UNIVERSITY OF BIRMINGHAM

School of Computer Science

EVALUATION METHODS AND STATISTICS

Prof. Chris Baber

Chair of Pervasive and Ubiquitous Computing

Inferential Statistics

- descriptive statistics summarises our data (central tendency, dispersion)
- To draw an inference from a set of data, we use inferential statistics
- Inferential statistics requires:
  - an estimate of a population parameter
  - a test of an hypothesis
- We are asking can we be confident that a statistic (data value) is near a given parameter, and what is the **probability** that the parameter is within a **range** that includes the statistic
  - Confidence Interval = range
  - Confidence Level = probability

So, I might claim that, in my lectures, I talk for between 20 and 25 minutes [confidence interval] (before asking you to do something) 95% of the time [confidence level]

Point Estimates

- Values that can be used to estimate the parameters of a sample, n, of a population, N
  - Proportion
  - Margin of error
  - Mean
  - Variance

Population Proportion

- What proportion of University students enjoy attending lectures?
  - How big is N?
    - All students in this module, all students in Computer Science, all postgraduate students at Birmingham, all students in the UK…?
    - Lets assume that N = 60 (students in this module)
  - How big is n?
    - How many students provide data?
    - Do we select these students at random?
    - What would that require?
    - Assume we have 37 students who admit that they enjoy attending lectures
    - So, n = 37

Population Proportion

P = n / N

= 37 / 60

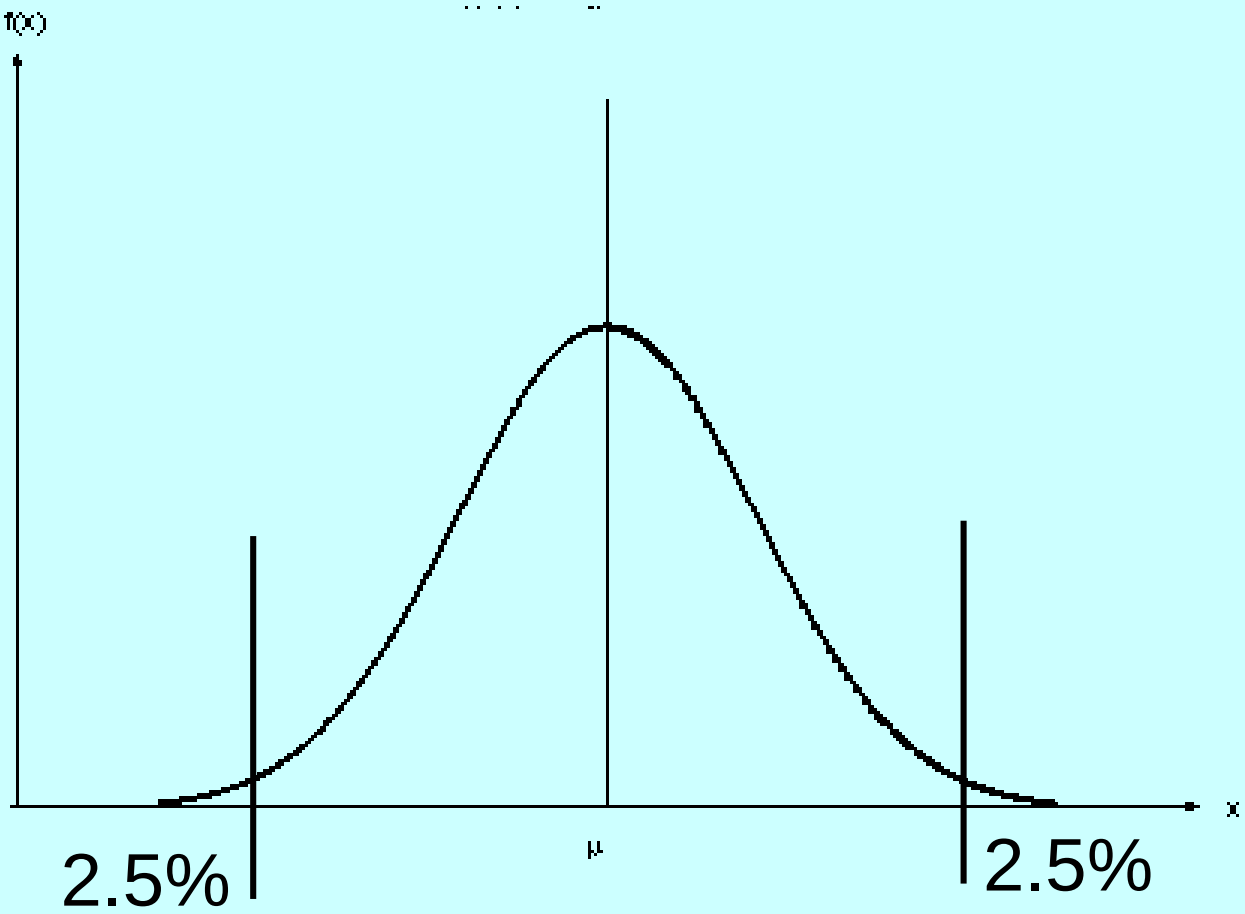
= 0.62

- Can we be confident that this value of 62% is a fair reflection?
- What is our margin of error?

Margin of Error

- To calculate a margin of error, we calculate upper and lower bounds of our estimate
- To do this we assume that our data follow a normal distribution…
- …and we assume that we can be 95% confident in the accuracy of our estimate
- This means that we accept that we will cut off the top and bottom of the normal distribution

95% confidence interval of the Normal Distribution



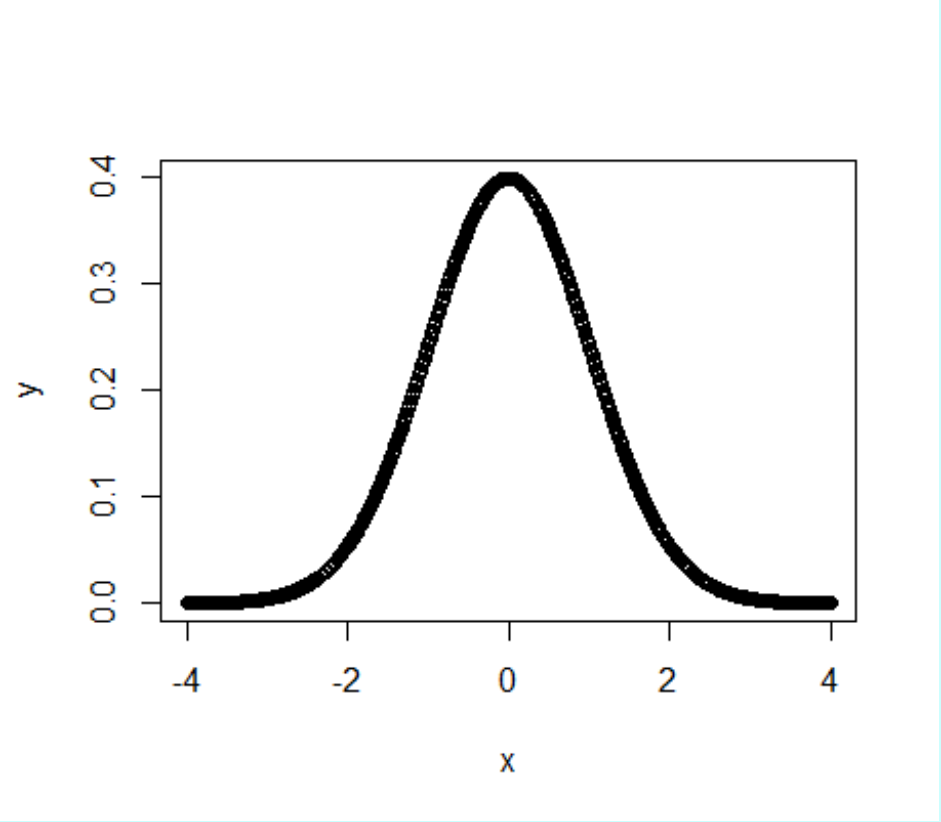
UNIVERSITY OF BIRMINGHAM

The Normal Distribution (again)

- Normal Distributions have the probability density:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{(2\sigma^2)}}$$

- Rather than calculate this each time, we assume that all normal distributions can approximate this Standardized Normal Distribution which assume $\mu = 0$, $\sigma = 1$…



UNIVERSITY OF BIRMINGHAM

Standardized Normal Variable

- The z (from 'standardized') variable allows us to determine the probability that a value, x, falls within a normal distribution

$$z = \frac{x - \mu}{\sigma}$$

- Assume a data set with mean 918 and sd 180.17. What is the probability of obtaining a x<750?

First, sketch the distribution and mark where you think x lies…

Second, z = (750-918)/180.17 = -0.93

So…x is around 1 sd below the mean (we can check this as 918-180.17 = 737.83)

Z tables

- Rather than calculate z each time, we can use Normal Distribution tables to describe the 'standardized normal distribution' of $\mu = 0$, $\sigma = 1$
- To use this z-table when confidence interval, α, is 95%
  - Confidence level of 95% excludes 5% of distribution
  - Assuming a two-tailed test means 2.5% for each tail of the distribution
  - So, area under half of distribution is 50%-2.5% = 47.5%
  - In the z-table, 0.475 is z = 1.9 plus 0.06 = 1.96

The normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4

Confidence Level	α	$\alpha/2$	Z
85%	15%	7.5%	1.435*
90%	10%	5%	1.645*
95%	5%	2.5%	1.96

UNIVERSITY OF BIRMINGHAM

Cumulative Z table (same procedure)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Assume 95% confidence interval.
This is 5% off each side of the distribution, so 2.5% of one side. This is 97.5.
We look up .975 on the z-table.
This is a z-score of 1.96

(for 90% confidence, this would be .95 and a z-score of 1.64)

UNIVERSITY OF BIRMINGHAM

Back to Margin of Error…

- Upper and Lower bounds are defined by the expected standard deviation (which I will explain later) around the central measure
- Upper bound= $\mu + z\sigma x$
  - Standard deviation, σ, is defined by SQRT(x-mean)/n)

    σ = SQRT(0.62*(1-62))/60)

    UB = 0.1 + 1.96*0.063

    =0.74
- Lower bound = $\mu - z\sigma x$ = 0.49
- Margin of Error = 0.74 – 0.62 = 0.25 (or 0.67 – 0.49 = 0.25)
- So, we can be 95% confident that between 49% and 74% of this group of students enjoy attending lectures
- BUT…think about validity and generalisability of this claim based on these data

UNIVERSITY OF BIRMINGHAM

The Expected Mean, Variance and Standard Deviation

The Expected **mean** is the weighted average of all possible values of a variable, x.

- If you throw a (fair) die, the probability of throwing any value (1 to 6) is 1/6
- From this, we can calculate the Expected mean from a set of throws as:

$E(x) = 1xp(x=1)+2xp(x=2)\ldots+6xp(x=6)$

$= 1/6+2/6+3/6+4/6+5/6+6/6 = 21/6$

$=7/2 = 3.5$

The **Variance** is the spread of possible value of a variable, x.

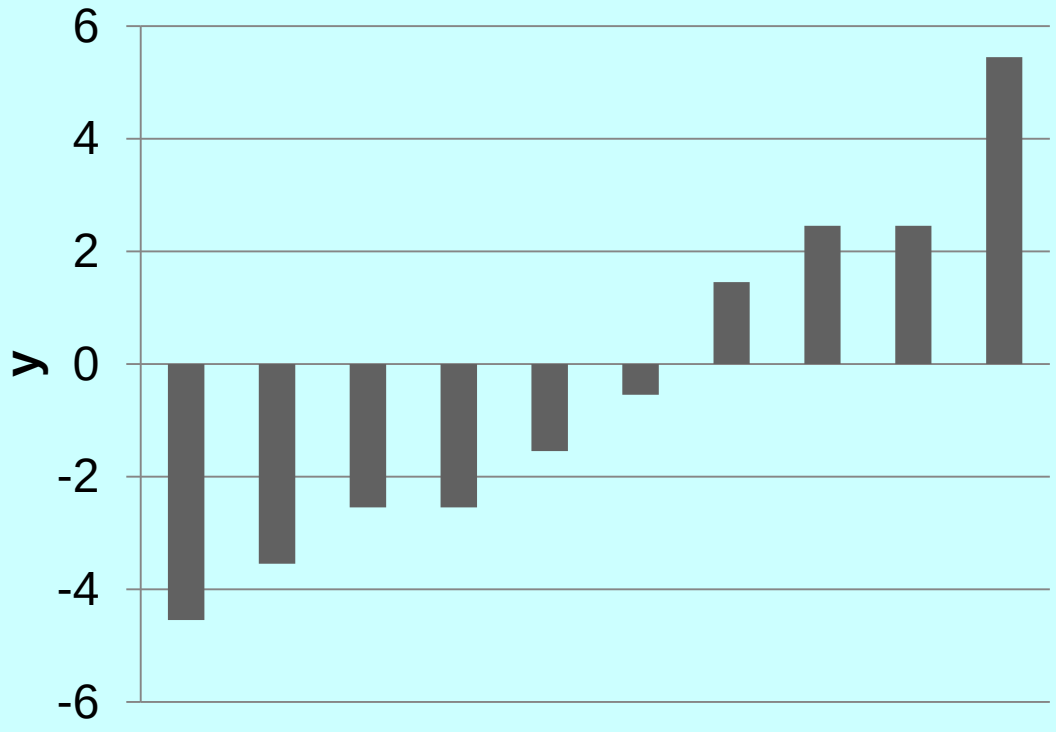
$Var(x) = E(x^2)- Ex^2$

- The Standard Deviation, σ, in the square root of Variance

	1	**2**	**3**	**4**	**5**	**6**	
E*(x)	0.167	0.333	0.5	0.667	.833	1.0	Sum, Σ = 3.5
E*(x^2)	0.167	0.667	1.5	2.667	4.167	6.0	Sum, Σ = 15.167
							Var(x) = 2.91
							Sd, σ = 1.71

UNIVERSITY OF BIRMINGHAM

Understanding variation



Y =
3.454
-2.545
-4.545
2.454
-1.545
5.454
-3.545
1.454
-0.545
-2.545
2.454
μ (mean) y = 0
Σ (sum) y = 0

UNIVERSITY OF BIRMINGHAM

Sum of squares

y	$(y-\mu y)$	$(y-\mu y)^2$
3.454	3.454	11.930
-2.545	-2.545	6.477
-4.545	-4.545	20.657
2.454	2.454	6.022
-1.545	-1.545	2.387
5.454	5.454	29.746
-3.545	-3.545	12.567
1.454	1.454	2.114
-0.545	-0.545	0.297
-2.545	-2.545	6.477
2.454	2.454	6.022
$\mu y = 0$		
$\Sigma y = 0$	$\Sigma(y-\mu y)= 0$	**$\Sigma (y-\mu y)^2 = 104.697$**

UNIVERSITY OF BIRMINGHAM

Degrees of Freedom

- More variability in the data results in larger sum of squares
- More data also results in larger sum of squares
- So, we could take an average (by dividing by the number of samples)
- But, simply dividing by n can underestimate population variance
- So (in statistics) d.f. is the number (N) of data points minus the number of parameter used in the calculation
- Unless otherwise stated, the number of parameters is 1
- So, d.f. = N – 1

Sample Variance

-

Sample Variance (s^2) = $\frac{\text{sum of squares}}{\text{degrees of freedom}}$

$$s^2 = \frac{\sum(x - \bar{x})^2}{N - 1}$$

s^2 = 104.697 / (11-1) = 10.4697

Sd = 3.235692

So, we could describe our data as 0 (± 3.24)

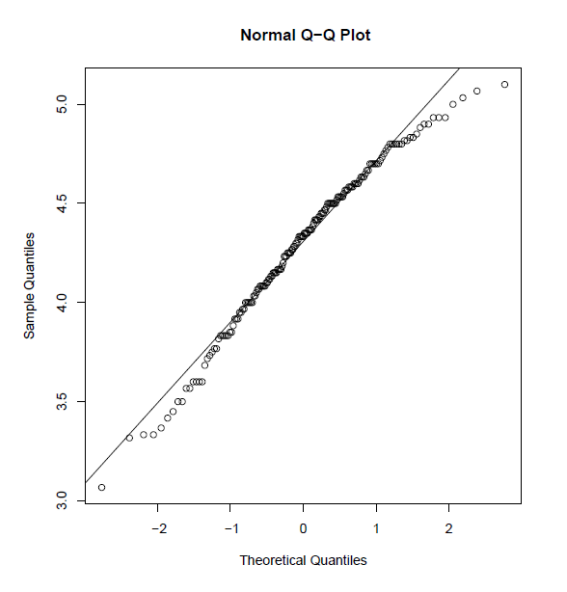
UNIVERSITY OF BIRMINGHAM

Testing for Normality

- We can plot the data in a histogram and visually check if this looks like a ‘bell curve’
- We can convert the data to a Q-Q plot
- We can apply a Shapiro-Wilk test

Q-Q plot

- Quantile-Quantile plot: sorts all data from experiment and plots one sample against the other. R as qqplot() to do this
- More usefully, you want to check thet relationship between samples follows a normal distribution. So, would apply the qqnorm() command in R.
- This plots samples to a hypothetical line. The closer the points are to the line, the more likely the data are normally distributed



UNIVERSITY OF BIRMINGHAM

Shapiro-Wilk test

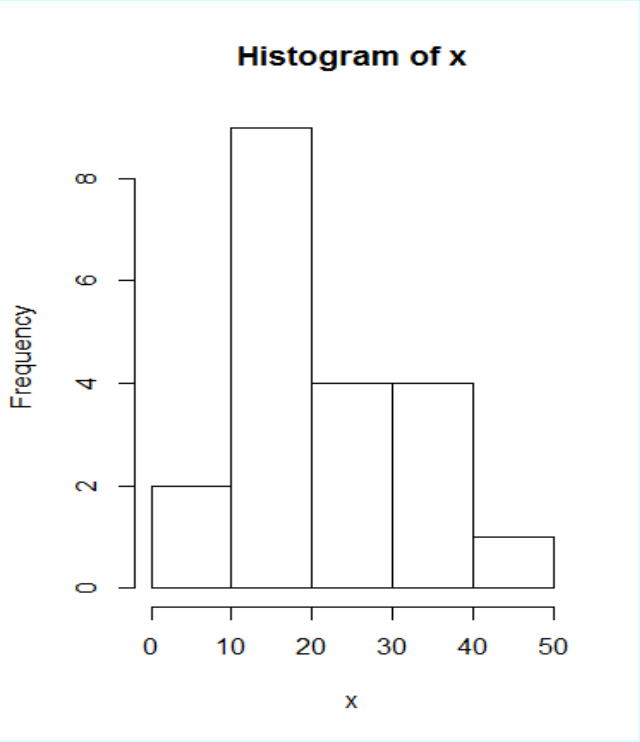
- Checks for correlation between samples and assumed normal distribution
- Compares the slope of the observed data against expected values normalised to the sum of square
- Values > 0.5 show normality

If the data are not normally distributed…

- Check for outliers
  - Calculate mean and standard deviation
  - Apply a cut-off of mean + 2sd
  - You should define these criteria prior to collecting data
- Remove outliers
  - Winsorization (next slide)
- Transformation

Trimming and Winsorization

- Suppose your data has many samples in the tails of the distribution…
- …this could lead to a skewed distribution.
- If you trimmed (i.e., removed samples from) the tails, this could make the distribution normal.
- Trimming
  - Remove a fixed percentage (say, 5%) of samples.
  - E.g., 20 samples, 5% trimming would 1 extreme sample from *each* tail
- Winsorization
  - Trim (as above) and then replace these with most extreme values remaining in each tail



3	**7**	**12**	**15**	**17**	**17**	**18**	**19**	**19**	**19**	**Mean = 22.55**	**Mean.wins. = 22.05**
20	**22**	**24**	**26**	**30**	**32**	**32**	**33**	**36**	**50**		

UNIVERSITY OF BIRMINGHAM

Z Transforming Data

- We can normalise the data by converting each value to its z score (see previous lectures)

Log Transforming Data

- This example is from Howell (2017, p.339)…

Note…

a. There is good correlation between means (r=.88)
b. but we do not have equal variance across levels of independent variable

Table 11.6 Original and transformed data from Conti and Musty (1984)

(a) Original Data

	Control	0.1 μg	0.5 μg	1 μg	2 μg	
	130	93	510	229	144	
	94	444	416	475	111	
	225	403	154	348	217	
	105	192	636	276	200	
	92	67	396	167	84	
	190	170	451	151	99	
	32	77	376	107	44	
	64	353	192	235	84	
	69	365	384		284	
	93	422			293	
Mean	109.40	258.60	390.56	248.50	156.00	
						$r = .88$
S.D.	58.50	153.32	147.68	118.74	87.65	
Variance	3421.82	23,506.04	21,806.78	14,098.86	7682.22	

Log Transforming Data

- To manage unequal variance, we could transform the data, e.g., through $\log_{10}$

Note…

a. There is modest correlation between means (r=.33)
b. but we have similar variance across levels of independent variable

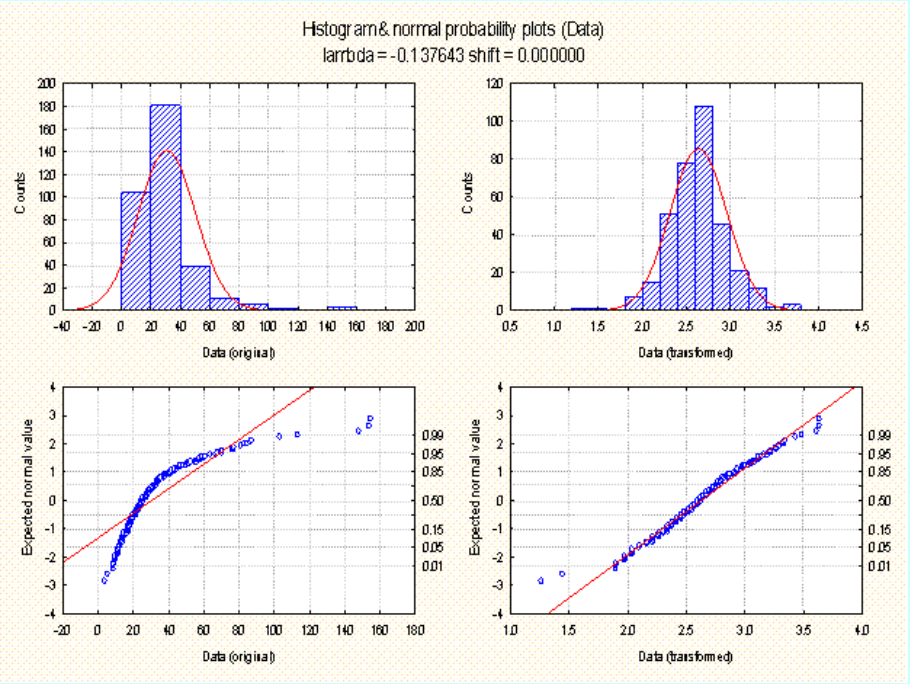
(b) Log Data

	Control	0.1 μg	0.5 μg	1 μg	2 μg	
	2.11	1.97	2.71	2.36	2.16	
	1.97	2.65	2.62	2.68	2.04	
	2.35	2.60	2.19	2.54	2.34	
	2.02	2.28	2.80	2.44	2.30	
	1.96	1.83	2.60	2.22	1.92	
	2.28	2.23	2.65	2.18	2.00	
	1.50	1.89	2.58	2.03	1.64	
	1.81	2.55	2.28	2.37	1.92	
	1.84	2.56	2.58		2.45	
	1.97	2.62			2.47	
Mean	1.981	2.318	2.557	2.353	2.124	
						$r = .33$
S.D.	0.241	0.324	0.197	0.208	0.268	
Variance	0.058	0.105	0.039	0.043	0.072	

UNIVERSITY OF BIRMINGHAM

Box-Cox transformation

- Named after George Box and Sir David Roxbee Cox in 1964
- Transforms non-normal data into a normal distribution by applying a power transform, lambda (-5 to 5, and most optimal selected)
- Examples of formula…
  - -lambda Y^-l = 1/y^3
  - +l = y^l



https://documentation.statsoft.com/STATISTICAHelp.aspx?path=Spreadsheets/Spreadsheet/UsingSpreadsheets/BoxCoxTransformations/BoxCoxTransformationOverviewandTechnicalNotes

UNIVERSITY OF BIRMINGHAM

If none of these approaches help…

- Reconsider the measurement scale you have used (and the quality of the data you have collected)
- Would it be appropriate to continue with statistical analysis?
- Apply non-parametric tests
  - Pairwise
    - Independent – Mann-Whitney
    - Paired - Wilcoxon
  - Factorial
    - One-way ANOVA – Kruskal-Wallis
    - Repeated measures ANOVA - Friedman

Defining Sample size, or how many Participants to use in a test?

- Central Limit Theorem suggests that a normal distribution appears after around 30 samples…but this is a crude (and wishful() way of defining sample size.
- Assuming that our data follows a normal distribution, we can set some limits to the level of confidence we have in avoiding type I errors…
  - if we set α to 0.05% (95% confidence)
  - we can assume that this will exclude the tails of the distribution
  - either side 0, we exclude 0.5 (of the distribution) - $\alpha/2 = 0.5 - 0.02=0.475$
  - 0.475 in the z tables gives z = 1.96
  - we can use this to help calculate sample size to produce this level of confidence from a **known distribution**
  - But we should also consider the **Power** that we will accept

Sample Size calculation

- $$n = \frac{2(za + z1 - \beta)^2 . \sigma^2}{\Delta^2}$$

za		
α-error	5%	1%
2-sided	1.96	2.5758
1-sided	1.65	2.33

$z1 - \beta$				
Power	80%	85%	90%	95%
	0.8416	1.0363	1.2816	1.6449

Δ = estimated effect size, i.e., difference between conditions
σ = standard deviation (of data set)

UNIVERSITY OF BIRMINGHAM

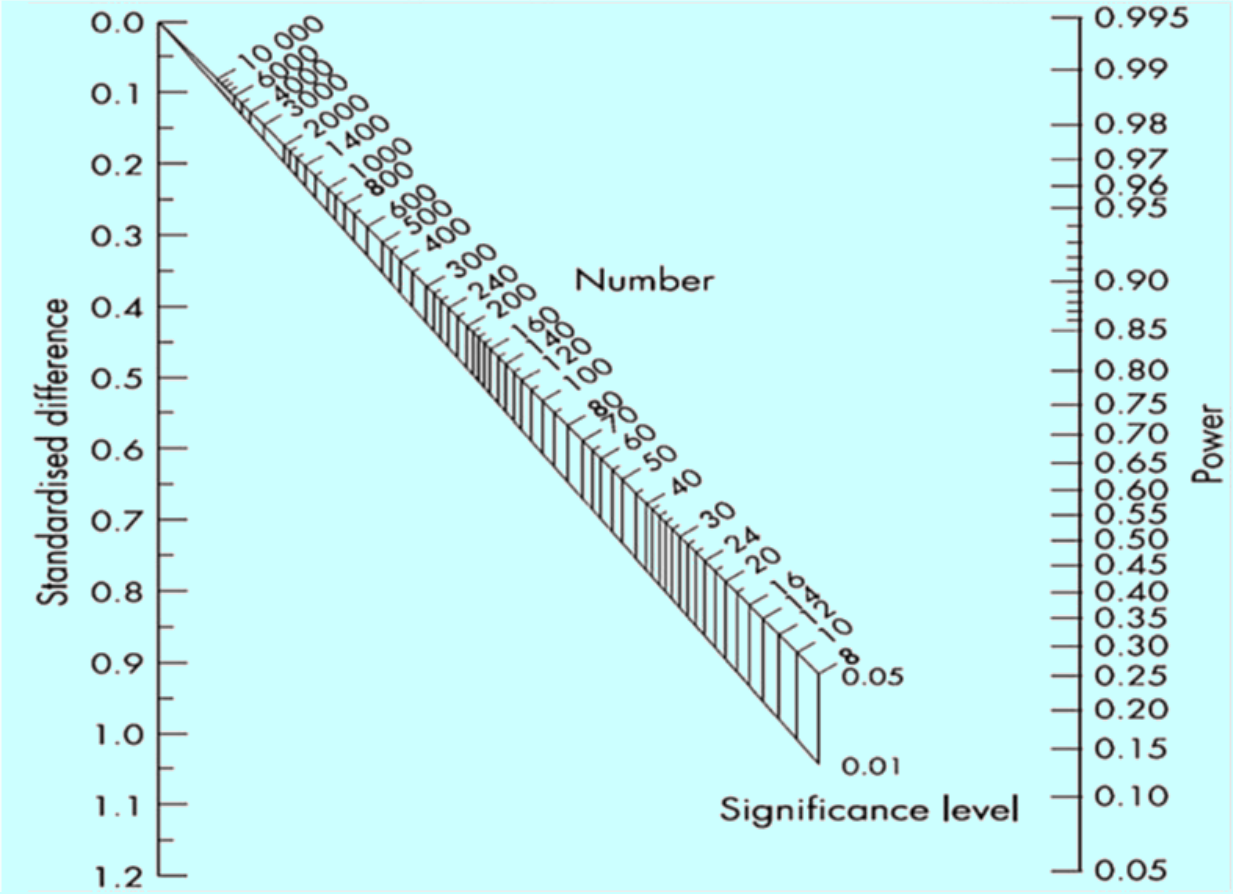
Example

Parameter	A	B	C	D
Power	80%	90%		
Standard deviation	0.6		0.7	
Effect size	60%			30%
Confidence interval	95%			
Number of Participants	16	21	29	114

$$nA = \frac{2(1.96 + 0.8416)^2(0.6)^2}{(0.6)^2}$$

UNIVERSITY OF BIRMINGHAM

Altman, D.G. (1991) Statistics for Medical Research: nomogram for sample size estimation



Standardized Difference = difference between conditions / standard deviation

UNIVERSITY OF BIRMINGHAM

Look-up Tables (Cohen, 1988)

two-tailed α = .05 or one-tailed α = .025

	d										
Power	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**1.0**	**1.20**	**1.40**
.25	332	84	38	22	14	10	8	6	5	4	3
.50	769	193	86	49	32	22	17	13	9	7	5
.60	981	246	110	62	40	28	21	16	11	8	6
2/3	1144	287	128	73	47	33	24	19	12	9	7
.70	1235	310	138	78	50	35	26	20	13	10	7
.75	1389	348	155	88	57	40	29	23	15	11	8
.80	1571	393	175	99	64	45	33	26	17	12	9
.85	1797	450	201	113	73	51	38	29	19	14	10
.90	2102	526	234	132	85	59	44	34	22	16	12
.95	2600	651	290	163	105	73	54	42	37	19	14
.99	3675	920	409	231	148	103	76	58	38	27	20

Effect size Small (0.2) Medium (0.5) Large (0.8)

$$\text{Effect size, } d = \frac{\text{Mean Treatment} - \text{Mean Control}}{\text{standard deviation (pooled)}}$$

$$\text{standard deviation (pooled)} = \sqrt{\frac{\sigma 1^2 + \sigma 2^2}{2}}$$

UNIVERSITY OF BIRMINGHAM