statstutor

community project

encouraging academics to share statistics support resources

Statistical Methods
8. Parametric Testing

Based on materials provided by Coventry University and Loughborough University under a National HE STEM Programme Practice Transfer Adopters grant

National HE STEM Programme

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Workshop outline

We will consider:

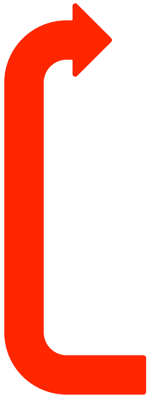
- ☐ Confidence intervals
- ☐ Parametric statistics
- ☐ Normality testing:
  - ➢ Skewness and kurtosis
  - ➢ Shapiro-Wilk test
- ☐ Paired-samples t-test
- ☐ Independent-samples t-test
- ☐ Assumptions
- ☐ Robustness

The statistical analysis process

- ☐ Make sure you have a good data set to start with
- ☐ Generally, use Excel (see Workshop 4) before you use SPSS (see Workshop 6)

1. First describe and present your data, e.g. frequency distributions in tables or charts
2. Calculate basic statistics where possible, e.g. means and standard deviations
3. Start to interpret your data – what might it mean?
4. Select specific items for closer attention (based on your research hypotheses)
5. Select and carry out the right kind of test
6. Look for statistical significance
7. Modify and repeat if necessary

Demonstrate that you are in control of the process!



Pulse data set

- 91 students were asked to take their pulse for a minute
- Each student was then asked to toss a coin:
  - If it came up heads they ran on the spot for a minute
  - If it came up tails they sat for a minute
- At the end of the minute they all took their pulse rates again for a minute
- They also supplied additional personal information:
  - Gender
  - Smoking habits
  - Normal exercise level
  - Height
  - Weight

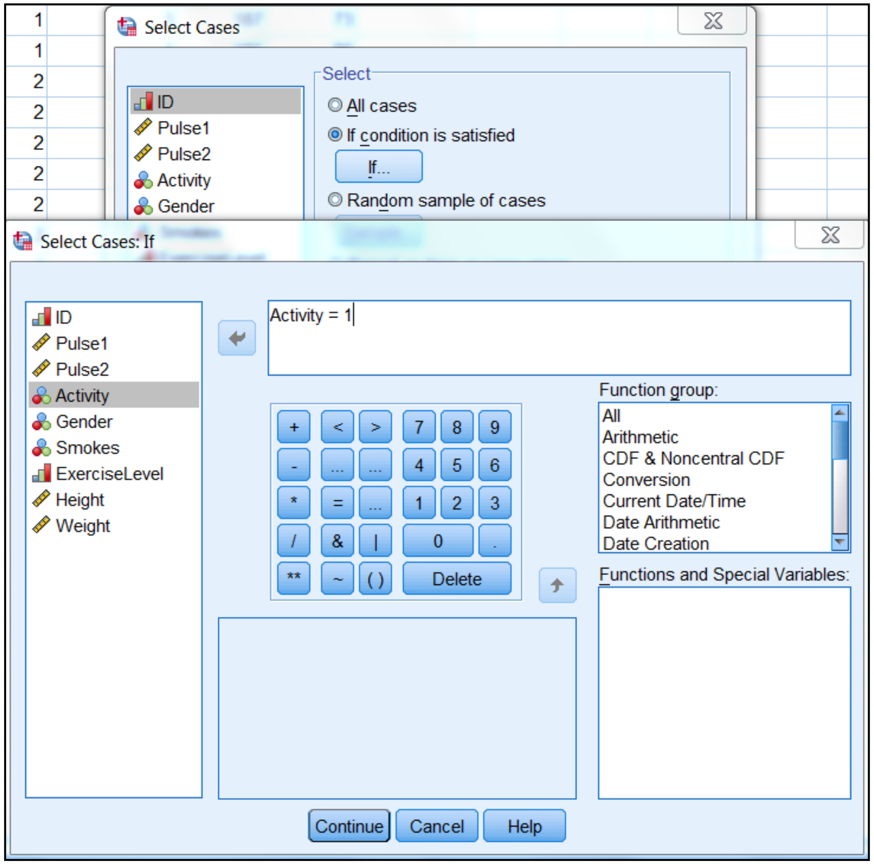
Research questions

1. Does running on the spot change your pulse rate?
2. Do regular smokers have a different pulse rate when sitting than non-/non-regular smokers?

Open the SPSS data file Pulse.sav which you created in Workshop 7.

Research Question 1

- We need to compare the pulse rate of the people who ran on the spot (*Activity* = 1) before they ran (*Pulse1*) and afterwards (*Pulse2*)
- This is known as **paired** or **related data**
- We first need to select only these cases
- Select *Data – Select Cases…*
- Select *If condition is satisfied* and click on *If…*
- Select the variable *Activity* then press = and *1*

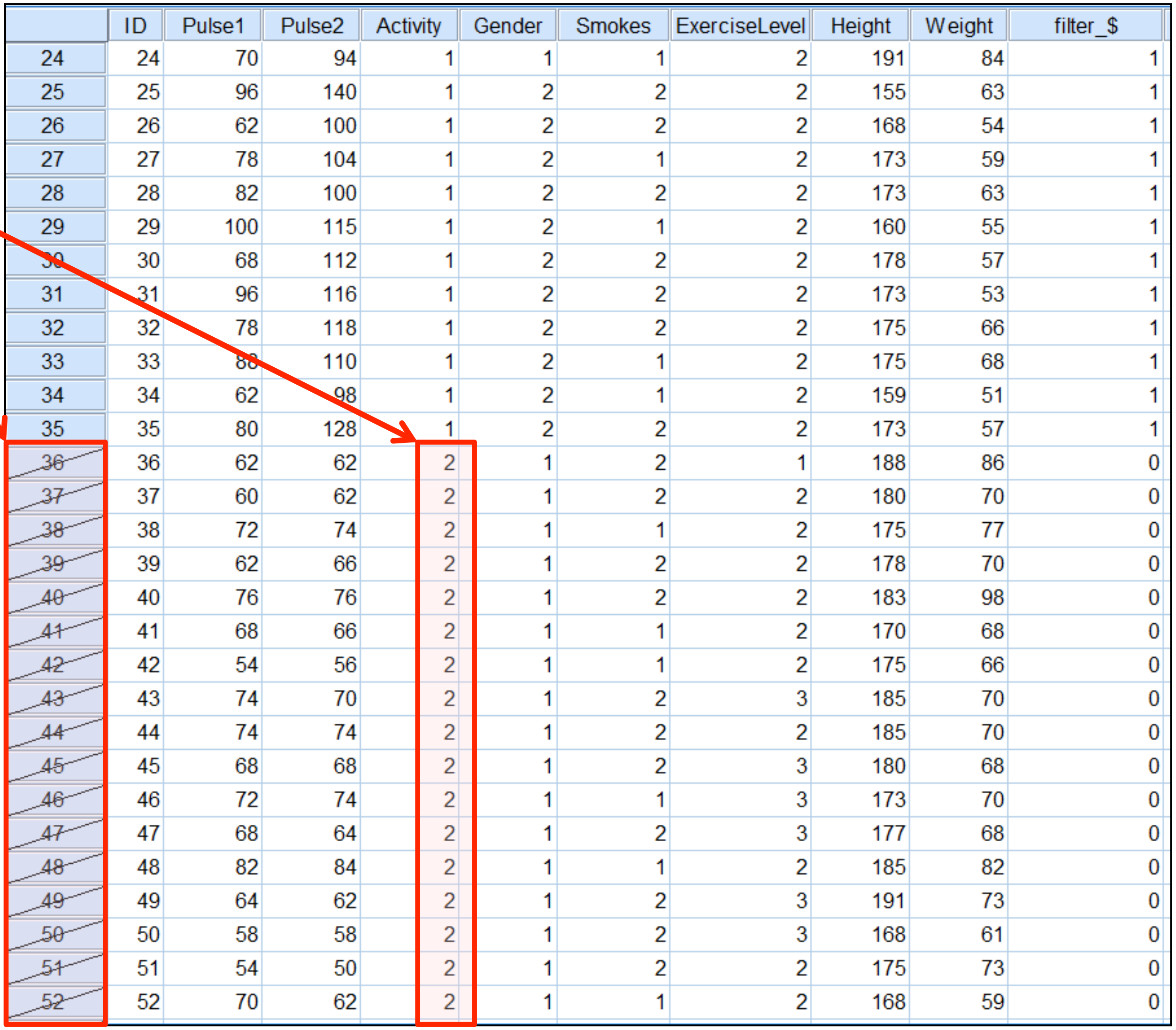


Select: ● If condition is satisfied

Activity = 1

This causes the cases with *Activity* = 2 to be struck out in the Data View:

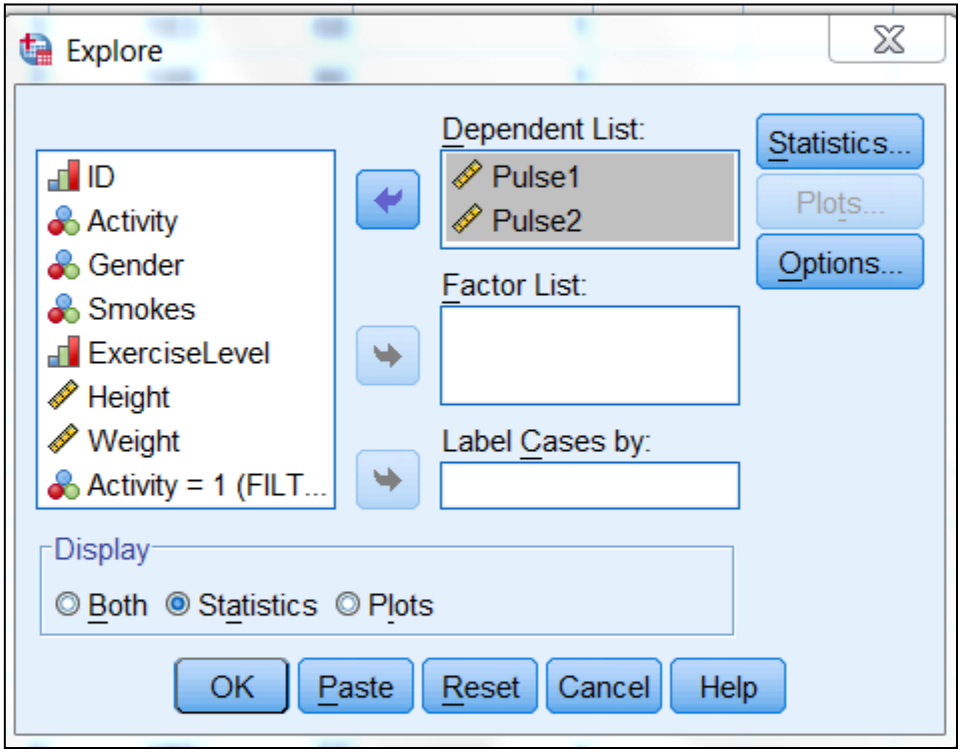
There are 35 cases remaining



	ID	Pulse1	Pulse2	Activity	Gender	Smokes	ExerciseLevel	Height	Weight	filter_$
24	24	70	94	1	1	1	2	191	84	1
25	25	96	140	1	2	2	2	155	63	1
26	26	62	100	1	2	2	2	168	54	1
27	27	78	104	1	2	1	2	173	59	1
28	28	82	100	1	2	2	2	173	63	1
29	29	100	115	1	2	1	2	160	55	1
30	30	68	112	1	2	2	2	178	57	1
31	31	96	116	1	2	2	2	173	53	1
32	32	78	118	1	2	2	2	175	66	1
33	33	88	110	1	2	1	2	175	68	1
34	34	62	98	1	2	1	2	159	51	1
35	35	80	128	1	2	2	2	173	57	1
~~36~~	36	62	62	2	1	2	1	188	86	0
~~37~~	37	60	62	2	1	2	2	180	70	0
~~38~~	38	72	74	2	1	1	2	175	77	0
~~39~~	39	62	66	2	1	2	2	178	70	0
~~40~~	40	76	76	2	1	2	2	183	98	0
~~41~~	41	68	66	2	1	1	2	170	68	0
~~42~~	42	54	56	2	1	1	2	175	66	0
~~43~~	43	74	70	2	1	2	3	185	70	0
~~44~~	44	74	74	2	1	2	2	185	70	0
~~45~~	45	68	68	2	1	2	3	180	68	0
~~46~~	46	72	74	2	1	1	3	173	70	0
~~47~~	47	68	64	2	1	2	3	177	68	0
~~48~~	48	82	84	2	1	1	2	185	82	0
~~49~~	49	64	62	2	1	2	3	191	73	0
~~50~~	50	58	58	2	1	2	3	168	61	0
~~51~~	51	54	50	2	1	2	2	175	73	0
~~52~~	52	70	62	2	1	1	2	168	59	0

Exploring the confidence intervals for the means

- ☐ Select Analyze – Descriptive Statistics – Explore
- ☐ Select *Pulse1* and *Pulse2* as dependent variables
- ☐ Under Display select Statistics



Dependent List: Pulse1, Pulse2

Factor List:

Label Cases by:

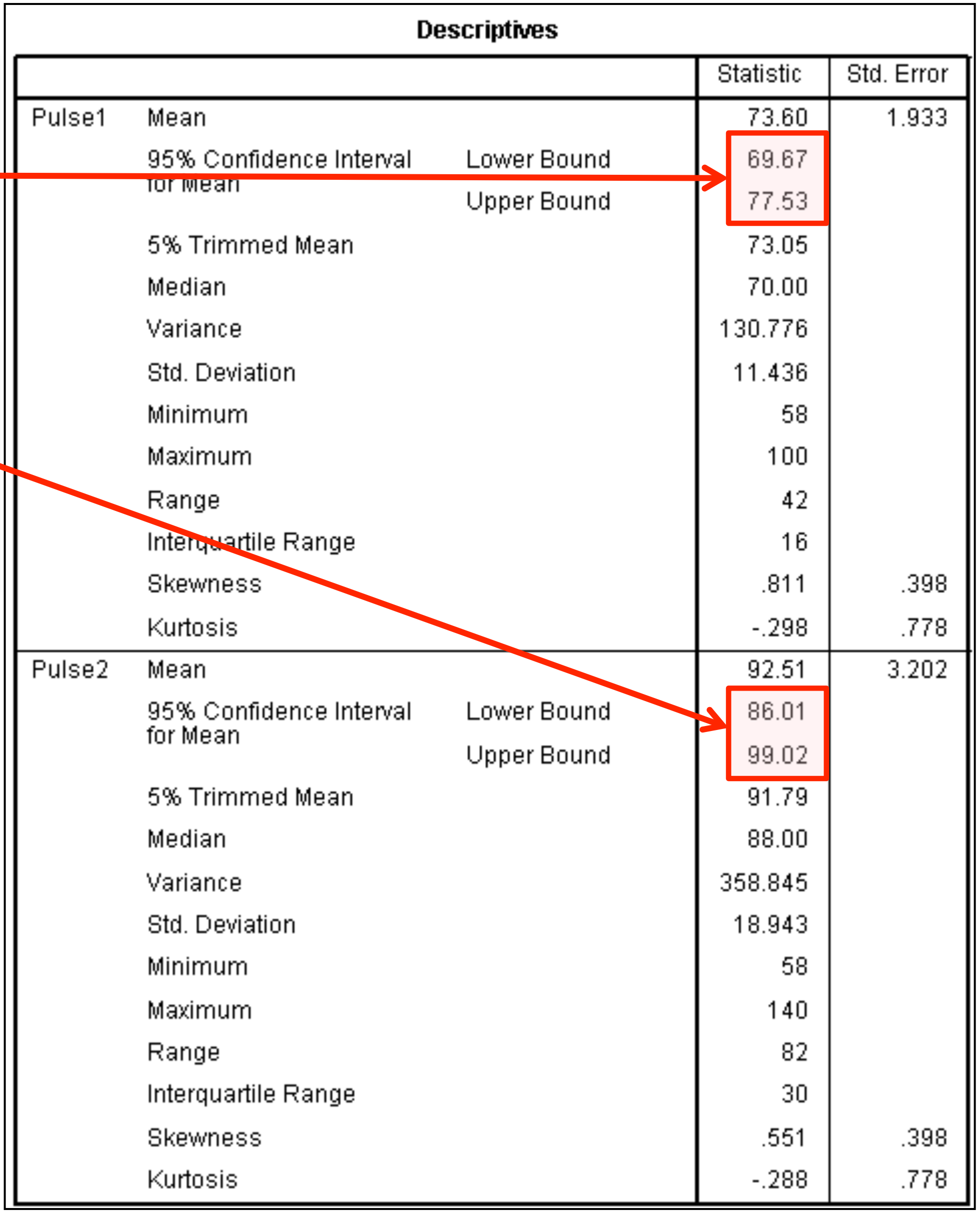
Display: ○ Both ● Statistics ○ Plots

95% confidence interval for *Pulse1* is between 69.67 and 77.53

95% confidence interval for *Pulse2* is between 86.01 and 99.02

As these intervals do not overlap we have evidence that the means are different.

However, we need a formal procedure for making this decision.

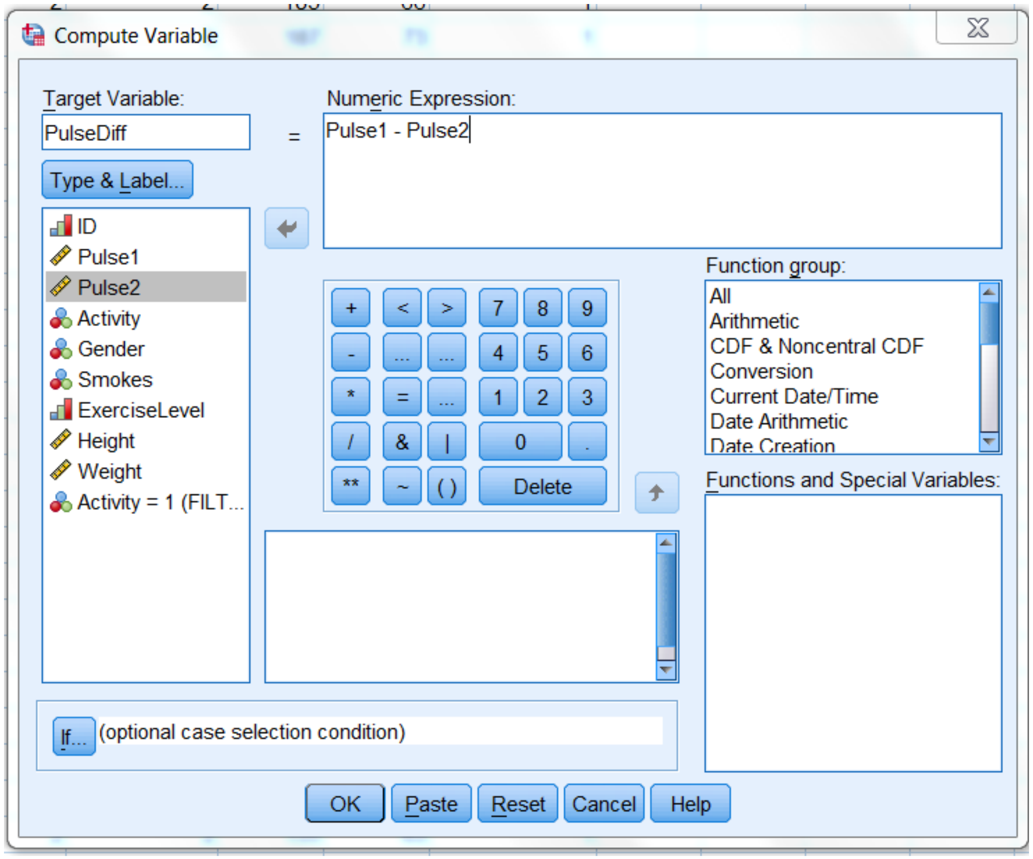


Descriptives

			Statistic	Std. Error
Pulse1	Mean		73.60	1.933
	95% Confidence Interval for Mean	Lower Bound	69.67	
		Upper Bound	77.53	
	5% Trimmed Mean		73.05	
	Median		70.00	
	Variance		130.776	
	Std. Deviation		11.436	
	Minimum		58	
	Maximum		100	
	Range		42	
	Interquartile Range		16	
	Skewness		.811	.398
	Kurtosis		-.298	.778
Pulse2	Mean		92.51	3.202
	95% Confidence Interval for Mean	Lower Bound	86.01	
		Upper Bound	99.02	
	5% Trimmed Mean		91.79	
	Median		88.00	
	Variance		358.845	
	Std. Deviation		18.943	
	Minimum		58	
	Maximum		140	
	Range		82	
	Interquartile Range		30	
	Skewness		.551	.398
	Kurtosis		-.288	.778

Comparing paired data

- It is better to compare paired data by subtracting one variable from the other
- Select Transform – Compute Variable…
- Enter *PulseDiff* as the Target Variable
- Enter *Pulse2 – Pulse1* as the Numeric Expression using the list of variables



Target Variable: PulseDiff

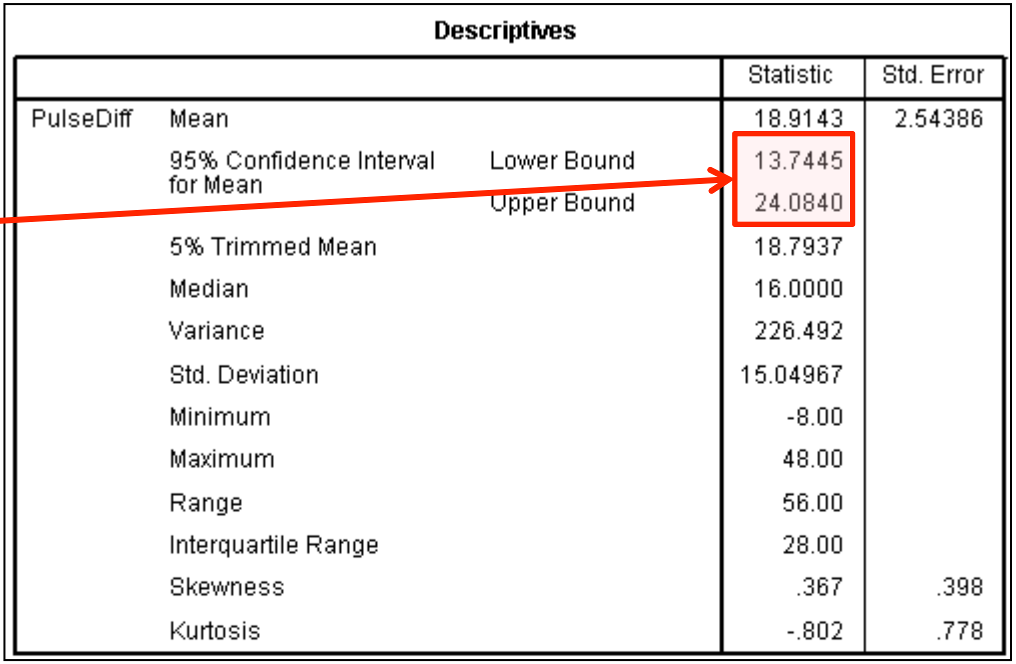
Numeric Expression: Pulse1 - Pulse2

Explore *PulseDiff*

We can now repeat the Explore process for this new variable:

95% confidence interval for *PulseDiff* is between 13.74 and 24.08.

This is clearly positive, meaning the pulse has increased, but it is still not a formal decision procedure.



Descriptives

			Statistic	Std. Error
PulseDiff	Mean		18.9143	2.54386
	95% Confidence Interval for Mean	Lower Bound	13.7445	
		Upper Bound	24.0840	
	5% Trimmed Mean		18.7937	
	Median		16.0000	
	Variance		226.492	
	Std. Deviation		15.04967	
	Minimum		-8.00	
	Maximum		48.00	
	Range		56.00	
	Interquartile Range		28.00	
	Skewness		.367	.398
	Kurtosis		-.802	.778

Plot a graph of *PulseDiff*

As in Workshop 7, we can create a histogram of *PulseDiff* with a normal curve superimposed:

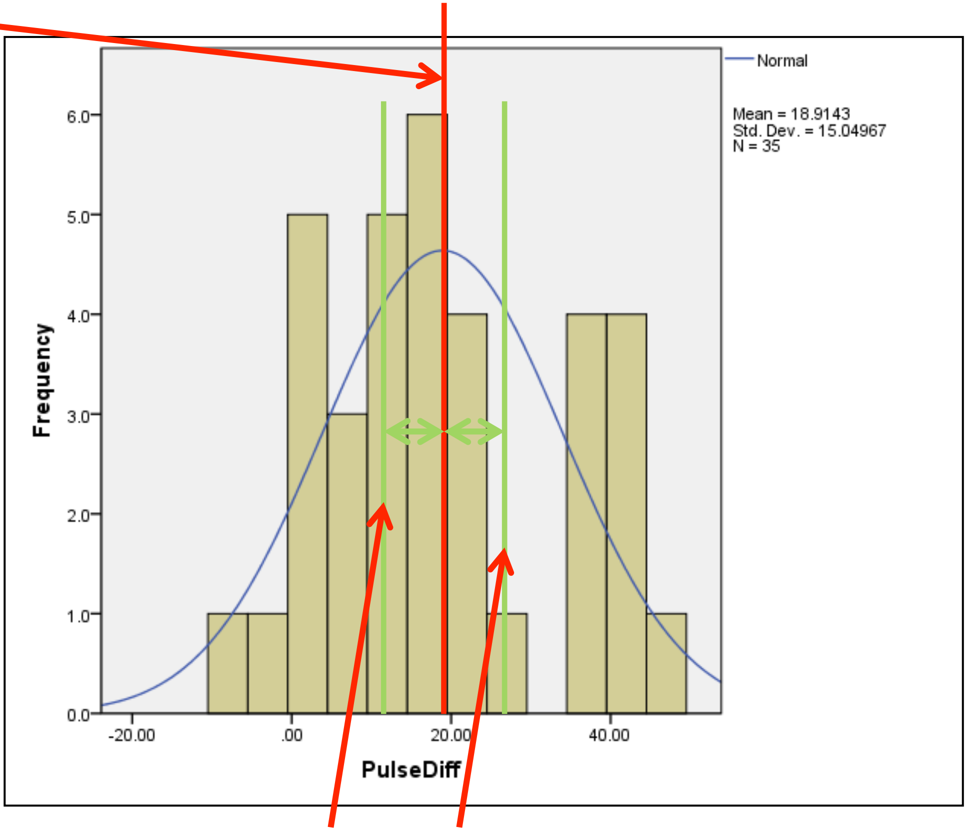
Sample mean = 18.91

Standard deviation = 15.05

Standard error of mean = *Standard deviation*$/\sqrt{N}$

$=15.05/\sqrt{35}\ =2.54$

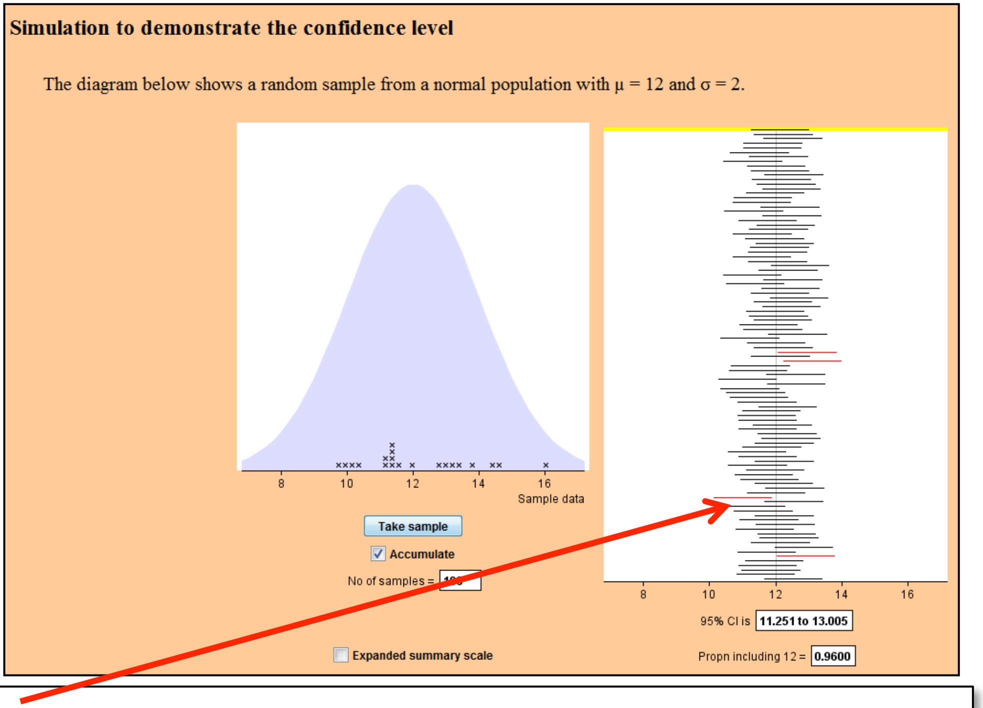
This estimates how much the **mean** is expected to vary if repeated samples were obtained



95% confidence interval for mean is ±1.96 × standard error

What does a 95% confidence interval mean?

- ☐ We **should not** say, “There is a 95% chance that the (population) mean of *PulseDiff* is between 13.74 and 24.08”
- ☐ But we **should** say: “If we carried out the same study 100 times, approximately 95 of the confidence intervals would cover the true population mean”
- ☐ See Section 9.3.3 of: http://cast.massey.ac.nz/core/index.html?



Simulation to demonstrate the confidence level

The diagram below shows a random sample from a normal population with μ = 12 and σ = 2.

☑ Accumulate

No of samples = 100

95% CI is 11.251 to 13.005

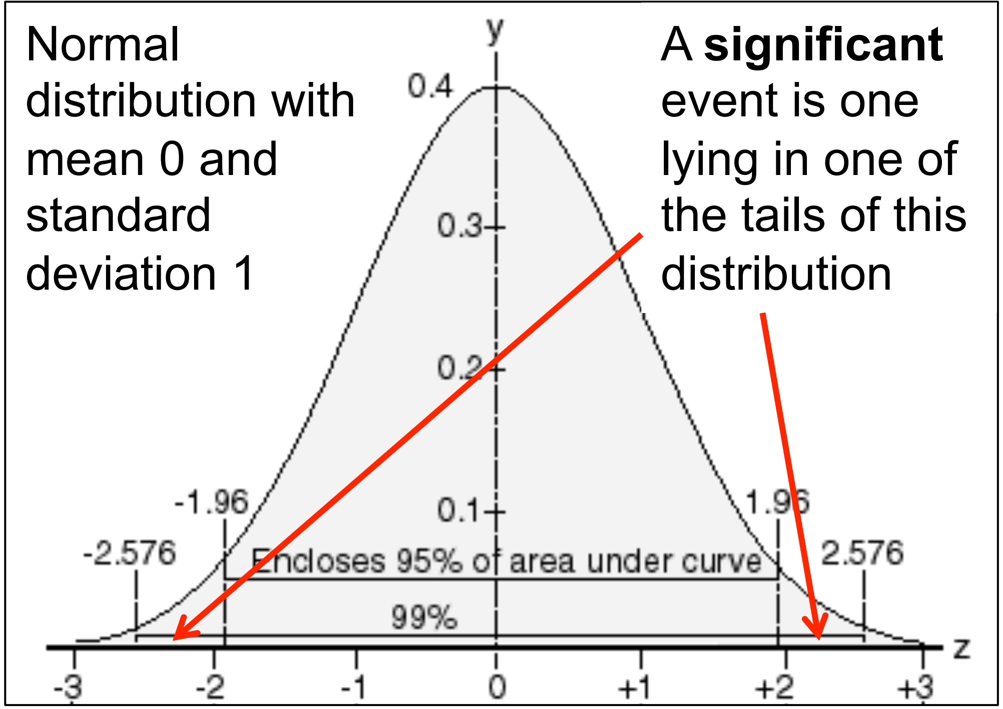
Propn including 12 = 0.9600

☐ Expanded summary scale

Confidence intervals in red do not contain the population mean

What is parametric statistics?

- A form of statistical testing (inference) which assumes data comes from a **distribution** (defined by parameters)
- Often this is the **normal distribution**, whose parameters are a **mean** and a **standard deviation**
- We may need to test for this first
- Normally only use numerical (scale) data
- Classic ‘mistake’: turning Likert values (strongly agree, etc.) into numbers and using a t-test

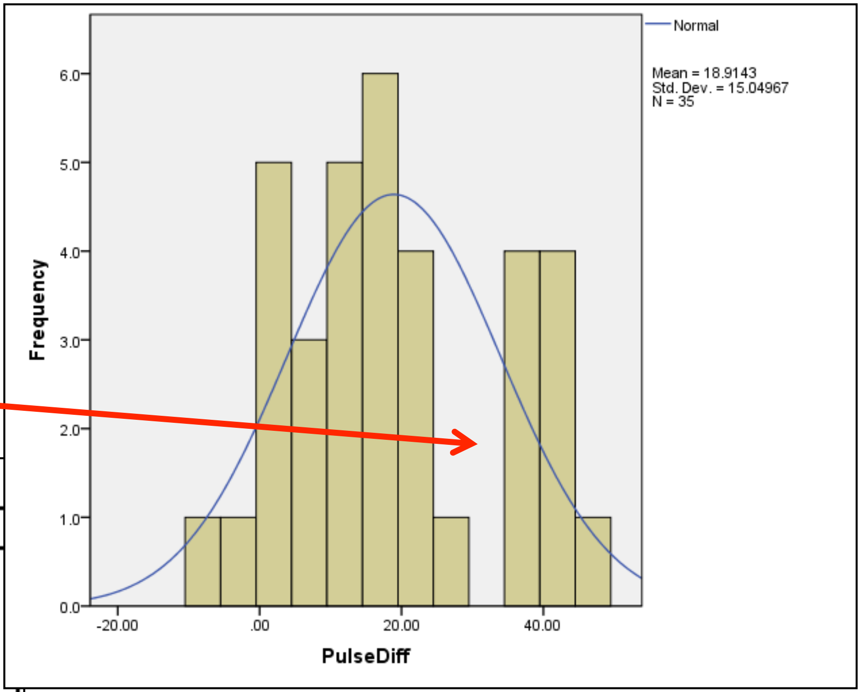


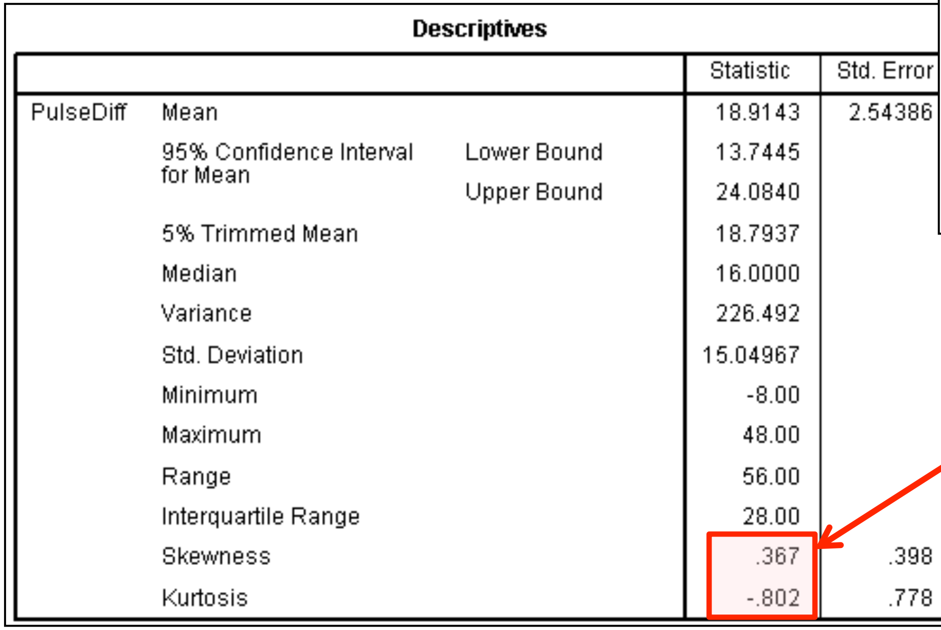
Normal distribution with mean 0 and standard deviation 1

A **significant** event is one lying in one of the tails of this distribution

Check *PulseDiff* for normaility

- ❑ Graph of *PulseDiff* seemed to be fairly flat (negative kurtosis)
- ❑ No sign of skewness
- ❑ Data gap is a worry





Descriptives

			Statistic	Std. Error
PulseDiff	Mean		18.9143	2.54386
	95% Confidence Interval for Mean	Lower Bound	13.7445	
		Upper Bound	24.0840	
	5% Trimmed Mean		18.7937	
	Median		16.0000	
	Variance		226.492	
	Std. Deviation		15.04967	
	Minimum		-8.00	
	Maximum		48.00	
	Range		56.00	
	Interquartile Range		28.00	
	Skewness		.367	.398
	Kurtosis		-.802	.778

Skewness is 0.367

Kurtosis in -0.802

The null and alternative hypotheses

- Statistical testing is about making a decision about the significance of a data feature. We usually assume that this feature was just a random event and then seek to measure how unlikely such an event was.
- The statement of this position is known as the **null hypothesis** and is written **H_0**
- With statistical testing we try to **reject** the null hypothesis **with a certain level of confidence** based on the probability (or ‘P-’) value of the test statistic
- This is like assuming an accused person in **innocent** then **convicting** them **beyond reasonable doubt**
- The logical opposite of the null hypothesis is known as the **alternative hypothesis**

Standard significance levels and the null hypothesis (H_0)

P-value of test statistic	Significant?	Formal action	Informal interpretation
> 0.1	No	Do not reject H_0	No evidence to reject H_0
Between 0.1 and 0.05	No	Do not reject H_0	**Weak evidence** to reject H_0
Between 0.05 and 0.01	Yes: at 95%	Reject H_0 at 95% confidence	**Evidence** to reject H_0
Between 0.01 and 0.001	Yes: at 99%	Reject H_0 at 99% confidence	**Strong evidence** to reject H_0
Less than 0.001	Yes: at 99.9%	Reject H_0 at 99.9% confidence	**Very strong evidence** to reject H_0

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

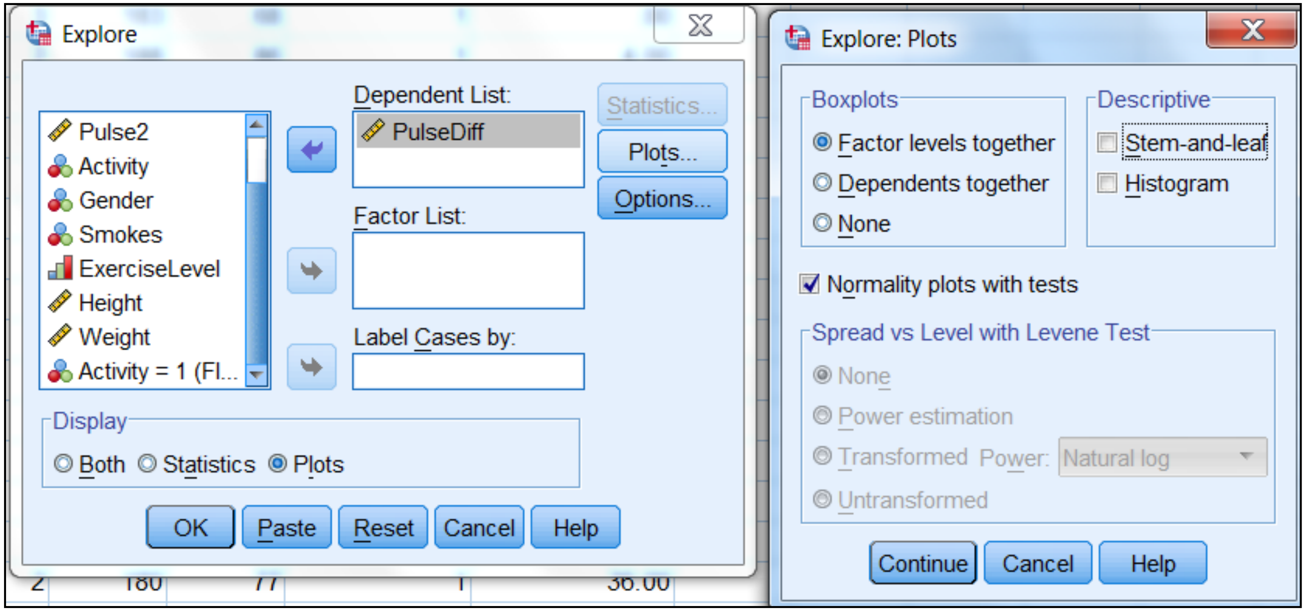
Things that can go wrong

	H_0 really true	**H_0 really false**
H_0 rejected	Type I error	Correct decision
H_0 accepted	Correct decision	Type II error

- Type I error is equivalent to **convicting the innocent**
- Type II error is equivalent to **acquitting the guilty**
- Reducing the chance of a Type I error by changing the significance threshold increases the chance of a Type II error
- The best solution is to **increase the sample size**
- The **power** of a test is **1 – Probability(Type II error)**
- More details in Workshop 13

First: test for normality using the Shapiro-Wilk test

- ❑ Use this test for sample sizes <2000; otherwise use the Kolmogorov-Smirnov test
- ❑ Select: *Analyze – Descriptive Tests – Explore*
- ❑ Select *PulseDiff* in the dependent variable list
- ❑ Select Plots… and Normality plots with tests
- ❑ H_0: The data **is** normally distributed


Explore

Dependent List: PulseDiff

Display: ○ Both ○ Statistics ● Plots

Explore: Plots

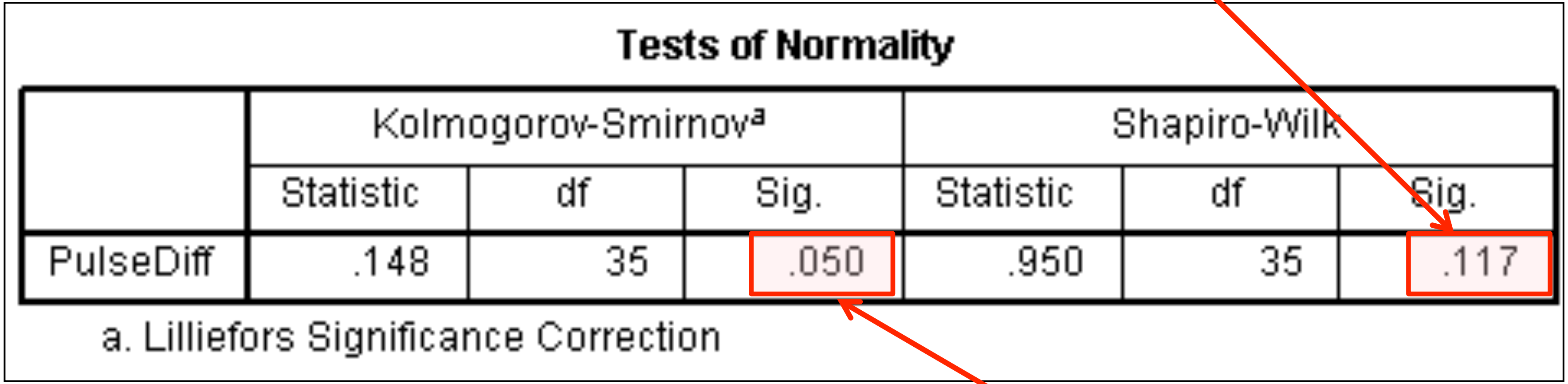
Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☐ Stem-and-leaf ☐ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

Probability value of Shapiro-Wilk test is not significant (>0.1): No evidence that this data set is **not** normally distributed



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
PulseDiff	.148	35	.050	.950	35	.117

a. Lilliefors Significance Correction

However, the Kolmogorov-Smirnov test probability value is lower: This test is more sensitive to gaps in the data

Which test?

Data type	Association or correlation (two variables)	Difference (comparison: one variable, two groups)	
		Related (same subjects)	Unrelated (different subjects)
Nominal	Chi-squared	N/A	Chi-squared
Ordinal	Spearman[1] or Chi-squared	Wilcoxon	Mann-Whitney U[1] or Chi-squared
Scale	Pearson/Linear regression[2] or Spearman	Paired samples t-test[2] or Wilcoxon	Independent samples t-test[2] or Mann-Whitney U

Note:

1. Use these tests when each variable has at least about 10 values, otherwise use the other test
2. These parametric tests should be used only when their assumptions are satisfied, otherwise use the other (nonparametric) test

www.statstutor.ac.uk

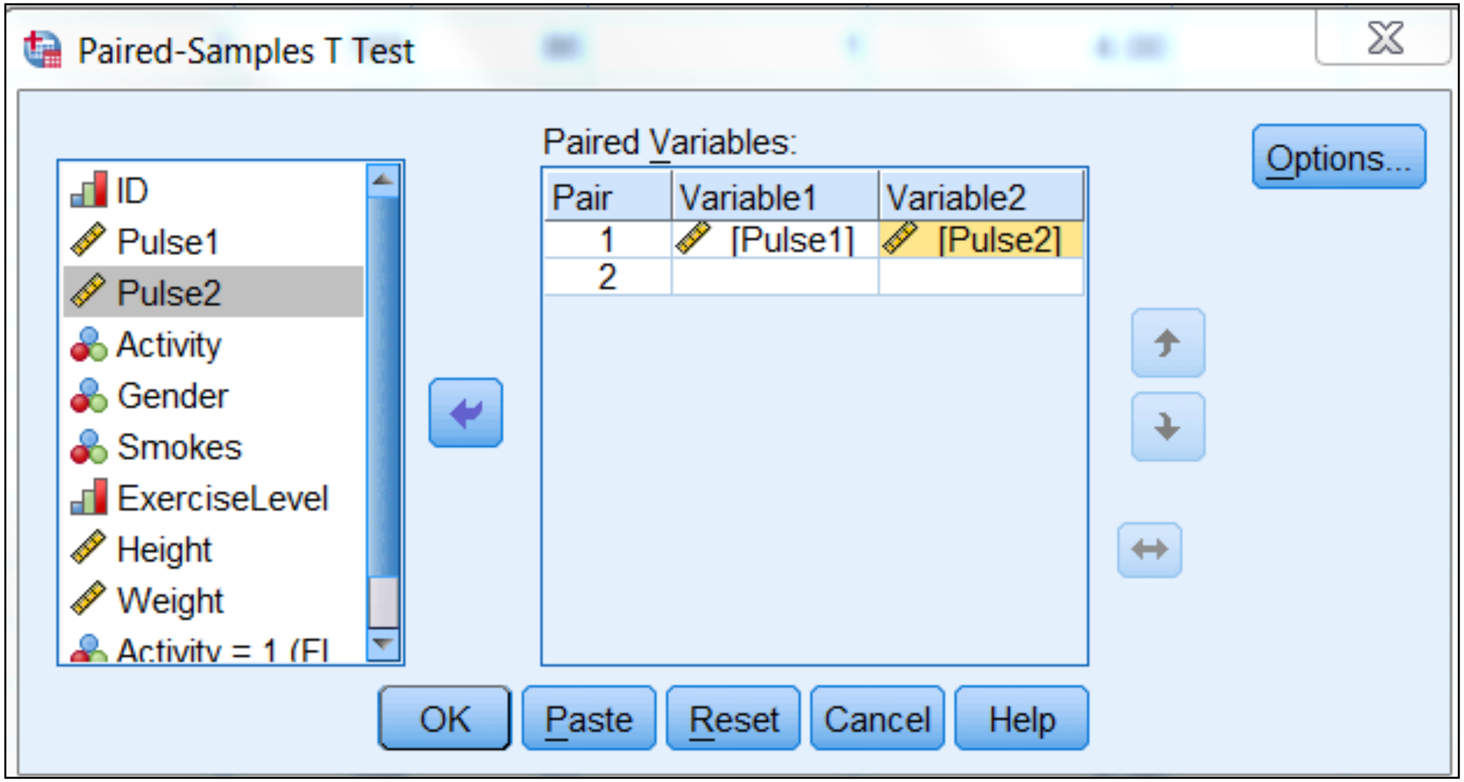
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Test for Research Question 1: Paired t-test

- Applies to the same subjects with two (scaled-based) data values ("within")
- Tests the difference between the means of the two samples
- Here: *Pulse1* and *Pulse2*
- Assumes *Pulse2 – Pulse1* is normally distributed – but is fairly robust to non-normal data sets (see later for details)
- H_0: The difference in the means of the two pulse rates is zero

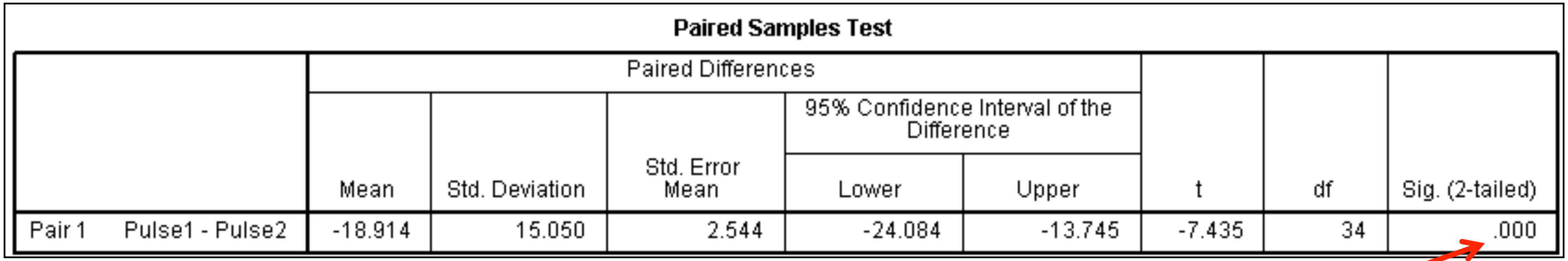
- Select: Analyze – Compare Means – Paired-Samples T Test
- Select *Pulse1* as *Variable1* and *Pulse2* as *Variable2* in Pair 1



Paired Variables:

Pair	Variable1	Variable2
1	[Pulse1]	[Pulse2]
2		

Note: SPSS subtracts *Variable2* from *Variable1* so the mean is now negative



Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Pulse1 - Pulse2	-18.914	15.050	2.544	-24.084	-13.745	-7.435	34	.000

Gives a significance value of “0.000”, meaning 0.000 to 3 decimal places, or less than 0.0005. Thus we reject the null hypothesis with 99.9% confidence and conclude there is very strong evidence that running on the spot changes the pulse rate.

Robustness

- ☐ Parameter-based statistical tests make certain assumptions in their underlying models
- ☐ However, they often work well in other situations where these assumptions are violated
- ☐ This is known as **robustness**
- ☐ **Note:** Statisticians have different opinions on robustness: the advice given here is 'middle of the road'

Robustness of paired t-test

The paired t-test is robust to deviations from normality (ours was OK as the Shapiro-Wilk test was not significant) provided:

- ☐ The sample size is not small (>30, ours was 35, so it is OK)
- ☐ The **effect size** (|*mean*| ÷ *standard deviation* of *Pulse1* – *Pulse2*) is not small (>0.3); here it is OK:

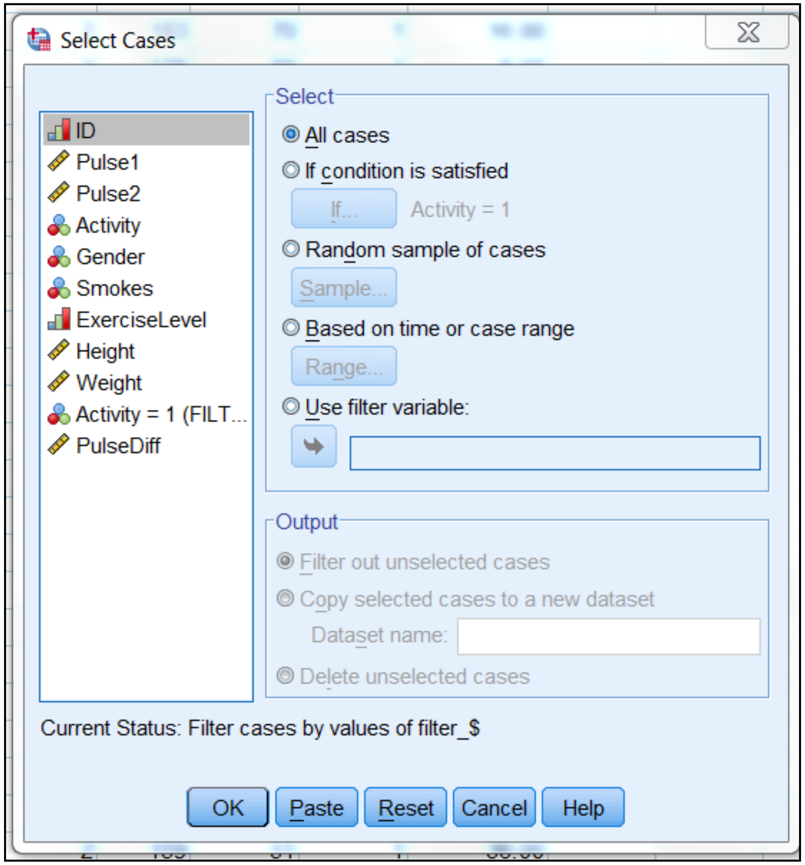
Mean	**|Mean|**	**Standard deviation**	**|Mean| ÷ standard deviation**
-18.91	18.91	15.05	1.26

Source:

Zumbo, B. D. and Jennings, M. J. (2002) The Robustness of validity and efficiency of the related samples t-Test in the presence of outliers, *Psicológica*, 23(2), pp. 415-450.

Research Question 2

- ☐ Do regular smokers have a different pulse rate when sitting than non-/non-regular smokers?
- ☐ In order to investigate this question we will need to compare the *Pulse1* values for smokers against non-/ non-regular smokers
- ☐ The first step is to select all the cases:
  - ➢ Data – Select cases… – All cases



Select Cases

Select

● All cases

○ If condition is satisfied

Activity = 1

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

Output

● Filter out unselected cases

○ Copy selected cases to a new dataset

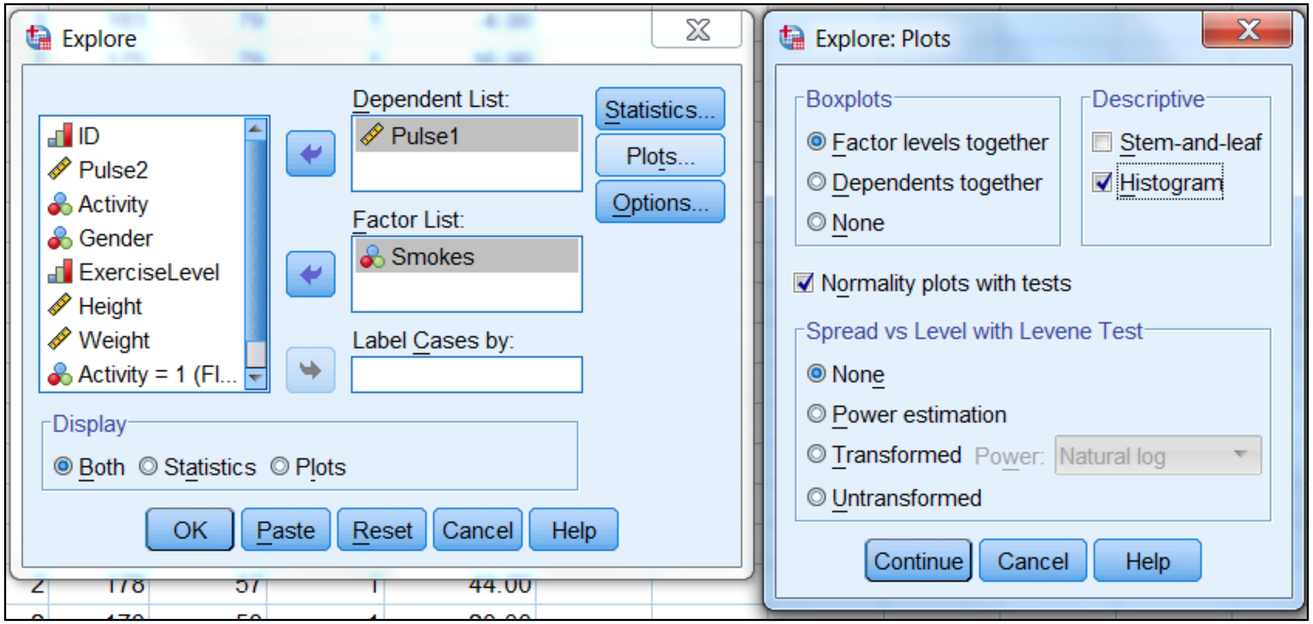
Dataset name:

○ Delete unselected cases

Current Status: Filter cases by values of filter_$

Explore *Pulse1* against *Smokes*

- ☐ Select Analyze – Descriptive statistics – Explore
- ☐ Select *Pulse1* on the Dependent List and Smokes on the Factor list
- ☐ Select Both under Display and click on Plots…
- ☐ Then select Histogram and Normality plots with tests



Explore

Dependent List: Pulse1

Factor List: Smokes

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Explore: Plots

Boxplots: ● Factor levels together ○ Dependents together ○ None

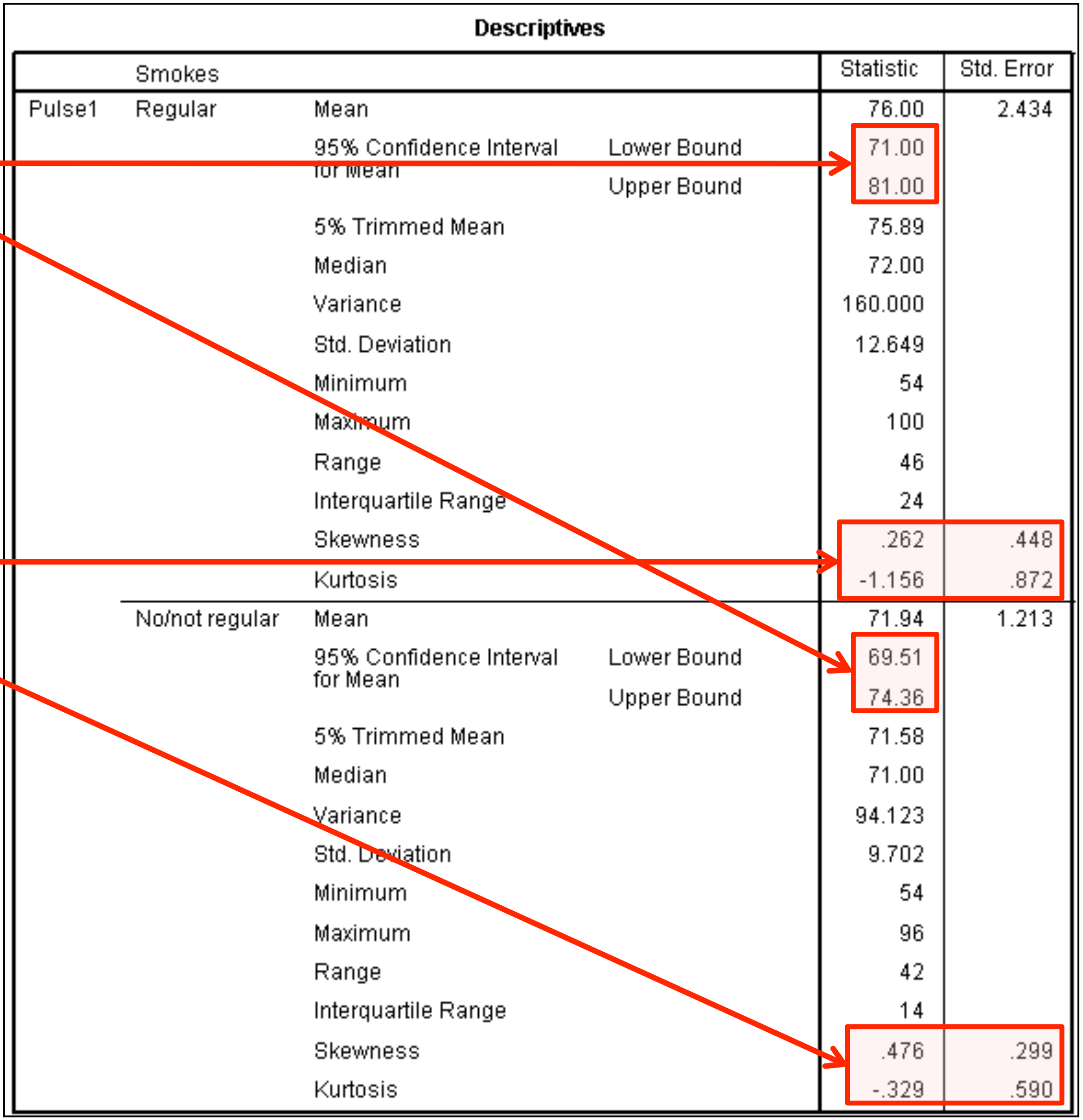
Descriptive: ☐ Stem-and-leaf ☑ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

Confidence intervals overlap so it seems unlikely that we will get a significant result

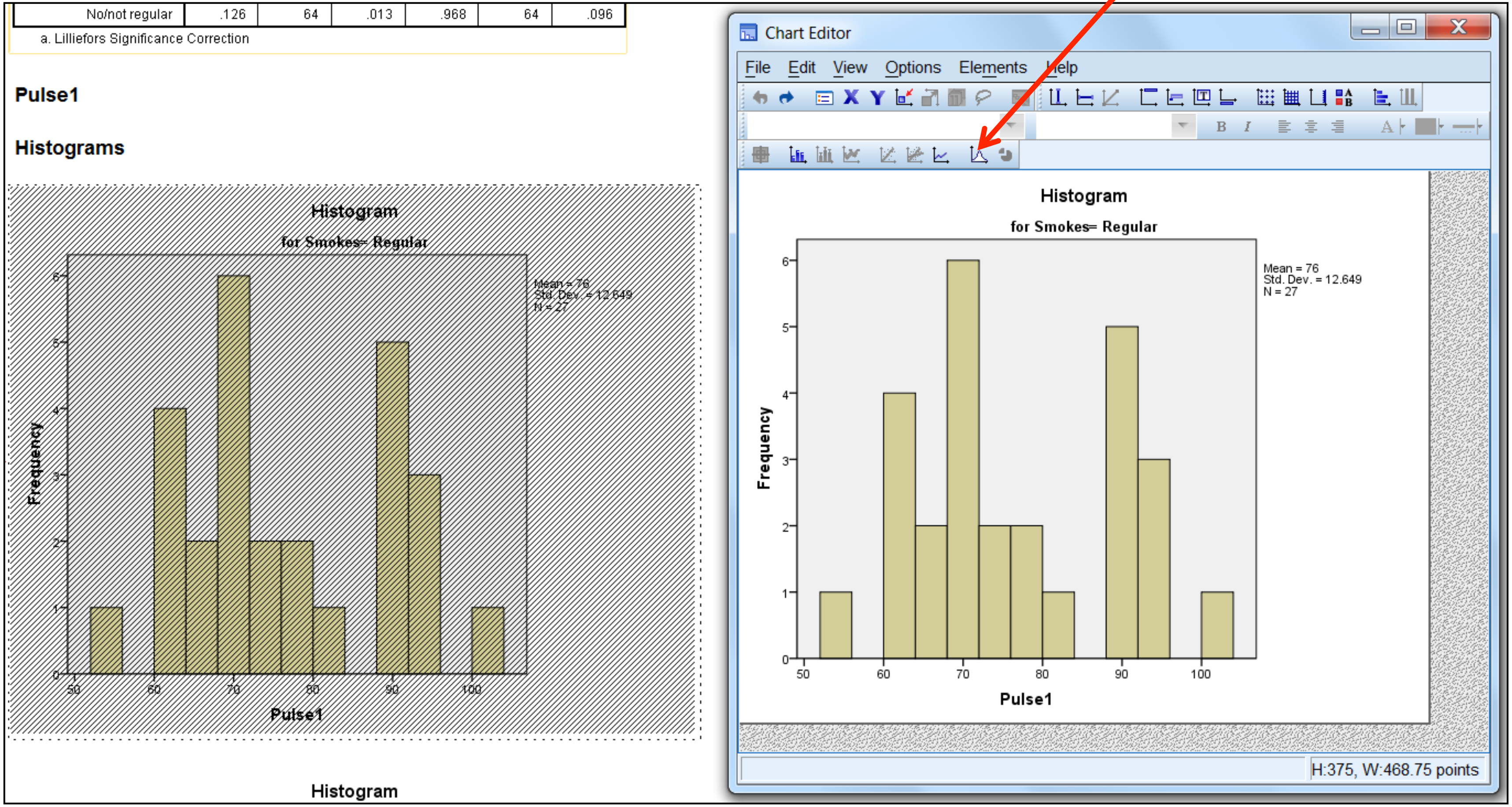
Smokes kurtosis quite low, *No/not regular* skewness a bit high, otherwise OK (absolute value should be less than 1.96 × standard error)



Descriptives

	Smokes			Statistic	Std. Error
Pulse1	Regular	Mean		76.00	2.434
		95% Confidence Interval for Mean	Lower Bound	71.00	
			Upper Bound	81.00	
		5% Trimmed Mean		75.89	
		Median		72.00	
		Variance		160.000	
		Std. Deviation		12.649	
		Minimum		54	
		Maximum		100	
		Range		46	
		Interquartile Range		24	
		Skewness		.262	.448
		Kurtosis		-1.156	.872
	No/not regular	Mean		71.94	1.213
		95% Confidence Interval for Mean	Lower Bound	69.51	
			Upper Bound	74.36	
		5% Trimmed Mean		71.58	
		Median		71.00	
		Variance		94.123	
		Std. Deviation		9.702	
		Minimum		54	
		Maximum		96	
		Range		42	
		Interquartile Range		14	
		Skewness		.476	.299
		Kurtosis		-.329	.590

Normal curve approximations can be added to the graphs by double-clicking on them and pressing this button:

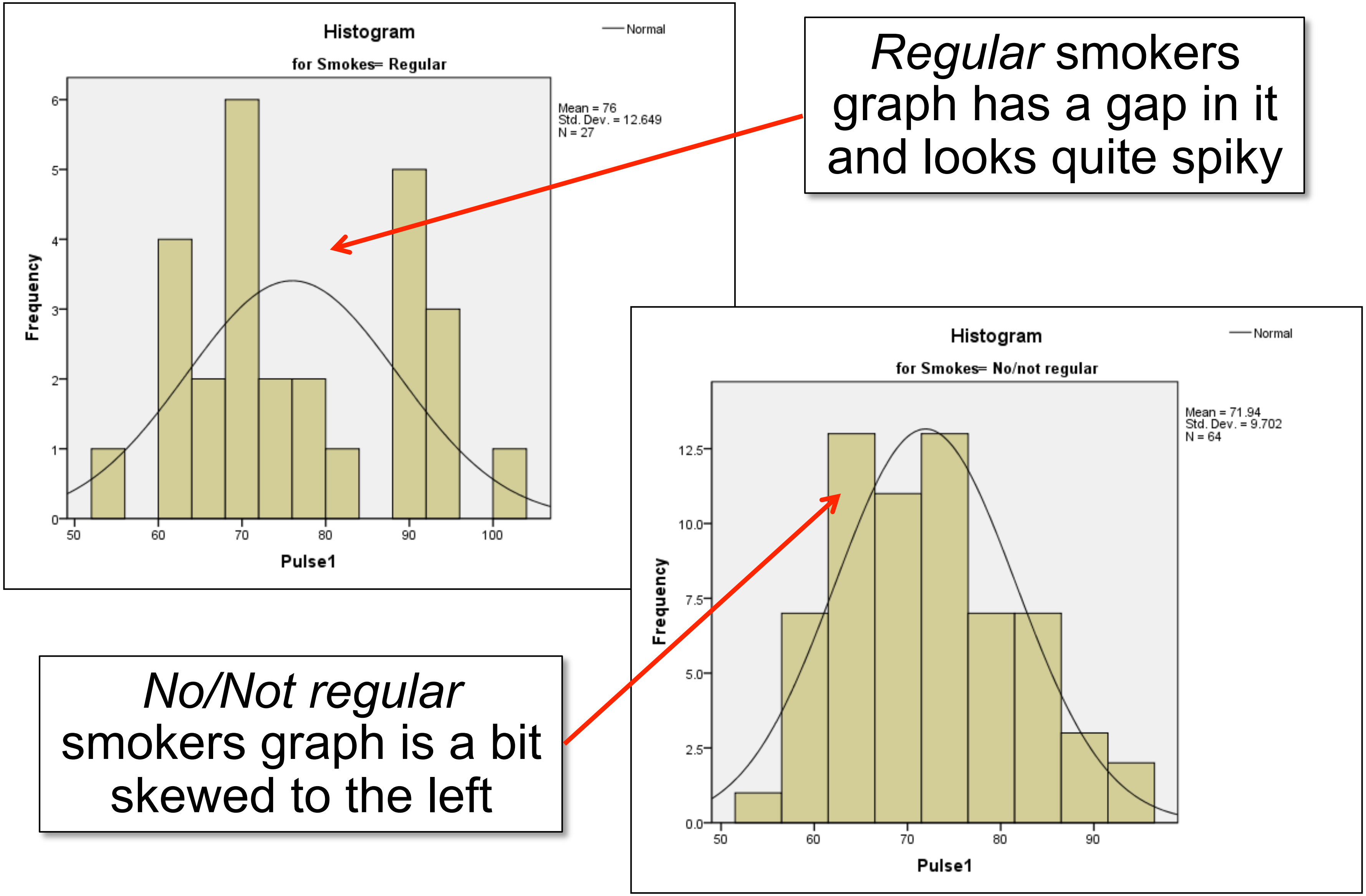


No/not regular	.126	64	.013	.968	64	.096

a. Lilliefors Significance Correction

Pulse1

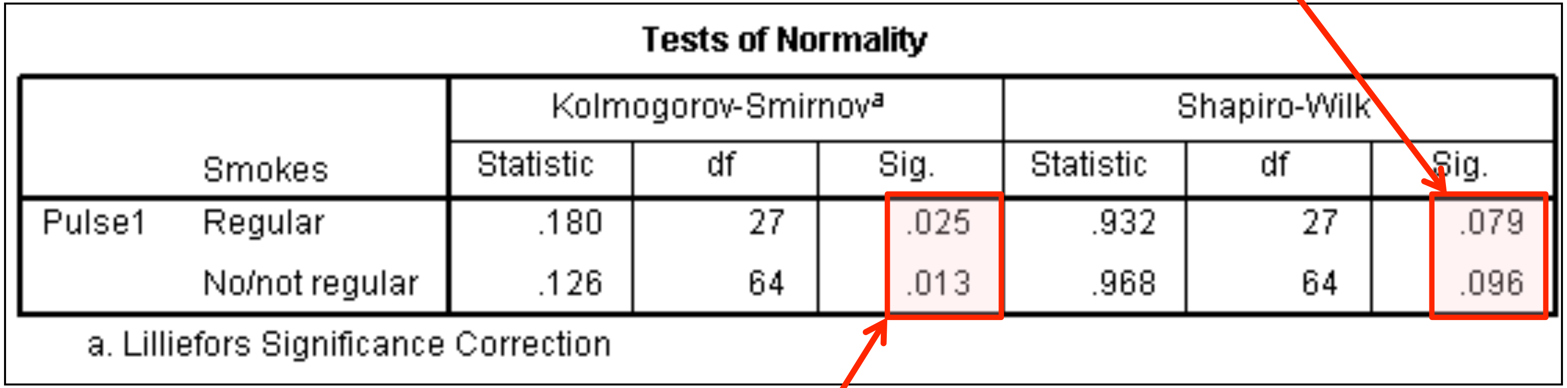
Histograms



Regular smokers graph has a gap in it and looks quite spiky

No/Not regular smokers graph is a bit skewed to the left

- Shapiro-Wilk significance value for both groups is between 0.05 and 0.1 – weak evidence that the data is not normally distributed



Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Smokes	Statistic	df	Sig.	Statistic	df	Sig.
Pulse1	Regular	.180	27	.025	.932	27	.079
	No/not regular	.126	64	.013	.968	64	.096

a. Lilliefors Significance Correction

- Kolmogorov-Smirnov values are even more significant – unusual as Shapiro-Wilk is generally a more powerful test for small samples

Which test?

Data type	**Association or correlation (two variables)**	**Difference (comparison: one variable, two groups)**	
		Related (same subjects)	**Unrelated (different subjects)**
Nominal	Chi-squared	N/A	Chi-squared
Ordinal	Spearman[1] or Chi-squared	Wilcoxon	Mann-Whitney U[1] or Chi-squared
Scale	Pearson/Linear regression[2] or Spearman	Paired samples t-test[2] or Wilcoxon	Independent samples t-test[2] or Mann-Whitney U

Note:

1. Use these tests when each variable has at least about 10 values, otherwise use the other test
2. These parametric tests should be used only when their assumptions are satisfied, otherwise use the other (nonparametric) test

www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Test for Research Question 2: Unpaired t-test

☐ Applies to the different subjects with one (scaled-based) data value each (“between”)

☐ Tests the difference between the means of the two samples

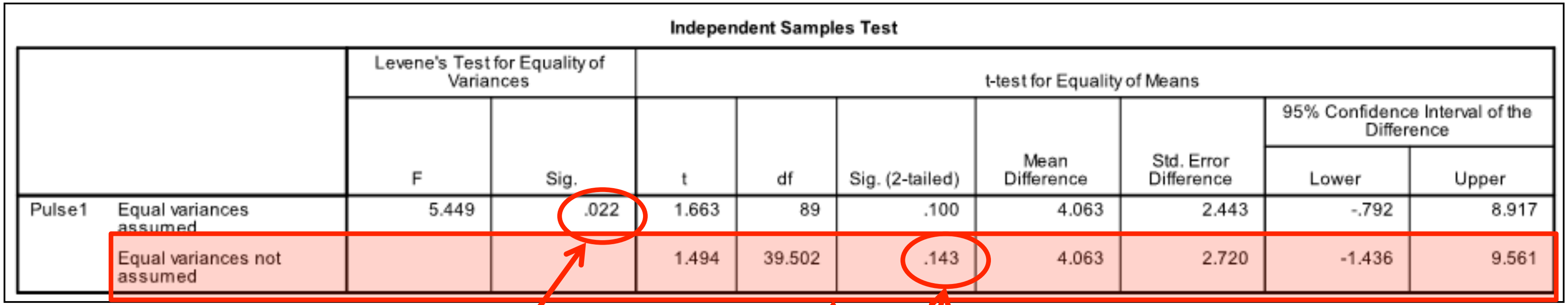
☐ Here: *Regular* and *No/not regular Smokes*

☐ Assumes both groups are normally distributed

☐ H_0: Regular smoking has no effect on *Pulse1*

☐ Two variants – depends whether or not variances can be assumed equal (use Levene’s test first, H_0: Variances are equal)

- Select: *Analyze – Compare Means – Independent-Samples T Test*
- Select *Pulse1* as the *Test Variable*
- Select *Smokes* as the *Grouping Variable*
- Select *Define Groups:* define *1* as the first group number and *2* as the second group number



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pulse1	Equal variances assumed	5.449	.022	1.663	89	.100	4.063	2.443	-.792	8.917
	Equal variances not assumed			1.494	39.502	.143	4.063	2.720	-1.436	9.561

- Automatically computes Levene's test and outputs both versions:
  - Significant at 95% (so we reject H_0 and conclude that we **cannot** assume the variance are equal
  - So we now look at the bottom row
- t-test not significant at 95% so we conclude that there is no evidence that regular smokers have a different pulse rate when sitting than non-/non-regular smokers

Robustness of unpaired t-test

This test is robust to deviations from normality (ours was borderline as there was weak evidence of non-normality) provided:

1. The sample sizes are equal (here they were unequal)
2. The sample sizes are 25 or more per group (here they were 27 and 64)

Here, although the robustness conditions are not met it might be best to do a non-parametric test as well (see Workshop 9) because of the borderline normality scores.

Source:

Sawilowsky, S. S. and Blair, R. C. (1992) A more realistic look at the robustness and Type II error properties of the t test to departures from population normality, *Psychological Bulletin*, 111(2), pp. 352–360.

Activity

Explore the effect of regular smoking on running on the spot:

☐ Formulate a research question
☐ Produce some descriptive statistics
☐ Make some initial informal observations
☐ Select an appropriate test
☐ Carry out the test
☐ Interpret the results
☐ Check the test and its robustness assumptions

Recap

- ☐ Confidence intervals
- ☐ Parametric statistics
- ☐ Normality testing:
  - ➢ Skewness and kurtosis
  - ➢ Shapiro-Wilk test
- ☐ Paired-samples t-test
- ☐ Independent-samples t-test
- ☐ Assumptions
- ☐ Robustness