \binom{n}{i}p^i(1-p)^{n-i}.$$

The success probability of a single Bernoulli trial is p and $i=0,1,\ldots,n$.

To say that the random variable X has the binomial distribution with parameters n and p we write $X \sim \mathcal{B}(n,p)$.

This defines a family of probability distributions, with each member characterized by a given value of the parameter p and the number of trials n.

Binomial distribution

Since p_i, $0 \leq i \leq n$ is a probability distribution we have the identity (which we will use later on)

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1$$

for any $0 \leq p \leq 1$.

A special case of the Binomial distribution is the Bernoulli distribution where $n = 1$ and

$$\mathrm{P}(X_i = i) = p^i (1-p)^{1-i}.$$

There is another special relationship between the Bernoulli distribution and the Binomial distribution.

If $X_i \sim$ Bernoulli(p) for $1 \leq i \leq n$ and $Y = \sum_{i=1}^{n} X_i$ then

$$Y \sim \mathcal{B}(n, p).$$

Example (Dice)**.** Roll a fair dice 9 times. Let X be the probability of sixes obtained. Then $X \sim \mathcal{B}(9, 1/6)$; that is

With your table calculator or with R:

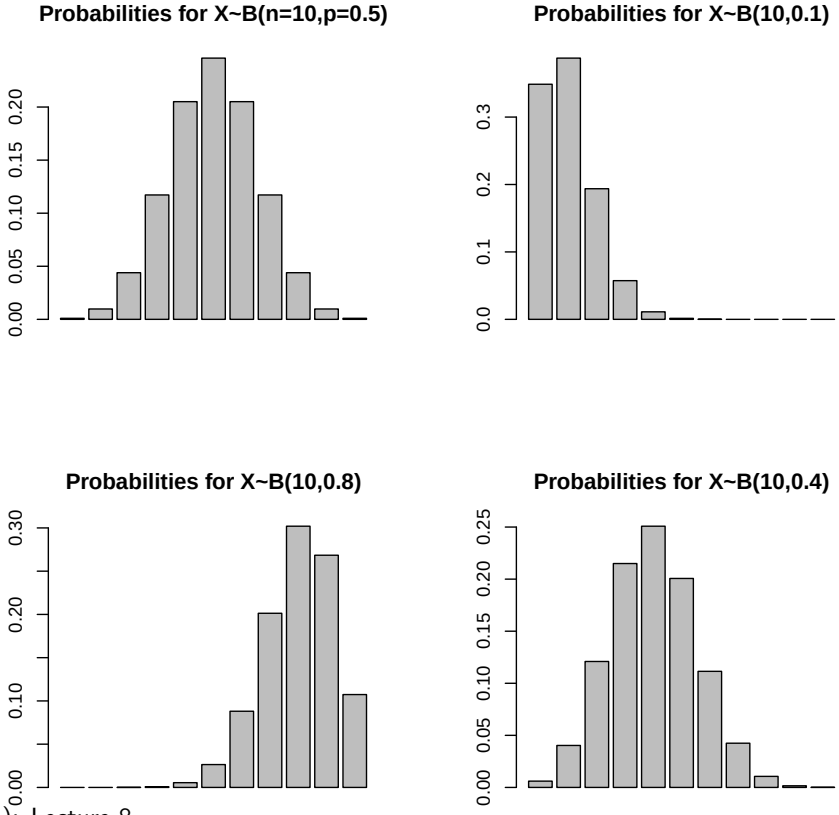
```
> n = 9;
> p = 1/6;
> round(dbinom(0:n,n,p),4) # dbinom for B(n,p) prob's
 [1] 0.1938 0.3489 0.2791 0.1302 0.0391
 [5] 0.0078 0.0010 0.0001 0.0000 0.0000
> pbinom(1,n,p) # for B(n,p) cumulative probabilities
[1] 0.5426588
```

Hence, $\mathrm{P}(X = 4) = 0.0391$ and $\mathrm{P}(X < 2) = F(1) = 0.5426588$.

Shape of the binomial distribution

- We get a binomial distribution if
  1. we are counting something over a fixed number of trials or repetitions,
  2. the trials are independent and
  3. the probability of the outcome of interest is constant across trials.
- The binomial distribution is centred at $n \times p$,
- the closer p to $1/2$ the more symmetric the distribution/histogram,
- the larger n the closer the shape to a bell (normal).

```
par(mfrow=c(2,2)); n =10 # and for n=50, 100, etc
barplot(dbinom(0:n,n,1/2))
title(main="Probabilities for X~B(n=10,p=0.5)")
barplot(dbinom(0:n,n,0.1))
title(main="Probabilities for X~B(10,0.1)")
barplot(dbinom(0:n,n,0.8))
title(main="Probabilities for X~B(10,0.8)")
barplot(dbinom(0:n,n,0.4))
title(main="Probabilities for X~B(10,0.4)")
```

Example. In a small pond there are 50 fish, 20 of which have been tagged. Seven fish are caught and X represents the number of tagged fish in the catch. Assume each fish in the pond has the same chance of being caught. Is X binomial

(a) if each fish is *returned* before the next catch?

(b) if the fish are *not returned* once they are caught?

Mean of a distribution

Definition 11. For a random variable X taking values $0, 1, 2, \ldots$ with

$$\mathrm{P}(X = i) = p_i \quad i = 0, 1, 2, \ldots$$

the mean or expected value of X is defined to be

$$\mu = \mathrm{E}(X) = \sum_i i \times p_i.$$

Interpretation of $\mathrm{E}(X)$

- Long run average of observations of X because $p_i \approx f_i/n$.
- Centre of balance of the probability density (histogram).
- Measure of location of the distribution.

Definition 12. For any function $g(X)$ we define the expected value $\mathrm{E}(g(X))$ by

$$\mathrm{E}(g(X)) = \sum_i g(i) \times p_i.$$

Expectation of a Dice Roll

Let $X = \{\text{Face showing from a dice roll}\}$ where $p_i = \mathrm{P}(X = i) = 1/6$ for $i = 1, 2, \ldots, 6$. Then

$$\begin{aligned} \mu &= \mathrm{E}(X) \\ &= \sum_{i=1}^{6} i \times p_i \\ &= \sum_i i \times 1/6 \\ &= 3.5. \end{aligned}$$

Note: the expected value in this case is not one of the observed values.

Mean of a distribution (cont)

Theorem 11. For constants a and b

$$\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b.$$

Proof.

□

Expectation of $X \sim \mathcal{B}(n, p)$

Theorem 12. The expectation of $X \sim \mathcal{B}(n, p)$ is $\mathrm{E}(X) = np$.

Proof.

□

Example (Multiple choice section in M1905 exam is worth 35%).
20 questions and each question has 5 possible answers. A student decides to answer the questions by selecting an answer at random.

(a) What is the expected number of correct responses?

(b) Probability that the student has more than 10 correct answers?

(c) If the student scores 4 for a correct answer but -1 for a wrong response, what is his expected score?

Monday, 27th August 2012

Lecture 9 - Content

- ☐ **Variance of a distribution**
- ☐ **More integer-valued distributions**
- ☐ **Probability generating functions**

Expectation of a distribution – Reminders

The expectation of a distribution (or expectation of a random variable) is the mean of the probability distribution (a measure of distribution location).

Note that

- $\mathrm{E}(X) = \sum_i i \times p_i = \sum_i i \times \mathrm{P}(X = i)$ and
- $\mathrm{E}(g(X)) = \sum_i g(i) \times p_i = \sum_i g(i) \times \mathrm{P}(X = i)$.

Variance of a distribution

Example. Suppose X (e.g. number of shoes in suitcase) takes the values 2, 4 and 6 with probabilities

i	2	4	6
p_i	0.1	0.3	0.6

Hence,

What is $\mathrm{E}(X^2)$?

Suppose X (e.g. number of shoes in suitcase) takes the values 2, 4 and 6 with probabilities

i	2	4	6
p_i	0.1	0.3	0.6

What is $\mathrm{E}(X^2)$?

Solution 1: $\mathrm{E}(X^2) \overset{\text{Def}}{=} \sum g(i)p_i = \sum i^2 p_i = 26.8 \neq 5^2$.
Solution 2: $i \mapsto i^2 = j$ and $X \mapsto X^2 = Y$, use $\mathrm{E}(Y) = \sum_j j p_j$

j	4	16	36
p_j	0.1	0.3	0.6

The distribution of Y can be hard to get (e.g. for continuous rvs).

Definition 13. The variance of the random variable X is defined by

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}(X-\mu)^2 = \mathrm{E}(X^2) - \mu^2,$$

where $\mu = \mathrm{E}(X)$ and σ^2 is also a measure of spread.

This is like the large sample limit of a sample variance.
The standard deviation of X is $\sigma = \sqrt{\sigma^2}$.

Variance of a Linear Transformation

Theorem 13. For any constants a and b

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

Proof.

□

Example. If $X \sim \mathcal{B}(n, p)$ then we'll show later that $\mathrm{Var}(X) = n \times p \times (1 - p)$.

- Hence, if $p = 0$ or 1 then the variance is 0.
- the variance is largest when $p = 0.5$ and in this case it is $\sigma^2 = n/4$.

More integer-valued distributions

Geometric distribution

The binomial random variable is just one possible integer-valued random variable. Suppose we have an infinite sequence of independent trials, each of which gives a success with probability p and failure with probability $q = 1 - p$.

Definition 14. The geometric distribution with parameter p (= success prob.) has probabilities for the number of failures X before the first success,

$$p_i = \mathrm{P}(X = i) = q^i p, \quad i = 0, 1, 2, \ldots.$$

Note the probabilities add to 1:

$$\mathrm{P}(X = 0) + p_1 + \ldots = p + qp + q^2 p + \ldots = p(1 + q + q^2 + \ldots) = p\left(\frac{1}{1-q}\right) = 1$$

Example. A fair die is thrown repeatedly until it shows a six.

(a) What is the probability that more than 7 throws are required?

```
1-pgeom(7,1/6)        1-sum(dgeom(0:7,1/6))
```

(b) Is it more likely that an odd number of throws is required or an even number?

The Poisson approximation to the Binomial

The Poisson distribution often serves as a first theoretical model for counts which do not have a natural upper bound.

Possible examples

- modeling of number of accidents, crashes, breakdowns
- modeling radioactivity measured by the Geiger counter
- modeling of so-called rare events (meteorite impacts, heart attacks)

The Poisson distribution can be seen as the limiting distribution of $\mathcal{B}(n, p)$:

Approximation is good if $n \cdot p^2$ is small!

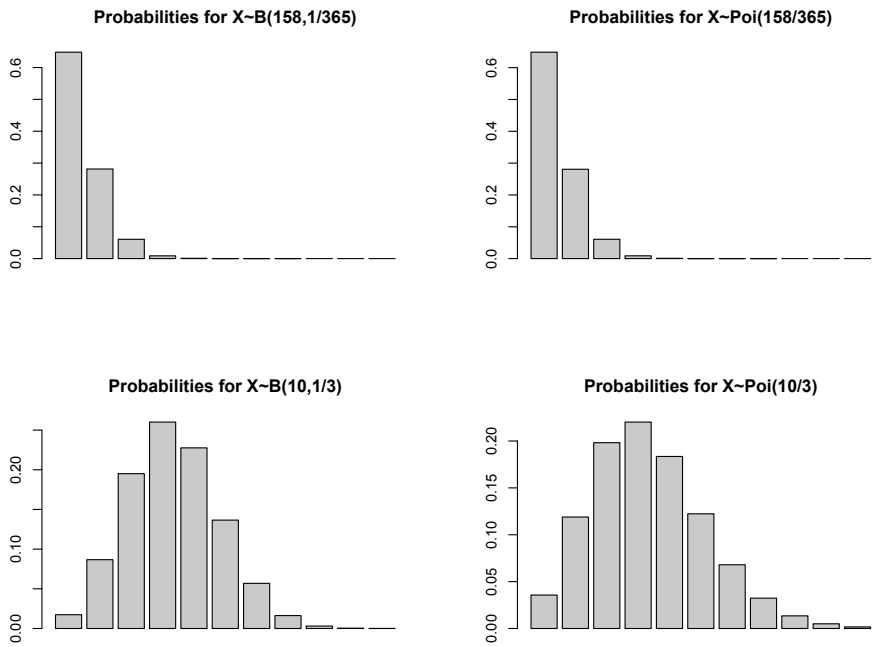
$X \sim \mathcal{B}(158, \frac{1}{365})$ and $n \cdot p^2 = 0.001186$:

```
> # What is the probability that of 158 people, exactly k have a birthday today?
> n = 158; p=1/365;
> round(dbinom(0:7,n,p),5);
[1] 0.64826 0.28139 0.06068 0.00867 0.00092 0.00008 0.00001 0.00000
> round(dpois(0:7,p*n),5);
[1] 0.64864 0.28078 0.06077 0.00877 0.00095 0.00008 0.00001 0.00000
```

But for $n = 10$

```
> n = 10; p=1/3;
> round(dbinom(0:4,n,p),5);
[1] 0.01734 0.08671 0.19509 0.26012 0.22761
> round(dpois(0:4,p*n),5);
[1] 0.03567 0.11891 0.19819 0.22021 0.18351
```


Probability distribution for $X \sim \mathcal{B}(n, p)$ and $X \sim \mathcal{P}(\lambda)$



Probability generating functions

Let $X \in \mathbb{N}$ and $p_i = \mathrm{P}(X = i)$, $i = 0, 1, 2, \ldots$

Definition 15. The probability generating function is defined as

$$\pi(s) = p_0 + p_1 s + p_2 s^2 + p_3 s^3 + \ldots.$$

Example. If X only takes a finite number of values (e.g. $X \sim \mathcal{B}(n, p)$) then $\pi(s)$ is a polynomial.

Alternatively (e.g. $X \sim \mathcal{P}(\lambda)$) $\pi(s)$ is a power series.

Properties of $\pi(s)$

Let $s \in [0, 1]$ then

- $0 \leq \pi(s) \leq 1$,
- $\pi(1) = p_0 + p_1 + \ldots = 1$,
- $\pi'(s) = p_1 + 2p_2 s + 3p_3 s^2 + \ldots \geq 0, \quad s \geq 0.$
- $\pi'(1) = p_1 + 2p_2 + 3p_3 + \ldots = \mathrm{E}(X)$ (if $\mathrm{E}(X)$ is finite),
- $\pi''(s) = 2p_2 + 6p_3 + 4 \cdot 3p_4 + \ldots$ at $s = 1$, so $\pi''(1) = \mathrm{E}(X(X-1))$ and

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}\, X)^2 = \pi''(1) + \pi'(1) - (\pi'(1))^2.$$

Example (Poisson distribution).

Example (Binomial distribution).

Tuesday, 28th August 2012

Lecture 10 - Content

- **Continuous random variables**
- **Chebyshev's inequality**

References from Phipps & Quine

- Section 2.2 pages 62-66.

Answer to Challenge Question

Show that

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

Let n be an integer. Then by the Binomial Theorem

$$(x+y)^n = \sum_{i=0}^{n}\binom{n}{i}x^i y^{n-i}.$$

Let $y=1$ and $x=-\frac{\lambda}{n}$ then

$$\begin{aligned}\lim_{n\to\infty}\left(1-\tfrac{\lambda}{n}\right)^n &= \lim_{n\to\infty}\sum_{i=0}^{n}\frac{n\times(n-1)\times\ldots\times(n-i+1)}{n^i}(-1)^i\frac{\lambda^i}{i!}\\ &= \lim_{n\to\infty}\sum_{i=0}^{n}(-1)^i\frac{\lambda^i}{i!}\end{aligned}$$

since

$$\lim_{n\to\infty}\frac{n\times(n-1)\times\ldots\times(n-i+1)}{n^i} = 1.$$

The last line is the Taylor series expansion for $e^{-\lambda}$.

Continuous random variables

Examples

Continuous random variables have images in $\mathbb{R}$, e.g.

- the speed of a car,
- the amount of alcohol in a person's blood after 4 standard drinks,
- the temperature at 1pm.

Distribution Function of a Continuous Random Variable

Definition 16. A distribution function, $F(x) = \mathrm{P}(X \leq x)$, is any function that satisfies

(i) $0 \leq F(x) \leq 1$ (F is a probability)

(ii) $F(x) \uparrow$, i.e. $F(x)$ is a monotonic increasing function of x.

(iii) If $a < b$ then $\mathrm{P}(a < X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) = F(b) - F(a)$.

(iv) $F(-\infty) = 0$, $F(+\infty) = 1$.

(v) $F(x)$ is right-continuous; i.e. for every number x^*, $\lim_{x \downarrow x^*} F(x) = F(x^*)$.

Key Property of Continuous Random Variables

Theorem 14. A continuous random variable X attains with probability zero any value of its image. That is

$$\mathrm{P}(X = x) = 0$$

for all real numbers $x \in \mathbb{R}$.

Proof. Note that the set $A = \{X = x\}$ is a subset of $B = \{x - \epsilon < X \leq x\}$ for any $\epsilon > 0$. Since, if $A \subset B$ then $\mathrm{P}(A) \leq P(B)$ we have

$$0 \leq \mathrm{P}(X = x) \leq P(x - \epsilon < X \leq x) = F(x) - F(x - \epsilon).$$

Due to the continuity of F we have

$$0 \leq \mathrm{P}(X = x) \leq \lim_{\epsilon \downarrow 0} F(x) - F(x - \epsilon) = 0.$$

□

Hence, if X is a continuous random variable then,

$$\mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X < b) = \mathrm{P}(a < X < b).$$

Probabilities for Continuous Random Variables

- Suppose that we focus on events $X \in (a, b]$, i.e. $(a, b]$ an interval of length $b - a > 0$.
- Dividing $(a, b]$ into n equal subintervals of width Δx; it follows that

$$\mathrm{P}(a < X \leq b) = \lim_{n \to \infty} \sum_{i=1}^{n} \widehat{f}(i; \Delta x) \times \Delta x.$$

where

$$\begin{aligned} \widehat{f}(i; \Delta x) &= \frac{\mathrm{P}(a + (i-1)\Delta x < X \leq a + i\Delta x)}{\Delta x} \\ &= \frac{F(a + i\Delta x) - F(a + (i-1)\Delta x)}{\Delta x} \end{aligned}$$

for $i = 1, \ldots, n$.

- Consider any sequence $i = i(n)$ such that

$$\lim_{n \to \infty} (a + i\Delta x) = x$$

for some $x \in (a, b]$ and let $f \geq 0$ be an integrable function in $\mathbb{R}$ such that

$$f(x) = \lim_{n \to \infty} \widehat{f}(i; \Delta x).$$

- Then

$$f(x) = \lim_{\Delta x \to 0} \frac{F(x) - F(x - \Delta x)}{\Delta x} = \frac{dF(x)}{dx}$$

and

$$\mathrm{P}(a < X \leq b) = \int_a^b \underbrace{f(x)}_{= \text{probability density function}} dx.$$

Probability density function

Definition 17. A probability density function or simply a probability density is any non-negative function $f(x) \geq 0$ such that

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Theorem 15. $F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$

As an immediate consequence of the Fundamental Theorem of Calculus,

$$f(x) = \frac{dF(x)}{dx}$$

as previously stated.

Indicator Functions

The following type of function appears quite frequently in Statistics when defining probability density functions.

Definition 18. The function $\mathbf{1}_A(x) = \mathbf{1}\{x \in A\}$ is called the indicator function of the set A. It has image 1 if $x \in A$ and image 0 if $x \notin A$.

(Although we have not yet defined the expectation of a continuous random variable it turns out that

$$E[\mathbf{1}_A(x)] = \mathrm{P}(A)$$

which is a useful property in certain contexts.)

Scaling of non-negative functions to construct density functions

Example. Find c s.t. the following non-negative function is a probability density of a random variable:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ cxe^{-4x^2} & \text{for } x > 0 \end{cases} = cxe^{-4x^2} \times \mathbf{1}_{(0,\infty)}(x).$$

Moments of continuous random variables

Definition 19. Let g be any continuous function. The expected value of $g(X)$ of a continuous random variable X having probability density f is defined by

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

The mean of X is given by

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

The kth moment about the mean of X is given by

$$\mu_k = \mathrm{E}[(X-\mu)^k] = \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx.$$

The variance of X is given by

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2.$$

Useful Results

The following results, which we showed hold for integer values random variables, also hold for continuous random variables:

$$E(aX + b) = aE(X) + b$$

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

Proof: Left as an exercise.

Uniform distribution

Definition 20. The uniform distribution, with parameters a and b, has

$$f(x) = \frac{1}{b-a} \cdot \mathbf{1}_{(a,b)}(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{elsewhere;} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x \geq b. \end{cases}$$

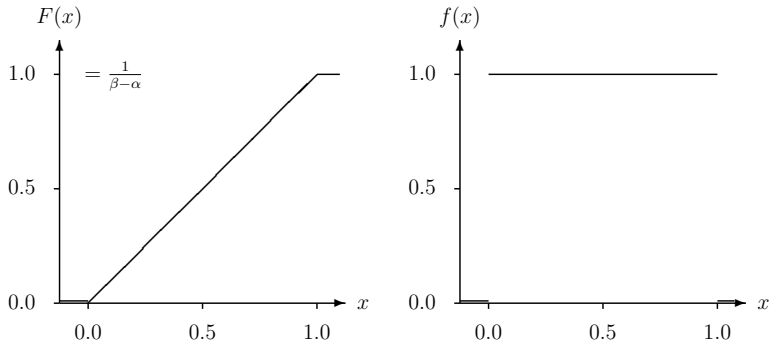
Short notation:

$$X \sim \mathcal{U}(a, b)$$

[The uniform distribution is potentially useful to model or to be applied in conjunction with rounding errors/effects, generating random variables, simulation studies.]

Uniform distribution

Example (Uniform distribution for $a = 0$ and $b = 1$)**.**



Uniform distribution – Expectation and Variance

Theorem 16. If $X \sim \mathcal{U}(a, b)$ then,

$$\mu = \mathrm{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = \mathrm{E}[(X-\mu)^2] = \frac{1}{12}(b-a)^2.$$

Proof.

Uniform distribution – R code

```
> n = 10000
> set.seed(1)
> x = runif(n)  # Generates Uniform(0,1) values
> hist(x)
> mean(x) # We should expect this value to be close to (0 + 1)/2
[1] 0.4990762
> var(x)  # We should expect this value to be close to (1 - 0)^2/12 = 1/12
[1] 0.08383338
> 1/12
[1] 0.08333333
```


Chebyshev's inequality

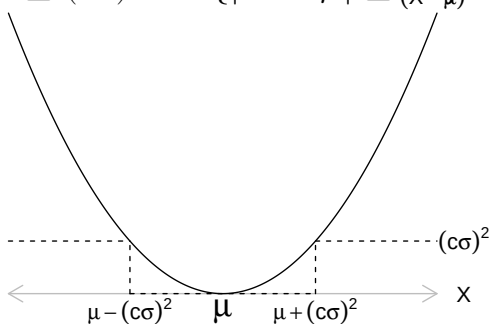
Links the three notions of probability, mean and variance.

Theorem 17. If a random variable X has mean μ and variance σ^2, then for any positive number c,

$$\mathrm{P}(|X - \mu| \geq c\sigma) \leq 1/c^2.$$

Proof. Note that

$$|X - \mu|^2 \geq (c\sigma)^2 \times \mathbf{1}\{|X - \mu| \geq c\sigma\} \quad (*)$$



From the definition of the expected value and the indicator function we have

$$\mathrm{E}[\mathbf{1}_A(X)] = \int_A f(x)dx = \mathrm{P}(A).$$

Hence, taking expectations on both sides of $(*)$ yields

$$\mathrm{E}[|X-\mu|^2] = \sigma^2 \geq (c\sigma)^2\,\mathrm{P}(|X-\mu| \geq c\sigma).$$

□

Examples

Example. Consider the IQ score where $\mu = E(X) = 100$ and $\sigma^2 = \mathrm{Var}(X) = 10^2$. What can we say about $P(X > 150)$?

Examples

Example. Suppose that $X \sim \mathcal{U}(0, 10)$. Use Chebyshev's inequality to bound the probability $\mathrm{P}(|X - 5| > 4)$.

Monday, 3rd September 2012

Lecture 11 - Content

- ☐ **Normal random variables**
- ☐ **Standardized random variables**
- ☐ **Pseudo-random numbers in R**

References from Phipps & Quine

- ☐ Section 2.3 pages 66-69.

Normal random variables

- The normal distribution or the normal probability density dates back to the 18th century.
- Abraham de Moivre (1667–1754) and Pierre-Simon Marquis de Laplace (1749–1827) find the normal distribution as an approximate distribution to the Binomial.
- Johann Carl Friedrich Gauss (1777–1852) assumed the normal distribution of errors in the context of the least squares method.

Alternative names for the normal

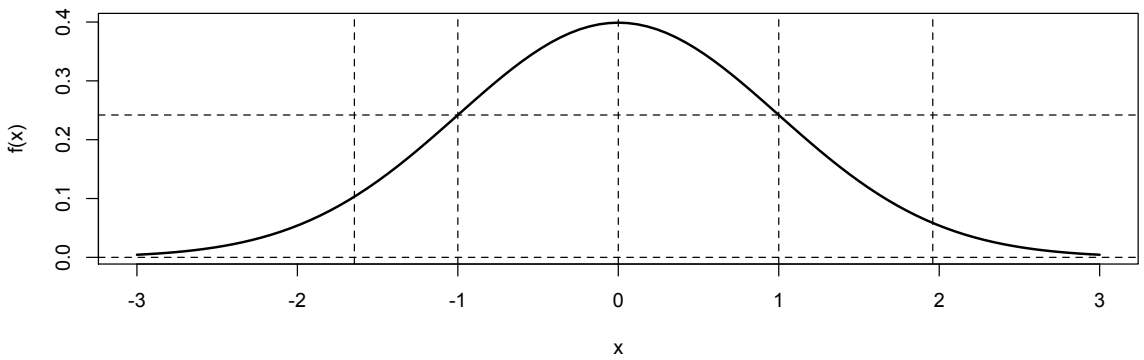
- Gaussian distribution,
- Bell distribution.

Normal probability density

Definition 21. The normal probability density is given by

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \Rightarrow X \sim \mathcal{N}(\mu,\sigma^2).$$

It has location parameter $\mu = \mathrm{E}(X)$ and scale parameter $\sigma^2 = \mathrm{Var}(X)$.



Some useful facts

- The density function of the normal distribution has the shape of a symmetric bell curve.
- Its maximum is at $x = \mu$ and it has inflection points at

Why is the normal distribution so famous?

- If $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow$ simple results and theorems!
- Central limit theorem: the mean of many independent random variables $X_1, X_2, \ldots$ (having finite variances) is approximately normally distributed

$$\sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right) \approx \mathcal{N}(0, 1).$$

- The distribution of measurement errors is often very similar to the normal distribution

Standard normal random variable

Definition 22. The normal with mean 0 and variance 1 is called the standard normal random variable and is generally denoted by Z. Thus

$$Z \sim \mathcal{N}(0, 1)$$

with

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty.$$

Remark

The integral $\int_{-\infty}^{\infty} e^{-z^2} dz$ is called the Euler-Poisson integral and equals $\sqrt{\pi}$. See additional slides at the end of this lecture.

Normal distribution function

The normal distribution function $F(x; \mu, \sigma^2)$ has no closed form, thus

$$\begin{aligned} F(x; \mu, \sigma^2) &= \mathrm{P}(X \leq x) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt. \end{aligned}$$

In practice the normal distribution function needs to be approximated numerically.

There are several nice ways of doing this, but they rely on transforming the integral into "standard form".

Standardised random variables

Theorem 18. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then the centred and standardised random variable

$$\begin{aligned} Z = \frac{X-\mu}{\sigma} &\sim \mathcal{N}(0,1) \\ \Rightarrow \mathrm{P}(Z \leq z) := \Phi(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt. \end{aligned}$$

Proof. The proof is left as an exercise. Begin with $F(x; \mu, \sigma^2)$, substitute $Z = g(X) = \frac{X-\mu}{\sigma}$, continue with calculus knowledge till you get $F(z; 0, 1)$. □

Thanks to the theorem it is sufficient to know the (tabulated) probabilities of the standard normal distribution e.g. from the formula sheet, software, or any other source.

Standardizing random variables

Definition 23. If X is any random variable with mean μ and variance σ^2 then

$$Z = \left(\frac{X - \mu}{\sigma}\right)$$

is called the standardized version of X.

Theorem 19. If $Z = \left(\frac{X-\mu}{\sigma}\right)$ with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$ then,

$$\mathrm{E}(Z) = 0 \quad \text{and} \quad \mathrm{Var}(Y) = 1.$$

Proof. Follows from the definitions of E and Var and from the identity

$$\mathrm{E}(a + bX) = a + b\,\mathrm{E}(X).$$

□

Useful identities for the normal

- $\phi(-z) = \phi(z)$ because of symmetry of ϕ.
- $\Phi(-z) = 1 - \Phi(z)$ because of symmetry of $\phi \geq 0$ and $\int \phi(t)dt = 1$.
- $\mathrm{P}(|Z| \leq z) = 2\Phi(z) - 1$ because

$$\mathrm{P}(|Z| \leq z) = \mathrm{P}(-z \leq Z \leq z) = \Phi(z) - \Phi(-z).$$

Example. $X \sim \mathcal{N}(3, 2^2)$. Find $\mathrm{P}(X \leq 4)$ and $\mathrm{P}(X < 1.24)$.

Example. $X \sim \mathcal{N}(5, 3^2)$. Find c such that $\mathrm{P}(X > c) = 0.1$.

Exponential distribution and friends

Definition 24. The exponential distribution, with parameter λ, has probability density

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

and distribution function given by

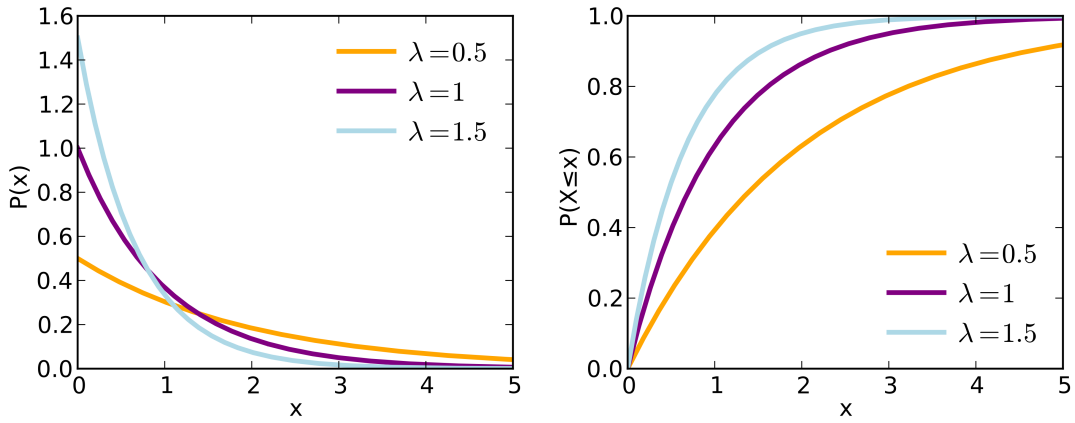
$$F(x) = 1 - e^{-\lambda x} = \int_0^x \lambda e^{-\lambda t} dt \quad t > 0.$$

To say that the random variable X has the exponential distribution with parameter $\lambda > 0$ we write

$$X \sim \mathcal{E}(\lambda).$$

Sometimes an alternative parameterisation is used where $\beta = 1/\lambda$ becomes the parameter of the distribution.

Plots of the Exponential Distribution



Properties and applications

- The mean and variance of $X \sim \mathcal{E}(\lambda)$ equals $\mathrm{E}(X) = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$.
- Waiting times X between two events, failure distribution with underlying constant failure rate, distance between roadkill on a street etc are often modelled by $X \sim \mathcal{E}(\lambda)$.
- The exponential distribution is memoryless

$$\mathrm{P}(X > t + h) = \mathrm{P}(X > t)\,\mathrm{P}(X > h), \quad t, h > 0,$$

and therefore

$$\mathrm{P}(X > t + h | X > t) = \mathrm{P}(X > h),\ t, h > 0$$

Example – Exponential Distribution

Example. Suppose that the amount of time one spends in a bank is exponentially distributed with mean 10 minutes, $\lambda = 1/10$. What is the probability that a customer will spend more than 15 minutes in the bank? What is the probability that a customer will spend more than 15 minutes in the bank given that he is still in the bank after 10 minutes?

Pseudo-random numbers in R

```
> # generating samples of 'independent' continous 'random' variables
> set.seed(010909)  # set random seed to 01 Sep 09
> n = 10            # choose sample size of 10
> rnorm(n)          # 10 pseudo-standard-normal random numbers
 [1] -1.6657 -0.1583 -0.2662 -0.9809 -1.0117 -1.2175  0.0986  0.7802  2.3596 -0.3192
> runif(10)         #       ...-uniform [0,1]
 [1] 0.1784 0.8924 0.7842 0.4014 0.7271 0.2366 0.1984 0.0003 0.7880 0.8027
> rexp(10)          #       ...-exponential with mean 1
 [1] 0.5629 0.4597 0.1792 0.5607 0.5740 0.7506 2.4387 0.7580 0.2380 0.0726
> # hence the r... in front of norm, unif, exp signifies drawing random numbers
> # the d signifies density, the p = P(X <= x), the q returns the quantile.
> curve(dnorm,from=-3,to=3)
> pnorm(95,mean=100,sd=10) # qnorm(0.3085375,mean=100,sd=10) = 95
[1] 0.3085375
> 1-pnorm(95,mean=100,sd=10)
[1] 0.6914625
> pnorm(95,mean=100,sd=10,lower.tail = FALSE)
[1] 0.6914625
```


The Gamma Distribution

A generalisation of the exponential distribution leads to the family of gamma distributions.

Definition 25. The gamma distribution, with parameters α and β, has probability density

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x, \alpha, \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

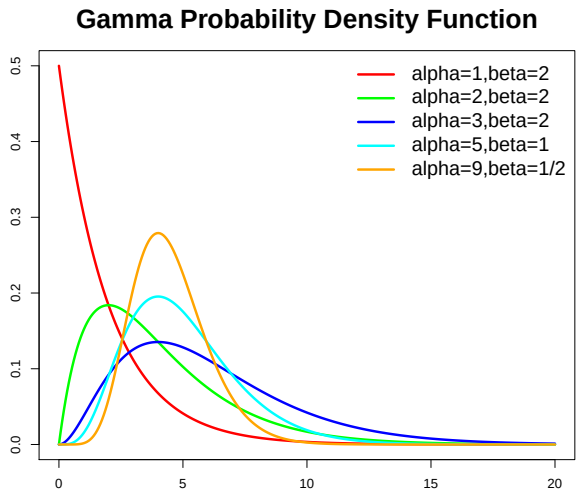
If $f(x)$ has the above density then we write $X \sim \text{Gamma}(\alpha, \beta)$.

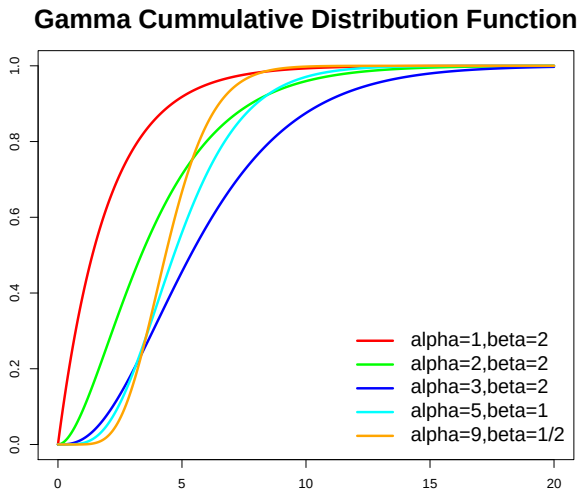
Note that $\Gamma(\alpha)$ is a value of the gamma function, defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

It is a generalisation of $n!$, $n \in \mathbb{N}$.

Plots of the Gamma Distribution





Friends of the Gamma Distribution

Depending on special choices of the parameters α and β the gamma distribution becomes

- for $\alpha = 1$ the exponential distribution (with $\beta = 1/\lambda$),
- for $\alpha = 1/2$ and $\beta^{-1} = \sigma^2/2$ the distribution of $Y = X^2$, if $X \sim \mathcal{N}(0, \sigma^2)$,
- for $\alpha = m/2$, $m \in \mathbb{N}$, and $\beta = 2$ the chi-square distribution.

Properties of the Gamma Distribution

Theorem. If $X \sim \text{Gamma}(\alpha, \beta)$ then, the mean and variance of X equals

$$\mu = E[X] = \alpha\beta \quad \text{and} \quad \sigma^2 = E[(X-\mu)^2] = \alpha\beta^2.$$

Proof. For the mean we have (proof for variance is similar):

$$\mu = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x \cdot x^{\alpha-1} e^{-x/\beta} dx$$

$$\overset{y=x/\beta}{\Rightarrow} \mu = \frac{\beta}{\Gamma(\alpha)} \underbrace{\int_0^\infty y^\alpha e^{-y} dy}_{=\Gamma(\alpha+1)=\alpha\Gamma(\alpha)} = \alpha\beta$$

□

The Gamma Function

Let

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

Using integration by parts shows that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for any $\alpha > 1$.
Remember: Integration by parts: $\int fG = FG - \int Fg$
For $f(x) = e^{-x}$ and $G(x) = x^\alpha$ it follows:

$$\begin{aligned}
\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x} dx = [-x^\alpha e^{-x}]_0^\infty - \int_0^\infty \alpha x^{\alpha-1}(-1)e^{-x} dx \\
&= -\lim_{x\to\infty} x^\alpha e^{-x} + \underbrace{0^\alpha e^{-0}}_{=0, \text{ since } \alpha>0} + \alpha \underbrace{\int_0^\infty x^{\alpha-1} e^{-x} dx}_{=\Gamma(\alpha)} \\
&= -\lim_{x\to\infty} \left(x^{-\alpha} \sum_{k=0}^\infty \frac{x^k}{k!} \right)^{-1} + \alpha \cdot \Gamma(\alpha) = \alpha \cdot \Gamma(\alpha)
\end{aligned}$$

Now we have the proof for $\Gamma(\alpha+1) = \alpha!, \ (\alpha \in \{1, 2, 3, \ldots\})$ if and only if $\Gamma(1+1) = 1\Gamma(1) = 1! = 1$.

That is easy:

$$\Gamma(1+1) = \int_0^\infty x e^{-x} dx = \int_0^\infty e^{-x} dx = [-e^{-x}]_0^\infty = 1.$$

On the Euler-Poisson or Gaussian integral

The Euler-Poisson integral is the improper integral

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

and exists because $\exp(-x^2)$ is continuous and bounded, i.e. $0 \leq e^{-x^2} < e^{-|x|+1}$ noting that $\int_{-\infty}^{\infty} e^{-|x|+1} dx = 2e$.

Instead of calculating I we show that

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right) \times \left(\int_{-\infty}^{\infty} e^{-y^2} dy\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2-y^2} dx dy = \pi.$$

For any point $(x, y) \in \mathbb{R}^2$ we have the alternative coordinate notation $x = r\cos\theta$ and $y = r\sin\theta$. Hence,

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} \times |J| dr d\theta,$$

where $|J|$ denotes the determinant of the Jacobi matrix, i.e. matrix of partial derivatives:

$$J = \begin{pmatrix} \partial x/\partial r & \partial y \partial r \\ \partial x/\partial\theta & \partial y/\partial\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix} \Rightarrow |J| = r(\cos^2\theta + \sin^2\theta) = r.$$

By substituting $r^2 = u$ we obtain,

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r \, dr d\theta = \int_0^{2\pi} \int_0^{\infty} \frac{1}{2} e^{-u} \, du d\theta = \int_0^{2\pi} \frac{1}{2} \, d\theta = \pi.$$

Tuesday, 4th September 2012

Lecture 12 - Content

- ☐ **Joint distributions**
- ☐ **Independent random variables**
- ☐ **Central limit theorem**

References from Phipps & Quine

- ☐ Section 2.4 pages 69-72.
- ☐ Section 3.2 pages 73-75.

Standard Normal Distribution

Let $Z \sim N(0,1)$ then

- The probability density function at z is given by

  ```
  > dnorm(z)
  ```

- The (cumulative) distribution function at z, $\Phi(z) = \mathrm{P}(Z < z)$, is given by

  ```
  > pnorm(z)
  ```

- The inverse (cumulative) distribution function at t, $\Phi^{-1}(t)$ or the value of z such that $\Phi(z) = t$, is given by

  ```
  > qnorm(t)
  ```

- To generate n random values from $Z \sim N(0,1)$ we use

  ```
  > rnorm(n)
  ```


Normal Distribution

Let $X \sim N(\mu, \sigma^2)$ then

- The probability density function at x is given by

  ```
  > dnorm(x,mu,sigma)
  ```

- The (cumulative) distribution function at x, $\Phi((x-\mu)/\sigma) = \mathrm{P}(X < x)$, is given by

  ```
  > pnorm(x,mu,sigma) # OR
  > pnorm( (x-mu)/sigma )
  ```

- The inverse (cumulative) distribution function at t, the value of x such that $\mathrm{P}(X < x) = t$, is given by

  ```
  > qnorm(t,mu,sigma)
  ```

- To generate n random values from $X \sim N(\mu, \sigma^2)$ we use

  ```
  > rnorm(n,mu,sigma)
  ```

Joint distributions

Independence of random variables

Let X be a real-valued random variable (e.g. normal, exponential, binomial) and $x \in \mathbb{R}$ any number, then

$$A = \{X \leq x\}$$

represents an event. Let Y be another real-valued random variable and

$$B = \{Y \leq y\}, \quad y \in \mathbb{R}.$$

Recall the definition of independence of events: A and B are independent iff

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

which is a special case of the general multiplication rule,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) = \mathrm{P}(B)\,\mathrm{P}(A|B) \quad \text{if} \quad \mathrm{P}(A), \mathrm{P}(B) \neq 0.$$

Definition 26. Two random variables X and Y are independent if and only if for any numbers x and y the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events.

Example.

- ($X =$ 'height', $Y =$ 'weight') from a random person are not independent.
- $X_1 =$ 'lottery numbers next draw' and $X_2 =$ 'lottery numbers in three weeks time' are
- $X_1 =$ 'todays rainfall' and $X_2 =$ 'tomorrows rainfall' are

From the above Definition 26 we easily get the joint cumulative distribution function and joint probability density function of independent random variables.

Joint distribution functions and densities

Definition 27. The joint cumulative distribution function of two random variables X and Y is

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y)$$

and the joint density function is denoted $f_{X,Y}(x, y)$.

Note that, if X and Y are continuous random variables, then $F_{X,Y}(x, y)$ and $f_{X,Y}(x, y)$ are related via

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

and

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) dt ds.$$

Marginal distribution functions and densities

Definition 28. If $F_{X,Y}(x, y)$ is the joint cumulative distribution function of two random variables X and Y then, $F_X(x)$ and $F_Y(y)$ are called the marginal cumulative distribution functions of X and Y, respectively.

For integer valued random variables the marginal probability mass functions can be calculated via

$$P(X = x) = \sum_y P(X = x, Y = y) \qquad \text{and} \qquad P(Y = y) = \sum_x P(X = x, Y = y)$$

while for continuous random variables the marginal density functions can be calculated via

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx.$$

From these the marginal cumulative distribution functions can be calculated in the usual way.

Expectations of Joint Distributions

Let $g(x, y)$ be a bivariate function and let X and Y be random variables with joint density function $f_{X,Y}(x, y)$.

If X and Y are discrete random variables then

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) f_{X,Y}(x, y).$$

If X and Y are continuous random variables then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)dxdy.$$

Independence

Definition 29. Let $F_X(x) = \mathrm{P}(X \leq x)$ and $F_Y(y) = \mathrm{P}(Y \leq y)$ be the cumulative distribution functions of the independent random variables X and Y then, the joint cumulative distribution function is

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x)\,\mathrm{P}(Y \leq y) = F_X(x)F_Y(y).$$

Definition 30. Let $f_X(x)$ and $f_Y(y)$ be the probability density functions of the independent random variables X and Y then, the joint probability density function is given by

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Independent random variables: rules for expectations and variances

Theorem 20 (Properties of E and Var)**.** Let X and Y be random variables then

1. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
2. if X and Y are independent then, $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$
3. if X and Y are independent then, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Note that for any two, not necessarily independent, random variables

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$$

where

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))].$$

Proof of 1. discrete case only:

Central limit theorem

Many observed phenomena can be modelled as the sum of several random variables:

- total weight of passengers in a lift,
- total of available funds

or means of random variables

- average class mark,
- average height and weight,
- average temperature in Sydney.

The central limit theorem is useful in these types of situations.

Some useful facts about the normal distribution

Theorem 21. Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, let X and Y be independent and let a and b be two real numbers. Then

$$Z = aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2).$$

Proof: Not in MATH 1905.
In general, let $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ be independent and a_i be real numbers for $1 \leq i \leq n$ then

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, a_i^2\sigma_i^2\right).$$

Example

Example (Mean and variance of the sample mean $\overline{X}$)**.** Let the n random variables $X_1, X_2, \ldots, X_n$ be pairwise independent and each have the same distribution with mean μ and variance σ^2. Then for the sample mean, that is $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have,

i) mean: $\mu_{\overline{X}} = E(\overline{X}) = \mu$

ii) variance: $\sigma_{\overline{X}}^2 = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$

Sums of normal random variables

Theorem 22. If all $X_i \sim \mathcal{N}(\mu, \sigma^2)$ then,

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

[This is only true for normal rvs; in STAT2911 moment generating functions are introduced that make a simple proof available].

Example. X_1, X_2, X_3 are independent random variables with

i	0	1	3
p_i	1/3	1/3	1/3

$\overset{T_2=X_1+X_2}{\Rightarrow}$

i	0	1	2	3	4	6
p_i	1/9	2/9	1/9	2/9	2/9	1/9

i	0	1	2	3	4	5	6	7	9
p_i	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{4}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$

(Note, the distribution of T_3 clusters around the mean $\mathrm{E}\,T_3 = 4$.)

Theorem 23 (CLT, central limit theorem)**.** If $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance $0 < \sigma^2 < \infty$ then,

$$\mathrm{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \le x\right) \to \Phi(x) = \mathrm{P}(Z \le x) \quad \text{as} \quad n \to \infty.$$

Proof. Postponed to second year... □

Thus for n large (here $n \ge 25$) the following are approximately true:

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$
$$\overline{X} = \frac{1}{n}T \sim \mathcal{N}(\mu, \sigma^2/n).$$

The closer the distribution of X_i is to the normal the better the approximation for small n values.

Example (PQ, p71)**.** Steel rods, made with diameter $R \sim \mathcal{N}(4.90, 0.03^2)$ (in cm), are to fit into sockets, made with diameter $S \sim \mathcal{N}(5.00, 0.04^2)$ (in cm). For a satisfactory fit the socket diameter should exceed the rod diameter, but by no more than 0.20 cm. If a rod and socket are taken at random, what is the probability that the fit is unsatisfactory?

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(i) What is the probability that the average length of 25 independent tibia lengths will be less than 7.6 mm?

Solution (i):

Because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(ii) What is the prob. that the average will differ from 7.8 by more than 0.1?

Solution (ii):

Note we can show that

$$\mathrm{P}(|L - 7.8| > 0.1) = 0.7414$$

so the average varies much less than the individual measurements.

Again, because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

(i) What is the probability that the average BP is greater than 125 mm hg?

Solution (i):

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

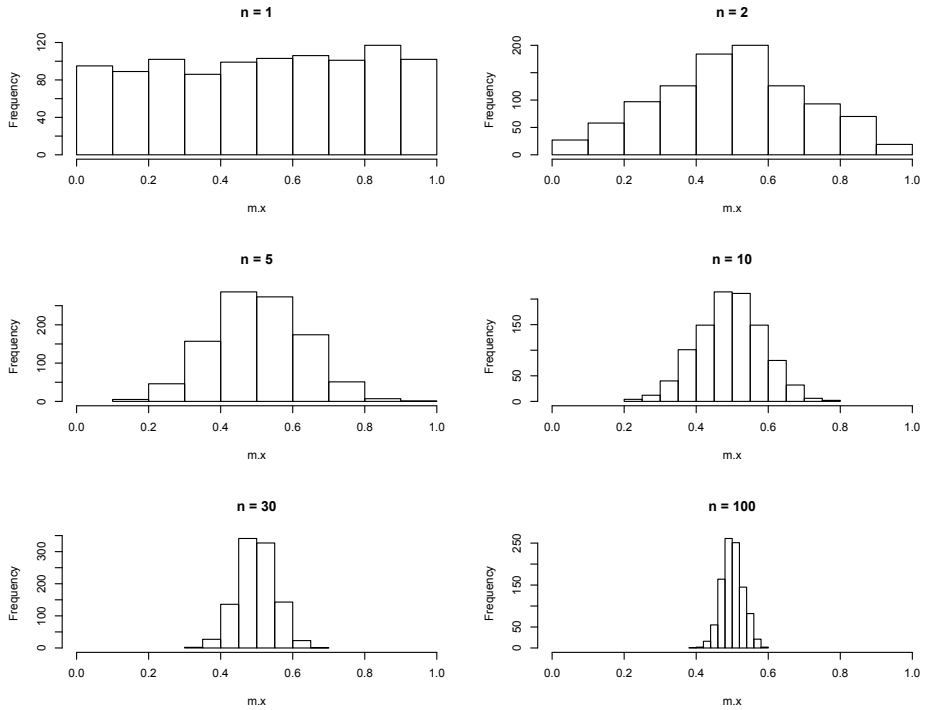
(ii) If the sample size increases to 40 how changes the answer to (i)?

Solution (ii):

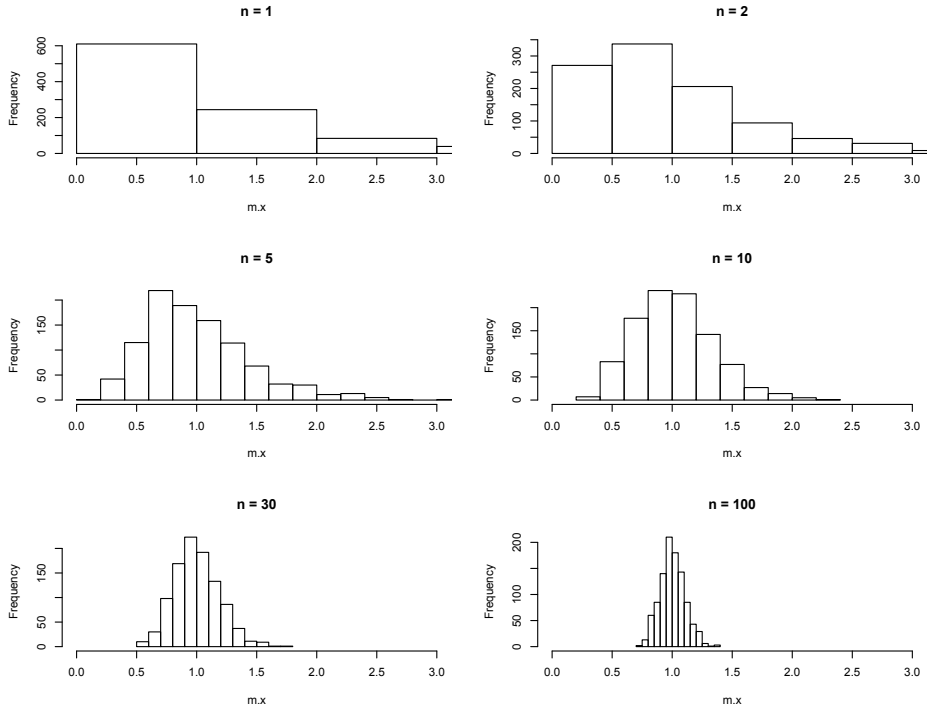
CLT in R

```
# Uniform distribution and CLT
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = runif(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,1),main="n = 1")


# Exponential distribution and CLT
par(mfrow=c(3,2))
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = rexp(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,3),main="n = 1")

# choose n in {1,2,5,10,30,100}

qqnorm(m.x)          # produces a Q-Q-plot
plot(density(m.x))   # produces a smooth estimated density
```


CLT for $X \sim \mathcal{U}$



CLT for $X \sim \mathcal{E}$



More on functions of random variables

In many applications interest focuses on some function $g(X)$ of the random variable X. E.g. change scale from meters to millimeters, logarithm of daily exchange rate changes or squared body height (BMI). In STAT2911 you will learn more. In the following I show some results that you are likely to understand with what you already know.

Theorem. (A simple version of the transformation theorem for densities) Let the random variable X have probability density f_X and let $g : \mathbb{R} \to \mathbb{R}$ be some monotone function and $h = g^{-1}$ be the inverse of g with

$$\frac{\partial h(y)}{\partial y} = h'(y).$$

Then, the density function of $Y = g(X)$ is given by $f_Y(y) = f_X(h(y)) \cdot |h'(y)| \cdot 1_{g(\mathbb{R})}(y)$.

Proof. From the definition of the probability density of Y and by applying the chain rule we get for a non-decreasing function g, that

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) = \frac{\partial}{\partial y} \mathrm{P}(Y \le y) \\ &= \frac{\partial}{\partial y} \mathrm{P}(g(X) \le y) = \frac{\partial}{\partial y} \mathrm{P}(X \le h(y)) \\ &= \frac{\partial}{\partial y} F_X(h(y)) = f_X(h(y)) h'(y) \\ &= f_X(h(y)) \cdot |h'|. \end{aligned}$$

For a non-increasing function $g(\cdot)$ the proof is essentially the same. □

Example. Let $X \sim \mathcal{U}(0,1)$ and $Y = g(X) = X^c, \quad c > 0$. The inverse of g equals $h(y) = y^{\frac{1}{c}}$, its derivative is $\frac{\partial h(y)}{\partial y} = \frac{1}{c} \cdot y^{\frac{1}{c}-1}$. From the transformation theorem it follows that

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)| = 1 \cdot \frac{1}{c} \cdot y^{\frac{1}{c}-1} \cdot 1_{(0,1)}(y).$$

Monday, 10th September 2012

Lecture 13 - Content

- Normal approximation to the Binomial
- Sampling distributions

References from Phipps & Quine

- Section 3.1 pages 72-73.
- Section 3.3 pages 75-78.

Normal approximation to the Binomial

Let X_i be independent random variables (outcomes of Bernoulli trials), defined as

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a S,} \\ 0 & \text{if the } i\text{th trial is a F,} \end{cases}$$

and let $p = \mathrm{P}(\mathsf{S})$ on the ith trial.

Theorem 24. Let $X = X_1 + \ldots + X_n \sim \mathcal{B}(n, p)$ with $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = n\,\mathrm{Var}(X_1) = np(1-p)$. Then, X is approximately $\mathcal{N}(np, np(1-p))$.

Proof. Postponed to second year... □

- The approximation is quite good if $np \geq 5$ and $n(1-p) \geq 5$!
- The closer p is to 0.5 the better the approximation for small n.

Example. $(X \sim \mathcal{B}(12, 0.5))$

$$\mathrm{P}(X = 3) = \binom{12}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^9 = \frac{12 \times 11 \times 10}{1 \times 2 \times 3} \cdot \frac{1}{2^{12}} = 0.0537.$$

Comparing to the area under the approximating normal curve, e.g.

$$X \simeq Y \sim \mathcal{N}(6, 3)$$

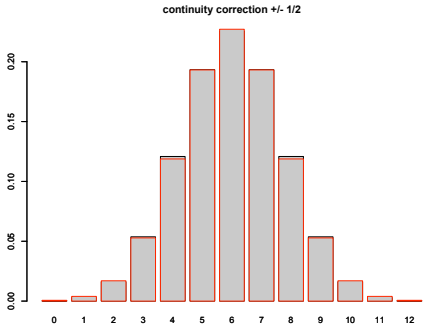
$$\mathrm{P}(X = 3) \simeq \mathrm{P}\left(3 - \lambda < Y < 3 + (1 - \lambda)\right); \quad \lambda \in [0, 1].$$

Most authors choose $\lambda = 1/2$ which in the above example is clearly closer to the true value of $\mathrm{P}(X = 3)$:

```
> mu=6; sd=sqrt(3);
> pnorm(4.0,mu,sd) - pnorm(3.0,mu,sd)
[1] 0.08247428
> pnorm(3.5,mu,sd) - pnorm(2.5,mu,sd)
[1] 0.05280327
> pnorm(3.0,mu,sd) - pnorm(2.0,mu,sd)
[1] 0.03117159
```


Example (cont)

- Note that `pnorm(3.5176,mu,sd) - pnorm(2.5176,mu,sd)` comes very close to `dbinom(3,12,0.5)` but is only optimal for this particular example.
- Overall performance of $\lambda = 1/2$ is best.



Continuity correction

- To approximate binomial probabilities using the normal consider areas of corresponding rectangles.
- Adjust the normal probability statement by adding or subtracting 0.5 to the constant to increase the area under the normal curve.

$$\begin{aligned} \mathrm{P}(X = x) &\simeq \mathrm{P}(x - 0.5 < Y < x + 0.5) \\ &= \mathrm{P}\left(\underbrace{\frac{x - 0.5 - \mu}{\sigma}}_{z_l} < Z < \underbrace{\frac{x + 0.5 - \mu}{\sigma}}_{z_u} \right) \\ &= \Phi(z_u) - \Phi(z_l). \end{aligned}$$

- For $\mathrm{P}(X \geq x)$ repeat the above step by noting:

$$\mathrm{P}(X \geq x) = \sum_{i \geq x} \mathrm{P}(X = i).$$

Example. If $X \sim \mathcal{B}(12, 0.5)$ find $\mathrm{P}(2 \leq X < 5)$.

```
> pnorm((4.5-6)/sqrt(3)) - pnorm((1.5-6)/sqrt(3))
[1] 0.1885507
> sum(dbinom(2:4,12,0.5))
[1] 0.1906738
```

Example (PQ, p80 Q21)**.** It is known that 80% of patients with a certain disease can be cured with a certain drug. What is the probability that amongst 150 patients with the disease, at most 37 of them cannot be cured with the drug.

Example. The proportion of children having a particular type of birth defect born to Pima Indian women is 0.05. Calculate the probability that in 785 independent births no more than 21 children have the birth defect.

Sampling distributions

- How do statistics vary across samples?
- Height for randomly selected $n = 4$ adult males.
- What is the distribution of $\overline{X}$ and S^2?

Model: Assume 4 independent readings of
Observations: X_1, X_2, X_3, X_4

$$\Rightarrow \overline{X} = \frac{1}{4}\sum_{i=1}^{4} X_i$$

The mean: because $\mathrm{E}\, X_i = 178$ and $\mathrm{Var}\, X_i = 8^2$ it follows $\overline{X} \sim \mathcal{N}(178, 4^2)$.
The sample variance: but $s^2 \not\sim \mathcal{N}$!

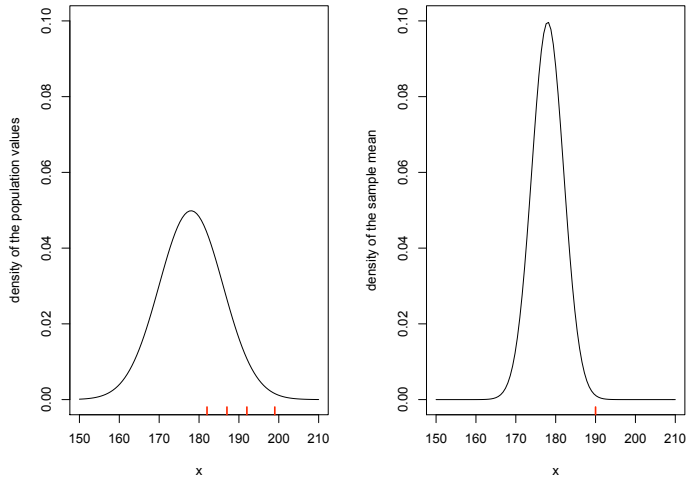
$$S^2 = \frac{1}{4-1}\Big((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + (X_4 - \overline{X})^2\Big).$$

Sampling distribution for S^2 and non-normal models

- Use CLT for large n and non-normal models.
- Knowing the sampling distribution helps identify unusual statistic values.
- E.g. if $\overline{X}$ was 190 (four basketball players):

```
# sampling distribution and extreme observations
x = c(182,187,192,199);
x.m = mean(x)
dnorm2 = function(x){
return(dnorm(x,mean=178,sd=8))}
dnorm3 = function(x){
return(dnorm(x,mean=178,sd=4))}
par(mfrow=c(1,2))
curve(dnorm2,from=150,to=210,ylim=c(0,0.1),ylab="density of the population values")
rug(x,col=2,lwd=2)
curve(dnorm3,from=150,to=210,ylim=c(0,0.1),ylab="density of the sample mean")
rug(x.m,col=2,lwd=2)
```

Distribution of population and mean



There must be something special with those 4 observations!

Sampling distributions – Movie 1

(Loading bootstrap.mp4)

Sampling distributions – Movie 2

(Loading bootstrap.mp4)

Semester 2, 2012 (Last adjustments: September 10, 2012)

Lecture Notes

MATH1905 Statistics (Advanced)

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod (at) sydney.edu.au

Semester 1, 2012 (Last adjustments: September 10, 2012)

Monday, 10th September 2012

Lecture 1 - Content

- ☐ **Joint distributions**
- ☐ **Independent random variables**
- ☐ **Central limit theorem**

References from Phipps & Quine

- ☐ Section 2.4 pages 69-72.
- ☐ Section 3.2 pages 73-75.

Standard Normal Distribution

Let $Z \sim N(0,1)$ then

- The probability density function at z is given by

  ```
  > dnorm(z)
  ```

- The (cumulative) distribution function at z, $\Phi(z) = \mathrm{P}(Z < z)$, is given by

  ```
  > pnorm(z)
  ```

- The inverse (cumulative) distribution function at t, $\Phi^{-1}(t)$ or the value of z such that $\Phi(z) = t$, is given by

  ```
  > qnorm(t)
  ```

- To generate n random values from $Z \sim N(0,1)$ we use

  ```
  > rnorm(n)
  ```


Normal Distribution

Let $X \sim N(\mu, \sigma^2)$ then

- The probability density function at x is given by

  ```
  > dnorm(x,mu,sigma)
  ```

- The (cumulative) distribution function at x, $\Phi((x-\mu)/\sigma) = \mathrm{P}(X < x)$, is given by

  ```
  > pnorm(x,mu,sigma) # OR
  > pnorm( (x-mu)/sigma )
  ```

- The inverse (cumulative) distribution function at t, the value of x such that $\mathrm{P}(X < x) = t$, is given by

  ```
  > qnorm(t,mu,sigma)
  ```

- To generate n random values from $X \sim N(\mu, \sigma^2)$ we use

  ```
  > rnorm(n,mu,sigma)
  ```

Joint distributions

Independence of random variables

Let X be a real-valued random variable (e.g. normal, exponential, binomial) and $x \in \mathbb{R}$ any number, then

$$A = \{X \leq x\}$$

represents an event. Let Y be another real-valued random variable and

$$B = \{Y \leq y\}, \quad y \in \mathbb{R}.$$

Recall the definition of independence of events: A and B are independent iff

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

which is a special case of the general multiplication rule,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) = \mathrm{P}(B)\,\mathrm{P}(A|B) \quad \text{if} \quad \mathrm{P}(A), \mathrm{P}(B) \neq 0.$$

Definition 1. Two random variables X and Y are independent if and only if for any numbers x and y the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events.

Example.

- ($X =$ 'height', $Y =$ 'weight') from a random person are not independent.
- $X_1 =$ 'lottery numbers next draw' and $X_2 =$ 'lottery numbers in three weeks time' are
- $X_1 =$ 'todays rainfall' and $X_2 =$ 'tomorrows rainfall' are

From the above Definition 1 we easily get the joint cumulative distribution function and joint probability density function of independent random variables.

Joint distribution functions and densities

Definition 2. The joint cumulative distribution function of two random variables X and Y is

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y)$$

and the joint density function is denoted $f_{X,Y}(x, y)$.

Note that, if X and Y are continuous random variables, then $F_{X,Y}(x, y)$ and $f_{X,Y}(x, y)$ are related via

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

and

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) dt ds.$$

Marginal distribution functions and densities

Definition 3. If $F_{X,Y}(x, y)$ is the joint cumulative distribution function of two random variables X and Y then, $F_X(x)$ and $F_Y(y)$ are called the marginal cumulative distribution functions of X and Y, respectively.

For integer valued random variables the marginal probability mass functions can be calculated via

$$P(X = x) = \sum_y P(X = x, Y = y) \qquad \text{and} \qquad P(Y = y) = \sum_x P(X = x, Y = y)$$

while for continuous random variables the marginal density functions can be calculated via

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx.$$

From these the marginal cumulative distribution functions can be calculated in the usual way.

Expectations of Joint Distributions

Let $g(x, y)$ be a bivariate function and let X and Y be random variables with joint density function $f_{X,Y}(x, y)$.

If X and Y are discrete random variables then

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) f_{X,Y}(x, y).$$

If X and Y are continuous random variables then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)dxdy.$$

Independence

Definition 4. Let $F_X(x) = \mathrm{P}(X \leq x)$ and $F_Y(y) = \mathrm{P}(Y \leq y)$ be the cumulative distribution functions of the independent random variables X and Y then, the joint cumulative distribution function is

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x)\,\mathrm{P}(Y \leq y) = F_X(x)F_Y(y).$$

Definition 5. Let $f_X(x)$ and $f_Y(y)$ be the probability density functions of the independent random variables X and Y then, the joint probability density function is given by

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Independent random variables: rules for expectations and variances

Theorem 1 (Properties of E and Var)**.** Let X and Y be random variables then

1. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
2. if X and Y are independent then, $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$
3. if X and Y are independent then, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Note that for any two, not necessarily independent, random variables

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$$

where

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))].$$

Proof of 1. discrete case only:

Central limit theorem

Many observed phenomena can be modelled as the sum of several random variables:

- total weight of passengers in a lift,
- total of available funds

or means of random variables

- average class mark,
- average height and weight,
- average temperature in Sydney.

The central limit theorem is useful in these types of situations.

Some useful facts about the normal distribution

Theorem 2. Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, let X and Y be independent and let a and b be two real numbers. Then

$$Z = aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2).$$

Proof: Not in MATH 1905.
In general, let $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ be independent and a_i be real numbers for $1 \leq i \leq n$ then

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, a_i^2\sigma_i^2\right).$$

Example

Example (Mean and variance of the sample mean $\overline{X}$)**.** Let the n random variables $X_1, X_2, \ldots, X_n$ be pairwise independent and each have the same distribution with mean μ and variance σ^2. Then for the sample mean, that is $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have,

i) mean: $\mu_{\overline{X}} = E(\overline{X}) = \mu$

ii) variance: $\sigma_{\overline{X}}^2 = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$

Sums of normal random variables

Theorem 3. If all $X_i \sim \mathcal{N}(\mu, \sigma^2)$ then,

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

[This is only true for normal rvs; in STAT2911 moment generating functions are introduced that make a simple proof available].

Example. X_1, X_2, X_3 are independent random variables with

i	0	1	3
p_i	$1/3$	$1/3$	$1/3$

$\overset{T_2=X_1+X_2}{\Rightarrow}$

i	0	1	2	3	4	6
p_i	$1/9$	$2/9$	$1/9$	$2/9$	$2/9$	$1/9$

i	0	1	2	3	4	5	6	7	9
p_i	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{4}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$

(Note, the distribution of T_3 clusters around the mean $\mathrm{E}\,T_3 = 4$.)

Theorem 4 (CLT, central limit theorem)**.** If $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance $0 < \sigma^2 < \infty$ then,

$$\mathrm{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \le x\right) \to \Phi(x) = \mathrm{P}(Z \le x) \quad \text{as} \quad n \to \infty.$$

Proof. Postponed to second year... □

Thus for n large (here $n \ge 25$) the following are approximately true:

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$
$$\overline{X} = \frac{1}{n}T \sim \mathcal{N}(\mu, \sigma^2/n).$$

The closer the distribution of X_i is to the normal the better the approximation for small n values.

Example (PQ, p71)**.** Steel rods, made with diameter $R \sim \mathcal{N}(4.90, 0.03^2)$ (in cm), are to fit into sockets, made with diameter $S \sim \mathcal{N}(5.00, 0.04^2)$ (in cm). For a satisfactory fit the socket diameter should exceed the rod diameter, but by no more than 0.20 cm. If a rod and socket are taken at random, what is the probability that the fit is unsatisfactory?

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(i) What is the probability that the average length of 25 independent tibia lengths will be less than 7.6 mm?

Solution (i):

Because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(ii) What is the prob. that the average will differ from 7.8 by more than 0.1?

Solution (ii):

Note we can show that

$$\mathrm{P}(|L - 7.8| > 0.1) = 0.7414$$

so the average varies much less than the individual measurements.

Again, because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

(i) What is the probability that the average BP is greater than 125 mm hg?

Solution (i):

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

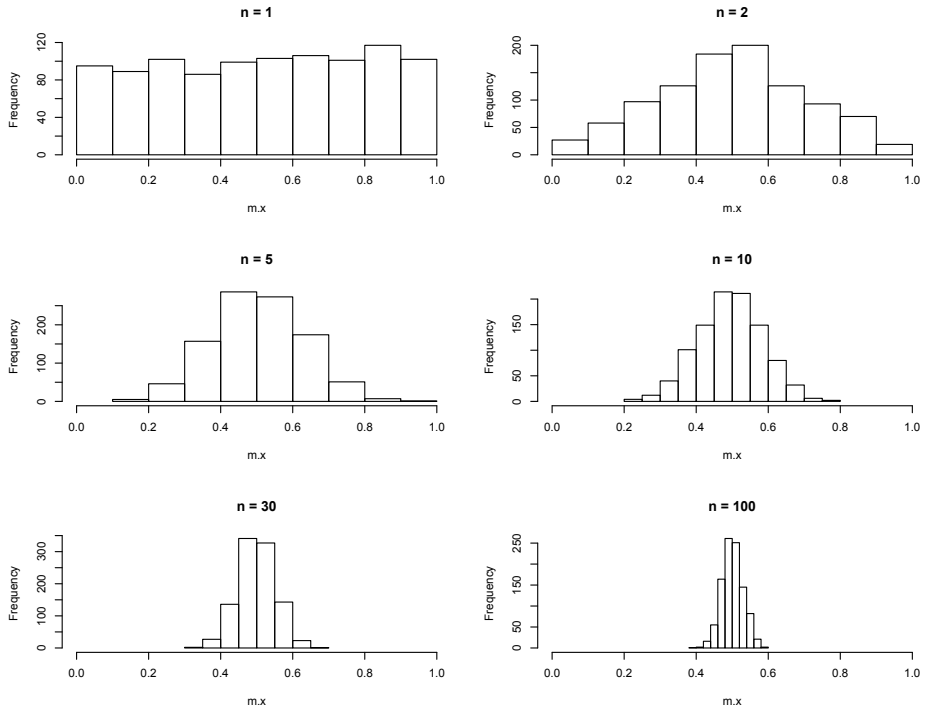
(ii) If the sample size increases to 40 how changes the answer to (i)?

Solution (ii):

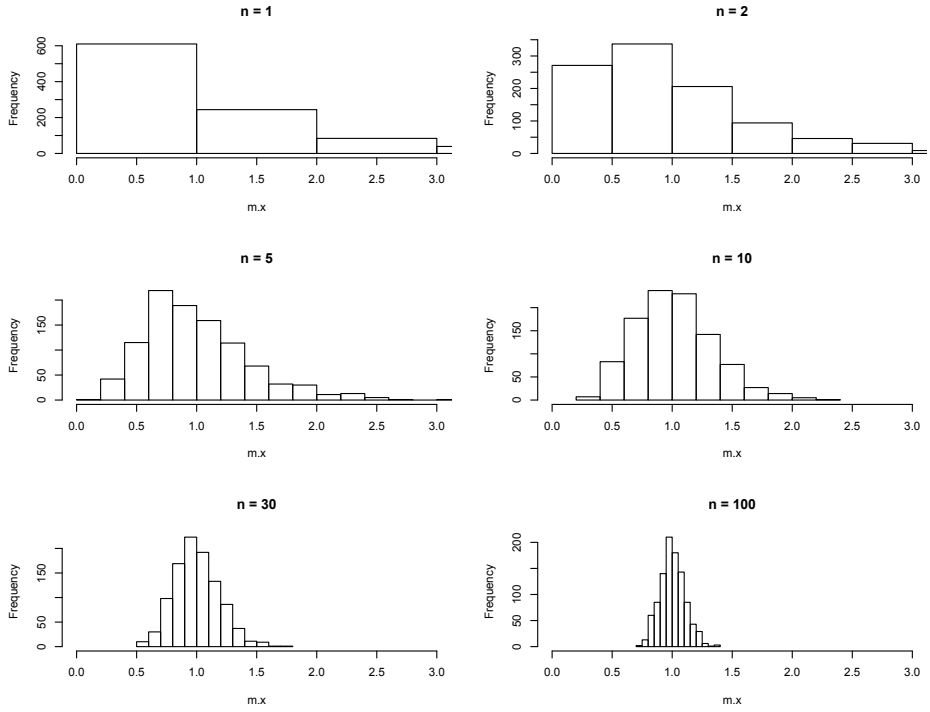
CLT in R

```
# Uniform distribution and CLT
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = runif(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,1),main="n = 1")


# Exponential distribution and CLT
par(mfrow=c(3,2))
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = rexp(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,3),main="n = 1")

# choose n in {1,2,5,10,30,100}

qqnorm(m.x)          # produces a Q-Q-plot
plot(density(m.x))   # produces a smooth estimated density
```


CLT for $X \sim \mathcal{U}$



CLT for $X \sim \mathcal{E}$



More on functions of random variables

In many applications interest focuses on some function $g(X)$ of the random variable X. E.g. change scale from meters to millimeters, logarithm of daily exchange rate changes or squared body height (BMI). In STAT2911 you will learn more. In the following I show some results that you are likely to understand with what you already know.

Theorem. (A simple version of the transformation theorem for densities) Let the random variable X have probability density f_X and let $g : \mathbb{R} \to \mathbb{R}$ be some monotone function and $h = g^{-1}$ be the inverse of g with

$$\frac{\partial h(y)}{\partial y} = h'(y).$$

Then, the density function of $Y = g(X)$ is given by $f_Y(y) = f_X(h(y)) \cdot |h'(y)| \cdot 1_{g(\mathbb{R})}(y)$.

Proof. From the definition of the probability density of Y and by applying the chain rule we get for a non-decreasing function g, that

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) = \frac{\partial}{\partial y} \mathrm{P}(Y \le y) \\ &= \frac{\partial}{\partial y} \mathrm{P}(g(X) \le y) = \frac{\partial}{\partial y} \mathrm{P}(X \le h(y)) \\ &= \frac{\partial}{\partial y} F_X(h(y)) = f_X(h(y)) h'(y) \\ &= f_X(h(y)) \cdot |h'|. \end{aligned}$$

For a non-increasing function $g(\cdot)$ the proof is essentially the same. □

Example. Let $X \sim \mathcal{U}(0,1)$ and $Y = g(X) = X^c, \quad c > 0$. The inverse of g equals $h(y) = y^{\frac{1}{c}}$, its derivative is $\frac{\partial h(y)}{\partial y} = \frac{1}{c} \cdot y^{\frac{1}{c}-1}$. From the transformation theorem it follows that

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)| = 1 \cdot \frac{1}{c} \cdot y^{\frac{1}{c}-1} \cdot 1_{(0,1)}(y).$$

Lecture 2 - Content

- Normal approximation to the Binomial
- Sampling distributions

References from Phipps & Quine

- Section 3.1 pages 72-73.
- Section 3.3 pages 75-78.

Normal approximation to the Binomial

Let X_i be independent random variables (outcomes of Bernoulli trials), defined as

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a S,} \\ 0 & \text{if the } i\text{th trial is a F,} \end{cases}$$

and let $p = \mathrm{P}(\mathsf{S})$ on the ith trial.

Theorem 5. Let $X = X_1 + \ldots + X_n \sim \mathcal{B}(n, p)$ with $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = n\,\mathrm{Var}(X_1) = np(1-p)$. Then, X is approximately $\mathcal{N}(np, np(1-p))$.

Proof. Postponed to second year... □

- The approximation is quite good if $np \geq 5$ and $n(1-p) \geq 5$!
- The closer p is to 0.5 the better the approximation for small n.

Example. $(X \sim \mathcal{B}(12, 0.5))$

$$\mathrm{P}(X = 3) = \binom{12}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^9 = \frac{12 \times 11 \times 10}{1 \times 2 \times 3} \cdot \frac{1}{2^{12}} = 0.0537.$$

Comparing to the area under the approximating normal curve, e.g.

$$X \simeq Y \sim \mathcal{N}(6, 3)$$

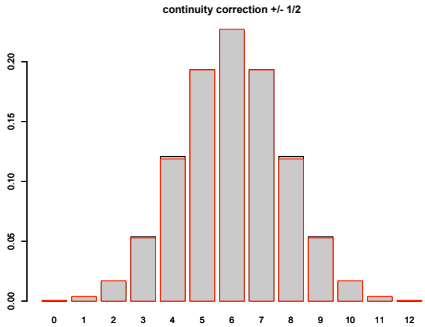
$$\mathrm{P}(X = 3) \simeq \mathrm{P}\left(3 - \lambda < Y < 3 + (1 - \lambda)\right); \quad \lambda \in [0, 1].$$

Most authors choose $\lambda = 1/2$ which in the above example is clearly closer to the true value of $\mathrm{P}(X = 3)$:

```
> mu=6; sd=sqrt(3);
> pnorm(4.0,mu,sd) - pnorm(3.0,mu,sd)
[1] 0.08247428
> pnorm(3.5,mu,sd) - pnorm(2.5,mu,sd)
[1] 0.05280327
> pnorm(3.0,mu,sd) - pnorm(2.0,mu,sd)
[1] 0.03117159
```


Example (cont)

- Note that `pnorm(3.5176,mu,sd) - pnorm(2.5176,mu,sd)` comes very close to `dbinom(3,12,0.5)` but is only optimal for this particular example.
- Overall performance of $\lambda = 1/2$ is best.



Continuity correction

- To approximate binomial probabilities using the normal consider areas of corresponding rectangles.
- Adjust the normal probability statement by adding or subtracting 0.5 to the constant to increase the area under the normal curve.

$$\begin{aligned} \mathrm{P}(X = x) &\simeq \mathrm{P}(x - 0.5 < Y < x + 0.5) \\ &= \mathrm{P}\left(\underbrace{\frac{x - 0.5 - \mu}{\sigma}}_{z_l} < Z < \underbrace{\frac{x + 0.5 - \mu}{\sigma}}_{z_u} \right) \\ &= \Phi(z_u) - \Phi(z_l). \end{aligned}$$

- For $\mathrm{P}(X \geq x)$ repeat the above step by noting:

$$\mathrm{P}(X \geq x) = \sum_{i \geq x} \mathrm{P}(X = i).$$

Example. If $X \sim \mathcal{B}(12, 0.5)$ find $\mathrm{P}(2 \leq X < 5)$.

```
> pnorm((4.5-6)/sqrt(3)) - pnorm((1.5-6)/sqrt(3))
[1] 0.1885507
> sum(dbinom(2:4,12,0.5))
[1] 0.1906738
```

Example (PQ, p80 Q21)**.** It is known that 80% of patients with a certain disease can be cured with a certain drug. What is the probability that amongst 150 patients with the disease, at most 37 of them cannot be cured with the drug.

Example. The proportion of children having a particular type of birth defect born to Pima Indian women is 0.05. Calculate the probability that in 785 independent births no more than 21 children have the birth defect.

Sampling distributions

- How do statistics vary across samples?
- Height for randomly selected $n = 4$ adult males.
- What is the distribution of $\overline{X}$ and S^2?

Model: Assume 4 independent readings of
Observations: X_1, X_2, X_3, X_4

$$\Rightarrow \overline{X} = \frac{1}{4}\sum_{i=1}^{4} X_i$$

The mean: because $\mathrm{E}\, X_i = 178$ and $\mathrm{Var}\, X_i = 8^2$ it follows $\overline{X} \sim \mathcal{N}(178, 4^2)$.
The sample variance: but $s^2 \not\sim \mathcal{N}$!

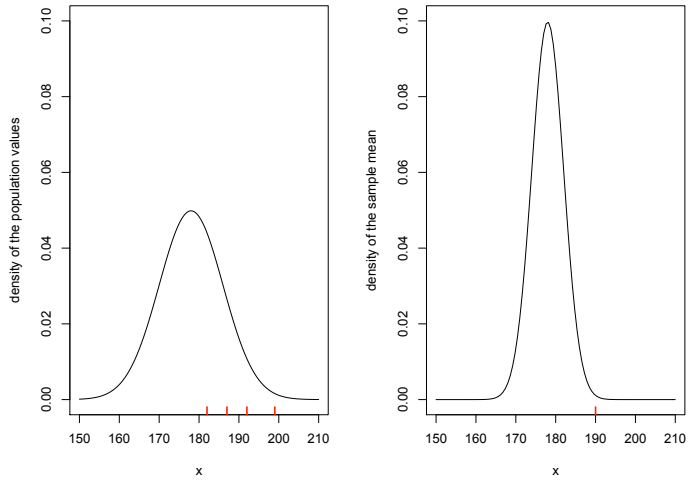
$$S^2 = \frac{1}{4-1}\Big((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + (X_4 - \overline{X})^2\Big).$$

Sampling distribution for S^2 and non-normal models

- Use CLT for large n and non-normal models.
- Knowing the sampling distribution helps identify unusual statistic values.
- E.g. if $\overline{X}$ was 190 (four basketball players):

```
# sampling distribution and extreme observations
x = c(182,187,192,199);
x.m = mean(x)
dnorm2 = function(x){
return(dnorm(x,mean=178,sd=8))}
dnorm3 = function(x){
return(dnorm(x,mean=178,sd=4))}
par(mfrow=c(1,2))
curve(dnorm2,from=150,to=210,ylim=c(0,0.1),ylab="density of the population values")
rug(x,col=2,lwd=2)
curve(dnorm3,from=150,to=210,ylim=c(0,0.1),ylab="density of the sample mean")
rug(x.m,col=2,lwd=2)
```

Distribution of population and mean



There must be something special with those 4 observations!

Sampling distributions – Movie 1

(Loading bootstrap.mp4)

Sampling distributions – Movie 2

(Loading bootstrap.mp4)

Lecture 3 - Content

- **Statistical inference**
- **Hypothesis testing**
- **One-sided tests for proportions**

Statistical inference

- Linking of observed data with possible statistical models or probability models.
- Based on some statistical model (i.e. assuming an underlying distribution, F, for observed data):
  - make decisions, e.g. in statistical hypothesis testing 'is the average measurement error equal to zero',
  - produce estimates, e.g. if the data is normal then use the mean to estimate the expected value,
  - make predictions, e.g. with time series, linear regression, and much more....

Random sample

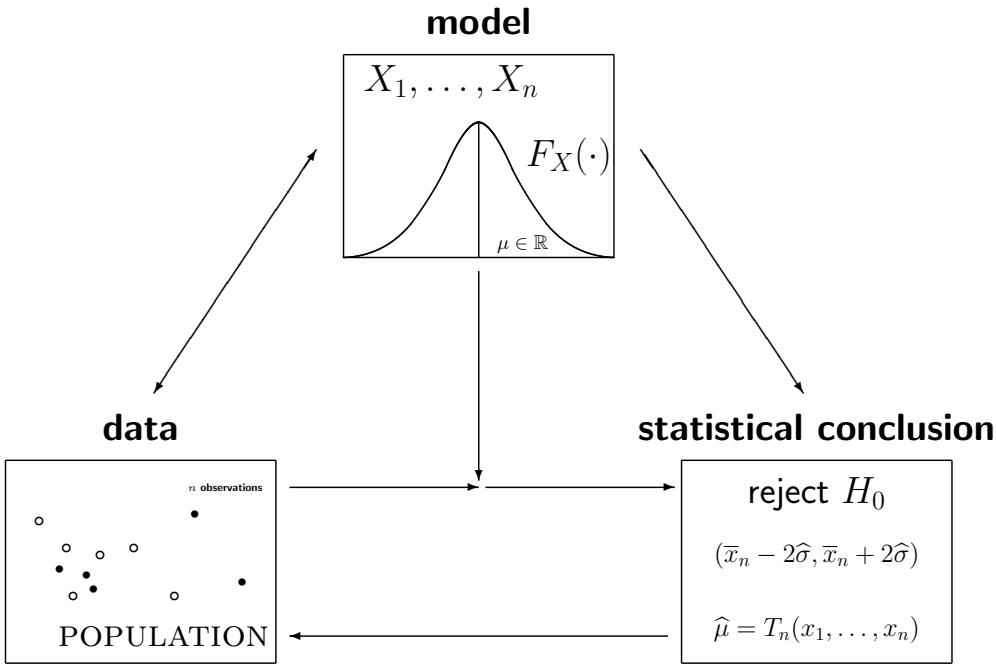
Statistical inference is inference about a population from a random sample drawn from it.

Definition 6. A set of observations (random variables) $X_1, \ldots, X_n$ constitutes a random sample of size n from the infinite population with cumulative distribution function $F(x) = \mathrm{P}(X \leq x)$ if:

- each X_i is a rv with identical CDF given by $F(x)$,
- these n random variables are independent.

Short notation: A sample $X_1, \ldots, X_n$ of length n is a set of n independent, identically distributed (iid) rvs with distribution F.

Statistical inference visualised



Three basic questions

1. Which parameter value serves based on the sample data as a *best guess* for an unknown model parameter?
   $\Rightarrow$ point estimation
2. Is there enough evidence based on the sample data to reject a pre-specified parameter value?
   $\Rightarrow$ hypothesis testing
3. Which possible parameter values of the statistical model are compatible with the sample data?
   $\Rightarrow$ interval estimation or confidence intervals

Hypothesis testing

Definition 7. A hypothesis, H, is a statement about an unknown parameter (e.g. μ) of the population.

This definition is vague by design.

Just about any kind of statement can count as a hypothesis, provided it is about a population parameter.

Hypothesis testing is the process of making a decision about a population parameter on the basis of statistics of an observed sample.

Definition 8. A null hypothesis, H_0, is a hypothesis set up to be nullified or refuted in order to support an alternative hypothesis, H_1.

In general the hypothesis test decides between two complementary hypotheses, H_0 and H_1. For example,

- H_0 may be a statement that the drug has no effect on controlling blood pressure and
- H_1 can be a statement that the drug has some effect on controlling blood pressure.

Typically H_0 is the simpler hypothesis, in the sense that it is about a parameter taking a specific value (rather than a range of values).

In hypothesis testing, one must decide either to accept H_0 as true or to reject H_0 as false and decide if H_1 is more plausible after observing the sample.

Definition 9. The critical region describes

- conditions under which H_0 should be rejected and
- conditions under which H_0 should be accepted.

General strategy:

- Find some statistic, τ (some function of our observed data).
- Find the distribution of τ assuming H_0 is true (called the null distribution).
- Calculate a corresponding P-value (defined below)
- Use the P-value to assess if data are consistent with H_0.

Definition 10. The P-value is the probability of getting an observed value of the test statistic or a more *unusual* value of the test statistic, under the assumption that H_0 is true.

Example

Most of these ideas can be illustrated by considering a coin toss example.

Let p, a parameter, be the probability of a head.

Assume the coin is 'fair so that at each toss we assume that $p = 0.5$. We call this the null hypothesis so that

$$H_0 \colon p = 0.5$$

and look for evidence against the null hypothesis H_0.

The only sensible alternatives are that:

- The coin is biased towards 'tails' in which case

$$H_1 : p < 0.5$$

- or the coin is biased towards 'heads' in which case $H_1 : p > 0.5$.

We look for evidence in favour of one of the alternatives by tossing the coin, say, 20 times and determine which of the hypotheses are most likely.

Example. Let X be the number of heads in 20 throws. Suppose we see 15 heads. Is the coin fair?

If the coin toss is fair then

$$X \sim \mathcal{B}(20, 0.5)$$

What is the chance of seeing exactly 15 heads?

$$\mathrm{P}(X = 15) = \text{dbinom(15,20,0.5)} = 0.01478577 \qquad \text{(which is small)}$$

(For continuous random variables analogous probabilities are zero, which is why we look for values of our test statistic as extreme or more extreme than what we observe).

What is the chance of seeing 15 heads or more?

$$P(X \geq 15) = \text{1 - pbinom(14,20,0.5)} = 0.02$$

which is still unlikely. Hence, H_0 is false or H_0 is true but we observed an unlikely outcome.

Example (Vaccination)**.** A flu vaccine is known to be 25% effective in the second year after inoculation. To determine if a new vaccine is more effective, 20 people are chosen at random and inoculated. If 9 of those receiving the new vaccine do not contract the virus in the second year after vaccination is the new vaccine superior to the old one?

- ☐
- ☐
- ☐
- ☐

- ☐

- ☐

Interpreting P-values

Uncertainty in the results: Because observations vary from sample to sample we can never say for sure whether H_0 is true or not.
Interpretation:

- Small P-values, for example a P-value of 0.01, means either
  - H_0 is true and the observed sample is improbable.
  - H_0 is not true.
- Large p-values, for example a P-value of 0.99 means either
  - the observed sample is consistent with H_0.
  - the observed sample comes from H_1, but by chance we are fooled into thinking the data comes from H_0.

The smaller the P-value, the stronger the evidence against H_0 in favour of H_1.

Some comments on the P-value

- If the P-value is small enough then we have evidence against H_0 in favour of the alternative hypothesis H_1.
- In the vaccination example we would conclude that the new vaccine is better.

- How small does the P-value have to be to decide in favour of H_1?
- There is no set value but

$$P\text{-value} \leq \alpha = 0.05 = 1/20$$

  is often used in practice. Other choices are: 0.1, 0.01, or 0.001 according to the 'innocent until proven guilty' principle.
- Under H_0, P-values have a uniform distribution or come very close to being uniform distributed!

Checklist for statistical tests

1. Hypotheses:
   - ☐ Null hypothesis, H_0.
   - ☐ Alternative hypothesis, H_1.
2. What is the test statistic, τ, and its sampling distribution if H_0 is true.
3. What is the critical region of the test statistic, i.e. which values of τ argue against H_0?
4. Observed test statistic (value of τ from the sample) and corresponding P-value.
5. Findings. If the P-value is small then either
   - ☐ H_0 is true and we have observed an unlikely event or
   - ☐ H_0 is false.

One-sided tests for proportions

Consider tests of

$$H_0 : p = p_0$$

against alternatives of the form

$$H_1 : p > p_0 \quad \text{or} \quad H_1 : p < p_0$$

for the distribution family $\mathcal{B}(n, p)$.

This situation occurs, say for example, when trying to determine (statistically) whether or not a coin is biased towards heads or tails.

Example

Example (Accid. Anal. and Prev. 1995:143-150)**.** A random sample of 319 front seat occupants involved in head-on collisions resulted in 95 who sustained no injuries. Does this support the claim that the proportion of uninjured occupants exceeds $1/3$?

Let $X =$ 'number of uninjured' in the sample and let

$$X \sim \mathcal{B}(319, p).$$

We wish to test $H_0 : p = 1/3$ against $H_1 : p > 1/3$.

Large values of X (our test statistic) argue for H_1.

Therefore the critical region will be the widest interval $[c_\alpha, \infty)$ such that

$$\mathrm{P}_{H_0}(X \geq c_\alpha) \leq \alpha.$$

The P-value is $\mathrm{P}(X \geq 95)$ calculated assuming H_0 is true.

Example (continued)**.**

- In R with `1-pbinom(94,319,1/3)` or

```
> binom.test(95,319,1/3,alt="greater")
Exact binomial test
data:  95 and 319
number of successes = 95, number of trials =
319, p-value = 0.9211
alternative hypothesis: true probability of success is greater than 0.3333333
95 percent confidence interval:
 0.255656 1.000000
sample estimates:
probability of success
             0.2978056
```

Example (continued).

- or using the CLT: under $H_0 : X \simeq Y \sim \mathcal{N}(np, np(1-p))$, i.e. the

$$\begin{aligned} P\text{-value} &= \mathrm{P}(X \geq 95) = 1 - \mathrm{P}(X \leq 94) \\ &\simeq 1 - \mathrm{P}\left(Z \leq \frac{94.5 - 106.33}{\sqrt{70.89}}\right) \\ &= 1 - \Phi(-1.41) = 0.92 \text{ with } \texttt{1-pnorm(-1.405454)} \end{aligned}$$

$\Rightarrow$ there exists not enough evidence to support the claim that $p > 1/3$ but there is for any $p_0 \leq 0.253$.

```
> prop.test(95,319,1/3,alt="greater")
1-sample proportions test with continuity correction
data:  95 out of 319, null probability 1/3
X-squared = 1.6556, df = 1, p-value = 0.9009
alternative hypothesis: true p is greater than 0.3333333
95 percent confidence interval: 0.2560441 1.0000000
sample estimates:    p
             0.2978056
[P-value is different because there are various ways of correcting for continuity.]
```


R code

The code demonstrates the how P-values are uniformly distributed.

```
> set.seed(1)
> B = 10000  # no simulation runs
> n = 319    # sample size
> p = 1/3    # parameter value under Ho
> tau = rbinom(B,n,p)
> pvalue = 1 - pbinom( tau - 1 , n , p )  # alternative is p > 1/3
> hist(pvalue,breaks = 10)
```

Tuesday, 18 September 2012

Lecture 4 - Content

- ☐ **Two-sided tests for proportions**
- ☐ **Sign test**

Checklist for statistical tests

1. Hypotheses:
   - ☐ Null hypothesis, H_0.
   - ☐ Alternative hypothesis, H_1.
2. What is the test statistic, τ, and its sampling distribution if H_0 is true.
3. What is the critical region of the test statistic, i.e. which values of τ argue against H_0?
4. Observed test statistic (value of τ from the sample) and corresponding P-value.
5. Findings. If the P-value is small then either
   - ☐ H_0 is true and we have observed an unlikely event or
   - ☐ H_0 is false.

Two sided tests

Previously we only looked for alternatives of the form

$$H_1 : p > p_0 \quad \text{or} \quad H_1 : p < p_0.$$

These are called one-sided tests because they only consider the parameter lying to one side of a hypothesised value, in this case p_0.

In general we may not know in advance which alternative to choose. in this case we need to consider the two-sided hypothesis

$$H_1 : p \neq p_0$$

and in some cases this may be the only feasible alternative hypothesis.

WARNING: It is a statistical no-no to choose H_1 based on observed data. Instead H_1 should be chosen to dispel some preconceived outcome or alternatively based on expert opinion.

Test for proportions

Consider the two-sided hypothesis

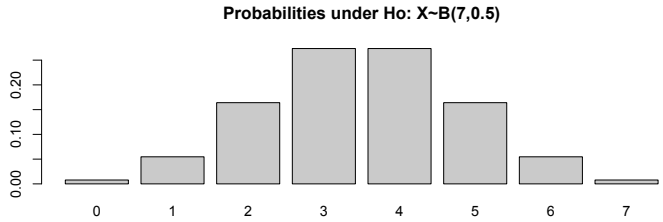
$$H_0 \colon p = p_0$$

where the general alternative is

$$H_1 \colon p \neq p_0.$$

Here we observe $X \sim \mathcal{B}(n, p)$, with $X \sim \mathcal{B}(n, p_0)$ under H_0:

$$\Rightarrow \text{large values of} \quad |X - np_0| \quad \text{argue against } H_0.$$



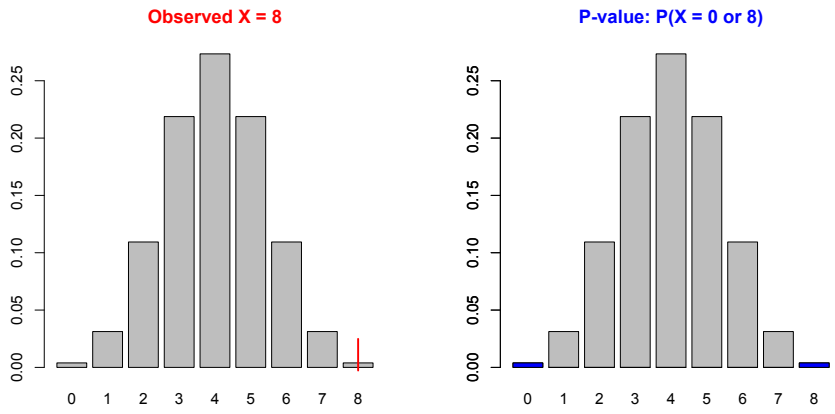
Example (Paul the octopus)**.** Is Paul the octopus guessing?

(http://en.wikipedia.org/wiki/Paul_the_octopus)

Paul correctly predicts 8 out of 8 winners in the 2010 World Cup!

- Let p denote the probability of correctly predicting the winner.
- Test: $H_0 : p = \frac{1}{2}$ against $H_1 : p \neq \frac{1}{2}$.
- Results: 8 of 8 winners in the 2010 World Cup were correctly predicted!
- Does this provide sufficient evidence against H_0?
- Test statistic: $X =$ 'no of correctly predicted winners in a sample of size n = '.
- Under H_0: $X \sim \mathcal{B}(8, 0.5)$; note $8 \times 0.5 < 5$, i.e. not yet with CLT.
- P-value: the values $X = 0$ and $X = 8$ are equally extreme or more extreme outcomes than the observed value of $X = 8$.

Example (cont).



- $\mathrm{P}(X \leq 0) + \mathrm{P}(X \geq 8) =$ `2*pbinom(0,8,0.5)` $=$ 0.0078125.
- Conclusion:

- Or much faster with `binom.test(8,8,1/2,alt="two.sided")`.

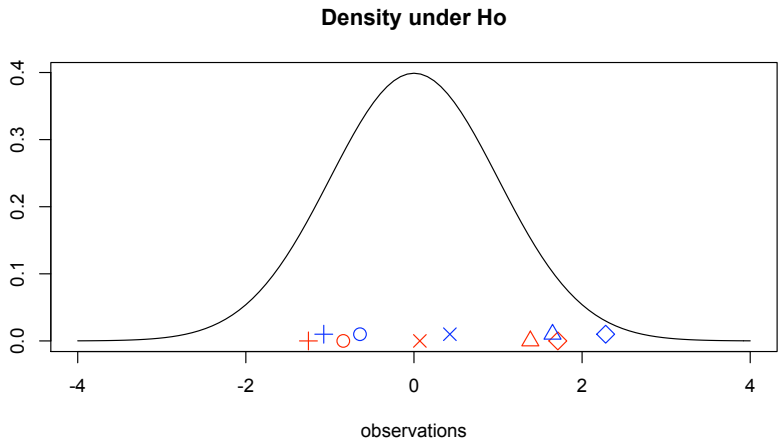
Example. A company claims that 93% of all items produced are non-defective. A random sample of 100 items is taken. If the observed number of defectives in the sample was 11 is there any reason to doubt the 93% claim?

- Test
- Under H_0
- P-value

```
> 2*(1-pnorm(1.37))
[1] 0.1706869
> prop.test(11,100,0.07,alt="two.sided")
[...edited output...]
p-value = 0.1701
95 percent confidence interval: 0.05886717 0.19223346
```


Sign test

Paired data are very common. For example before/after trials, studies on twins, left/right arm freckles count.
Are the two samples from populations with the same distribution?



Analyse differences!

Theorem 6. If X and Y are iid with distribution function F then the distribution of $D = X - Y$ is symmetric with symmetry centre 0, i.e. $\mathrm{P}(D \leq -d) = \mathrm{P}(D \geq d)$ for all $d \in \mathbb{R}$.

Proof.

□

Constructing a simple test...

- Base a test on the number of positive differences.
- Hence, use the sign of the differences and ignore their magnitude

  $\Rightarrow$ test reduces to simple test of proportions.

Note, the simple test of proportions is for data with two possible outcomes only (yes/no, S/F, etc). Thus, we will discard differences which are exactly zero.

Example (Rats). A biochemical substance is believed to have an inhibitive effect on muscular growth. Ten laboratory rats of similar types are selected. For each rat

- one hind leg was regularly injected with the biochemical substance.
- The corresponding muscle on the other hind leg was regularly injected with a harmless placebo.
- At the end of 6 months the weights of the muscles were measured (in gms) and recorded as follows:

Rat	1	2	3	4	5	6	7	8	9	10
Bioch.	1.7	2.0	1.7	1.5	1.6	2.4	2.3	2.4	2.4	2.6
Placebo	2.1	1.8	2.2	2.2	1.5	2.9	2.9	2.4	2.6	2.5

- Analyse the data to determine whether this experiment provides evidence of a significant inhibitive effect.
- Why is this a good design for the study?

Example (cont).

- □
- □
- □
- □
- □
- □
- □
- □

Example (cont).

```
> # rat example
> x = c(1.7, 2.0, 1.7, 1.5, 1.6, 2.4, 2.3, 2.4, 2.4, 2.6)
> y = c(2.1, 1.8, 2.2, 2.2, 1.5, 2.9, 2.9, 2.4, 2.6, 2.5)
> d = y-x
> d
> plot(x,y,xlim=c(1.5,3),ylim=c(1.5,3))
> abline(0,1)
> text(2.75,1.5,"negative differences")
> text(1.75,3,"positive differences")
> points(c(1.8,2),c(1.8,1.8),type="l",lty=2,col="red")
> text(1.9,1.7,"y-x = -0.2")
> s = sign(d)[sign(d)!=0]
> table(s)
> binom.test(table(s),p=0.5,alt="less")
```

Example (Paint). A paint supplier claims that a new additive will reduce the drying time of acrylic paint. To test this claim 10 panels of wood are painted: one half with the original paint formula and one half with the paint having the new additive. The drying times in hours are given below.

```
> panel  = 1:10
> npaint = c(6.4,5.8,7.4,5.5,6.3,7.8,8.6,8.2,7.0,4.9)
> rpaint = c(6.6,5.9,7.8,5.7,6.0,8.4,8.8,8.4,7.3,5.8)
> d = rpaint - npaint
> d
 [1]  0.2  0.1  0.4  0.2 -0.3  0.6  0.2  0.2  0.3  0.9
```

- Can we conclude that the new additive is effective in reducing the drying time of the paint?
- Same steps as in previous example... but P-value $= 0.0107$.

Example (cont).

- The sign test can be used to test the hypothesis that the differences are scattered around 0.
- If the differences have a distribution that is symmetric about 0 then the probability of getting a positive difference, p_+, is 0.5.
- There are 10 non-zero differences.
- Test $H_0 : p_+ = \frac{1}{2}$ against $H_1 : p_+ > \frac{1}{2}$.
- Let X denote the number of positive differences. Large values of X support H_1. There are $m = 10$ non-zero differences. Thus if H_0 is true then $X \sim \mathcal{B}(10, 0.5)$.
- We observe 9 positive differences out of the $m = 10$ non-zero ones. P-value $= P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.9893 = 0.0107$. Since P is small we conclude that the new additive is effective in reducing the drying time of the paint.

Remarks

- Note the sign test ignores a lot of the information in the sample but it can be applied in quite general situations.
- Does not depend on the distribution of the data! For this reason sometimes these types of tests are called non-parametric.
- The sign test can be used to test if a single sample is taken from a continuous distribution that is symmetric about its population mean μ.

Monday, 1 October 2012

Lecture 5 - Content

☐ No lecture due to Labour Day holiday

Tuesday, 2 October 2012

Lecture 6 - Content

- **Tests for the mean μ**
- **Z-tests**

Reminder of Binomial/Sign Tests

For binomial/sign tests we have $\tau = X \sim \mathcal{B}(n, p)$.

For some fixed and known value p_0, or null hypothesis is

$$H_0 \colon p = p_0.$$

Under the assumption of H_0 we have $\tau = X \sim \mathcal{B}(n, p_0)$. We test H_0 against one of the following alternative hypotheses (with P-values),

$$H_1 \colon \begin{cases} p < p_0 & P\text{-value} = \mathrm{P}(X \leq x) \\ p > p_0 & P\text{-value} = \mathrm{P}(X \geq x) \\ p \neq p_0 & P\text{-value} = \mathrm{P}(|X - np_0| \geq |x - np_0|) \end{cases}$$

Reminder of P-values

Reminder: under H_0 the P-value is approximately $\mathcal{U}(0,1)$.

If the P-value is less than or equal to α (usually 5%) reject H_0. State there is statistical evidence against H_0 in favour of H_1.

If the P-value is greater than α accept H_0. State there is not sufficient statistical evidence to refute H_0 or the data is consistent with H_0. (DO NOT SAY THAT H_0 IS TRUE!!!).

Tests for the mean μ

Statistical tests can be developed to test claims about the population mean.

Assumption 0: Identically Distributed Since we are drawing samples from a particular population we implicitly assume that the samples are drawn from the same population, i.e. samples are identically distributed.

Assumption 1: Independence Assume that samples drawn from the population are selected independently, i.e. draws from the population do not depend on previous selections from the population

Assumption 2: Normal Samples (Stronger than Assumption 0) The population we are interested in has a Normal distribution, $\mathcal{N}(\mu, \sigma^2)$.

Tests for the mean μ

Suppose we have independent $X_1, \ldots, X_n$ with

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

An obvious test statistic to use for making inference about the mean μ is $\tau = \overline{X}$, the sample mean.

Two scenarios

At this point it is important to distinguish between two situations

- σ is known (e.g. IQ-test)
- σ is unknown, which is in general the case.

The distribution of $\tau = \overline{X}$ depends on whether σ is known or whether σ is unknown and needs to be estimated in some way.

Assumption 3: σ is known

The Z-test is constructed under the assumption that σ is known.

If the population variance, σ^2, is known the sampling distribution of the sample average is also known based on results stated in previous lectures.

If σ is known then the distribution of $\overline{X}$ is

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

where n is the sample size.

One-sided Z-test

- Test $H_0\colon \mu = \mu_0$ against $H_1\colon \mu > \mu_0$, where μ_0 is a given value.
- If H_0 is true then $\mu = \mu_0$ and so

$$\overline{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

- Large values of $\overline{X}$ argue for H_1 (and against H_0).
- If the observed sample average is $\overline{x}$ the P-value is

$$P\text{-value} = \mathrm{P}(\overline{X} \geq \overline{x}) = \mathrm{P}\left(Z \geq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right), \quad \text{where} \quad Z \sim \mathcal{N}(0,1).$$

Definition 11. The Z-value is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

and its corresponding test is called the Z-test.

Normal distributed data and n small

Example (Birthweights). The birthweights of a random sample of $n = 14$ boys born to mothers who smoked heavily during pregnancy were recorded (in ounces). The data are:

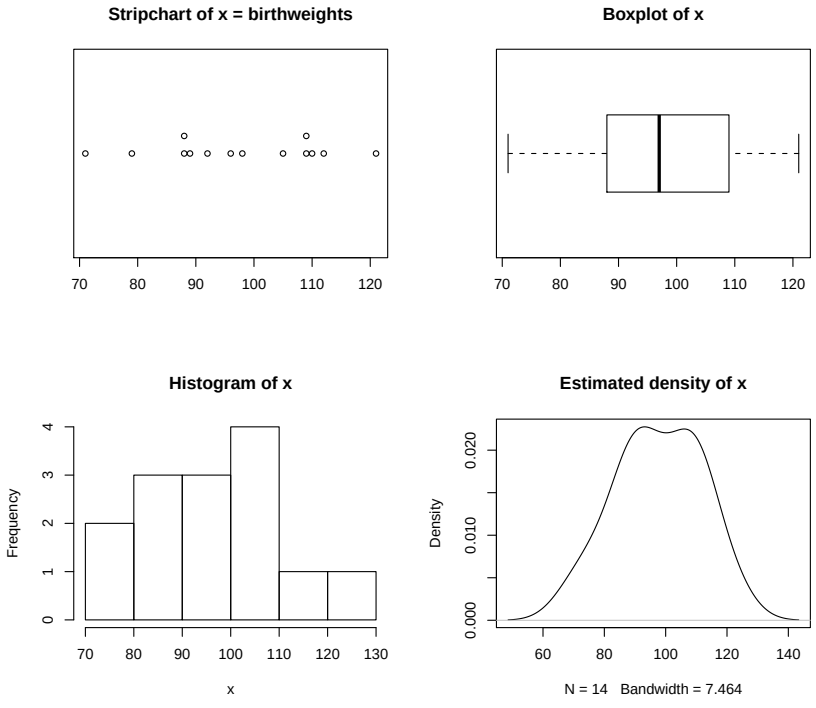
$$79,\ 92,\ 88,\ 98,\ 109,\ 109,\ 112,$$
$$88,\ 105,\ 89,\ 121,\ 71,\ 110,\ 96.$$

- It is believed that on average, boys born to mothers who smoke have a lower birthweight than the national average of 109 ounces (3.09kg).
- Is it reasonable to assume that birthweight has a normal distribution?
- Use R to explore ...

Example (cont)

```
> x = c(79,92,88,98,109,109,112,88,105,89,121,71,110,96)
> par(mfrow=c(2,2))
> stripchart(x, method="stack",offset=1, pch=1)
> title(main="Stripchart of x = birthweights")
> boxplot(x,range=1,horizontal=TRUE)
> title(main="Boxplot of x")
> hist(x)
> plot(density(x),main="Estimated density of x")
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.00   88.25   97.00   97.64  109.00  121.00
> IQR(x)
[1] 20.75
> sd(x)
[1] 14.05816
```


Example (cont)



Example (cont)

- Hence, we assume that the population of birthweights for boys born to mothers who smoke is modelled by
$$W \sim \mathcal{N}(\mu, 15^2).$$
- Test $H_0 \colon \mu = 109$ against $H_1 \colon \mu < 109$.
- The sample size is $n = 14$.
- Small values of $\overline{W}$ support H_1.
- If H_0 is true then the sampling distribution of $\overline{W}$ is
$$\overline{W} \sim \mathcal{N}\left(109, \frac{15^2}{14}\right).$$
- The observed value is $\overline{w} = \overline{x} = 97.64$ and $s = 14.05816$.

Example (cont)

-

- strong evidence against H_0.

Sample size n is large, normal or non-normal data

Example (SIDS victims)**.** In a random sample of 128 arterioles taken from SIDS (sudden infant death syndrome) victims the mean muscle thickness as a percentage of total arteriole diameter was 9.10.

- Assume that percentage muscle thickness can be modelled by

$$X \sim \mathcal{N}(\mu,\ 2.15^2).$$

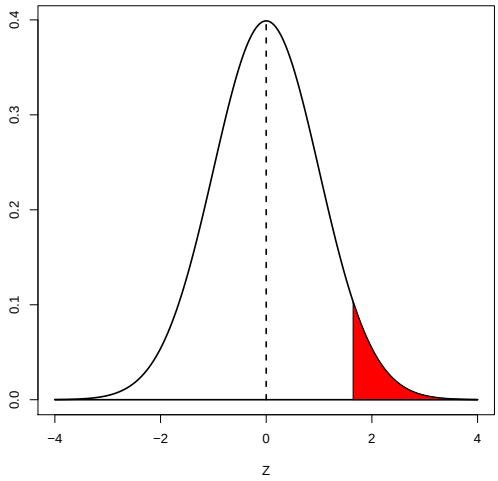
- For normal children of the same age $\mu = 6.04$.
- Is there evidence that the muscle thickness is greater in SIDS victims?

Example (cont)

- Test $H_0\colon \mu = 6.04$ against $H_1\colon \mu > 6.04$.
- Base the test on $\overline{X}$,

$$\overline{X} \sim \mathcal{N}(6.04, 2.15^2/128) \text{ if } H_0 \text{ is true.}$$

- Large values of $\overline{X}$ support H_1.



$$P\text{-value} = \mathrm{P}(\overline{X} \geq 9.10) = \mathrm{P}\left(Z \geq \frac{9.10 - 6.04}{2.15/\sqrt{128}}\right) = \mathrm{P}(Z \geq 16.10) < 10^{-4}$$

- Thus, the P-value is *very* small and so there is strong evidence against H_0.

Conclusions

- In the previous example the sample size was very large ($n = 128$).
- In such cases we know that the Central Limit Theorem (CLT) states that

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{(approx.)}$$

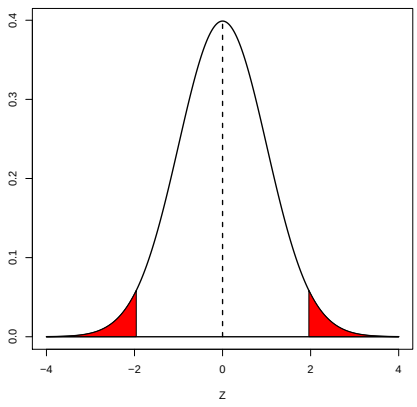
  whether the population is normal or not.
- Thus if the sample size is large then the CLT will enable us to calculate approximate P-values for tests of hypotheses about the mean regardless of the distribution of the underlying population provided σ is known.

Two-sided Z-tests

Example (Breaking strengths). A new synthetic fishing line is marketed with a manufacturer's claim that the mean breaking strength is 8 kgs with an s.d. of 0.5 kgs. Test this claim if a random sample of 50 lines is tested and the average of the sample of breaking strengths is $\overline{x} = 7.85$ kg.

- Here we have no reason to assume the true mean breaking strength is above or below 8 kgs if the claim is not true.
- Assume that the breaking strength can be modelled by $X \sim \mathcal{N}(\mu, 0.5^2)$.
- Test $H_0 \colon \mu = 8$ against $H_1 \colon \mu \neq 8$.
-
-

-



□

Conclusions from the previous three examples

- In all of the above examples we have been given the value for the population standard deviation, σ.
- In practice σ is generally unknown.
- In these cases how do we proceed?
- Recall the Z-test statistic is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- We can estimate σ by using the sample standard deviation,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Monday, 8 October 2012

Lecture 7 - Content

- **One-sample t-tests**

One sample t-test

- In all examples in the last lecture(s) we were given the value for the population standard deviation σ,
- In practice σ is generally unknown!
- Estimate σ^2 by the sample variance s^2,

$$s = \sqrt{\frac{S_{xx}}{n-1}}$$

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum x_i^2 - n(\overline{x}^2). \end{aligned}$$

Theorem 7. If $\overline{X}$ is the mean of a sample of size n taken from a normal distribution having the mean μ and the variance σ^2, then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \text{ is a random variable}$$

having the t distribution with $\nu = n - 1$ degrees of freedom.

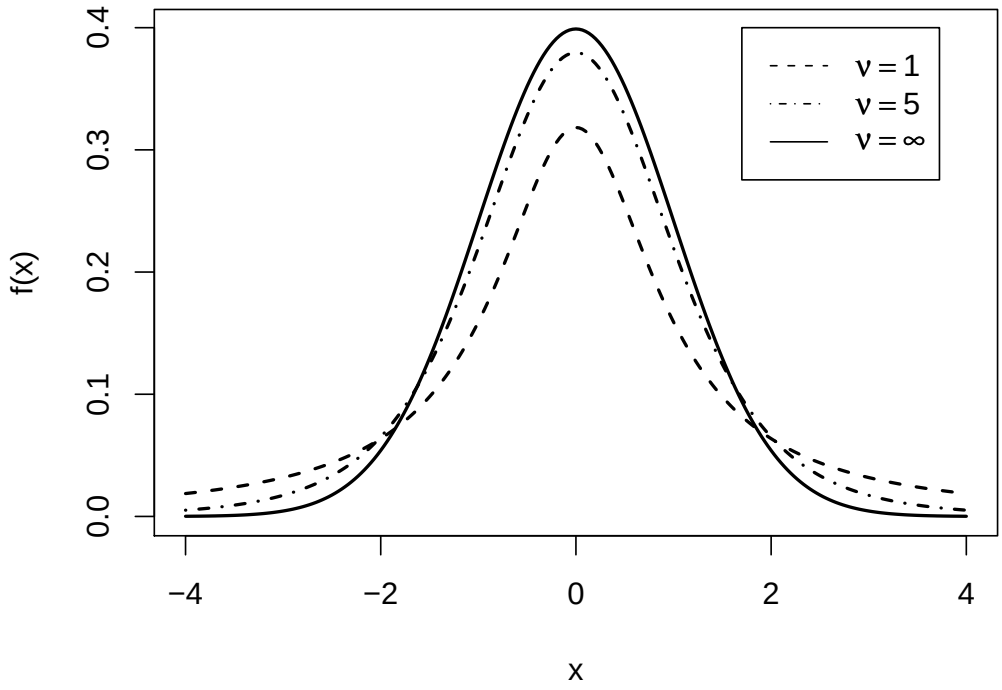
(Note that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.)

The t distribution

- The proof of the previous theorem will be shown in second year (need to show how to determine the distribution of a transformation of random variables).
- William S. Gosset (1908) (pen name: Student; statistician at Guinness)

- The density of the t distribution is symmetric and gets closer to the normal when $\nu = n - 1$ gets larger.
- Thicker tails of the t distribution takes into account the additional variability due to the estimation of σ by s.

The t distribution



The pdf of the t-distribution

Definition 12. A random variable having the t distribution with parameter $\nu = n - 1$ (degrees of freedom) has pdf (probability density function)

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

To say that the random variable T has the t distribution with $\nu \in \mathbb{N}$ df we write $T \sim t(\nu)$.

Remember: The Γ-function is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$$

and has the following properties (can be proved by partial integration):

$$\begin{aligned}\Gamma(\alpha+1) &= \alpha\Gamma(\alpha) \Rightarrow \Gamma(n+1) = n!;\ n \in \mathbb{N},\\ \Gamma(1/2) &= \sqrt{\pi}.\end{aligned}$$

Reminder of Assumptions

Assumption 0: Identically Distributed Since we are drawing samples from a particular population we implicitly assume that the samples are drawn from the same population, i.e. samples are identically distributed.

Assumption 1: Independence Assume that samples drawn from the population are selected independently, i.e. draws from the population do not depend on previous selections from the population

Assumption 2: Normal Samples (Stronger than Assumption 0) The population we are interested in has a Normal distribution, $\mathcal{N}(\mu, \sigma^2)$.

Z-tests assume that σ^2 is known.

Assumption 3: population is normal but σ^2 is unknown

- We can use a t-test when the population we are sampling from is normal but σ^2 is unknown.
- Check t-tables (formula sheet). Unlike the normal and binomial, t-tables are based on

$$\mathrm{P}(t_\nu > t) = p,$$

  where ν is the degree of freedom (row), p is the upper tail probability (column) and t is given in the body of the table.
- In R the following functions are helpful:
  - PDF: `dt(x,df=nu)`
  - CDF: `pt(q,df=nu)`
  - quantiles (critical values): `qt(p,df=nu)`
  - random numbers: `rt(n,df=nu)`

Example.

Example (Birthweights revisited)**.** Birthweights of boys to mothers who smoked:

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.00   88.25   97.00   97.64  109.00  121.00
> sd(x)
[1] 14.05816
```

- The 14 observations look like they could come from a normal distribution.
- Also, $\overline{w} = 97.643$ and $s = 14.058$.
- Test $H_0 \colon \mu = 109$ against $H_1 \colon \mu < 109$ using a t-test.
- Test statistic:

$$\tau = T = \frac{\overline{w} - 109}{s/\sqrt{14}},$$

  small values of τ support H_1.

Example (cont)

□

□

□

Example (cont)

```
> t.test(x,mu=109,alt="less")

One Sample t-test

data:  x
t = -3.0228, df = 13, p-value = 0.0049
alternative hypothesis: true mean is less than 109
95 percent confidence interval:
     -Inf 104.2966
sample estimates:
mean of x
 97.64286
```

One-sample t-tests continued

- Given a sample $X_1, \ldots, X_n$ from populations $\mathcal{N}(\mu, \sigma^2)$.
- Test $H_0 : \mu = \mu_0$ based on the test statistic

$$\tau = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}; \quad \text{where } s^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2 \text{ estimates } \sigma^2.$$

- If H_0 is true then,

$$\tau \sim t_\nu, \quad \text{where } \nu = \text{degrees of freedom.}$$

Example (Lubricants). The contents (in litres) of a random sample of 9 containers of lubricant are given:

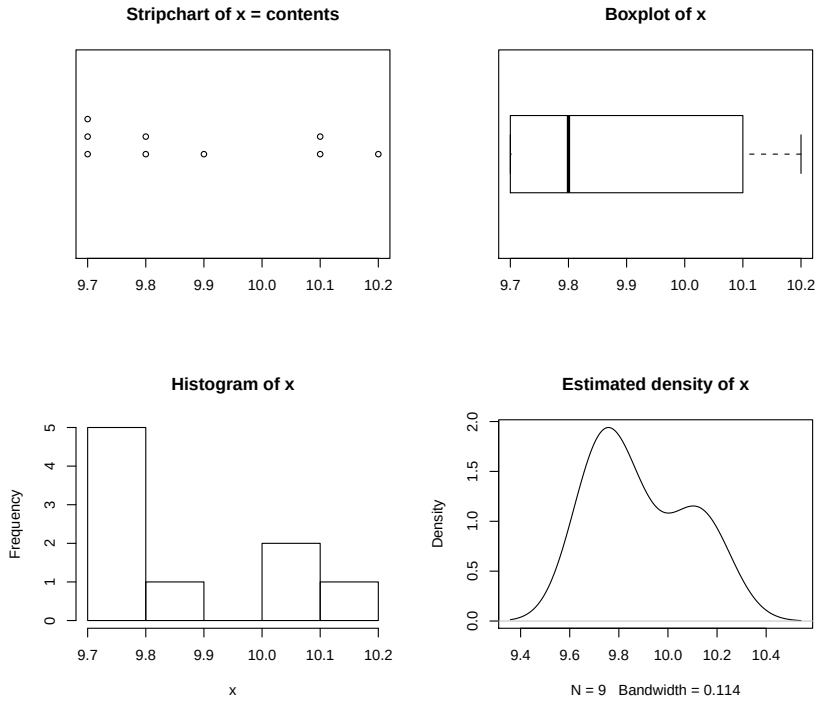
$$10.2,\ 9.7,\ 10.1,\ 9.7,\ 10.1,\ 9.8,\ 9.9,\ 9.8,\ 9.7.$$

Use these data to test the hypothesis that the (population) average content is 10 litres against the alternative that the true average contents is less than 10 litres.

- Can you assume that the contents $X \sim \mathcal{N}(\mu, \sigma^2)$? With R:

```
x = c(10.2,9.7,10.1,9.7,10.1,9.8,9.9,9.8,9.7)
stripchart(x, method="stack",offset=1, pch=1)
boxplot(x,range=1,horizontal=TRUE)
hist(x)
plot(density(x),main="Estimated density of x")
t.test(x,mu=10,alternative ="less")
```

Example (cont)



Example (cont)

- The sample average is $\overline{x} = 89/9 = 9.8889$.
- The sample standard deviation with R or by hand is

```
> sd(x) [...]    [1] 0.1964971
```

Tuesday, 9 October 2012

Lecture 8 - Content

- **One-sample t-tests continued**
- **Paired t-tests**

References from Phipps & Quine

- Section 3 pages 96–100.

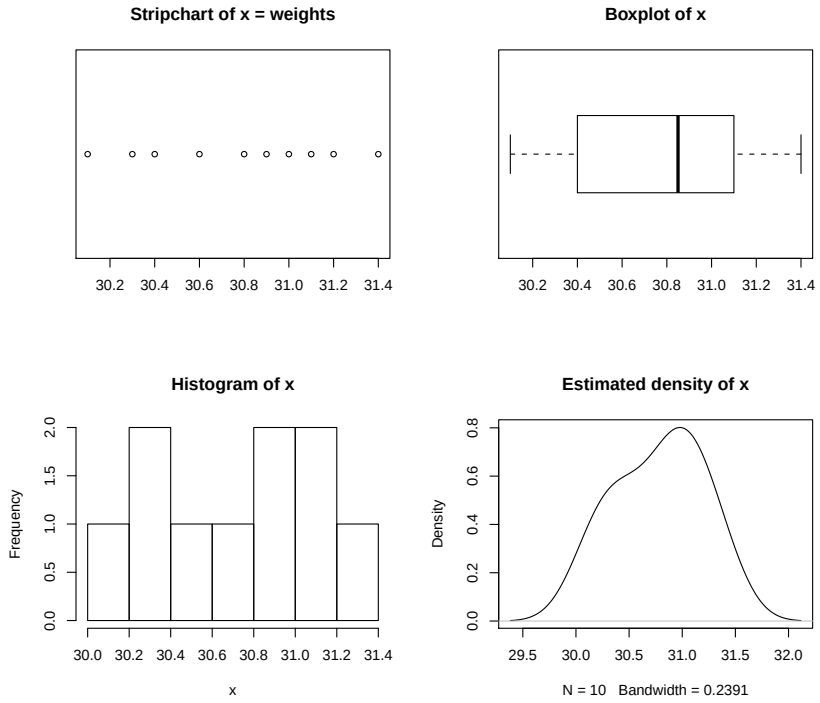
Example (Tablets). Ten tablets are weighed giving the weights (in mgs):

```
> x= c(31.0,31.4,30.4,30.1,30.6,31.1,31.2,30.9,30.3,30.8)
```

The machine producing these is set to give a mean weight of 30 mg. Is there evidence that the setting is incorrect?
Assume the weights are normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.10   30.45   30.85   30.78   31.08   31.40
> sd(x)
[1] 0.4211096
> stripchart(x, method="stack",offset=1, pch=1)
> boxplot(x,range=1,horizontal=TRUE)
> hist(x)
> plot(density(x))
```

Example (cont)



Example (cont)

- Sample size is small ($n = 10 < 25$), exploratory data analysis suggests normality may be reasonable (difficult to test for small sample sizes).
- The sample average is $\overline{x} = 30.78$.
- The sample standard deviation is $s = 0.4211096$.
- We wish to test,

$$H_0 \colon \mu = 30 \text{ against } H_1 \colon \mu \neq 30.$$

- Because sample size is very small, base the test on

$$\tau = \frac{\overline{X} - 30}{S/\sqrt{n}}.$$

- Either small values or large values of τ support H_1.

Example (cont)

- Under the assumption that the null hypothesis is true (along with independence and normality) the null distribution of the test statistic is $t_{n-1} = t_9$.

Example (cont)

- Alternatively, using the R commands,

```
> 2*pt(-5.857327,9)
[1] 0.000241544
> 2*pt(5.857327,9,lower.tail=F)
[1] 0.000241544
```

we get the exact P-value of 0.000241544.

Example (cont)

```
> t.test(x,mu=30,alternative ="two.sided")

One Sample t-test

data:  x
t = 5.8573, df = 9, p-value = 0.0002415
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 30.47876 31.08124
sample estimates:
mean of x
    30.78
```

Thus, there is strong evidence against H_0.

Summary of Z-tests and t-tests

- Assume the samples are independent and normally distributed.
- For some fixed and known value μ_0 the null hypothesis is $H_0\colon \mu = \mu_0$.
- If σ is unknown then $\tau = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$ and

$$H_1\colon \begin{cases} \mu < \mu_0 & P\text{-value} = \mathrm{P}(t_{n-1} \leq \tau) \\ \mu > \mu_0 & P\text{-value} = \mathrm{P}(t_{n-1} \geq \tau) \\ \mu \neq \mu_0 & P\text{-value} = 2\,\mathrm{P}(t_{n-1} \geq \tau) \end{cases}$$

- If σ is known then $\tau = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ or if σ is unknown and n is large ($n > 25$ so that the CLT applies) then $\tau = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim N(0,1)$ and

$$H_1\colon \begin{cases} \mu < \mu_0 & P\text{-value} = \mathrm{P}(Z \leq \tau) \\ \mu > \mu_0 & P\text{-value} = \mathrm{P}(Z \geq \tau) \\ \mu \neq \mu_0 & P\text{-value} = 2\,\mathrm{P}(Z \geq \tau) \end{cases}$$

Paired data

- Paired data are very common,
  - before/after trials
  - studies on twins
  - left arm vs right arm or left eye vs right eye experiments
- We can test if the two (paired) samples come from populations with the same mean by focusing on differences.
- Have differences zero mean?

Paired data - Assumptions

We have data of the form

X	X_1	X_2	$\ldots$	X_n
Y	Y_1	Y_2	$\ldots$	Y_n
D	D_1	D_2	$\ldots$	D_n

where $D_i = X_i - Y_i$. To perform a t-test we needed to assume

- Normality (hence identically distributed)
- Independence

where we do not know the variance of the data.

Paired data - Assumptions

- For the Paired t-test we assume that the differences $D_1, \ldots, D_n$ are independent normally distributed random variables.
- We not make assumptions on the Xs or Ys except that the Xs and Ys are *not independently obtained*, i.e. there is a natural pairing of the data.
- Later for the two-sample t-test (another test involving two sets of data) we assume that X and Y are *independently obtained*.
- The paired t-test is similar in spirit to the sign-test. However, for the sign-test we assume symmetry while for the paired t-test we make the stronger assumption that the differences are normally distributed.
- Also, the sign-test removes zero differences, whereas the paired t-test uses all available observations.

Example (Rats, PQ p125 and L16)**.** Does a biochemical substance have an inhibitive effect on muscular growth? For each of 10 rats:

- one hind leg was regularly injected with the biochemical substance.
- The corresponding muscle on the other hind leg was regularly injected with a harmless placebo.
- At the end of 6 months the weights of the muscles were measured (in gms) and recorded as follows:

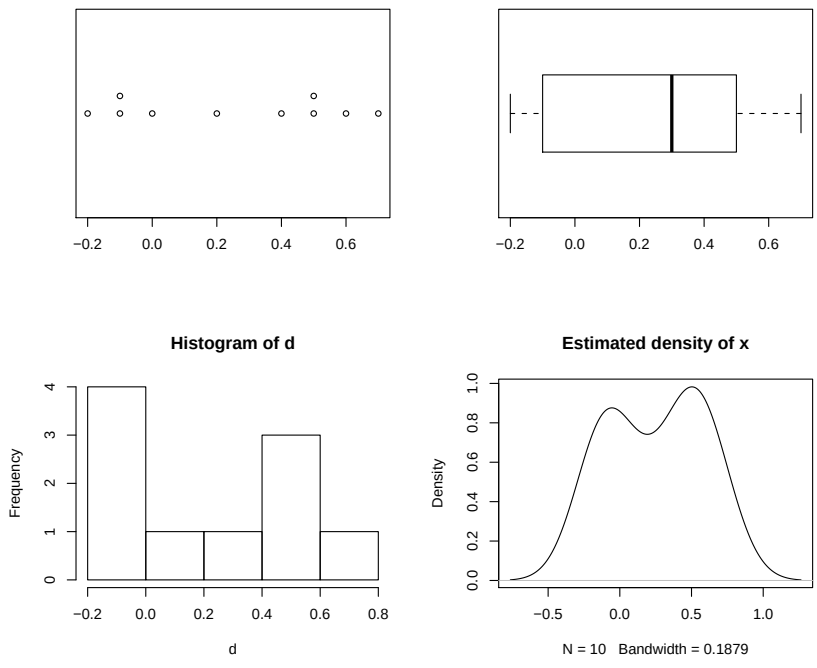
Rat	1	2	3	4	5	6	7	8	9	10
Bioch.	1.7	2.0	1.7	1.5	1.6	2.4	2.3	2.4	2.4	2.6
Placebo	2.1	1.8	2.2	2.2	1.5	2.9	2.9	2.4	2.6	2.5

- Analyse the data to determine whether this experiment provides evidence of a significant inhibitive effect.

Example (cont)

```
> x = c(1.7, 2.0, 1.7, 1.5, 1.6, 2.4, 2.3, 2.4, 2.4, 2.6)
> y = c(2.1, 1.8, 2.2, 2.2, 1.5, 2.9, 2.9, 2.4, 2.6, 2.5)
> d = y-x
> par(mfrow=c(2,2))
> stripchart(d, method="stack",offset=1, pch=1)
> boxplot(d,range=1,horizontal=TRUE)
> hist(d)
> plot(density(d),main="Estimated density of x")
```


Example (cont)



- Sample size is small ($n = 10 < 25$), exploratory data analysis suggests normality may be reasonable (again, difficult to test for small sample sizes).

```
> mean(d)
[1] 0.25
< sd(d)
[1] 0.3308239
```

- The sample average is $\overline{x} = 0.25$ and the sample standard deviation is $s = 0.3308239$.
- We wish to test,

$$H_0 \colon \mu_d = 0 \quad \text{against} \quad H_1 \colon \mu_d > 0.$$

- Again, because sample size is very small, base the test on

$$\tau = \frac{\overline{X}}{S/\sqrt{n}}.$$

- Either small values or large values of τ support H_1.

Example (cont)

- Under the assumption that the null hypothesis is true (along with independence and normality) the null distribution of the test statistic is $t_{n-1} = t_9$.
-
-
-

Example (cont)

- Alternatively, using the R commands,

```
> 1- pt(2.389699,9)
[1] 0.02028870
> pt(2.389699,9,lower.tail=F)
[1] 0.02028870
```

we get the exact P-value of 0.02028870.

Example (cont)

- Alternatively, using R:

```
> t.test(d,mu=0,alternative ="greater")

        One Sample t-test

data:  d
t = 2.3897, df = 9, p-value = 0.02029
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.05822761        Inf
sample estimates:
mean of x
     0.25
```

Example (cont)

- Via a sign test we obtain

```
> s = sign(d)[sign(d)!=0]
> table(s)
s
-1  1
 3  6
> binom.test(c(6,3),p=0.5,alt="greater")

        Exact binomial test

data:  c(6, 3)
number of successes = 6, number of trials = 9, p-value = 0.2539
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.3449414 1.0000000
sample estimates:
probability of success
             0.6666667
```

- In this case the t-test and the sign-test give conflicting results. This is not uncommon when the sample size is small.

Example – Paint

Example (Paint, continued from L16)**.** A paint supplier claims that a new additive will reduce the drying time of acrylic paint. To test this claim 10 panels of wood are painted: one half with the original paint formula and one half with the paint having the new additive. The drying times in hours are given below.

```
> panel  = 1:10
> npaint = c(6.4,5.8,7.4,5.5,6.3,7.8,8.6,8.2,7.0,4.9)
> rpaint = c(6.6,5.9,7.8,5.7,6.0,8.4,8.8,8.4,7.3,5.8)
> d = rpaint - npaint
> d
 [1]  0.2  0.1  0.4  0.2 -0.3  0.6  0.2  0.2  0.3  0.9
```

- Can we conclude that the new additive is effective in reducing the drying time of the paint?
- Same steps as in previous example.

Example (cont)

Monday, 15 October 2012

Lecture 9 - Content

- ☐ **Two-sample t-tests**
- ☐ **Confidence intervals**

Two-sample t-tests

Assumptions

Two independent samples with n_x observations $x_1, \ldots, x_{n_x}$ from one population and n_y observations $y_1, \ldots, y_{n_y}$ from another. We assume that the populations can be modelled by $\mathcal{N}(\mu_x, \sigma^2)$ and $\mathcal{N}(\mu_y, \sigma^2)$:

(i) Two independent samples from

(ii) normal populations with

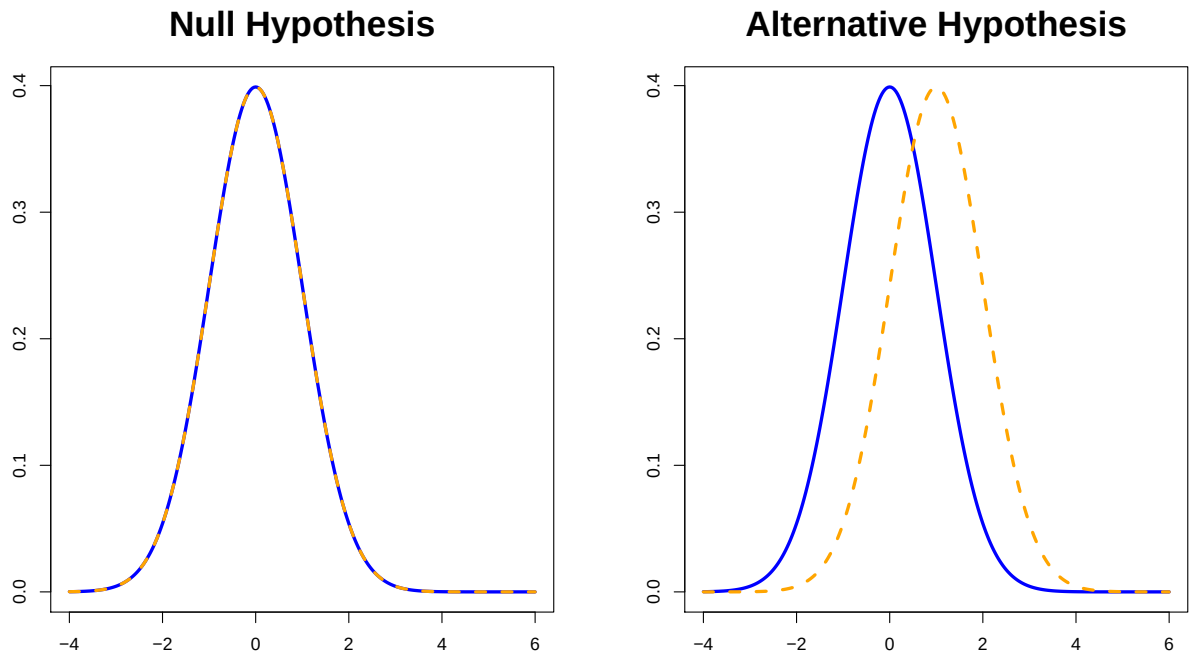
(iii) common variance.

Example (Height and gender: http://en.wikipedia.org/wiki/Human_height).

Country/Region	Average male height (m)	Average female height (m)	Age range	Method	Year
Argentina	1.735	1.608	17	Measured	1998-2001
Australia	1.748	1.635	18+	Measured	1995
Austria	1.796	1.671	21-25	Self Reported	1997-2002
Azerbaijan	1.718	1.654	16+	Measured	2005
Bahrain	1.651	1.542	19+	Measured	2002
Belgium	1.795	1.678	21-25	Self Reported	1997-2002
Bolivia	1.600	1.422	20-29	Measured	1970
Brazil	1.707	1.588	18+	Measured	2008-2009

Average height of Australians (to 0 d.p.): $\mu_x = 164$ and $\mu_y = 175$ with standard deviation typically in the range of $\sigma \in (6.5\text{cm}, 7.5\text{cm})$

Two Sample t-test



Testing equality of population means

How do we test

$$H_0 : \mu_x = \mu_y \quad \text{against} \quad H_1 : \mu_x \neq \mu_y \, ?$$

Available information:

- sample sizes: n_x and n_y
- sample means: $\overline{x}$ and $\overline{y}$
- sample variances: s_x^2 and s_y^2

Test statistic: If σ^2 is known then the differences of the means has distribution

$$\overline{X} - \overline{Y} \quad \overset{\text{if } H_0 \text{ is true}}{\Rightarrow} \quad \overline{X} - \overline{Y} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}\right)$$

Two-sample Z- and t-test statistics

Hence, if $H_0 : \mu_x = \mu_y$ is true and σ^2 is known,

$$\frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = Z \sim \mathcal{N}(0, 1)$$

and more generally, if σ^2 is unknown and can be estimated by the pooled variance

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

thus,

$$\tau = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{(n_x+n_y-2)},$$

i.e. a t-distribution with degrees of freedom equal $\nu = n - 2 = n_x + n_y - 2$.

Example (Height and gender: http://en.wikipedia.org/wiki/Human_height (cont)). Suppose that in a particular MATH1905 tutorial we have $\overline{x} = 164$, $\overline{y} = 175$, $s_x = 6.8$, $s_y = 7.2$, $n_x = 8$, $n_y = 9$ and we want to test whether males are taller than females in the MATH1905 tutorial.

- The null and alternative hypotheses are

$$H_0 \colon \mu_x = \mu_y \quad \text{versus} \quad H_1 \colon \mu_x < \mu_y.$$

- The pooled variance is given by

$$s_p^2 = \frac{(8-1) \times 6.8^2 + (9-1) \times 7.2^2}{(8+9-2)} = 49.22667.$$

- The observed value of the test statistic is

$$\tau = \frac{164 - 175}{\sqrt{49.22667} \times \sqrt{\frac{1}{8} + \frac{1}{9}}} = -3.226519$$

- Large (negative) values provide evidence for H_1.

- Assuming independent normal observations with common variance and under the null hypothesis the null distribution of the test statistic is $t_{(n_x+n_y-2)} = t_{15}$.
- The P-value for this hypothesis is given by

$$P(t_{15} < -3.226519) = \text{pt(-3.226519, 15)} = 0.002824303$$

- Hence, we reject the null hypothesis (that male and female heights in the MATH1905 tutorial are equal) in favour of the alternative hypothesis (that men and taller than women in the MATH1905 tutorial are equal).

Example (Fusion of Ice)**.** Two methods, A and B were used in the determination of the latent heat of fusion of ice. The investigators wished to find out whether the methods differed. The following table gives the change in total heat from ice at $-0.72°$C to water at $0°$C in calories per gram.

A: 79.98 80.04 80.02 80.04 80.03 80.03 80.04
80.05 80.03 80.02 80.00 80.02 79.97

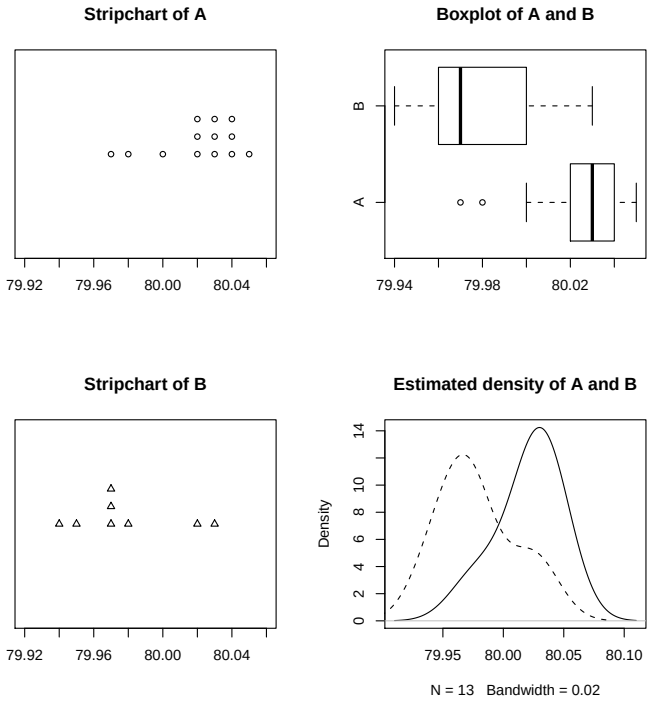
B: 80.02 79.94 79.98 79.97
79.97 80.03 79.95 79.97

- Assume the change in total heat values can be modelled in each case by a normal distribution.
- Do you agree?

Example (cont)

```
> A = c(79.98,80.04,80.02,80.04,80.03,80.03,80.04,80.05,
        80.03,80.02,80.00,80.02,79.97)
> B = c(80.02,79.94,79.98,79.97,79.97,80.03,79.95,79.97)
> par(mfrow=c(2,2))
> stripchart(A, method="stack",offset=1, pch=1,xlim=c(79.92,80.06))
> title(main="Stripchart of A")
> boxplot(c(A,B)~c(rep("A",13),rep("B",8)),range=1,horizontal=TRUE)
> title(main="Boxplot of A and B")
> stripchart(B, method="stack",offset=1, pch=2,xlim=c(79.92,80.06))
> title(main="Stripchart of B")
> plot(density(A,bw=0.02),main="Estimated density of A and B")
> points(density(B,bw=0.02),type="l",lty=2)
> c(length(A),length(B)) ... edited ...   [1] 13     8
> c(mean(A),mean(B)) ... edited ...        [1] 80.02 79.98
> c(sd(A),sd(B)) ... edited ...            [1] 0.024 0.031
```

Example (cont)



Example (cont)

A: $n_x = 13 \quad \overline{x} = 80.0208 \quad s_x = 0.02397$

B: $n_y = 8 \quad \overline{y} = 79.9788 \quad s_y = 0.03137$

- ☐
- ☐
- ☐
- ☐
- ☐
- ☐

In R with `pt()` command or by `t.test(A,B,mu=0,var.equal=TRUE)`.

```
        Two Sample t-test

data:  A and B
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```


Summary of Hypothesis Testing

1. Tests for Proportions: $X \sim \mathcal{B}(n, p)$

   $$H_0 : p = p_0$$

   Base the test on X and use binomial tables or the normal approx. to get the P-value of the **binomial test**.

2. Tests of the Mean - Single Sample:

   $$H_0 : \mu = \mu_0.$$

   (i) Population is symmetric
   Use the **sign test** which is based on the test for proportions and the number of positive signs with $p_0 = 0.5$.

2. Tests of the Mean - Single Sample (cont):

(ii) Population is $\mathcal{N}(\mu, \sigma^2)$ with σ *known*.
Use the **Z-test**

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

(iii) Population is $\mathcal{N}(\mu, \sigma^2)$ with σ *unknown* and n is small ($n < 25$).
Use the **t-test**.

$$\tau = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

(iv) Population is $\mathcal{N}(\mu, \sigma^2)$ with σ *unknown* and n is large ($n > 25$).
Then **Z-test** approx. **t-test**.

$$\tau = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim N(0,1)$$

3. Tests of Means - Two Samples:

Are the data paired?

(a) Yes - Calculate the differences.

(i) Differences have a symmetric distribution about μ
Use the **sign test** to test $H_0 : \mu = 0$.

(ii) Differences have a $\mathcal{N}(\mu, \sigma^2)$ distribution
Use the **t-test** to test $H_0 : \mu = 0$.

(b) No - Are the samples independent?
Are the populations Normal with common variance? If 'yes', use the

2 sample t-test to test $H_0 : \mu_x = \mu_y$

$$\tau = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}.$$

Confidence intervals

- Given a sample $X_1, \ldots, X_n$ from a normal population $X \sim \mathcal{N}(\mu, \sigma^2)$ how do we estimate μ?
- The best estimate in the least squares or maximum-likelihood sense is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- $\overline{X}$ is close to the true μ but with probability one wrong, i.e.

$$\mathrm{P}(\overline{X} = \mu) = 0 \quad \text{since the normal is continuous.}$$

- Known result: If σ is known then,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = Z \sim \mathcal{N}(0, 1)$$

and thus $\mathrm{P}(-1.96 \leq Z \leq 1.96) = 0.95 \Rightarrow$ substitute Z and solve for μ.

95% CI for μ if σ is known

Thus,

95% CI for μ if σ is known

$$0.95 = \mathrm{P}\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

We can interpret this equation as saying:

> If we were to repeat the experiment over and over again (with the same sample size) and recalculate the confidence interval each time then 95% of the calculated confidence intervals will contain the true value of μ.

Using statistical jargon we say the random interval

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

covers μ with probability 0.95.

Another of John's pet hates

The wrong interpretation is:

> There is a 95% chance that the population mean is between 165cm and 189cm.

The correct interpretation is:

> For 95and 189cm covers the population mean.

Note that the "randomness" is on the fact that samples are drawn from a particular population, **not in the parameter of interest!**

$100(1-\alpha)\%$ CI for μ if σ is known

Definition 13. The $100(1-\alpha)\%$ CI for μ is given by

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and is constructed by finding $z_{\alpha/2}$ such that

$$1-\alpha = \mathrm{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

and solving for μ.

Example (Cholesterol Levels)**.** Consider the distribution of serum cholesterol levels for all males in the United States who are hypersensitive and who smoke. The distribution is normal with an unknown mean and a known variance of 46 mg/100 ml (based on historical records). Suppose that we draw a random sample of size $n = 12$ from the population of interest which has sample average $\overline{x} = 217$ mg/100 ml. What is the 95% confidence interval the population mean μ?

□

□

Tuesday, 16 October 2012

Lecture 10 - Content

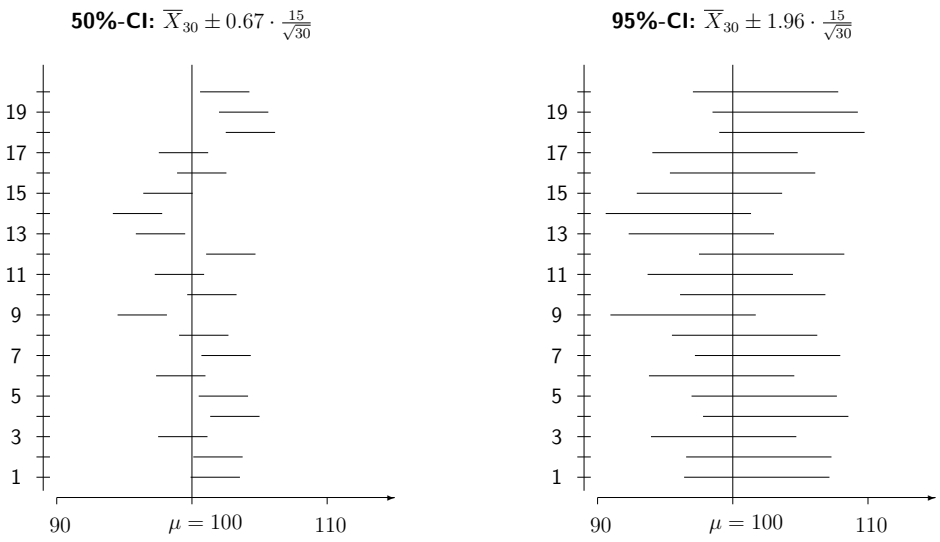
- **Confidence intervals continued**

Confidence intervals (cont)

Properties of CIs

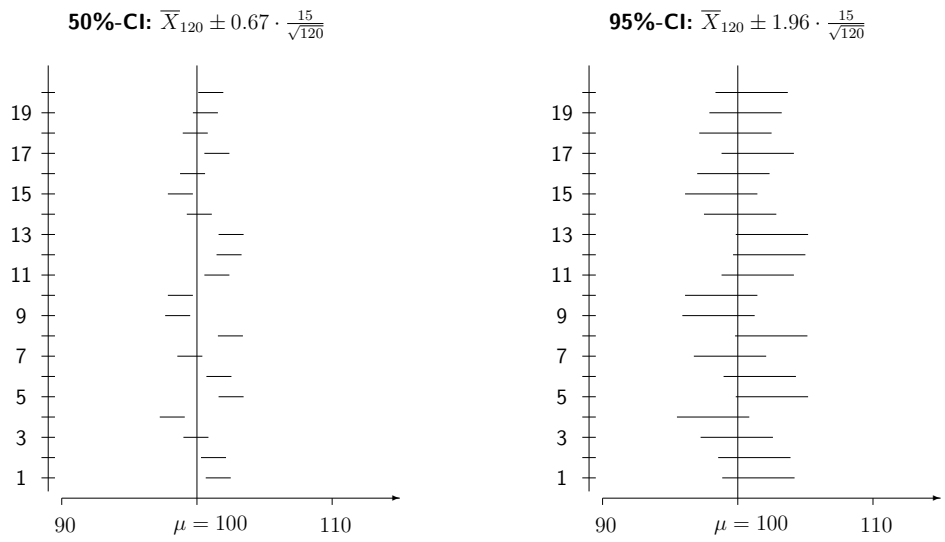
- Cover the true μ value with relative frequency approximately $(1-\alpha)$;
- as you increase n the CI gets narrower;
- as you increase the confidence level, i.e. make $(1-\alpha)$ larger, the CI gets wider.

Simulated CIs for IQ tests, $n = 30$:



Here, population variance is $\sigma^2 = 15^2$ and population mean $\mu = 100$.

Simulated CIs for IQ tests, $n = 120$:



Here, population variance is $\sigma^2 = 15^2$ and population mean $\mu = 100$.

Example (Birthweight). Use the following data to construct a 90% and 99% CI for the average birthweight of a term baby (37 - 41 weeks gestation) if it is known that the birthweight (in kgs) is $W \sim \mathcal{N}(\mu, 0.525^2)$.

$$2.853,\ 3.127,\ 3.159,\ 3.800,\ 2.656,\ 3.245,\ 3.510,\ 3.082$$

- $\overline{x} = 25.432/8 = 3.179$.
- 90% CI for μ: Find z such that $0.90 = \mathrm{P}(-z \leq Z \leq z)$, that is, $\mathrm{P}(Z > z) = 0.05$. From t-tables with $\nu = \infty$, z-tables or with R: $z = 1.645$.
- C.I. calculates to $3.179 \pm 1.645 \times \frac{0.525}{\sqrt{8}} = 3.179 \pm 0.305 = (2.874,\ 3.484)$.
- 99% C.I. for μ: $0.99 = \mathrm{P}(-z_1 \leq Z \leq z_1) \Rightarrow z_1 = 2.576$ and CI is $3.179 \pm 2.576 \times \frac{0.525}{\sqrt{8}} = 3.179 \pm 0.478 = (2.701,\ 3.657)$.

$100(1-\alpha)\%$ CI for μ if σ is unknown

Base the CI on the t-statistic,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

where S^2 the sample variance.

Definition 14. A $100(1-\alpha)\%$ CI for μ of a normal population with unknown variance σ^2 is given by

$$\overline{X} \pm t' \frac{s}{\sqrt{n}},$$

where t' is from the t-tables or from R such that

$$1-\alpha = \mathrm{P}(-t' \leq t_{n-1} \leq t').$$

Example. Consider the distribution of serum cholesterol levels for all males in the United States who are hypersensitive and who smoke. The distribution is normal with an unknown mean and a unknown variance. Suppose that we draw a random sample of size $n = 12$ from the population of interest which has sample average of 217 mg/100 ml and sample variance of 46. What is the 95% confidence interval the population mean μ?

□

□

- Note that when we assumed $\sigma = s = \sqrt{46}$ we obtained the confidence interval

$$(213.16, 220.84)$$

- Notice the confidence intervals are slightly wider taking into account the uncertainty when estimating σ by s.

Example (Paint). The 10 values below are the first sample of values on paint primer thickness that were collected as part of an ongoing process of monitoring the performance of an industrial system.

$$1.30,\ 1.10,\ 1.20,\ 1.25,\ 1.05,$$

$$0.95,\ 1.10,\ 1.16,\ 1.37,\ 0.98$$

- Assume the primer thickness can be modelled by $X \sim \mathcal{N}(\mu, \sigma^2)$.
- $\overline{x} = 1.146, \quad s = 0.1363.$
- A 95% C.I. for μ is $\overline{x} \pm t' \frac{s}{\sqrt{10}}$, where $0.95 = \mathrm{P}(-t' \leq t_9 \leq t')$.
- $\mathrm{P}(t_9 > t') = 0.025$ thus, $t' = 2.262$.
- The CI is $1.146 \pm 2.262 \times \frac{0.1363}{\sqrt{10}} = 1.146 \pm 0.097$ or $(1.049,\ 1.243)$.

CIs for proportions

Data: n independent trials and the probability of success at each trial is p, X denotes the number of successes,

$$\Rightarrow X \sim \mathcal{B}(n,p), \quad \mathrm{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p).$$

Standardized scores: Calculate standardized number of successes,

$$Z' = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}},$$

where $\widehat{p} = X/n$ is the sample proportion (estimated proportion).

- If n is large: $Z' \simeq \mathcal{N}(0,1) \Rightarrow$ use Z' to obtain approximate CIs for p.
- However, the variance depends also on the unknown parameter p!
- $\mathrm{Var}\, X/n = p(1-p)/n \approx \widehat{p}(1-\widehat{p})/n \leq \frac{1}{2}\left(1-\frac{1}{2}\right)/n = \frac{1}{4n}$.

Definition 15. An approximate $100(1-\alpha)\%$ CI for p is obtained from

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

and a conservative CI for p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{4n}}.$$

Example. What sample size is necessary to give a 95% C.I. for a proportion with width ± 0.03? [Note that as a convention, width is the same as the length of the CI, i.e. ± 0.03 corresponds to a length = width = 0.06]

-
- 1068

Example. A new type of photoflash bulb was tested to estimate the probability, p, of producing the required light output at the appropriate time. The sample of 1000 bulbs were tested and 810 were observed to function according to specifications. Estimate p and find and approximate 95% confidence interval for p.

- □
- □
- □

Comments on opinion polls

- ACNeilsen and others poll typically about $1,000$ people.
- Why?
- The conservative $\pm$ factor for a 95% C.I. is

$$1.96/\sqrt{4 \times 1000} = 0.031$$

  hence the margin of error is about 3 percent.
- As a rough guide the margin of error is

$$\frac{1.96}{\sqrt{4n}} \simeq \frac{1}{\sqrt{n}}.$$

Example (Sample sizes in surveys)**.** A survey is to be conducted to determine the proportion of a population with a certain attribute.

(i) What sample size is necessary to ensure the sample proportion is within 0.03 of the true population proportion with probability at least 0.9?

Solutions:

(i) We want n such that $\mathrm{P}(|\widehat{p} - p| < 0.03) \geq 0.90$. $\widehat{p}$ is approximately normally distributed with variance $p(1-p)/n$ so we want $P\left(|Z| < \frac{0.03 \times \sqrt{n}}{\sqrt{p(1-p)}}\right) \geq 0.90$

$$\frac{0.03 \times \sqrt{n}}{\sqrt{p(1-p)}} \geq 1.645 \overset{\text{solve for } n}{\Rightarrow} n \geq \left(\frac{1.645}{0.03}\right)^2 \times p(1-p).$$

If we replace $p(1-p)$ by $\frac{1}{4}$ then we have $n \geq 751.67$ so a sample of size 752 will certainly suffice.

Example (Sample sizes in surveys)**.** A survey is to be conducted to determine the proportion of a population with a certain attribute.

(ii) What sample size is needed so that a 95% C.I. has width no more than 0.04 (i.e. the $\pm$ term is less than 0.02)?

Solutions:

(ii) We use the conservative version of the C.I. and recall the 95% C.I. $\pm$ factor is always less than

$$1.96\sqrt{\frac{1}{4n}}.$$

Solve

$$\frac{1.96}{2\sqrt{n}} \leq 0.02 \Rightarrow \frac{1.96}{2 \times 0.02} \leq \sqrt{n} \Rightarrow 2401 \leq n.$$

Thus a sample of 2401 observations is needed.

Example (Sample sizes in surveys)**.** A survey is to be conducted to determine the proportion of a population with a certain attribute.

(iii) As in (ii) but assuming that the true proportion will be less than 30%?

Solutions:

(iii) Because $p \leq 0.3$ we get a smaller conservative bound of $\operatorname{Var} Z' \leq 0.3 \times 0.7/n$. Hence, for the 95% CI the $\pm$ factor is always less than

$$1.96\sqrt{\frac{0.21}{n}}.$$

Solve

$$\frac{1.96 \times \sqrt{0.21}}{\sqrt{n}} \leq 0.02 \;\Rightarrow\; \frac{1.96 \times \sqrt{0.21}}{0.02} \leq \sqrt{n} \;\Rightarrow\; 2016.84 \leq n.$$

Thus a sample of 2017 observations is needed.

Summary of Confidence Interval

We have covered the following cases:

- Normal/Constant $\sigma^2 = \sigma_0^2$ case: $\overline{x} \pm z^* \times \frac{\sigma_0}{\sqrt{n}}$
- Normal/Unknown $\sigma^2/n < 30$ case: $\overline{x} \pm t^* \times \frac{s}{\sqrt{n}}$
- Normal/Unknown $\sigma^2/n \geq 30$ case: $\overline{x} \pm z^* \times \frac{s}{\sqrt{n}}$
- Proportions: $\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
- Proportions (Conservative): $\widehat{p} \pm z^* \sqrt{\frac{1}{4n}}$

where $P(|Z| \leq z^*) = 1 - \alpha$, $P(|t_{n-1}| \leq t^*) = 1 - \alpha$ and α is typically 5%.

Monday, 22 October 2012

Lecture 11 - Content

- χ^2 **goodness of fit tests**
- **Further applications of** χ^2 **GoF tests**

χ^2 Goodness of fit tests

Motivational setting

- Suppose we have n independent trials with X successes and $n - X$ failures: $X \sim \mathcal{B}(n, p)$.
- Test $H_0 : p = p_0$ against $H_1 : p \neq p_0$. If H_0 is true then

	Success	Failure	Total
Observed	$O_1 = X$	$O_2 = n - X$	n
Expected	$E_1 = np_0$	$E_2 = n(1 - p_0)$	n

- Large values of $|X - np_0|$ support H_1. Thus large values of

$$\tau = \frac{(X - np_0)^2}{np_0(1 - p_0)}$$

support H_1.

Motivational setting (cont)

- Note,

$$\begin{aligned}(O_2 - E_2)^2 &= [(n - X) - (n - np_0)]^2 \\ &= (X - np_0)^2 = (O_1 - E_1)^2.\end{aligned}$$

- Also,

$$\frac{1}{np_0} + \frac{1}{n(1 - p_0)} = \frac{1}{np_0(1 - p_0)}.$$

- Thus,

$$\begin{aligned}\tau &= \frac{(X - np_0)^2}{np_0(1 - p_0)} = (X - np_0)^2[\frac{1}{np_0} + \frac{1}{n(1 - p_0)}] \\ &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}.\end{aligned}$$

- This is a special case of Pearson's χ^2 statistic.

Pearson's X^2 GoF test

Assume we have g categories, not just success/failure and H_0 specifies a model giving expected frequencies for each category.

Definition 16. Pearson's χ^2 test-statistic is

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed frequency in the ith category and E_i is the expected frequency if H_0 is true.

Alternative Calculation Formula for X^2

Pearson's X^2 GoF test (cont)

- Thus, the easiest form for calculation purposes of X^2 is,

$$X^2 = \sum_{i=1}^{g} \frac{O_i^2}{E_i} - n.$$

- We reject the model if the X^2-statistic is too large.
- The sampling distribution of the statistic has (asymptotically) a chi-squared distribution with $g-1$ degrees of freedom.

$$P\text{-value} = \mathrm{P}(\chi^2_{g-1} \geq \text{observed } X^2 \text{ value}).$$

  - Note that $\chi^2_1 = Z^2$, where $Z \sim \mathcal{N}(0,1)$. In R with `pchisq()` or tables.
  - The X^2 test should only be used when the expected frequencies, E_i, are greater than 5.
    (Recall this corresponds to $np \geq 5$ for the normal approximation to the binomial!)

The χ^2 distribution

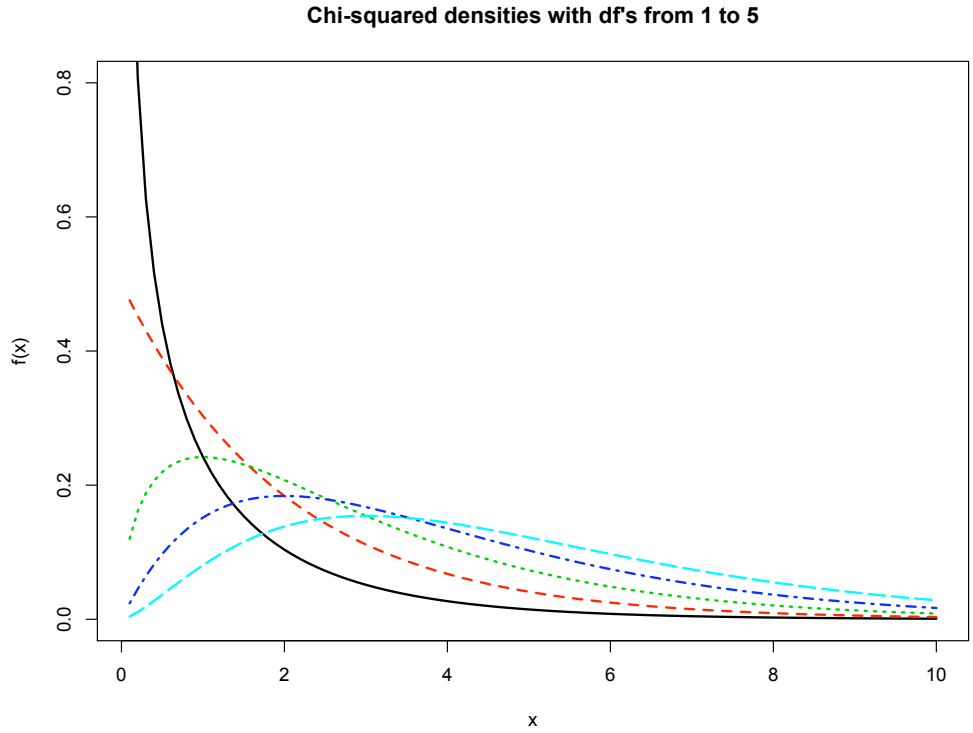
- The χ^2 r.v. can only take non-negative values. The distribution is not symmetric but is right skewed. Tables typically give

$$\mathrm{P}(\chi^2_\nu > x) = p = \texttt{1-pchisq(x,nu)}$$

  for particular d.f. ν and p values.
- With R we can visualise the densities:

```
> x = 1:100/10
> plot(x,dchisq(x,1),type="l",ylim=c(0,0.8))
> points(x,dchisq(x,2),type="l",lty=2,col=2)
> points(x,dchisq(x,3),type="l",lty=3,col=3)
> points(x,dchisq(x,4),type="l",lty=4,col=4)
> points(x,dchisq(x,5),type="l",lty=5,col=5)
```

The χ^2 distribution (cont)



Example.

(i) $\mathrm{P}(\chi^2_1 > 3.841) = ?$

- From tables the points tabulated below are x, where $P(\chi^2_\nu > x) = p$

	p					
ν	0.25	0.15	0.10	0.05	0.025	0.01
1	1.323	2.072	2.706	3.841	5.024	6.635

- Using R:

```
> 1-pchisq(3.841,1)
[1] 0.05001368
> pchisq(3.841,1,lower.tail=FALSE)
[1] 0.05001368
```

(Note also that $1.96^2 = 3.841,\ P(|Z|^2 > 1.96^2) = 0.05$.)

(ii) $\mathrm{P}(\chi^2_{10} > 20) = ?$

- From tables

			p			
ν	0.25	0.15	0.10	0.05	0.025	0.01
10	12.549	14.534	15.987	18.307	20.483	23.209

  So that $0.025 < \mathrm{P}(\chi^2_{10} > 20) < 0.05$.

- Via R

```
> pchisq(20,10,lower.tail=FALSE)
[1] 0.02925269
```

(iii) $\mathrm{P}(13.848 < \chi^2_{24} < 39.364) = ?$

- Can't be obtained because tables only have $1 \leq \nu \leq 10$.
- Via R

```
> pchisq(39.364,24) - pchisq(13.848,24)
[1] 0.9250087
```


Example (Phenotypes, PQ p117)**.** In an experiment involving a dihybrid cross of flies, 148 progeny were classified by phenotype as follows.

AB	Ab	aB	ab	Total
87	31	25	5	148

- Genetic theory predicts a ratio 9:3:3:1 for AB:Ab:aB:ab.
- Do the data support the theory?
-

Example (cont).

-

-

```
pchisq(2.775,3,lower.tail=FALSE)
```

-

Example (Accidents). The number of fatal accidents on NSW roads in the July month between 2004 - 2010 were:

2004	2005	2006	2007	2008	2009	2010
44	50	34	41	34	27	29

Test the claim that the accident rate didn't change over this seven year period:

- p_i denotes the probability that a fatal accident is 'allocated' to month i.
- Model: $p_i = \frac{1}{7}, \quad i = 1, \ldots, 7.$
- The total number of accidents is 259. Thus $E_i = \frac{259}{7} = 37.$

  The test statistic is $X^2 = \sum_{i=1}^{7} \frac{O_i^2}{E_i} - 259 = 11.24.$

- Thus, the P-value $= 0.081 \Rightarrow$ no rejection of the claim that the accident rate is constant across the years based on these data.

Further applications of χ^2 GoF tests

- Recall that for known parameters,

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{g} \frac{O_i^2}{E_i} - n \overset{\text{under } H_0}{\sim} \chi^2_{g-1}.$$

- If we want to check the fit of a model that involves unknown parameters we first have to estimate the parameters.
- Since we use the same data to estimate the parameters and test the fit we find the sampling distribution of the X^2 statistic has to be adjusted.
- The distribution is still χ^2 but the dfs are reduced to $g - k - 1$, where
  - g is the number of categories and
  - k is the smallest number of parameters that need to be estimated using the data.

Example (More on phenotypes)**.** In a backcross experiment to investigate the genetic linkage between two factors A and B in a species of flower, some researchers classified 400 offspring by phenotype as follows:

AB	Ab	aB	ab
128	86	74	112

(i) Under the no linkage model, the four phenotypes are equally likely.
Show that this model is a poor fit.

(ii) If linkage is in the coupling phase, the probabilities of the four phenotypes are

AB	Ab	aB	ab
$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

where p is the 'recombination fraction' and is estimated by the overall proportion of Ab and aB.
Show that this model fits the data well.

Example (cont; (i)). Model says that all categories are equally likely.

- The total number of observations is $n = 400$.
- Observed O_i : 128 86 74 112
  Expected E_i : 100 100 100 100
- $X^2 = \sum_i \frac{O_i^2}{E_i} - n = 18.$
- $P = \mathrm{P}(\chi_3^2 \geq 18) < 0.01.$
- Thus the data are not consistent with the model.

Example (cont; (ii)). Here, $\widehat{p} = (86 + 74)/400 = 0.4$.

- The expected frequencies are

$$E_1 = E_4 = 400 \times 0.3 = 120$$

$$E_2 = E_3 = 400 \times 0.2 = 80.$$

- $X^2 = 1.97$ and $g - k - 1 = 2$
- $P = \mathrm{P}(\chi_2^2 \geq 1.97) > 0.10.$
- The data are consistent with this model.

Tuesday, 23 October 2012

Lecture 12 - Content

- **Further applications of χ^2 GoF tests (cont)**

Example (Infections)**.** 200 groups of 5 insects each were inspected. For each group the number of infected insects (x) was counted giving:

$$3,\ 2,\ 5,\ 1,\ 0,\ ..,\ 2.$$

The data were condensed into the table below, writing x_i for the number infected and f_i for the corresponding frequency:

x_i :	0	1	2	3	4	5	Total
f_i :	20	62	55	38	20	5	200

Does the binomial model fit the data?

- The null hypothesis is that $X \sim \mathcal{B}(5, p)$. We need to estimate p.
- There were $5 \times 200 = 1000$ insects in total and 391 of these were infected, i.e. an estimate is

$$\widehat{p} = \frac{391}{1000} = 0.391.$$

Example (cont)**.**

i	0	1	2	3	4	5	Total
p_i	0.0837	0.2689	0.3453	0.2217	0.0712	0.0091	1.00
E_i	16.754	53.783	69.061	44.340	14.234	1.828	200
O_i	20	62	55	38	20	5	200

Notice that the E_i value falls below 5 for the last group.

Testing the fit of a normal model

Given a data set $x_1, x_2, .., x_n$ we want to test if the data come from a $\mathcal{N}(\mu, \sigma^2)$ population.

(a) First calculate the sample mean, $\overline{x}$, and the sample variance, s^2.

(b) Form a grouped frequency table summary of the data with (ideally) 5 to 10 categories. Aim to have at least 5 values in each category.

(c) To check the normal claim work out the expected frequencies for each category by fitting $\mathcal{N}(\overline{x}, s^2)$.

(d) Use X^2 as the test statistic. To calculate the P-value use $(g - 2 - 1)$ df.

Example (Rainfall). We have $n = 30$ observations corresponding to Sydney's annual rainfall (in inches) from 1980-2009 (from http://www.bom.gov.au):

```
> y = c( 956,1083,1499, 994, 816, 995,1200, 860,1359, 822,
+       1470,1649,1078,1149,1230, 907, 913,1282,1121,1977,
+       1526,1862,1313,1225,1217,1801,1346, 838,1038, 736)
> x = 2009:1980
```

Test if the rainfall follows a normal distribution.

(a) $\overline{y} = 1208.733$ and $s^2 = 105762.8$.

(b) Grouping the data into a frequency table:

Interval	Frequency
$y \leq 900$	5
$900 < y \leq 1200$	11
$1200 < y \leq 1500$	9
$y > 1500$	5

Example (cont).

(c) We now calculate the expected frequencies using

$$Y \sim \mathcal{N}(1208.733, 325.212^2)$$

$$P(Y \leq 900) = \mathrm{P}\left(Z \leq \frac{900 - 1208.733}{325.212}\right) = \mathrm{P}(Z \leq -0.95) = 0.1712.$$

Thus $E_1 = 30 \times 0.1712 = 5.137$.

$$\begin{aligned} P(900 < Y \leq 1200) &= P(-0.95 < Z \leq -0.03) \\ &= 0.318 \end{aligned}$$

$$E_2 = 30 \times 0.318 = 9.542.$$

Similarly,
$E_3 = 30 \times 0.3254 = 9.765$ and
$E_4 = 30 - 5.137 - 9.542 - 9.765 = 5.556$.

Example (cont). (d) Our table is

Interval	Frequency	Expected
$y \leq 900$	5	5.137
$900 < y \leq 1200$	11	9.542
$1200 < y \leq 1500$	9	9.765
$y > 1500$	5	5.556

The test statistic is

$$X^2 = \frac{5^2}{5.137} + \frac{11^2}{9.542} + \frac{9^2}{9.765} + \frac{5^2}{5.556} - 30 = 0.342.$$

Here $g = 4$ and $k = 2$ so we have 1 d.f.

The P-value is $P(\chi^2_1 \geq 0.342) = 0.559,$ with R.

Thus the data are consistent with the normal model.

Tests for independence

If we have data classified according to two attributes then we can construct a contingency table which is just a convenient way of presenting the group frequencies. For example, we have data on 422 drivers and motorcyclists killed in NSW in 1988. We classify the people by blood alcohol level and gender.

Alc (g/100ml)	0	$(0, 0.08)$	$[0.08, 0.15)$	≥ 0.15	Total
Male	206	37	35	76	354
Female	53	5	4	6	68
Total	259	42	39	82	

Test the claim that gender is independent of blood alcohol level.

A probability model for contingency tables

- Let p_{ij} denote the probability of a victim being gender i and alcohol level group j then the independence model says:

$$p_{ij} = p_i^g p_j^a, \quad \text{where}$$

  - p_i^g is the prob. of being of gender i,
  - p_j^a is the prob. of being in alcohol group j.

- We estimate p_i^s and p_j^a by the marginal proportions,

$$\widehat{p}_1^g = \frac{354}{422} \Rightarrow \widehat{p}_2^g = 1 - \widehat{p}_1^g = \frac{68}{422};$$

$$\widehat{p}_1^a = \frac{259}{422}, \ \widehat{p}_2^a = \frac{42}{422}, \ \widehat{p}_3^a = \frac{39}{422}, \ \text{and} \ \widehat{p}_4^a = \frac{82}{422}.$$

- The expected frequency under the independence model in the Male/Alcohol 0 group is

$$422 \times \widehat{p}_{11} = 422 \times \frac{259}{422} \times \frac{354}{422} = 217.265.$$

- For a general table with entries x_{ij} the expected frequencies under independence are
$$E_{ij} = n \times \frac{x_{i\bullet}}{n} \times \frac{x_{\bullet j}}{n} = \frac{x_{i\bullet} \times x_{\bullet j}}{n},$$
where $x_{i\bullet}$ denotes the sum of the ith row and $x_{\bullet j}$ denotes the sum of the jth column.
- The expected frequencies for the accident data are

```
217.265   35.232   32.716   68.787
 41.735    6.768    6.284   13.213
```

-

-

```
1-pchisq(9.859,3)
```

-

- In general if we have a table with r rows and c columns then the test statistic for testing independence will have
$$(r-1)(c-1) \text{ degrees of freedom.}$$

Tests for Symmetry in Tables

The following data for 205 married people were reported in Yule (1900).

	Wife		
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Here there are $g = 9$ groups. Suppose we wish to test for table symmetry, i.e. the probability of Tall Men marrying Short Women and Short Men marrying Tall Women are roughly equal.

The table corresponding to a symmetric table model would have $k = 5$ parameters

	Wife		
Husband	Tall	Medium	Short
Tall	p_1	p_2	p_3
Medium	p_2	p_4	p_5
Short	p_3	p_5	$1 - p_1 - \cdots - p_5$

Tests for Symmetry in Tables

	Wife		
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Under the symmetric model the expected values of the table entries would be

	Wife		
Husband	Tall	Medium	Short
Tall	18	24	13
Medium	24	51	26.5
Short	13	26.5	9

Tests for Symmetry in Tables

The value of the test statistic would be

$$X^2 = \frac{(18-18)^2}{18} + \frac{(28-24)^2}{24} + \cdots + \frac{(9-9)^2}{9} = 1.656991$$

The degrees of freedom would be $\nu = g - k - 1 = 9 - 5 - 1 = 3$.

The P-value would then be

$$P(\chi^2_3 > 1.656991) = \texttt{1-pchisq(1.656991,3)} = 0.6465376$$

So we would accept the null hypothesis for table symmetry.

Note that for a test for independence the test statistic is 2.907 (check yourself!) with $\nu = 4$ degrees of freedom and a P-value of 0.5735. Thus, a null hypothesis of independent rows and columns is also consistent with the data!

THE END !!!