Statistics Teaching

Adam Kane

4 November 2016

```
## Warning: package 'HH' was built under R version 3.3.2
```

```
## Loading required package: lattice
```

```
## Loading required package: grid
```

```
## Loading required package: latticeExtra
```

```
## Loading required package: RColorBrewer
```

```
## Loading required package: multcomp
```

```
## Warning: package 'multcomp' was built under R version 3.3.2
```

```
## Loading required package: mvtnorm
```

```
## Warning: package 'mvtnorm' was built under R version 3.3.2
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.3.2
```

```
## Loading required package: TH.data
```

```
## Warning: package 'TH.data' was built under R version 3.3.2
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
## Loading required package: gridExtra
```

Variables

A variable is something that can take on different values e.g. height is a variable. The opposite of variables are constants e.g. the gravitational constant which has one value only.

Types of variables (NOIR)

In statistics we can consider 4 variable types as set out by Stevens (1946):

Nominal variables

are variables that have two or more categories, but which do not have an intrinsic order. For example, classifying where people live in the USA by state. In this case there will be 50 ‘levels’ of the nominal variable.

```
nominalVariables <- c("Alaska", "Florida", "New York", "Washington", "Texas")
```

Ordinal variables

are variables that have two or more categories just like nominal variables only the categories can also be ordered or ranked. So if you asked someone if they liked the policies of the Democratic Party and they could answer either “Not very much”, “They are OK” or “Yes, a lot” then you have an ordinal variable. Why? Because you have 3 categories, namely “Not very much”, “They are OK” and “Yes, a lot” and you can rank them from the most positive (Yes, a lot), to the middle response (They are OK), to the least positive (Not very much). However, whilst we can rank the levels, we cannot place a “value” to them; we cannot say that “They are OK” is twice as positive as “Not very much” for example.

```r
ordinalVariables <- c("OK", "Not very much", "OK", "Yes, a lot", "Not very much")
```

Interval variables

are variables for which their central characteristic is that they can be measured along a continuum and they have a numerical value (for example, temperature measured in degrees Celsius or Fahrenheit). So the difference between 20C and 30C is the same as 30C to 40C. However, temperature measured in degrees Celsius or Fahrenheit is NOT a ratio variable. In interval scales, addition and subtraction make sense, but multiplication and division do not. That is, 70C is not “twice as hot”" as 35C. If this is confusing, think what a negative temperature would mean, or a 0 temperature! 30C is -1 times as hot as -30C? It doesn’t make sense!

```r
intervalVariables <- c(30,31,29,30,29,33,34,35)
```

Ratio variables

are interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. So, temperature measured in degrees Celsius or Fahrenheit is not a ratio variable because 0C does not mean there is no temperature. However, temperature measured in Kelvin is a ratio variable as 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever. Other examples of ratio variables include height, mass, distance and many more. Ratio responses mean that not only is there order and spacing, but that multiplication makes sense as well. Two common examples are height and weight. A person who weighs 200 pounds weighs double what a person who weighs 100 pounds weighs.

```r
ratioVariables <- c(0:10)
```

Problematic Percentages

So, are percentages nominal, ordinal, interval or ratio? Technically, they are not even ratio - you cannot double a percentage without distorting the meaning

Levels of measurement

In general it is advantageous to treat variables as the highest level of measurement for which they qualify. That is, we could treat education level as a categorical variable, but usually we will want to treat it as an ordinal variable. This is because treating it as an ordinal variable retains more of the information carried in the data. If we were to reduce it to a categorical variable, we would lose the order of the levels of the variable. By using a higher level of measurement, we will have more options in the way we analyze, summarize, and present data.

Parametric Vs Non Parametric Tests

There is a lot of confusion about parametric vs. non-parametric statistics and tests. Some of the literature that explains the difference gets pretty technical. Here is a layman's description that might not be 100% technically accurate but that will let you understand the difference.A parameter is a characteristic of a population. We often estimate parameters with statistics that come from samples. Some common parameters and statistics are the mean, the median, the standard deviation and so on.

Some tests use these parameters. For example, every variety of the t-test uses means and standard deviations. Therefore, the t-test is called a parametric test. On the other hand, some tests do not use these parameters. For example, the Mann Whitney U test uses no parameters. Therefore, it is called a non-parametric test.

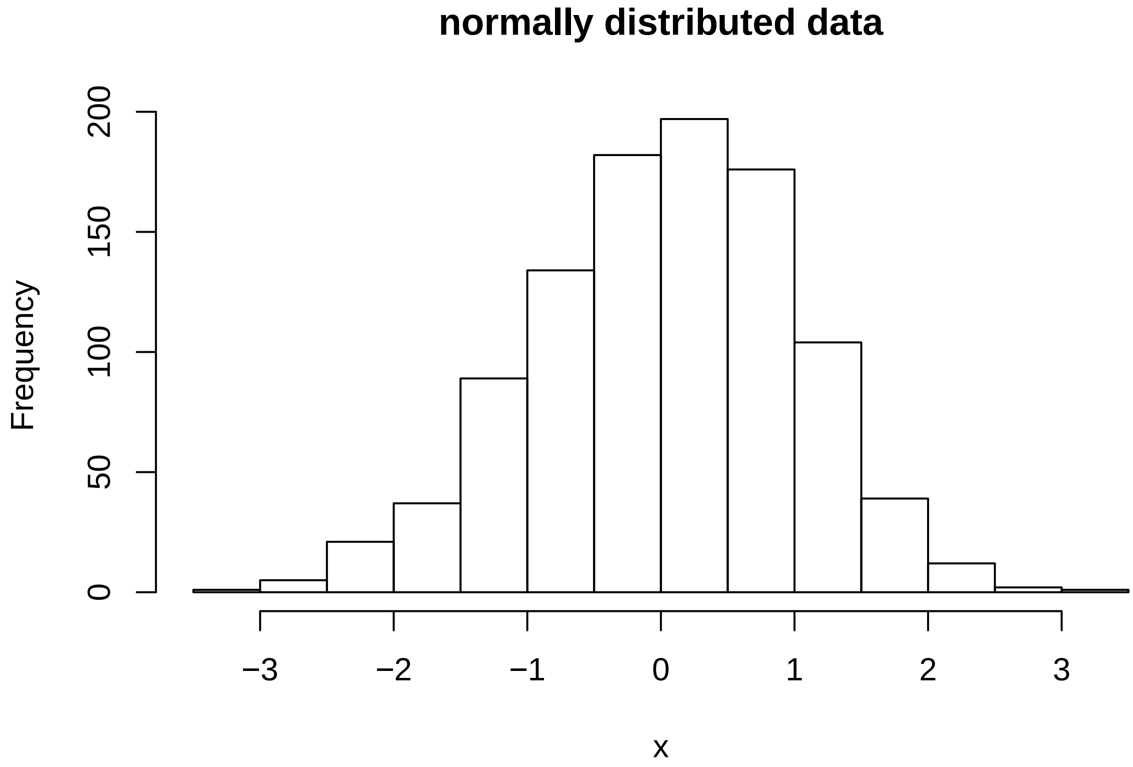
If you want to tell if a test is parametric or not, look at the formulas used in calculating it. Do they contain parameters/statistics?

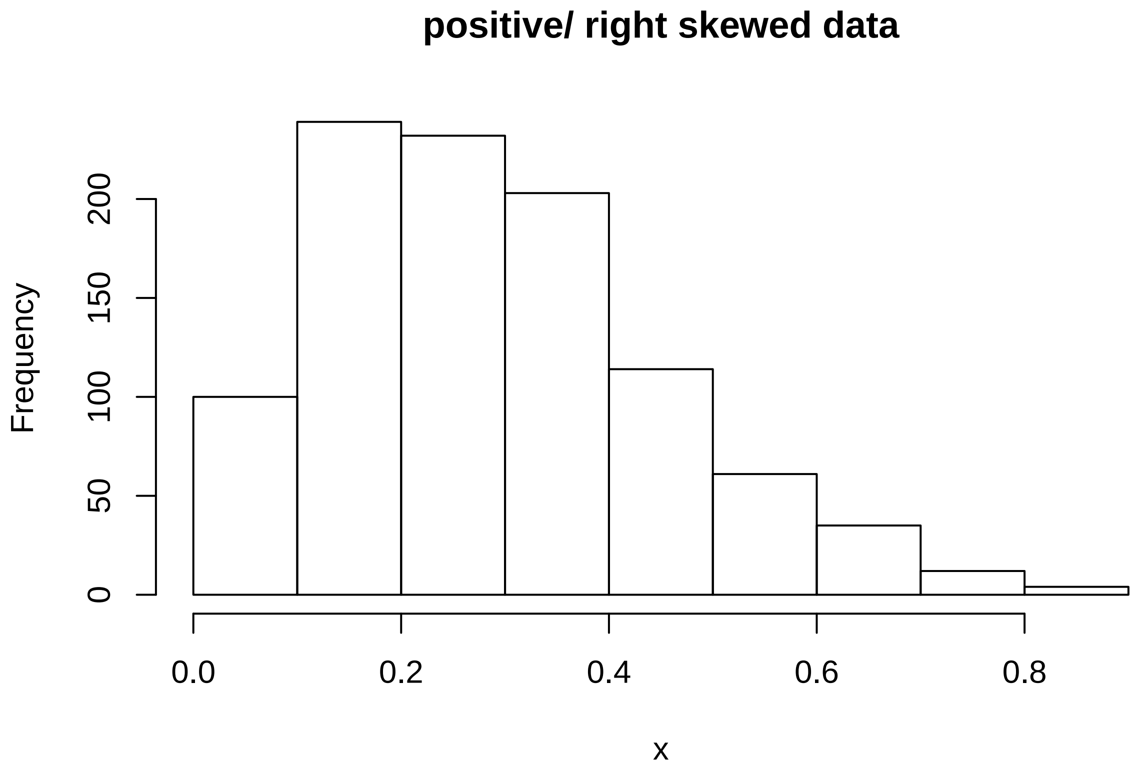
If your measurement scale is nominal or ordinal then you use non-parametric statistics

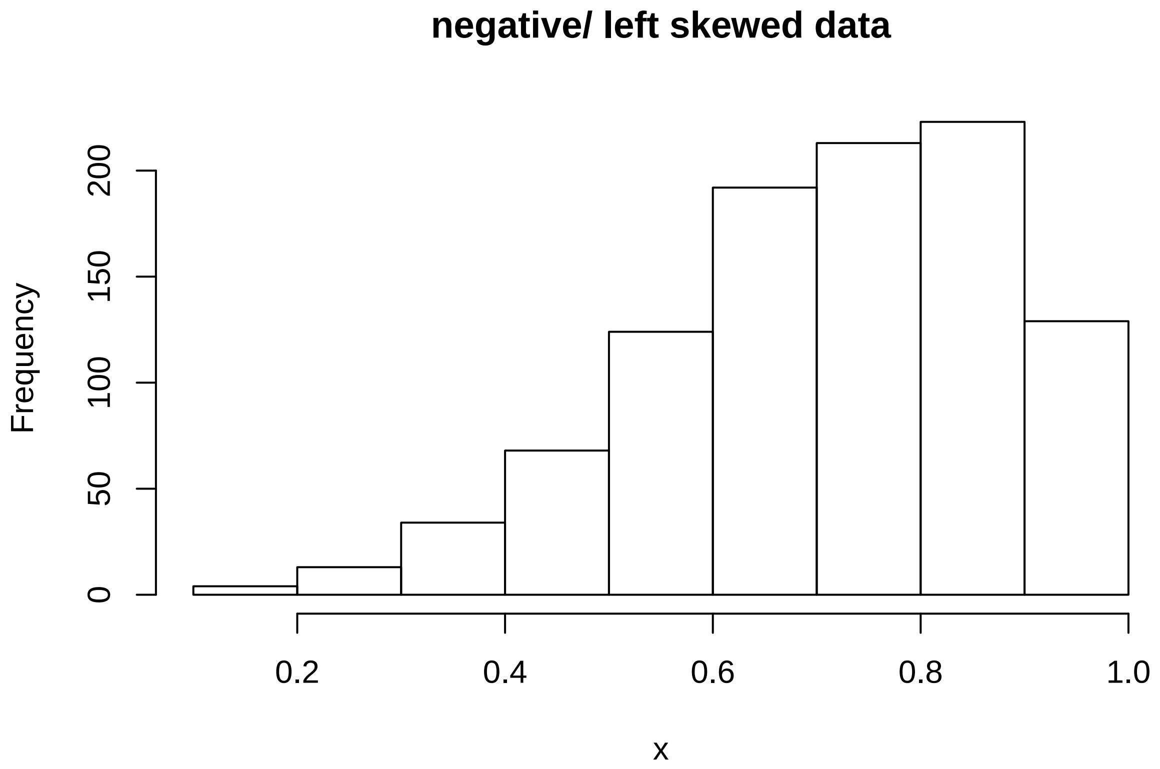
If you are using interval or ratio scales you use parametric statistics.

Histograms

A histogram is a type of graph used to display a distribution. It helps us to overcome the natural tendency to rely on summary information, such as an average. Histograms can reveal information no captured by summary statistics.




positive/ right skewed data
Frequency
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
x



Summary Statistics

Central Tendency

The mean

is a measure of central tendency this describes the middle or centre point of a distribution $mean = M = \frac{1}{n}\sum_{n}^{i=1} x$

The median

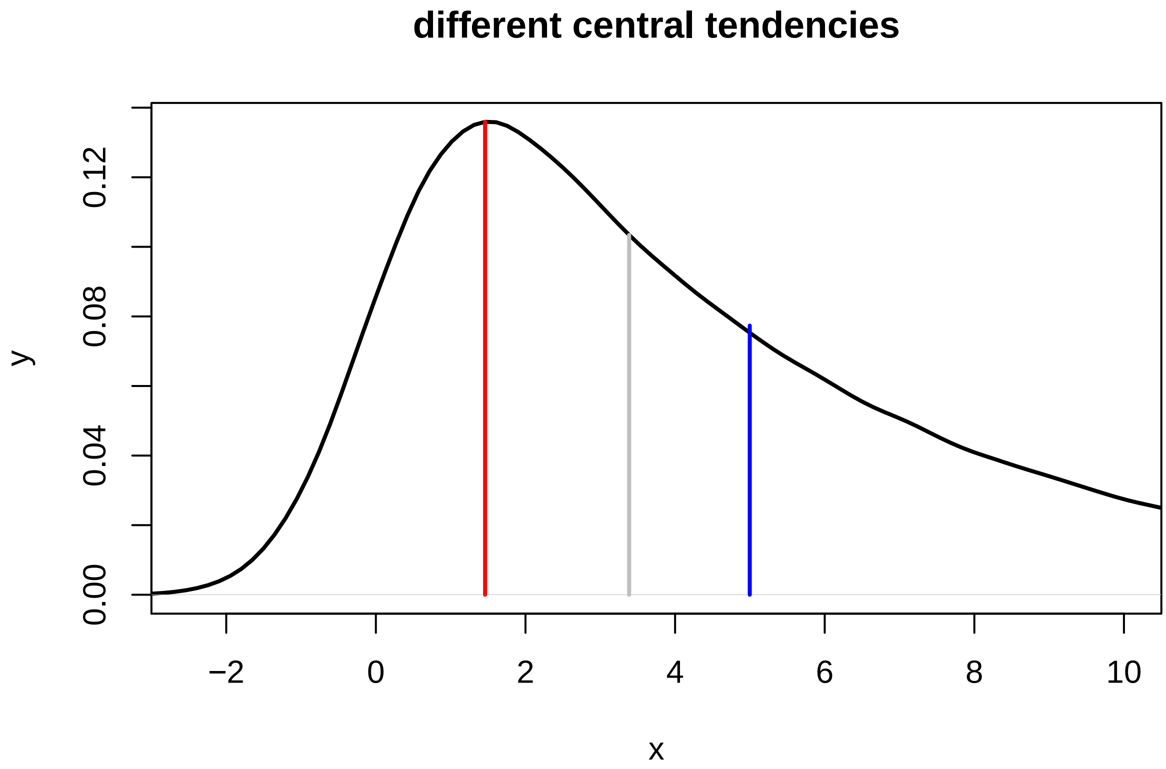
is the middle score (the score below which 50% of the distribution falls) preferred when there are extreme scores in the distribution

The mode

is the score that occurs most often in the distribution, useful for nominal variables

How distribution can affect measures of central tendency

Differing distribution may mean these three measures do not overlap Here the mean is blue, the median grey and the mode red.



Measures of Variability

a measure that describes the range and diversity of scores in a distribution

standard deviation (SD)

the average deviation from the mean in a distribution

$SD = \sqrt{\frac{[\Sigma(X-M^2)]}{N}}$ is used for descriptive statistics

$SD = \sqrt{\frac{[\Sigma(X-M^2)]}{N-1}}$ is used for inferential statistics

variance (SD^2)

sum of squared deviation scores = sum of squares are divided by the sample size

$SD^2 = \frac{[\Sigma(X-M^2)]}{N}$ is used for descriptive statistics

$SD^2 = \frac{[\Sigma(X-M^2)]}{N-1}$ is used for inferential statistics

this is also known as the mean squares

Example of Linsanity

Jeremy Lin was a basketball player who went on a scoring streak for the New York Knicks. We can calculate some summary statistics for his games.

here are the points he scored for the games he played:

```
pointsPerGame<-c(28,26,10,27,20,38,23,28,25,2)
```

we take the sum of those values and the sample size i.e. number of games he played to get the mean

```
sum(pointsPerGame)
```

```
## [1] 227
```

```
length(pointsPerGame)
```

```
## [1] 10
```

so the mean is

```
sum(pointsPerGame)/length(pointsPerGame)
```

```
## [1] 22.7
```

then the deviation scores show how much he deviated from the mean for each game i.e. it is the difference between a raw score and the mean.

```
pointsPerGame - mean(pointsPerGame)
```

```
##  [1]   5.3   3.3 -12.7   4.3  -2.7  15.3   0.3   5.3   2.3 -20.7
```

we can't get the average for the deviation scores because they sum to zero

```
deviationScores<- pointsPerGame - mean(pointsPerGame)
devsum <- sum(deviationScores)
```

```
## [1] 0
```

```
devsum/length(pointsPerGame)
```

```
## [1] 0
```

instead we square the deviation scores, sum them and divide by N to give us a score for variance.

That is to say we calculate mean squares because it is the sums of squares divided by N.

```
(pointsPerGame - mean(pointsPerGame))^2
```

```
##  [1]  28.09  10.89 161.29  18.49   7.29 234.09   0.09  28.09   5.29 428.49
```

```
devSq <- (pointsPerGame - mean(pointsPerGame))^2
devSumSq <- sum(devSq) ; devSumSq
```

```
## [1] 922.1
```

```
variance <- devSumSq/length(pointsPerGame) ; variance
```

```
## [1] 92.21
```

Squaring however does have a problem as a measure of spread and that is that the units are all squared, where as we'd might prefer the spread to be in the same units as the original data (think of squared points scored). Hence the square root allows us to return to the original units which is the standard deviation.

```r
sqrt(variance)
```

```
## [1] 9.602604
```

Standardised Scales

Z-scores

In statistics there is a standard scale the Z scale. Any score from any scale can be converted to Z scores

$Z = \frac{(X-M)}{SD}$

X = raw score, the score on the original scale

M = mean

SD = standard deviation

The mean Z-score is Z = 0

Positive Z scores are above average

Negative Z scores are below average

For example

```r
X = 99.6 # body temp for one person
M = 98.6 # the mean for the group
SD = 0.5 # the standard deviation for the group
Z=(X-M)/SD; Z
```

```
## [1] 2
```

This value of 2 means their score is 2 standard deviations above the mean

Percentile rank

The percentage of scores that fall at or below a score in a distribution Assume a normal distribution If Z = 0 then the percentile rank = 50th 50% of the distribution falls below the mean

Correlation

A statistical procedure used to measure and describe the relationship between two variables

Correlations can range between +1 and -1

+1 is perfect positive correlation

0 is no correlation (independence)

-1 is perfect negative correlation

When two variables, let's call them X and Y, are correlated, then one variable can be used to predict the other variable.

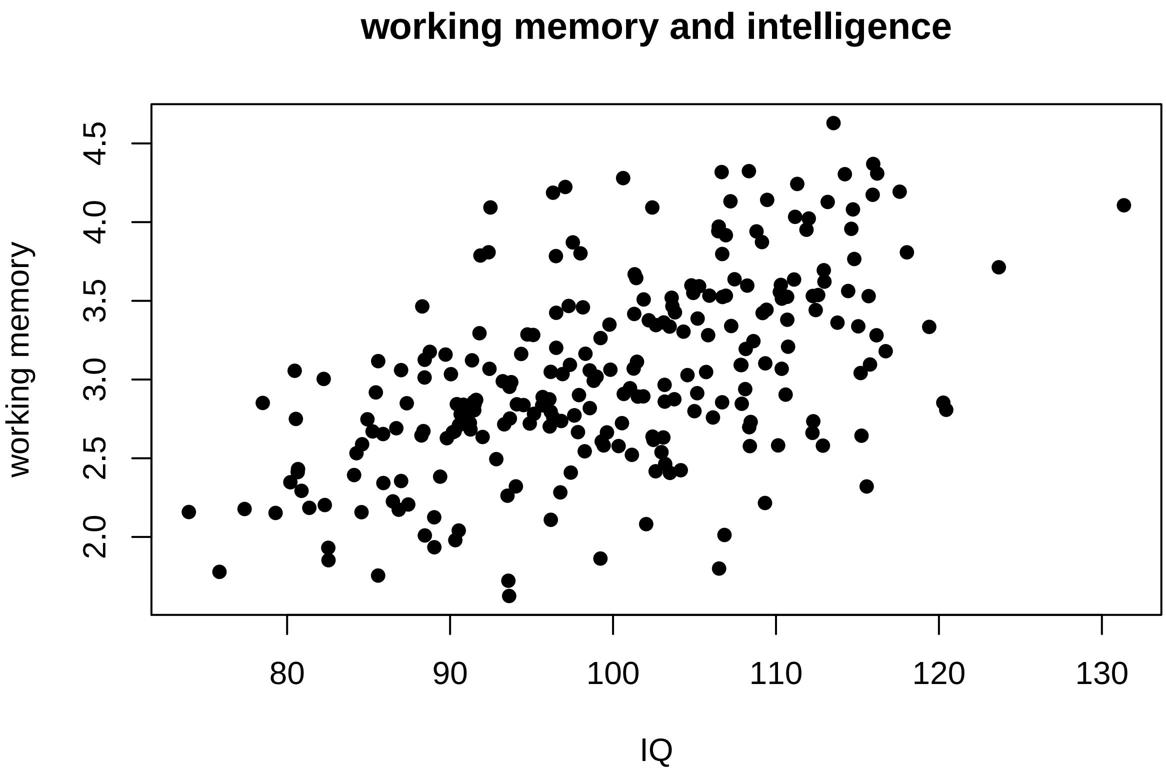
More precisely, a person's score on X can be used to predict his or her score on Y.

For example, working memory capacity is strongly correlated with intelligence, or IQ, in healthy young adults

So if we know a person's IQ then we can predict how they will do on a a test of working memeory.

We can see in this scatterplot that there is a positive correlation in our data, which is verified by the value we get for our correlation.

```
IQ <- rnorm(250, mean = 100, sd = 10)
workingMemory <- IQ*rnorm(250, mean = 3, sd = 0.5)/100
df = data.frame(IQ, workingMemory)
plot(df$workingMemory~df$IQ, xlab="IQ" , ylab="working memory", pch = 16, main = "working memory and in
```



```
cor.test(df$IQ, df$workingMemory)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$IQ and df$workingMemory
## t = 10.539, df = 248, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4641429 0.6363393
## sample estimates:
##       cor
## 0.5561822
```

Warnings about correlation

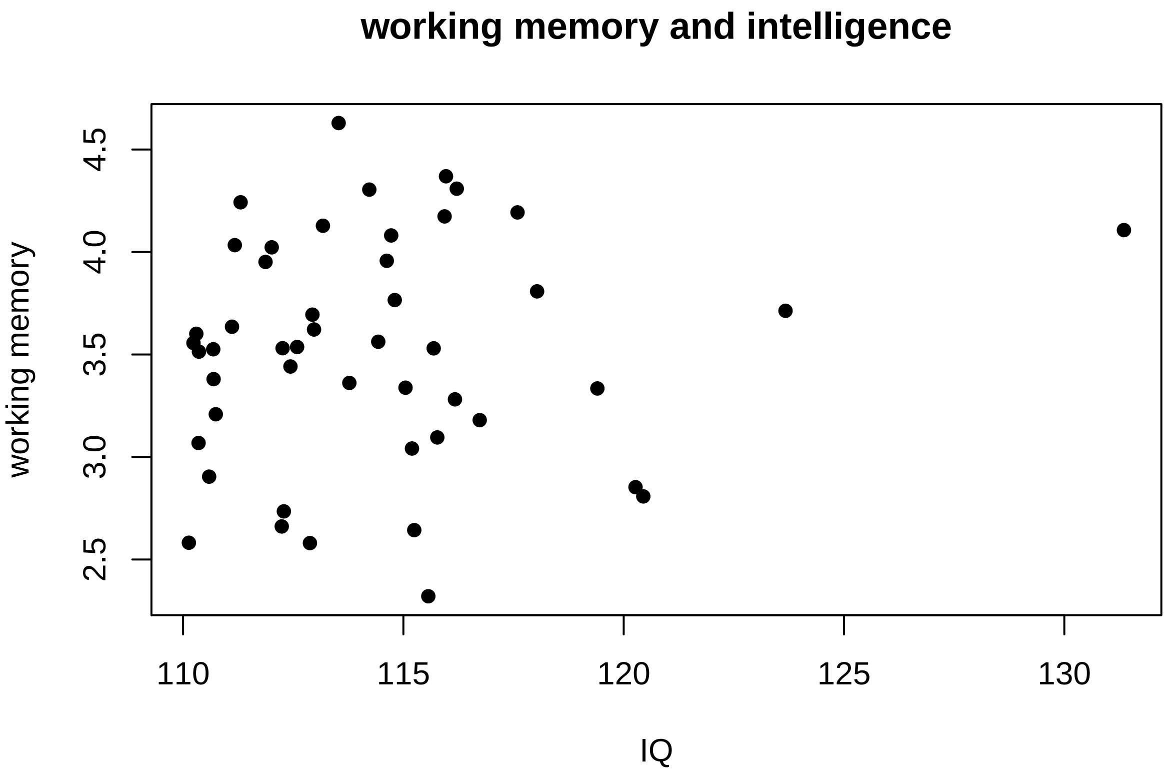
But we have to remember that correlation does not imply causation. In our example, working memory does not cause IQ and vice versa, rather there are lots of intervening variables.

The magnitude of a correlation is influenced by many factors, including: sampling (random and representative?), and the measurement of X & Y (are your measures of IQ reliable?).

When you fail to get a representative sample you can get attenutation of correlation due to a restriction of range in one of your variables. For instance, if you only select college graduates, you have preselected for higher IQ and this can reduce the correlation.

This restriction of range essentially restricts variance ultimately impacting our ability to discern covariance. In the following scatterplot and correlation measure you can see this effect.

```
dfAttenuated <- df[df$IQ >110, ]
plot(dfAttenuated$workingMemory~dfAttenuated$IQ, xlab="IQ" , ylab="working memory", pch = 16, main = "wo
```



```
cor.test(dfAttenuated$IQ, dfAttenuated$workingMemory)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dfAttenuated$IQ and dfAttenuated$workingMemory
## t = 0.74476, df = 46, p-value = 0.4602
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1805830  0.3814565
## sample estimates:
```

```
##       cor
## 0.1091523
```

Finally, a correlation coefficient is a sample statistic just like the mean and won't be representative unless the correlation coefficient is 1.

Types of Correlation

There are several types of correlation coefficients, for different variable types.

The Pearson product-moment correlation coefficient (r)

This is used when both variables, X & Y, are continuous.

The Point bi-serial correlation

This is used when 1 variable is continuous and 1 is dichotomous.

The Phi coefficient

When both variables are dichotomous

Spearman rank correlation

When both variables are ordinal (ranked data)

Focus on Pearson correlation

r = the degree to which X and Y vary together, relative to the degree to which X and Y vary independently.

r = (covariance of X & Y)/(variance of X & Y)

There are a number of ways to calculate r e.g.

the raw score formula and

the Z-score formula

Remember from our calculation for variance

$variance = SD^2 = MS = (SS/N)$

To calculate SS:

For each row, calculate the deviation score

$(X - M_x)$

Square the deviation scores

$(X - M_x)^2$

Sum the squared deviation scores

$SS_x = \Sigma[(X - M_x)^2] = \Sigma[(X - M_x) * (X - M_x)]$

Sum of Cross Products

We need to calculate the sum of cross products (SP) to get r for our correlation

for each row, calculate the deviation score on X $(X - M_x)$

For each row, calculate the deviation score on Y $(X - M_y)$

Then, for each row, multiply the deviation score on X by the deviation score on Y

$(X - M_x) * (Y - M_y)$

Then sum the "cross products"

$SP = \Sigma[(X - M_x) * (Y - M_y)]$

Covariance

Covariance measures the relationship between two variables.

Covariance = COV = SP/N

The formula for covariance is: $cov(X, Y) = \frac{\Sigma(X - M_x)(Y - M_y)}{N}$

or for inferential statistics where the denominator becomes n-1

$cov(X, Y) = \frac{\Sigma(X - M_x)(Y - M_y)}{N - 1}$

Covariance is not scaled, so it can't tell you the strength of that relationship. To account for this, correlation takes covariance and scales it by the product of the standard deviations of the two variables.

```r
cov(df$IQ,df$workingMemory) # covariance score
```

```
## [1] 3.583232
```

```r
cor.test(df$IQ,df$workingMemory) # correlation score
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$IQ and df$workingMemory
## t = 10.539, df = 248, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4641429 0.6363393
## sample estimates:
##       cor
## 0.5561822
```

Raw score formula

The raw score formula is thus:

$r = \frac{SP_{xy}}{\sqrt{(SS_x * SS_y)}}$

remember this r value is the degree to which X and Y vary together, relative to the degree to which X and Y vary independently.

In longer form:

$SP = \Sigma[(X - M_x) * (Y - M_y)]$

$SS_x = \Sigma[(X - M_x)^2] = \Sigma[(X - M_x) * (X - M_x)]$

$SS_y = \Sigma[(Y - M_y)^2] = \Sigma[(Y - M_y) * (Y - M_y)]$

So the raw score formula to calculate the correlation coefficient r can be written out in two ways:

$r = \frac{SP_{xy}}{\sqrt{(SS_x * SS_y)}}$

or,

$r = \frac{\Sigma[(X-M_x)*(Y-M_y)]}{\sqrt{(\Sigma(X-M_x)^2*\Sigma(Y-M_y)^2)}}$

Z-score formula

is the sum of the product of the Z-scores divided by N

$r = \frac{\Sigma(Z_x * Z_y)}{N}$

first we need to calculate the Z-scores

$Z_x = \frac{(X-M_x)}{SD_x}$

$Z_y = \frac{(Y-M_y)}{SD_y}$

where

$SD_x = \sqrt{\frac{(\Sigma(X-M_x)^2}{N}}$

$SD_y = \sqrt{\frac{(\Sigma(Y-M_y)^2}{N}}$

Proof of equivalence

here the denominator is the standard deviation

$Z_x = \frac{(X-M_x)}{\sqrt{\frac{(\Sigma(X-M_x)^2}{N}}}$

$Z_y = \frac{(Y-M_y)}{\sqrt{\frac{(\Sigma(Y-M_y)^2}{N}}}$

unpacked to it's full long form

here we have Z_x multiplied by Z_y divided by N

$$r = \frac{\frac{(X-M_x)}{\sqrt{\frac{(\Sigma(X-M_x)^2}{N}}} * \frac{(Y-M_y)}{\sqrt{\frac{(\Sigma(Y-M_y)^2}{N}}}}{N}$$

we can pack all this back together using some algebra

$r = \frac{\Sigma[(X-M_x)*(Y-M_y)]}{\sqrt{(\Sigma(X-M_x)^2*\Sigma(Y-M_y)^2)}}$

which can be simplified further to:

$r = \frac{SP_{xy}}{\sqrt{(SS_x * SS_y)}}$

which is the raw score formula.

Variance and covariance

Variance = MS = SS/N

Covariance = COV = SP/N

Correlation is standardised covariance

it's standardised so the value is in the range -1 to +1

Note on the denominators

Correlation for descriptive statistics

Divide by N

Correlation for inferential statistics

Divide by N-1

Assumptions of Correlation

let's consider Pearson correlation

Assumptions when interpreting r:

Normal distributions for X and Y

- how to detect violations?

Plot histograms and examine summary stats

Linear relationship between X and Y

- how to detect violations?

Examine scatterplots

Homoscedasticity

- how to detect violations?

Examine scatterplots

Reliability of X and Y

Validity of X and Y

Random and representative sampling

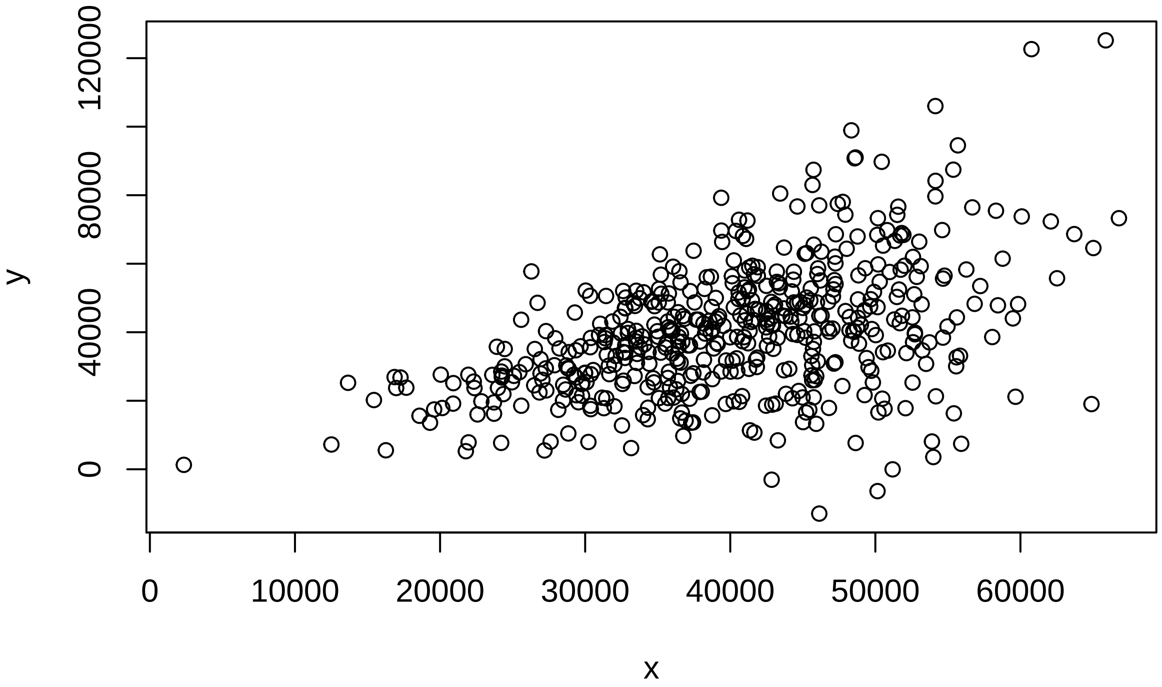
Homoscedasticity and heteroscedasticity

In a scatterplot the vertical distance between a dot and the regression line reflects the amount of predicition error (known as the "residual")

The idea of Homoscedasticity is that those residuals are not related to X. The residuals should be chance errors and not systematic.

If the residuals are related to X then we suspect some sort of confound in our study. This is termed Heteroscedasticity.

A classic example of heteroscedasticity is that of income versus expenditure on meals. As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals. Those with higher incomes display a greater variability of food consumption.



Anscombe's Quartet

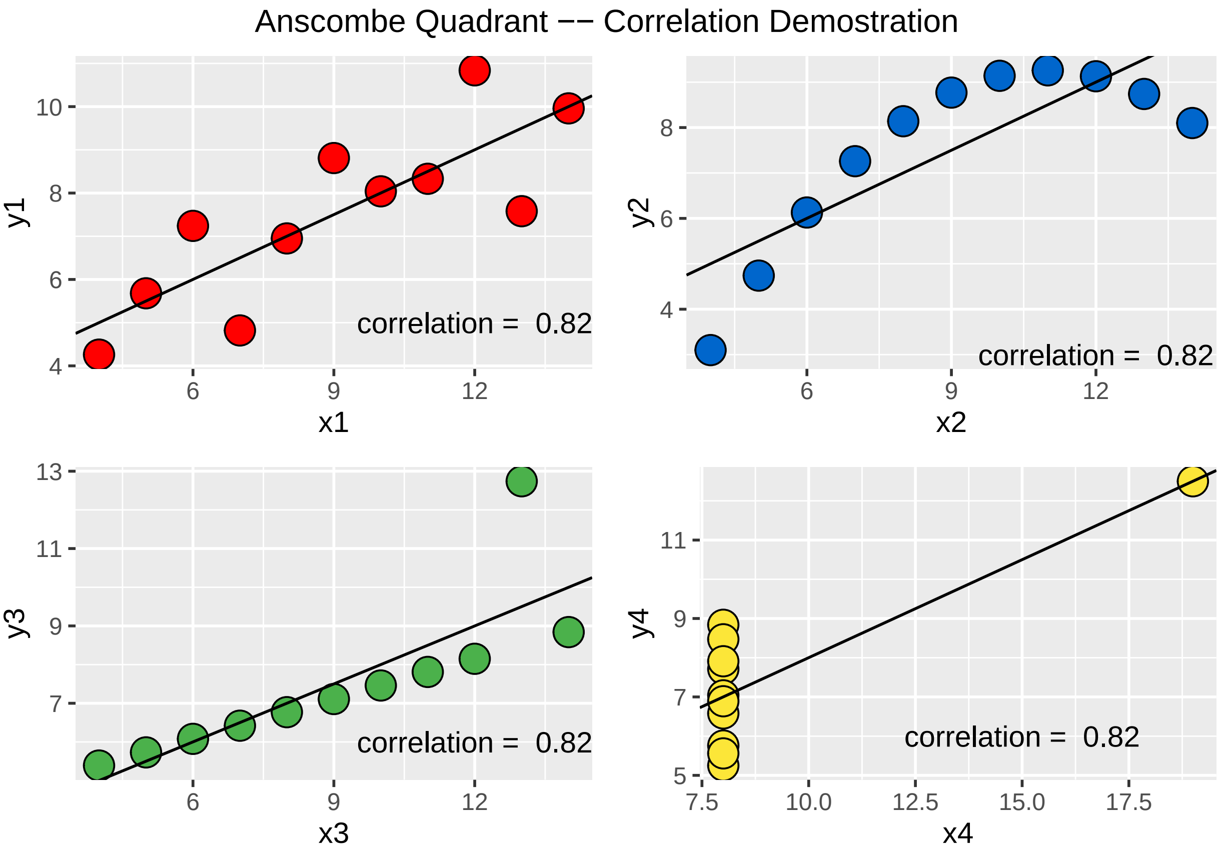
Why it's critical to look at your scatterplots.

Four datasets where the correlation is exactly the same. The datasets also have the same variance. But clearly there are differences in these datasets.

```
##        x1              x2              x3              x4
##  Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8
##  1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8
##  Median : 9.0   Median : 9.0   Median : 9.0   Median : 8
##  Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9
##  3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8
##  Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19
##        y1               y2              y3              y4
##  Min.   : 4.260   Min.   :3.100   Min.   : 5.39   Min.   : 5.250
##  1st Qu.: 6.315   1st Qu.:6.695   1st Qu.: 6.25   1st Qu.: 6.170
##  Median : 7.580   Median :8.140   Median : 7.11   Median : 7.040
##  Mean   : 7.501   Mean   :7.501   Mean   : 7.50   Mean   : 7.501
##  3rd Qu.: 8.570   3rd Qu.:8.950   3rd Qu.: 7.98   3rd Qu.: 8.190
##  Max.   :10.840   Max.   :9.260   Max.   :12.74   Max.   :12.500
## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:latticeExtra':
##
##     layer
```



Measurement

Reliability - do we have reliable measurements?

If I step on a scale multiple times do I get the same weight?

But some values are harder to evaluate for reliability. Raw scores are imperfect, e.g. body temperature is suscpetible to systematic bias and chance error.

Classical test theory states that, in a perfect world, it would be possible to obtain a "true score" rather than a "raw score" (X)

X = true score + bias + error

A measure (X) is considered reliable as it approaches the true score

The problem is we don't know the true score so we estimate reliability

Methods to estimate reliability

Test/re-test

Parallel tests

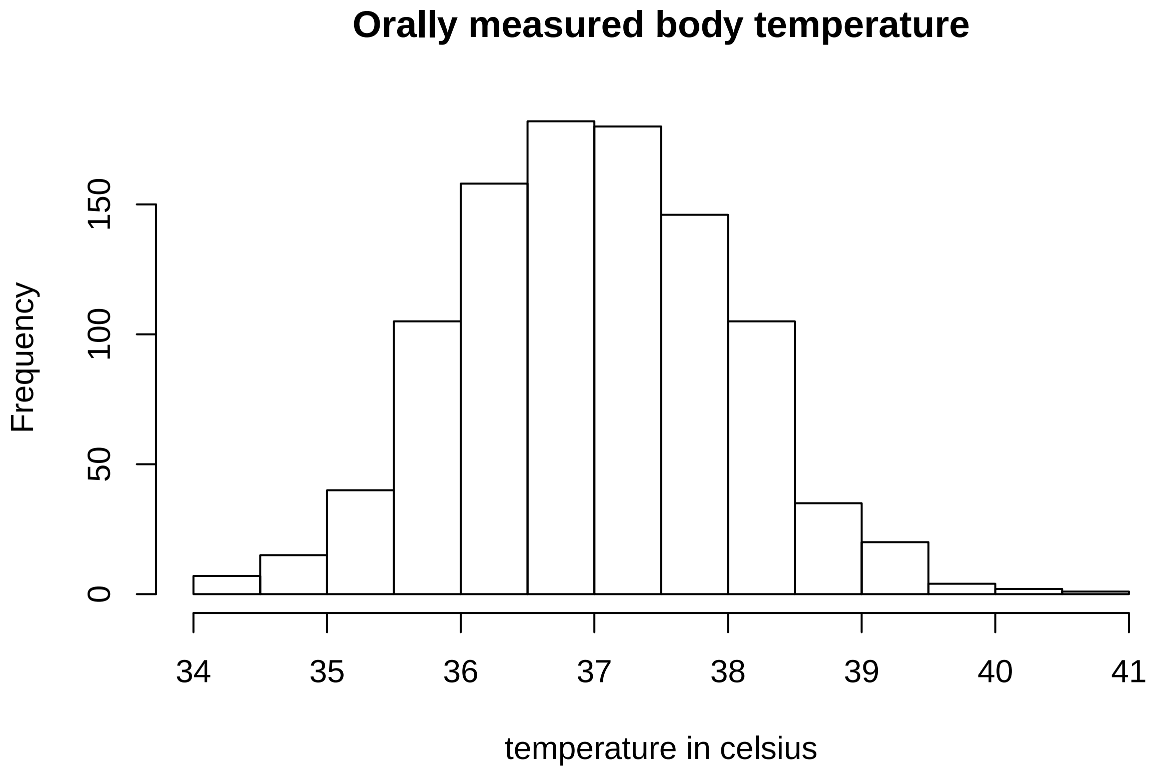
Inter-item reliability

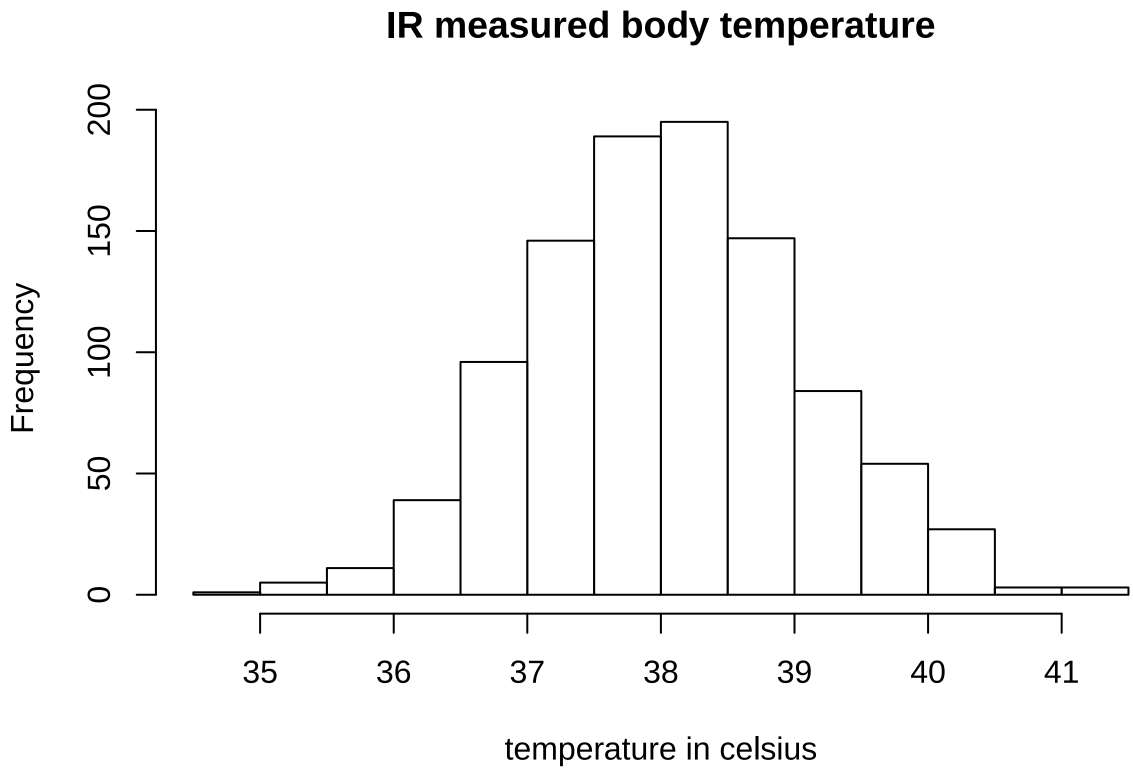
Example of body temperature

Can be measured in 3 ways,

Orally, internally, IR wand.

The IR wand has a systematic bias in that it always tends to record a higher temperature.





Test/re-test

One way to get a reliability estimate

Measure everyone twice so we'll have data for X1 and X2

Should be a strong correlation between the two measures otherwise you don't have a reliable measure

However, if the bias is uniform then we won't detect it with the test/re-test method. So in the case of the IR thermometer reading a little high, test/re-test won't work. The correlation will be high even though there is a bias.

Parallel tests

Measure body temp with the wand (X1) and with the oral thermometer(X2)

The correlation betweeen X1 and X2 is an estimate of reliability.

AND, now the bias of the wand will be revealed because you have two tests rather than one.

Inter-item

the most commonly used method in the social sciences because the focus is usually on human subjects who are difficult to work with.

Test/re-test and parallel tests are time consuming

Inter-item is therefore more cost efficient.

For example, suppose a 20 item survey is designed to measure extraversion

Randomly select 10 items to get sub-set A (X1)

The other items become sub-set B (X2)

Now we have two assessments of extraversion built into one overall survey.

If they're all getting at one personality trait then there should be a correlation between X1 and X2 which would represent an estimate of reliability. #Measurement ##Validity What is a construct?

An ideal "object" that is not directly observable as opposed to "real" observable objects

For example, "intelligence" is a construct.

How do we operationalise a construct?

The process of defining a construct to make it observable and quantifiable e.g. intelligence tests.

Construct validity

How do we assess the validity of a construct?

Let's take an example of a construct: verbal ability in children

We might operationalise this construct by using a vocabulary test.

Content validity

In the case of the vocabulary test we would ask does the test consist of words that children in the population and sample know?

Convergent validity

Does the test correlate with other, established measures of verabal ability e.g. with reading comprehension.

Divergent validity

Does the test correlate less well with measures designed to test a different type of ability e.g. spatial ability, or even more extreme, the height of the student where there should very little correlation.

Nomological validity

Are scores on the test consistent with more general theories, e.g. for child development and neuroscience. In that case a child with neural damage or disease to brain regions associated with language development should score lower on the test.

Sampling

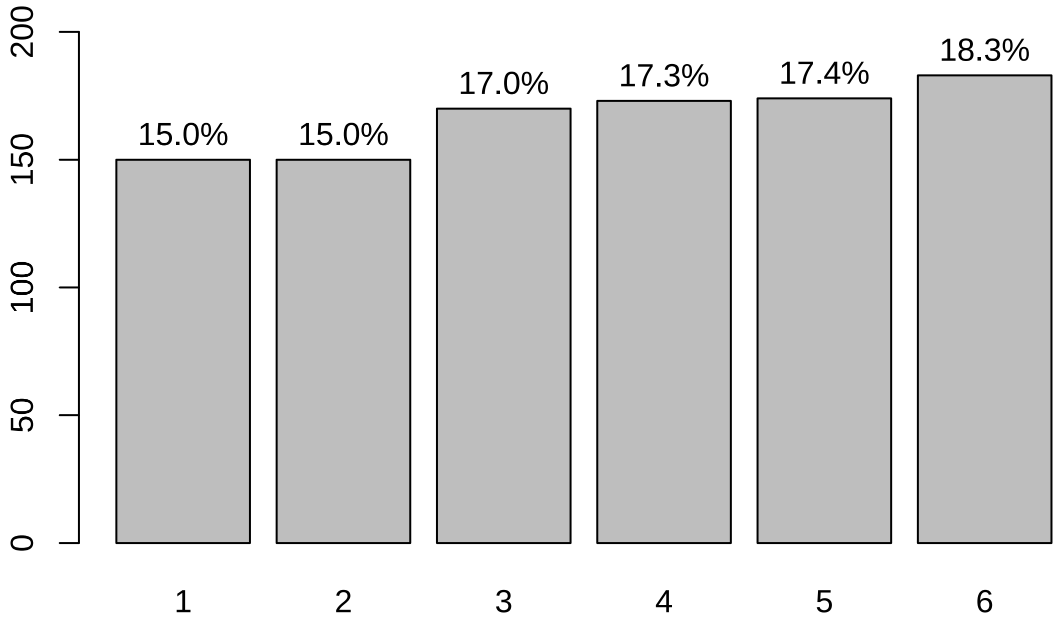
Remember, we want our sample to be random and representative.

Let's set up a die and roll it a 1000 times, then plot the results on a histogram

```
p.die <- rep(1/6,6)
sum(p.die)
```

```
## [1] 1
```

```
die <- 1:6
s <- table(sample(die, size=1000, prob=p.die, replace=T))
lbls = sprintf("%0.1f%%", s/sum(s)*100)
barX <- barplot(s, ylim=c(0,200))
text(x=barX, y=s+10, label=lbls)
```



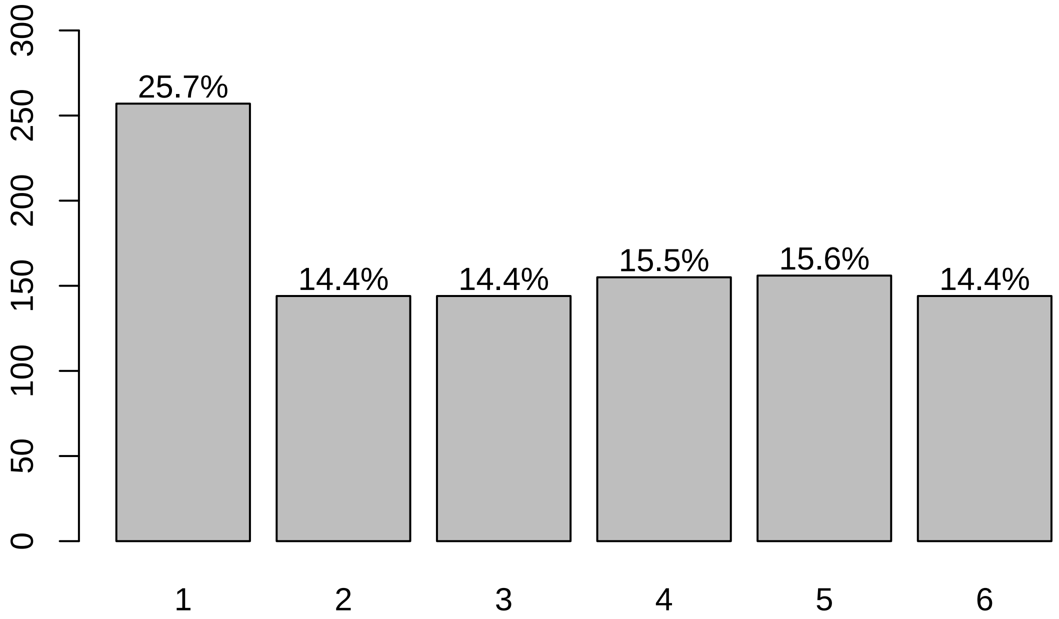
As expected each of the 6 sides came up approximately the same number of times.

We can do something similar but with a loaded die

```
p.die <- c(0.25,0.15,0.15,0.15,0.15,0.15)
sum(p.die)
```

```
## [1] 1
```

```
die <- 1:6
s <- table(sample(die, size=1000, prob=p.die, replace=T))
lbls = sprintf("%0.1f%%", s/sum(s)*100)
barX <- barplot(s, ylim=c(0,300))
text(x=barX, y=s+10, label=lbls)
```



Sampling error

The difference between the population and the sample.

This is important because we can't get everyone/everything in the popualtion.

Notice that the random histogram is not perfectly random

There is some fluctuation due to sampling error.

Problem

We don't know the population parameters

So, how do we estimate sampling error?

Estimating sampling error

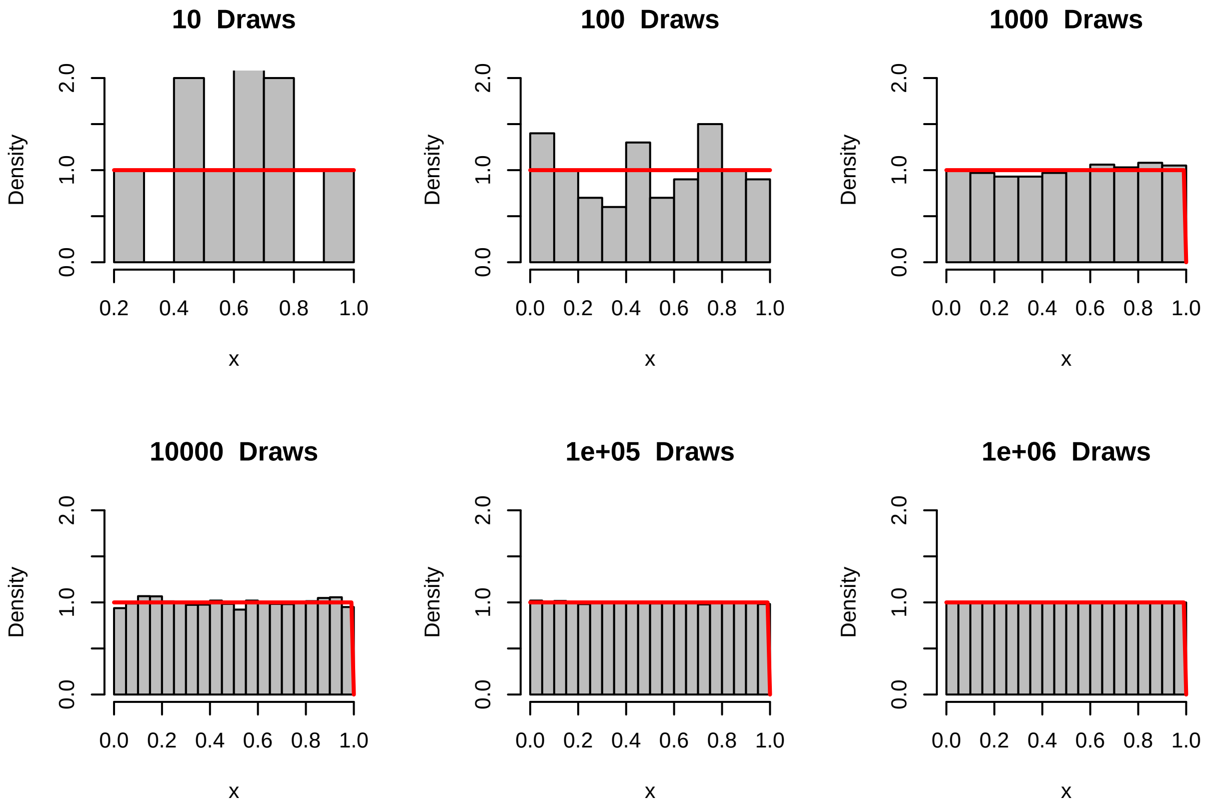
Sampling error mainly depends on the size of the sample, relative to the size of the population.

as sample size increases, sampling error decreases.

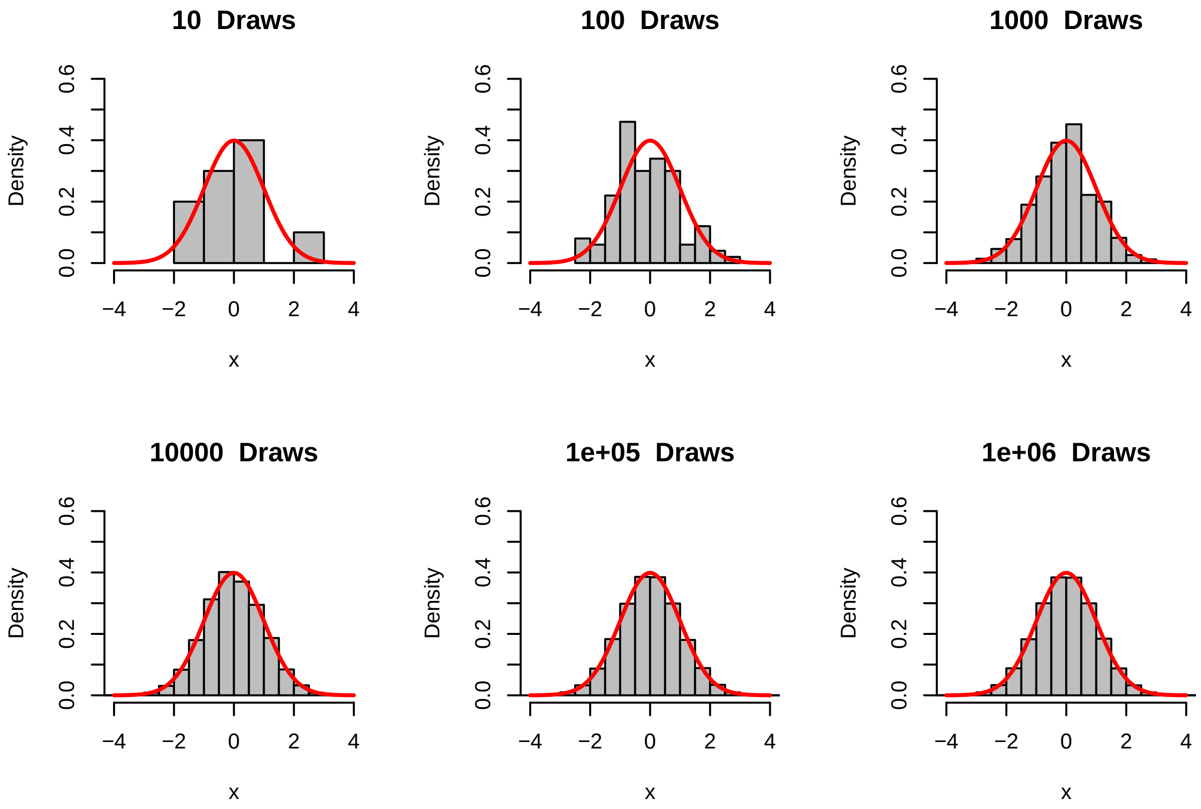
It also depends on the variance in the population (which we don't know).

as variance increases, sampling error increases.

The following histogram shows how the distribution becomes more uniform as the sample size increases illustrating the effect of sample size on sampling error.



We can check out the same effect by sampling from the normal distribtion using increasing sample sizes. As we can see the distribution becomes more normal as the sample size goes up.



Estimating sampling error

Sampling error is estimated from the size of the sample and the variance in the sample. This is under the assumption that the sample is random and representative of the population.

Standard error

Standard error is an estimate of the average amount of sampling error

$SE = \frac{SD}{\sqrt{N}}$

SE = standard error

SD = standard deviation of the sample

N = Size of the sample

We can see from here that as the sample size in the denominator increases then the standard error is going to rise.

By contrast, as the standard deviation increases (which should reflect the population standard deviation) so too will the standard error.

Here's an example of taking a sample from a population that is not very variable. Note the low standard error.

```
population<-rnorm(1000,mean=0,sd=1)
sd(population)
```

```
## [1] 1.001635
```

```
sampleDraw<-sample(population,10)
sd(sampleDraw)
```

```
## [1] 1.077219
```

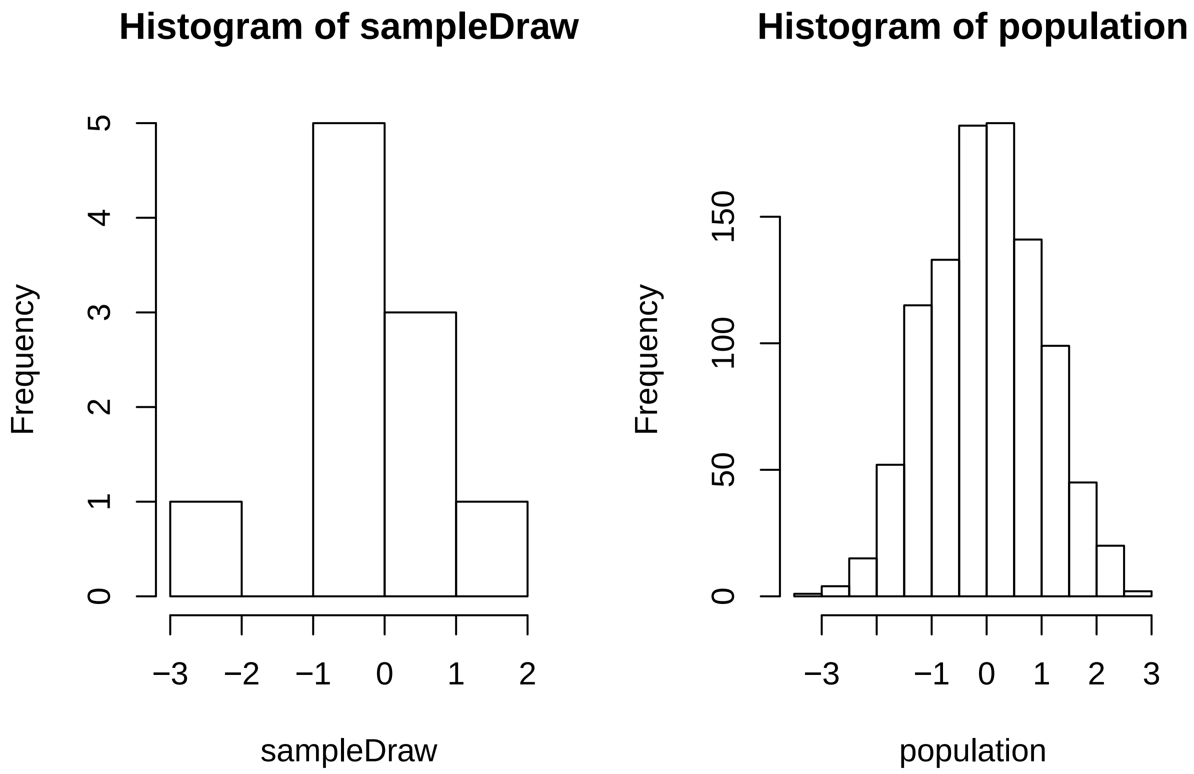
```
length(sampleDraw)
```

```
## [1] 10
```

```
sd(sampleDraw)/sqrt(length(sampleDraw))
```

```
## [1] 0.3406467
```

```
par(mfrow=c(1,2))
hist(sampleDraw)
hist(population)
```



Here's an example of taking a sample from a population that is variable. Note the higher standard error.

```
population<-rnorm(1000,mean=0,sd=10)
sd(population)
```

```
## [1] 9.71732
```

```
sampleDraw<-sample(population,10)
sd(sampleDraw)
```

```
## [1] 9.436989
```

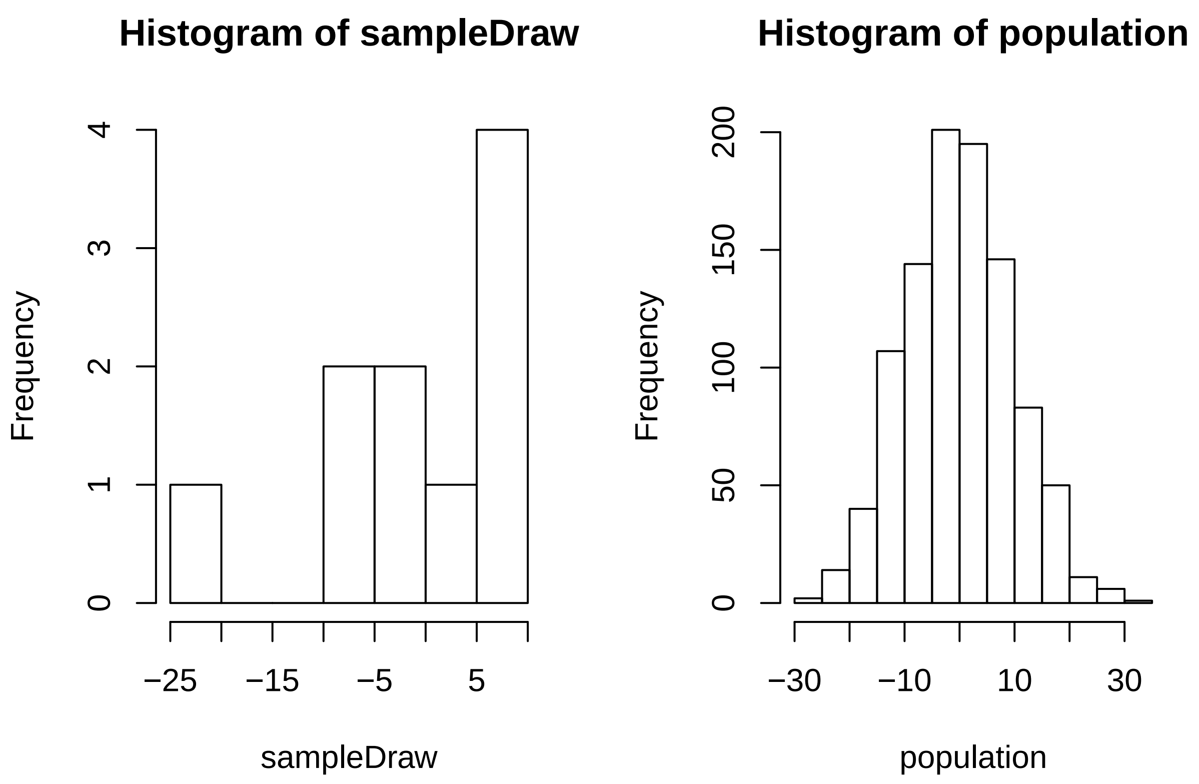
```
length(sampleDraw)
```

```
## [1] 10
```

```
sd(sampleDraw)/sqrt(length(sampleDraw))
```

```
## [1] 2.984238
```

```
par(mfrow=c(1,2))
hist(sampleDraw)
hist(population)
```



Regression

A regression is a statistical analysis used to predict scores on an outcome variable, based on scores on one or multiple predictor variables

simple regression: one predictor variable

multiple regression: multiple predictor variables

Regression equation

$Y = m + bX + e$ this is the equation of a line

Y = a linear function of X m = intercept b = slope e = residual error other notation, more commonly used for statistics $Y = B_0 + B_1 X_1 + e$

Y = a linear function of X_1 $B - 0$ = intercept = regression constant B_1 = slope = regression coefficient e = residual error

Model R and R^2

R = multiple correlation coefficient

R = $r_{y'y}$

The correlation between the predicted scores and the observed scores

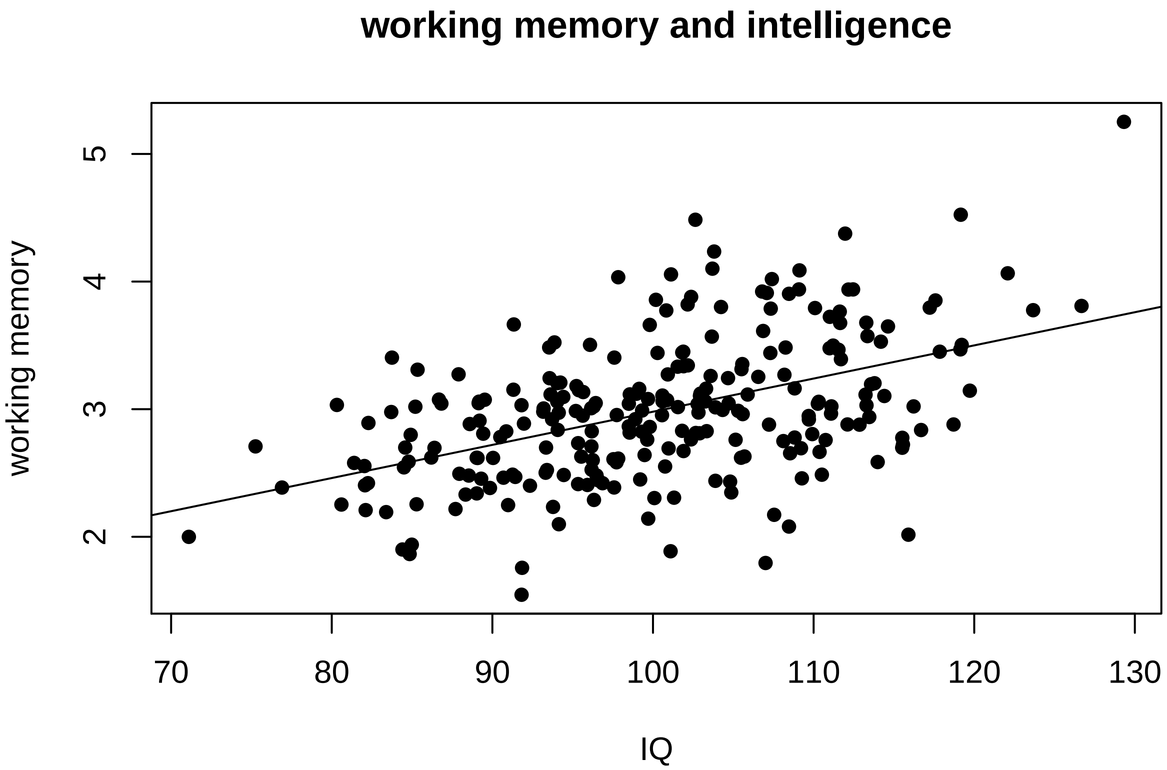
R^2 the percentage of variance in Y explained by the model.

Simple regression example

```r
set.seed(20)
IQ <- rnorm(250, mean = 100, sd = 10)
workingMemory <- IQ*rnorm(250, mean = 3, sd = 0.5)/100
df = data.frame(IQ, workingMemory)
plot(df$workingMemory~df$IQ, xlab="IQ" , ylab="working memory", pch = 16, main = "working memory and in
cor.test(df$IQ, df$workingMemory)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$IQ and df$workingMemory
## t = 8.6622, df = 248, p-value = 6.037e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3806483 0.5718293
## sample estimates:
##       cor
## 0.4819546
```

```r
model1 <- lm(df$workingMemory~df$IQ)
abline(model1)
```



```r
summary(model1)
```

```
##
## Call:
## lm(formula = df$workingMemory ~ df$IQ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37683 -0.32246 -0.00195  0.29800  1.50948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.38105    0.30251   1.260    0.209
## df$IQ        0.02599    0.00300   8.662 6.04e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4886 on 248 degrees of freedom
## Multiple R-squared:  0.2323, Adjusted R-squared:  0.2292
## F-statistic: 75.03 on 1 and 248 DF,  p-value: 6.037e-16
```

The estimate of the slope is 0.02599, so with every one unit increase in X there is a 0.02599 increase in Y.

The estimate of the intercept is 0.38105, the predicted score on Y when X = 0.

Thus the regression equation is

$Y = 0.38105 + 0.02599(X)$

In a simple regression the R^2 value 0.2323 is equal to the correlation coefficient 0.4819546 to the power of 2.

```r
0.4819546^2
```

```
## [1] 0.2322802
```

The goal with regression is to produce better models so we can generate more accurate predictions

Add more predictor variables and/or

develop better predictor variables.

Multiple Regression

Add in another predictor variables

$Y = B_0 + B_1Y_1 + B_2Y_2 + e$

Now we need to solve for B_0 & B_1 & B_2

```r
set.seed(20)
IQ <- rnorm(250, mean = 100, sd = 10)
workingMemory <- IQ*rnorm(250, mean = 3, sd = 0.5)/100
salary <- IQ*rnorm(250, mean = 500, sd = 0.1)
df = data.frame(IQ, workingMemory, salary)
model2 <- lm(df$workingMemory~df$IQ+df$salary)
summary(model2)
```

```
## 
## Call:
## lm(formula = df$workingMemory ~ df$IQ + df$salary)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.41027 -0.31297 -0.03154  0.29565  1.46744 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  0.434903   0.303143   1.435   0.1527  
## df$IQ        2.410854   1.428697   1.687   0.0928 .
## df$salary   -0.004771   0.002858  -1.669   0.0963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4868 on 247 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2347 
## F-statistic: 39.18 on 2 and 247 DF,  p-value: 1.663e-15
```

The linear combination of two predictors can do better at predicting the outcome than any one predictor by itself.

Calculation of regression coefficients

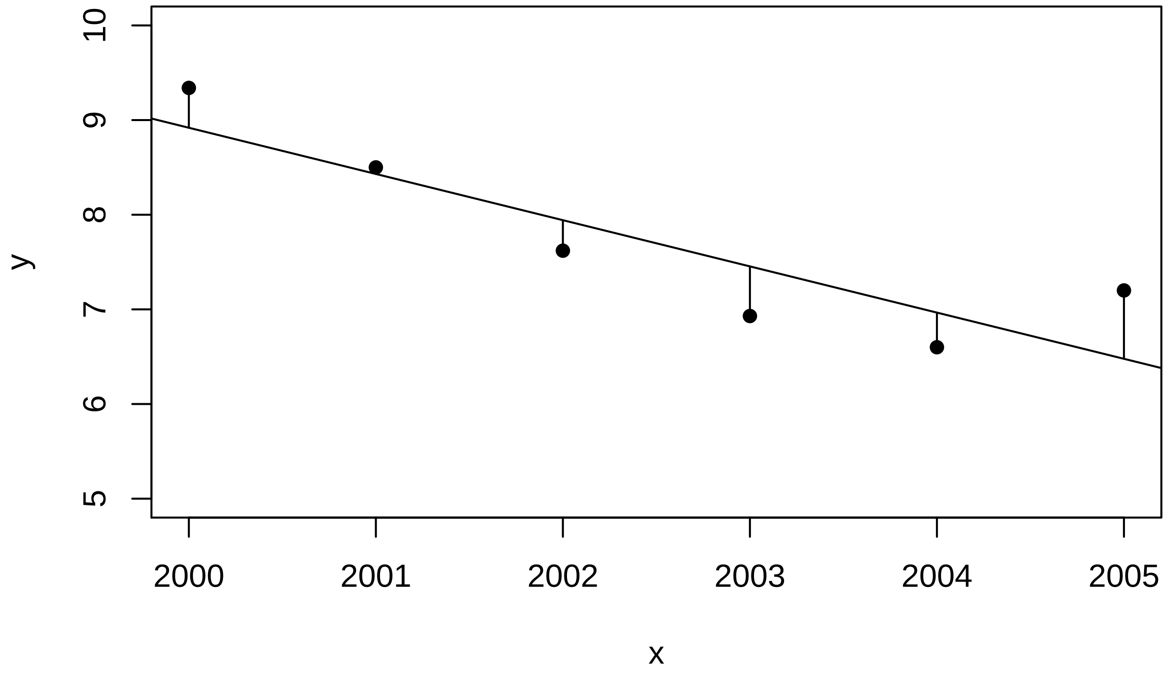
Regression equation $Y = B_0 + B_1X_1 + e$

$\hat{Y} = B_0 + B_1X_1$ where $\hat{Y}$ is the predicted score for Y

$Y - \hat{Y} = e(residual)$

The values of the coefficients (e.g. B_1) are estimated such that the regression model yields optimal predictions. What we want to do is minimise the residuals i.e. minimise the prediction error.

```
x<- c(2000,2001,2002,2003,2004,2005)
y<-c(9.34,8.50,7.62,6.93,6.60,7.2)
m1<-lm(y~x)
fitted<-predict(lm(y~x))
plot(x,y, pch =16, xlim = c(2000,2005), ylim = c(5,10))
abline(m1)
for (i in 1:6) lines(c(x[i],x[i]),c(y[i],fitted[i]))
```



ordinary least squares estimation

Minimise the sum of the squared (SS) residuals

SS.Residual = $\Sigma(Y - \hat{Y})^2$

The best fit slope is found by rotating the line until the SS.Residual is minimised. This gets the maximum likelihood estimate of the slope.

Visual approach

We have the sum of squared deviation scores (SS) in variable Y = SS.Y

We also have the sum of squared deviation scores (SS) in variable Y = SS.Y

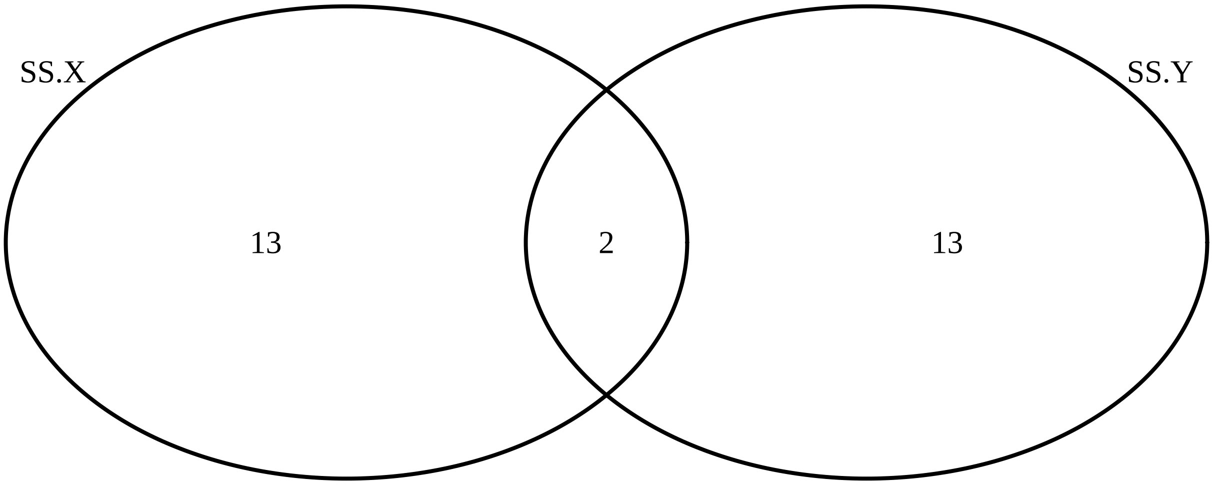
The overlap between these two is the sum of cross products between X and Y i.e. SP.XY. So the degree to which the two variables correlate will be a measure of how much overlap there is between the two.

Here's an example with low overlap and thus low correlation.

```
## Warning: package 'VennDiagram' was built under R version 3.3.2
```

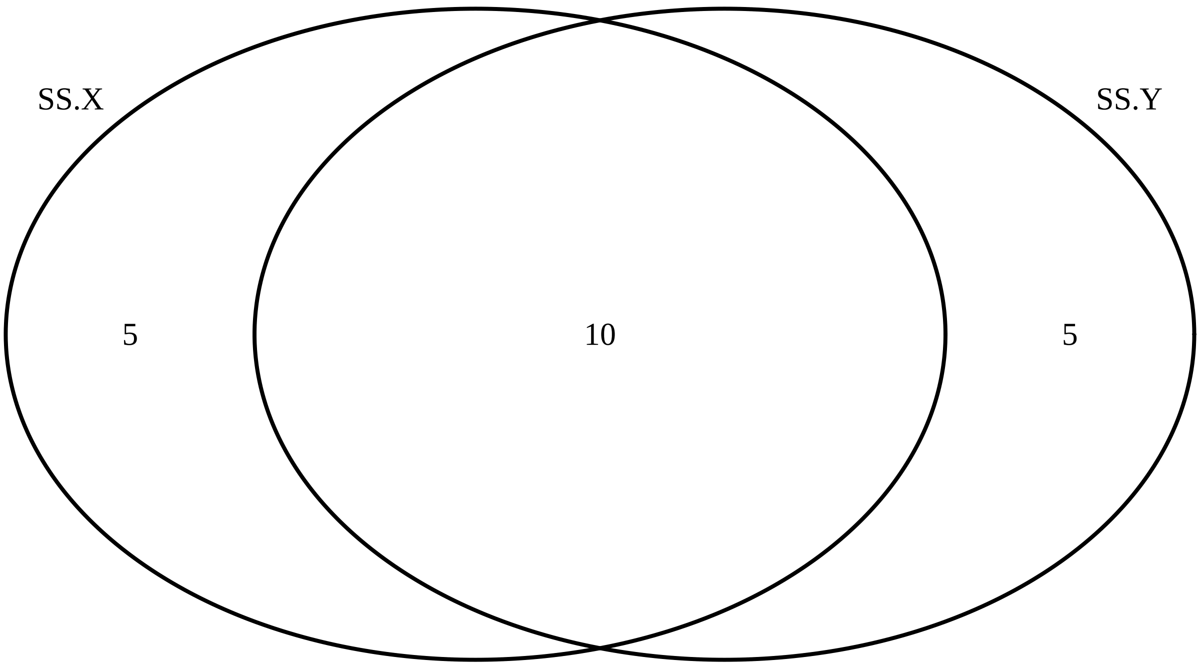
```
## Loading required package: futile.logger
```

```
## Warning: package 'futile.logger' was built under R version 3.3.2
```



```
## (polygon[GRID.polygon.176], polygon[GRID.polygon.177], polygon[GRID.polygon.178], polygon[GRID.polyg
```

Here's an example with high overlap and thus high correlation.



```
## (polygon[GRID.polygon.185], polygon[GRID.polygon.186], polygon[GRID.polygon.187], polygon[GRID.polyg
```

SP.XY can be thought of as the sum of squares of the model i.e. sum of cross products = SS of the model = SP.XY = SS.Model

Some of the variance in Y is explained by the model and some of it is unexplained, that's the residual. SS.Residual = (SS.Y - SS.Model)

Formula for the unstandardised coefficient

In a simple linear regression

$B_1 = r * (\frac{SD_y}{SD_x})$

where r is the correlation coefficient.

We divide by the standard deviations because we need to take into account the scale of Y and the scale of X. Y may be much more variable than X for instance so this division deals with that.

Formula for the standardised coefficient

Where everything is in Z-scores

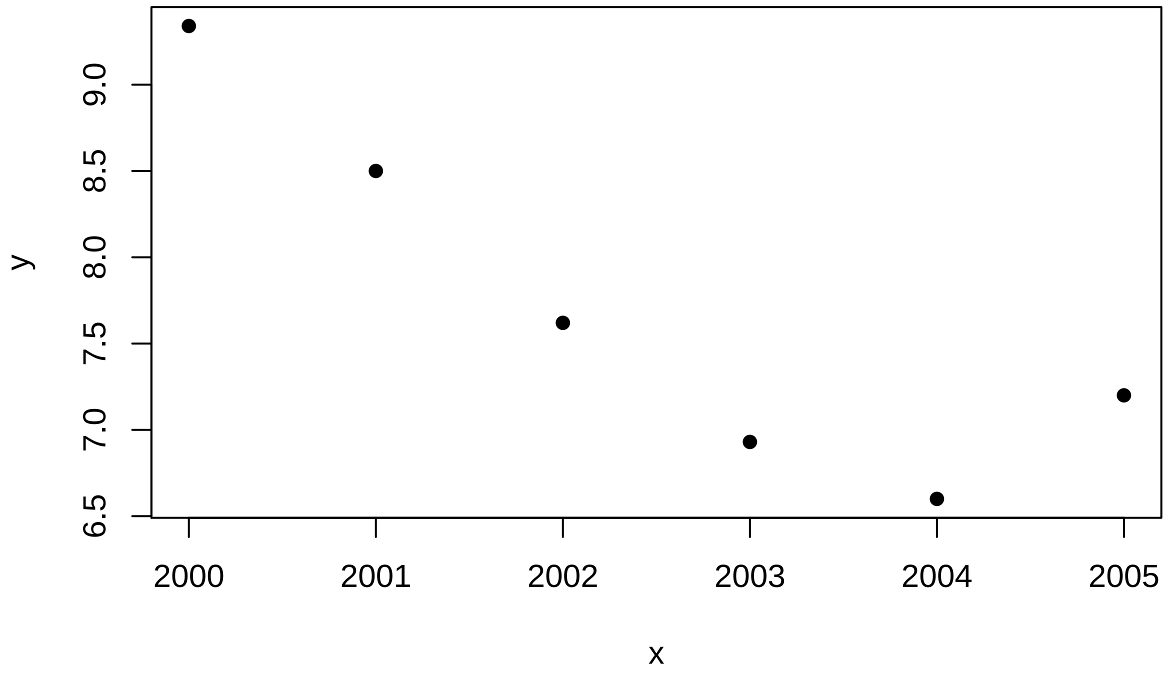
$SD_y = SD_x = 1$

$B = r * (\frac{SD_y}{SD_x})$

$\beta = r$

This is only true for simple linear regression.

```
x<- c(2000,2001,2002,2003,2004,2005)
y<-c(9.34,8.50,7.62,6.93,6.60,7.2)
plot(x,y, pch = 16)
```



```
# take the Z scores for x and y
Zx <- (x-mean(x))/sd(x)
Zy <- (y-mean(y))/sd(y)

# get the correlation for X and Y
cor(Zx,Zy)
```

```
## [1] -0.8799209
```

```
m1<-lm(Zy~Zx)
# Get the slope of y as a function of x
coef(m1)[2]
```

```
##         Zx
## -0.8799209
```

```
# also recall from earlier that the R squared value of a simple linear regression is equal to the corre
summary(m1)$r.squared
```

```
## [1] 0.7742609
```

```
cor(Zx,Zy)^2
```

```
## [1] 0.7742609
```

The correlation gives you a bounded measurement that can be interpreted independently of the scale of the two variables. The closer the estimated correlation is to ±1, the closer the two are to a perfect linear relationship. The regression slope, in isolation, does not tell you that piece of information.

The regression slope gives a useful quantity interpreted as the estimated change in the expected value of Y for a given value of X. Specifically, $\hat{\beta}$ tells you the change in the expected value of Y corresponding to a 1-unit increase in X. This information can not be deduced from the correlation coefficient alone.

Assumptions of linear regression

Normally distributed residuals

Linear relationship between X and Y

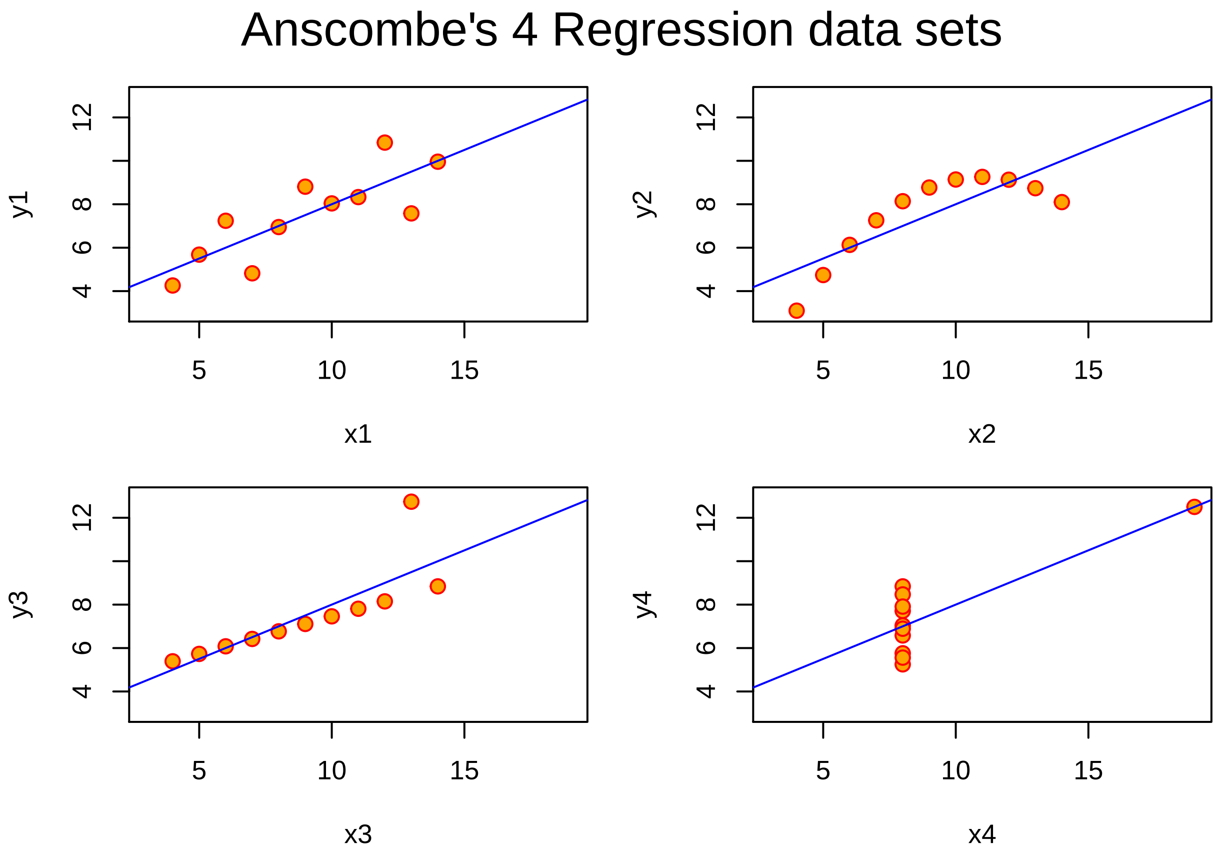
Homoscedasticity

Return to Anscombe's Quartet

The following graphs and their regression coefficients show how similar the four data sets are even for linear regression models.The regression equation for each is approximately:

$\hat{Y} = 3 + 0.5(X_1)$

```
##                   lm1      lm2       lm3       lm4
## (Intercept) 3.0000909 3.000909 3.0024545 3.0017273
## x1          0.5000909 0.500000 0.4997273 0.4999091
```



But only the top left panel looks suitable for a linear regression model. The others look like they don't satisfy the various assumptions of linear regression.

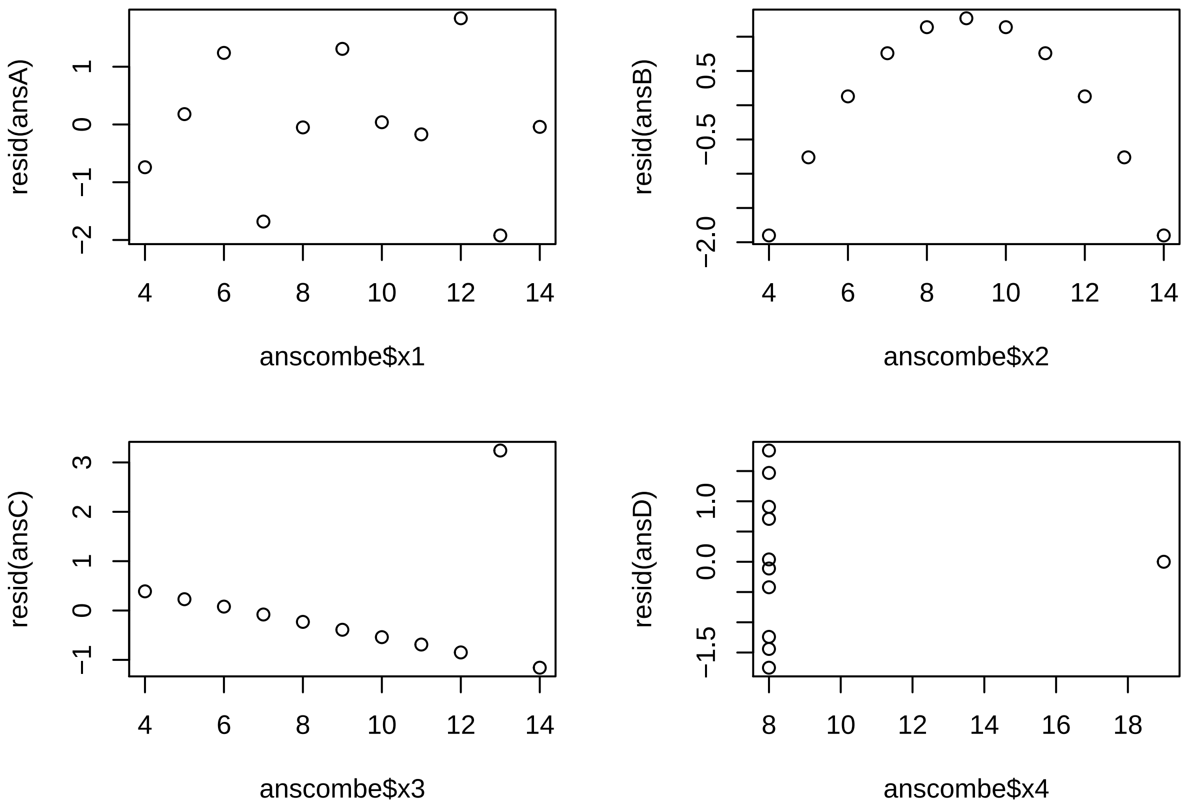
In order to test this we can save the residuals of each model.

$Y = B_0 + B_1 + e$

where

$e = (Y - \hat{Y})$

We can then look at the residuals as a function of the X predictor variable. Then we examine a scatterplot with the X variable on the X-axis and the residuals on the Y-axis.



The plot in the top left is what we are looking for. We don't want any pattern in our residual plot and this suggests no systematic error. The other plots suggest there is a relationship between X and the residual.

Null Hypothesis Significance Testing (NHST)

NHST is a procedure for hypothesis testing. We can consider NHST as a game where step 1 is the identification of the nully hypothesis and the alternative hypothesis.

Step 1

for a correlational study

H_0 = null hypothesis, e.g. r = 0

H_A = alternative hypothesis, e.g. r > 0

where r is the correlation coefficient

or for a regression model

H_0 = null hypothesis, e.g. B = 0

H_A = alternative hypothesis, e.g. B > 0

where B is the slop of the regression model

If the alternative hypothesis predicts the direction of the relationship between X & Y (positive Vs negative) it is termed a directional test (aka a one-tailed test)

Alternatively we could be agnositc and not have any idea about the direction of the relationship. In this case it would be a non-directional test (aka a two-tailed test)

The non-directional test for a regression model would be set up like this:

H_0 = null hypothesis, e.g. B = 0

H_A = alternative hypothesis, e.g. B != 0

Step 2

Assume H_0 is true, then calculate the probability of observing data with these characteristics, given that H_0 is true. This can be confusing because it's the opposite way you'd approach a study. For instance, Jonas Salk didn't predict his vaccination would have no effect.

$p = P(D|H_0)$

The probability of the data given the null hypothesis is true. This is the p-value. If the p-value is very low, then reject H_0, else retain H_0

4 possible outcomes of NHST

Either the null is true or it's false and then, as scientists, we have to successfully pick this out.

```
##                 Retain Null                  Reject Null
## Null is true    Correct decision             Type 2 error (miss)
## Null is false Type 1 error (false alarm) Correct decision
```

NHST Overview

$p = P(D|H_0)$

Given that the null hypothesis is true, the probability of these, or more extreme data, is p.

This does not mean the probability of the null hypothesis being true in p.

In other words, $P(D|H_0)! = P(H_0|D)$

NHST so far

for correlation - is the correlation significantly different from zero?

B - is the slope of the regression line for X significantly different from zero?

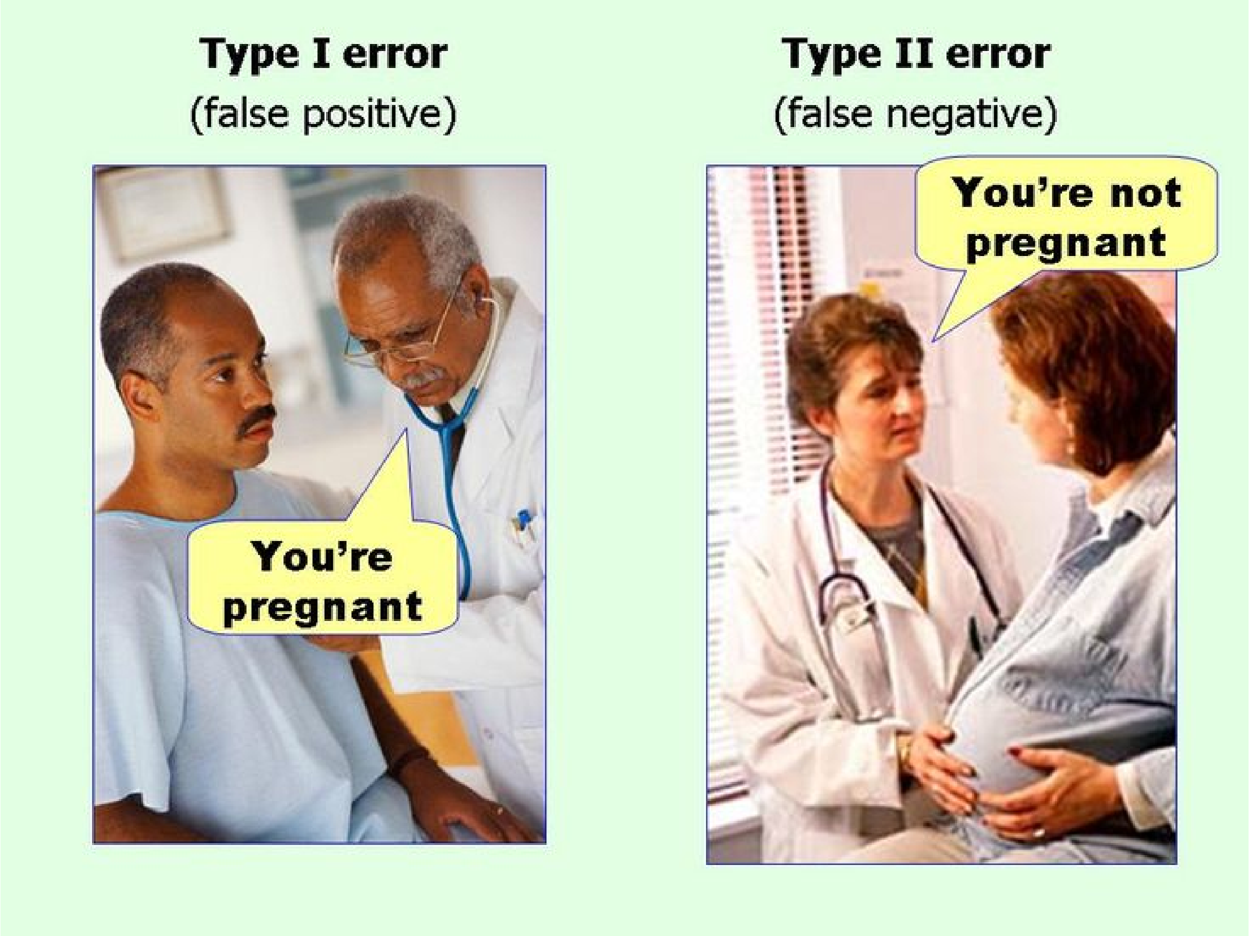

Type I error
(false positive)
You're pregnant
Type II error
(false negative)
You're not pregnant

Figure 1:

NHST for B - the slope of the regression line

t = B/SE

B = the unstandardised regression coefficient

SE = standard error

SE = $\sqrt{\frac{SS.Residual}{N-2}}$

NHST Problems and Remedies

1 - Biased by Sample Size

The p-value you get is based on a t-value and this t -value is affected by the standard error. N is in the denominator of the standard error and standard error is in the denominator of the t-value. Thus, if sample size goes up the standard error will go down, and if the standard error goes down the t-value will go up regardless of the slope value.

t = B/SE

B = the unstandardised regression coefficient

SE = standard error

SE = $\sqrt{\frac{SS.Residual}{N-2}}$

```
x<-rnorm(10000,mean=100,sd=10)
y<-rnorm(10000,mean=100,sd=10)*x
df<-data.frame (x,y)
smallSample<-df[sample(nrow(df), 10), ]
m1<-lm(smallSample$y~smallSample$x)
pVal <- anova(m1)$'Pr(>F)'[1];pVal
```

```
## [1] 0.150482
```

```
bigSample<-df[sample(nrow(df), 100), ]
m2<-lm(bigSample$y~bigSample$x)
pVal <- anova(m2)$'Pr(>F)'[1];pVal
```

```
## [1] 2.122148e-15
```

2 - Arbitrary Decision Rule

The cut-off value (alpha) is arbitrary

$p < 0.05$ is considered standard but still arbitrary.

Problems arise when p is close to 0.05 but not less than 0.05. p-hacking etc.

3 - Yokel local test

Many researchers use NHST because it#s the only approach they know. NHST encourages weak hypothesis testing.

4 - Error prone

Type 1 errors - The probability of Type 1 errors increases when researchers conduct multiple NHSTs, especially when it's on the same dataset. Have to correct for these multiple tests.

Type 2 errors - Many fields of research are plagued by a large degree of sampling error because we can only get a relatively small sample relative to the population, which makes it difficult to detect an effect, even when the effect exists.

Shady Logic

Modus tollens operates like this:

If p then q

Not q

Therefore, not p

Equivalently, in the language of statistics:

If the null hypothesis is correct, then these data can not occur

The data have occurred

Therefore, the null hypothesis is false

But in NHST the language is more probabilistic than this:

If the null is correct, then these data are highly unlikely.

These data have occurred

Therefore, the null is highly unlikely

To take an equivalent example:

If a person plays football, then he or she is probably not a professional player

This person is a professional player

Therefore, he or she probably does not play football.

NHST Remedies

Remedy for Bias by sample size

Supplement all NHSTs with estimates of effect size to get at the magnitude of the effect. For example, in regression, report standardised regression coefficients and the model R-squared.

Remedy for Arbitrary decision rule

Again supplement all NHSTs with estimates of effect size to get at the magnitude of the effect. Also, avoid phrases such as "marginally signifcant" or "highly significant".

Remedy for Yokel local test

Learn other forms of hypothesis testing. Consider multiple alternative tests and use model selection.

Remedy for NHST being error prone

Replicate significant effects to avoid long-term impact of type 1 errors

Obtain large and representative samples to avoid type 2 errors.

Remedy for Shady logic

Simply remember, $p = P(D|H_0)$

Or avoid NHST, and instead,

Report confidence intervals only or,

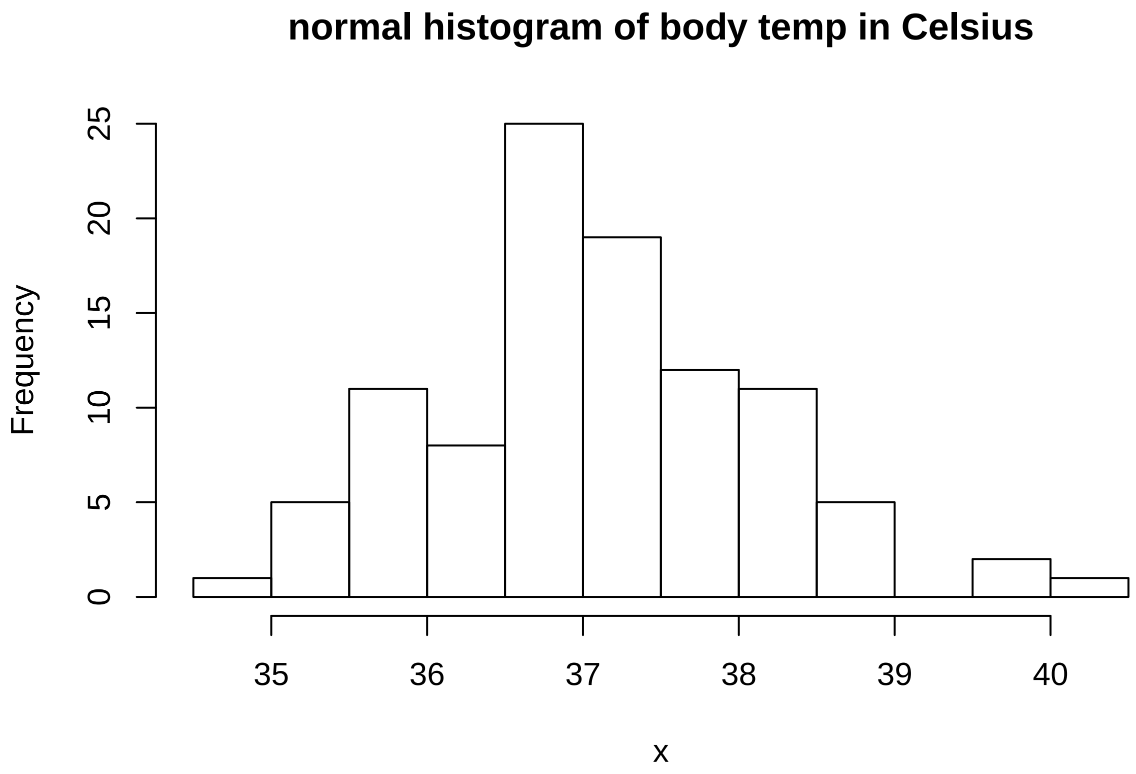
Apply Bayesian inference

Central Limit Theorem

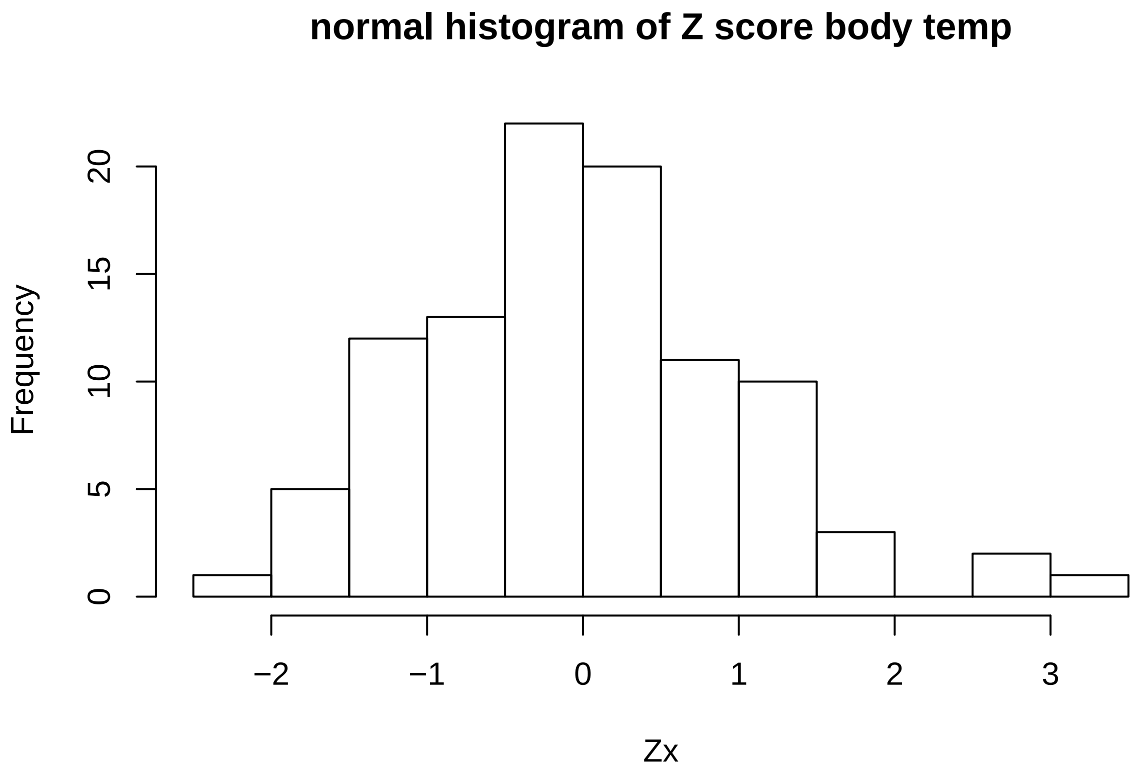
Review of histograms

Histograms are used to display distributions e.g. the body temperature of a random sample of healthy people.

```
x <- rnorm(100, mean = 37,sd=1)
Zx <- (x-mean(x))/sd(x)
hist(x, main = "normal histogram of body temp in Celsius")
```

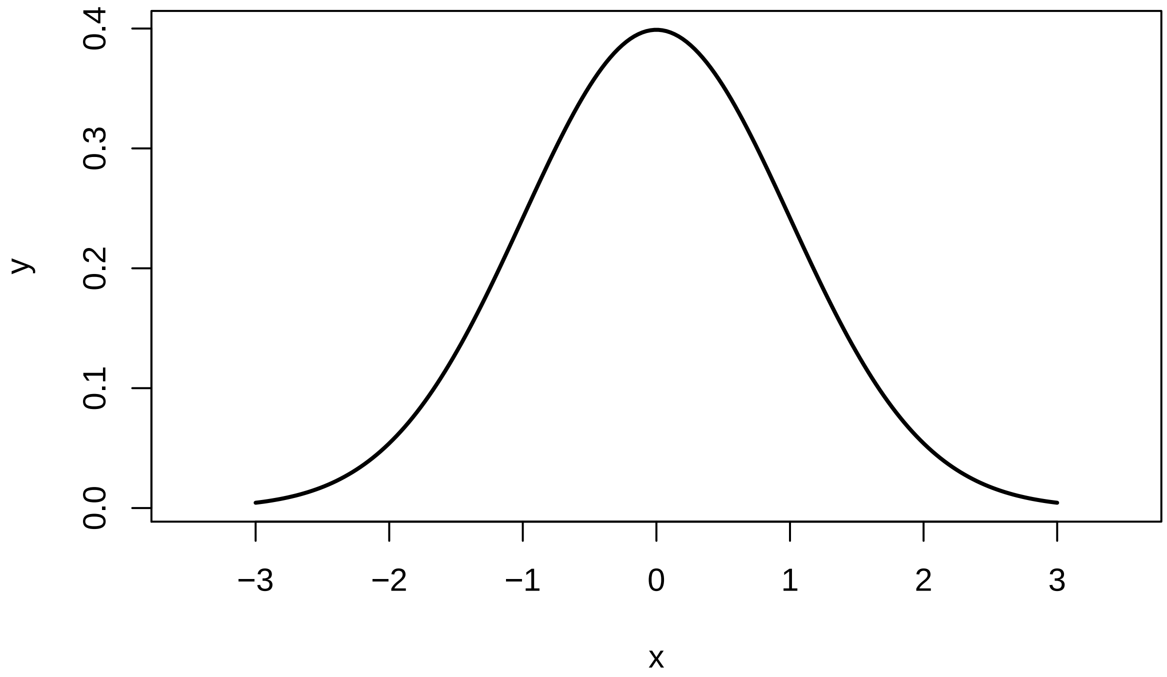


```
hist(Zx, main = "normal histogram of Z score body temp")
```



If a distribution is perfectly normal then the properties of the distribution are known.

```
x<-seq(-3,3,length=200)
s = 1
mu = 0
y <- (1/(s * sqrt(2*pi))) * exp(-((x-mu)^2)/(2*s^2))
plot(x,y, type="l", lwd=2, col = "black", xlim = c(-3.5,3.5))
```



Things to note about a normal distribution.

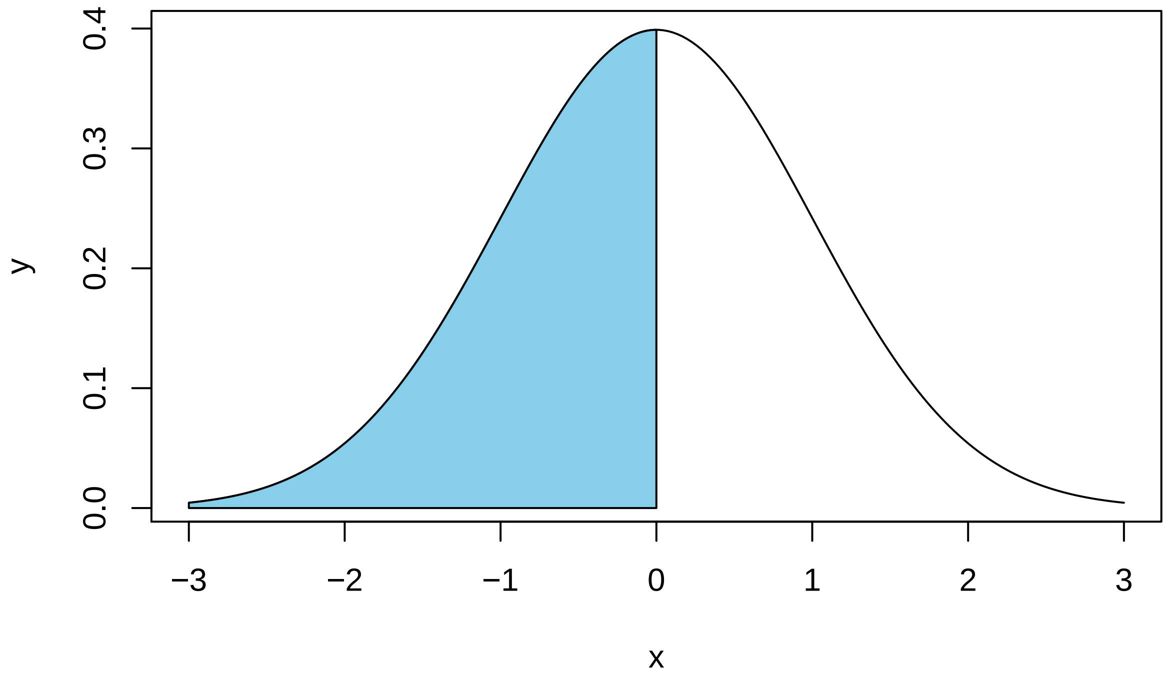
50% of the data fall above the mean and 50% below.

The majority of the data fall between 2 standard deviations above and below the mean. If you are higher or lower than these values then you are an extreme point.

This allows for predictions about the distribution because we know predictions aren't certain rather they are probabilistic.

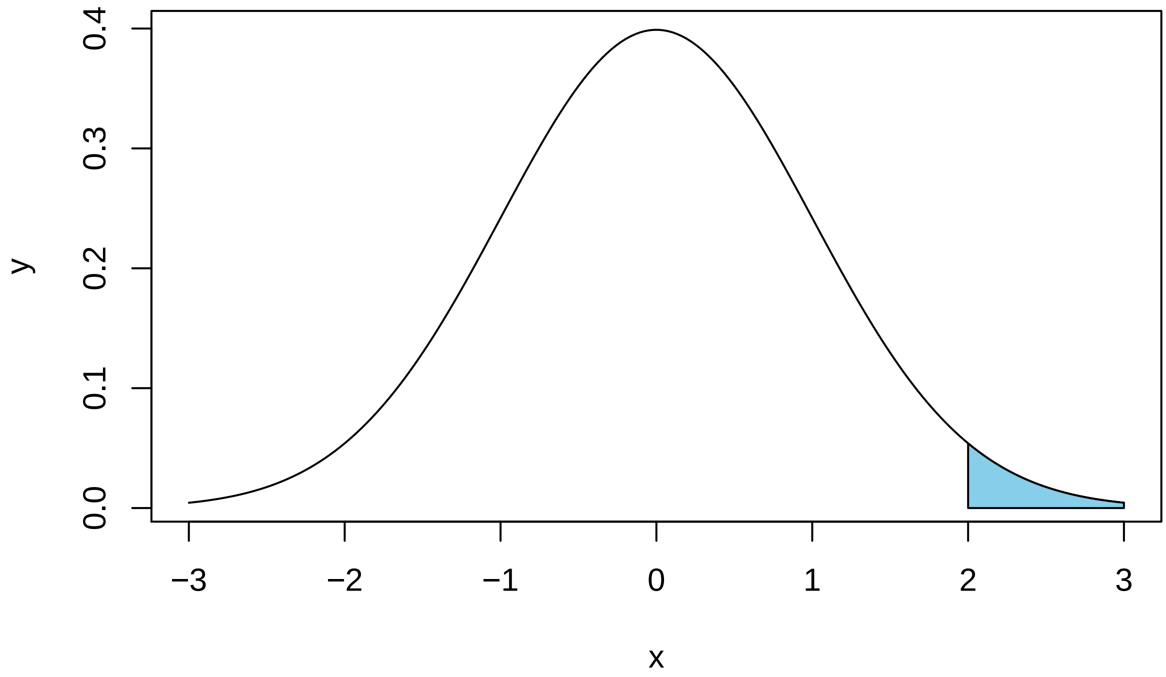
For example if one person is randomly selected from the sample, what is the probability that his or her body temperature is less than Z = 0 (or X = 37 for degrees celsius)? It's, p = 0.50.

```r
x=seq(-3,3,length=200)
y=dnorm(x,mean=0,sd=1)
plot(x,y,type="l")
x=seq(-3,0,length=100)
y=dnorm(x,mean=0,sd=1)
polygon(c(-3,x,0),c(0,y,0),col="skyblue")
```



Alternatively, if one person is randomly selected from the sample, what is the probability that his or her body temperature is greater than Z = 2 (or X = 38 for degrees celsius)? It’s, p = 0.20.

```
x=seq(-3,3,length=200)
y=dnorm(x,mean=0,sd=1)
plot(x,y,type="l")
x=seq(2,3,length=100)
y=dnorm(x,mean=0,sd=1)
polygon(c(2,x,3),c(0,y,0),col="skyblue")
```



If this sample is healthy, then no one should have a fever.

I detected a person with a fever

Therefore, this sample is not 100% healthy.

Sampling Distributions

But rather than using NHST on distributions of individuals we use NHST on distributions of sample statistics.

A sampling Distributions is a distribution of sample statistics, obtained from multiple samples, for example, a distribution of sample means, of sample correlations, of sample regression coefficients.

Sampling distributions are important in statistics because they provide a major simplification on the route to statistical inference. More specifically, they allow analytical considerations to be based on the sampling distribution of a statistic, rather than on the joint probability distribution of all the individual sample values.

It's important to realise that a sampling distribution is hypothetical. Instead we will only have a single sample.

Let's assume a mean is calculated from a sample, obtained randomly from the population.

Assume a certain sample size, N

Now assume we had multiple random samples (which is not what we do in practice), all of size N, and therefore many sample means. These would all differ a little bit because of sampling error.

Collectively, they form a sampling distribution.

Marrying sampling distributions and probability

If one sample is obtained from a normal healthy population, what is the probability that the sample mean is less than Z = 0? Again, it's p = 0.50.

If one sample is obtained from a normal healthy population, what is the probability that the sample mean is less than Z = 2? Again, it's p = 0.20.

In the latter case, if this population is healthy, then no one sample should have a high mean body temperature.

I obtained a very high sample mean.

Therefore, the population is not healthy.

Central Limit Theorem

Three principles

1. The mean of a sampling distribution is the same as the mean of the population
2. The standard deviation of the sampling distribution is the square root of the variance of the sampling distribution $\sigma^2 = \frac{\sigma^2}{N}$
3. The shape of a sampling distribution is approximately normal if either (a) N >= 30 or (b) the shape of the population is normal.

NHST and the Central Limit Theorem

Multiple regression

Assume the null is true

Conduct a study

Calculate B, SE and t

where t = B/SE

the p-value is a function of t and sample size

Conceptually, the t-value is a ratio of what we observed (e.g. the slope of the regression line) relative to what we would expect due to chance (e.g. the slope is zero). A ratio of 1 would be something around the null.

If the null hypothesis is true, then no one sample should have a very low or very high slope. Thus, if I obtain a very high slope I should reject the null. But what does 'very high' mean?

The very low or very high values depend on the normal distribution. Remember, the shape of a sampling distribution is approximately normal if either (a) N >= 30 or (b) the shape of the population is normal.

That means I can make probability judgements about the outcome. Note, that the third principle of the central limit theorem didn't say you get a normal distribution, rather it said you approximate one.

Instead, we get a t-distribution which comes from a family that are dependent on the sample size. As your sample size gets smaller your t-distribution gets a little wider which means you need a larger t-value to get out into the extremes to get a low p-value.

This all means that our ideas of 'very high' or 'very low' come from p being < 0.05.

Remember, that sampling error, and therefore standard error, is largely determined by sample size.

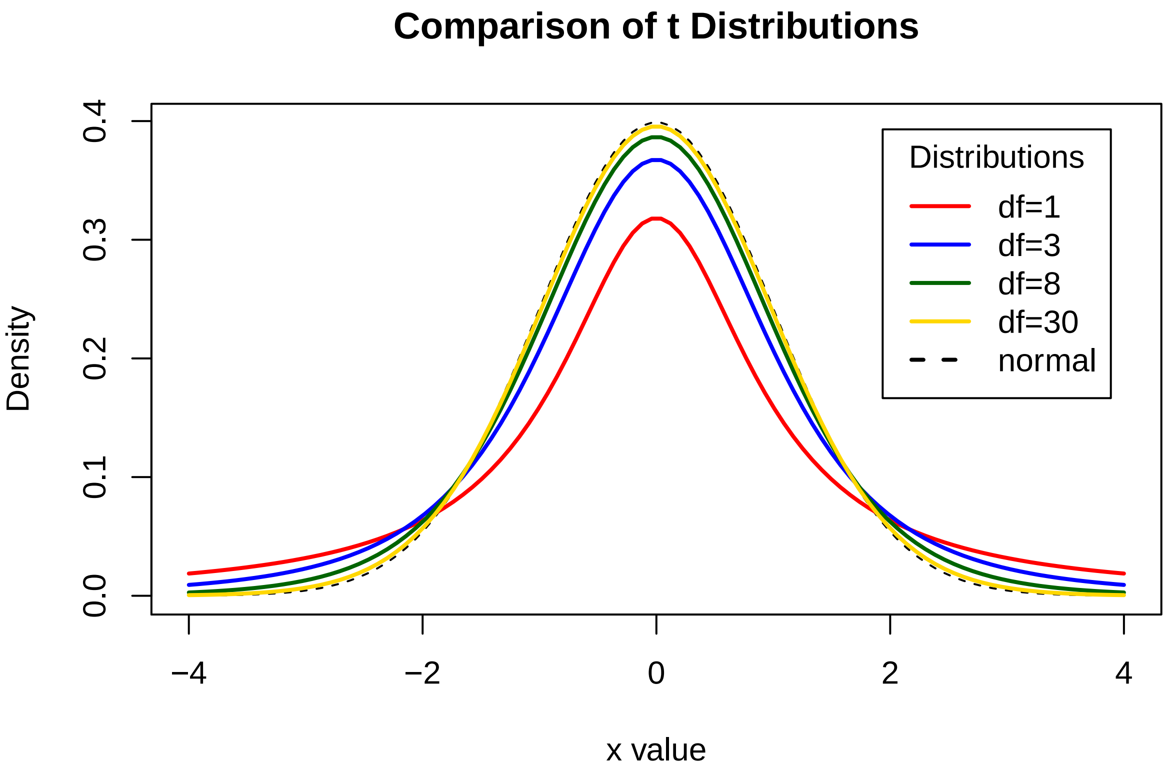
Standard error is the standard deviation of the sampling distribution. As samples get larger they're going to squeeze in around the mean and the standard error will decrease. It's important to note that NHST is biased by sample size.

t = B/SE

B = the unstandardised regression coefficient

SE = standard error

SE = $\sqrt{\frac{SS.Residual}{N-2}}$



As sample size increases, the actual mean approximates zero and the standard error shrinks as a function of sample size.

```
meanSamplingDist <- vector(length=length(i))
SESamplingDist <- vector(length=length(i))
for(i in 1:6){
  x <- rnorm(10**i)
  Zx <- (x-mean(x)/sd(x))
  SE <- sd(Zx)/sqrt(length(Zx))
  meanSamplingDist[i]<-(mean(Zx))
  SESamplingDist[i]<- (SE)
}

meanSamplingDist # the mean of the sampling distribution
```

```
## [1]  7.563941e-03 -2.984371e-03  1.063952e-03  2.671753e-05  3.730276e-06
## [6]  2.679761e-07
```

```
SESamplingDist # the standard error of the sampling distribution
```

```
## [1] 0.330676323 0.101355430 0.032241270 0.010105143 0.003155695 0.001000965
```

Confidence Intervals

Confidence intervals around sample means

All sample statistics e.g. a sample mean, are point estimates i.e. one point from a sample distribution. More specifically, a sample mean represents a single point in a sampling distribution. Any one sample will never be perfect.

The logic of confidence intervals is to report a range of values, rather than a single value. In other words, report an interval estimate rather than a point estimate.

We can define a confidence interval as an interval estimate of a population parameter, based on a random sample. The degree of confidence, e.g. 95%, represents the probability that the interval captures the true population parameter.

The main argument for interval estimates is the reality of sampling error. Sampling error implies that point estimates will vary from one study to the next (we use standard error to get a measure of sampling error). A researcher will therefore be more confident about accuracy with an interval estimate.

Below we can see that if we take a sample mean from a normal distribution we get different values each time due to sampling error. However, because our sample size is relatively big (30), this is quite small.

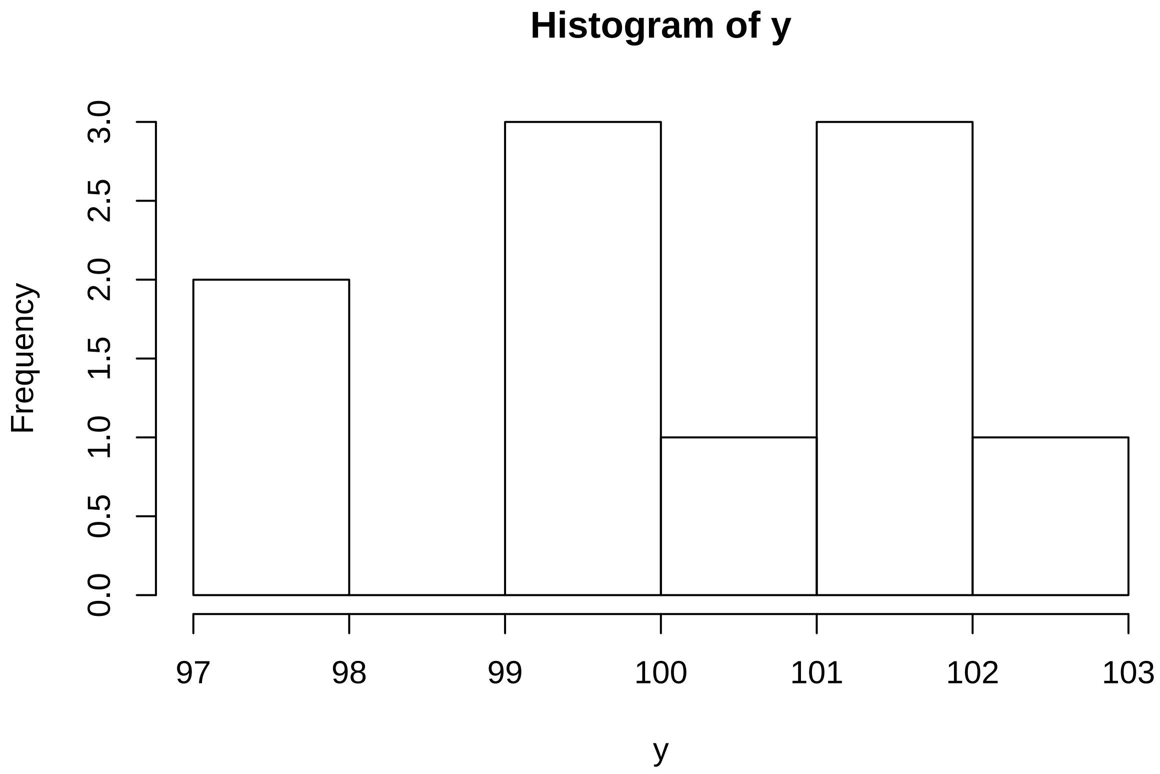
```
x<-rnorm(10000,100,10)
mean(x)
```

```
## [1] 99.88437
```

```
y <- replicate(10, {
  mm <- sample(x,30)
  mean(mm)
  print(mean(mm))
  })
```

```
## [1] 97.37197
## [1] 99.4733
## [1] 101.094
## [1] 99.44923
## [1] 99.12837
## [1] 101.6704
## [1] 102.8188
## [1] 101.1469
## [1] 100.2482
## [1] 97.0105
```

```
hist(y)
```



But if we reduce sample size there's more fluctuation in the point estimates i.e. the sample means.

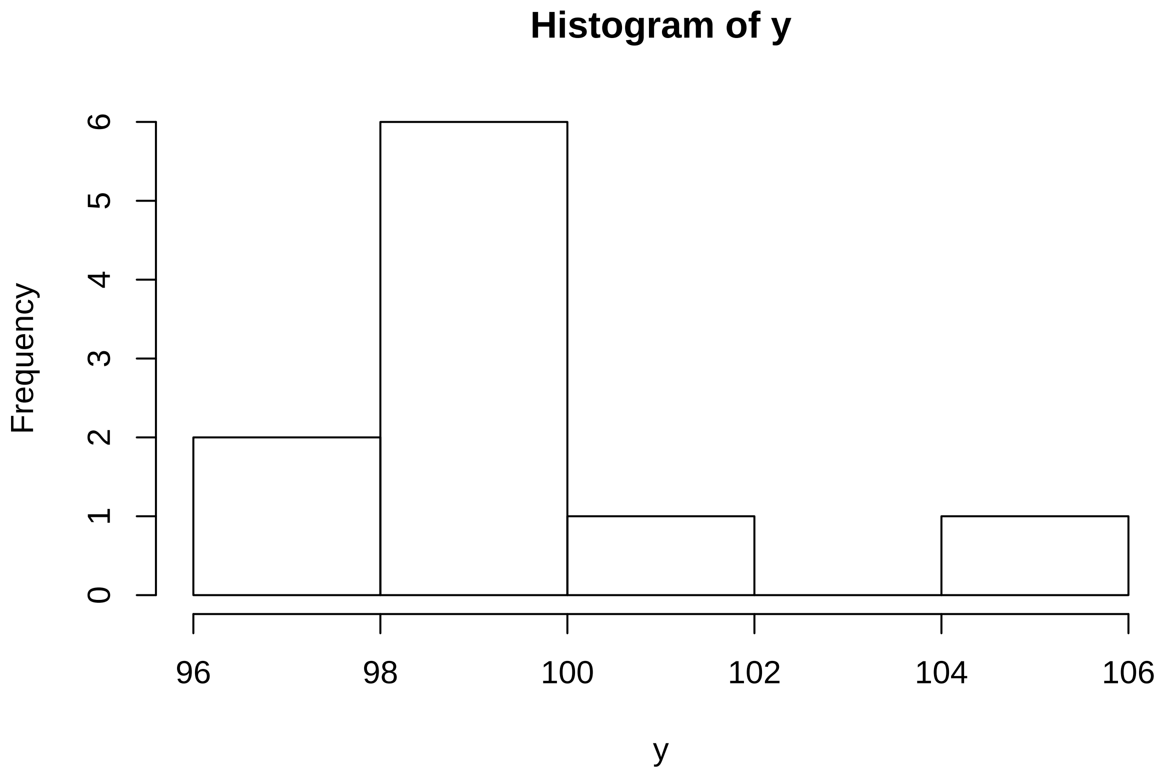
```
x<-rnorm(10000,100,10)
mean(x)
```

```
## [1] 99.97348
```

```
y <- replicate(10, {
  mm <- sample(x,10)
  mean(mm)
  print(mean(mm))
  })
```

```
## [1] 99.95794
## [1] 101.0147
## [1] 98.89737
## [1] 99.2869
## [1] 96.23901
## [1] 96.96515
## [1] 105.7327
## [1] 98.71632
## [1] 98.62212
## [1] 99.19111
```

```
hist(y)
```



#F-tests F tests are most commonly used for two purposes:

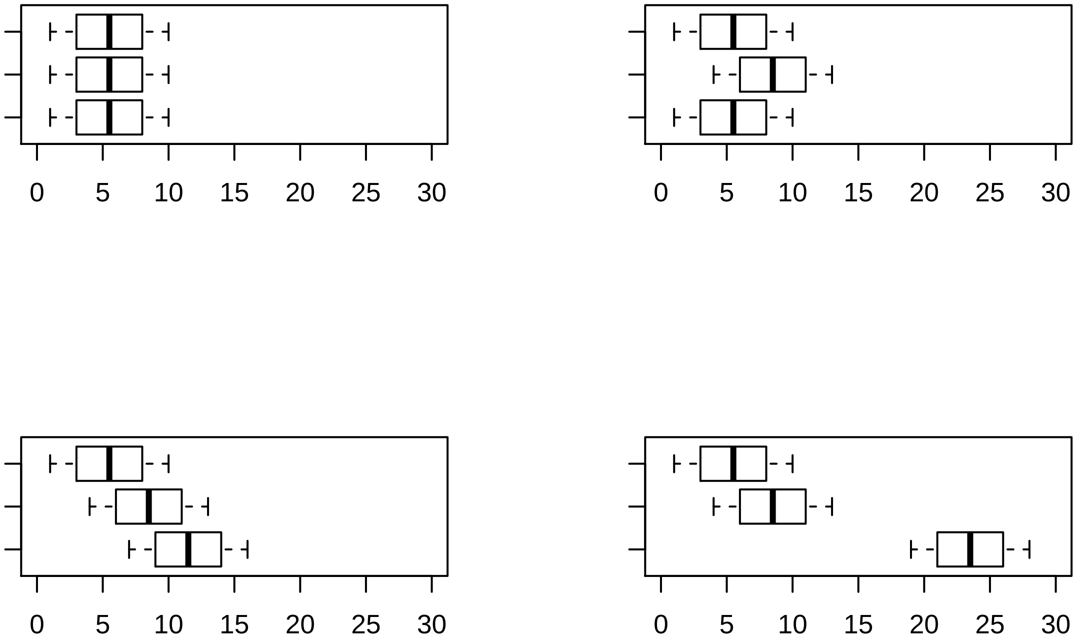
1. in ANOVA, for testing equality of means (and various similar analyses); and
2. in testing equality of variances

```
##        value        numdf        dendf
## 1.592065e-30 2.000000e+00 2.700000e+01
```

```
##     value     numdf     dendf
##  3.272727  2.000000 27.000000
```

```
##     value     numdf     dendf
##  9.818182  2.000000 27.000000
```

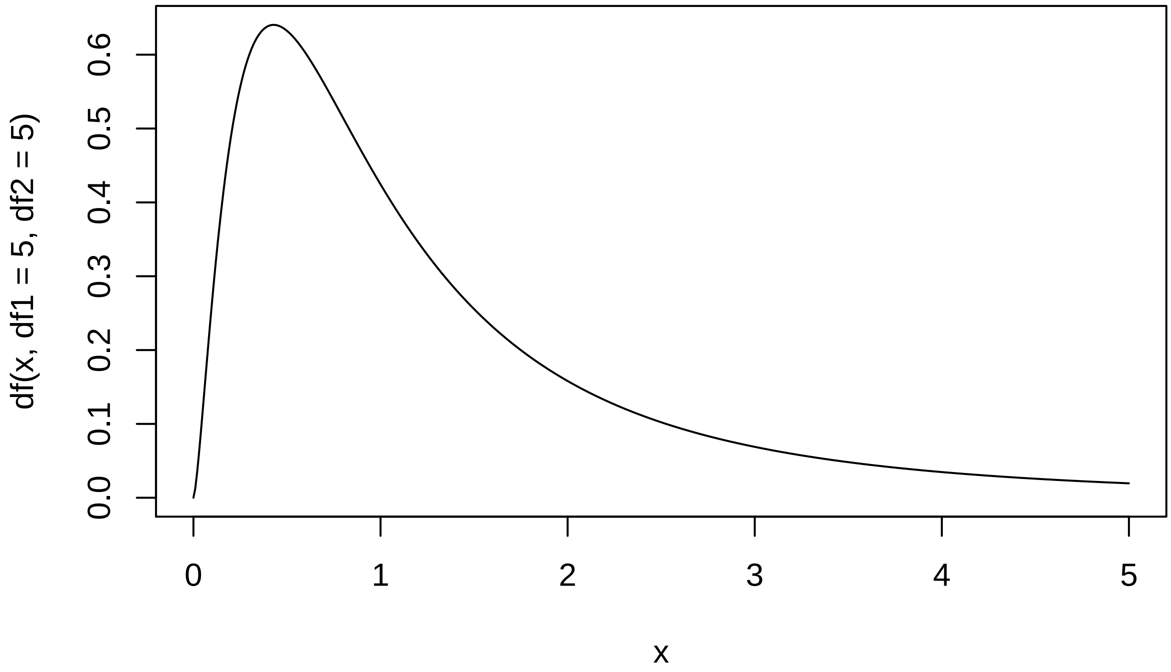
```
##    value    numdf    dendf
## 101.4545   2.0000  27.0000
```



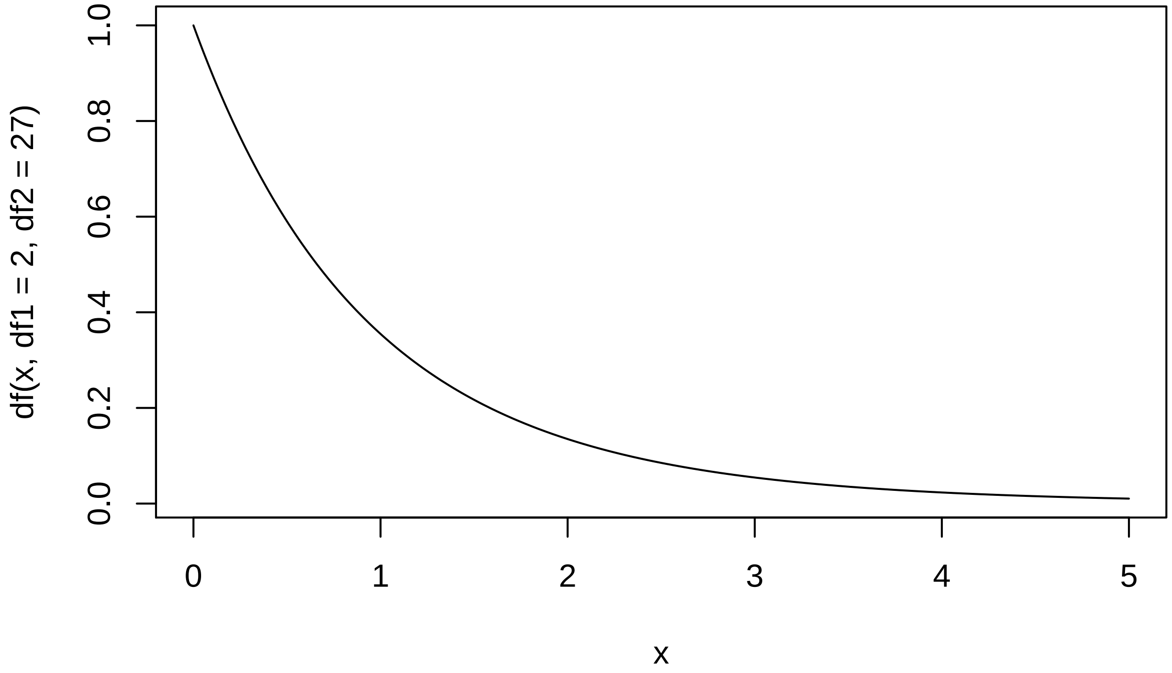
If the null hypothesis (equality of population means) were true, you'd expect some variation in sample means, and would typically expect to see F ratios roughly around 1. Smaller F statistics result from samples that are closer together than you'd typically expect, so you aren't going to conclude the population means differ.

That is, for ANOVA, you'll reject the hypothesis of equality of means when you get unusually large F-values and you won't reject the hypothesis of equality of means when you get unusually small values (it may indicate something, but not that the population means differ).

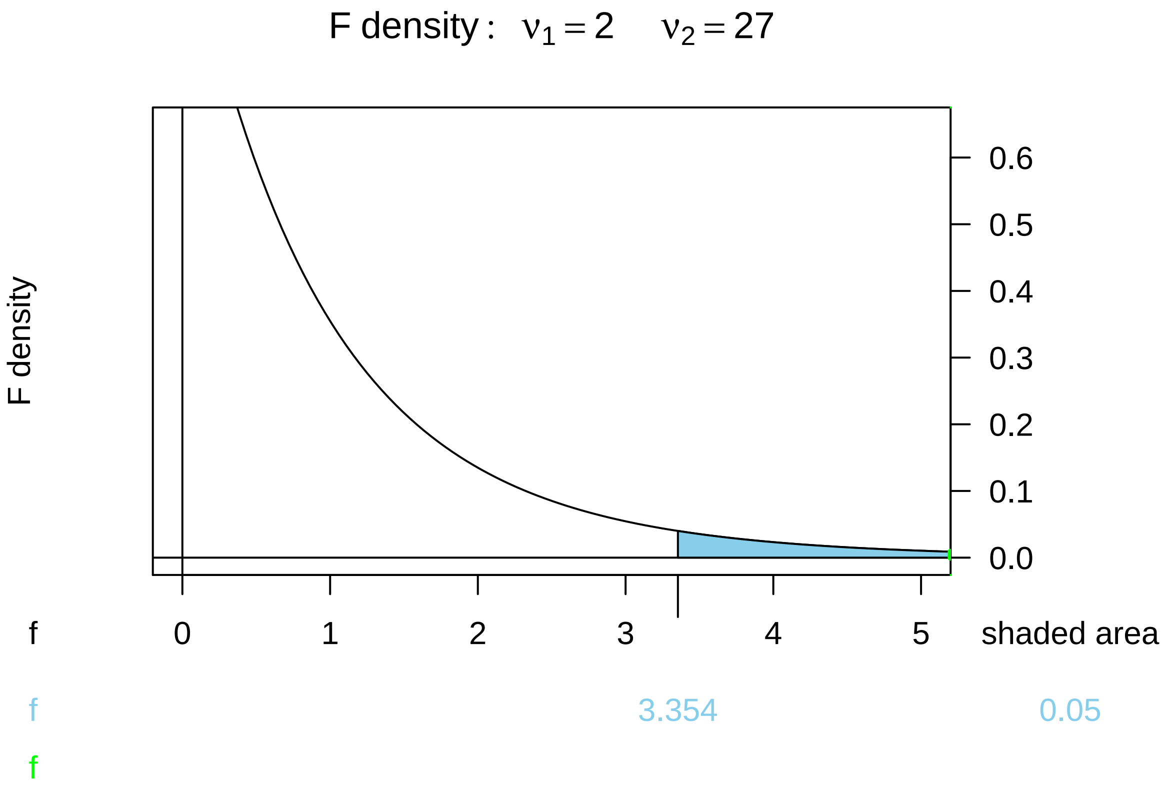
```
x<-seq(0,5,0.01); plot(x,df(x,df1=5,df2=5), type="l")
```



```r
x<-seq(0,5,0.01); plot(x,df(x,df1=2,df2=27), type="l")
```



This illustration shows that we only want to reject when F is in its upper tail.



2) F tests for equality of variance* (based on variance ratios). Here, the ratio of two sample variance estimates will be large if the numerator sample variance is much larger than the variance in the denominator, and the ratio will be small if the denominator sample variance is much larger than variance in the numerator.

That is, for testing whether the ratio of population variances differs from 1, you'll want to reject the null for both large and small values of F.

- (Leaving aside the issue of the high sensitivity to the distributional assumption of this test (there are better alternatives) and also the issue that if you're interested in suitability of ANOVA equal-variance assumptions, your best strategy probably isn't a formal test.)

Probability Density Functions

The "bell curve" shape is governed by the PDF. However, the actual "y"-value of this curve is itself more or less meaningless. The integral of the PDF $f(x)$ gives the probability that your random variable is less than some value:

$P(x < X) = \int_{\infty}^{X} f(x)dx$.

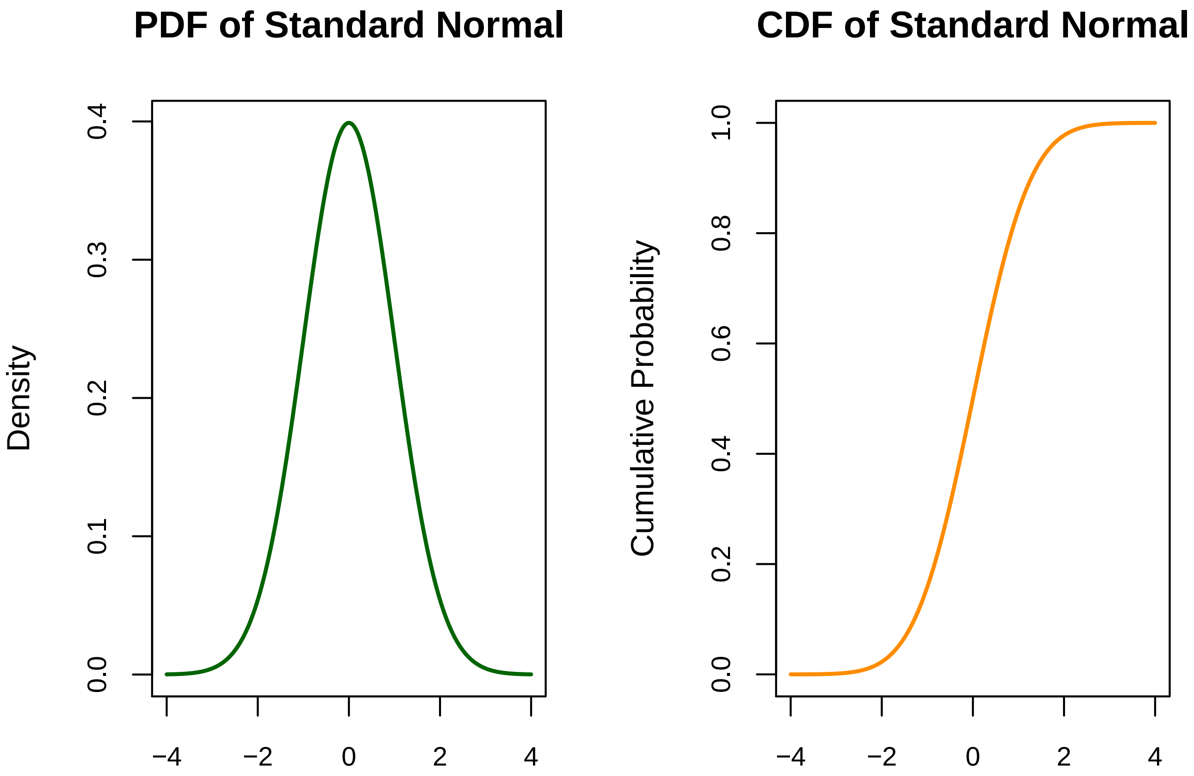
This is known as the CDF, or cumulative distribution function. By the fundamental theorem of calculus, the PDF is then the derivative of the CDF; that is, the PDF is the derivative of a function that returns a probability. So what is that intuitively? Honestly… it's not really anything. The "units" of the vertical axis in the PDF plot don't lead to anything intuitive; they are meaningful, but only in a derived, mathematical sense.

The area under the pdf equals 1. If the pdf value $f(x)$ exceeds 1 for some and indeed many values of x, that is perfectly fine: but $f(x)$ cannot exceed 1 for all x in an interval I of length exceeding 1. If the latter condition were to hold, then:

$\int_I f(x)dx =$ area under pdf in interval I > 1

in violation of the constraint that the total area is 1. The value of $f(x)$ is not a probability. The units of $f(x)$ are probability per unit length and you must multiply by length (more generally, find an area) to get a probability.

As a consequence, some people wish to think that $f(X)$ is the probability that $x = X$, but this is untrue for continuous distributions ($P(x = X) = 0$). However, for the PDF's discrete analog, the Probability Mass Function (PMF), this statement is quite true.



t-test

t-Test to compare the means of two groups under the assumption that both samples are random, independent, and come from normally distributed population with unknown but equal variances

To solve this problem we must use to a Student's t-test with two samples, assuming that the two samples are taken from populations that follow a Gaussian distribution (if we cannot assume that, we must solve this problem using the non-parametric test called Wilcoxon-Mann-Whitney test). Before proceeding with the t-test, it is necessary to evaluate the sample variances of the two groups, using a Fisher's F-test to verify the homoskedasticity (homogeneity of variances). In R you can do this in this way:

```r
a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)
b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)
```

```
var.test(a,b)
```

```
##
##  F test to compare two variances
##
## data:  a and b
## F = 2.1028, num df = 9, denom df = 9, p-value = 0.2834
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5223017 8.4657950
## sample estimates:
## ratio of variances
##           2.102784
```

We obtained p-value greater than 0.05, then we can assume that the two variances are homogeneous. Indeed we can compare the value of F obtained with the tabulated value of F for alpha = 0.05, degrees of freedom of numerator = 9, and degrees of freedom of denominator = 9, using the function qf(p, df.num, df.den):

```
qf(0.95, 9, 9)
```

```
## [1] 3.178893
```

Note that the value of F computed is less than the tabulated value of F, which leads us to accept the null hypothesis of homogeneity of variances.NOTE: The F distribution has only one tail, so with a confidence level of 95%, p = 0.95. Conversely, the t-distribution has two tails, and in the R's function qt(p, df) we insert a value p = 0975 when you're testing a two-tailed alternative hypothesis. Then call the function t.test for homogeneous variances (var.equal = TRUE) and independent samples (paired = FALSE: you can omit this because the function works on independent samples by default) in this way:

```
t.test(a,b, var.equal=TRUE, paired=FALSE)
```

```
##
##  Two Sample t-test
##
## data:  a and b
## t = -0.94737, df = 18, p-value = 0.356
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.93994   4.13994
## sample estimates:
## mean of x mean of y
##     174.8     178.2
```

We obtained p-value greater than 0.05, then we can conclude that the averages of two groups are significantly similar. Indeed, the value of t-computed is less than the tabulated t-value for 18 degrees of freedom, which in R we can calculate:

```
qt(0.975, 18)
```

```
## [1] 2.100922
```

This confirms that we can accept the null hypothesis H_0 of equality of the means.

Interactions

Interactions allow us assess the extent to which the association between one predictor and the outcome depends on a second predictor.