Statistics Teaching

Adam Kane

4 November 2016

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:

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.

```r
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.

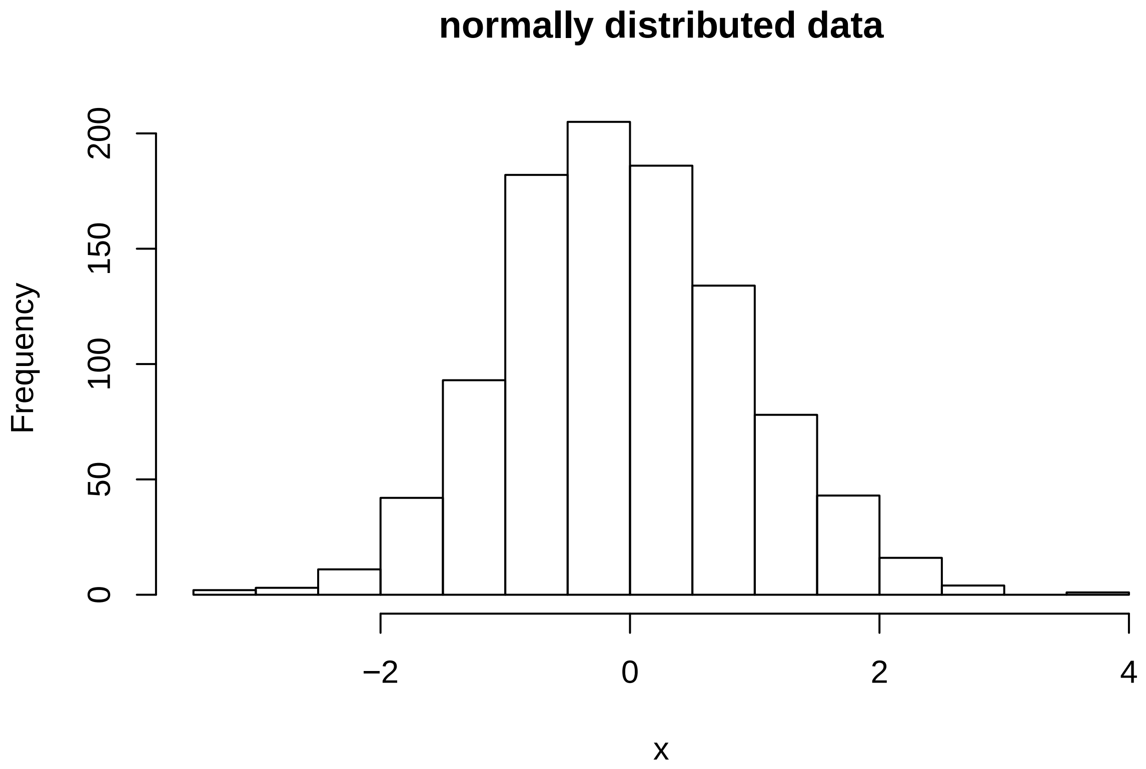
If you want to tell if a test is parametric or not, look at the formulas used in calculating it. Do they contain parameters/statistics?

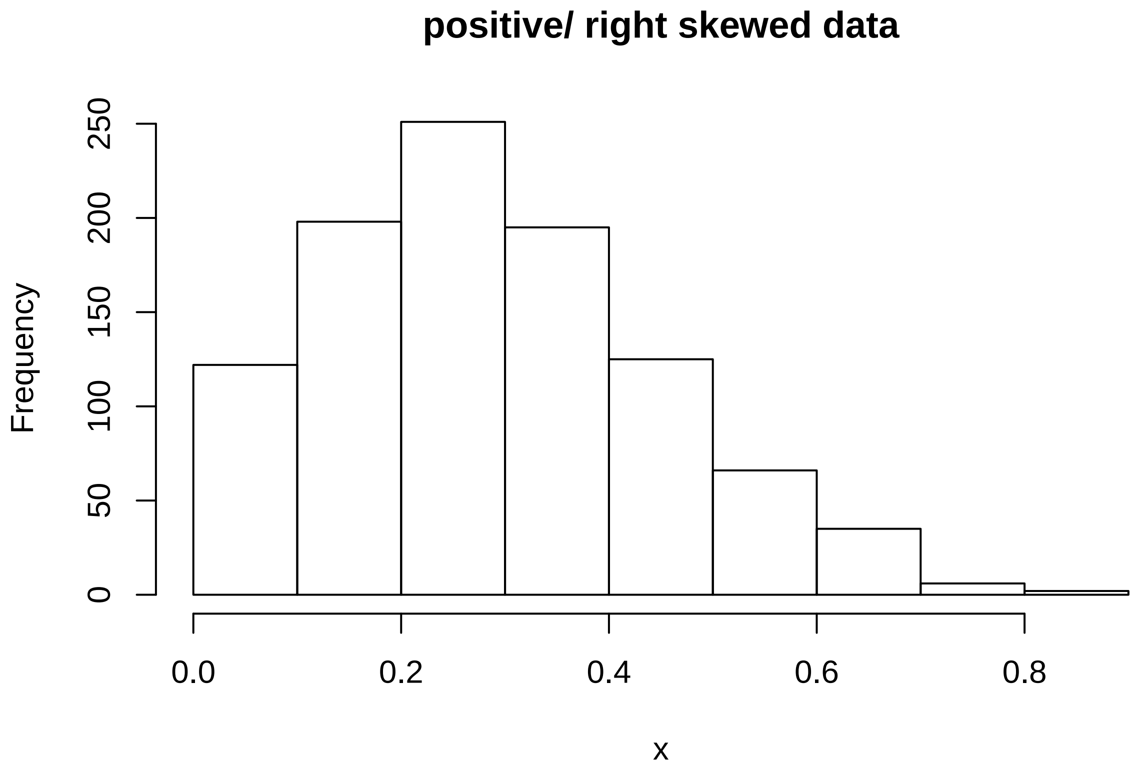
If your measurement scale is nominal or ordinal then you use non-parametric statistics

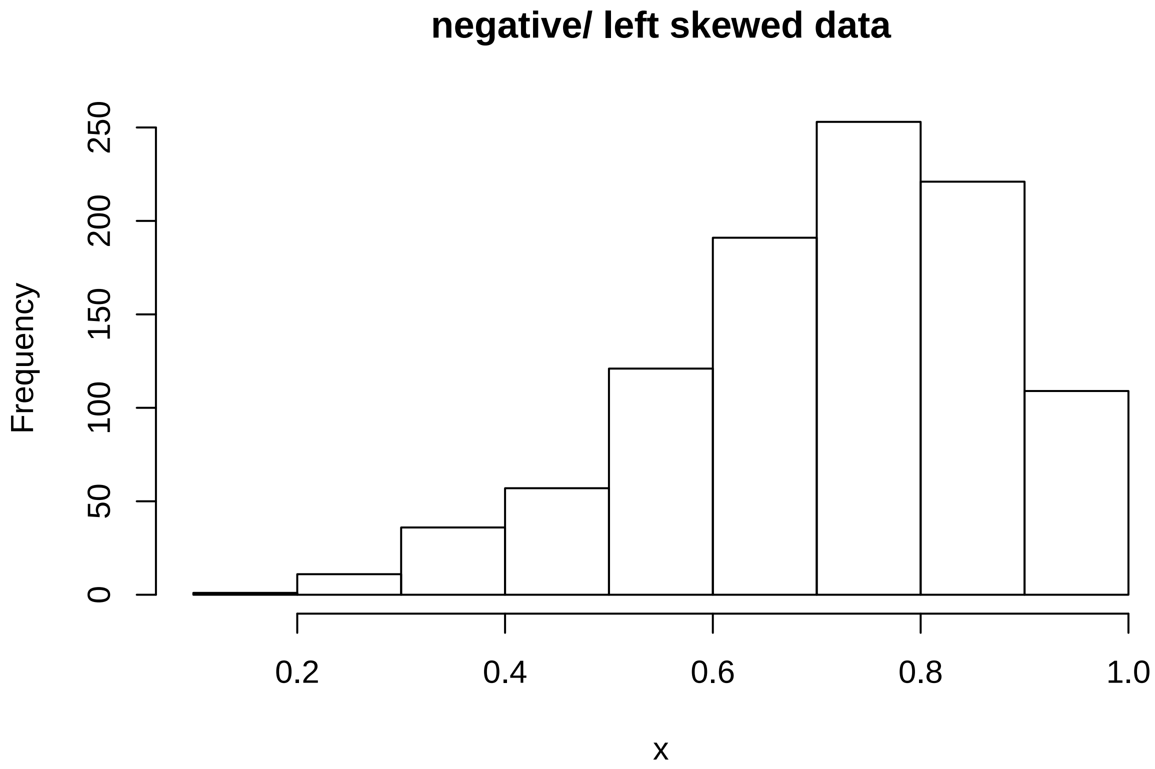
If you are using interval or ratio scales you use parametric statistics.

Histograms

are used to plot the distribution of the data


normally distributed data
Frequency
0
50
100
150
200
−2
0
2
4
x


positive/ right skewed data
Frequency
0
50
100
150
200
250
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

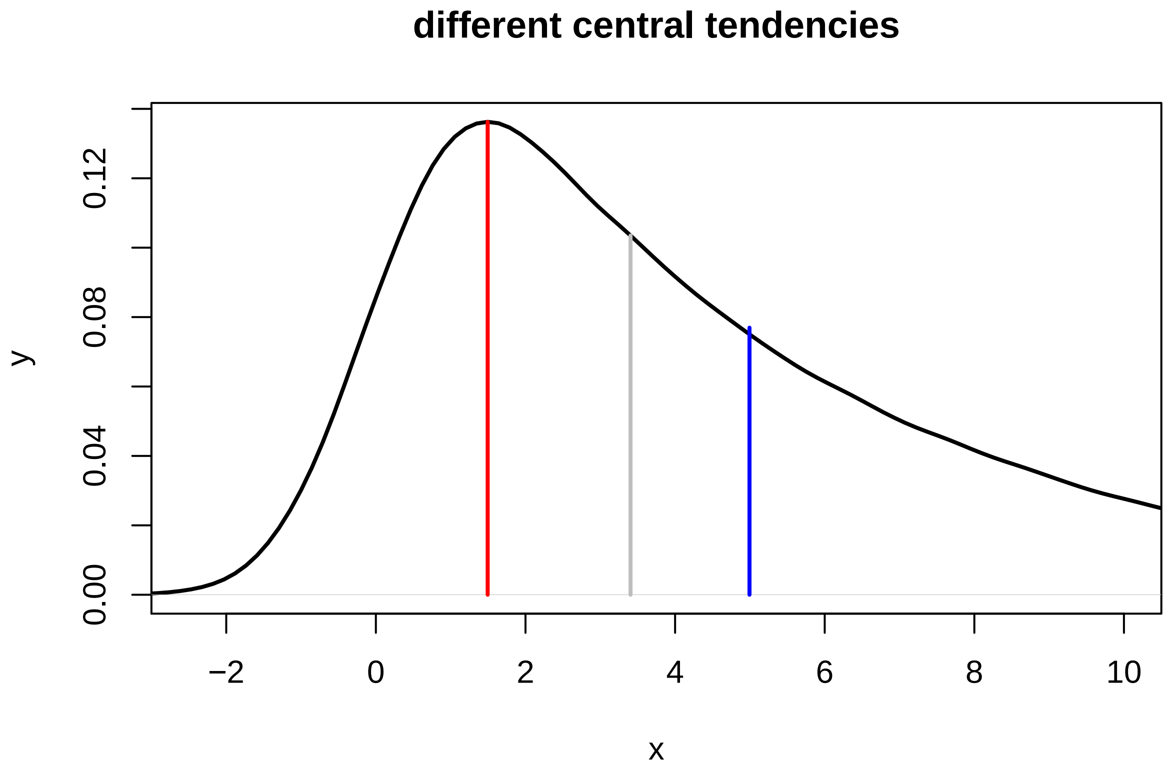
is the middle score (the score below which 50% of the distribution falls) preferred when there are extreme scores in the distribution

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:

```r
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

```r
sum(pointsPerGame)
```

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

```r
length(pointsPerGame)
```

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

so the mean is

```r
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.

```r
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

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

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

```r
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.

```r
(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
```

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

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

```r
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.

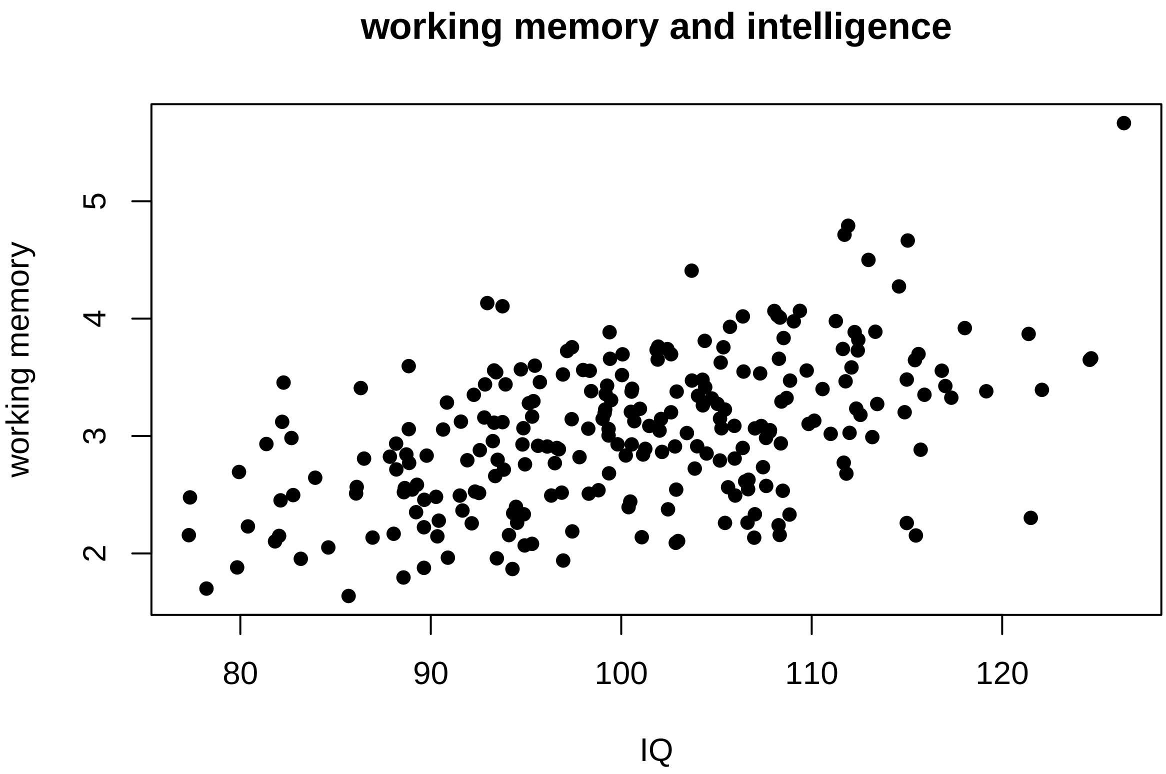
More precisely, a person’s score on X can be used to predict his or her score on Y.

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 = 8.8617, df = 248, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3900717 0.5792303
## sample estimates:
##       cor
## 0.4904055
```

Warnings about correlation

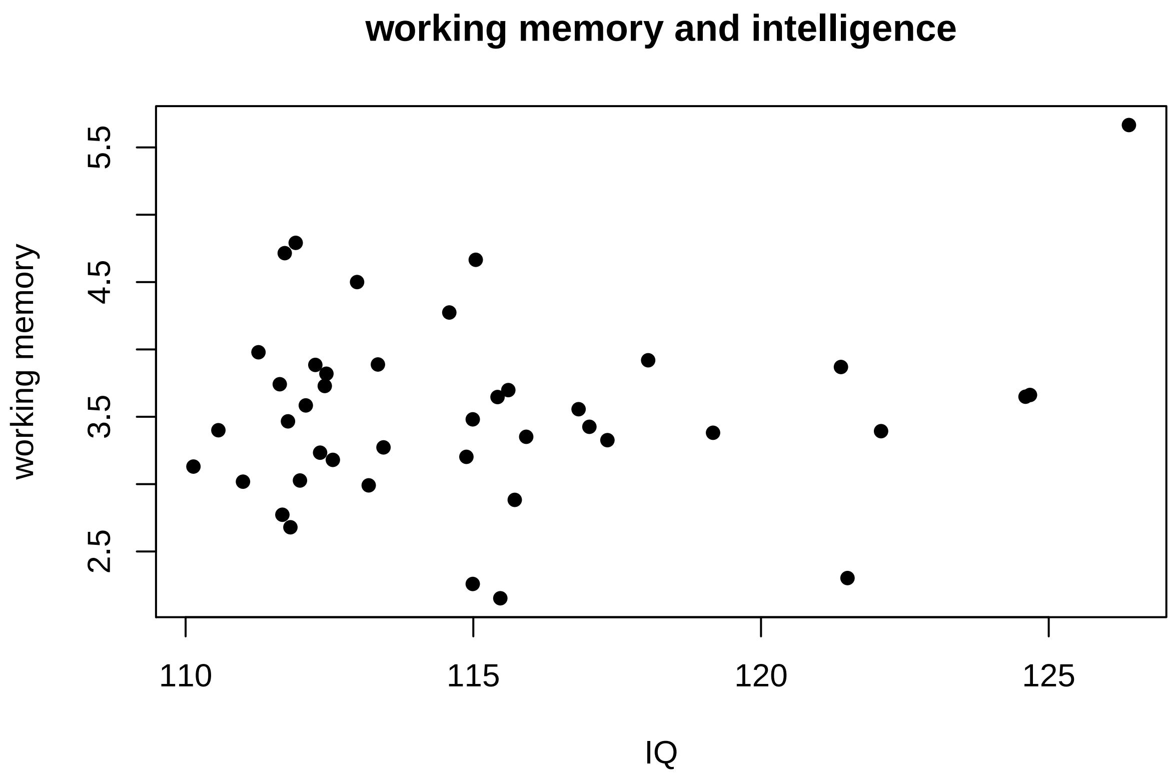
But we have to remember that correlation does not imply causation. In our example, working memory does not cause IQ and vice versa, rather there are lots of intervening variables.

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 = "w
```



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

```
##
##  Pearson's product-moment correlation
##
## data:  dfAttenuated$IQ and dfAttenuated$workingMemory
## t = 0.98995, df = 40, p-value = 0.3281
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1566523  0.4379875
## sample estimates:
```

```
##       cor
## 0.1546418
```

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"

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

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

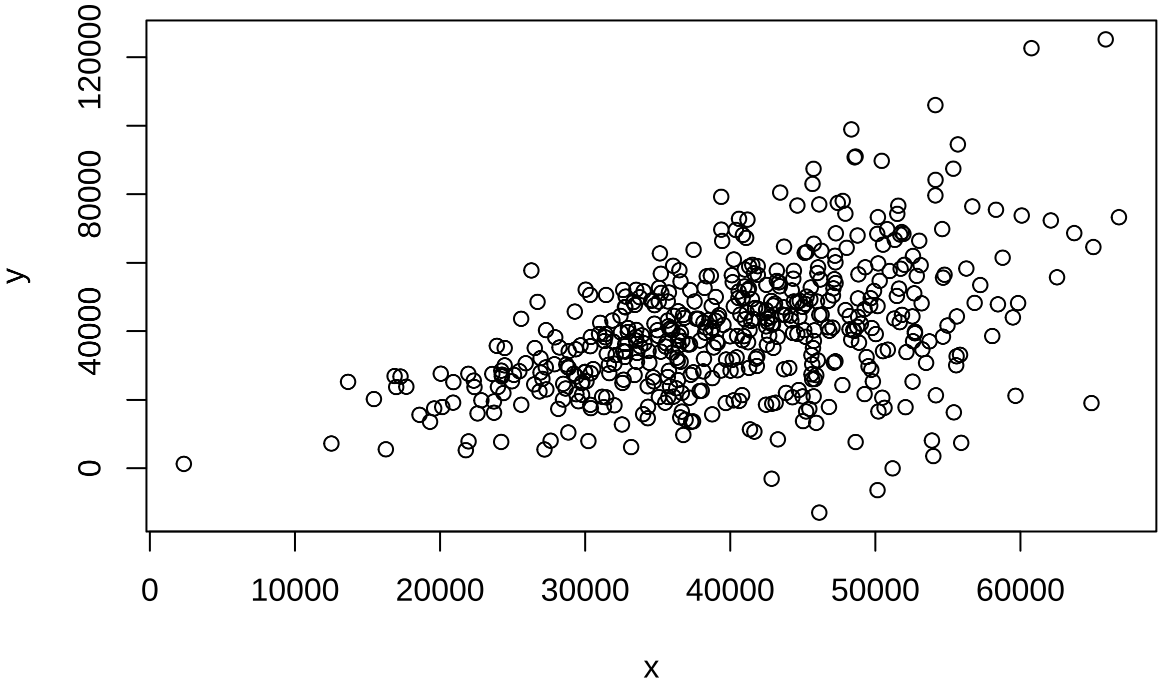
Homoscedasticity and heteroscedasticity

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

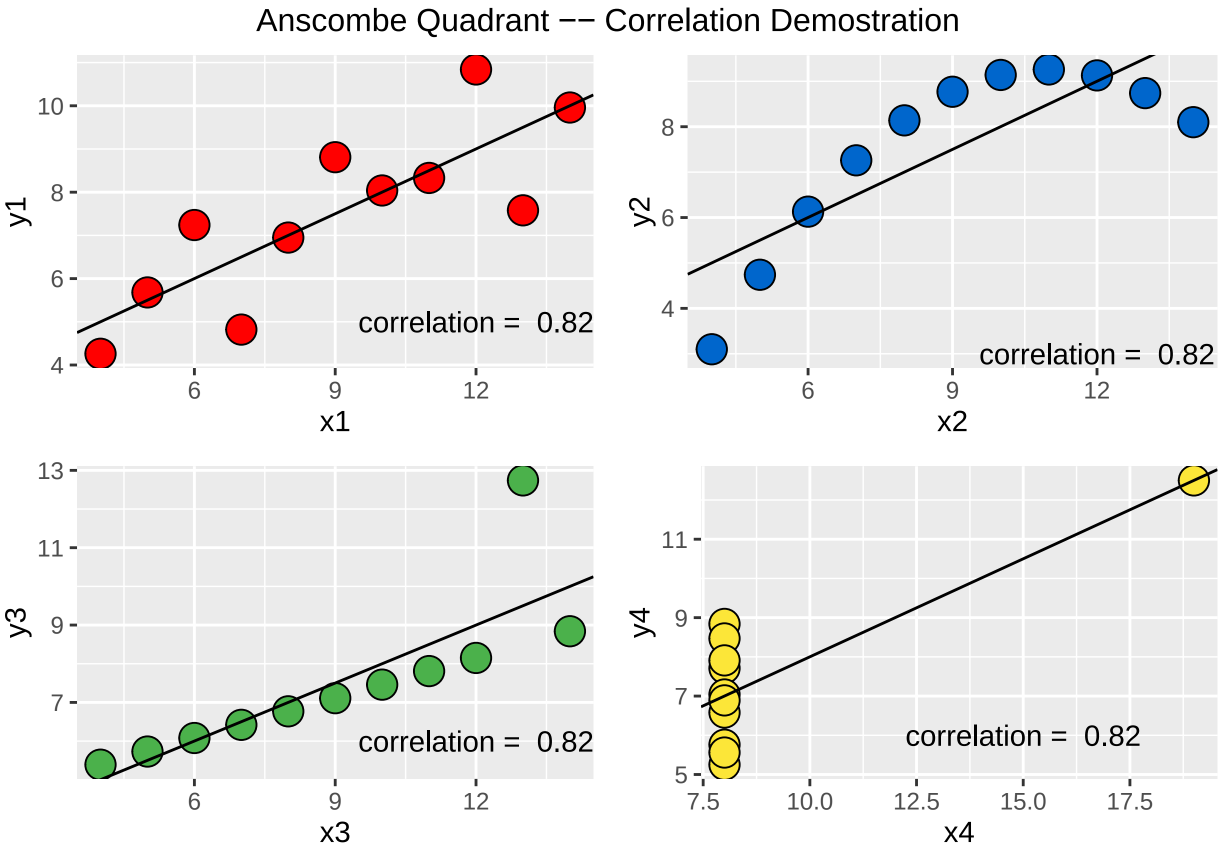
Why it's critical to look at your scatterplots.

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
```

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



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

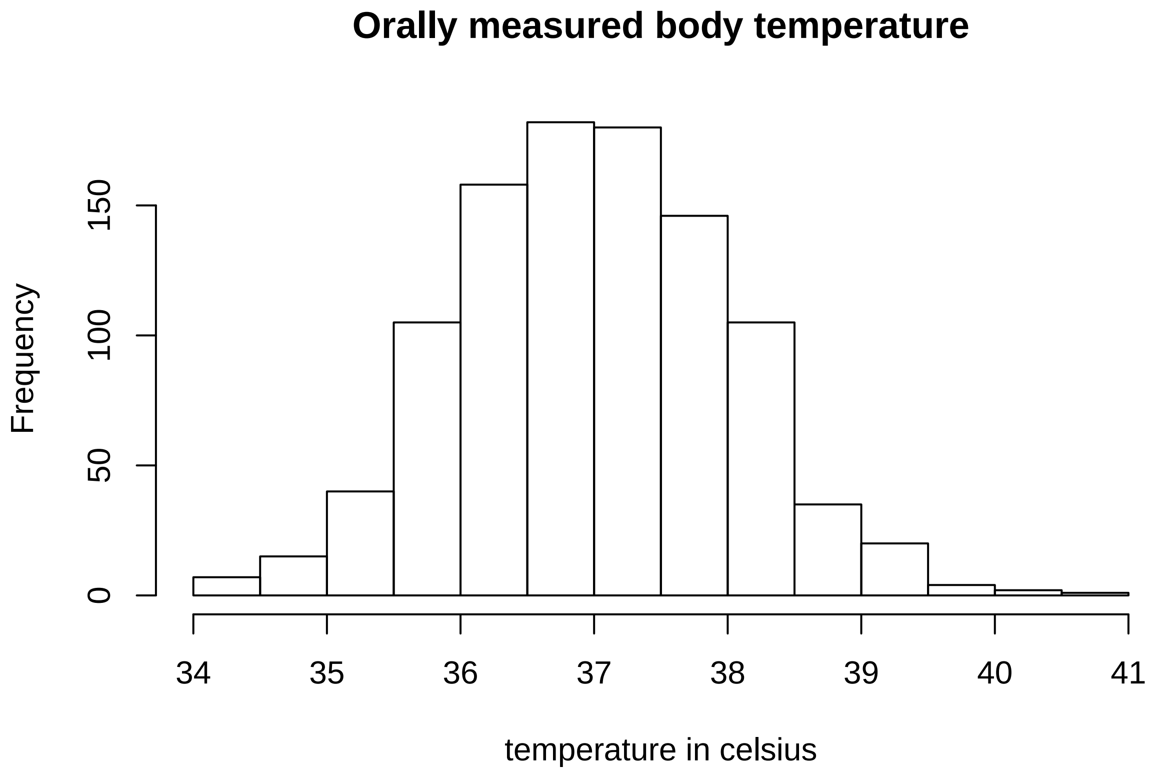
Inter-item reliability

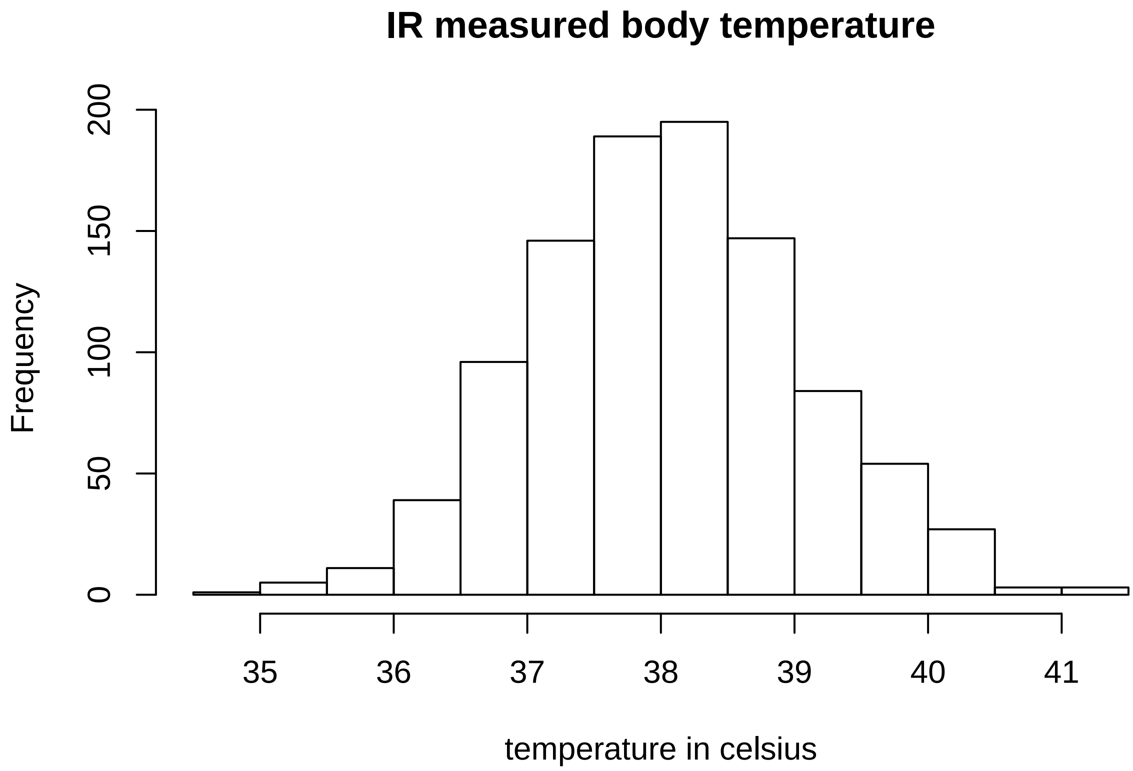
Example of body temperature

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.