BUILDING SCIENCE
IN AFRICA

AIMS VIDEO COURSES
SUPPORTING BOOKLET

PROBABILITY & STATISTICS

WITH
PROF DAVID SPIEGELHALTER

AIMS
SOUTH AFRICA

AIMS
SOUTH AFRICA

African Institute for Mathematical Sciences

6 MELROSE ROAD | MUIZENBERG | CAPE TOWN 7945 | SOUTH AFRICA
TEL: +27 (0)21 787 9320 | FAX: +27 (0)21 787 9321
EMAIL: info@aims.ac.za | WEB: www.aims.ac.za

AIMS Online Courses

The mission of the AIMS academic programme is to provide an excellent, advanced education in the mathematical sciences to talented African students in order to develop independent thinkers, researchers and problem solvers who will contribute to Africa's scientific development.

Teaching at AIMS is based on the principle of learning and understanding, rather than simply listening and writing, during classes, and on creating an atmosphere of increasing our knowledge through class discussions, through small group discussions, by formulating conjectures and assessing the evidence for them, and sometimes going down wrong paths and learning from the mistakes that led us there. The essential features of the classes at AIMS are that, in contrast to formal lecture courses, they are highly interactive, where the students engage with the lecturer throughout the class time, are encouraged to learn together in a journey of questioning and discovery, and where lecturers respond to the needs of the class rather than to a pre-determined syllabus. AIMS teaching philosophy is to promote critical and creative thinking, to experience the excitement of learning from true understanding, and to avoid rote learning directed only towards assessment.

Leading international and local experts offer the courses at AIMS, which are three weeks long (each module consisting of 30 hrs) and collectively form the coursework for a structured masters degree which also includes a research component. The advertised content is a guide, and the lecturers are encouraged, and indeed expected, to adapt daily to meet the current needs of the students.

Over the past ten years AIMS has achieved international recognition for this innovative and flexible approach. It has been the starting point for the remarkable success of our students and alumni and we all benefit from the support of many who have "witnessed the AIMS-magic and keep coming back for more."

This year we have decided to film selected courses and to make them available to a larger audience as an online facility. African universities may choose to use these courses to supplement and enhance their own postgraduate programmes. We believe this would be best achieved through engagement with AIMS. One way for this to happen, would be for AIMS to suggest or nominate a specialist tutor to spend time at the university, guiding students who follow the online programme. Where possible expert lecturers who have taught at AIMS may visit the university to give a short introduction to the course. We would welcome this interaction as well as the contribution our online courses will make to the growth of the mathematical sciences ecosystem in Africa.

Barry Green
Director & Professor of Mathematics
African Institute for Mathematical Sciences
January 2013

PROBABILITY & STATISTICS

2012

PROF DAVID SPIEGELHALTER

DAY 8

AIMS

SOUTH AFRICA

Idea of confidence intervals

Assume we are to observe a set of independent and identically distributed random variables $X_1, .., X_n$ each with probability density $f(x_i|\theta)$ for some unknown parameter θ.

We are going to calculate a MLE $\hat{\theta}(\underline{X})$

Assume we know $\mathbb{V}[\hat{\theta}(\underline{X})]$, the variance of the MLE, to be V

$\sqrt{V}$ is known as the *standard error* of $\hat{\theta}$)

Assume

$$\hat{\theta}(\underline{X}) \sim \text{Normal}(\theta_0, V).$$

Suppose we are going to calculate a random interval

$$\hat{\theta}(\underline{X}) \pm c\sqrt{V}$$

for some constant c.

Confidence intervals

Then the probability this interval will include the true value θ_0 is

$$P(\theta_0 \in \hat{\theta}(\underline{X}) \pm c\sqrt{V}) = P(\hat{\theta}(\underline{X}) - c\sqrt{V} < \theta_0 < \hat{\theta}(\underline{X}) + c\sqrt{V})$$

$$P(\theta_0 - c\sqrt{V} < \hat{\theta}(\underline{X}) < \theta_0 + c\sqrt{V}) = P(-c < \frac{\hat{\theta}(\underline{X}) - \theta_0}{\sqrt{V}} < c)$$

But $\frac{\hat{\theta}(\underline{X}) - \theta_0}{\sqrt{V}} \sim \text{Normal}(0,1)$, and so the

$$P(\text{the random interval includes the true parameter value})$$

$$= \Phi(c) - \Phi(-c) = 2\Phi(c) - 1$$

.

For a 95% interval, $c = 1.96$, and for a 68% interval, $c = 1$.

In practice, we calculate a single interval, say, $\hat{\theta}(\underline{x}) \pm 1.96\sqrt{V}$ based on a sample $\underline{x}$

Then we say we are '95% confident' the true value lies in the interval.

Approximate and exact confidence intervals

If $\hat{\theta}(\underline{X}) = \overline{X}$, then

- $V = \sigma^2/n$ where σ is the standard deviation of X.
- If σ is unknown and has to be estimated, then there are more precise confidence intervals based on 't-distributions'
- Only important for small samples

In general have to approximate $V \approx \frac{1}{n\hat{I}(\theta)}$, using the Fisher Information approximation for large samples

Average height of group

1. The standard deviation of heights in the class is 10.1 - you can assume this is a known σ
2. Take $n = 5$ heights at random
3. Calculate their mean $\overline{x}$
4. Calculate their standard error $\sigma/\sqrt{n} = 10.1/\sqrt{5} = 4.52$
5. Calculate a 68% interval $\overline{x} \pm \sigma/\sqrt{n} = \overline{x} \pm 4.52$
6. Draw your interval on the board
7. How many of the intervals include the 'true' value?

Confidence intervals for differences in means

Suppose we are to observe data

- $X_1, .., X_n$ from a distribution with unknown mean θ_X and known variance σ_X^2
- $Y_1, .., Y_m$ from a distribution with unknown mean θ_Y and known variance σ_Y^2

We want to estimate $\theta_X - \theta_Y$ and provide a confidence interval.

We will estimate $\theta_X - \theta_Y$ by $\overline{X} - \overline{Y}$

Now by the Central Limit Theorem, $\overline{X} \sim \text{Normal}\left(\theta_X, \frac{\sigma_X^2}{n}\right)$ and similarly for $\overline{Y}$

So $\overline{X} - \overline{Y} \sim \text{Normal}\left(\theta_X - \theta_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$

So a 95% confidence interval for $\theta_X - \theta_Y$ would be

$$\overline{X} - \overline{Y} \pm 1.96\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

Comparing two proportions

If we assume X is Binomial(θ_X, n), and Y is Binomial(θ_Y, m)

A 95% confidence interval for $\theta_X - \theta_Y$ would be

$$\frac{X}{n} - \frac{Y}{m} \pm 1.96\sqrt{\frac{\theta_X(1-\theta_X)}{n} + \frac{\theta_Y(1-\theta_Y)}{m}}$$

which can be estimated by

$$\frac{X}{n} - \frac{Y}{m} \pm 1.96\sqrt{\frac{\frac{X}{n}(1-\frac{X}{n})}{n} + \frac{\frac{Y}{m}(1-\frac{Y}{m})}{m}}$$

Comparing two proportionss

In the class, we observed $X = 14$ out of $n = 18$ (78%) females could roll their tongues, compared to $Y = 22$ out of $m = 31$ (71%) males.

Assume this class is a random sample from the population of the earth (!!)

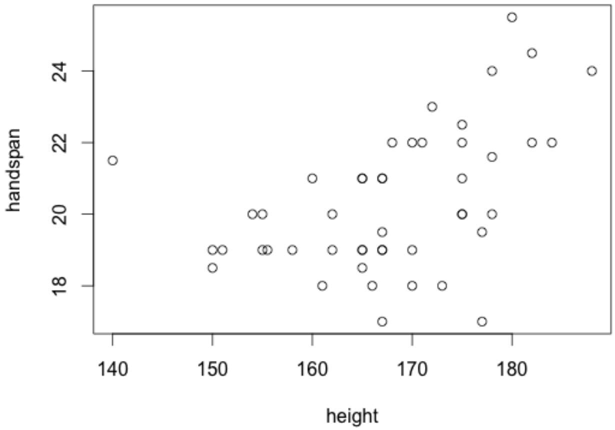
What is the difference between the % of all men who roll their tongues, and the % of all women?

Give a 95% interval.

$$\frac{X}{n} - \frac{Y}{m} \pm 1.96\sqrt{\frac{\frac{X}{n}(1-\frac{X}{n})}{n} + \frac{\frac{Y}{m}(1-\frac{Y}{m})}{m}}$$

Relating two quantities

Are height and hand-span related?



Relating two quantities

We observe a set of points $(x_1, y_1), .., (x_n, y_n)$ and we want to fit a straight line through them

We assume the x's have been centralised around their mean $\overline{x}$, and assume the model

$$y_i = a + b(x_i - \overline{x}) + \epsilon_i,$$

where the 'error' ϵ_i is assumed to have mean 0 and constant variance.

We want to estimate the coefficients and fit a line

$$y = \hat{a} + \hat{b}(x - \overline{x}).$$

$\hat{y}_i$ is the 'fitted' value for y_i, given by $\hat{y}_i = \hat{a} + \hat{b}(x_i - \overline{x})$.

$y_i - \hat{y}_i$ is the 'residual' - the distance of the data-point from the straight line

The 'least squares' solution (Gauss, 1820s), minimises the Residual Sum of Squares (RSS), where

$$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2,$$

Least squares

Therefore RSS $= \sum_i [y_i - \hat{a} - \hat{b}(x_i - \overline{x})]^2$, and to minimise we set the derivatives to 0.

$\frac{\delta}{\delta a}\mathrm{RSS} = -2\left(\sum_i [y_i - \hat{a} - \hat{b}(x_i - \overline{x})]\right) =$
$-2\left(\sum_i y_i - n\hat{a} - \hat{b}\sum_i (x_i - \overline{x})]\right)$.

The last term is 0, and setting the derivative to 0 gives $\hat{a} = \overline{y}$.

$\frac{\delta}{\delta b}\mathrm{RSS} = -2\left(\sum_i [y_i - \hat{a} - \hat{b}(x_i - \overline{x})](x_i - \overline{x})\right) =$
$-2\left(\sum_i (y_i - \overline{y})(x_i - \overline{x}) - \hat{b}\sum_i (x_i - \overline{x})^2]\right)$.

Setting the derivative to 0 gives $\hat{b} = \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.

Least squares

Interpretation: $\hat{a} = \overline{y}$ is the intercept at $\overline{x}$; $\hat{b}$ is the fitted gradient

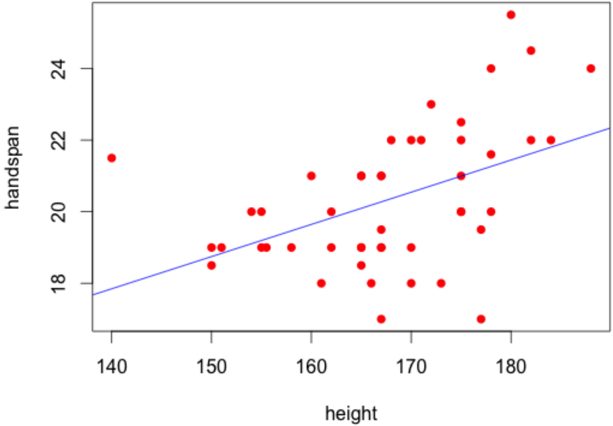
This solution does not require an assumption of normal errors

The standard deviation σ has MLE $\hat{\sigma} = \sqrt{\frac{\text{RSS}}{n}}$

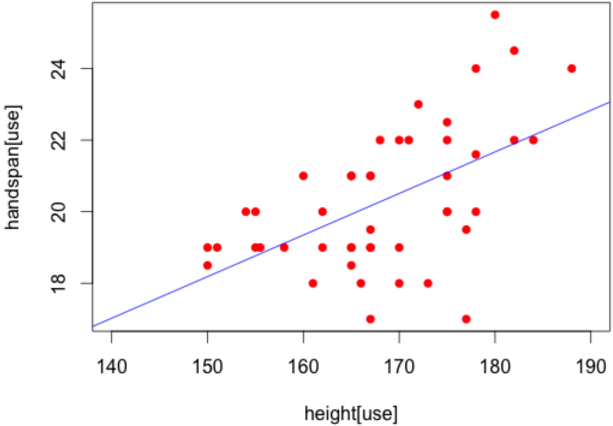
In statistics programs an 'unbiassed' estimate $\tilde{\sigma} = \sqrt{\frac{\text{RSS}}{(n-2)}}$ will be given and called the 'residual standard error'

Can get standard errors for all these estimates, and so obtain approximate intervals (can get more accurate from t distributions)

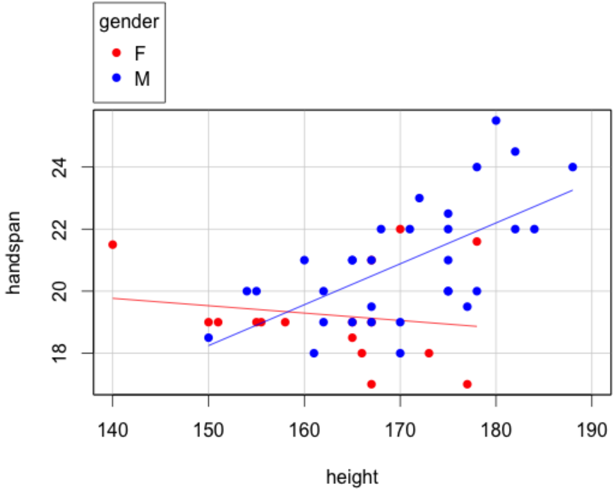
Fitted line



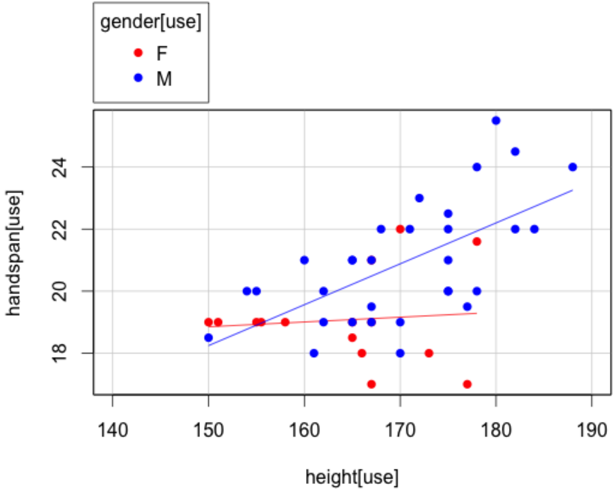
An influential datapoint?



Influence of gender?



Fitted line



gender[use]
● F
● M
handspan[use]
18
20
22
24
140
150
160
170
180
190
height[use]

Transformation to linearity

Gapminder example.

Stats practical on confidence intervals and fitting lines

You are recommended to type commands into the script window (top left) and then run them by selecting the line (you just need then cursor on the line, no need to highlight it)

1. Download `class-data3.R` into RStudio
2. Download `class-data.csv` and read it in
3. Check you understand what `sum(is.na(handspan))` is doing [use help, or Quick R, or Reference Card, to find about `is.na`]
4. Look at `is.na(handspan)` by just typing `is.na(handspan)`
5. After fitting the line, check the statistics from `summary(fitted)`. Create approximate confidence intervals for the gradient from 'estimate +/- 1.96 standard error' (more exact intervals are possible)
6. Check you understand how `use` works to select observations [creates an array of TRUE/FALSE]
7. What change is there in the gradient by removing the outlying data-point?
8. See how `scatterplot` works
9. Try changing FALSE to TRUE for smoother
10. `fitted$residuals` contains the residuals from the fitted line - draw a histogram of these. Do they look Normal?
11. Now fit a line for height against breath.
12. See if a straight line looks better with height against log(breath)