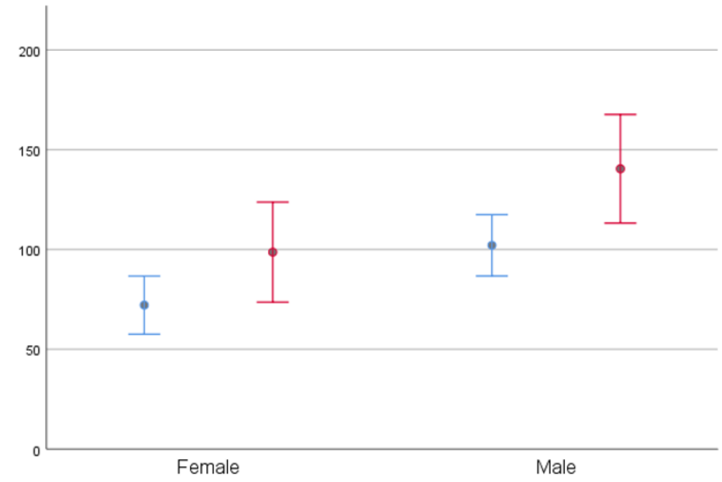
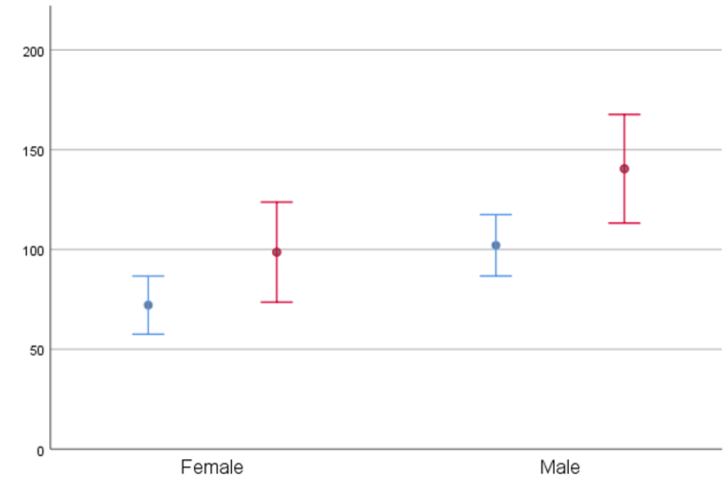


How to interpret ‘significance tests’

Jarlath Quinn – Analytics Consultant

Just waiting for all attendees to join…

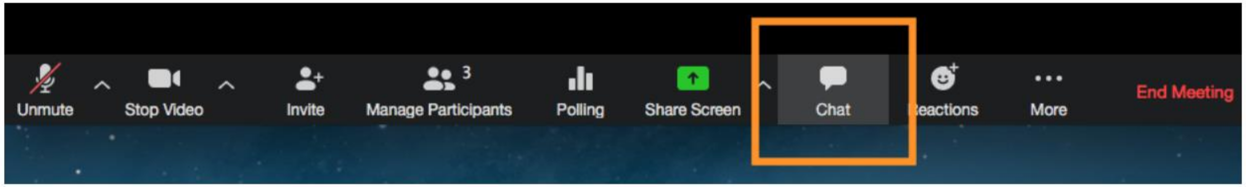


How to interpret 'significance tests'

Jarlath Quinn – Analytics Consultant

FAQ's

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we'll email links to download materials after the session has ended.
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the chat panel – if we run out of time we will follow up with you.



- Gold accredited partner to IBM, Predictive Solutions and DataRobot specialising in advanced analytics & big data technologies
- Work with open source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced and predictive analytics industry

- Deep experience of applied advanced analytics applications across sectors
  - Retail
  - Gaming
  - Utilities
  - Insurance
  - Telecommunications
  - Media
  - FMCG

Agenda

- What do we mean by 'Significance Testing'?
- Inferential Statistics and Significance Testing
- How to interpret P values
- How does a Chi Square test actually work?
- Correlations - When is 'significant' not that *significant*?
- Interpreting confidence intervals correctly

‘Significance’ is a troublesome term

Sir Ronald Fisher

- Introduced the phrase "…tests of significance"*
- Unfortunately, saying that something is 'statistically significant' is likely to be misunderstood
- In this context, 'significance' does not mean **notable**
- Sometimes it doesn't even mean it's **reliable**
- Even the term '**test**' is slightly problematic
- Significance Tests are more properly described as <u>probability of evidence estimates</u>

*From the book 'Statistical Methods for Research Workers" (1925)

The Usual Suspects

R.A. Fisher

Karl Pearson

Egon Pearson

Jerzy Neyman

Alpha level

Principal component analysis

Confidence Intervals

Reject / Accept

0.05

Probability distribution

Significance

ANOVA

Maximum Likelihood

Exact Tests

F Test

Chi Square

Null Hypothesis

95%

Design of experiments

Z tests

Stratified sampling

The Usual Suspects

R.A. Fisher

Karl Pearson

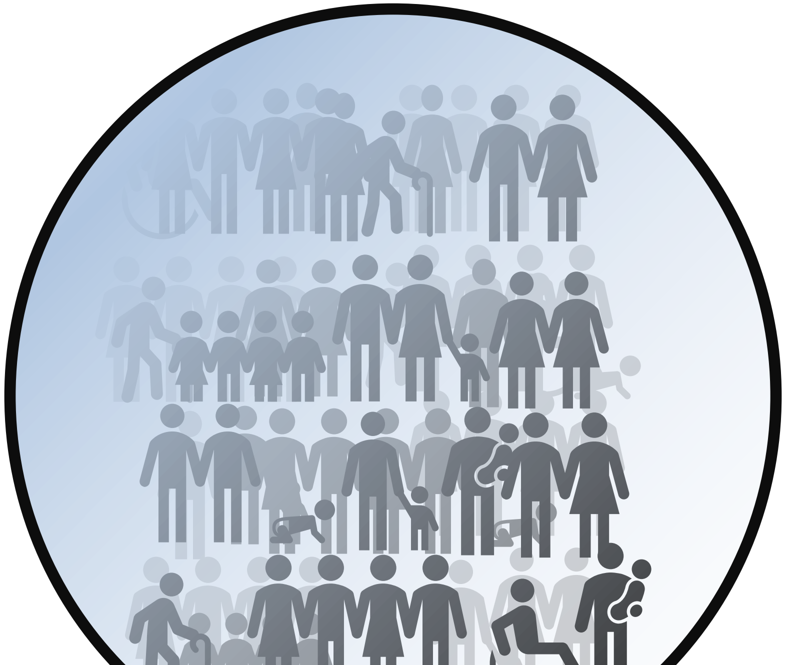
Egon Pearson

Jerzy Neyman

This approach to statistical analysis is referred to as the ‘Frequentist’ tradition

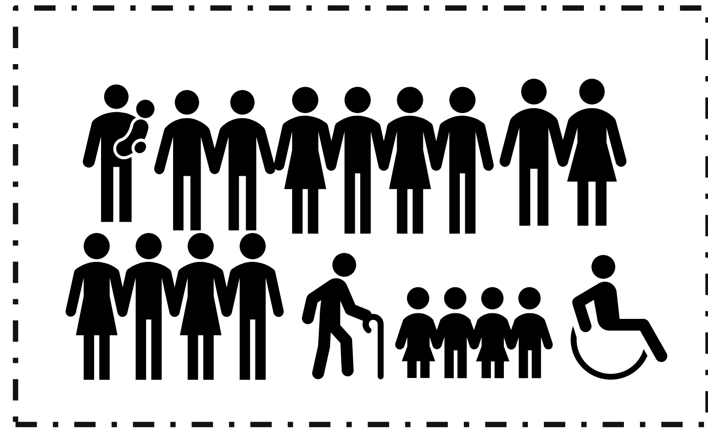
Inferential Statistics and Significance Tests

Populations vs Samples

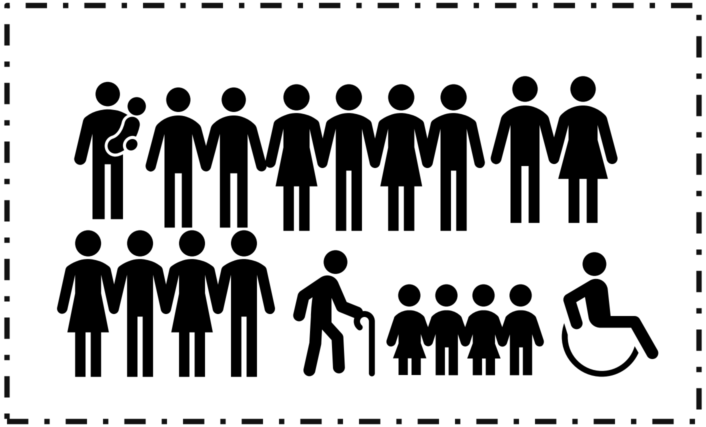


A **parameter** is a characteristic of a **population**.

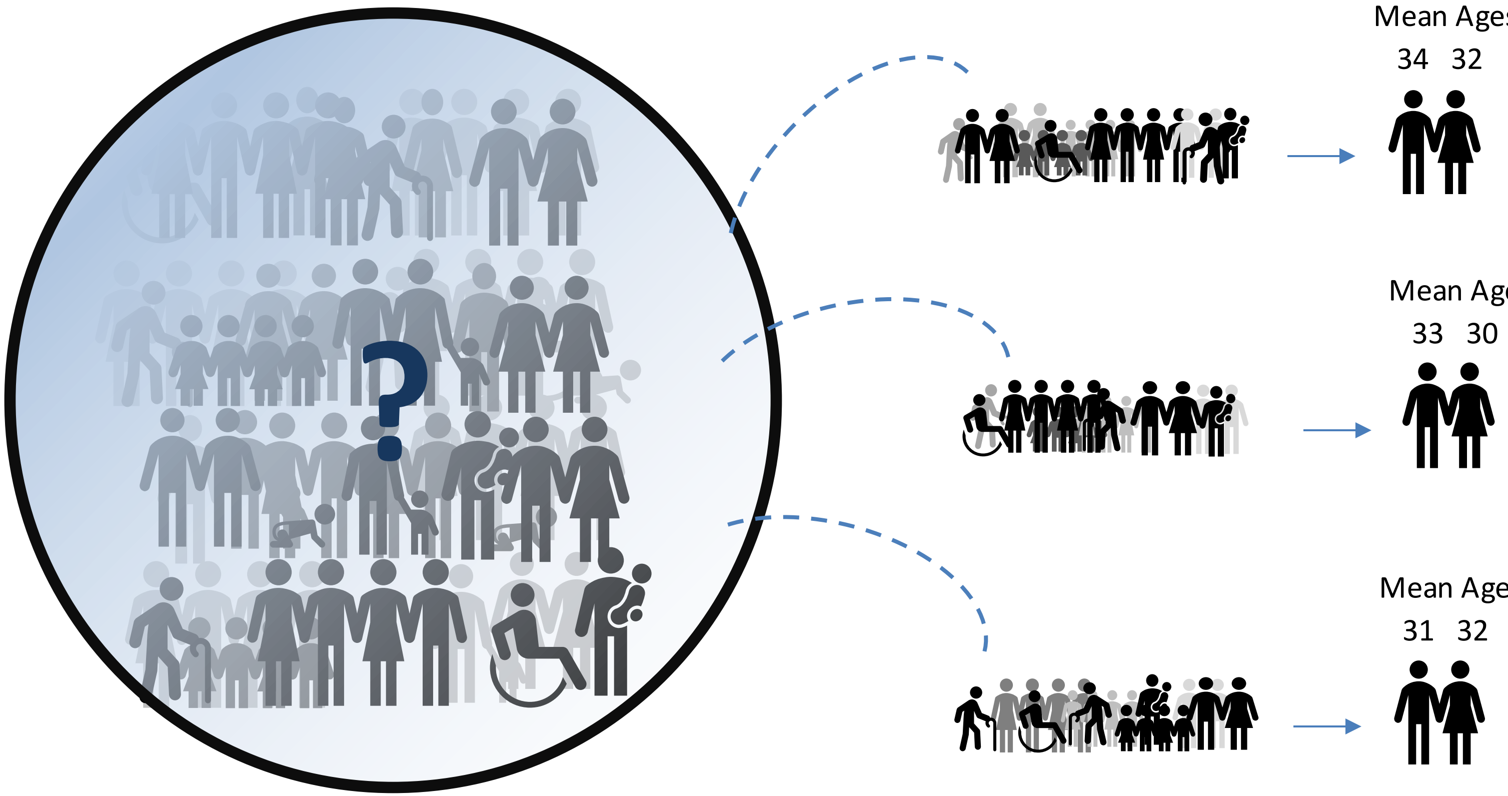
A **statistic** is a characteristic of a **sample**.



But how can we *infer* the parameter value of a population from a sample statistic?



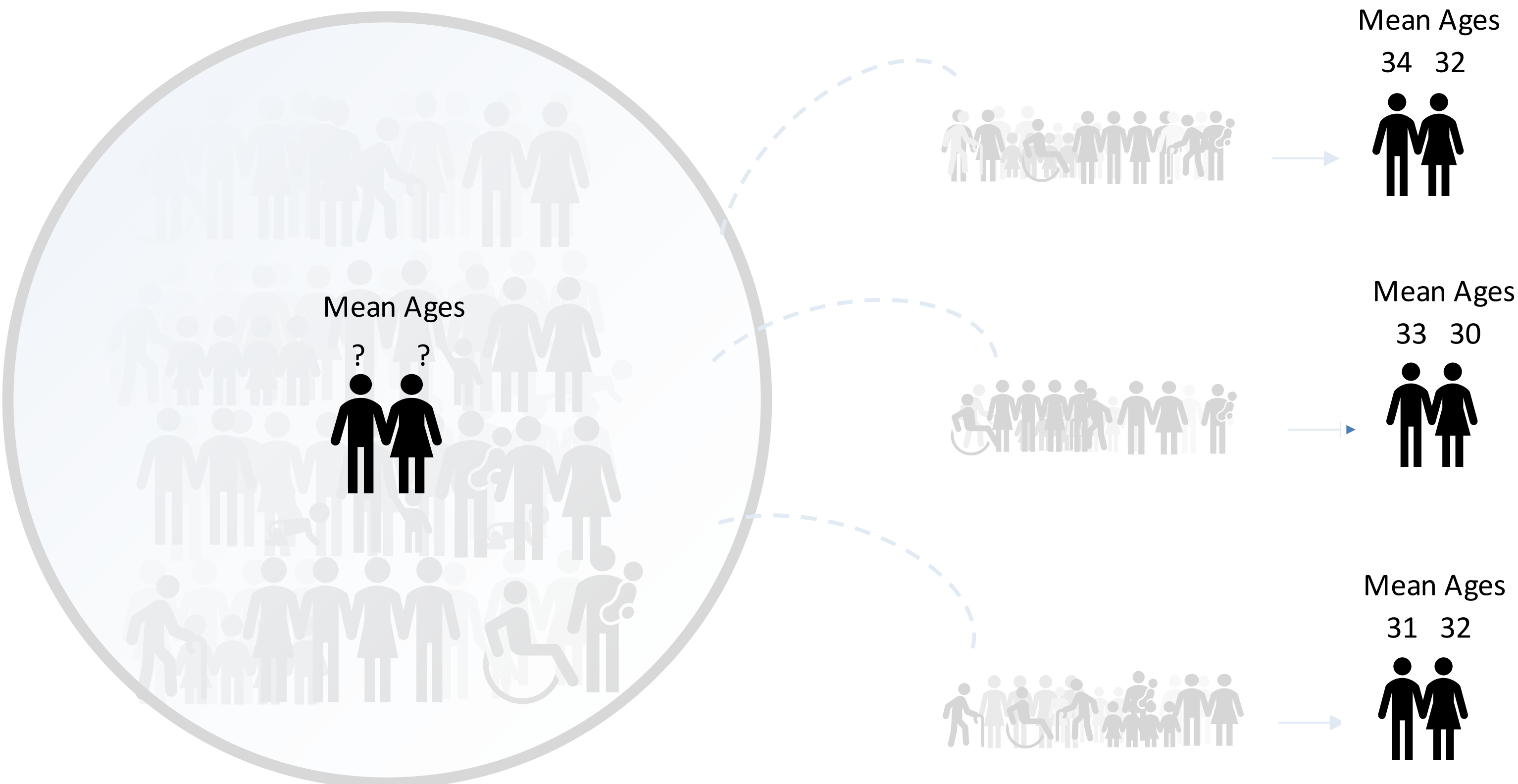
Especially when different samples from <u>the same population</u> *produce different statistics?*



They vary from one another

Statistics is *the science of variation*

Is it possible that the mean ages *are actually the same* in the population?



So where does the 'significance' come from?

- 'Significant' does not mean *noteworthy* - instead these 'tests' simply might indicate…
- *That the evidence does not support the assumption there's no difference between certain groups in the population*
- This is usually measured by a probability value:
  - Sig: 0.001 — 0.1%
  - P = 0.172 — 17.2%
  - Prob = 0.000 — Too small to display
  - Asymptotic Significance: 0.034 — 3%

If we assume *there is no difference* (or *no relationship*) between the groups…

NB: 'T tests' are used to compare pairs of mean values

Significance Testing = Hypothesis Testing

- If you are investigating a relationship with statistics, it's usually because you are curious to see if there is a difference or pattern
- In other words….you have a hypothesis….
- E.g. I think that there is a difference between the average age of men and women in my population - this is called the *Alternative Hypothesis (or 'H1' for short*)
- But in reality, you are testing to see if your data supports the idea that there is *no difference* between the average ages of men and women in the population of interest - this is called the *Null Hypothesis (or 'H0' for short*)
- The relevant test then calculates a probability that indicates:
  1. **If the null hypothesis is true…**
  2. **How often would we get a result as extreme as the one we observe?**

Accepting or Rejecting the Null Hypothesis

- How small (or how large) does this probability value need to be before we decide that the data doesn’t provide enough evidence to support the null hypothesis?
- P= 0.32 ? Does 32% seem to be a reasonable threshold?
- What about 0.0001 %?
- Clearly, we have to choose *some* threshold…
- In fact, threshold is known as the Alpha Level and it’s fairly arbitrary
- RA Fisher suggested 5% or P= 0.05….or a 1 in 20 chance
- On this basis, if the probability is less than P=0.05 ….we should therefore *reject* the null hypothesis on the basis that there is insufficient evidence to support it

Accepting or Rejecting the Null Hypothesis

- The Alternative Hypothesis (H1) is there for reference only. We never accept or reject it based on our analysis.
- In fact, we never **accept** the Null Hypothesis. The most we can do is not find enough evidence to support it and therefore we **reject** it.
- Some analysts argue this is a real limitation of classical ‘frequentist’ statistics.
- A completely alternative approach to hypothesis testing can be found in *Bayesian* statistics where we compare how much evidence there is for the Null Hypothesis vs the Alternative Hypothesis.
- For more information check out our recent video series on Bayesian analysis:
- https://www.sv-europe.com/blog/performing-bayesian-analyses-in-spss/

Interpreting 'P' values

- Contrary to a lot of statistical teaching, the 'P' value **does not** indicate:
  - The probability that the null hypothesis is true
  - The probability that the data were produced by random chance
- What P values **can** do, is indicate:
  - How compatible/incompatible the data are with a null hypothesis

This might sound like semantics but it's not…

- For a start, the probability of **x _given_ y** is not the same as the probability of **y _given_ x**
  - E.g. What's the probability of someone having a full driving license given they are aged 17 or over?
- Versus:
  - What's the probability of someone being aged 17 or over given that they have a full driving license?
- Therefore:
  - The probability of null hypothesis being true **given the evidence**
- Is *not* the same as:
  - The probability of getting that evidence **assuming the null hypothesis is true**

The null hypothesis is pretty important in these ‘tests of significance’

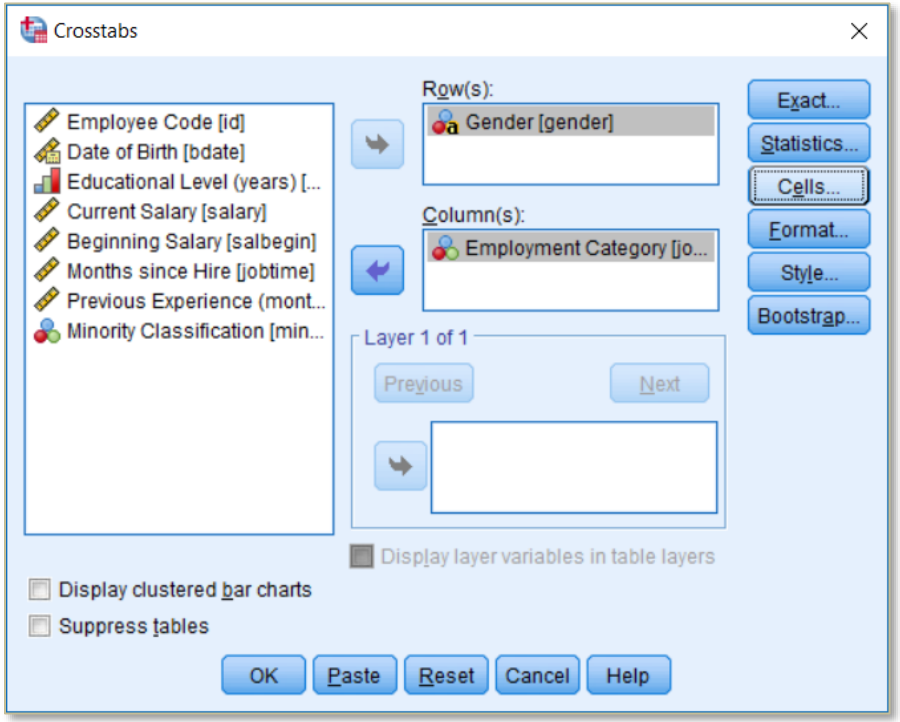
To correctly interpret any 'significance test' you must know the Null Hypothesis associated with it

- The Null Hypothesis for a Chi Square test is that the two variables *are independent of one another* in the population (i.e. they are not related)
- The Null Hypothesis for a T test or an F Test is that the group *means are the same* in the population (i.e. they are the same value)
- The Null Hypothesis for a Pearson's Correlation is that the correlation value is *actually zero* in the population (i.e. no linear relationship)
- The Null Hypothesis for Levene's Test of Equality of Variance is that the groups have the *same spread* (or standard deviation values) in the population
- The Null Hypothesis for KS-Lilliefors test is that the variable *is normally distributed* in the population

Crosstabs and Chi Square Tests

Crosstabs and Chi Square

- Crosstabs are a powerful way to examine relationships between categories
- The Chi-Square test is commonly used with crosstabs as an associated statistical test



Crosstabs and Chi Square

- Crosstabs are a powerful way to examine relationships between categories
- Crosstabs normally display actual frequency counts
- But they are hard to interpret if the group sizes are different

Gender * Employment Category Crosstabulation

Count

		Employment Category			Total
		Clerical	Custodial	Manager	
Gender	Female	206	0	10	216
	Male	157	27	74	258
Total		363	27	84	474

Crosstabs and Chi Square

- So they are often shown with row and/or column percentages
- In this example we have row percentages so we can compare gender in terms of employment category

Gender * Employment Category Crosstabulation

			Employment Category			Total
			Clerical	Custodial	Manager	
Gender	Female	Count	206	0	10	216
		% within Gender	95.4%	0.0%	4.6%	100.0%
	Male	Count	157	27	74	258
		% within Gender	60.9%	10.5%	28.7%	100.0%
Total		Count	363	27	84	474
		% within Gender	76.6%	5.7%	17.7%	100.0%

Crosstabs and Chi Square

- But Crosstabs can also display *expected counts*

$$\frac{\text{Row Total x Column Total}}{\text{Grand Total}} = \textbf{Expected count}$$

$$\frac{216 \times 363}{474} = \textbf{165.4}$$

Gender * Employment Category Crosstabulation

			Employment Category			Total
			Clerical	Custodial	Manager	
Gender	Female	Count	206	0	10	216
		Expected Count	165.4	12.3	38.3	216.0
	Male	Count	157	27	74	258
		Expected Count	197.6	14.7	45.7	258.0
Total		Count	363	27	84	474
		Expected Count	363.0	27.0	84.0	474.0

Crosstabs and Chi Square

- The differences between the **observed** and **expected** counts are called the ***residuals***

Gender * Employment Category Crosstabulation

			Employment Category			Total
			Clerical	Custodial	Manager	
Gender	Female	Count	206	0	10	216
		Expected Count	165.4	12.3	38.3	216.0
		Residual	40.6	-12.3	-28.3	
	Male	Count	157	27	74	258
		Expected Count	197.6	14.7	45.7	258.0
		Residual	-40.6	12.3	28.3	
Total		Count	363	27	84	474
		Expected Count	363.0	27.0	84.0	474.0

Crosstabs and Chi Square

- Performing a calculation based on the sum of the **squared residuals**, allows us to reach a value where the larger the number, the less likely it is that the variables are unrelated to each other in the population.
- This value is the **Pearson Chi-Square** statistic which in turn allows us to calculate a probability value

Gender * Employment Category Crosstabulation

			Employment Category			Total
			Clerical	Custodial	Manager	
Gender	Female	Count	206	0	10	216
		Expected Count	165.4	12.3	38.3	216.0
		% within Employment Category	56.7%	0.0%	11.9%	45.6%
	Male	Count	157	27	74	258
		Expected Count	197.6	14.7	45.7	258.0
		% within Employment Category	43.3%	100.0%	88.1%	54.4%
Total		Count	363	27	84	474
		Expected Count	363.0	27.0	84.0	474.0
		% within Employment Category	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	79.277[a]	2	.000
Likelihood Ratio	95.463	2	.000
N of Valid Cases	474		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.30.

Crosstabs and Chi Square

- The *null hypothesis* associated with Pearson Chi-square *test is that the variables are unrelated*
- A small probability value (less than 0.05) indicates that differences as large as the ones observed, will only occur quite rarely if we assume this null hypothesis is true
- We therefore *reject* the null hypothesis

Gender * Employment Category Crosstabulation

			Employment Category			Total
			Clerical	Custodial	Manager	
Gender	Female	Count	206	0	10	216
		Expected Count	165.4	12.3	38.3	216.0
		% within Gender	95.4%	0.0%	4.6%	100.0%
	Male	Count	157	27	74	258
		Expected Count	197.6	14.7	45.7	258.0
		% within Gender	60.9%	10.5%	28.7%	100.0%
Total		Count	363	27	84	474
		Expected Count	363.0	27.0	84.0	474.0
		% within Gender	76.6%	5.7%	17.7%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	79.277[a]	2	.000
Likelihood Ratio	95.463	2	.000
N of Valid Cases	474		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.30.

Crosstabs and Chi Square

- Reporting a chi-square result:
- "The Chi Square test indicates that the probability of obtaining a value as extreme as the one observed, is less than 0.01 assuming the variables gender and employment category are unrelated."
- "Given this result, there appears to be insufficient evidence to support the null hypothesis and it is therefore rejected."

Gender * Employment Category Crosstabulation

			Employment Category			Total
			Clerical	Custodial	Manager	
Gender	Female	Count	206	0	10	216
		Expected Count	165.4	12.3	38.3	216.0
		% within Gender	95.4%	0.0%	4.6%	100.0%
	Male	Count	157	27	74	258
		Expected Count	197.6	14.7	45.7	258.0
		% within Gender	60.9%	10.5%	28.7%	100.0%
Total		Count	363	27	84	474
		Expected Count	363.0	27.0	84.0	474.0
		% within Gender	76.6%	5.7%	17.7%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	79.277[a]	2	.000
Likelihood Ratio	95.463	2	.000
N of Valid Cases	474		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.30.

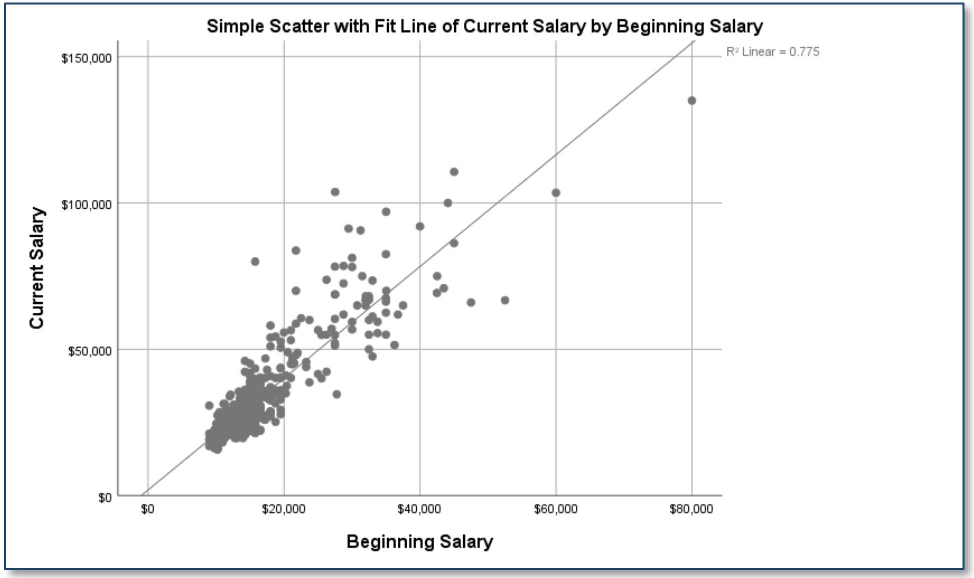
Correlation Coefficients

(and when is significant not that *significant*)

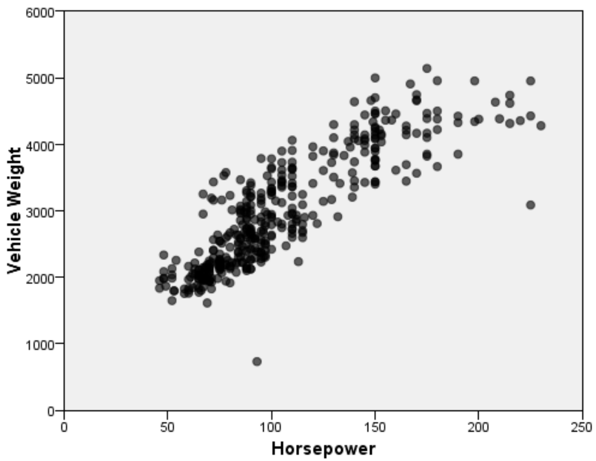
Scatterplots to are used to illustrate relationships between continuous variables

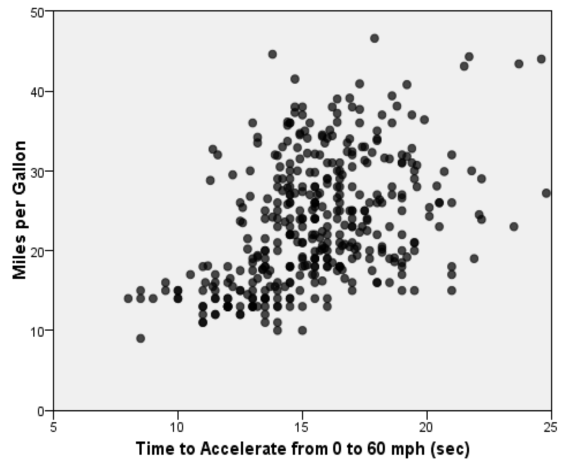
- Here, the scatterplot shows a strong linear relationship between salary and beginning salary

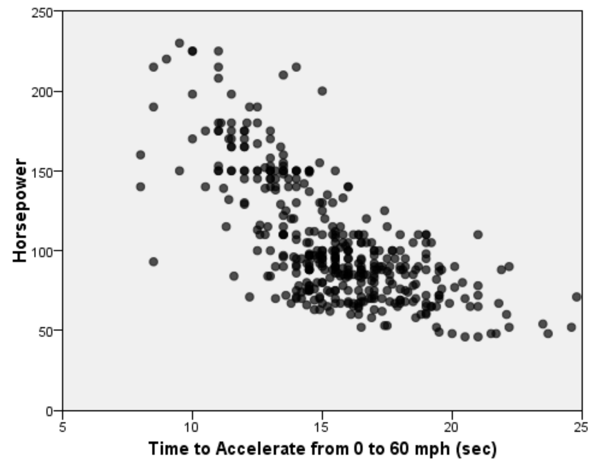
- Correlation values allow us to summarise these relationships



Correlations measure the strength of *linear* relationships





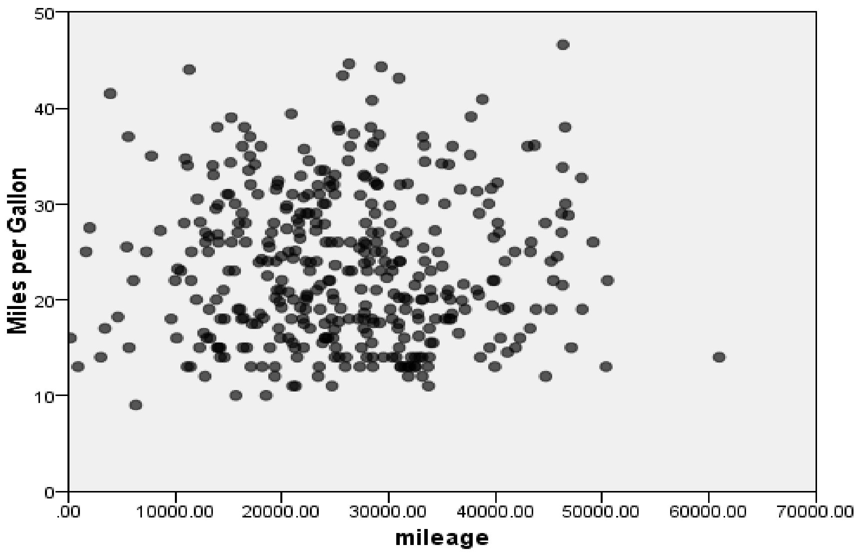


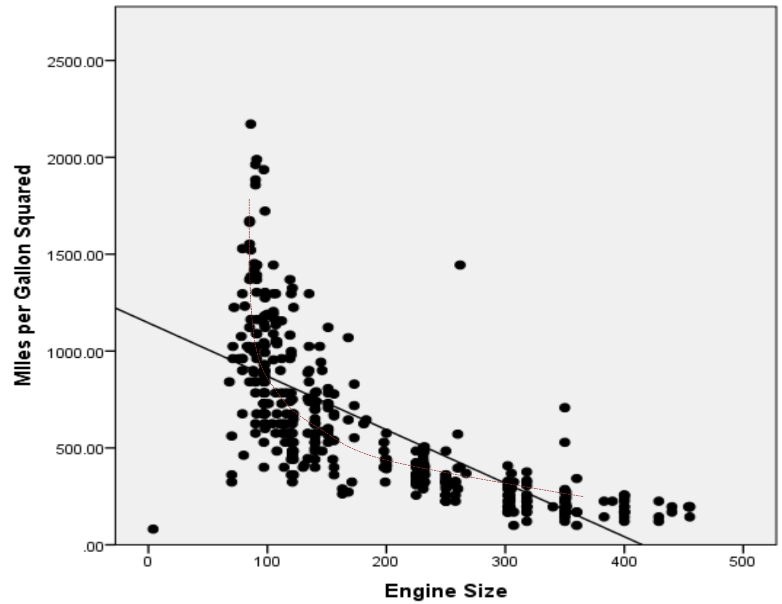
0.859 **0.434** **-.701**

Pearson Correlation Values*

* AKA Pearson's r

Non-linear relationships are not accounted for





-0.005

-.671

Pearson Correlation Values

Correlation Matrix

- Here are the correlation values for each pair of relationships between four variables…

Correlations

Pearson Correlation

	Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Beginning Salary	**.880**			
Months since Hire	.084	-.022		
Previous Experience (months)	-.097	.046	.002	
Educational Level (years)	**.661**	**.633**	.047	-.252

Correlation Matrix

- But which ones are 'statistically significant' ?

Correlations

Pearson Correlation

	Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Beginning Salary	**.880**			
Months since Hire	.084	-.022		
Previous Experience (months)	-.097	.046	.002	
Educational Level (years)	**.661**	**.633**	.047	-.252

Remember this?

- The Null Hypothesis for a Chi Square test is that the two variables are independent of one another in the population (i.e. they are not related)
- The Null Hypothesis for a T test or an F Test is that the group means are the same in the population (i.e. they are the same value)
- The Null Hypothesis for a Pearson's Correlation is that the correlation value is *actually zero* in the population (i.e. no linear relationship)
- The Null Hypothesis for Levene's Test of Equality of Variance is that the groups have the *same spread* (or standard deviation values) in the population
- The Null Hypothesis for KS-Lilliefors test is that the variable *is normally distributed* in the population

Correlation Matrix

- The null hypothesis is that the actual correlations in the population are zero i.e. there is no relationship between the pairs of variables

Correlations

Pearson Correlation

	Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Beginning Salary	**.880**			
Months since Hire	.084	-.022		
Previous Experience (months)	-.097	.046	.002	
Educational Level (years)	**.661**	**.633**	.047	-.252

- Let's add the 'significance' values

Correlation Matrix

- Look at the highlighted cell. The correlation is -0.097 i.e. extremely weak
- But the significance value is 0.034 (3.4%) i.e. below P = 0.05

Correlations

		Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Beginning Salary	Pearson Correlation	**.880**			
	Sig. (2-tailed)	<.001			
	N	474			
Months since Hire	Pearson Correlation	.084	-.022		
	Sig. (2-tailed)	.067	.626		
	N	474	475		
Previous Experience (months)	Pearson Correlation	-.097	.046	.002	
	Sig. (2-tailed)	.034	.319	.974	
	N	474	475	475	
Educational Level (years)	Pearson Correlation	**.661**	**.633**	.047	-.252
	Sig. (2-tailed)	<.001	<.001	.310	<.001
	N	474	475	475	475

Correlation Matrix

- Is this statistically significant? Yes – because the probability of getting a result as 'extreme' as -0.097, assuming there is no relationship between the two variables in the population is still only 0.034 (3.4%)

Correlations

		Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Beginning Salary	Pearson Correlation	**.880**			
	Sig. (2-tailed)	<.001			
	N	474			
Months since Hire	Pearson Correlation	.084	-.022		
	Sig. (2-tailed)	.067	.626		
	N	474	475		
Previous Experience (months)	Pearson Correlation	-.097	.046	.002	
	Sig. (2-tailed)	.034	.319	.974	
	N	474	475	475	
Educational Level (years)	Pearson Correlation	**.661**	**.633**	.047	-.252
	Sig. (2-tailed)	<.001	<.001	.310	<.001
	N	474	475	475	475

Correlation Matrix

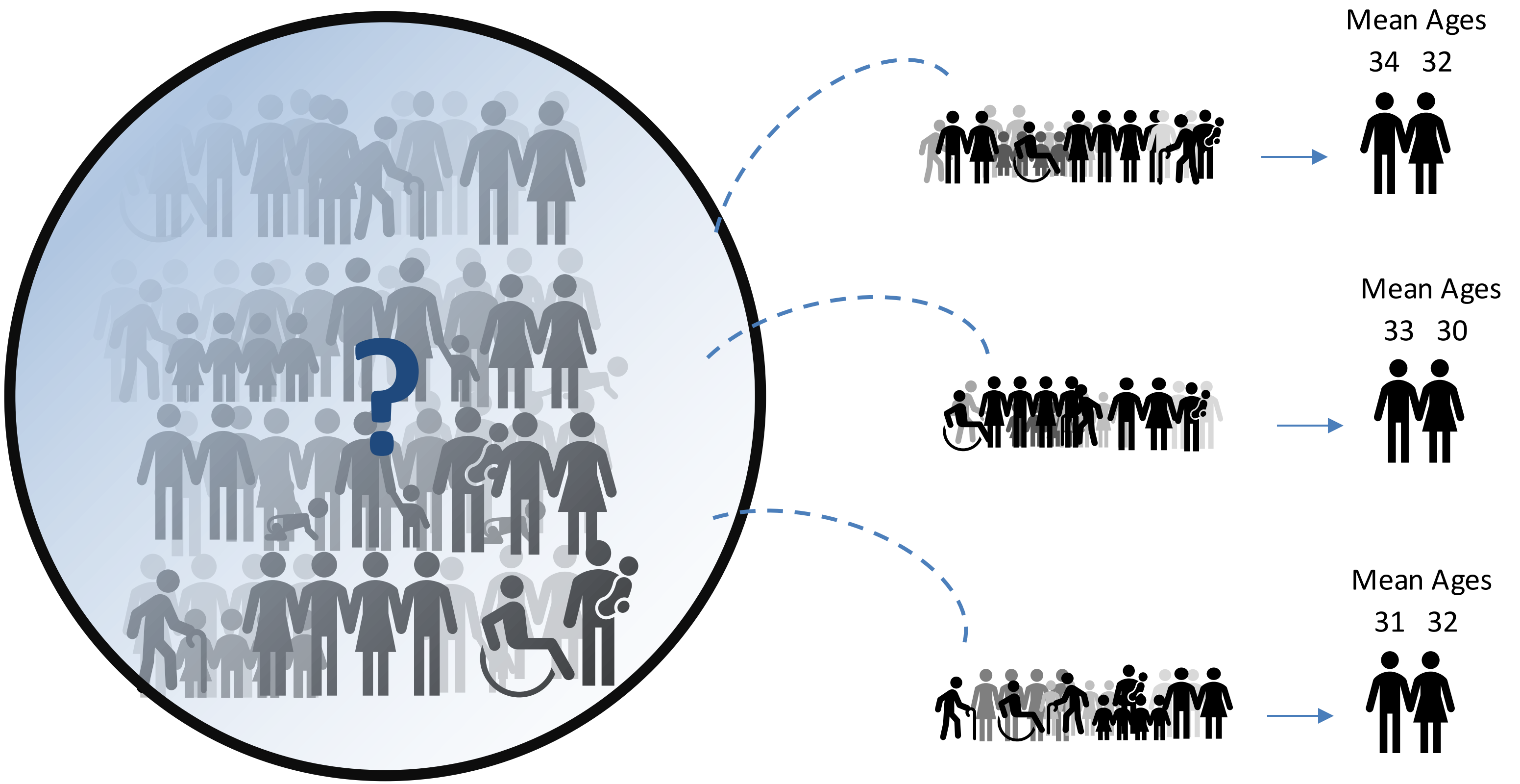
- Does that mean it is notable or worth reporting? No, not particularly. But then that's not what the null hypothesis for a Pearson's correlation value is directed at. It only states that there is no relationship there at all in the population.

Correlations

		Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Beginning Salary	Pearson Correlation	**.880**			
	Sig. (2-tailed)	<.001			
	N	474			
Months since Hire	Pearson Correlation	.084	-.022		
	Sig. (2-tailed)	.067	.626		
	N	474	475		
Previous Experience (months)	Pearson Correlation	-.097	.046	.002	
	Sig. (2-tailed)	.034	.319	.974	
	N	474	475	475	
Educational Level (years)	Pearson Correlation	**.661**	**.633**	.047	-.252
	Sig. (2-tailed)	<.001	<.001	.310	<.001
	N	474	475	475	475

How to interpret confidence intervals correctly

Different samples from the same population *vary*



Different samples give different results

- Repeat a survey or a project and you won't get *exactly* the same results - calculate a statistic and the results vary from one sample to the next.
- No statistical calculation can tell you what the *actual* value of a population parameter is and we don't have the luxury of repeating samples over and over again
- But *we can estimate a range* of values that it is likely to lie within
- We can do this by requesting *Confidence Intervals*
- Confidence Intervals can be shown graphically as *Error Bars*

Confidence Intervals

- The 'Explore' procedure in SPSS produces lots of summary measures…
- Here it says the mean age for employees in this data sample was **34.74**
- But it also provides 95% **Confidence Intervals**

Descriptives

			Statistic	Std. Error
Age of Employee	Mean		34.74	.542
	95% Confidence Interval for Mean	Lower Bound	33.67	
		Upper Bound	35.80	
	5% Trimmed Mean		34.06	
	Median		29.00	
	Variance		139.034	
	Std. Deviation		11.791	
	Minimum		20	
	Maximum		62	
	Range		42	
	Interquartile Range		18	
	Skewness		.859	.112
	Kurtosis		-.573	.224

Confidence Intervals

- So in this example, the mean age *happens* to be **34.74**
- Of course, we don't know what the mean age for employees is in the population that the sample was drawn from, but the confidence intervals indicate that it is likely to be contained in a range like **33.67 to 35.8**
- As we know, drawing another sample would probably result in a slightly different mean value
- But ***<u>all our statistics</u>*** are likely to vary from one sample to the next - not just the mean

Confidence Intervals

- So, if we drew another comparable sample and recalculated the confidence intervals, *they too would probably be different*
- In other words, we can't get too hung-up on the values of the confidence intervals themselves, any more than we can fixate on the precise value of the mean
- What we can say however, is that even if the intervals vary from one sample to the next, they are likely to be broad enough that on 95% of occasions, they will contain the population mean (i.e. the parameter)

Confidence Intervals

- What we **should not** say is:
- “We are 95% confident that the population parameter is between X and Y”
- **Because**:
  - The intervals themselves will vary from sample to sample
  - There’s no scientific basis to the phrase ‘we are 95% confident’
  - It is the **procedure** that on 95% of occasions will capture the population mean

How are confidence intervals calculated?

- The *special statistic* that drives confidence intervals is a called **a standard error**
- When calculating confidence intervals for a *mean* value, we use the **standard error of the mean**
- But there are other standard error statistics such as the standard error of the median, the standard error of the difference, the standard error of the correlation etc.
- These are useful for when we need to calculate confidence intervals for other statistics

How are confidence intervals calculated?

- A standard error is based on a *standard deviation*
- Indeed, just as a standard deviation measures variation *within* a sample, the standard error measures variation *between* samples
- So the standard error of a mean tells us on average how much we would expect a sample mean to vary from one sample to another

Descriptives

			Statistic	Std. Error
Age of Employee	Mean		34.74	.542
	95% Confidence Interval for Mean	Lower Bound	33.67	
		Upper Bound	35.80	
	5% Trimmed Mean		34.06	
	Median		29.00	
	Variance		139.034	
	Std. Deviation		11.791	
	Minimum		20	
	Maximum		62	
	Range		42	
	Interquartile Range		18	
	Skewness		.859	.112
	Kurtosis		-.573	.224

How are confidence intervals calculated?

- To calculate 95% confidence intervals…
- We multiply a standard error by roughly two (or 1.96 to be exact)
  - 0.542 x 1.96 = 1.062
- We then add and subtract this value from the mean to get our Confidence Intervals
- 34.74 – 1.062 = **33.678**
- 34.74 + 1.062 = **35.802**

Descriptives

			Statistic	Std. Error
Age of Employee	Mean		34.74	.542
	95% Confidence Interval for Mean	Lower Bound	33.67	
		Upper Bound	35.80	
	5% Trimmed Mean		34.06	
	Median		29.00	
	Variance		139.034	
	Std. Deviation		11.791	
	Minimum		20	
	Maximum		62	
	Range		42	
	Interquartile Range		18	
	Skewness		.859	.112
	Kurtosis		-.573	.224

Wide intervals vs narrow intervals

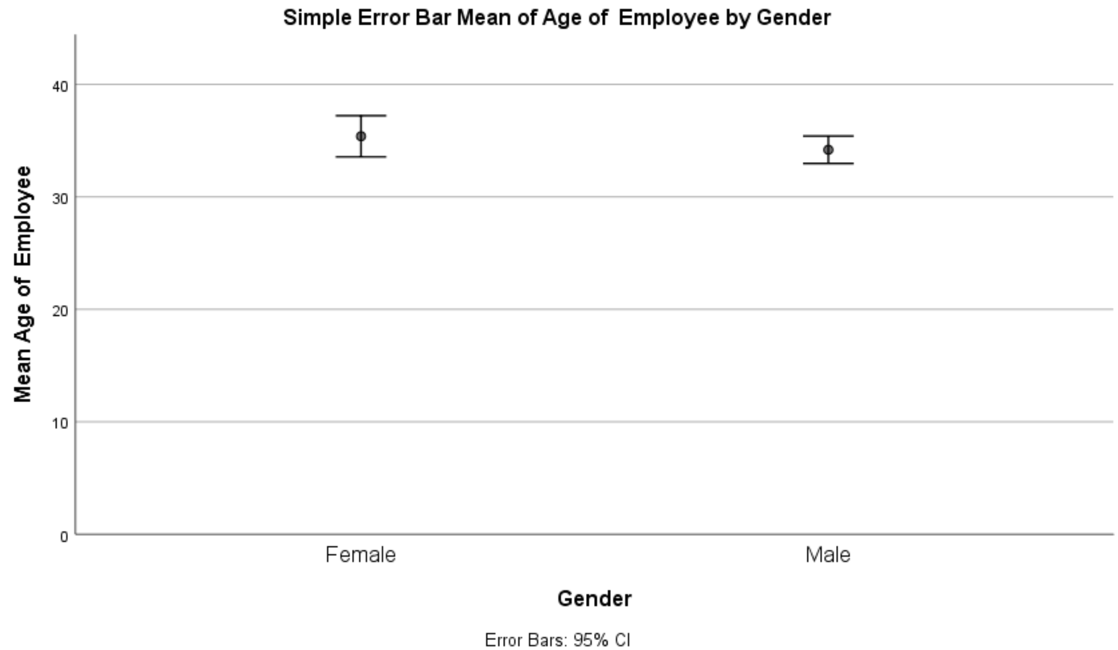
- **Because the standard error is based on the standard deviation and the sample size…**
- The width of the confidence intervals is affected by two values:
    1. The size of the sample
    2. The spread in data (the standard deviation of the variable)

- So confidence intervals move further apart:
    1. The smaller the sample
    2. The greater the spread (i.e. the larger the standard deviation)

- And confidence intervals move closer together:
    1. The larger the sample
    2. The lesser the spread (i.e. the smaller the standard deviation)

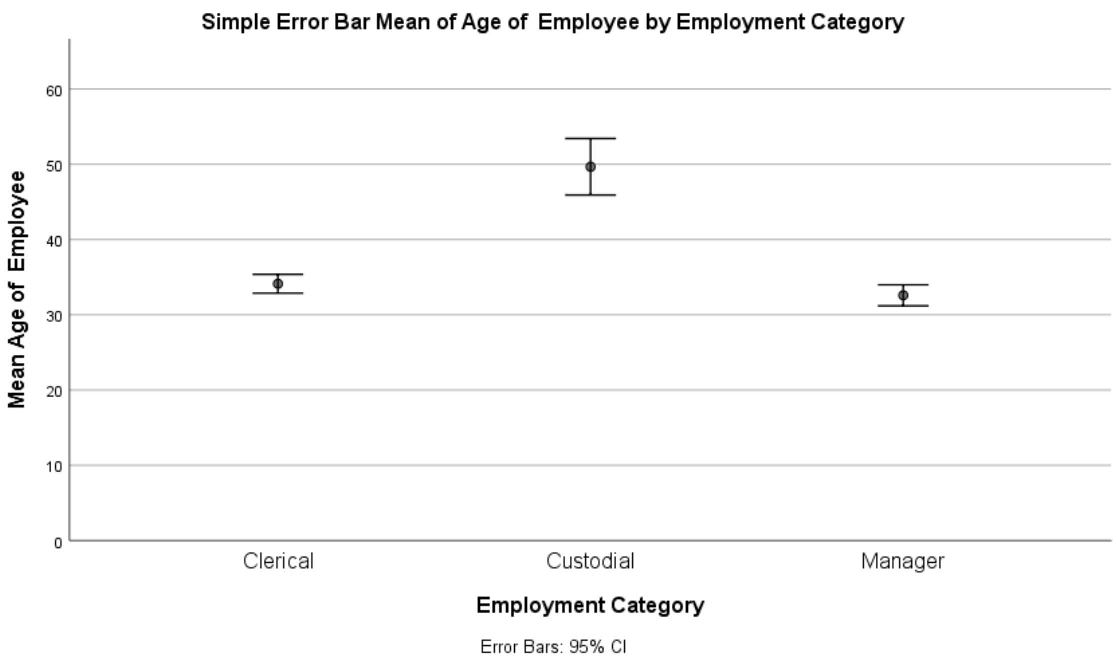
Graphing Confidence Intervals with Error Bars

- The dot in the middle represents the mean while the upper and lower bars represent the upper and lower confidence intervals
- Note that the bars don't quite cover the same range of values
- Note that the error bars overlap



Graphing Confidence Intervals with Error Bars

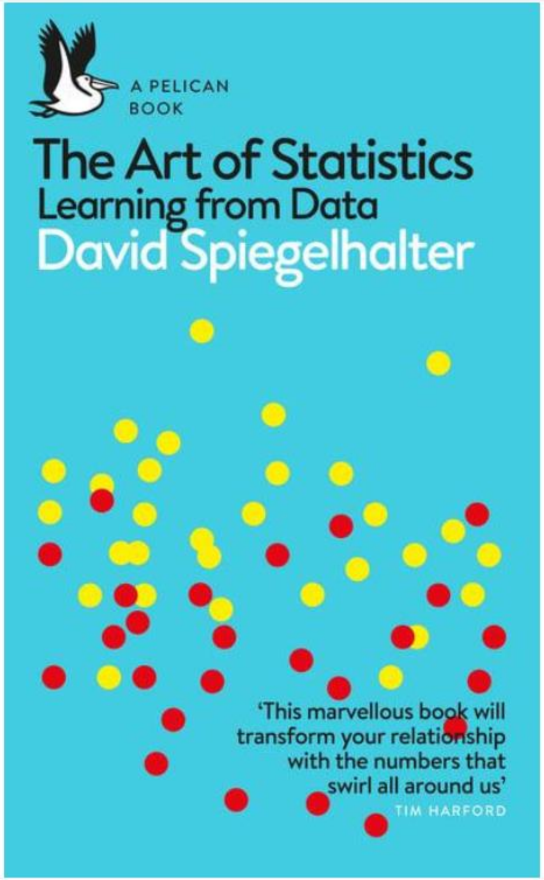
- Contrast the previous example with the variable Employment Category
- Note that in the Custodial categories the CI intervals **do not** overlap with the other two groups



Further Reading: the 'Eat Your Greens' blog series

1. Just because something is statistically significant doesn't mean it's practically significant
2. Testing versus inferring
3. What's 'standard' about a standard deviation?
4. Finding normality – why is the normal distribution so important when we so rarely encounter it in real life?
5. The gateway to inference – the standard error and confidence intervals
6. Understanding correlation
7. Making sense of significance tests
8. Introduction to Power Analysis

Further Reading: Recommended



The Art of Statistics Learning from Data - Pelican Books

D. J. Spiegelhalter (author)

Working with Smart Vision Europe

- **Consulting Services**

Project Support

Purchase 1-2 days of consultancy time to have an expert work alongside you on your own project

Analytics Advice

Give us 3-5 days to investigate your data & analytical strategy and we'll present our recommendations re: improvements & alternatives

Analytical Deep-Dive

Let us explore your data landscape to test hypotheses, identify problem areas, find key outcome drivers or develop new applications

Working with Smart Vision Europe

- **Sourcing Software**
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - http://www.sv-europe.com/buy-spss-online/
- **Training**
  - Formal classroom/virtual training
  - Custom course development
  - Informal 'bite-size' training split over time
- **Advice and Support**
  - 'No strings attached' technical and business advice relating to analytics
  - Tracked technical support services around the IBM SPSS product line

Contact us:

+44 (0)207 786 3568
info@sv-europe.com
Twitter: @sveurope
Follow us on Linked In
Sign up for our Newsletter

Thank you