
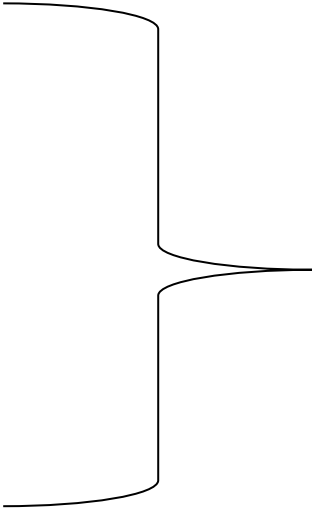

Q-Step Training Workshop

Data Analysis (intermediate level)

(see accompanying practical)

Mark Brown

Overview

- 
- 
- 1. Basic familiarity with SPSS
 - 2. Describing single variables
 - 3. Crosstabulation
 - 4. Sampling error / confidence intervals
 - 5. Scatterplots and Correlation
 - 6. Simple regression
- Assume ok with this

Part 1

Sampling Error

Calculating a confidence interval

2017 BSA Survey finds 36% would vote to leave if had another go!

If you were given the chance to vote again, how would you vote - to remain a member of the EU, to leave the EU, or would you not vote?

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---------|---------------------------------------|-----------|---------|---------------|--------------------|
| Valid | Remain a member of the European Union | 1515 | 38.0 | 50.9 | 50.9 |
| | Leave the European Union | 1069 | 26.8 | 35.9 | 86.9 |
| | I would not vote | 222 | 5.6 | 7.5 | 94.3 |
| | Prefer not to say/don't know | 169 | 4.2 | 5.7 | 100.0 |
| | Total | 2975 | 74.6 | 100.0 | |
| Missing | Item not applicable | 1013 | 25.4 | | |
| Total | | 3988 | 100.0 | | |

Sampling error

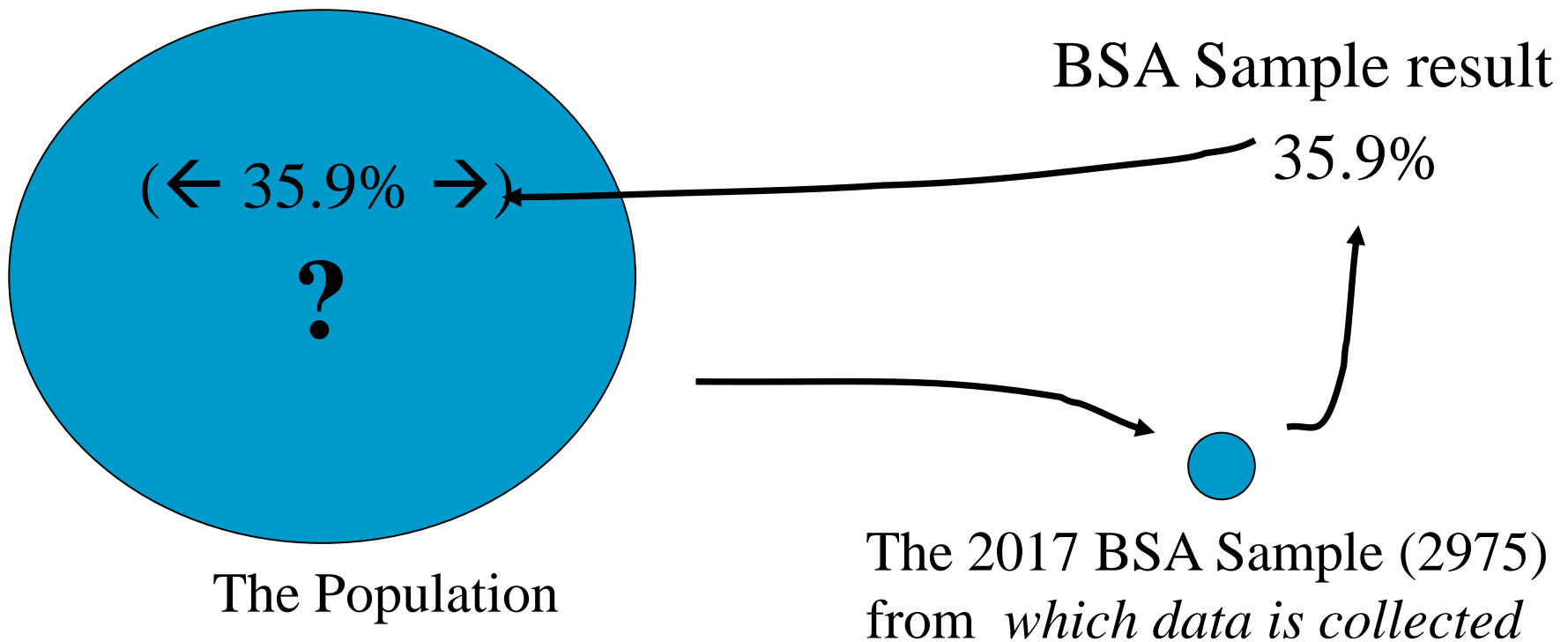
‘Survey found 36% of population would vote to leave the EU’

Or more accurately...

Survey found 36% of the sample would vote to leave the EU

Survey data from a sample includes random (sampling) error

We know there is a margin of error with all sample estimates..
Represented by a **confidence interval** (the range within which we are confident the true population figure lies)



What determines amount of **sampling error** (size of confidence interval?)

- Assuming you have a **random sample** (otherwise it is not strictly possible to calculate sampling error or confidence intervals)
 1. **Size** of the sample (bigger sample = smaller sampling error)
 2. Natural **variation** in thing you are estimating (easier to accurately estimate something where there is not much variation (less variation = smaller sampling error))

Focusing on sampling error.. But don't forget other types of error

- For the rest of this example we will focus on how to measure sampling error and use it to calculate confidence intervals around our survey estimates
- But remember that accuracy of survey data is also effected by non-sampling error (all other causes of error e.g. measurement error)

To calculate a confidence interval...

1. Calculate the **standard error** (this is a measure of how much sampling error there is)
 2. Use the standard error to build a **confidence interval**
- **95% confidence interval** = estimate + or – 2 x standard error



Calculating sampling error: 2 worked examples

- The way we calculate the standard error is slightly different for...

- **categorical variables** (estimating a population percentage)

- e.g. What % would vote to leave the European Union*

- **interval variables** (estimating a population mean)

- e.g. What is the mean starting salary of Manchester graduates?*

Show you how to calculate sampling error (and confidence intervals) for both these examples...

Standard error of a percentage

$$SE = \sqrt{\frac{\text{sample \%} * (100 - \text{sample \%})}{n}}$$

n = sample size

-
- Use the formula to calculate confidence intervals around our estimate of
 - 35.9% of people would vote to leave Europe

Calculating confidence intervals around a percentage estimate (35.9% of people want to leave Europe)

- **Step 1 calculate the standard error**

$$SE = \sqrt{\frac{sample \% * (100 - sample \%)}{n}}$$

$$SE = \sqrt{\frac{35.9*(100-35.9)}{2975}} = \mathbf{0.8795}$$

Calculating confidence intervals around a percentage estimate (35.9% of people would vote to leave Europe)

- **Step 2 use the standard error to calculate the confidence interval around estimate (35.9%)**

95% confidence intervals = $\pm 2 \times$ standard error

So.. = $35.9 \pm (2 \times 0.8795)$

= between **34.2% and 37.7%**

(can be confident that % who would vote to leave EU in population in population lies within that range in 95 out of 100 samples)

Big differences by age ... but are they statistically significant?

- Could the differences we see between age groups in our sample just be due to sampling error?

Crosstab

| | Age of respondent(grouped)<6 category> dv | | | | | | Total |
|---------------------------------------|---|--------------|--------------|--------------|--------------|--------------|---------------|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65+ | |
| Remain a member of the European Union | 214 66.7% | 323 60.1% | 251 53.3% | 248 46.9% | 202 46.7% | 276 40.4% | 1514 50.9% |
| Leave the European Union | 45 14.0% | 128 23.8% | 147 31.2% | 215 40.6% | 187 43.2% | 346 50.6% | 1068 35.9% |
| I would not vote | 36 11.2% | 49 9.1% | 47 10.0% | 33 6.2% | 28 6.5% | 29 4.2% | 222 7.5% |
| Prefer not to say/don't know | 26 8.1% | 37 6.9% | 26 5.5% | 33 6.2% | 16 3.7% | 33 4.8% | 171 5.7% |
| Total | 321 100% | 537 100% | 471 100% | 529 100% | 433 100% | 684 100% | 2975 100% |

Are differences statistically significant?

2 approaches

1. Run a Chi Square test of independence (are the variables related in the population)
2. Calculate confidence intervals for specific estimates

Crosstab

| | Age of respondent(grouped)<6 category> dv | | | | | | Total |
|---------------------------------------|---|--------------|--------------|--------------|--------------|--------------|---------------|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65+ | |
| Remain a member of the European Union | 214 66.7% | 323 60.1% | 251 53.3% | 248 46.9% | 202 46.7% | 276 40.4% | 1514 50.9% |
| Leave the European Union | 45 14.0% | 128 23.8% | 147 31.2% | 215 40.6% | 187 43.2% | 346 50.6% | 1068 35.9% |
| I would not vote | 36 11.2% | 49 9.1% | 47 10.0% | 33 6.2% | 28 6.5% | 29 4.2% | 222 7.5% |
| Prefer not to say/don't know | 26 8.1% | 37 6.9% | 26 5.5% | 33 6.2% | 16 3.7% | 33 4.8% | 171 5.7% |
| Total | 321 100% | 537 100% | 471 100% | 529 100% | 433 100% | 684 100% | 2975 100% |

1. Run a chi square test

1. The Pearson Chi-square test in SPSS

- Run as part of the crosstab procedure in SPSS..

Chi-Square Tests

| | Value | df | Asymp. Sig. (2-sided) |
|------------------------------|----------------------|----|-----------------------|
| Pearson Chi-Square | 193.643 ^a | 15 | .000 |
| Likelihood Ratio | 204.500 | 15 | .000 |
| Linear-by-Linear Association | 6.621 | 1 | .010 |
| N of Valid Cases | 2975 | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 18.45.

- It includes the Chi-square value (193.643)
- and a 'P value' which is the probability (so takes a value between 0 and 1) of getting the observed results (in your table) if the two variables were in fact not related (independent) in the population. (i.e. if just due to sampling error)
- ASK is the P VALUE less than 0.05? If so (which in this case it is), it means there is less than 5% chance we could have got those results if the variables were actually unrelated in the population and we therefore declare the relationship 'statistically significant' (can generalise to population)

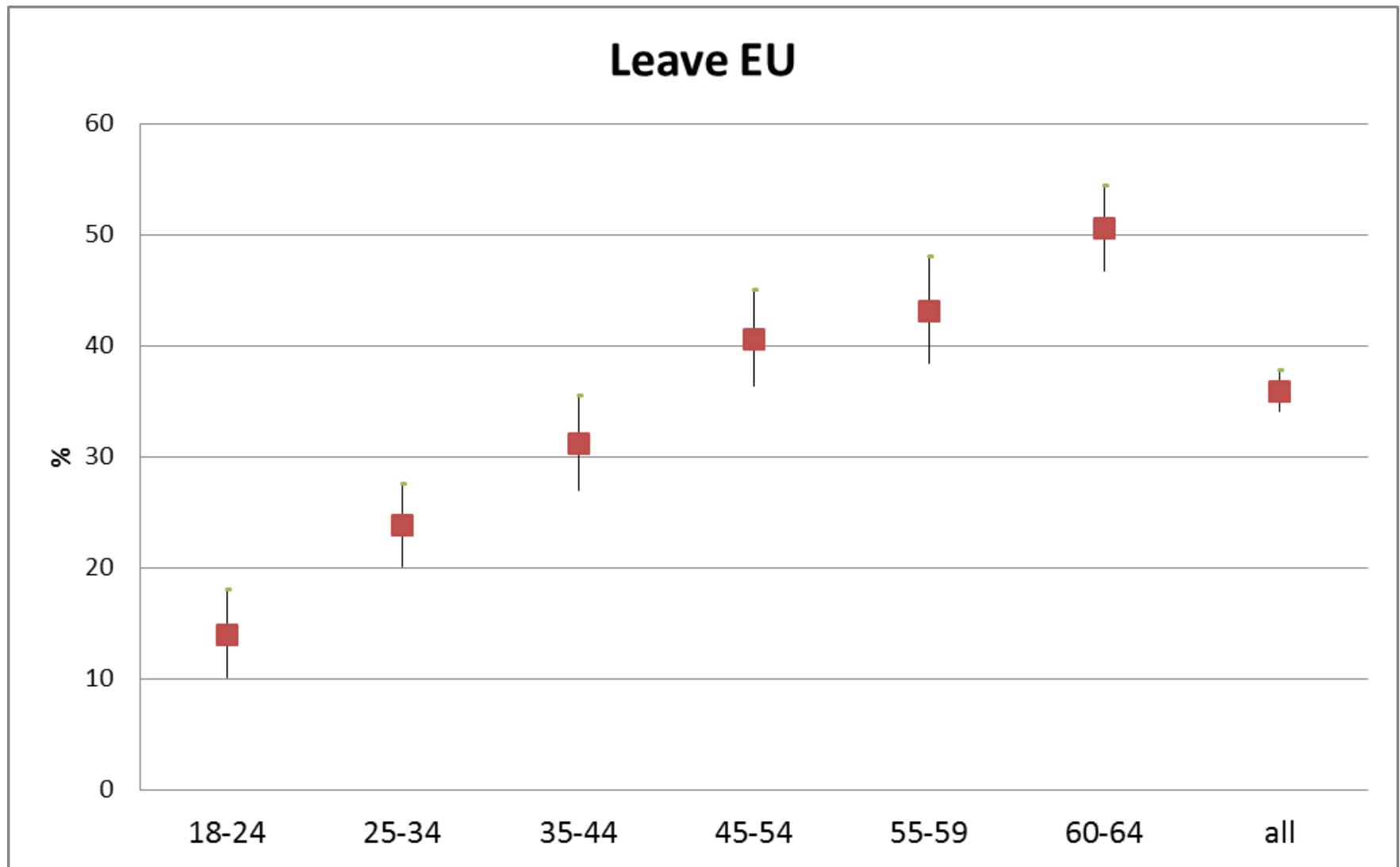
2. Calculate confidence intervals

e.g. For aged 18-24

- 1. sample estimate = **14% would vote to leave**
- 2. Standard error

$$SE = \sqrt{\frac{14*(100-14)}{321}} = \mathbf{1.9367}$$

- 3. Confidence interval = 14% +/- (2 x 1.9367)
= **between 10.2% and 17.8%**



Note where the confidence intervals overlap we can not conclude that the difference between ages is statistically significant

WORKED EXAMPLE 2

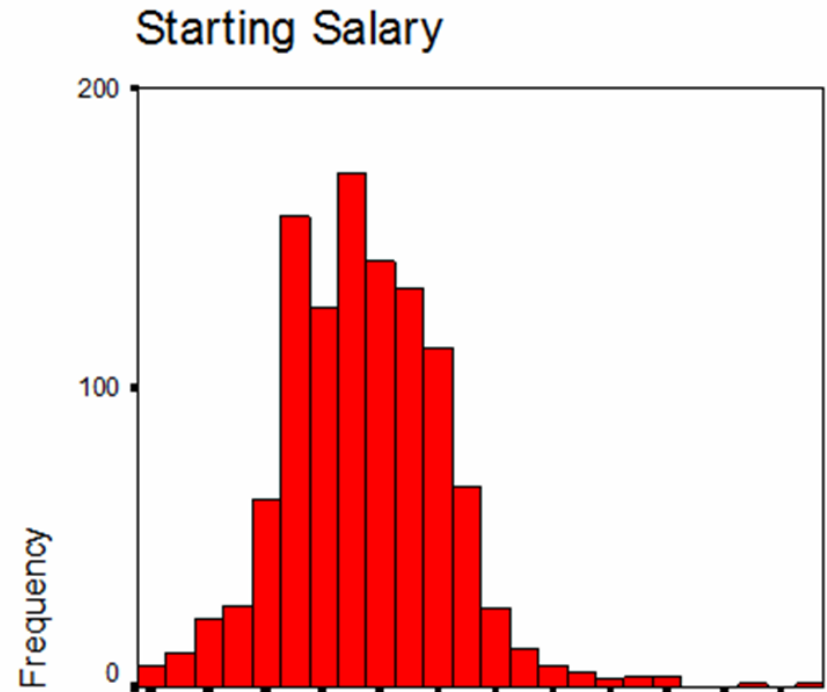
How to calculate a confidence interval if our estimate is a mean rather than a percentage

Calculating sampling error for an estimate of mean starting salary among Manchester graduates

Mean = £16,000

Standard deviation = £5,000

Sample size = 500



How to calculate the **standard error of the mean** for an interval variable

The formula...

$$\text{Standard Error of mean} = \frac{\text{standard deviation}}{\text{square root of sample size}}$$

Calculating Standard Error: starting salary of Manchester graduates

- In our example....

Standard Error =

$$\frac{5,000}{\sqrt{500}}$$

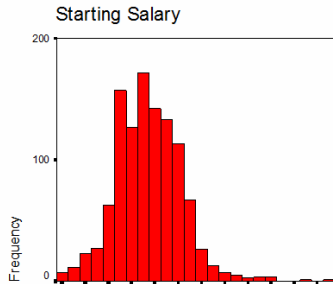
Standard deviation for starting salary in our sample

Symbol for square root

= 224 (£)

Sample size

Mean Starting Salary of Manchester Graduates



Mean = £16,000

Standard deviation = £5,000

Sample size = 500

$$\text{Standard Error} = \frac{5,000}{\sqrt{500}}$$

$$= 224 (\text{£})$$

95% confidence intervals = +/- 2 x standard error

$$\text{So..} = \text{£16,000} \pm (2 \times \text{£224})$$

$$= \text{between } \textbf{£15,552 and £16,448}$$

(95% probability that true mean starting income in population of Manchester Graduates lies in that range)

Calculating confidence intervals

So... In our example

95% confidence intervals for mean starting income...

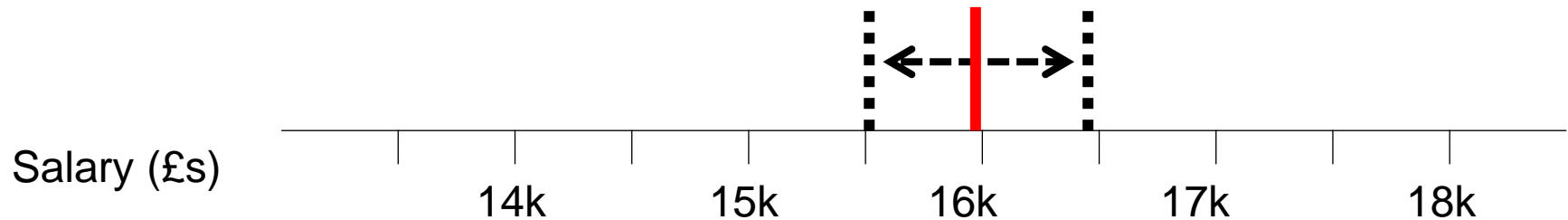
$$= \text{£}16,000 \text{ +/- } (2 \times \text{£}224)$$

$$= \text{between } \text{£}15,552 \text{ and } \text{£}16,448$$

(95% confident that true mean lies in that range)

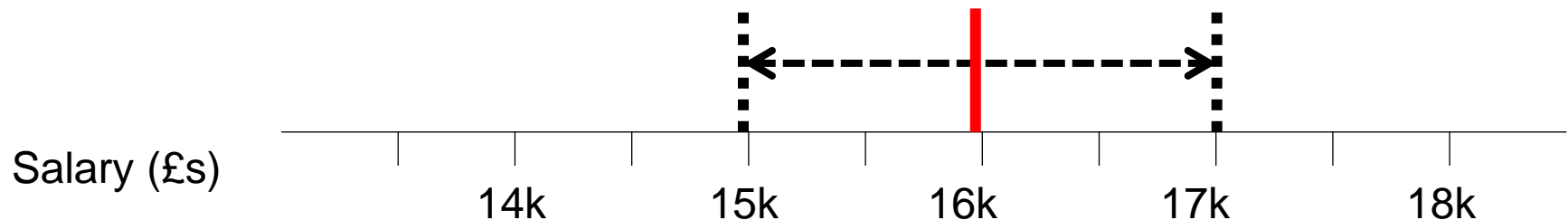
Influence of **sample size** on confidence intervals

We found if **Sample = 500**; mean = £16,000; st deviation = £5,000
95% confidence interval= **£15,552 to £16,448**



□ **A SMALLER SAMPLE** will widen confidence interval...

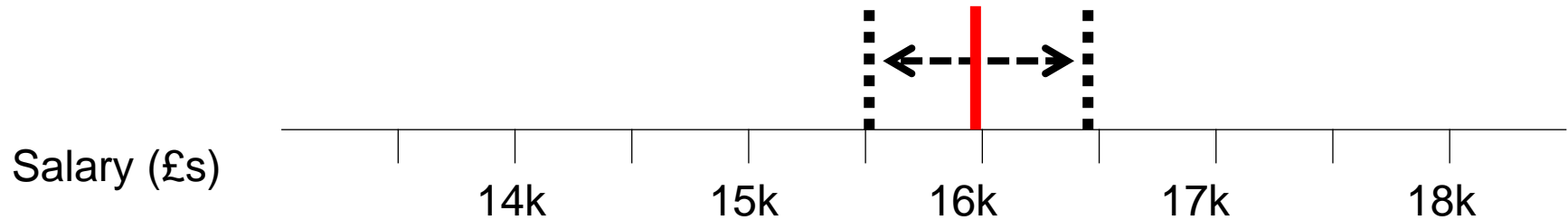
If **Sample = 100**, mean = £16,000; st deviation = £5,000
95% confidence interval= **£15,000 to £17,000**



Influence of greater **variation** on confidence intervals

We found if **Sample = 500**; **mean = £16,000**; **st deviation = £5,000**

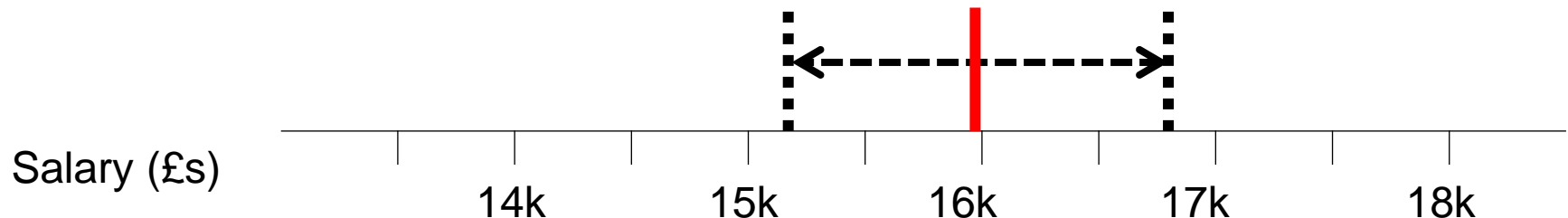
95% confidence interval= **£15,552 to £16,448**



□ **A BIGGER STANDARD DEVIATION** *will widen confidence interval...*

If Sample = 500, mean = £16,000; **st deviation = £8,000**

95% confidence interval= **£15,284 to £16,716**



Influence of sample size and variation on confidence intervals

So...

- A smaller sample or a bigger standard deviation will **widen confidence interval**...

Or

- A bigger sample or a smaller standard deviation will **narrow confidence interval**...

Recap

Confidence Intervals

- We use information from the survey about **sample size** and **variation** in a simple formula to calculate **the standard error**
- Once calculated, the standard error can be used as a building block to calculate **'confidence intervals'** around our survey estimate
- Typically in social science we work with **'95% confidence intervals'**
- Essentially these equal the range within which we can be 95% confident the true value for the population lies
- 95% confidence interval = sample estimate + or - 2 x standard error

NOTE: THE THEORY BEHIND THIS REQUIRES THAT THE SAMPLE WAS A RANDOM SAMPLE

Part 2

Exploring relationships in survey data

(when our data is **interval level**)

First things first...

- **What is Interval level data?**
- **Levels of Measurement**
 - Important concept to grasp because the level of measurement of a variable determines the techniques we use to analyse it

A recap...

Where **both** variables are **categorical**...

Crosstabulation...

Crosstab

| | Age of respondent(grouped)<6 category> dv | | | | | | Total |
|---------------------------------------|---|--------------|--------------|--------------|--------------|--------------|---------------|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65+ | |
| Remain a member of the European Union | 214 66.7% | 323 60.1% | 251 53.3% | 248 46.9% | 202 46.7% | 276 40.4% | 1514 50.9% |
| Leave the European Union | 45 14.0% | 128 23.8% | 147 31.2% | 215 40.6% | 187 43.2% | 346 50.6% | 1068 35.9% |
| I would not vote | 36 11.2% | 49 9.1% | 47 10.0% | 33 6.2% | 28 6.5% | 29 4.2% | 222 7.5% |
| Prefer not to say/don't know | 26 8.1% | 37 6.9% | 26 5.5% | 33 6.2% | 16 3.7% | 33 4.8% | 171 5.7% |
| Total | 321 100% | 537 100% | 471 100% | 529 100% | 433 100% | 684 100% | 2975 100% |

When one variable is **categorical** and the other is **interval**

Suppose we want to look at relationship between..

- levels of alcohol consumption (measured in units of alcohol, an **interval** variable)

and

- gender (**categorical** variable)

(D) Total units of alcohol/week

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|-------|-----|-----------|---------|---------------|--------------------|
| Valid | .00 | 849 | 15.8 | 16.3 | 16.3 |
| | .03 | 58 | 1.1 | 1.1 | 17.5 |
| | .04 | 49 | .9 | .9 | 18.4 |
| | .06 | 2 | .0 | .0 | 18.4 |
| | .06 | 74 | 1.4 | 1.4 | 19.9 |
| | .07 | 6 | .1 | .1 | 20.0 |
| | .07 | 29 | .5 | .6 | 20.5 |
| | .09 | 45 | .8 | .9 | 21.4 |
| | .09 | 1 | .0 | .0 | 21.4 |
| | .10 | 8 | .1 | .1 | 21.6 |
| | .12 | 17 | .3 | .3 | 21.9 |
| | .12 | 36 | .7 | .7 | 22.6 |
| | .12 | 1 | .0 | .0 | 22.6 |
| | .13 | 6 | .1 | .1 | 22.7 |
| | .14 | 3 | .0 | .1 | 22.8 |
| | .15 | 13 | .2 | .3 | 23.0 |
| | .16 | 5 | .1 | .1 | 23.1 |
| | .16 | 4 | .1 | .1 | 23.2 |
| | .17 | 18 | .3 | .4 | 23.5 |
| | .17 | 4 | .1 | .1 | 23.6 |
| | .17 | 13 | .2 | .2 | 23.9 |
| | .19 | 1 | .0 | .0 | 23.9 |
| | .19 | 2 | .0 | .0 | 23.9 |
| | .20 | 4 | .1 | .1 | 24.0 |
| | .20 | 2 | .0 | .0 | 24.0 |

□ Outliers...

| | | | | | |
|---------|-------------------|------|-------|-------|-------|
| | 203.12 | 1 | .0 | .0 | 99.8 |
| | 217.20 | 1 | .0 | .0 | 99.8 |
| | 224.00 | 1 | .0 | .0 | 99.9 |
| | 226.63 | 1 | .0 | .0 | 99.9 |
| | 232.00 | 1 | .0 | .0 | 99.9 |
| | 237.25 | 2 | .0 | .0 | 99.9 |
| | 378.06 | 1 | .0 | .0 | 100.0 |
| | 381.49 | 1 | .0 | .0 | 100.0 |
| | 385.00 | 1 | .0 | .0 | 100.0 |
| | 595.00 | 1 | .0 | .0 | 100.0 |
| | Total | 5160 | 96.7 | 100.0 | |
| Missing | No answer/refused | 94 | 1.8 | | |
| | Don't know | 82 | 1.5 | | |
| | Total | 176 | 3.3 | | |
| Total | | 5337 | 100.0 | | |

-
- Option 1:
 - Recode the alcohol consumption variable into groups (categorical variable) and use in a crosstab

(D) Alcohol units per week grouped * Sex Crosstabulation

| | | Sex | | Total |
|----------|-----------------------------------|----------------|----------------|----------------|
| | | Male | Female | |
| totalwug | Non-drinker/not in last 12 months | 287 11.7% | 503 18.7% | 790 15.4% |
| | Non-zero, but under 1 | 202 8.3% | 436 16.2% | 638 12.4% |
| | 1-7 | 664 27.1% | 893 33.2% | 1557 30.3% |
| | Over 7-10 | 185 7.6% | 207 7.7% | 392 7.6% |
| | Over 10-14 | 235 9.6% | 163 6.1% | 398 7.7% |
| | Over 14-21 | 293 12.0% | 191 7.1% | 484 9.4% |
| | Over 21-28 | 208 8.5% | 122 4.5% | 330 6.4% |
| | Over 28-35 | 131 5.4% | 69 2.6% | 200 3.9% |
| | Over 35-50 | 114 4.7% | 55 2.0% | 169 3.3% |
| | Over 50 | 127 5.2% | 52 1.9% | 179 3.5% |
| | Total | 2446 100.0% | 2691 100.0% | 5137 100.0% |

Units of alcohol by sex
(recode of units of
alcohol)

-
- Option 2
 - Leave alcohol consumption as an interval variable – **compare means** between men and women

Comparing means for different groups

- Mean units of alcohol per week by sex

Report

totalwu

| Sex | Mean | N | Std. Deviation |
|--------|---------|------|----------------|
| Male | 15.0234 | 2459 | 24.48427 |
| Female | 7.9853 | 2701 | 18.12036 |
| Total | 11.3392 | 5160 | 21.67538 |

Where **BOTH VARIABLES ARE INTERVAL...**

- We can view the relationship graphically using **scatter graphs**
- We can determine the direction and strength of relationship using a measure of **correlation**
- We can use a technique called **regression** to describe the relationship (and make predictions) using a **statistical model**

N.B. while it is possible to turn interval variables into categorical variables and just use cross-tabulation it is better to use correlation and regression if you have interval level data as they make full use of the detail available (e.g. recoding income (interval) into income groups (categorical) inevitably loses some of the original 'detail' contained within the income variable)

SPSS Practical

Today's data set...

- World Bank Development Indicators:
- **'WorldBank2017.sav'**
- Note the cases in this dataset are 'countries' not 'individual people'
-
- So variables contain information about each country e.g. the U5 mortality rate and the % of the country urban
- We'll use U5 Mortality rate as our dependent variable in rest of the practical
-

WorldBank2017

| | Variable Name | Variable Label |
|----|--------------------------|--|
| 1 | CountryName | Country Name |
| 2 | maternalmortality | Maternal mortality ratio (modeled estimate, per 100,000 live births) |
| 3 | U5mortality | Mortality rate, under-5 (per 1,000 live births) [|
| 4 | Infantmortality | Mortality rate, infant (per 1,000 live births) |
| 5 | Adolescentfertility | Adolescent fertility rate (births per 1,000 women ages 15-19) |
| 6 | TotalFertilityRate | Fertility rate, total (births per woman) |
| 7 | Lifeexpectancy | Life expectancy at birth, total (years) |
| 8 | Urbanpopulation | Urban population (% of total) |
| 9 | basicdrinkingwater | People using at least basic drinking water services (% of population) |
| 10 | basicsanitation | People using at least basic sanitation services (% of population) |
| 11 | SkilledBirthAttendant | Births attended by skilled health staff (% of total) |
| 12 | Healthspend_percentofGDP | Current health expenditure (% of GDP) |
| 13 | Healthspend_percapita | Current health expenditure per capita (current US\$) |
| 14 | womeninparliament | Proportion of seats held by women in national parliaments (%) |
| 15 | InternetUse | Individuals using the Internet (% of population) |
| 16 | EmployedInServices | Employment in services (% of total employment) (modeled ILO estimate) |
| 17 | EmployedInAgriculture | Employment in agriculture (% of total employment) (modeled ILO estimate) |
| 18 | Doctors | Physicians (per 1,000 people) |
| 19 | NursesMidwives | Nurses and midwives (per 1,000 people) |
| 20 | Electricity | Access to electricity (% of population) |
| 21 | Popgrowth | Population growth (annual %) |
| 22 | UrbanPopGrowth | Urban population growth (annual %) |
| 23 | GNIpercapitagrowth | GNI per capita growth (annual %) |
| 24 | GDPpercapitagrowth | GDP per capita growth (annual %) |
| 25 | GNIpercapita | GNI per capita, Atlas method (current US\$) |
| 26 | NationalDebt | Present value of external debt (% of GNI) |
| 27 | Fem_sced | Educational attainment, at least completed lower secondary, population 25+, female (%) |
| 28 | Sced_percentfemale | Secondary education, general pupils % female |
| 29 | CompusoryEd | Compulsory education, duration (years) |

Mortality rate, under-5 (per 1,000 live births) [

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|-------|------|-----------|---------|---------------|--------------------|
| Valid | 2.10 | 2 | 1.1 | 1.1 | 1.1 |
| | 2.30 | 1 | .6 | .6 | 1.7 |
| | 2.60 | 3 | 1.7 | 1.7 | 3.4 |
| | 2.70 | 2 | 1.1 | 1.1 | 4.5 |
| | 2.80 | 2 | 1.1 | 1.1 | 5.6 |
| | 3.10 | 1 | .6 | .6 | 6.2 |
| | 3.30 | 3 | 1.7 | 1.7 | 7.9 |
| | 3.40 | 1 | .6 | .6 | 8.5 |
| | 3.50 | 3 | 1.7 | 1.7 | 10.2 |
| | 3.60 | 2 | 1.1 | 1.1 | 11.3 |
| | 3.70 | 3 | 1.7 | 1.7 | 13.0 |
| | 3.80 | 1 | .6 | .6 | 13.6 |
| | 3.90 | 1 | .6 | .6 | 14.1 |
| | 4.20 | 3 | 1.7 | 1.7 | 15.8 |
| | 4.30 | 3 | 1.7 | 1.7 | 17.5 |
| | 4.50 | 1 | .6 | .6 | 18.1 |
| | 4.60 | 1 | .6 | .6 | 18.6 |
| | 4.70 | 1 | .6 | .6 | 19.2 |
| | 5.10 | 1 | .6 | .6 | 19.8 |
| | 5.30 | 2 | 1.1 | 1.1 | 20.9 |
| | 5.40 | 1 | .6 | .6 | 21.5 |
| | 5.60 | 1 | .6 | .6 | 22.0 |
| | 5.70 | 2 | 1.1 | 1.1 | 23.2 |
| | 6.40 | 1 | .6 | .6 | 23.7 |
| | 6.60 | 1 | .6 | .6 | 24.3 |
| | 7.20 | 1 | .6 | .6 | 24.9 |
| | 7.30 | 1 | .6 | .6 | 25.4 |
| | 7.40 | 3 | 1.7 | 1.7 | 27.1 |
| | 7.50 | 1 | .6 | .6 | 27.7 |
| | 7.60 | 2 | 1.1 | 1.1 | 28.8 |
| | 7.80 | 2 | 1.1 | 1.1 | 29.9 |
| | 7.90 | 2 | 1.1 | 1.1 | 31.1 |

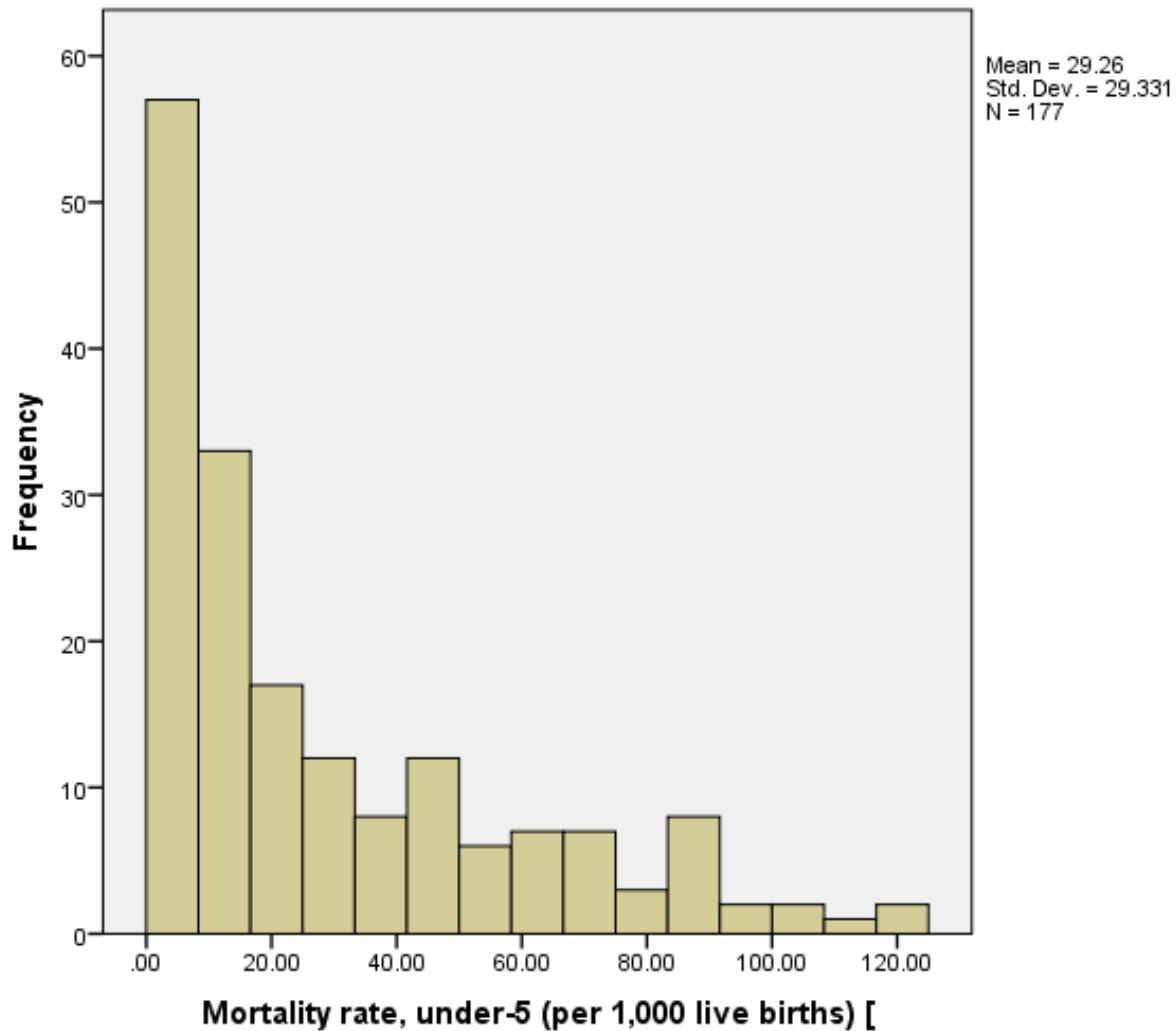
U5 Mortality Rate

(interval level variable)

A better way to show
the distribution?

In SPSS

- Exercise 1



U5 Mortality Rate

(interval level
variable)

Histogram

Statistics

Mortality rate, under-5 (per 1,000 liv

| | | |
|--------------------|---------|-------------------|
| N | Valid | 177 |
| | Missing | 3 |
| Mean | | 29.2598 |
| Std. Error of Mean | | 2.20468 |
| Median | | 16.5000 |
| Mode | | 2.60 ^a |
| Std. Deviation | | 29.33138 |
| Variance | | 860.330 |
| Range | | 121.10 |
| Minimum | | 2.10 |
| Maximum | | 123.20 |

U5 Mortality Rate Summary Statistics

a. Multiple modes exist. The
smallest value is shown

Exploring the relationship between two interval level variables...

- graphically using **scatter graphs**

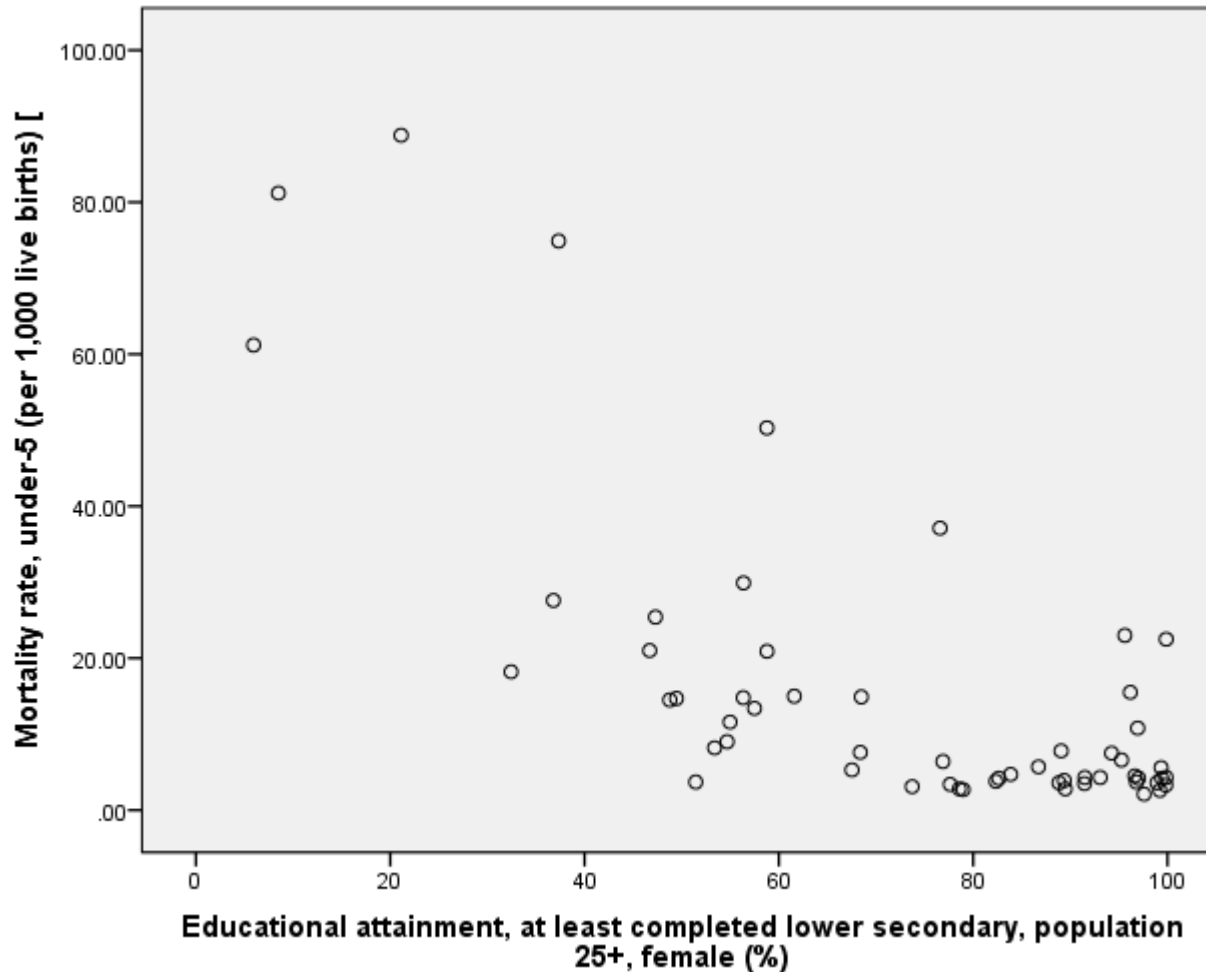
WorldBank2017

| | Variable Name | Variable Label |
|----|--------------------------|--|
| 1 | CountryName | Country Name |
| 2 | maternalmortality | Maternal mortality ratio (modeled estimate, per 100,000 live births) |
| 3 | U5mortality | Mortality rate, under-5 (per 1,000 live births) [|
| 4 | Infantmortality | Mortality rate, infant (per 1,000 live births) |
| 5 | Adolescentfertility | Adolescent fertility rate (births per 1,000 women ages 15-19) |
| 6 | TotalFertilityRate | Fertility rate, total (births per woman) |
| 7 | Lifeexpectancy | Life expectancy at birth, total (years) |
| 8 | Urbanpopulation | Urban population (% of total) |
| 9 | basicdrinkingwater | People using at least basic drinking water services (% of population) |
| 10 | basicsanitation | People using at least basic sanitation services (% of population) |
| 11 | SkilledBirthAttendant | Births attended by skilled health staff (% of total) |
| 12 | Healthspend_percentofGDP | Current health expenditure (% of GDP) |
| 13 | Healthspend_percapita | Current health expenditure per capita (current US\$) |
| 14 | womeninparliament | Proportion of seats held by women in national parliaments (%) |
| 15 | InternetUse | Individuals using the Internet (% of population) |
| 16 | EmployedInServices | Employment in services (% of total employment) (modeled ILO estimate) |
| 17 | EmployedInAgriculture | Employment in agriculture (% of total employment) (modeled ILO estimate) |
| 18 | Doctors | Physicians (per 1,000 people) |
| 19 | NursesMidwives | Nurses and midwives (per 1,000 people) |
| 20 | Electricity | Access to electricity (% of population) |
| 21 | Popgrowth | Population growth (annual %) |
| 22 | UrbanPopGrowth | Urban population growth (annual %) |
| 23 | GNIpercapitagrowth | GNI per capita growth (annual %) |
| 24 | GDPpercapitagrowth | GDP per capita growth (annual %) |
| 25 | GNIpercapita | GNI per capita, Atlas method (current US\$) |
| 26 | NationalDebt | Present value of external debt (% of GNI) |
| 27 | Fem_sced | Educational attainment, at least completed lower secondary, population 25+, female (%) |
| 28 | Seced_percentfemale | Secondary education, general pupils % female |
| 29 | CompusoryEd | Compulsory education, duration (years) |

In SPSS

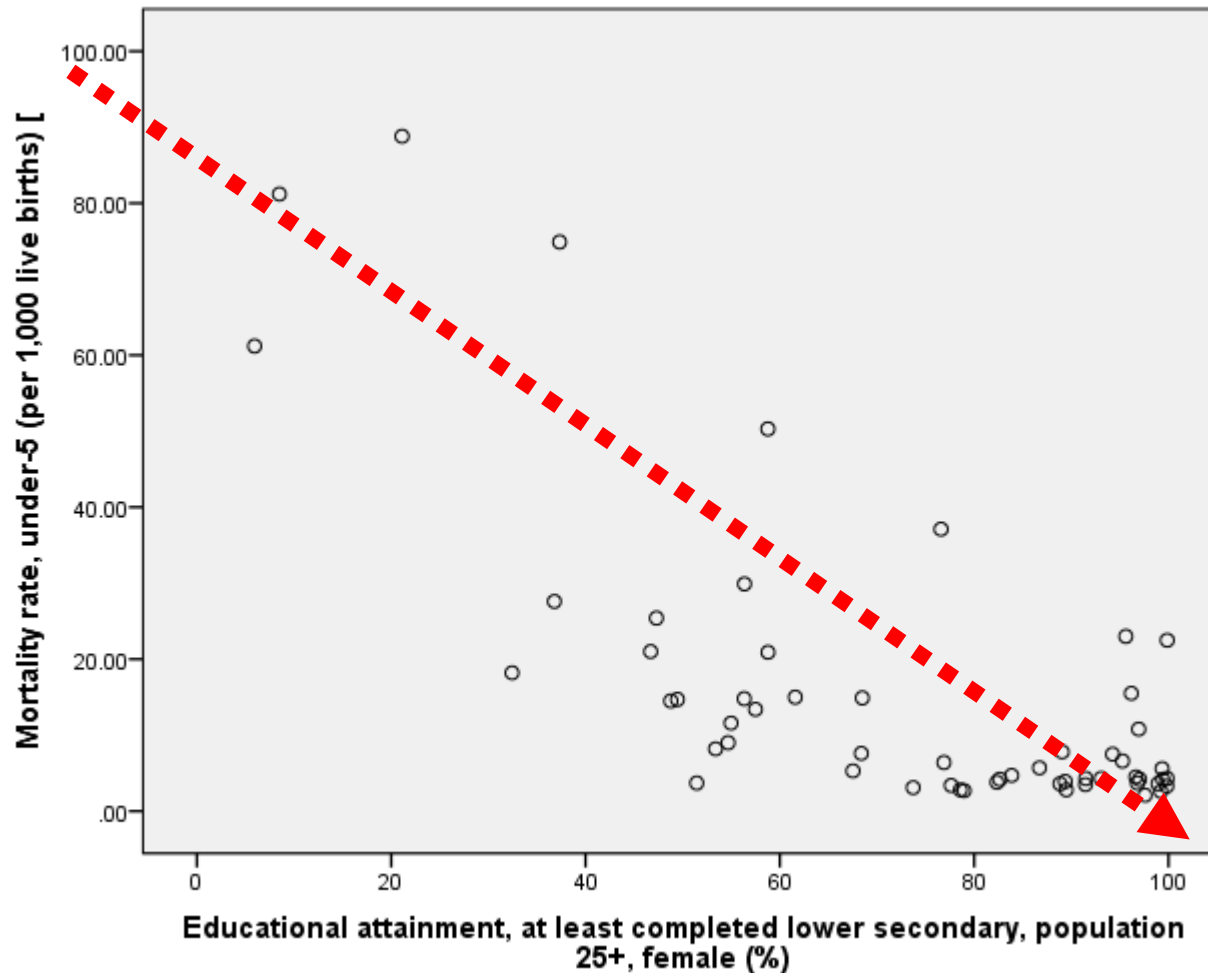
- Exercise 2

U5 Mortality and % of women with lower secondary education



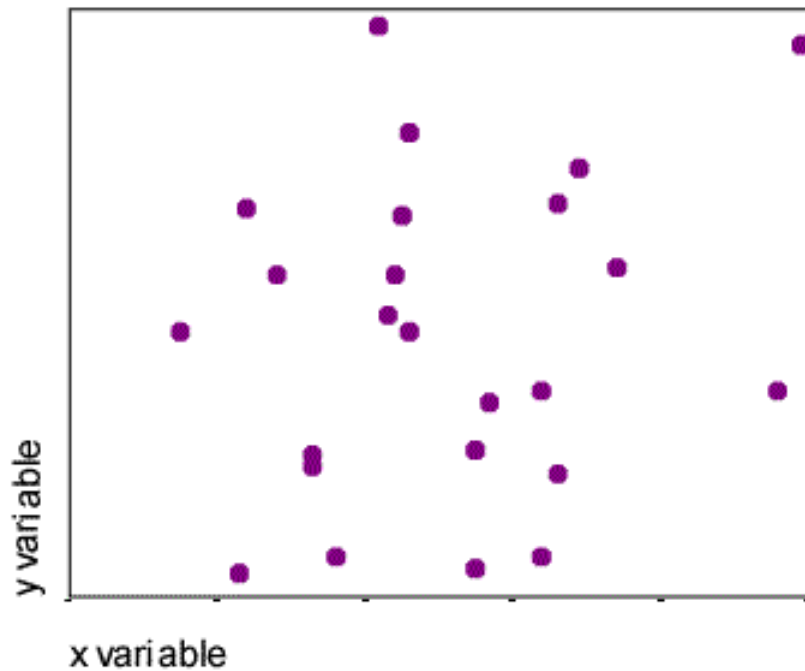
The graph allows a subjective assessment of direction and strength of a relationship, whether it is straight (linear) or curved. Also reveals the presence of any ‘outliers’

U5 Mortality and % of females in secondary education

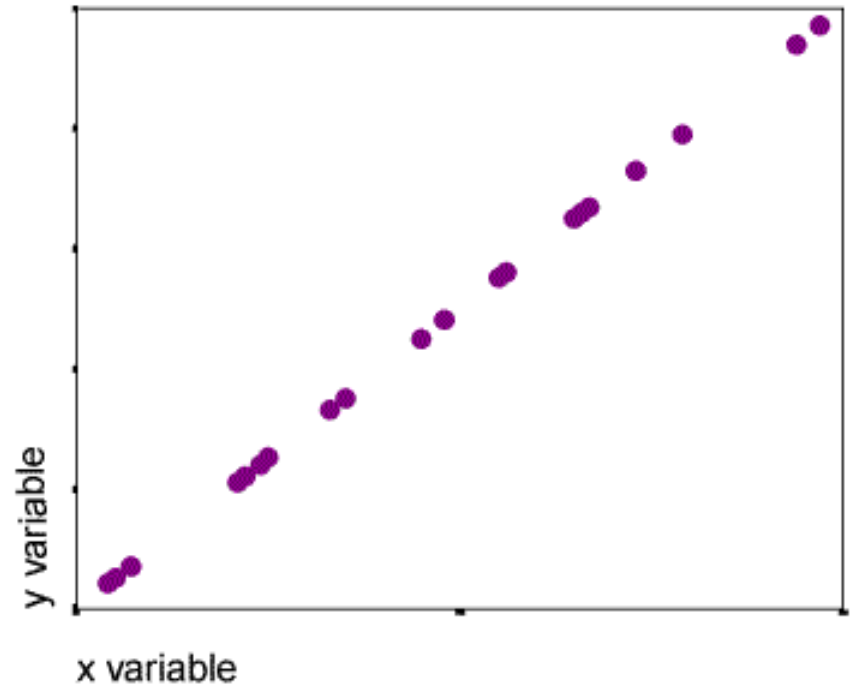


The graph allows a subjective assessment of direction and strength of a relationship, whether it is straight (linear) or curved. Also reveals the presence of any ‘outliers’

Using scatterplots to recognise relationships

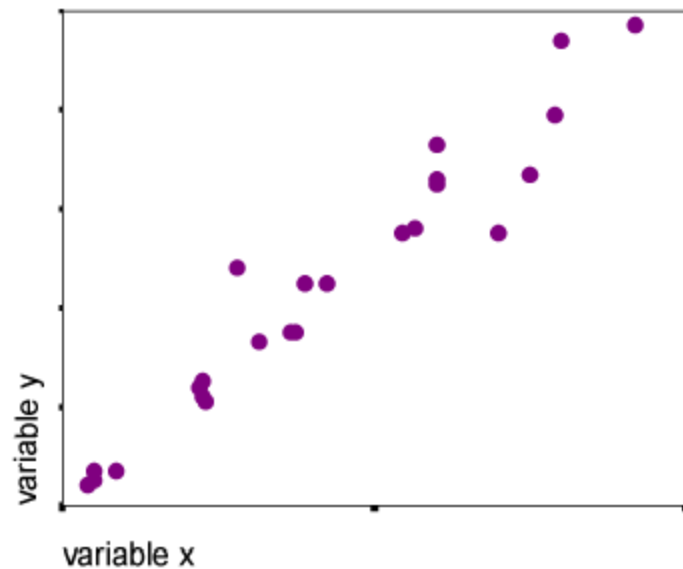


Points randomly scattered = No relationship

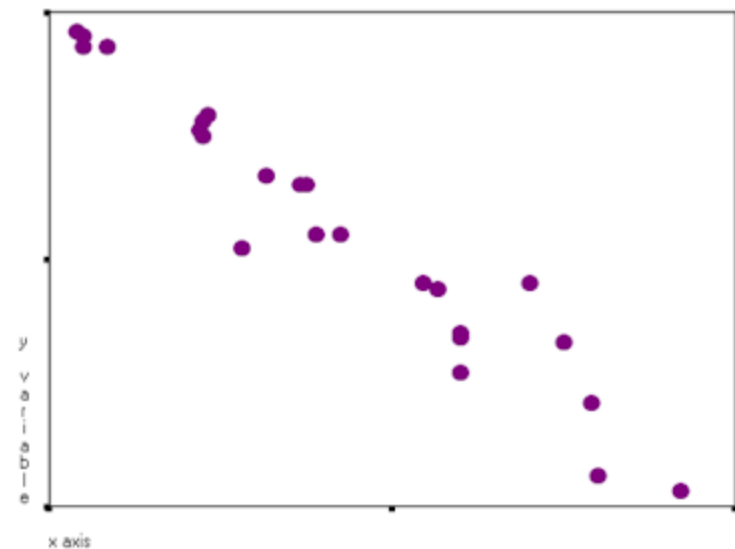


Points on a straight line
= Perfect *linear* relationship
(rare in social science)

Social data is very unlikely to be perfectly related...

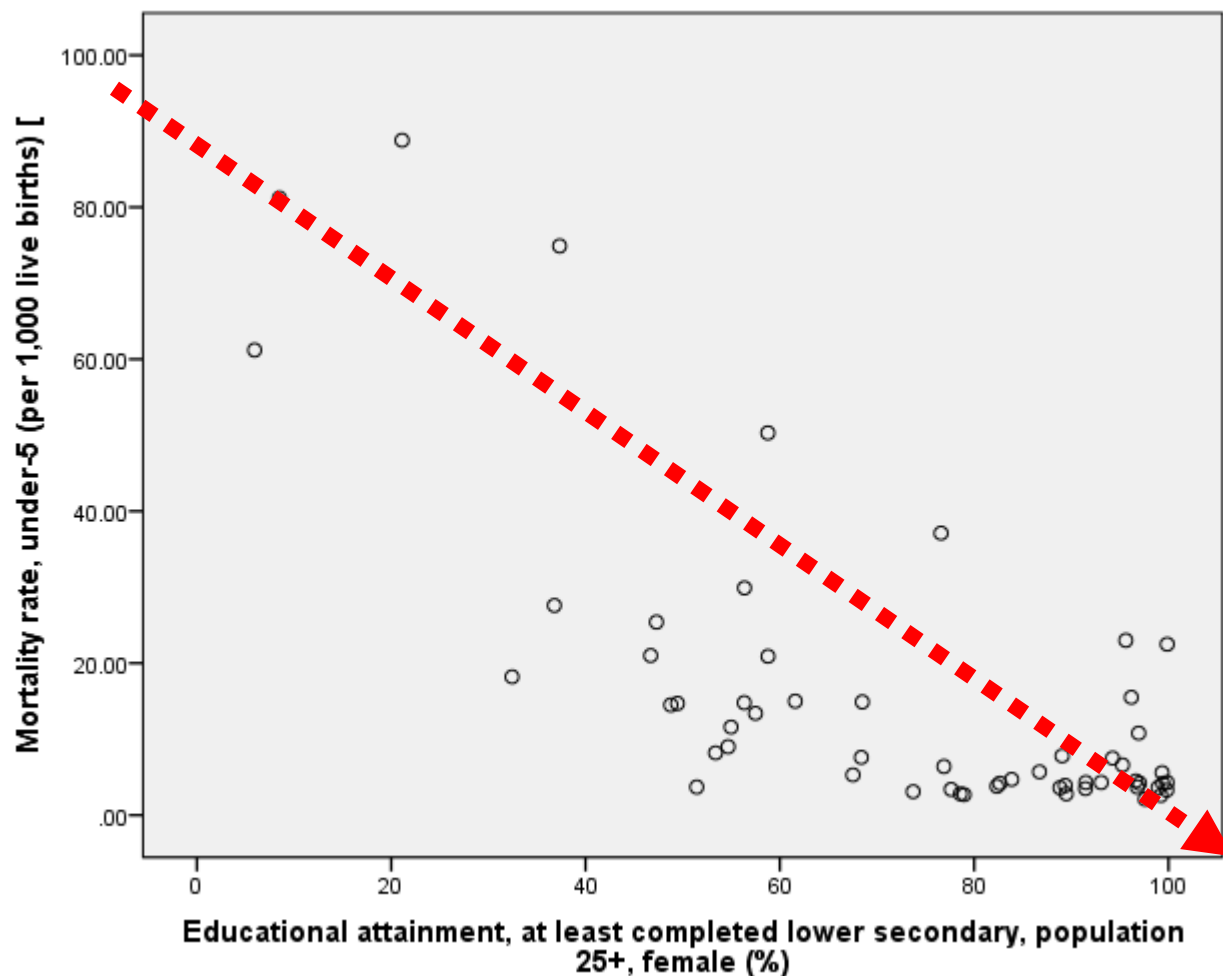


Strong positive relationship



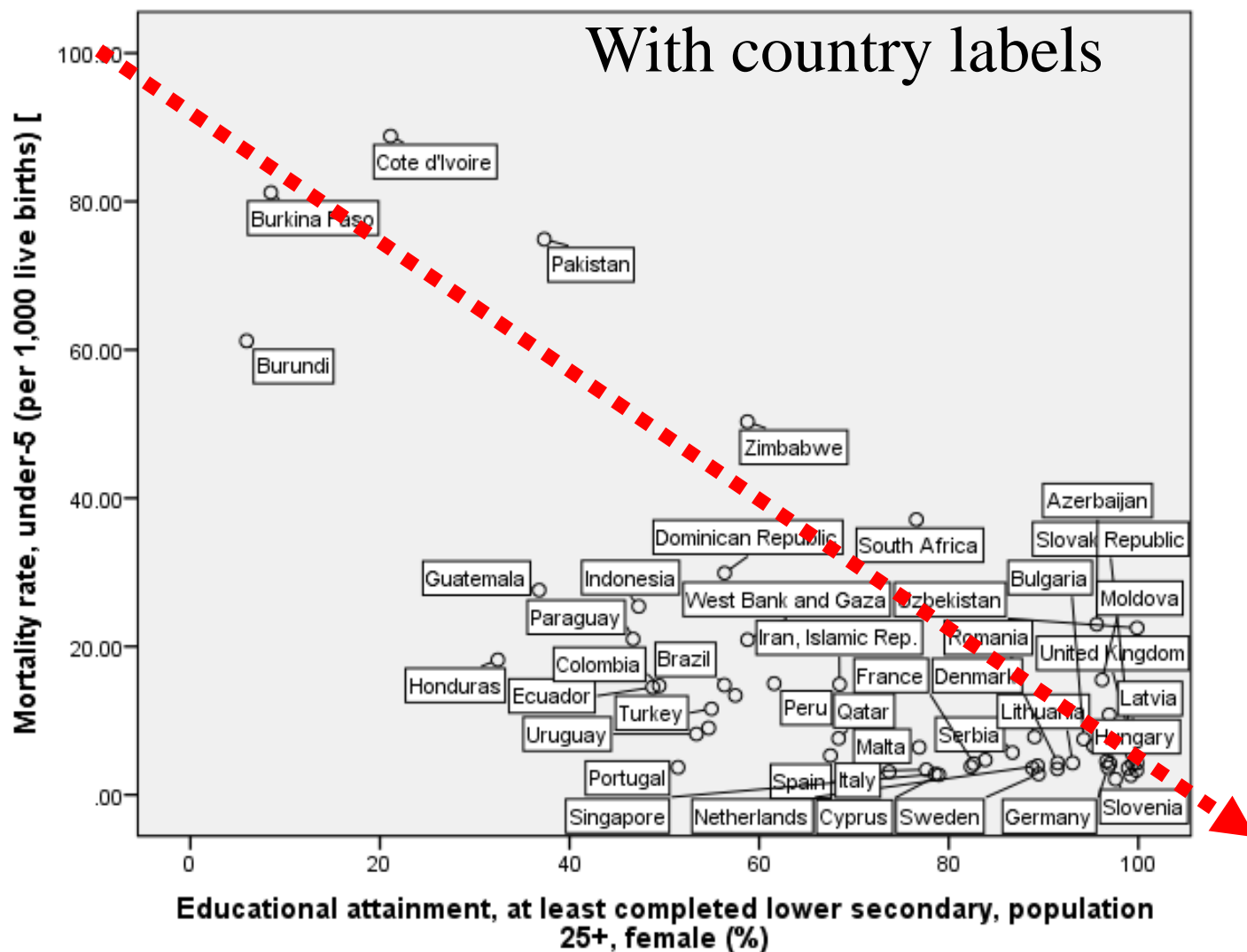
Strong negative relationship

Under 5 mortality and % females in secondary education

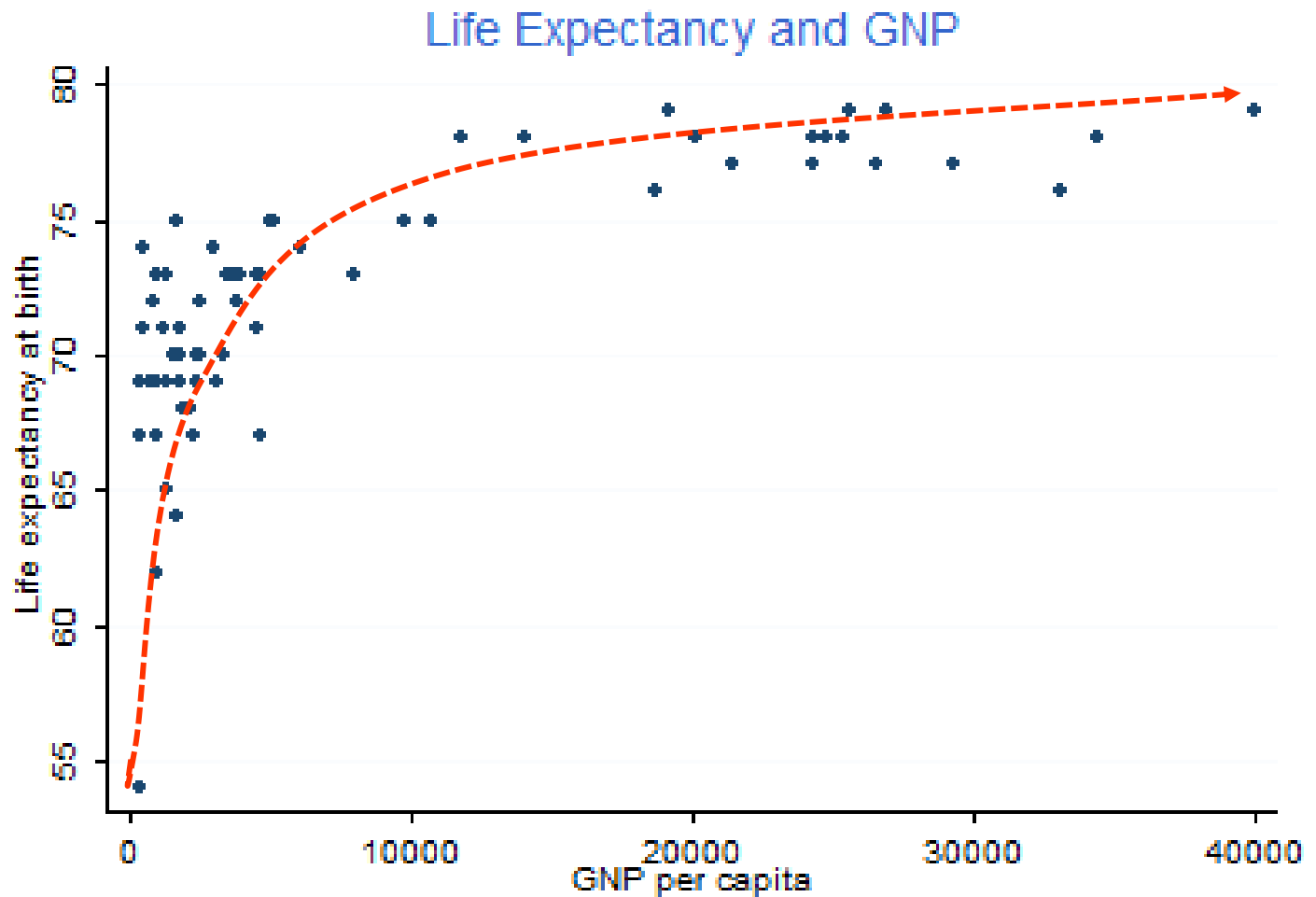


Strong negative
linear relationship

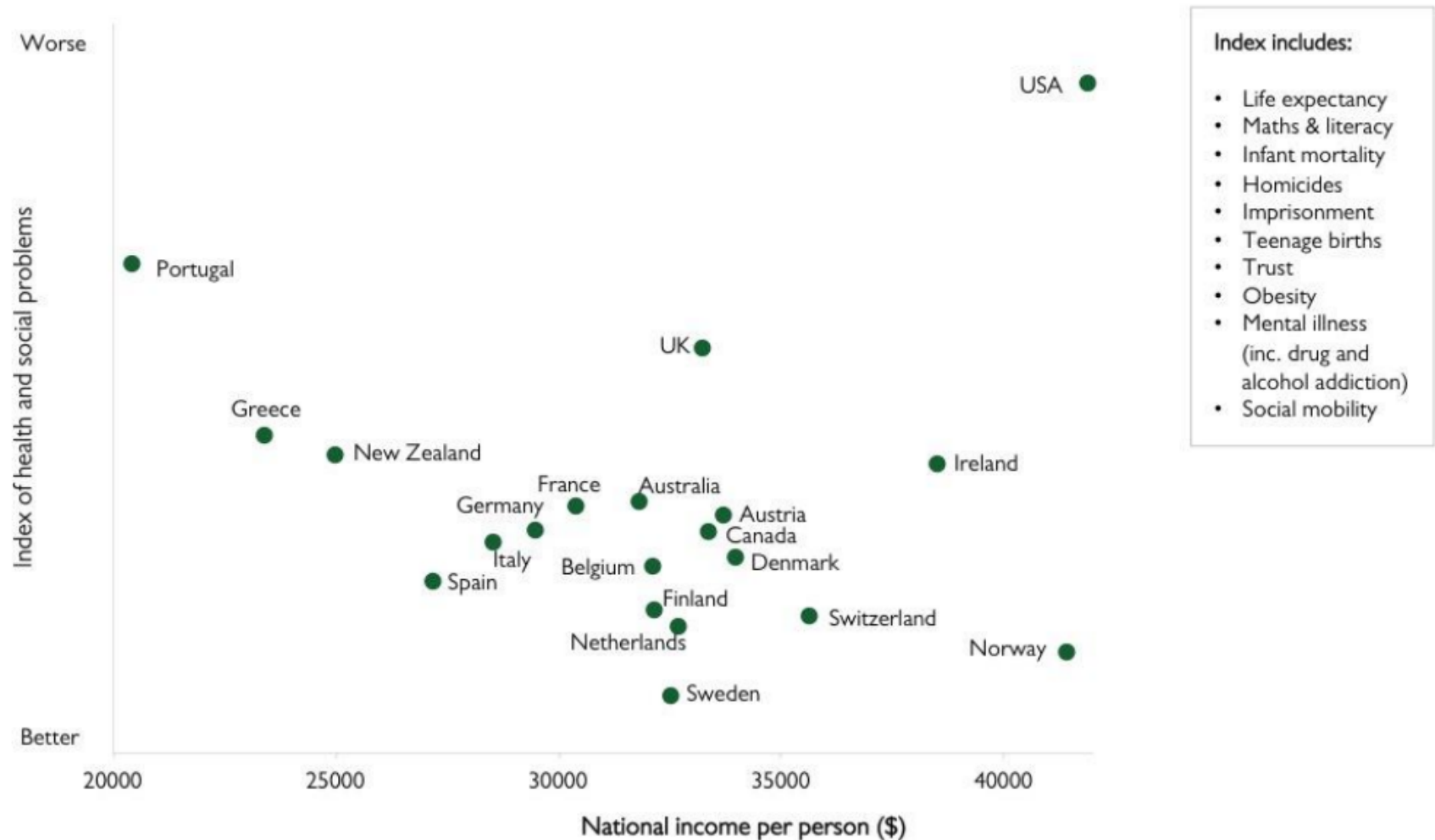
Under 5 mortality and % females in secondary education



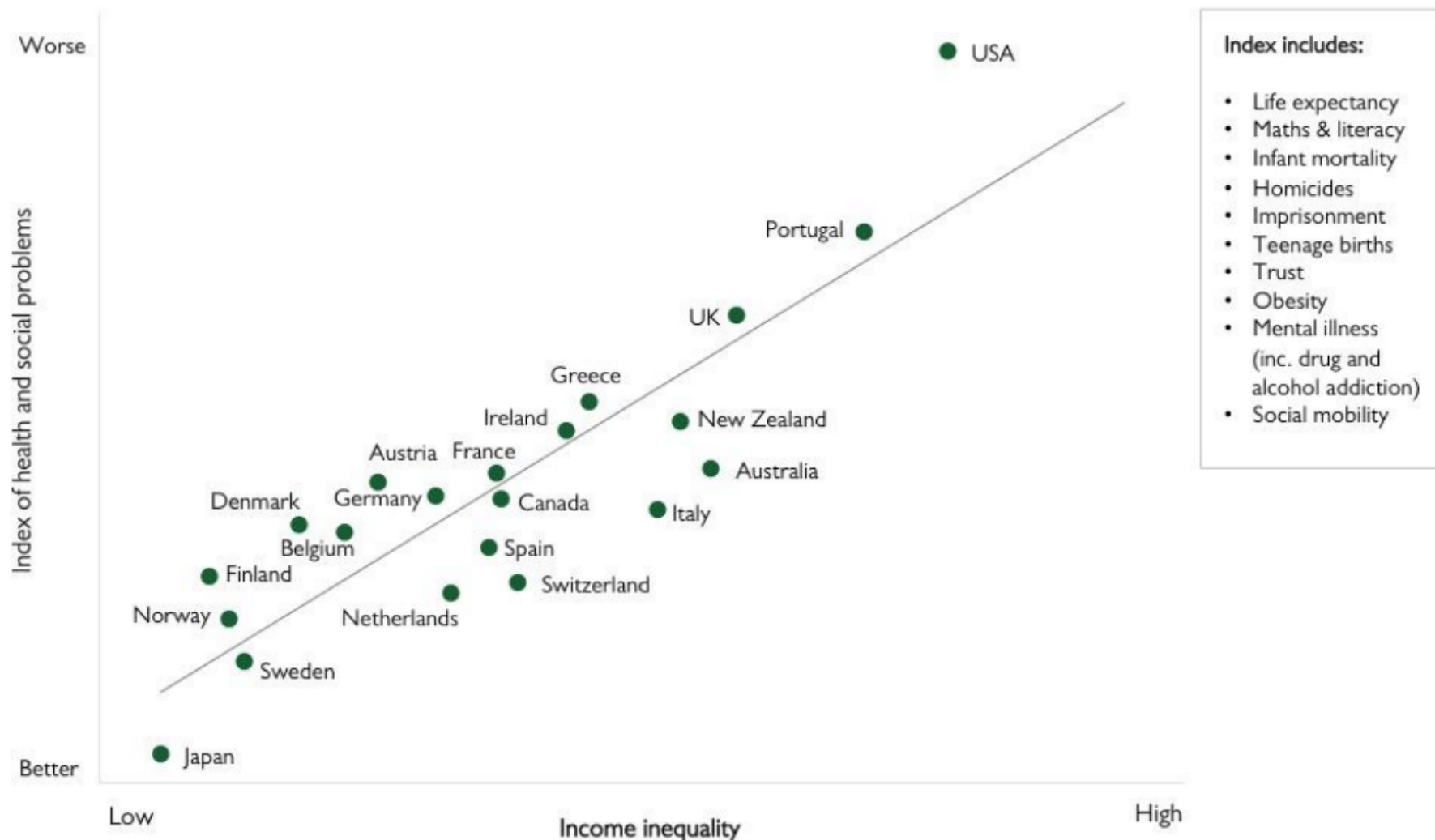
Some relationships are non-linear



Health and social problems are not related to average income in rich countries



Health and social problems are worse in more unequal countries



<http://www.gapminder.org/>

GAPMINDER

a fact-based worldview

GAPMINDER WORLD

VIDEOS

DOWNLOADS

TEACH

IGNORANCE

DATA

Search...



Refresh your world

Pour the sparkling fresh numbers into your eyes and upgrade your worldview.

EXAMPLES:

Wealth & Health of Nations ▶

CO₂ emissions since 1820 ▶

Africa is not a country! ▶

Is child mortality falling? ▶

Where is HIV decreasing? ▶

BUBBLE CHART ▶



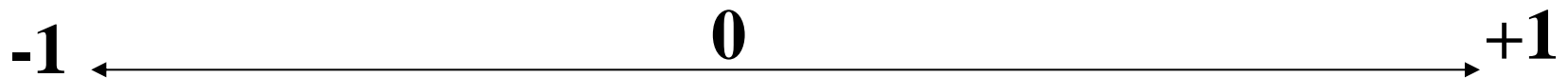
<http://www.gapminder.org/videos/the-joy-of-stats/>

A more objective measure?

- Scatterplots provide a very useful visual representation of the relationship between two variables
- But interpretation is subjective
- We can use statistical measures to give a more objective measure of the relationship:
- **Correlation**
- **Regression**

Correlation to here

- Correlation describes the **strength and direction** of a relationship between two interval variables.
- There are various measures of correlation used. We will focus on **Pearson's Correlation Coefficient (r)** which measures the strength of **linear** (straight line) relationship between two variables
- The coefficient takes a value between -1 and 1, where -1 indicates a perfect negative linear relationship, +1 indicates a perfect positive linear relationship and 0 indicates no linear relationship at all





Perfect negative

Perfect positive

N.B. Pearson's correlation coefficient is only measuring the presence of a linear relationship – it is therefore not a good measure to use if the relationship is non-linear (e.g. a curve line)

Pearsons Correlation (r): interpreting the value

| Negative Range | Description | Positive Range |
|---|---------------|---|
|  | |  |
| 0.00 | None | 0.00 |
| -0.19 - -0.01 | 'Very weak' | 0.01 - 0.19 |
| -0.39 - -0.20 | 'Weak' | 0.20 - 0.39 |
| -0.69 - -0.40 | 'Modest' | 0.40 - 0.69 |
| -0.89 - -0.70 | 'Strong' | 0.70 - 0.89 |
| -0.99 - -0.90 | 'Very strong' | 0.90 - 0.99 |
| -1.00 | Perfect | 1.00 |

extent
to
which
points
cluster
tightly
round
a
straight
line



Calculation of Pearson Correlation

- It is straightforward to request a correlation coefficient in SPSS. So do I need to know how to calculate it?
- What is most important is that you understand when to use it, what it is measuring and how to interpret it. You will rarely have to calculate the statistic by hand
- HOWEVER, understanding how it is calculated can help your understanding of its meaning and is important if you anticipate going on to use more advanced statistics (a fuller explanation is provided in a longer version of these slides) .

Pearson Correlation Assumptions

- Variables must be interval level
- Relationship must be linear
- Variables should really be normally distributed (for our purposes we won't be too strict about this one, but where variable distributions are very skewed it's worth considering Spearman's correlation – see next)

In SPSS

- Exercise 3

PEARSON CORRELATION

Correlations

| | | Mortality rate, under-5 (per 1,000 live births) [| Educational attainment, at least completed lower secondary, population 25+, female (%) |
|--|---------------------|---|--|
| Mortality rate, under-5 (per 1,000 live births) [| Pearson Correlation | 1 | -.717** |
| | Sig. (2-tailed) | | .000 |
| | N | 177 | 56 |
| Educational attainment, at least completed lower secondary, population 25+, female (%) | Pearson Correlation | -.717** | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 56 | 57 |

** . Correlation is significant at the 0.01 level (2-tailed).

Is our measure of correlation statistically significant?

- With any analysis based on a **sample** we should ask whether the correlation value can be generalised to the population
- i.e. we need to do a test of statistical significance to see if we can reject the 'null hypothesis' (null hypothesis = that there is no relationship between the two variables in the population i.e. that the correlation value is = zero)

Pearson Correlation

Correlations

| | | Mortality rate, under-5 (per 1,000 live births) [| Educational attainment, at least completed lower secondary, population 25+, female (%) |
|--|---------------------|---|--|
| Mortality rate, under-5 (per 1,000 live births) [| Pearson Correlation | | -.717** |
| | Sig. (2-tailed) | | .000 |
| | N | 177 | 56 |
| Educational attainment, at least completed lower secondary, population 25+, female (%) | Pearson Correlation | -.717** | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 56 | 57 |

The value for r (0.717)

** . Correlation is significant at the 0.01 level (2-tailed).

The p value (<0.000)

- Is correlation significantly different from 0?
- Null hypothesis that $r = 0$
- To be statistically significant we are looking for a p value < 0.05 (5%)
- In this example we see $p < 0.05$ so conclude that r is not equal to 0, therefore U5 Mortality and % women with sec ed are correlated

An alternative measure: Spearman's Correlation

- Use where a scatterplot suggests the relationship may be curvilinear
- Can also be used for ordinal level data.
- Better than Pearson if variables have skewed distributions
- Calculation is based on **ranks** (the position of a data point once the data has been ordered from lowest to highest) not the actual value .
- The statistic computes the Pearson correlation for the ranks (rather than the actual values).
- Like Pearson correlation it takes a value between -1 and 1, with the same interpretation of magnitudes and direction.

Spearman's

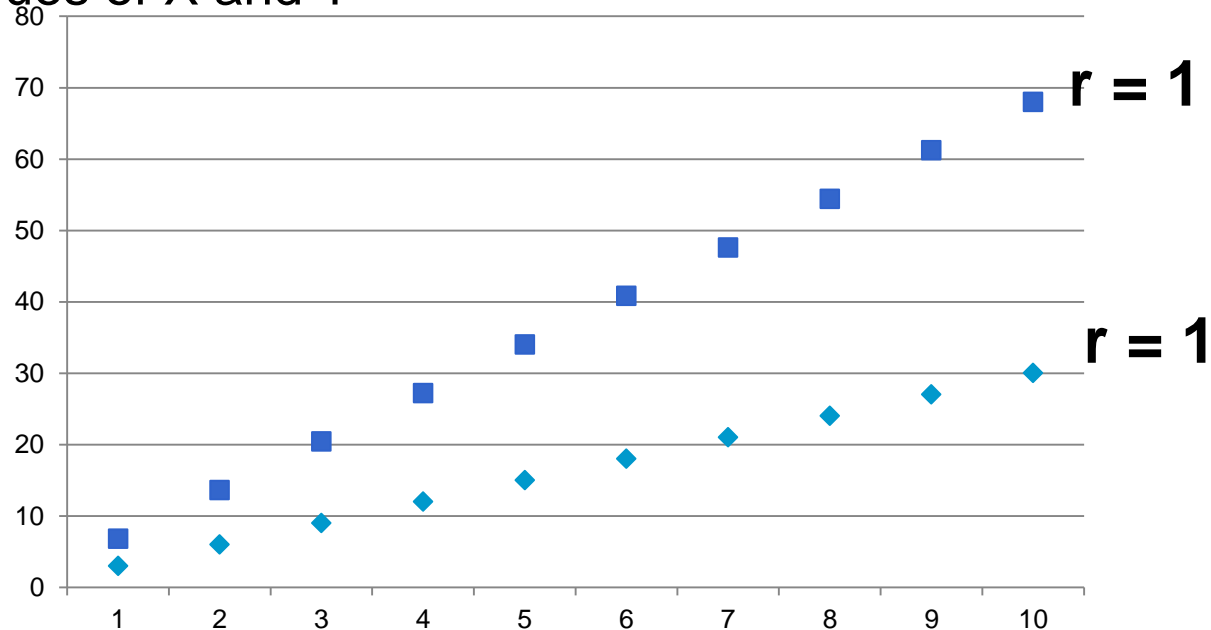
Correlations

| | | | Mortality rate, under-5 (per 1,000 live births) [| Educational attainment, at least completed lower secondary, population 25+, female (%) |
|----------------|--|-------------------------|---|--|
| Spearman's rho | Mortality rate, under-5 (per 1,000 live births) [| Correlation Coefficient | 1.000 | -.587** |
| | | Sig. (2-tailed) | . | .000 |
| | | N | 177 | 56 |
| | Educational attainment, at least completed lower secondary, population 25+, female (%) | Correlation Coefficient | -.587** | 1.000 |
| | | Sig. (2-tailed) | .000 | . |
| | | N | 56 | 57 |

** . Correlation is significant at the 0.01 level (2-tailed).

The limits of correlation

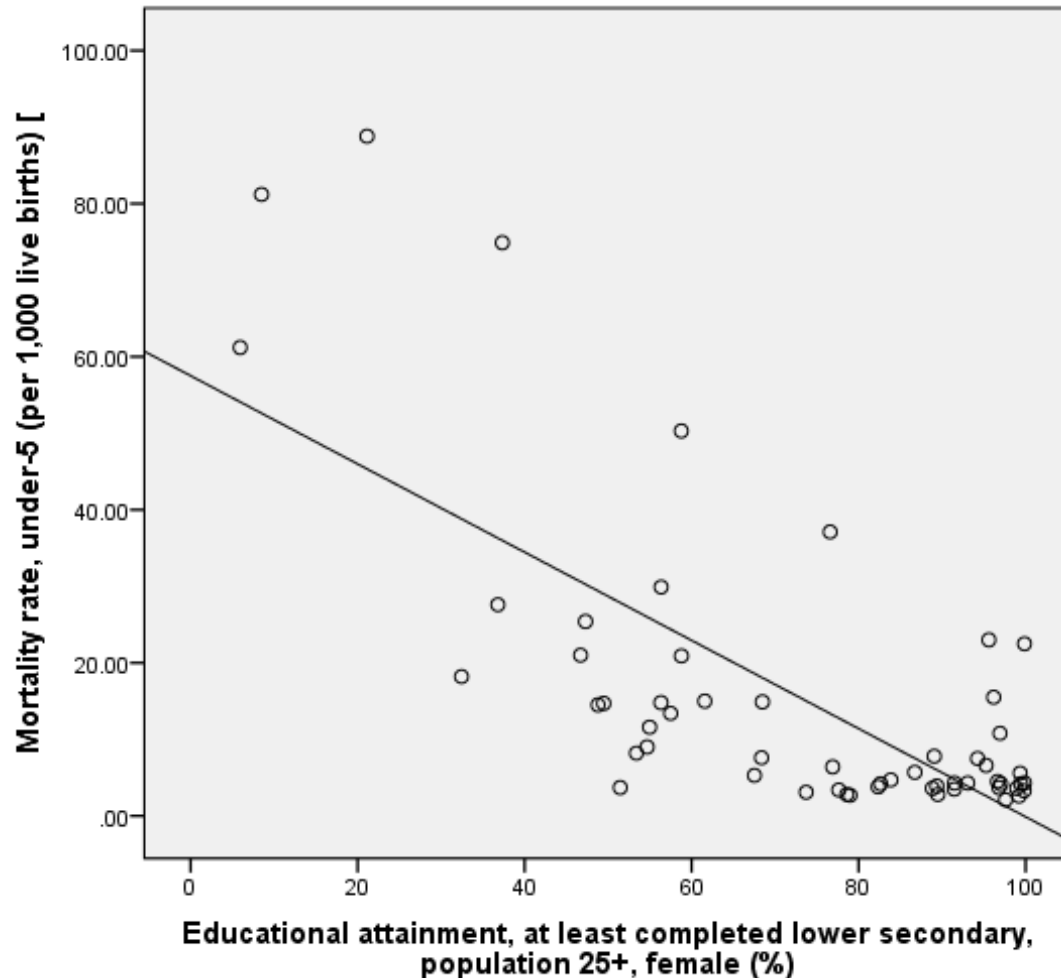
- Correlation provides a measure of how strong a linear relationship is (between 0 and 1) but it doesn't tell you about the actual relationship is in terms of the values of each variable (how the value of y changes with x)
- E.g. the graph shows two relationships - both have a Pearson correlation value of 1.0 (perfect correlation), but they are different in the relationship between values of X and Y



Simple Linear Regression

- Use simple regression to describe the linear relationship between an independent (explanatory) variable and a dependent (response) variable.
- The independent (explanatory) variable is used to **predict** values for the dependent (response) variable.
- Procedure involves generating a statistical model which describes the nature of the relationship between the two variables (by fitting a straight line of best fit through the data points).

Simple Linear Regression (overview)



□ With simple regression we fit a line to the data to show relationship between a dependent variable and a predictor

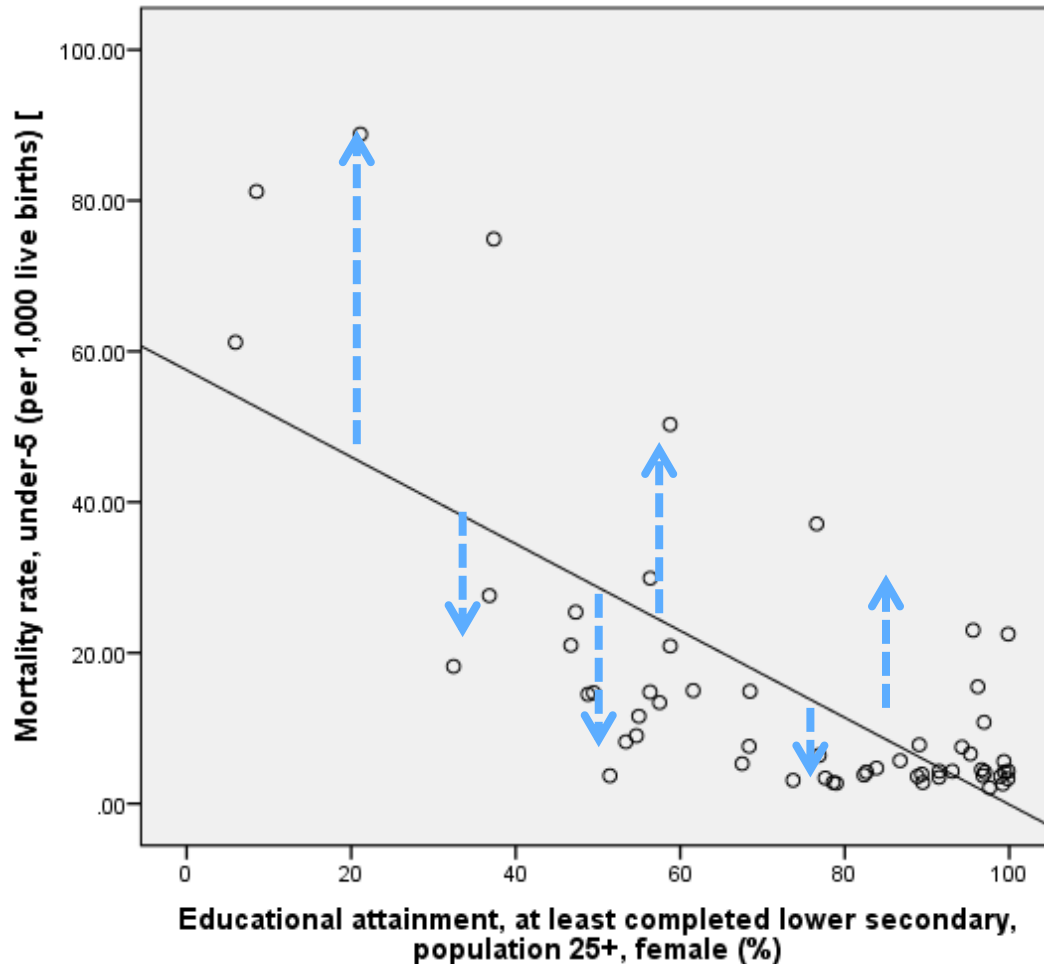
The line represents a simple statistical 'model' of the relationship going on in the real world

We can use that line (model) to 'make predictions' about the value of the dependent variable based on a value of the predictor..

...although in social science the predictions are rarely very good with one predictor – need 'multiple regression' (allows inclusion of more than one predictor variable) ... covered in more advanced courses

The line of best fit

The regression line solution

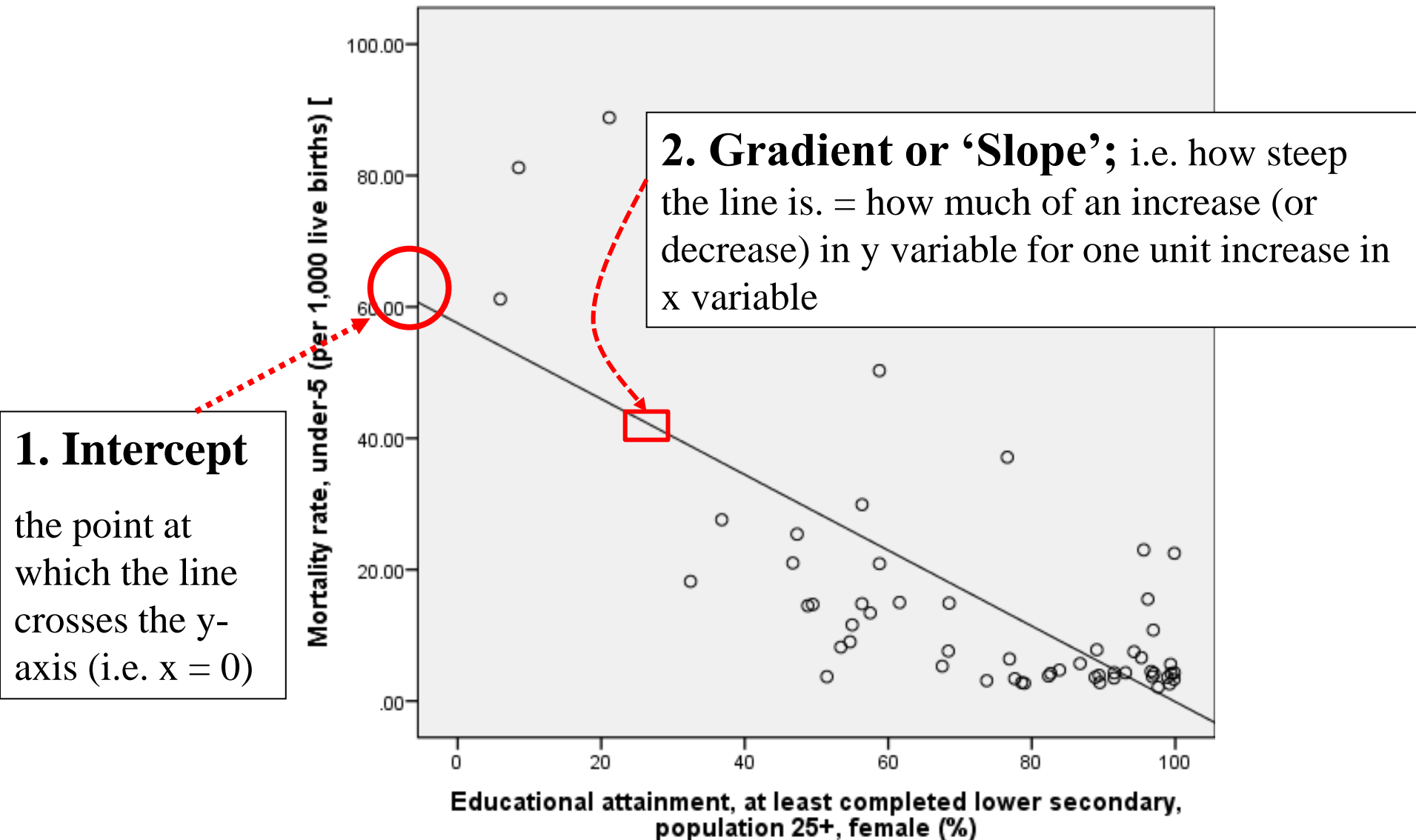


- In regression - the line is placed to minimise the distance between data points and the line (sum of the squared errors).
- This method is called **ordinary least squares**
- The error is the difference between the observed and predicted value (e).

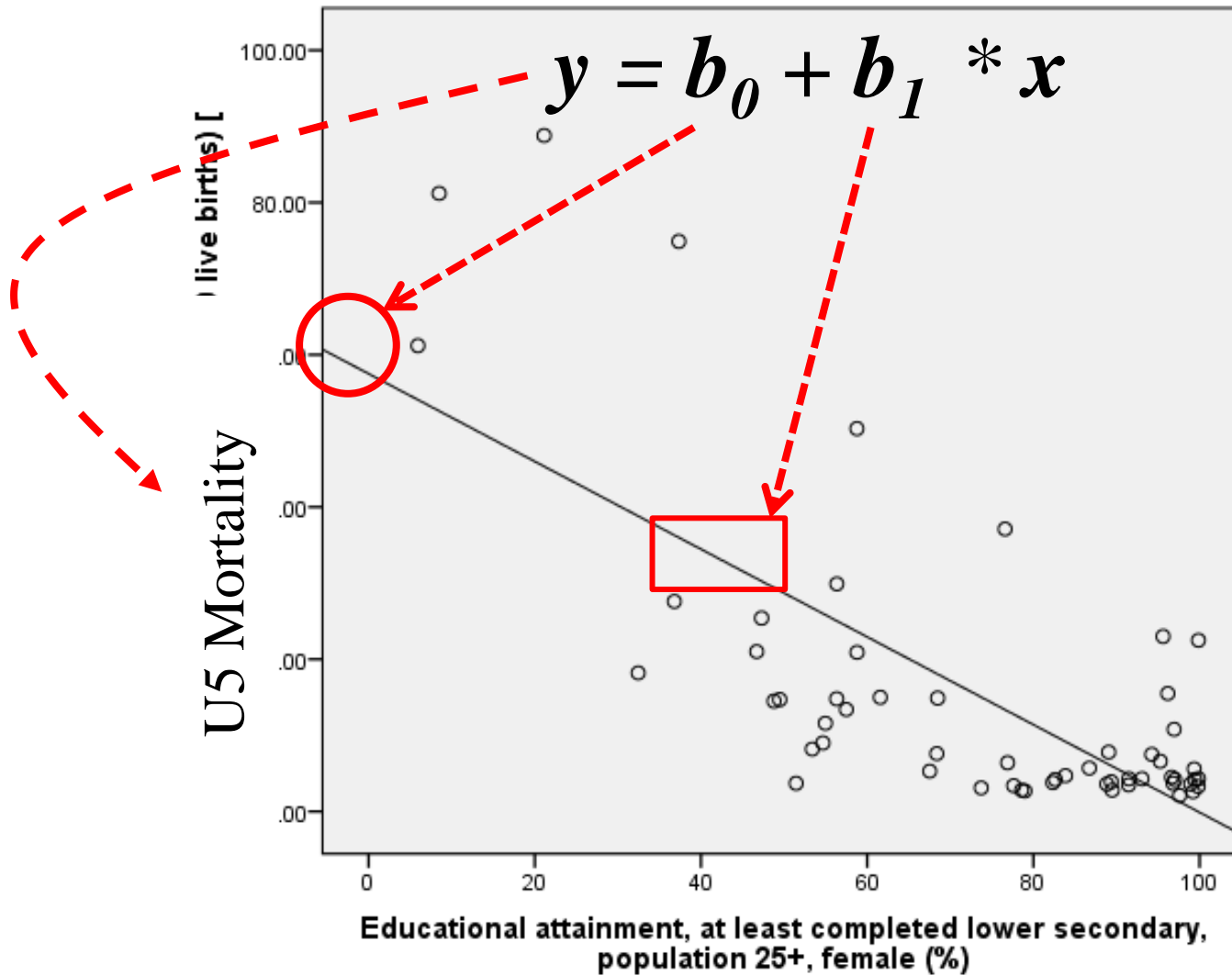
Running a regression in SPSS

- When you run a regression (in SPSS) it calculates the line of best fit..
- ...and produces various tables and outputs that include the information we need to describe the regression line with an equation

We need to know **TWO** things to describe a straight line on a graph:




Expressed as an equation:



The regression equation
enables us to predict y for a given value of x

$$y = b_0 + b_1 x$$

- 
- If we know the value of these two coefficients (b_0 and b_1) we can then estimate values of y for a given value of x
 - Because there are residuals or errors, the estimate will not be perfect

The regression equation

The regression equation could be calculated by hand.
However, we let SPSS do it!

$$y_i = b_0 + b_1 x_i$$



$$\text{U5 Mortality} = 57.509 + -0.576 * x (\% \text{ fem with sec ed})$$

b_0 and b_1 are called regression '**coefficients**'

- b_0 is the intercept (also called the 'constant')
- b_1 is the gradient which is the increase in y (U5 mortality) for a unit increase in x (% females with sec ed) – so mortality rate drops 0.576 deaths per 1000 for every extra one percent of females in sec ed
- The equation is often called a '**model**'

Expressed as an equation:

$$y = 57.509 + (-0.576 * x)$$

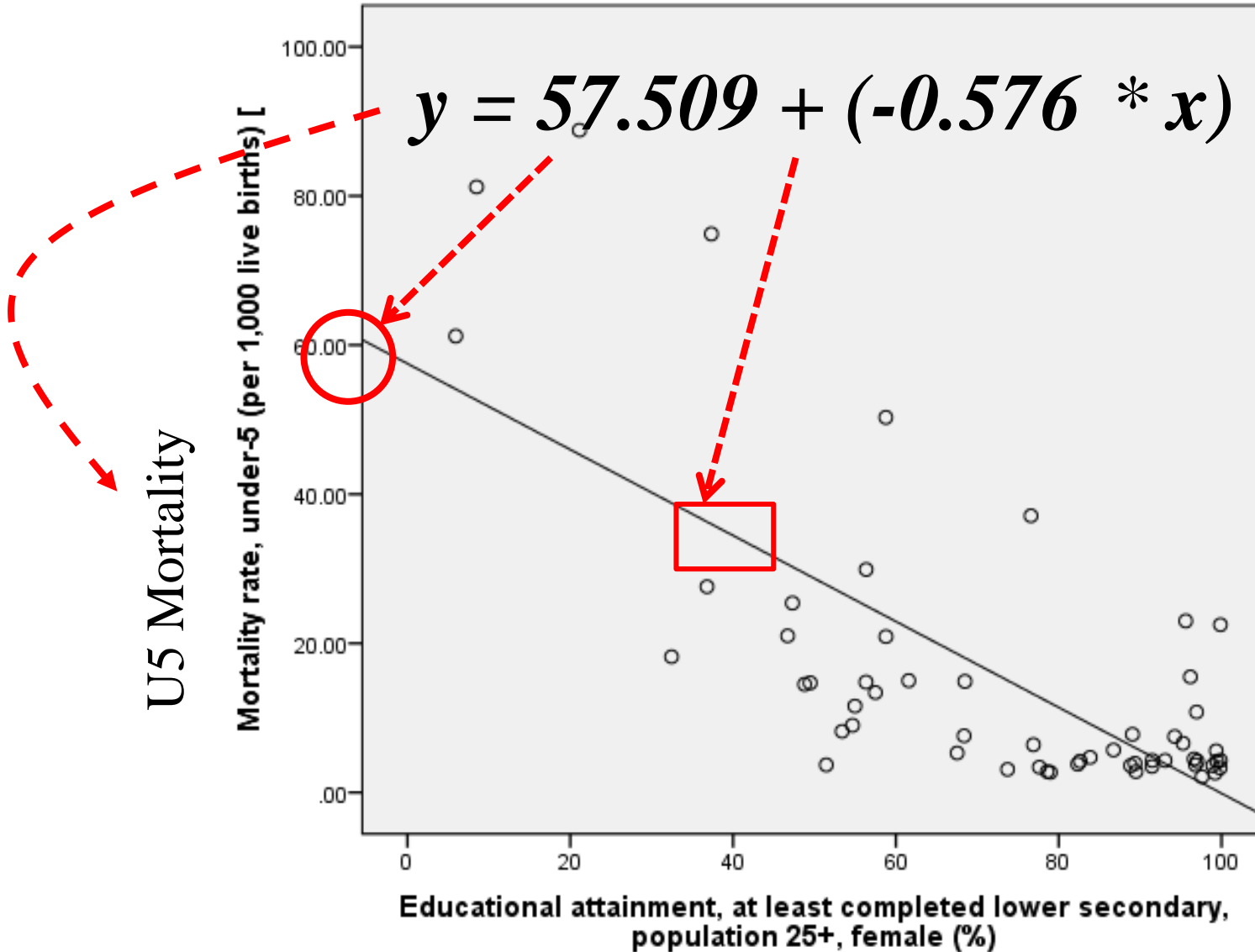
U5 Mortality

Mortality rate, under-5 (per 1,000 live births) [

100.00
80.00
60.00
40.00
20.00
.00

Educational attainment, at least completed lower secondary,
population 25+, female (%)

0 20 40 60 80 100



Using the model to predict y

$$y_i = b_0 + b_1 * x_i$$

$$\text{U5 mortality} = 57.509 + - 0.576 * \% \text{ fem with sec ed}$$

Using the values for b_0 and b_1 we can estimate predicted values for y for any given value of x .

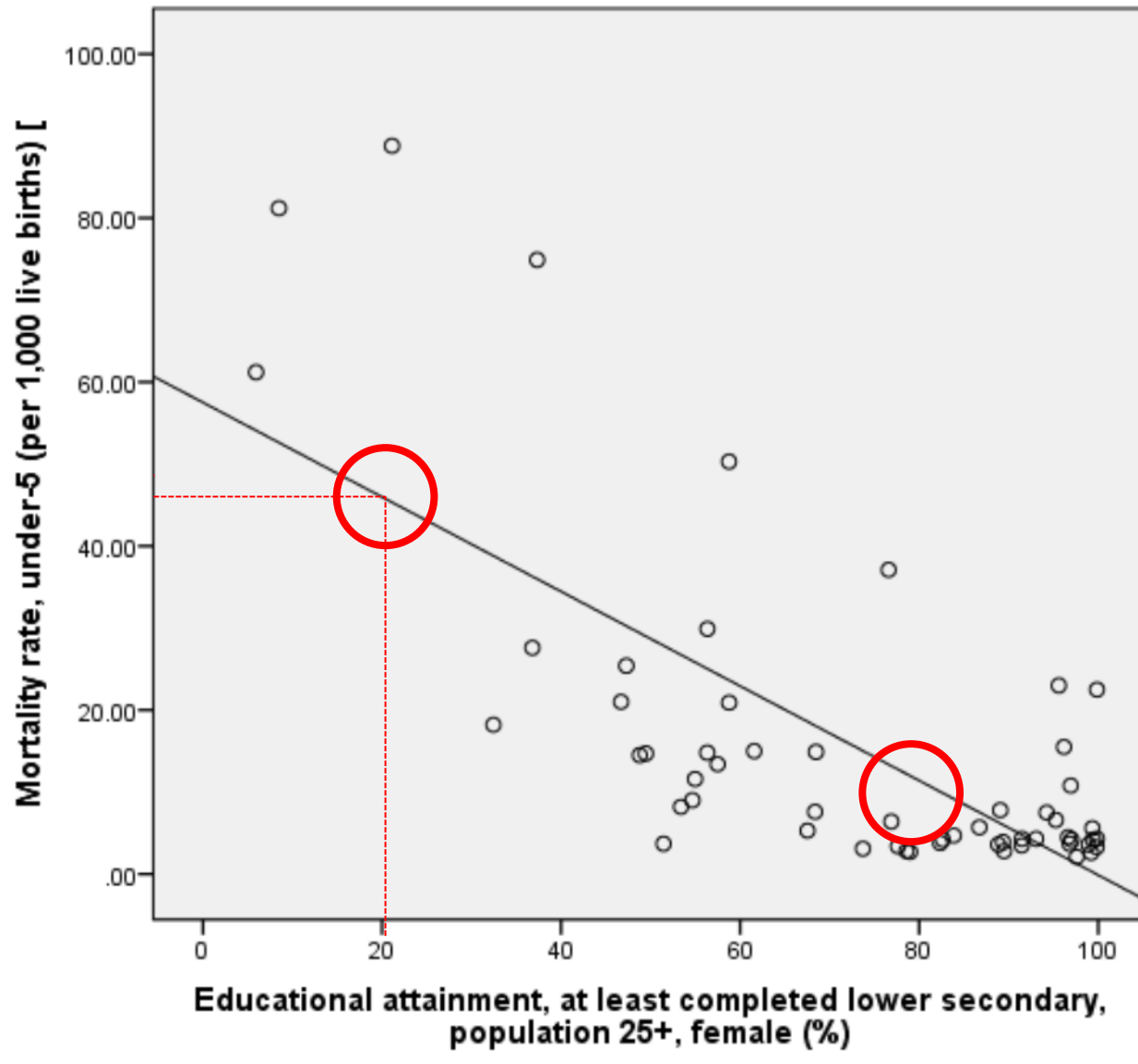
E.g.1: estimated U5 mortality rate for a country with **80%** of females in secondary education :

$$= 57.509 + (-0.576 \times 80) = 57.509 - 46.08 = \underline{11.429} \text{ deaths per 1000}$$

E.g.2: estimated U5 mortality rate for a country with **20%** of females in secondary education :

$$= 57.509 + (-0.576 \times 20) = 57.509 - 11.52 = \underline{45.989} \text{ deaths per 1000}$$

Using regression model to predict U5 for a given level of % female with sec education



Doing it in SPSS

- Getting SPSS to run a simple regression is fairly straightforward
- The challenge is making sense of the results (it gives you a few tables!)
- The output tables crucially include
 - **the values for b_0 and b_1 that we need for our equation**
 - **The value for R and R^2 (R^2 measures the proportion of the variation in the y variable explained by your model i.e. in our example how much is variation in course mark explained by attendance)**
 - And some other stuff

In SPSS

- Exercise 4

Testing the regression output

Statistical significance

- Remember where the data on which the model is based is **sample data** we need to consider sampling error (the reliability of the model will depend on size of the sample).
- Tests can be carried out for the different parts of the model (including b_0 , and b_1) and the overall fit of the model
- SPSS will do this automatically
- We need to look at the p-values for each test.

The results in SPSS

The coefficients (b_0 and b_1)

| Coefficients ^a | | | | | |
|---------------------------|--|-----------------------------|------------|---------------------------|------|
| Model | | Unstandardized Coefficients | | Standardized Coefficients | Sig. |
| | | B | Std. Error | Beta | |
| 1 | (Constant) | 57.509 | 5.880 | | .000 |
| | Educational attainment, at least completed lower secondary, population 25+, female (%) | -.576 | .076 | -.717 | .000 |

a. Dependent Variable: Mortality rate, under-5 (per 1,000 live births) [

- Top row of results gives b_0 (constant) = 57.509
- Second row gives b_1 (slope) = -0.576
- The regression model is:
Blood pressure = 57.509 + (0.576 x % of fem with sec ed)
- Last column gives the p values – note b_1 significantly different from 0 ($p < 0.05$) so education variable has a statistically significant relationship with U5 mortality

Model Fit (R and R²)

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .717 ^a | .514 | .505 | 13.95755 |

a. Predictors: (Constant), Educational attainment, at least completed lower secondary, population 25+, female (%)

- R (0.717) is the Pearson correlation coefficient
- **R²** measures the proportion of the variation in the y variable explained by the model (we actually use the adjusted **R²** as a model based on sample data tends to overestimate goodness of fit in population)
- So the adjusted R² for our example is **0.505**
- Interpret this as **50.5%** of variation in U5 mortality is explained by % fem with sec ed.. (though this is not necessarily a causal relationship)

Model Fit: Overall significance of the model

ANOVA^a

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 11127.953 | 1 | 11127.953 | 57.121 | .000 ^b |
| | Residual | 10519.915 | 54 | 194.813 | | |
| | Total | 21647.868 | 55 | | | |

a. Dependent Variable: Mortality rate, under-5 (per 1,000 live births) [

b. Predictors: (Constant), Educational attainment, at least completed lower secondary, population 25+, female (%)

- The right-hand column here gives p value for the overall significance of the model (measured by F statistic)
- If the significance value of the F statistic is smaller than 0.05 then can conclude the explanatory variable does help explain variation in the dependent variable

Taking it further

- In this session we go no further than **simple regression** (one explanatory variable and one dependent variable)
- This limit of one explanatory variable makes simple regression of limited use in most data analysis in the social sciences (predictions of an outcome are rarely very good based on one predictor – we know most social outcomes are influenced by many factors)
- However, we can turn to ‘**multiple regression**’, an extension of simple regression that allows inclusion of more than one predictor variable at the same time) ... a bit more complex but a very powerful tool when used on the right questions with the right data
- Lots of versions of regression (including **logistic regression** used where the outcome variable is categorical rather than interval)

A great website on regression

<http://www.restore.ac.uk/srme>

Using Statistical Regression Methods in Education Research

Home

Using the site

Modules

Resources

Glossary

About the Authors

Welcome!

This website aims to teach statistical regression methods for use in educational and social research. We try to avoid going into the technical details (we're not statisticians!) to help you to learn these methods and to interpret your research findings using [SPSS](#). If you're new to this website then why not view the [introduction video](#) or have a look at the information about [using the site](#)? Alternatively you could plunge straight in to our [Foundation Module](#).

List of Modules:

▶ [Module 1 - Foundation](#)

▶ [Module 2 - Simple Regression](#)

▶ [Module 3 - Multiple Regression](#)

▶ [Module 4 - Logistic Regression](#)

▶ [Module 5 - Ordinal Regression](#)

