Data Analysis: Spreadsheets: Typical Statistics Functions

iT Centre Learning

iT services

UNIVERSITY OF OXFORD

The small print

Prerequisites

Time in the workshop is precious – it is an opportunity for you to interact with the workshop leader and other participants through questions and discussions and to share your experiences and concerns. To make the most of this time we sometimes ask you to carry out learning activities ahead of the workshop so that everyone comes into the class with the same basic knowledge. We keep this prior learning to a minimum and often make use of online videos. Online videos provided through 'Molly' can be accessed by University members anytime, anywhere, through a browser or app.

Your course booking will tell you if any prior learning activity is required. If you don't have an environment where you can do this learning, you can come along to one of our 'quiet' sessions. These are scheduled every week in normal term-time, and are a quiet space where you can work through 'Molly' videos or other workshop resources.

If you turn up for a workshop without having done the prior learning, the workshop leader may suggest that you come back on another session.

Copyright

About the workshop designer

Graham Addis started his first technology role in 1978 and has gathered decades of practical experience in industry. He has always been passionate about passing on his knowledge and undertook his first formal teaching position as a Customer Training Specialist for Intel back in 1984. Since that time his career has combined extensive real world experience with teaching and mentoring. In 2017 he joined the academic world at the University of Oxford and currently specialises in teaching spreadsheets, databases and programming.

Revision history

Version	Date	Author	Comments
2.0	May 2020	Graham Addis	Convert to online format.
1.3	November 2019	Graham Addis	Minor edits and small print updates
1.2	August 2016	Duncan Young	Small print edits
1.1	November 2016	Duncan Young	Adapted to new course design
1.0	June 2014	Steven Albury	Initial version

About this workshop

This workshop will give you an insight into some of the techniques that can be used to work with statistics in your research.

We will include pointers to other workshops and further resources that will help you go on later to analyse and organise your data.

What you will learn

This session provides an introduction to statistical functionality in Excel where you will learn how to use various tools to perform a core set of statistical operations including mean, standard deviation, frequency, goodness of fit, t tests, ANOVA, correlation and regression and then display the results in charts such as histograms.

What you need to know

The ideas and techniques covered in this workshop will apply to a range of tools. We will demonstrate using *Excel for Windows*, which is widely available. However, the concepts will be the same, whatever spreadsheet software you decide to use.

I will assume that you are reasonably confident in using the tool you have chosen to use to create your spreadsheets. With your chosen tool, you will need to be able to:

- open and navigate around a workbook using the mouse and scrollbars, save a workbook
- add data to cells, and select and amend such data
- create a formula that calculates using values found in other cells
- Navigate the commands and menus, using Help as necessary

If you need to review these activities, Molly is a great place to get guidance. There is an activity with relevant Molly videos in the IT Learning Portfolio: visit skills.it.ox.ac.uk/it-learning-portfolio and search for “Spreadsheets: Typical statistics functions (Activity)”.

The resources you need

Sample documents that you can use to experiment with will be made available, but you may like to bring along your own.

Unless you have been told otherwise, in classroom workshops there will be a computer available for you to use with *Excel for Windows* installed.

You can use your own computer with your preferred app installed if you want to – just bear in mind that I am not an expert in every app (although I am sure that between us we will be able to sort out most problems!).

Learning Objectives

This workshop has the following learning objectives:

Learning Objective One - ANOVA

Learning Objective Two - Working with Central Tendency

Learning Objective Three - The Shape of Data

Learning Objective Four - Hypothesis Testing

Learning Objective Five - Multi Sample Hypothesis Testing

Learning Objective Six - Correlation and Regression

Learning Objective Seven - Kolmogorov-Smirnov test

Learning Objective One - ANOVA

This section explains the how to access, and run, data analysis add-ins.

Ensure the **Analysis toolPak** add-in is enabled:
File->Options->Add-ins Select **Analysis ToolPak**
Choose **Manage "Excel Add-ins"**

The group **Data->Analysis** should now be available in the **Data** ribbon.

In workbook **CT1 (Student).xlsx** worksheet **ANOVA**: select the **Data Analysis** tool **Data->Analysis-> Data Analysis**. Choose the option **Anova: Single Factor** and perform the Anova analysis on the data in the worksheet.

Learning Objective Two - Working with Central Tendency

This section explains the importance of central tendency when working with research data.

We will look at a recommended general approach and then consider specific tactics that make use of Excel's capabilities to calculate various aspect of central tendency.

In the workbook **CT1 (Student).xlsx** worksheet **Start 1** enter the formula to obtain the indicated results.

Complete examples are available in the worksheet **End 1** for reference.

Learning Objective Three - The Shape of Data

The shape of the data you are working with has a significant bearing on the statistics that you can use with it.

In the workbook **CT1 (Student).xlsx** worksheet **Start 2** enter the formula to obtain the indicated results.

Complete examples are available in the worksheet **End 2** for reference.

Learning Objective Four - Hypothesis Testing

Most research statistics are used to test hypotheses about populations by examining data about samples of those populations

In the workbook **CT1 (Student).xlsx** worksheets **Start 3**, and **CHISQ DICE** enter the formula to obtain the indicated results.

Complete examples are available in the worksheet **End 3** for reference.
Further examples illustrated in **End 3 (Alt)**.

Learning Objective Five - Multi Sample Hypothesis Testing

Real life data finds us often needing to test more than one hypothesis at a time.

In the workbook **CT1 (Student).xlsx** worksheets **CHISQ DICE**, **Start 4**, and **Ox Temps** enter the formula to obtain the indicated results.

Complete examples are available in the worksheet **End 4** for reference.

Learning Objective Six - Correlation and Regression

Real life data sets contain patterns which can be compared and, from which, predictions can be inferred.

In the workbook **CT1 (Student).xlsx** worksheets **Correlation 1** enter the formula to obtain the indicated results.

In the workbook **CT1 (Student).xlsx** worksheets **Correlation 2**, **Regression**, and **Regression 2** use the **Correlation** and **Regression** tools from the **Data Analysis** tool **Data->Analysis-> Data Analysis**.

Learning Objective Seven - Kolmogorov-Smirnov test

Real life data sets contain which follows various distributions, how do we tell if they are ‘normal’, or close to?

In the workbook **CT1 (Student).xlsx** worksheet **KS** there is a worked example of the Kolmogorov-Smirnov test.

Follow the analysis and attempt to replicate the results in a separate area of the worksheet

Further information

Getting extra help

Course Clinics

The IT Learning Centre offers bookable clinics where you can get pre- or post-course advice. Contact us using courses@it.ox.ac.uk.

Study Videos from Molly

Molly is our collection of self-service courses and resources. This includes providing LinkedIn Learning video-based courses free to all members of the University. Visit skills.it.ox.ac.uk/molly and sign in with your Single Sign-On (SSO) credentials.

Some courses recommend pre- and/or post-course activities to support your learning. You can watch these online videos anywhere, anytime, and even download them onto a tablet or smartphone for off-line viewing.

If you need a quiet place to work through learning activities away from distractions, the IT Learning Centre offers 'quiet' sessions where you can book a place. These are scheduled frequently during normal term times.

About the IT Learning Portfolio online

Many of the resources used in the IT Learning Centre courses and workshops are made available as Open Educational Resources (OER) via our Portfolio website at skills.it.ox.ac.uk/it-learning-portfolio.

Find the pre-course activity for this course in the IT Learning Portfolio: visit skills.it.ox.ac.uk/it-learning-portfolio and search for "Spreadsheets: Typical statistics functions (Activity)".

About the IT Learning Centre

The IT Learning Centre delivers over 100 IT-related teacher-led courses, which are provided in our teaching rooms and online, and we give you access to thousands of on-line self-service courses through Molly (powered by LinkedIn Learning).

Our team of teachers have backgrounds in academia, research, business and education and are supported by other experts from around the University and beyond.

Our courses are open to all members of the University at a small charge. Where resources allow, we can deliver closed courses to departments and colleges, which can be more cost-effective than signing up individually. We can also customize courses to suit your needs.

Our fully equipped suite of seven teaching and training rooms are usually available for hire for your own events and courses.

For more information, contact us at courses@it.ox.ac.uk

About IT Customer Services

The IT Learning Centre is part of the Customer Services Group. The group provides the main user support services for the department, assisting all staff and students within the University as well as retired staff and other users of University IT services. It supports all the services offered by IT Services plus general IT support queries from any user, working in collaboration with local IT support units.

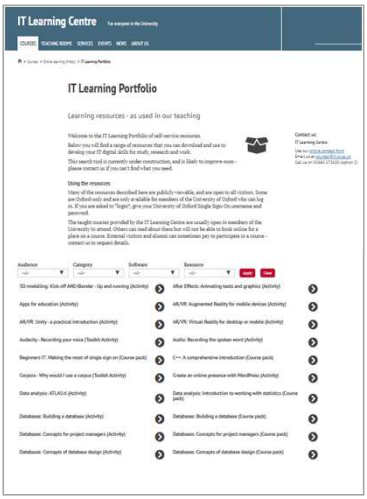
The Customer Services Group also offers a data back-up service; an online shop; and a PC maintenance scheme. Customer Services is further responsible for desktop computing services – for staff and in public/shared areas – throughout UAS and the Bodleian Libraries.

Spreadsheets:
Typical statistics functions

Graham Addis
graham.addis@it.ox.ac.uk

Find the resources for the workshop in our IT Learning Portfolio

Download the files (and more) from the IT Learning Portfolio at
https://skills.it.ox.ac.uk/it-learning-portfolio



Resources for your learning

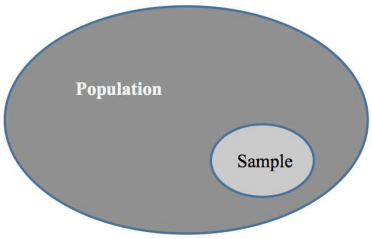
- **Activities** for you to practice today
  In the course handbook
  Work at your own pace!
  Be selective
- **Videos** with today's topics in Molly
- Follow-up work
  Continue with exercises after the session
  Bookable Course Clinics later

Introduction: What are statistics?

This course introduces some of the basic concepts in statistics and shows you how some of these ideas are applied in Excel.

Sample vs Population

"You cannot study everyone everywhere doing everything" (Miles and Huberman 1994)



We infer from sample to population
- We try to get representative samples

Characteristics of the sample are 'statistics'
- Characteristics of the population are 'parameters'

Why Do Statistical Analysis?

Question –> problem -> data -> conclusions.

We don't do statistics for statistics sake but to answer **questions**.
A report produced for the sake of it will go unread and be pointless.

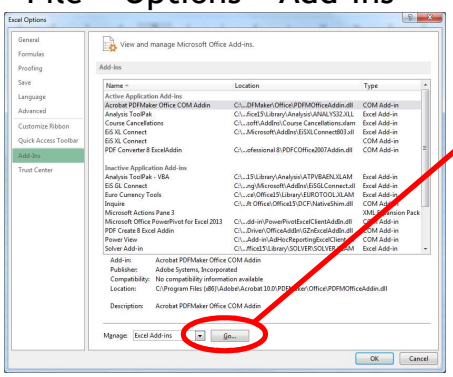
The Fundamental assumptions of statistics:
- Variation/diversity/noise/error is everywhere
- There is never a signal without noise/error
- **Statistics deals in probability not certainty**

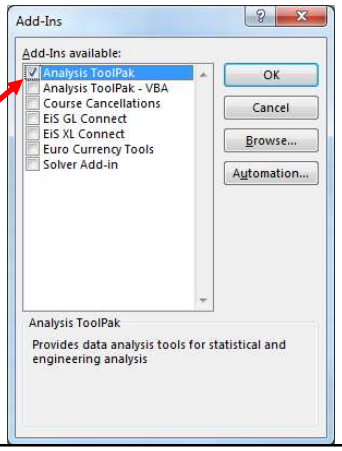
Popular Statistical Tools

Correlation	how much do two sets of data interact
Regression	predict results on one set of data from a related set of data
T Test	is there any difference between these two sets of data or one set and a norm
Chi Squared	is there any difference in variance between these two sets of data
ANOVA	is there a difference between any two of these multiple sets of data
Kolmogorov Smirnov	is my data normally distributed

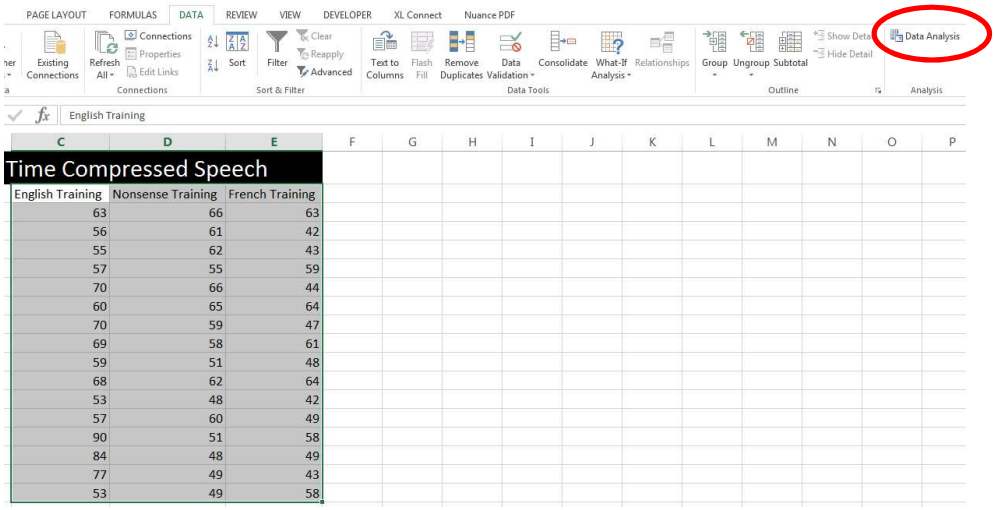
I hear ANOVA is important. How do I run it?

- You'll need the Analysis Toolpak Add-In
- File > Options > Add-Ins



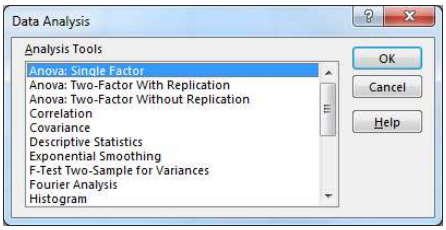


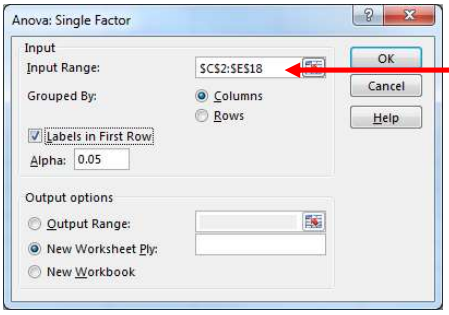
I hear ANOVA is important. How do I run it?



Time Compressed Speech		
English Training	Nonsense Training	French Training
63	66	63
56	61	42
55	62	43
57	55	59
70	66	44
60	65	64
70	59	47
69	58	61
59	51	48
68	62	64
53	48	42
57	60	49
90	51	58
84	48	49
77	49	43
53	49	58

I hear ANOVA is important. How do I run it?





Time Compressed Speech		
English Training	Nonsense Training	French Training
63	66	63
56	61	42
55	62	43
57	55	59
70	66	44
60	65	64
70	59	47
69	58	61
59	51	48
68	62	64
53	48	42
57	60	49
90	51	58
84	48	49
77	49	43
53	49	58

I hear ANOVA is important. How do I run it?

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
English Training	16	1041	65.0625	124.4625
Nonsense Training	16	910	56.875	44.78333
French Training	16	834	52.125	73.05

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1370.542	2	685.2708	8.484721	0.000747	3.204317
Within Groups	3634.438	45	80.76528			
Total	5004.979	47				

If F > F crit

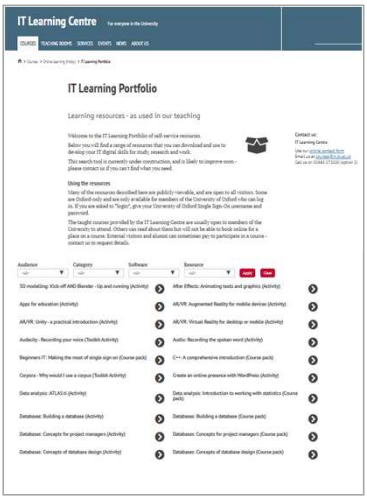
OR

0.05 > P-value…

Result is statistically significant!

Find the resources for the workshop in our IT Learning Portfolio

Download the files (and more) from the IT Learning Portfolio at https://skills.it.ox.ac.uk/it-learning-portfolio



Practical Session 1

Learning Objective	Workbook	Worksheet
One	**CT1 (Student).xlsx**	**ANOVA**

GREAT! What does that mean?

It suggests that the data in at least one pair of columns is more different than can be explained by luck, noise or error

Excellent. Which pair?!

It doesn't tell you.

WHAT?!?!

It's answering a question about general variance

More specific conclusions need further tests

So which tests will I need?

Fundamentally, some version of a t-test

There are different schools of thought on the exact procedure

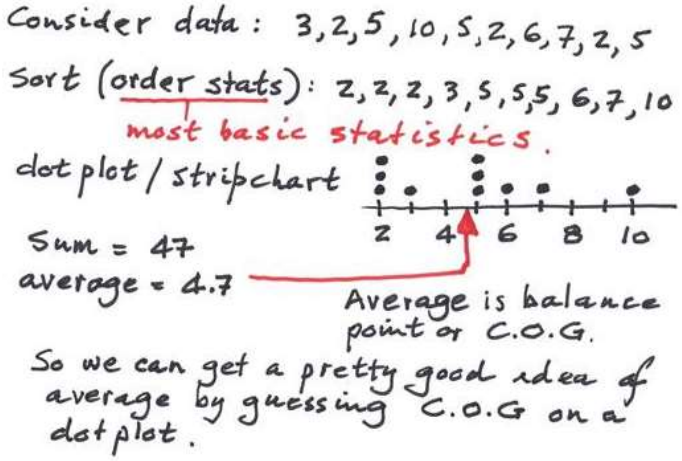
And how do I run a t-test?

You calculate a t-value, then see if it is significant

So how do I calculate a t-value?

(sample mean-expected mean)/(sample SD/SQRT(no. of samples))

Centre of Gravity



AVERAGE()

Used to identify the arithmetic mean, the central ‘balance’ point.

AVERAGEIF[S]() allows one or more conditions to select items to be included in the average.

AVERAGEA() –FALSE (or “”) = 0, TRUE = 1

TRIMMEAN() – exclude a percentage of outliers

MEDIAN(), MODE()

The Median is the data item where exactly half the data lies either side – very useful if data has outlying values at the top or bottom.

The Mode is the most common item (MODE.SNGL) or items (MODE.MULT) in a data set. It is the only ‘average’ you can use for category data (colours, days of the week).

Average, Median & Mode show Central Tendency

Practical Session 2

Learning Objective	Workbook	Worksheet
Two	**CT1 (Student).xlsx**	**Start 1** (Hints in '**End 1**')

Deviation from the Mean

50, 50, 50, 50, 50 – no deviation from the mean
40, 45, 50, 55, 60 – some deviation from the mean
10, 30, 50, 70, 90 – lots of deviation from the mean

Could calculate average deviation from the mean

Negatives cancel out positives – always zero!

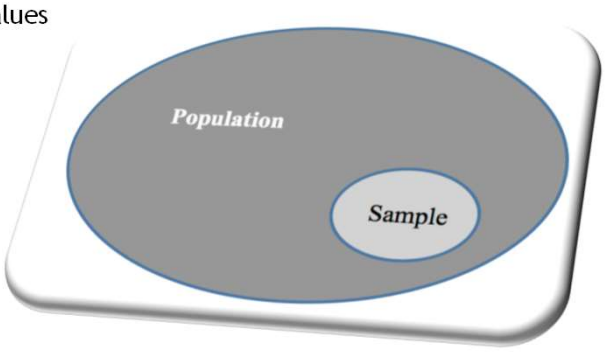
Square the deviations, then average them…

VARIANCE!

Variance

Sample variance cannot be an overestimate of population variance

Unlikely that sample gives all extreme values



Variance

Variance of a population – VAR.P and VARPA

Sample mean can estimate population mean

Sample variance MUST underestimate (or very rarely equal) population variance

Bessel's Correction – Number of samples minus 1

Variance of a sample – VAR.S and VARA

The Standard Deviation

The square root of the variance

Summarises variability of dataset in one number

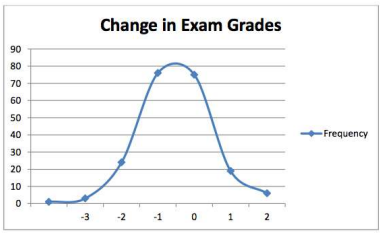
More spread out the scores, larger the SD

With normal distribution:

68% within one SD of mean

95% within two SDs

99% within three



Even outside normal distribution, percentages don't tend to stray far from these

Standard Deviation Functions

STDEV (old), STDEV.S and STDEVA

STDEVP (old) STDEV.P and STDEVPA

No STDEVIF[S]

Would need to filter with IF, then apply STDEV

Practical Session 3

Learning Objective	Workbook	Worksheet
Three	**CT1 (Student).xlsx**	**Start 2** (Hints in '**End 2**')

How do I calculate a t-value?

t value

(sample mean-expected mean)/(sample SD/SQRT(no. of samples))

How do I know if my value of t is significant? T tests

T value	1.63
T.DIST	0.940495

T.DIST() function

=T.DIST(t value, no. of samples -1, TRUE)

So how do I know if my T.DIST result is significant?

You need to examine your **hypothesis**…

Some Terminology

Experiment

- Process resulting in one of at least two distinct results
- Often 'treatment' group v control group
- 'Treatment' is independent variable (a drug, a type of training…)
- What is being measured is dependant variable (performance, growth…)
- Consists of one or more Trials

Trial

- Each time you go through the process of the experiment

Hypothesis

- Predicted answer to a research question

Null Hypothesis (H_0)

"We have observed nothing new or out of the ordinary"

Alternative Hypothesis (H_1)

"We have observed something new and significant"

Process is to reject, or not, the null hypothesis

- We never accept the null hypothesis

Legal Trial

Null Hypothesis – defendant did not commit the crime

Alternative Hypothesis – defendant did commit the crime

Data – testimony and evidence

Do we:

reject the null hypothesis (proven guilty), or
not reject the null hypothesis (not proven guilty)?

Null Hypothesis Errors

Type I error – reject null hypothesis when you think you've found something unusual, but you haven't

- Jury finds someone guilty who was innocent
- Doctor diagnoses an illness patient doesn't have

ALPHA is the probability of a Type I error

- We set this at the beginning of a study
- Typically it is set at 0.05 (5%)

Null Hypothesis Errors

Type II error – not rejecting a null hypothesis when you should

Record companies deciding the Beatles were nothing out of the ordinary

Deciding Leicester City could not win the Premiership

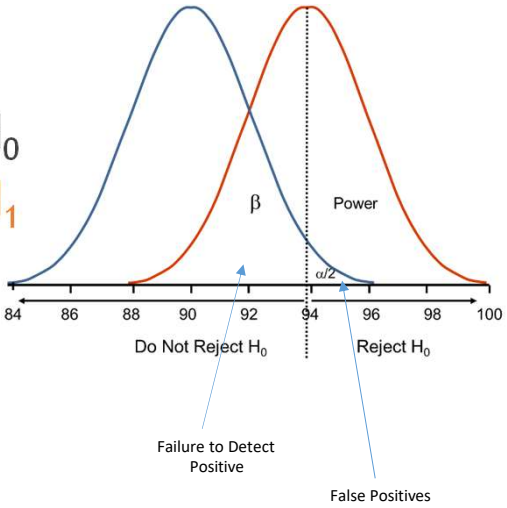
BETA is the probability of a Type II error

More evidence is the best defence

One Sample Hypothesis Testing

Possible sample means for H_0

Possible sample means for H_1



X-axis shows sample means

Is the mean of your sample in H_0 or H_1?

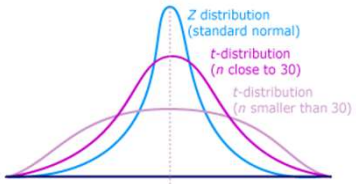
http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Power/BS704_Power_print.html

One Sample Hypothesis Testing

Use Z.TEST() if sure of a standard normal distribution

Otherwise, use T tests

T.DIST() or T.DIST.RT()



http://ci.columbia.edu/ci/premba_test/c0331/s7/s7_4.html

Returns probability of obtaining a t-value at least as high as yours if H_0 is true.

If probability > (1 – Alpha) *[probability<Alpha for RT]* reject H_0

T.INV()

Returns critical t-value for a given probability

Are there one or two tails?

If you can't predict whether H_1 scores are higher or lower than H_0 scores, you are working with 2 tails rather than 1.

The typical 5% of Alpha now becomes 2.5% at each end

Standard normal cutoff 5% value for 1 tail = 1.645

Standard normal cutoff 5% value for 2 tail = 1.96

Harder to reject H_0 with 2 tailed test

T.DIST.2T() and T.INV.2T()

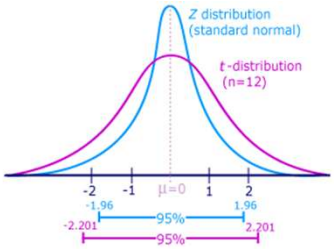
Z Test v T Test

Critical T values are higher than critical Z values

Smaller sample, greater uncertainty

Harder to prove difference is significant

Confidence level	Z score
90%	1.645
95%	1.960
98%	2.326
99%	2.576



Interpreting t Test Results

Our data has 21 samples

This gives 20 Degrees of Freedom (df)

1 Tailed Test

95% confidence required

Alpha = 0.05

T value	1.63
T.DIST	0.940495

	Level of Significance						
2 Tailed	0.40	0.30	0.20	0.15	0.10	0.05	0.02
1 Tailed	0.20	0.15	0.10	0.075	0.05	0.025	0.01
df							
1	1.376	1.963	3.133	4.195	6.320	12.69	31.81
2	1.060	1.385	1.883	2.278	2.912	4.271	6.816
3	0.978	1.250	1.637	1.924	2.352	3.179	4.525
4	0.941	1.190	1.533	1.778	2.132	2.776	3.744
5	0.919	1.156	1.476	1.699	2.015	2.570	3.365
6	0.906	1.134	1.440	1.650	1.943	2.447	3.143
7	0.896	1.119	1.415	1.617	1.895	2.365	2.999
8	0.889	1.108	1.397	1.592	1.860	2.306	2.897
9	0.883	1.100	1.383	1.574	1.833	2.262	2.822
10	0.879	1.093	1.372	1.559	1.813	2.228	2.764
11	0.875	1.088	1.363	1.548	1.796	2.201	2.719
12	0.873	1.083	1.356	1.538	1.782	2.179	2.682
13	0.870	1.079	1.350	1.530	1.771	2.160	2.651
14	0.868	1.076	1.345	1.523	1.761	2.145	2.625
15	0.866	1.074	1.341	1.517	1.753	2.131	2.603
16	0.865	1.071	1.337	1.512	1.746	2.120	2.584
17	0.863	1.069	1.333	1.508	1.740	2.110	2.567
18	0.862	1.067	1.330	1.504	1.734	2.101	2.553
19	0.861	1.066	1.328	1.500	1.729	2.093	2.540
20	0.860	1.064	1.325	1.497	1.725	2.086	2.529

If T value > lookup value…

or T.DIST > (1 – Alpha) then result is significant

Central Limit Theorem

The sampling distribution of the mean is approximately normal for a large sample size

- "Large" means 30 or more
- If population distribution normal, would be normal anyway
- CLT – if sample size large, population distribution doesn't matter

SD of the sampling distribution of the mean equals SD of the population / SQRT (sample size)

Confidence

Statistics are about probability

- Want to establish upper & lower boundaries for population on this variable
- 95% confidence has become the standard
  - 'Alpha' = 0.05
  - 2 SDs of the mean for normal distribution

CONFIDENCE[.NORM() or .T()]

- Needs Alpha, SD and sample size
- Returns distance from mean of the boundary of Alpha

One Sample Hypothesis Testing

Chi-Square tests – hypothesis testing for variances

Does the process vary more than we think it does/should?

CHISQ.DIST() and CHISQ.DIST.RT()

CHISQ.INV() and CHISQ.INV.RT()

Practical Session 4

Learning Objective	Workbook	Worksheet
Four	**CT1 (Student).xlsx**	**Start 3** (Hints in '**End 3**' and '**End 3 (Alt)**')
Four	**CT1 (Student).xlsx**	**CHISQ DICE**

Two Sample Hypothesis Testing

Comparing samples from different populations

Are differences due to chance (H_0) or not (H_1)?

Sample sizes between populations don't have to be equal

Sample sizes within populations do

Central Limit Theorem still applies

Z-Test can still be used for standard normal

Again, real life rarely that neat

T.TEST() and t-Test tool

Two Sample Hypothesis Testing

Testing two variances, you divide one by the other rather than subtract – F test (*R. Fisher*)

F-ratio determines whether use equal or unequal variances version of the t test

F.TEST()

F.DIST() and F.DIST.RT()

F.INV() and F.INV.RT()

F Test Tool

If F > F Critical one-tail…

Reject H_0
Use Unequal Variances

Otherwise, use Equal Variances

First Variance MUST BE greater than second – switch if necessary…

F-Test Two-Sample for Variances

	English Training	*Nonsense Training*
Mean	65.0625	56.875
Variance	124.4625	44.78333333
Observations	16	16
df	15	15
F	2.779214738	
P(F<=f) one-tail	0.028236997	
F Critical one-tail	2.403447071	

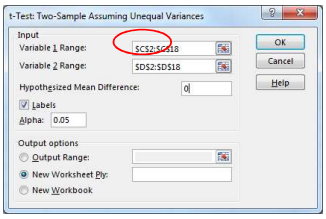
t Test Tool

Hypothesized (H_0) mean difference

Typically zero

If P(T<=t) *[one/two]* tail < Alpha…

Reject H_0



t-Test: Two-Sample Assuming Unequal Variances

	English Training	*Nonsense Training*
Mean	65.0625	56.875
Variance	124.4625	44.78333333
Observations	16	16
Hypothesized Mean Difference	0	
df	25	
t Stat	2.517400485	
P(T<=t) one-tail	0.009300496	
t Critical one-tail	1.708140761	
P(T<=t) two-tail	0.018600992	
t Critical two-tail	2.059538553	

Three or More Hypothesis Testing

More populations > higher alpha > more errors

3 populations =.14, 4=.26, 5=.40, 6=.54, 7=.66…

Need to consider variances rather than means

Analysis of Variance – ANOVA

ANOVA tool (as seen earlier)

RANK(),LARGE(),SMALL(),MAX(),MIN()

These are ordering statistics – they tell us something about the range of the data.

RANK() gives the position of a number in a set of numbers: .EQ and .AVG deal with ties.

LARGE() and SMALL() are useful for example to find the 2nd/3rd/11th, etc. item in terms of size

MIN() and MAX() identify the lower and upper bounds of a range

MINA() / MAXA() - FALSE (or text) = 0, TRUE = 1

PERCENTILE, QUARTILE, PERCENTRANK

PERCENTILE() – the score at a given percentile
QUARTILE() – the score at five preset percentiles
    0=lowest, 1=25%, 2=50%, 3=75%, 4=100%
PERCENTRANK() – the percentile of a given score

.EXC – greater than
.INC – greater than or equal to

FREQUENCY

Specify intervals called ‘bins’
    5, 10, 15, 20…
    20, 40, 60, 80…
Show how many items fall into each ‘bin’

FREQUENCY()
    Array function (Ctrl+Shift+Enter)
Histogram tool

Practical Session 5

Learning Objective	Workbook	Worksheet
Five	**CT1 (Student).xlsx**	**Start 4** (Hints in '**End 4**')
Five	**CT1 (Student).xlsx**	**Ox Temps**

Correlation

Two things vary together

Relationship doesn't mean causality

Correlation can be negative

One gets higher as the other gets lower

r^2 - co-efficient of determination – RSQ()

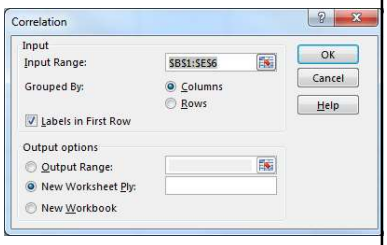
CORREL() and PEARSON()

Use TDIST and FISHER() to test correlation hypotheses

Correlation and Covariance tools

Correlation Tool

Measurement	Set 1	Set 2	Set 3	Set 4
A	11	10	11	-10
B	7	8	3	-8
C	9	10	6	-7
D	4	4	4	-4
E	12	11	12	-11



	Set 1	*Set 2*	*Set 3*	*Set 4*
Set 1	1			
Set 2	0.95389	1		
Set 3	0.88447	0.70971	1	
Set 4	-0.93865	-0.88252	-0.80418	1

Closer to 1 or -1, stronger the correlation

Regression

Slightly counter-intuitively, regression is about prediction

Use data on one variable to predict value of another

Line shows relationship between independent (x) and dependant (y) variable

Regression co-efficients

Where does the line intercept (a) the y axis?

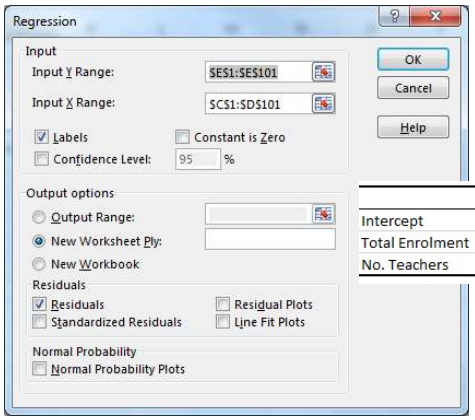
What is the slope (b) of the line?

$y = a + bx$

Regression

Variance -> *Residual* Variance
Standard Deviation - > Standard *Error of Estimate*
SLOPE(), INTERCEPT()
STEYX() – standard error of estimate
FORECAST[.LINEAR]
TREND() – predicts y for given x
LINEST()
Regression tool

Regression Tool



	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	628.9887569	0.984598555	638.8276256	1.3762E-177
Total Enrolment	-0.00086682	0.002108563	-0.411095221	0.681909526
No. Teachers	0.020881825	0.043864294	0.476055197	0.635105544

Practical Session 6

Learning Objective	Workbook	Worksheet
Six	**CT1 (Student).xlsx**	**Correlation 1**
Six	**CT1 (Student).xlsx**	**Correlation 2**
Six	**CT1 (Student).xlsx**	**Regression**
Six	**CT1 (Student).xlsx**	**Regression 2**

Standard Scores (z-scores)

Converting two sets of scores to the same scale

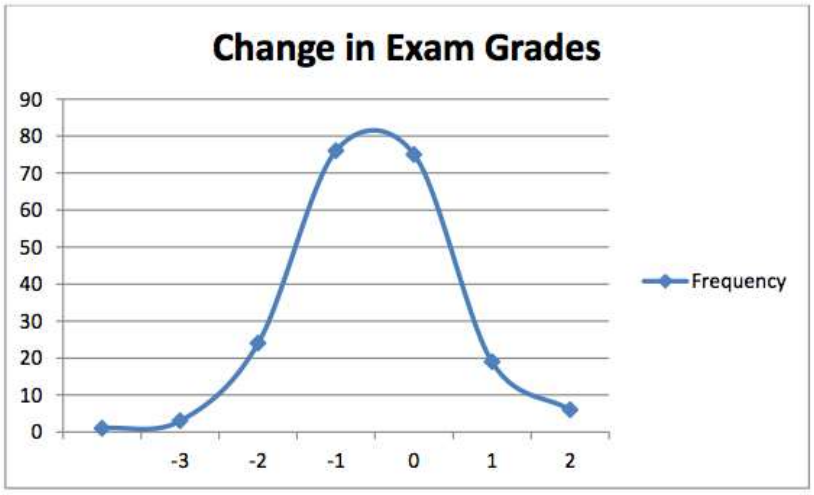
- Use the mean of a set as its zero point
- Use the SD as its unit of measure
- Divide a score's deviation by the SD

STANDARDIZE()

Uses: assigning exam grades, comparing sport eras, IQ scores (convert to T score to get only positives)

The Shape of Data

The way data is distributed can tell us a lot about what it means – the basis is the normal (Gaussian) distribution



Normal Distribution

'Area under the curve' – what percentage of scores are between 60 and 70?

NORM.DIST()

Needs mean, SD and a score

CUMULATIVE = TRUE -> percent from 0 to score

CUMULATIVE = FALSE -> percent at score

Use NORM.DIST(,,,FALSE) for both 60 and 70

Subtract 60 score from 70 score

Normal Distribution

NORM.INV()
- Needs mean, SD and a cumulative probability
- Result is the score at that point

NORM.S.DIST() and NORM.S.INV()
- Mean is 0, SD is 1

PHI()
- Height of std normal distribution at that point

Is the data normally distributed?

Kolmogorov – Smirnov test
- Start with mid point FREQUENCY table
- Column for cumulative frequency at each bin
- Column dividing those by sample size
- Column using STANDARDIZE to make z-scores
- Column applying NORM.S.DIST to z-scores
- Column subtracting red col from blue col (ABS)
  - Result is 0 if normal distribution

Is the data normally distributed?

Kolmogorov-Smirnov test continued

Consult critical value in K-S table

http://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/

Assume you will use the 0.05 column (95% confidence)

Find critical value for number of values you have

Compare with largest value in blue-red result column

If critical value > largest blue-red result, data is good fit for normal distribution

Practical Session 7

Learning Objective	Workbook	Worksheet
Seven	**CT1 (Student).xlsx**	**KS**

Stats Functions

A list of all the statistical functions in Excel can be found at:

https://support.office.com/en-gb/article/Statistical-functions-reference-624dac86-a375-4435-bc25-76d659719ffd

Excel for Mac 2011 DOES NOT include the Analysis Toolpak

Microsoft recommends downloading 3rd party tools

https://support.office.com/en-gb/article/I-can-t-find-the-Analysis-ToolPak-d678dc08-bdc2-4eda-8b94-08755fa4b55a

Excel for Mac 2016 DOES include the Analysis Toolpak

Further courses

- Statistical concepts for researchers
- Good practice with Pivot Tables
- Good practice in spreadsheet design
- Dealing with that difficult spreadsheet

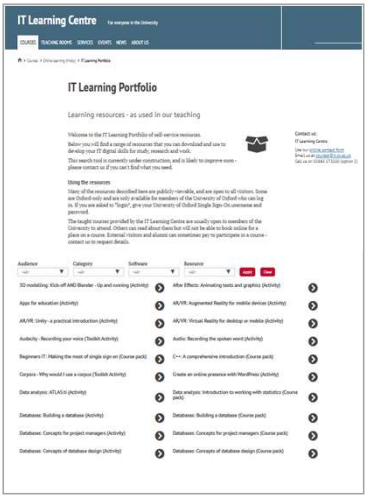
iT Centre Learning

iT services

UNIVERSITY OF OXFORD

Find the resources for the workshop in our IT Learning Portfolio

Download the files (and more) from the IT Learning Portfolio at https://skills.it.ox.ac.uk/it-learning-portfolio



Graham.Addis@it.ox.ac.uk