



Excel, an Alternative to SPSS

Ahmed A. Mirza, PhD

King Fahad Research Center

KAU

Learning Outcomes

By the end of this workshop, participants will be able to:

- Explain p-value as it relates to the null and alternative hypotheses
- Explain and employ the different components of descriptive analysis according to data type:
 - Central tendency (mean, median and mode)
 - Spread (standard deviation, interquartile range and quartile deviation)
- Judge the normality of data distribution using central tendency, spread values, and online resources
- Judge the variance of data using Leven's test and online resources
- Employ the correct data analysis tools from excel to analyze data

Excel Functions

- Descriptive
Average/median/mode/stddev/skew/quartile
ftest/ttest/chi
- Data tool pack
 - Descriptive (including confidence Interval)
 - Percentile
 - Histogram
 - Leven's test of variance (ftest)
 - T-test (ttest)
 - Correlation
 - ANOVA

Review

- Differentiate between the null and alternative hypotheses, one- or two- tailed.
- Judge the identity of a variable as independent or dependent
- Differentiate amongst the types of data (continuous – interval, ratio, or discrete; and categorical – nominal, dichotomous or ordinal)
- Judge the data to be parametric or nonparametric

What is a p-value?

- The p-value is defined as the **probability** of obtaining a result equal to or "more extreme" than what was actually observed, when the **null hypothesis is true**.
- The probability that the observed data exhibits the effects of the independent on the dependent variables as it is when the null hypothesis is true.
- **Null hypothesis states that there is no association or no effect**

Types of errors

- Type 1 error: a false positive finding (alpha)
 - Telling a patient he has the disease, when he in fact does not (incorrectly rejecting the null hypothesis)
- Type 2 error: a false negative finding (beta)
 - Misdiagnosing a patient as healthy, when he actually has cancer (not rejecting a false null hypothesis)

What is a p-value?

- If the p-value is lower or equal to the cut-off (alpha) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value greater than the cut-off (alpha) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.

Degrees of Freedom (df)

- The number of factors in a calculation that we can vary and still achieve a specific outcome
- It is needed for best approximation in certain statistical tests (to build models that resembles the data)
- Example
 - Using a dice, need to choose five numbers that will add up to 12
 - You can only choose 4
 - Need to choose three numbers that will be averaged to 5
 - You can only 2

Parametric vs. Nonparametric

- All categorical data are **nonparametric** (dichotomous, nominal and ordinal)
- Continuous (ratio and interval) data is **parametric** only if it has a normal distribution, if it is not normal then it is **nonparametric**. (normal distribution exhibits a bell shaped curve)

Ratio, interval or discrete

- Ratio: all whole numbers where zero means no measurement or no value (e.g., weight, volume)
- Interval: all whole numbers where zero is meaningful (e.g., temp, altitude [below or above sea level])
- Discrete: numbers with no fractions (e.g., number of cars in an accident)

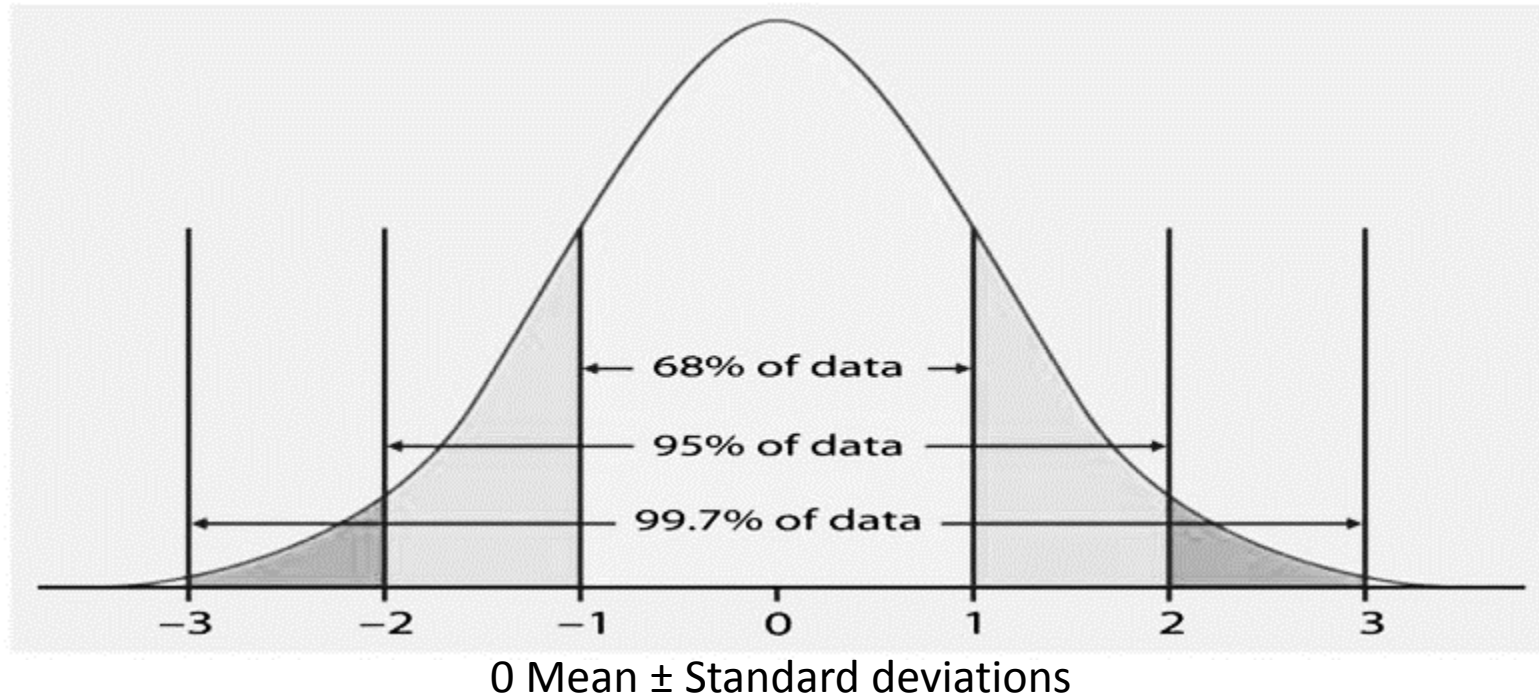
Descriptive stats

Values that are used to describe the shape of the data so to aid in deciding on the statistical model to be tested

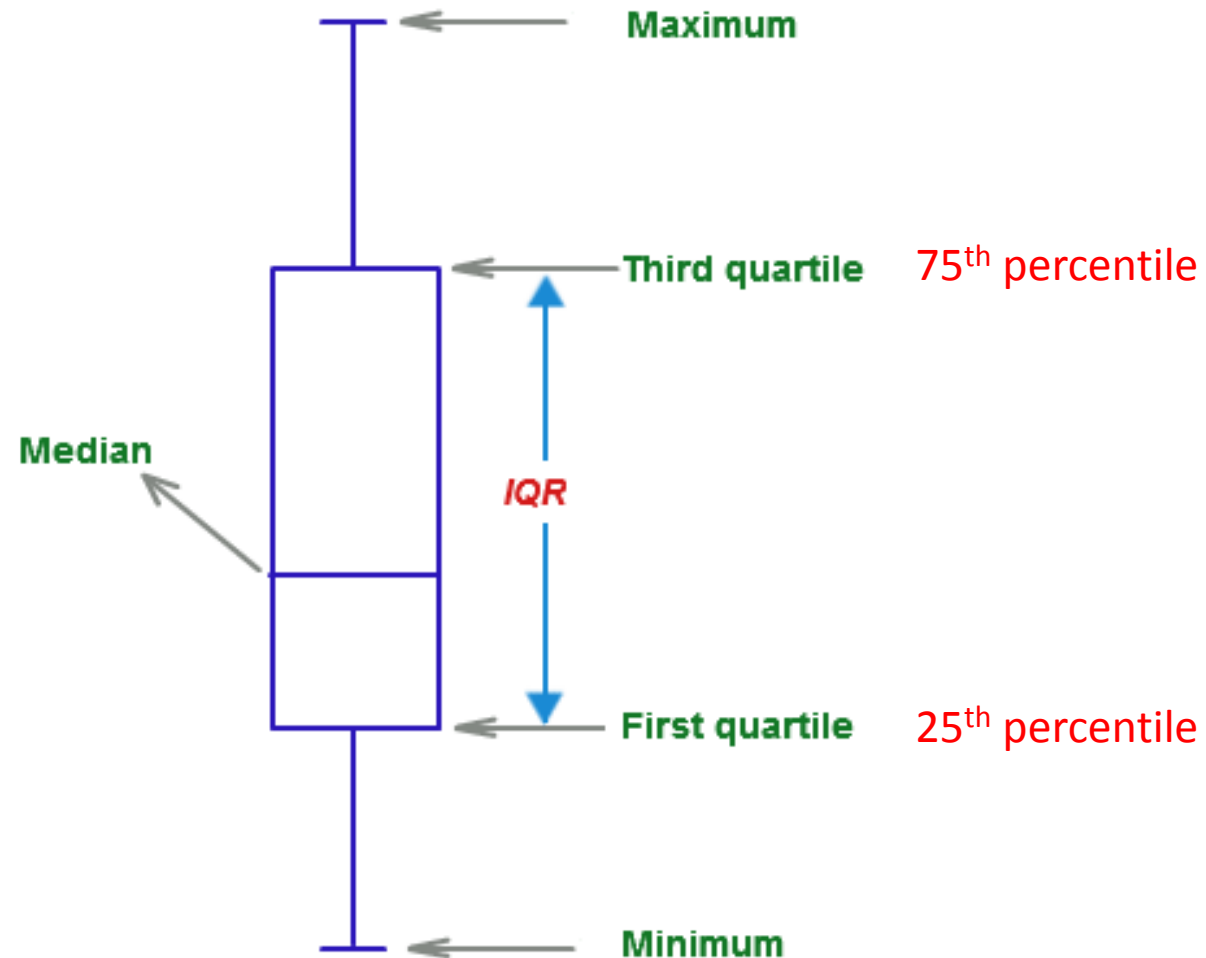
- Central tendency:
 - Describes the center of samples and its localization
 - mean, median and mode
- Spread (variability):
 - Describes the dispersion and variation of samples
 - Standard Deviation, interquartile range and quartile deviation

Standard deviation

The mean \pm 3 standard deviations cover 99.7% of the data under a bell-shaped curve



Box (whisker) plots



Normality of data distribution

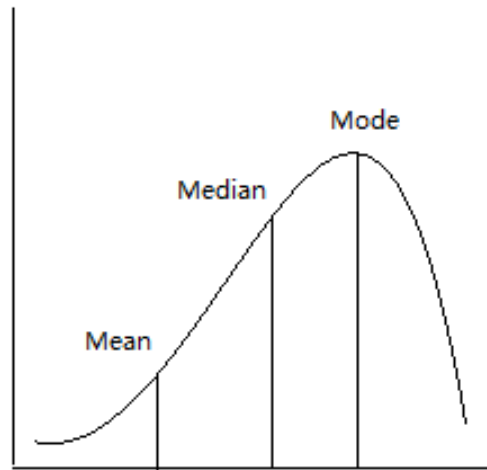
- The shape of the data is a key in determining statistical models to be tested
- The normal bell shaped curve occurs naturally
- Most continuous variables follow a normal distribution
- Many statistical tests require that data is normally distributed

Signs of normal distribution

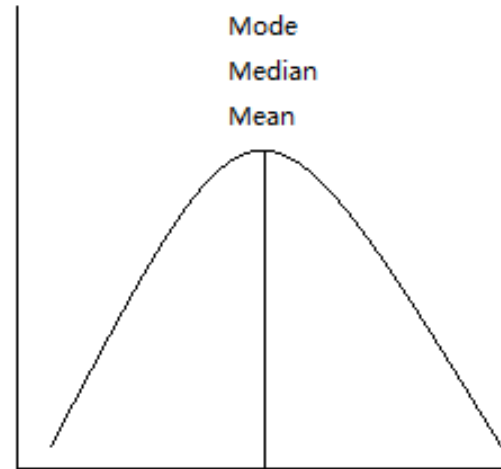
- Unimodal – the values of the 3 central tendency measures are identical (mean=median=mode)
- Symmetric in shape – equal number of data points
- Almost all data points under the curve are represented within 3 standard deviation (sd) above and below the mean as such:
 - 68.2% of data points are within 1 sd \pm mean
 - 95.4% of data points are within 2 sd \pm mean
 - 99.7% of data points are within 3 sd \pm mean

Skewedness

Negative skew

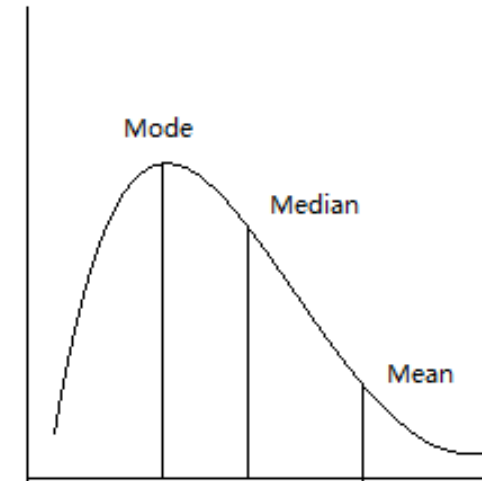


Left skew



Normal Distribution

Positive skew



Right skew

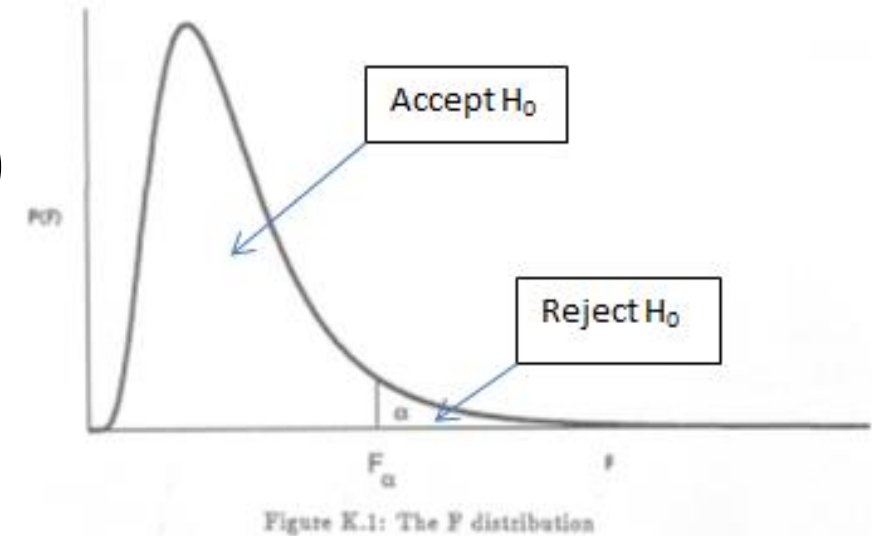
Skewedness is determined by comparing the mean's relative position to the median, either more positive or more negative than the center of the data (median)

Quick test of Normality (Central tendency)

- Check the mean, median and mode:
 - Equal → normal distribution
 - Not equal → non-normal distribution
- Plot data as histogram and visually inspect the shape of the curve
 - Symmetric bell shaped curve → normal distribution
 - Asymmetric, non-bell-shaped curve → non-normal distribution
- Use online resources
 - Shapiro-Wilk online calculator

Testing for equal variance (spread)

- The F-distribution follows a positively skewed curve (unlike normal distribution)
- Variance is a positive value that comes from the squared value of the standard deviation s^2
- The null hypothesis states that all data groups are homogenous (of equal variance) and the alternative hypothesis states that data are heterogeneous (of unequal variance)



What is Chi-Square (χ^2)

- A calculated Chi statistic based on a Chi distribution
- Asks if the observed counts (frequency) in the sample are different from counts expected in the population
- Similar to a one-sample t-test or z-scores
- It is a “goodness-of-fit” test
- In other words, how well does the sample data fit the population (based model) or is it associated with each other

Calculate Expected values

Observed (O)			cigarettes	hookah	none	
	men		11	9	3	23
	women		3	7	8	18
			14	16	11	41
Expected (E)			cigarettes	hookah	none	
	men		=D\$5*\$G3/\$G\$5	8.97561	6.170732	
	women		6.146341463	7.02439	4.829268	

then calculate chi

			cigarettes	hookah	none			chitest
	Obsereved (O)	men	11	9	3	23		C10:E11)
		women	3	7	8	18		
			14	16	11	41		
	expected (E)		cigarettes	hookah	none			
0		men	7.85365854	8.97560976	6.17073171			
1		women	6.14634146	7.02439024	4.82926829			
2								

Exercise to do at Home

Two-sample t-test example

Problem:

Testing sleep deprivation on academic test performance. Two groups of students were evaluated twice using the same quiz, the first group while being sleep deprived and the second group while being well-rested. Are the two quiz scores significantly different from each other at $\alpha = 0.05$?

Sleep deprived	Well-rested
4.7	4.7
4.2	5.3
4.2	4.7
0.7	0.7
2.7	2.7
3.3	3.3
3.3	3.3
1.3	1.3
2.7	2.7
3.6	4.0
3.0	3.3
2.0	2.0
3.0	6.0
4.7	4.7
4.5	4.7
0.6	2.7
0.0	2.7
1.8	2.7
2.7	3.3
2.7	2.7

Problem solving

Frist: state the null and alternative hypotheses, decide one direction, and calculate descriptive analysis

- H_0 : Sleep deprivation does not affect test performance
- H_1 : Sleep deprivation affect test performance
- One-tailed

	sleep deprive	well-rested
	4.7	4.7
	4.2	5.3
	4.2	4.7
	0.7	0.7
	2.7	2.7
	3.3	3.3
	3.3	3.3
	1.3	1.3
	2.7	2.7
	3.6	4.0
	3.0	3.3
	2.0	2.0
	3.0	6.0
	4.7	4.7
	4.5	4.7
	0.6	2.7
	0.0	2.7
	1.8	2.7
	2.7	3.3
	2.7	2.7
Mean	2.8	3.4
sd	1.4	1.3
Median	2.9	3.3
Mode	2.7	2.7
variance	1.9	1.8

Second: test Normality for both groups

- Very complex in excel
- Very easy in SPSS
- Can be done online
- Google Shapiro-Wilk test

Paste data here: (results below)

5.3
4.7
0.7
2.7
3.3
3.3
1.3
2.7
4.0
3.3
2.0
6.0
4.7
4.7
2.7
2.7
2.7
3.3
2.7

Calculate Clear all

Results:

n = 20
Mean = 3.3749999999999999
SD = 1.3400216023396196
W = 0.9545171696544339

Threshold (p=0.01) = 0.8679999709129333 --> HO accepted
Threshold (p=0.05) = 0.9049999713897705 --> HO accepted
Threshold (p=0.10) = 0.9200000166893005 --> HO accepted

--> Your data seems normal

Third: determine if the sample is homogenous (of equal variance)

- In Excel:
 - Perform an F-test using the function (FTEST) and select the two groups
 - If the F-value is bigger than α (0.05) the sample is HOMOgenous (equal variance)
 - If the F-value is smaller than α (0.05) the sample is HETEROgenous (unequal variance)
 - F-value is 0.905
 - Greater than $\alpha \rightarrow$ failure to reject the null \rightarrow groups are homogenous and of equal variance

fx		=FTEST(F3:F22,G3:G22)	
E		FTEST(array1, array2)	
	sleep deprived	well-rested	
	4.7	4.7	
	4.2	5.3	
	4.2	4.7	
	0.7	0.7	
	2.7	2.7	
	3.3	3.3	
	3.3	3.3	
	1.3	1.3	
	2.7	2.7	
	3.6	4.0	
	3.0	3.3	
	2.0	2.0	
	3.0	6.0	
	4.7	4.7	
	4.5	4.7	
	0.6	2.7	
	0.0	2.7	
	1.8	2.7	
	2.7	3.3	
	2.7	2.7	
Mean	2.8	3.4	
sd	1.4	1.3	
Median	2.9	3.3	
Mode	2.7	2.7	
variance	1.9	1.8	
F-Test	0.905284		

Fourth: Determine if the two means are significantly different using T-Test

- Perform an T-test using the function (T-TEST) and select the two groups
- If the T-test value is smaller than α (0.05) the two means are significantly different
- If the T-test value is larger than α (0.05) the two means are not significantly different
- T-test value is 0.1807

larger than $\alpha \rightarrow$ failure to reject the null and accept the alternative \rightarrow the two means are NOT different

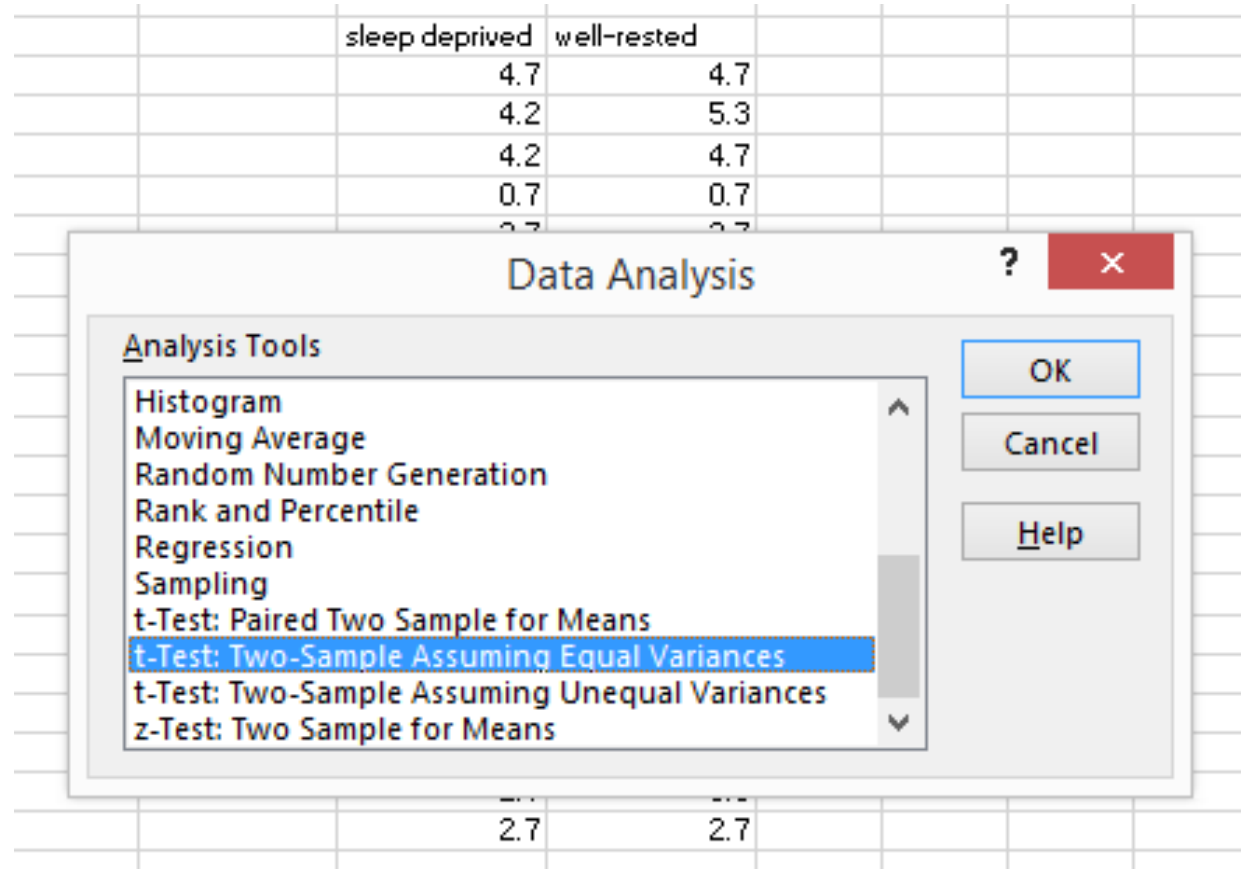
Note: this function does not provide a t-stat

Excel formula bar: `=TTEST(F3:F22,G3:G22,2,2)`

	sleep deprived	well-rested
	4.7	4.7
	4.2	5.3
	4.2	4.7
	0.7	0.7
	2.7	2.7
	3.3	3.3
	3.3	3.3
	1.3	1.3
	2.7	2.7
	3.6	4.0
	3.0	3.3
	2.0	2.0
	3.0	6.0
	4.7	4.7
	4.5	4.7
	0.6	2.7
	0.0	2.7
	1.8	2.7
	2.7	3.3
	2.7	2.7
Mean	2.8	3.4
sd	1.4	1.3
Median	2.9	3.3
Mode	2.7	2.7
variance	1.9	1.8
F-Test	0.905284	
T-TEST	0.1807386	

Using Excel to automate the process

- Once decided on variance and hypothesis direction
- From Data analysis tool pack
 - Select appropriate test
 - Highlight data accordingly



Results displayed in Excel

t-Test: Two-Sample Assuming Equal Variances		
	<i>Variable 1</i>	<i>Variable 2</i>
Mean	2.78	3.366666667
Variance	1.902385965	1.8
Observations	20	20
Pooled Variance	1.851192982	
Hypothesized Mean D	0	
df	38	
t Stat	-1.363532725	
P(T<=t) one-tail	0.090369316	
t Critical one-tail	1.68595446	
P(T<=t) two-tail	0.180738631	
t Critical two-tail	2.024394164	

Note: this method provides a t-stat

Another method to test homogeneity

- The F-test is extremely sensitive and requires that assumption of normality to be very true
- Any slight deviation from normality can cause wrong conclusion
- Levene's test is much more robust test, applicable to routine use and can easily be performed in SPSS and Excel with little extra work.

Levene's test in Excel

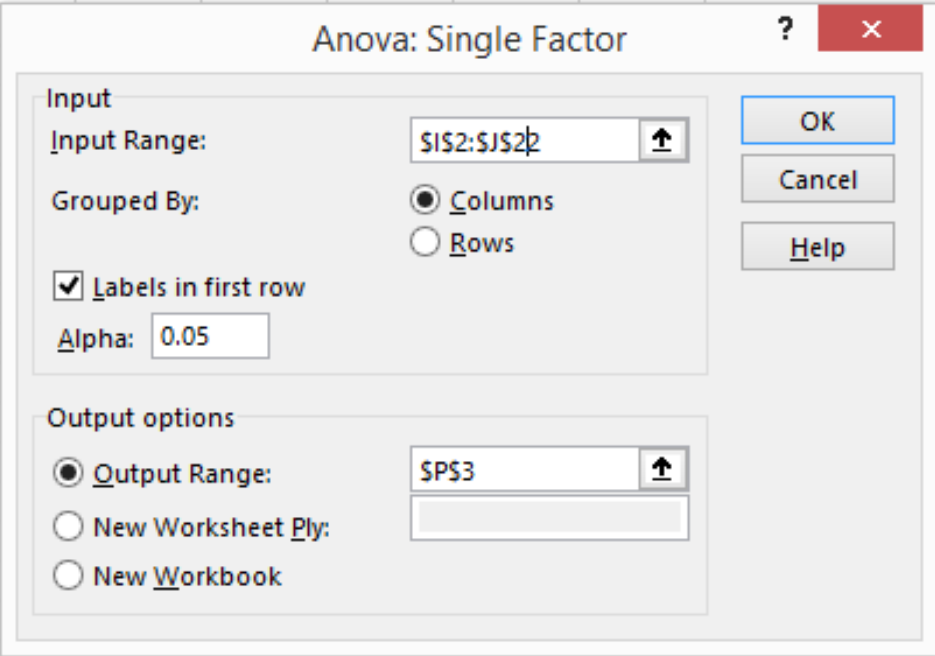
- First: Calculate the ABSOLUTE difference between subject score and the mean of each group independently from each other
- Second: select from analysis Tool pack “ANOVA Single Factor”

=ABS(F3-\$F\$24)					
ABS(number)	F	G	H	I	J
	sleep deprived	well-rested		slp dep diff	well diff
	4.7	4.7		=ABS(F3-\$F\$24)	1.3
	4.2	5.3		1.4	2.0
	4.2	4.7		1.4	1.3
	0.7	0.7		2.1	2.7
	2.7	2.7		0.1	0.7
	3.3	3.3		0.6	0.0
	3.3	3.3		0.6	0.0
	1.3	1.3		1.4	2.0
	2.7	2.7		0.1	0.7
	3.6	4.0		0.8	0.6
	3.0	3.3		0.2	0.0
	2.0	2.0		0.8	1.4
	3.0	6.0		0.2	2.6
	4.7	4.7		1.9	1.3
	4.5	4.7		1.7	1.3
	0.6	2.7		2.2	0.7
	0.0	2.7		2.8	0.7
	1.8	2.7		1.0	0.7
	2.7	3.3		0.1	0.0
	2.7	2.7		0.1	0.7
Mean	2.8	3.4			
sd	1.4	1.2			

In the ANOVA dialog box

- Select the difference of both groups
- Be sure to label each group and check the “labels in the first row” box
- Input desired Alpha value
- Select any empty cell for the output range or a new worksheet

slp dep diff	well diff
1.9	1.3
1.4	2.0
1.4	1.3
2.1	2.7
0.1	0.7
0.6	0.0
0.6	0.0
1.4	2.0
0.1	0.7
0.8	0.6
0.2	0.0
0.8	1.4
0.2	2.6
1.9	1.3
1.7	1.3
2.2	0.7
2.8	0.7
1.0	0.7
0.1	0.0
0.1	0.7



The image shows the 'Anova: Single Factor' dialog box in Microsoft Excel. The dialog box is titled 'Anova: Single Factor' and has a question mark icon and a close button (X) in the top right corner. It contains several sections: 'Input' with 'Input Range' set to '\$I\$2:\$J\$20' and a selection icon; 'Grouped By' with radio buttons for 'Columns' (selected) and 'Rows'; a checked box for 'Labels in first row'; an 'Alpha' field set to '0.05'; and 'Output options' with radio buttons for 'Output Range' (selected, set to '\$P\$3'), 'New Worksheet Ply', and 'New Workbook'. On the right side of the dialog box are three buttons: 'OK', 'Cancel', and 'Help'.

Levene's test output and interpretation in Excel

- If P-Value > alpha then the null hypothesis is **not** rejected and so the two groups have equal variance
- Levene's test can be used to test the variance of more than two groups applicable for later statistical tests

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
slp dep diff	20	21.4	1.07	0.69722807		
well diff	20	20.86666667	1.043333333	0.654163743		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.007111111	1	0.007111111	0.010524129	0.918830084	4.098171731
Within Groups	25.67644444	38	0.675695906			
Total	25.68355556	39				

- **Last:** Report the statistical analysis finding.

A group of students were evaluated using the same quiz while sleep deprived or well-rested (Mean = 2.8, SD = 1.4 vs. Mean = 3.4, SD = 1.3) Results of an independent sample *t*-test indicates a non-significant difference in the two test scores $t(38) = -1.36$, $P = 0.1807$. The assumption of normality and equal variance were met with no evidence of outliers.

Repeated Measures T-Test

- The three Golden Assumptions
 - Continuous data
 - Normal distribution
 - ~~Equal variance~~
- No need for variance test because it is a repeated measure of the same sample (group)
- Select the “paired” option in excel