BST 215

Quantitative Research Methods

(Term 1)

John Doyle
Aberconway E22
DoyleJR@cf.ac.uk

DESCRIPTIVE STATISTICS

Notation

Given a set of numbers (*observations*): {5, 6, 2, 1, 19, 9, 7}
n = 7, the number of observations
$x_1 = 5$, $x_2 = 6$, …, $x_7 = 7$; also $x_n = x_7$
x_i is the ith observation: if i = 1, it is the first one.
∑ is the Greek letter capital S, which stands for "sum".

Hence $\sum_{i=1}^{7} x_i$ is to sum all the numbers from x_1 to x_7 (= 49)

More generally, $\sum_{i}^{n} x_i$ is to sum all n numbers, or even more succinctly: $\sum x_i$

Occasionally we don't want to sum everything. Then we need to be precise:

$\sum_{i=2}^{6} x_i$ is to sum all but the first and last numbers (= 37)

$\sum_{i\neq 3}^{7} x_i$ is to sum all numbers except x_3 (= 47)

The Greek letter ∏ can be used in exactly the same way as ∑ but with the meaning to multiply numbers rather than sum them. Hence, $\prod_{i=1}^{n} x_i$ = 5×6×2×1×19× 9×7 (= 71,820)

Ordered data: A number of notations exist for writing down the numbers or observations in size order: 1, 2, 5, 6, 7, 9, 19. We will use $x_{(i)}$ to refer to the ith number when written down in ascending order. So, $x_{(5)} = 7$.

| x_i | is the absolute value of x_i. That is, if $x_i = -5$, then | x_i | = +5. Turn everything positive!

In statistics we can go a long way on knowing three facts: (i) **location**: where are the observations centred? (ii) **dispersion**: how spread out are the observations? and (iii) **n**: how many observations are there? There are a number of ways of measuring (i) and (ii). The smaller the dispersion in the observations, and the greater the number of observations, the more certainty we have that a measure of location is correctly located. This is a fundamental principle in statistics, as we see later.

Measures of LOCATION (a.k.a. CENTRAL TENDENCY)

► The **mean**, average, or arithmetic average are all words for the same familiar concept, which is:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (add them up and divide by n). Note, the mean occurs so frequently in statistics that it is given the special symbol, $\bar{x}$ **(= 7.00).**

► The **median** is the middle number of our numbers (*once put in order of size*), so in our case, since n=7, it is $x_{(4)} = 6$. It is *not* $x_4 = 1$. The rule is that median is found at rank (n+1)/2. So here, (7+1)/2 = 4. **(Md = M = 6).**

But if there were just 6 numbers, it would be found at rank (6+1)/2 = 3.5. But there is no number at rank 3.5 (the 3.5th smallest). Instead, we average the middle two numbers at rank 3 and rank 4: ($x_{(3)} + x_{(4)}$) / 2. So, for the numbers $x_{(i)}$ = 4, 6, 19, 21, 33, 34, we average 19 and 21, to get median = 20. If n is odd, median is middle number; if n is even, median is average of middle two numbers.

► The **mode** is the most frequently occurring number. In the n=7 example, all numbers are the same. Therefore the mode occurs at all numbers. This already gives us a clue that the mode, although a useful theoretic concept, is rarely of much practical use.

► The **geometric mean,** *g,* is defined as the nth root of the multiplied numbers. In algebraic notation:

g = $\sqrt[n]{\prod_{i}^{n} x_i}$ = $\left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}}$ **(= 4.94)**. Note: the numbers must be strictly positive. Geometric mean is useful when dealing in ratios, proportions, or growths. An investment grows by 10% in year 1 then falls by 10% in year 2. The overall growth is given by the multipliers 1.10 × .90 = .99, which is annualized to √.99 = 0.99499. Hence decline of 0.501% per annum. Arithmetically averaging +10% and -10%, or even 1.10 and .90 will not give the right answer.

► The **harmonic mean,** *h*, is defined as: $\frac{1}{h} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$ **(= 3.22)**

It is easier to see what is going on with the arithmetic, geometric and harmonic means as follows:
$\bar{x}+\bar{x}+\bar{x}+\bar{x}+\bar{x}+\bar{x}+\bar{x}$ = 5+6+2+1+19+9+7 ($\bar{x}$ best represents the numbers when adding them
g×g×g×g×g×g×g = 5×6×2×1×19×9×7 (g best represents the numbers when multiplying them)
$\frac{1}{h}+\frac{1}{h}+\frac{1}{h}+\frac{1}{h}+\frac{1}{h}+\frac{1}{h}+\frac{1}{h} = \frac{1}{5}+\frac{1}{6}+\frac{1}{2}+\frac{1}{1}+\frac{1}{19}+\frac{1}{9}+\frac{1}{7}$ (h best represents the numbers when reciprocating & adding them).

► **Trimmed mean** removes extreme low and high values before calculating the mean. It's precise definition depends on the amount of trimming that is used. 10% trimming means that the highest 10% and the lowest 10% of observations are removed. Here is trimming that removes the top and bottom two observations:

TM = $\frac{1}{n-4}\sum_{i=3}^{n-2} x_{(i)}$. A 25% trim, thus averaging over the middle 50%, is known as the **midmean**. As trimming approaches 50%, we end up with the median. On the same theme of averaging selected data, **trimean** = (Q1+2Q2+Q3)/4; **Q123** = (Q1+ Q2+Q3)/3; and **midrange** = ($x_{(1)}$+$x_{(n)}$)/2 = average of max and min.

Measures of DISPERSION (a.k.a. SCALE, or SPREAD)

► **Variance** = $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is the mean square deviation of the numbers from their mean. Note, $x_{(5)}$ contributes much more than any other number to the variance, which makes variance sensitive to extreme observations, an attribute that it passes on to the standard deviation (s.d.). In statistics the word variance has a precise meaning. Never use it more loosely to mean "variability".

► **Standard deviation:** s = $\sqrt{Variance}$.

There are versions of variance and s.d. that divide by (n-1) rather than n, which are used when we want to generalize from samples to populations. More later.

► **MAD$_1$** (first definition – mean absolute deviation from the mean): $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

► **MAD$_2$** (second definition – median absolute deviation from the median)

			Var	**MAD$_1$**			**MAD$_2$**
i	**x_i**	**$x_i - \bar{x}$**	**$(x_i - \bar{x})^2$**	**$\lvert x_i - \bar{x}\rvert$**	**$x_{(i)}$**	**$x_{(i)}$-M**	**$\lvert x_{(i)}$-M$\rvert$**
1	5	-2	4	2	1	-5	5
2	6	-1	1	1	2	-4	4
3	2	-5	25	5	5	-1	1
4	1	-6	36	6	6	0	0
5	19	12	144	12	7	1	1
6	9	2	4	2	9	3	3
7	7	0	0	0	19	13	13
			30.57	**4**			**3**

s.d. = √30.57 = 5.53

► **Range:** maximum – minimum = $x_{(n)} - x_{(i)}$. Is sensitive to extreme values (obviously).

► **Quartiles.** *[Not themselves measures of dispersion, but see IQR]* Intuitively speaking, quartiles divide the ordered numbers into 4 equally numerous groups, just as the median divided the numbers into 2 equally numerous groups. Unfortunately, several different definitions of how to do this exist (e.g. Excel and SPSS differ, as does the following simple definition found in Silver's book): the lower quartile Q1 is found at rank (n+1)/4; the upper quartile Q3 is found at rank 3(n+1)/4. Median is Q2, found at rank 2(n+1)/4. Q1 = $x_{(2)}$, and Q3 = $x_{(6)}$.

Life is simple when n+1 is divisible by 4, because we find the quartiles at integer (whole-number) ranks, thus giving actual values of $x_{(i)}$. But if n+1 is not divisible by 4, as in the n=6 example [$x_{(i)}$ = 4, 6, 19, 21, 33, 34], then Q1 is found at a fractional rank: (6+1)/4, i.e. rank 1.75; and Q3 is at rank 3(6+1)/4 = 5.25. The median is at rank 3.5, and the definition told us to average 19 and 21. This is equivalent to dividing the gap between 19 and 21 in half and locating the median there. Rank 1.75 is clearly closer to rank 2 than to rank 1, so we locate Q1 at ¾ of the way across the gap between $x_{(1)}$ and $x_{(2)}$. The gap is 6 – 4 = 2, so we move on 1.5 (¾ of 2) from $x_{(1)}$ = 4. Hence Q1 = 4 + 1.5 = 5.5. Similarly Q3, being located at rank 5.25 is ¼ of the way across the gap between $x_{(5)}$ and $x_{(6)}$, which is 34-33 = 1. So, Q3 = 33 + .25.

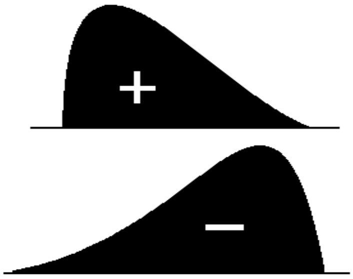
► **IQR (inter-quartile range).** Q3 – Q1 = $x_{([n+1]/4)} - x_{(3[n+1]/4)}$. This is the middle 50% of observations, and is not at all influenced by extreme values. It is therefore said to be a *robust* measure of dispersion – unlike the range of s.d. Later we see that the IQR is an integral component in the construction of boxplots.

Standard deviations, ranges, and IQRs are all measured in the same units that the observations were measured in (dollars, metres, kilograms, numbers of bananas, etc.). Variance is measured in dollars-squared, bananas-squared, and so on. Don't let this be your last thought as you drift off to sleep. Here is a "dimensionless" measure of variation that is sometimes useful to compare variation across different situations:

► **Coefficient of variation**: standard deviation / mean. In the n=7 example it is: 5.53 / 7 = .79. In the n=6 example it is: 11.67 / 19.5 = .60. Changing units of the observations (e.g. from pounds to pence, or from litres to pints) will change the mean and the standard deviation, but it will not change the coefficient of variation. Data must be "ratio scale" (more on levels of measurement later): for instance, changing from ºCentigrade to ºFahrenheit will give different coefficients of variation, implying that neither should have been calculated in the first place.

Measuring SKEW & KURTOSIS

It is usually important to know something about the general shape of the distribution (of observations), which then suggests the kinds of testing that should be done. First, **skewness** tells us whether the observations are symmetrically distributed. Positive skewness occurs when the distribution comes to an abrupt end for smaller values than for larger ones, and vice versa for negative skew.



Three measures to capture this quality are:

The **inter-quartile measure of skew**.

This relies on the idea that

= (Q1 + Q3 – 2Q2) / IQR

Pearson's coefficient of skewness:

= 3(Mean – Median) / sd

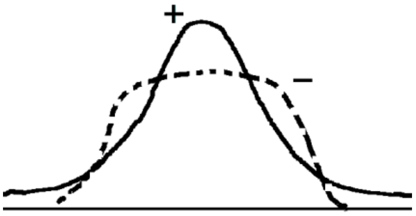
The **moment measure of skew**.

Define the k^{th} *moment* about the mean to be:

$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ (The variance is m_2.)

Skew = $m_3 / (m_2)^{3/2} = m_3 / s^3$, where s is s.d.

Finally, kurtosis tells us whether distributions end abruptly (negative kurtosis), or have stretched tails at both ends (positive kurtosis), illustrated in the following distributions.



Although subtle, and sometimes hard to see by eye, positive kurtosis in particular can have a profound effect on statistical testing. The usual way to measure kurtosis is using moments:

$K = m_4 / (m_2)^2 - 3$.

The subtracted 3 comes from the fact that a normal (Gaussian) distribution, which is often used as a reference distribution, has a moment kurtosis of 3. More precisely, this is known as "excess kurtosis". If the 3 is not subtracted, it is called Fisher kurtosis (I think). Unfortunately, most statistical packages do not make this distinction clear. SPSS uses excess kurtosis.

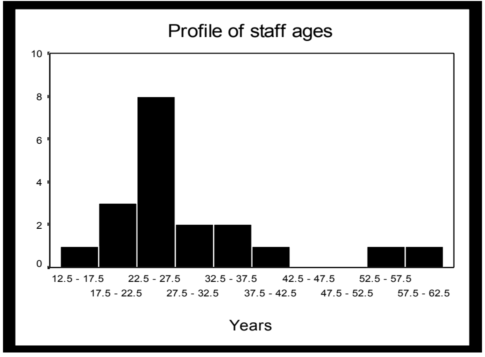
Excel has the following relevant functions: sum(), average(), geomean(), harmean(), median(), quartile(), varp(), stdevp(), avedev(), skew(), and kurt(). The last two are variants on the moment definitions given here.

Question 1. *Given the following observations (people's age), calculate the descriptive statistics you think are most important: {14, 15, 15, 16, 19, 22, 27, 28, 32, 62}.*

GRAPHICAL DEVICES

Here are the ages of 19 people in a department: 17, 18, 21, 21, 24, 24, 25, 25, 26, 26, 27, 27, 30, 32, 34, 34, 38, 57, 58. What does the distribution of ages look like?

Histograms



Note, software has chosen that the intervals are [12.5, 17.5), [17.5, 22.5), [22.5, 27.5), and so on, where [a, b) means a ≤ x < b (up to but not including b). The default is usually is to have intervals of equal width (here 5 years). Histograms with unequal width categories need to have the height of the bar adjusted.

Stem & leaf diagrams

The same data may be displayed so that we get the same qualities as a histogram but can recover the actual numbers. In the following, the number 17 is broken down into 15+2, and 58 is 55+3. The starting value here is 15, with intervals of 5.

```
55| 2 3
50|
45|
40|
35| 3
30| 0 2 4 4
25| 0 0 1 1 2 3
20| 1 1 4 4
15| 2 3
```

57 and 58 seem to be outliers.

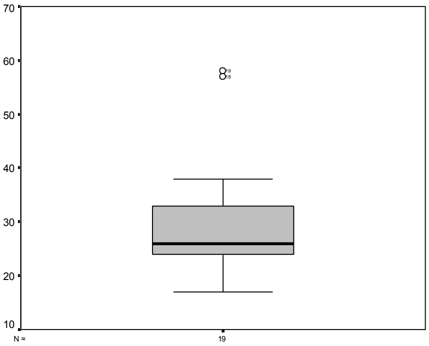
Here is the same data with starting value of 16, and intervals of 8.

```
56| 1 2
48|
40|
32| 0 2 2 6
24| 0 0 1 1 2 2 3 4 6
16| 1 2 5 5
```

It's a good idea to try out alternatives. Both too many intervals and too few are poor at revealing interesting features of distributions. There are issues when numbers are not positive integers.

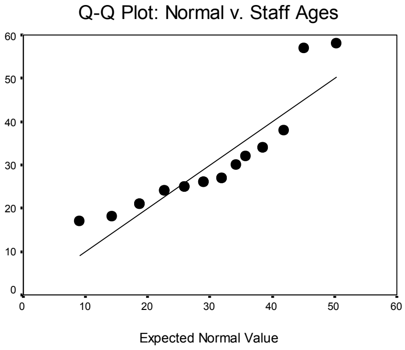
Here is the same data viewed as a boxplot. Marked are: max and min that are not outliers, Q1, Q2, Q3, and outliers with their IDs. ID18 is 57, ID19 is 58. Outliers defined as lying further than 1.5*IQR beyond Q1 or Q3.

Question 2. *Using the same data, calculate all the other descriptive statistics you now know.*

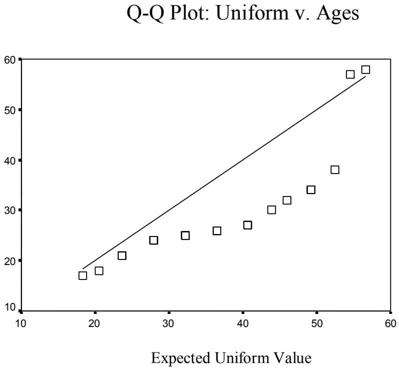


QQ (quantile-quantile) Plots

A special kind of scatterplot. Plot one distribution against another. Here, whether the data are normally distributed (should lie on line). Note, I have swapped the axes from those that SPSS uses as default.



Does it look like a uniform distribution (should lie on line)?



Levels of measurement

The standard scheme is due to Stevens (1946). It constrains the kinds of algebraic operations and statistics that are considered legitimate.

Level	Notes & Ops	Stats
Nominal / Categorical	Can count instances in the categories, but the categories have no ordering.	Mode, binomial (sign test), χ^2, loglinear.
Ordinal	Can order or rank observations. Tied ranks are allowed.	Median, quartiles, Mann-Whitney U, Wilcoxon Signed Rank, Kendall's tau
Interval	Can subtract one observation from another to form differences.	Mean, standard deviation, IQR, Regression, Pearson's R, Spearman's rho, t-tests, ANOVA
Ratio	Can divide observations, hence forming *ratios.* Implies a true zero where there is an absence of the variable.	Coefficient of variation, geometric mean.

Ratio. Money, weight, numbers sold are all ratio: it is possible to think of A being *twice* as heavy as B, costing twice as much as B, selling 10% more than B, etc. Also, zero sales / money / weight really means zero.

Interval. We can subtract degrees centigrade to say that today (20°C) is 5°C hotter than yesterday (15°C), and it is the same difference at all points of the scale (50°C is still 5°C hotter than 45 °C). But we cannot form meaningful ratios: 20°C is not twice as hot as 10°C. Years are interval: each year 2011, 2012, 2013 has the same interval, but 2013/2012 is a meaningless ratio. Note, there is an absolute zero of temperature (degrees Kelvin) 0°K = -273°C which is a ratio measurement: zero means absence of molecular motion and hence no heat. Ratio and interval are sometimes grouped together as "metric" data, because we have *measured* some underlying quality for each observations. Interval measurement is often ratio with a man-made, somewhat arbitrary zero.

Ordinal. I can order objects by size without ever having to measure them (hence ordinal is non-metric). Sometimes the ranking contains many ties (small dogs, medium dogs, large dogs). Sometimes the ranking is unique (who swiped themselves in first, second, third,.., last). Sometimes in-between. You will often see interval data operations performed on ordinal data.

Categorical / nominal. Here we just have a bunch of categories without order, and how many belong to them. For instance, numbers voting for different political parties (in UK system, at least). Note, two categories are considere not to have an order: numbers passing versus failing and exam; numbers of Male versus Female in department.

There is a big decrement in *statistical efficiency* and hence *power* when moving from Ordinal to Categorical data and tests – given idealized conditions apply. Not so much from interval to ordinal. But when distributions are not idealized the power of different tests to detect effects may even be reversed. Therefore, sometimes we may *choose* to treat interval data as ordinal, or even categorical, because fewer assumptions are made about our actual data conforming to an idealized distributions such as the normal distribution.

More complete schemes. Some statisticians have pointed out that Steven's classification does not do justice to other kinds of data we come across. For instance: partial orderings; numbers which are bounded (e.g. probabilities only exist between 0 and 1); ordered measurement but with inherent circularity, e.g., months of the year, distances round the world; graded category membership (e.g., when exactly does red become orange become yellow?), and so on.

The role of the Null Hypothesis in testing theories.

(A Socratic dialogue between me and my stooge)

"I think this coin is biased in favour of heads."

[This is the experimental hypothesis, aka the alternate hypothesis, H_a or H_1]

"I think the coin is not biased in favour of heads. Prove what you say!"

"OK. For the moment I'm going to assume that you are right and I am wrong: In that case the coin would not favour heads."

[This is the Null hypothesis, or H_0]

My stooge and I go away and toss the coin a pre-decided number of times (say 100) and come back with the results.

[The experiment / empirical investigation]

"I've found there are many more heads than tails. I've been to a statistician who tells me that finding this many heads (*or more*) in 100 tosses of an unbiased coin will occur, on average, exactly one time in 500. In other words the probability of this happening is equal to .002 (p = .002)."

"So?"

"This leads me to reject what you claim (the hypothesis that the coin is not biased towards heads)."

[Rejecting the Null hypothesis]

"But what precisely does the statistician mean by '…on average, exactly one time in 500'?"

"If I were to go and repeat my experiment 500 times using an unbiased coin I would expect to find such an extreme number of heads exactly once."

[We appeal to the argument of repeatedly re-sampling the population (at least in theory). The argument of repeated re-sampling allows us to generalise from our sample (of 100 tosses) to the population. The population, in this case, is a rather abstract one: it is all the tosses of the coin, as yet unmade.] We hope to decide between: Is my sample from population X (tosses with an unbiased coin) or population Y (tosses with a biased coin)?][1]

"So, by disproving me you think you've proved your little theory? It doesn't always follow that if I am wrong you must be right."

"But in this case it does. Your so-called theory was merely that my theory was wrong. So, if you are wrong that I am wrong, it follows that I must be right! Two wrongs do sometimes make a right."

"I think you've overstated your case, smarty-pants. Suppose your little experiment with the 100 tosses of the coin had just happened to be the 1 in 500. You can't know with 100% certainty that you are right."

"Yes, I suppose I did get carried away with my success. Let me be more precise, then. I reject your Null hypothesis, accepting a probability of .002 that, given my data, I am wrong to reject H_0."

In technical jargon falsely rejecting H_0 is known as a Type I error. When we reject the Null Hypothesis, we always do so accepting a certain probability of committing a Type I error (i.e. that we might be wrong). This is very important though usually it is usually left implied in research papers. To counter this sloppiness, remember the following phrase:

When I reject the Null Hypothesis it is always with an estimated probability of making a Type I error that we compute as the p-value.

It is custom and practice to "reject H0" when the probability of making a Type I error falls below .05 (i.e. less than 5 in 100). People then talk about it as being a "statistically significant result", or more loosely a "significant result".

Another example in brief: I want to know whether boys or girls weigh more at birth. Ha is that the mean weight for girls will not equal the mean weight for boys. H0 is that the mean weights for boys and girls will be the same. I sample the population of boys and girls, and find that, say, my sample of boys weigh more, but only to a small degree. My friendly statistician tells me that, based on this evidence, if I took another sample it is not unlikely that the girls in it would weigh more than the boys. I conclude that I cannot reject H0. This is not to say that there is no difference in reality - if you think about it, there is bound to be some tiny difference. It's just that I haven't got enough evidence to decide with sufficient certainty which way it goes… which leads me on to Type II errors.

Type II errors & Power

A Type I error occurs whenever someone rejects H0 when it shouldn't have been rejected. Since researchers are usually interested in rejecting H0 (and thereby 'proving' their own pet theory), there is a tendency for them to want to reject H0 when they shouldn't. To these people statistics may seem like a ball and chain placed around their necks by kill-joy statisticians, whose sole purpose on this earth is to prevent them from having fun in the big world of theory-building. As researchers we all suffer from this tendency, wanting to see things in our data that aren't there, wanting to read the future in tea-leaves. It is therefore natural that statistics as first taught should emphasise avoiding Type I errors.

[1] Suppose I used a different coin that was a bit bent. I theorised that because of the asymmetry it would be biased, but I wasn't prepared to say whether it would favour heads or tails, just that it would favour one of them. In this case we are in the realm of 2-tailed testing, where previously it had been 1-tailed.

However, another kind of error (Type II error) occurs when someone fails to reject H0 when it should have been rejected. That is, they miss something. The logic is as follows: There is some theoretically important difference that actually does exist in the population; the researcher samples the population to see if it is there, but their sample happens not to exhibit that difference, or not to a sufficient degree to permit H0 to be rejected.

		Real state of affairs	
		H_0 true	**H_0 false**
Decision from data	**Reject H_0**	**Type I Error α**	OK
	Don't reject H_0	OK	**Type II Error β**

Type I error: You reject H0 when you shouldn't. Your data leads you *to see something that isn't really there* (in the population). (= Illusions; crying "Wolf!"; astrology; false positives.)

Type II error: You fail to reject H0 when you should have. Your data leads you *to fail to spot something that is really there* (in the population)

The probability of making a Type I error is usually signified by the Greek letter alpha (α); and the probability of making a Type II error by the Greek letter beta (β).

People seem to be 'trigger-happy' towards rejecting H0. That is, we undertake a piece of research in order to demonstrate a difference, show a correlation, or whatever. If it doesn't turn up we feel we have been wasting our time and effort. Thus we are not entirely dispassionate: we usually *want* H0 to be false. Recognising this bias among researchers, statistics books and teachers place great emphasis on the sins of committing Type I errors. So much so that statistics may often seem to be a super-ego, preventing you from having fun with wild and speculative reasoning.

But we are also lazy. Being lazy we design investigations that have insufficient power to allow us to reject H0. The greater the power of an investigative design, the less chance there is of making a Type II error. Power can be defined as the probability of *not* making a Type II error, which is just 1 - β. Power increases if, among other things:

(a) the difference (or correlation, or 'effect') we are looking for is large,
(b) the variability in the population is small,
(c) we use bigger samples,
(d) we use more powerful investigative designs (e.g. repeated measures rather than independent measures, where possible and appropriate),
(e) we use more powerful statistical tests (e.g. tests fitted for interval data rather than ordinal, or nominal, where appropriate): (also removing outliers, de-skewing distributions, etc.)
(f) we screen out unwanted, definable, sources of variability: e.g. measure people's height without shoes, at a specific time (why?); it's almost always worth recording whether respondents are M or F, etc. Better sampling could also be included here.

Of these, (a) is outside our control. The apparent population variability may be smaller than we think: the better job we do of (f), the less variability may be left in (b). (d) and (e) require technical, statistical appreciation. In my estimation (e) may sometimes reduce β to as little as 50% of what it might have been, but usually to only 90%-95%, though this improvement may be achieved in almost all cases (my estimation, again). (d) could reduce β to 25% of original, but it can't always be applied (the design might not be adjustable). (f) requires the most knowledge of your topic area. Big improvements via (f) often imply an improved understanding of the problem area, whether the understanding causes or is a consequence of the improvement. (c) is the most mechanical, most sure-fire way of improving power, with unlimited potential, given large enough n, where possible. However, there are only so many EC countries, electrical utilities, etc. Also there is a cost.

(g) Power also increases the more generous (larger) alpha is. That is, if we hold (a) – (f) constant, we can always improve power by increasing alpha. But this means we will be committing more Type I errors, which is usually considered undesirable. However, the trade-off between α and β does depend on the circumstance. For instance, you don't want to miss detecting the possible occurrence of a large and tsunami, even if it leads to too many false alarms when they don't occur.

Sometimes, of course, researchers are really interested in showing that there is no difference. For instance, that wonder-drug X is no such thing, and that people recover just as quickly without it as with it. It should be clear that the claim "we found no difference" is often because the design is hopelessly under-powered (e.g. a couple of people took X but Fred didn't - look how quickly Fred recovered). In this special case the onus should really be on the researchers to demonstrate that they have used a sufficiently powerful design to detect practically important differences, were they do exist. The logic of hypothesis testing does not lend itself so easily to this special case.

In the first example we predicted that there should be more heads than tails: this is a directional prediction, thus we were engaged in what is known as 1-

tailed testing. In the second example our hypothesis was sufficiently loose that (a) finding boys to be heavier would be a result, and (b) finding girls to be heavier would also be a result. This is an example of 2-tailed testing. Generally speaking, we like to be in a position to make 1-tailed tests, even if we opt for the more conservative 2-tailed. This is not just because we can halve the p-value of a 2-tailed test and thus reject more H0s, but because in order to make a case for a 1-tailed test, we need to be able to make more focused predictions based on past research or the force of logic itself. This all give the reader the impression that the researcher is in control and knows where he / she is going. *What is not acceptable is to have a peek at your data after it has been collected, then seeing which way the data is going, to make a tenuous post hoc case for a 1-tailed test.*

Occasionally, we are in the happy position that 2-tailed testing is the best of all options. A classic use of 2-tailed testing is to arbitrate between competing scientific (social scientific) theories. First, find or engineer a situation in which the two theories make opposite predictions. For instance, theory X predicts boys will be heavier, whereas theory Y predicts girls will be heavier. Find out which are heavier (with a sufficiently small probability of making a Type I error). Ditch the theory that doesn't predict. This head-to-head confrontation is rarely done, but according to Popper, theory falsification is the most efficient way to proceed.

I may have a complicated theory that makes me predict, in the first example, that heads will come up more often on this particular coin. However, every coin will be biased to some tiny degree (the wind resistance will differ on either face, the amount of friction, the prominence of the features, etc, etc.). So, at the outset we must assume it's 50-50 whether Ha will be correct or not. In other words, there's a p of .5 that my theory will receive confirming evidence. Once this confirming evidence is revealed, how much have we moved on? Not much. So what is the p=.002? Don't be confused into thinking that it has anything to do with the probability that my theory is wrong.

If I have a powerful enough design (e.g. a million tosses) I could determine which way the bias goes for any given coin. Let's say 57% of coins have a tiny bias to heads and 43% a tiny bias to tails (NOT that heads will come up 57% of the time). Do not be confused into thinking that if I find a statistically significant result consistent with my theory that this necessarily says much about my theory.

Null hypothesis testing is about establishing what are the facts. It is not about establishing theories, only insofar as facts support theories! THIS IS IMPORTANT.

I have a theory about wearing a bobble-hat, which I put to a battery of tests…

Prediction 1: Exercising (+ hat on) will make me fit: prior likelihood v.high - no information in correct prediction;
Prediction 2: I'll be able to do tasks better with my right hand while wearing hat. ditto
Prediction 3: Eating more (+hat on) will make me slimmer: Highly diagnostic.

The more unexpected the prediction(s) (if they turn out to be so), the more likely we are to believe the theory.

Categorisation of simple statistical tests

		Differences	Correlations
Categorical		Binomial, Sign Test, Chi-square test	Phi
Ordinal	Repeated Measures	Wilcoxon's Signed Rank Test	Spearman's rho (Kendall's tau)
Ordinal	Independent Measures	Mann-Whitney U test (aka Wilcoxon's Rank-Sum test)	
Interval	Repeated Measures	repeated measures t-test	Pearson's
Interval	Independent Measures	independent measures t-test (equal var) independent measures t-test (uneq var)	

Independent measures occur when there is no connection between the observations in groups A and B - for instance, 30 French manufacturers compared with 28 British manufacturers; or 43 women's salaries compared with 29 men's. Note, A and B may have different N.

Repeated measures (also known as "matched pairs") occur when each observation in group A can be paired with an equivalent observation in Group A. This occurs in two ways:

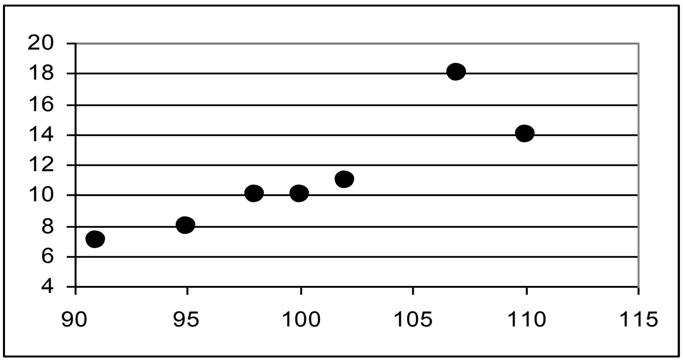
(i) People, organizations, etc can be paired as being highly similar – the ultimate (which occurs only in psychological studies, unfortunately) is to study, say, twenty pairs of twins where one twin gets treatment A, and the other gets treatment B. Except for the case of twins, however, usually quite a bit of work has to go into ensuring that the pairs have been properly matched.

(ii) The other way occurs when it is the same person / organization that gets both treatment A and treatment B. For instance, a sales force may work under incentive scheme A, and then the same people may work under scheme B. The problem here is that there may be carry-over effects from A to B. The person who has undergone A is no longer the same person they were before A began. A partial solution is to have half the force working under A then under B, while half work under B then A.

Given the difficulties of matching up observations in group A with those in group B, why not treat A and B as independent measures? The reason is that successful and appropriate repeated-measures designs are VERY much more powerful that independent-measures equivalents. This is because repeated-measures designs factor out individual differences among the observations – for instance, this person might be a supersalesperson, so that which group they get assigned to (A or B) might have a bearing on the statistical outcome. Repeated-measures designs allow us to compare supersalesperson doing A with supersalesperson doing B.

CORRELATION AND SIMPLE REGRESSION

Scatterplot of (x_i, y_i), n=7.



Calculating regression line and correlation coefficient

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
91	**7**	-9.43	-4.14	88.90	17.16	39.06
95	**8**	-5.43	-3.14	29.47	9.88	17.06
98	**10**	-2.43	-1.14	5.90	1.31	2.78
100	**10**	-0.43	-1.14	0.18	1.31	0.49
102	**11**	1.571	-0.14	2.47	0.02	-0.22
107	**18**	6.571	6.857	43.18	47.02	45.06
110	**14**	9.571	2.857	91.61	8.16	27.35
100.4	11.1			37.39	12.12	18.80
$\bar{x}$	$\bar{y}$			**Var(x)**	**Var(y)**	**Cov(x,y)**

b	0.503
a	-39.3
R	0.883
t(5)	4.204
P	**.008456**

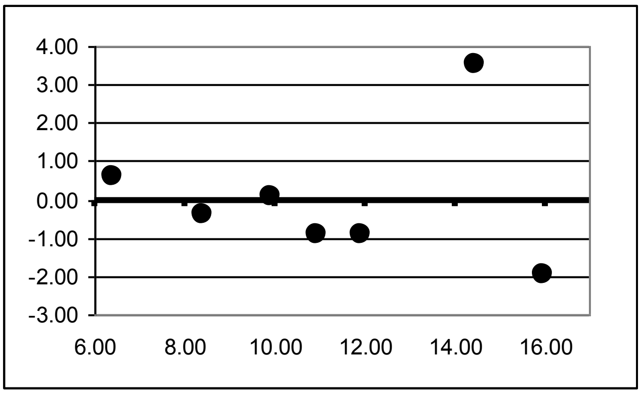
b = $Cov(x,y)/Var(x)$.

Line of best fit goes through ($\bar{x}$, $\bar{y}$), so $\bar{y}$ = a+b$\bar{x}$; knowing $\bar{x}$, $\bar{y}$, and *b* we can calculate *a*.

(Pearson's) R = $Cov(x,y)\big/\sqrt{Var(x)Var(y)}$.

$$t(n-2) = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \ .$$

Examining the residuals and checking for outliers
Scatterplot of estimated (x-axis) vs. residual (y-axis).



Are there patterns still in the residuals?

Checking for outliers using 1.5 * IQR

$\hat{y}_i$	Residuals $y_i - \hat{y}_i$	Ranked Residuals	
6.40	0.60	-1.955	
8.41	-0.41	-0.933	**Q1**
9.92	0.08	-0.927	
10.93	-0.93	-0.414	**Median**
11.93	-0.93	0.078	
14.45	**3.55**	0.597	**Q3**
15.95	-1.95	**3.553**	

Q1	-0.933
Q3	0.597
IQR	1.530
1.5*IQR	2.295
Lower Fence	-3.228
Upper Fence	2.892

$\hat{y}_i$ is estimate of y_i

(x_5,y_5) is a regression outlier in the residuals, lying beyond upper fence (=2.89). Note, however that neither x_i nor y_i is itself an outlier.

Spearman's correlation (Rho)

Although there is a special formula for Spearman's Rho, it is easiest to think of it as a Pearson's correlation performed on ranked data. Here is the same data, but treated as ordinal (ranked) data. SPSS returns exactly the same R and probability as if you had done a Pearson's on ranks.

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	1	-3	-3	9.00	9.00	9.00
2	2	-2	-2	4.00	4.00	4.00
3	3.5	-1	-0.5	1.00	0.25	0.50
4	3.5	0	-0.5	0.00	0.25	0.00
5	5	1	1	1.00	1.00	1.00
6	7	2	3	4.00	9.00	6.00
7	6	3	2	9.00	4.00	6.00
4	4			4.00	3.93	3.79
$\bar{x}$	$\bar{y}$			**Var(x)**	**Var(y)**	**Cov(x,y)**

b	0.946429
a	0.214286
Rho	0.954994
t(5)	7.199067
p	**.000806**

In this case, Spearman's Rho > Pearson's R (.95 vs .88) and was the more significant correlation (p≈.0008 vs p≈.008) – *because of the outlier* which affects Rho less than R.

Finally, R^2 (literally squaring the correlation coefficient) turns out to be the proportion of variance explained by knowing about the relationship between x and y.

Excel has the following useful functions *(presuming data is in a1:a7 and b1:b7)***:**
average(a1:a7), median(a1:a7). varp(a1:a7), cov(a1:a7,b1:b7), correl(a1:a7,b1:b7) , rank(a1, a$1:a$7), tdist(t,df,tails) e.g., tdist(7.199,5,2), which translates ts into ps.

Kendall's tau is another rank-based correlation coefficient - but not in BST215.

Some statistical tests

Before we get into this one small, niggly point. If we take a sample from a population the mean of that sample will be an *unbiased* estimate of the population mean. That is to say the sample mean will neither systematically over-estimate or under-estimate the population mean. Unfortunately the sample variance (and standard deviation) will systematically under-estimate the real population mean. Fortunately, it does this in a systematic way.

Instead of using $\Sigma (X-x_i)^2 / n$ as our formula for the variance, we need to use

$\Sigma (X-x_i)^2 / (n-1)$. In both cases the standard deviations are just the square root of the variance. The first version of the standard deviation is usually referred to as s_n , and the second one as s_{n-1}.

Repeated measures (matched pairs) t-test

To determine whether the mean for group A is bigger/smaller/different from the mean for group B, when group A contain either the same people as group B, or they can be paired off in some convincing way. Examples: do people find text in Goudy Old Style harder to read than in Arial? Time each person reading passages first in one font, then in the other. (Problems of carry-over.) Are people more touch-sensitive with their left hand than their right? Do doctors earn more than dentists across the world, or vice versa? Get a sample of countries, and for each one measure the average earnings for doctors and dentists. Are first-born twins taller than second-born twins? (See below). They should all have this general form:

	Goudy Old Style	Arial	Difference
Person 1	**153**	**159**	**6**
Person 2	209	222	13
..			
Peson n	166	170	4

or

	Doctors	Dentists	Difference
UK	45000	50000	5000
Germany	66000	64000	-2000
…….			
Country n	34000	24000	-10000

We test the mean of the differences against zero.

Step 1: Calculate the mean difference (call it X).
Step 2: Calculate the standard deviation of the sample mean (aka standard error):
standard error = $s_{n-1} / \sqrt{n}$. (This comes from the Central Limit Theorem). Note: n is the number of PAIRS.
Step 3: Divide X by standard error to the t-statistic.
Step 4: use (n-1) degrees of freedom to look this (called t) up in t-tables.

This turns out to be $t = X / \text{stderror} = X / (s_{n-1} / \sqrt{n}) = (X \sqrt{n}) / s_{n-1}$.

The greater t is, the greater surety we have in rejecting H0. But note: t gets big if (i) X is big, (ii) n is big; (iii) s_{n-1} is small. These are the three things I've being going on to you about since Lecture 1!

Independent measures t-tests

If you cannot pair each member of group A with a member of group B, then you need this form of the t-test. The formula this time is

$t = X / \surd(s_1^2 / n_1 + s_2^2 / n_2)$, where s_1 and s_2 are the standard deviations of groups A and B, respectively, and n_1 and n_2 are the numbers in groups A and B, respectively. The square root is of all of that calculation in brackets. Although the algebra is not quite as simple as for the repeated measures t-test, it is possible to see the same three factors increasing power in the test: Bigger X; bigger n_1 and/or n_2; and smaller dispersions (s_1 and s_2).

Example

Consider the following example. In the right hand side it is analysed using repeated measures (aka matched-pairs) t-testing. In the second we use independent measures t-tests (imagine I lost the pairing of twins).

Ha: First-born twin will be taller than second-born
H0: The first-born will not be taller (may be smaller or same height)

Example: Heights of 8 pairs of twins (in cm)

First-born Twin	Second Twin	Height Difference
179	177	2
179	178	1
155	156	-1
199	196	3
169	168	1
133	130	3
143	142	1
192	190	2

Independent measures t-test			Repeated measures t-test	
Average	168.625	167.125	Average	1.5
$S_{(n-1)}$	23.286645	23.012031	$S_{(n-1)}$	1.3093073
	67.783482	66.194196		
s.e.	11.574873		s.e.	0.46291
T	0.13		T	3.24
df	14		df	7
P (1-tail)	0.449		p (1-tail)	0.007
Conclude:	Don't reject H0		*Conclude:*	Reject H0

This example has been chosen to show up the possible increase in power from using repeated measures designs (aka Within-Subjects designs) compared with independent-measures designs (aka Between-Subjects designs). It overstates the case, though not necessarily by very much.

In journals we may see this kind of result written up succinctly. For instance as: Mean heights for first-born and second-born twins were 168.63cm and 167.13cm, respectively. This difference in means was significant by a repeated measures t-test, $t(7) = 3.24$, $p<.007$, 1-tailed. OR The difference in means was non-significant by an independent measures t-test, $t(14) = 0.13$, $p>.05$, 1-tailed. Thus we cannot reject the null hypothesis.

Finite population correction

Suppose that our group of 8 twins was sampled from a particular race of people living on a remote island, and suppose that there were only 10 twins alive on that island. It is clear that if we were to re-sample the population, we would have to have at least 6 of the original twins in our sample. Thus, the outcome cannot surely be very different from the one we obtained. This is the case of having a finite population to generalise to. It is taken care of by deflating the standard error by a factor $\sqrt{((N-1)/(N-n))}$. (See Silver, 2nd Ed., p.187), where N is the size of the population, and n the size of the sample. However, think carefully about who you wish to generalise to: the population alive, or a more abstract population of all possible twins who might be born to members of that island race.

Non-parametric alternatives to t-test (*optional*)

To parallel our distinction between Pearson'r r and Spearman's Rho, Wilcoxon's Signed Rank test uses ordinal data. First rank the difference data, disregarding signs. (Lowest difference gets smallest rank), to give row 2, below. Return the signs (row 3, below). Sum:

Differences	2	1	-1	3	1	3	1	2
Ranks (abs vals)	5.5	2.5	2.5	7.5	2.5	7.5	2.5	5.5
Signed Ranks (=Ri)	5.5	2.5	-2.5	7.5	2.5	7.5	2.5	5.5

$W = \Sigma R_i$ (= 31 in this case)

The figure of W may be looked up in special tables. In this case we get $p < .025$, (1-tail). This is not quite as small as the probability obtained using the repeated measures t-test ($p < .01$, 1-tail), which bears ut my point in (e), above, that the right test can aid power.

Alternatively, for $n \geq 10$, the null distribution of W is approximately normal, so we can calculate a z-score:

$$z = W / \sqrt{(\Sigma R_i^2)}$$

There is an equivalent rank-based test (the Mann-Whitney U Test, aka Wilcoxon's Rank-Sum Test) that parallels the independent-measures t-test.

Binomial

Problems such as coin-tossing may be modelled using the Binomial distribution. These problems consist of a number of trials (e.g. tosses of a coin, rolling a dice), with two outcomes on each trial (e.g. Heads or Tails; throwing less than 3 vs 3 or more). If the known probability of one outcome is p, the probability of the other is 1-p (usually called q). For coin-tossing $p = q = .5$; if a "success" is throwing less than 3, $p = 1/3$ and $q = 2/3$. Each trail must be independent of all others. That is, knowing the outcome of any past trial or trials gives you no additional information about how the current trial will turn out.

For instance, what is the probability of a dice thrown 10 times landing on 6 exactly three times? There are formulae that govern this, but far simpler is to use Excel, which has a function called binomdist. To calculate the above problem we would enter into an Excel spreadsheet cell the following formula:

=binomdist(3,10,1/6,0).

and it will tell you the answer is .155. It should be obvious where the first three figures come from. If we wanted to calculate what the probability is of throwing 3 or fewer (that is we want to accumulate probabilities up to 3) we would type in:

=binomdist(3,10,1/6,1).

The fourth number directs the function to accumulate (1) or not (0). The probability it calculates is .930.

One last example. You send out 100 questionnaires. What is the probability of receiving 24 or fewer replies to your survey if the probability of any one coming back is .2? Answer:

=binomdist(24,100,0.2,1) … which gives .869

Thus, the probability of getting back 25 or more is 1 minus this figure, namely .131. I bet you thought it was more than this.

If you don't use a spreadsheet, most calculators will do this for you these days. If n is large enough we may use the normal distribution to approximate the binomial (Rule is: When both np and nq ≥ 5). It has a mean of np, and a variance of npq. Suppose the expected proportion of successes is p, but we obtain P, we can calculate a z-score:

$z = (nP - np) / \sqrt{}(npq)$,

Looking z up in our normal distribution tables shows the cumulative probability of obtaining at least as extreme a P as we did. E.g. What is the probability of getting 540 or more heads from 1000 tosses of a coin? Here p = .5, P = .54, n = 1000. Hence,

$z = (540 - 500) / \sqrt{}(1000 * .5 * .5) = 2.53$; The probability of obtaining a z-score of this or greater is a mere .006.

A little bit of algebra on our formula reveals a familiar pattern. Divide top and bottom of the formula by √n, to get:

$z = ((P - p) \sqrt{}n) / pq$

So, z increases as n increases, as (P-p) increases, and as pq decreases. And as z increases, the more we are convinced we are not looking at a mere chance fluctuation. We know that getting 54 heads out of 100 is no news, but seemingly getting 540 out of 1000 (the same proportion) is news. That is because n is 10 times bigger, so z is √10 times bigger. (Same old story.)

Confidence Intervals

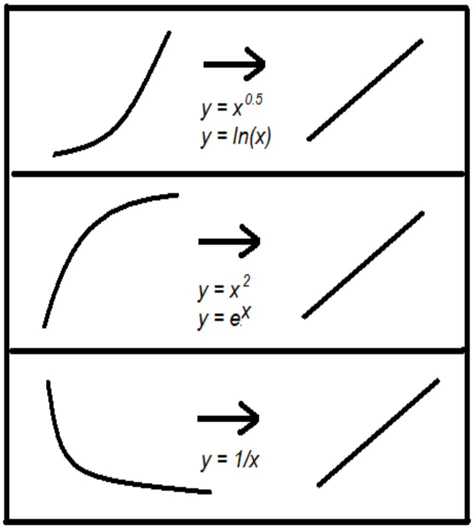
If you don't like null hypothesis testing, confidence intervals is an alternative, much recommended, though less frequent, way of arguing the case. Return to our twins problem:
We calculated that the mean difference in height between first and second born twins is 1.5 cm, with a standard error (= standard deviation of the mean of 8 twins) of 0.46291. Look up t values in tables, with 7 df. Say, p = .05 for a 2-tailed test. We read off that t = 2.3646. That says that if we took lots of different samples of twins (in groups of 8), 95% or the mean differences in height would lie 2.3646 standard errors from our obtained mean of 1.5. Since we know what the standard error is, we can calculate this distance. It is 2.3646 * .46291 = 1.095. So, we can be 95% confident that the true mean lies between 1.5 - 1.095 and 1.5 + 1.095, i.e. between .405 and 2.595. Our *95% Confidence Interval* lies between .405 and 2.595. Now, since our confidence interval does not include zero, less than 5% of samples will show the second born twin to be taller. In fact, we can be more confident than this. Samples that are above our confidence interval (i.e. means greater than 2.595) still demonstrate that the first-born twin is taller than the second one. We could, therefore, use a 1-tailed test. For p = .01, 1-tailed, with 7 df we read off that t = 2.998. Multiplying this by the standard error, as above, we get 1.388. In other words we anticipate that 99% of our samples of 8 twins would have a mean difference of 1.5 - 1.388 (= .112) or above.

Transforming data

Curves to lines

There are several reasons why we may wish to transform data: to remove skew from a distribution; to make a scatterplot more linear in x with y, so that we may calculate the (linear) regression equation; to “stabilize variance” (one assumption often made in standard analyses is that variance is approximately equal over the entire range of data – *homoskedasticity*). Sometimes, however, it is either more practical or more logical to work with transformed data – for instance, if a fleet of delivery lorries has to cover a certain number of miles in a week, then using gallons per mile for each vehicle may more easily reveal likely fuel costs than using gallons per mile (a simple example of the reciprocal transformation).

The curves on the left-hand-side of the figure may be straightened by using the following transformations. Square-rooting[2] and logging (ln(x)) also work to remove positive skew in distributions, whereas squaring and exponentiation (e^x) work to remove negative skew. These transformations only work on positive data

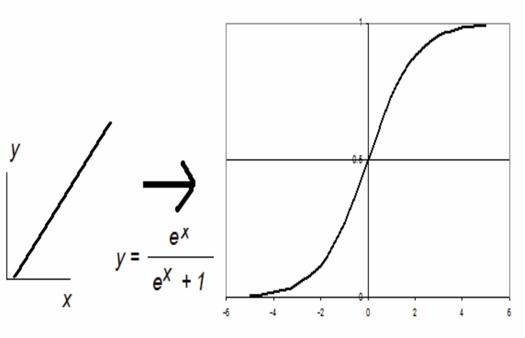


Lines to curves

Sometimes we wish to transform a straight line into something else – for instance, here a “lazy S” that lies between 0 and 1. This is useful for modeling probabilities that must lie between 0 and 1, as in the *logit* function (as shown in the figure).

The highly similar *probit* transform transforms the normal deviate z onto a probability value. Can use p=NORMSDIST(z) in Excel.

Both logit and probit may be used as “link functions” in regression when the dependent variable is categorical (usually 0 or 1), as in didn’t / did go bankrupt.



Data to ranks (OK for negative numbers)

A special kind of transformation occurs when we transform interval data into ordinal or ranked data. We do this when using Spearman’s rather than Pearson’s correlation coefficient. It turns out that many well-known non-parametric tests can be seen as interval tests performed on data transformed that has been transformed to ranks. For instance, Spearman’s rho is a Pearson’s correlation on ranked data, Mann-Whitney U is an independent measures t-test on ranked data, χ^2 is a Pearson’s correlation on categorical data…)[3].

Ranking and beyond (Van der Waerden[4])

We can go further down this road, because we can think of the transform-to-ranks as mapping our original data onto a uniform distribution. Why not map it onto other distributions – most obviously the normal distribution? This is in fact exactly what a van der Waerden transformation does.

Data	Rank	Quantile	z
-14	1	0.167	-0.967
-5	2	0.333	-0.431
6	3	0.500	0.000
8	4	0.667	0.431
20	5	0.833	0.967

Compute quantiles (percentiles) from ranks using p=i/(n+1), and interpret as probability points (p) of a normal distribution to give z values using =NORMSINV(p) in Excel[5]. Use the z-values instead of raw data in Pearson’s, t-tests, etc. Note, Spearman’s, Mann-Whitney, etc is equivalent to using the quantiles in Pearson’s, t-tests, etc. Note p → z reverses the probit z → p, mentioned above.

[2] Other powers are possible: the smaller the power the more ‘unbending’ is done, with zero effectively a log transform. . A generalisation of these is the Box-Cox transformation: $(x^k – 1) / k$. The researcher (or computer) chooses k to transform the data to achieve some desirable quality, say normality. When k=.5 it is the square root transform; when k→0, it is log; when k=1 it is linear (i.e., no transformation); k < 0 is also possible.

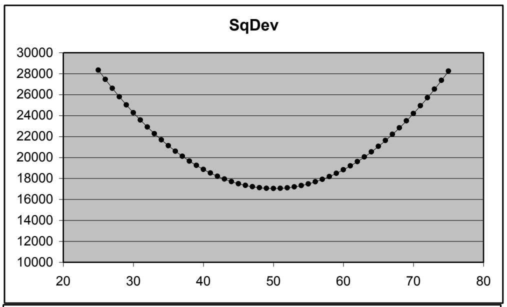
[3] See Conover, W.J. & Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, 35(3), 124–129.

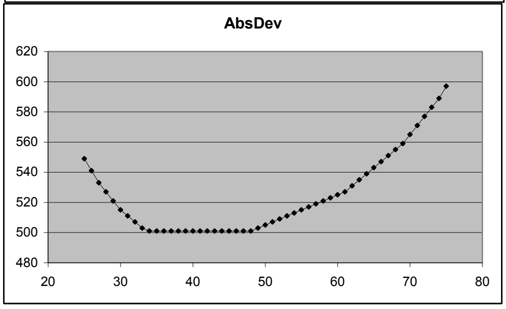
[4] See Wikipedia.

[5] An alternative formula due to Tukey is p=(i–⅓)/(n+⅓), giving quantiles .125, .3125, .5, .6875, and .875, with corresponding z values of -1.150, -.489, 0, +.489, and 1.150. This is a small n issue. For large values of n these is no practical difference.

L1, L2, and Lk measures.

Xi	
2	
13	
20	
20	
21	
27	
30	
33	
34	
48	
61	
69	
74	
79	
84	
91	
97	
98	
50.06	**Mean**
41	**Median**





The **Mean** is the value of x that *minimises* the **sum of the squared deviations** of all x_i from that value. Sometimes referred to the least squares criterion, or **L2 criterion**.

The **Median** is the value (are the values) of x that *minimises* the **sum of the absolute deviations** of all x_i from that value. Sometimes referred to as the absolute deviation criterion, or **L1 criterion**. Note that this criterion (sum of absolute deviations) is flat across the middle two observations, 34 and 48. The median is usually taken to be located at the middle of the flat portion (i.e., the average of the middle two observations).

It is possible to construct measures of location which minimise $S_k = \sum |x_i - \bar{x}|^k$ for a general k, with k=1 and k=2 being the special case of the median and mean. 1<k<2 will be a compromise between mean and median. You can do this using the Excel Solver. First, make a guess about where your new measure of location will be. Then calculate S_k. Now do *Tools / Solver* and *Minimize* the cell that S_k is in *By Changing* the cell that your initial guess is in. The solution will be in that same cell. First of all, try it for k=2, and check it does give you the mean.

Trimming

10% trimming removes the upper 10% and lower 10% of observations and constructs a mean from those that remain. 0% trimming is the usual mean, and as trimming approaches 50%, we approximate the median. 25% trimming has a special name – the midmean. Therefore, an m% trimmed mean is another kind of compromise between mean and median.

Percentiles (Quantiles)

The lower quartile, median, and upper quartile are the 25th, 50th, and 75th percentiles, respectively. In Excel: =percentile(a1:a30,0.57) for the 57th percentile for the numbers in a1:a30.

Weighted means

A weighted mean is defined to be:

$$\bar{x}_{weighted} = \sum_{i=1}^{n} w_i x_i$$

Where w_i are the weights applied to corresponding x_i. For the usual mean, $w_{i\,=\,1}$ for all i. It might be that certain observations deserve more weight. Suppose I have the mean salary of EU countries, and want to form an idea of the mean salary of the EU itself, wouldn't it make sense to weight Germany more than Luxembourg?

Iteratively re-weighted means

With an iteratively re-weighted mean we:

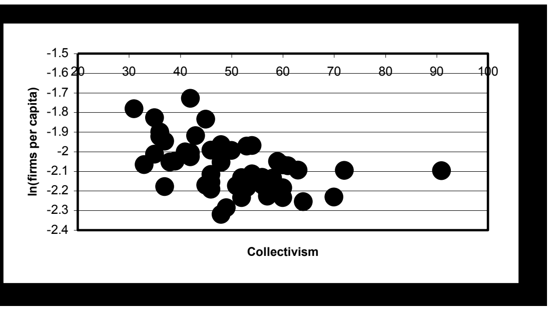
1. Start with a guess about where the measure of location lies (typically the median is used to start).
2. Then observations closest to that point are given greatest weight and distant points are given increasingly less weight (to a point where zero weight is given to what might therefore be defined as outliers.
3. A weighted mean is taken. Then back to step 2.

Repeat steps 2 and 3 until the weighted mean in step 3 is as close to that in step 2 as required.

This is advanced stuff.

Synopsis of tests *(needs verbal explanation)*

State	Collect-ivism	Ln(Firms per capita)	Categ. (collect)	Catec. t(firms)
Hawaii	91	-2.097	1	0
Louisiana	72	-2.096	1	0
South Carolina	70	-2.231	1	0
Mississippi	64	-2.255	1	0
Maryland	63	-2.095	1	1
Utah	61	-2.072	1	1
California	60	-2.067	1	1
Georgia	60	-2.185	1	0
Virginia	60	-2.235	1	0
New Jersey	59	-2.050	1	1
Texas	58	-2.136	1	0
Alabama	57	-2.227	1	0
Indiana	57	-2.209	1	0
North Carolina	56	-2.169	1	0
Tennessee	56	-2.131	1	0
Delaware	55	-2.131	1	0
Arkansas	54	-2.113	1	0
Florida	54	-1.969	1	1
Kentucky	53	-2.183	1	0
New York	53	-1.972	1	1
Illinois	52	-2.132	1	0
Nevada	52	-2.235	1	0
Pennsylvania	52	-2.182	1	0
New Mexico	51	-2.175	1	0
Connecticut	50	-1.994	1	1
Arizona	49	-2.286	0	0
Alaska	48	-1.964	0	1
Rhode Island	48	-2.054	0	1
West Virginia	48	-2.320	0	0
Massachusetts	46	-1.992	0	1
Michigan	46	-2.157	0	0
Missouri	46	-2.116	0	0
Wisconsin	46	-2.192	0	0
Maine	45	-1.835	0	1
Ohio	45	-2.172	0	0
New Hampshire	43	-1.919	0	1
Idahoe	42	-2.005	0	1
Oklahoma	42	-2.026	0	1
Vermont	42	-1.729	0	1
Minnesota	41	-2.001	0	1
Iowa	39	-2.049	0	1
Kansas	38	-2.054	0	1
North Dakota	37	-1.947	0	1
Washington	37	-2.179	0	0
Colorado	36	-1.899	0	1
South Dakota	36	-1.924	0	1
Nebraska	35	-2.013	0	1
Wyoming	35	-1.828	0	1
Oregon	33	-2.067	0	1
Montana	31	-1.781	0	1
medians	**49.5**	**-2.096**		
skew	0.925	0.586		
kurtosis	2.168	0.072		



		Collectivism		
log		**Interval**	**Ordinal**	**Categor.**
Firms	**Interval**	R	xxxxxx	t
per	**Ordinal**	xxxxxx	rho	MW
cap	**Categor.**	t	MW	Chi2

		Correlation	p
Pearson's	R	-0.474035	0.000506
Spearman'	rho	-0.537774	0.000056
Kendall's	tau	-0.360582	0.000253

Chi2: Analyse / Descriptive Statistics / Crosstabs

Observed	7	18	25
	18	7	25
	25	25	50

Expected by H0	12.5	12.5
	12.5	12.5

(O-E)^2/E	2.42	2.42
	2.42	2.42

Chi2 (without cc) = 9.68

p 0.001863

phi -0.44

Chi2 with continuity correction (only for 2x

(|O-E|-0.5)^2/E

2	2
2	2

Chi2 (**with** cc) = 8

p 0.004678

	1-tailed	2-tailed
Fisher Exact Test	0.004199	0.0021

Mann-Whitney U test

	U	**p**
categorise by firmsize	135.5	0.000403
categorise by collectivisn	153	0.001613

Independent Samples t-Test (categorised 0/1 by firmsize)

Levene's Test (H0: Vars are equal)			t-test (H0: means eq)		Sig (2-tailed)
F	Sig.		t	df	p
0.042886	0.836817	var eq?	3.802691	48	0.000404
		var ne?	3.802691	47.1	0.000411

95% C I = 5.09 to 16.51

95% CI = .045 to .182

Indep samples t-test categorised by collectivis

		Levene's Test for Equality of Varian		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
ln(Firms p.cap)	Equal variances assumed	5.549	0.023	3.34	48	0.001617
	Equal variances not assumed			3.34	37.43	0.001896

STATISTICAL FORMULAE

(N.B. this is not intended to be a complete, or even required, list)

Definitions:

$\bar{x}$ sample mean
s sample standard deviation
μ population mean
σ population standard deviation

Mean of n numbers x_i $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Variance of n numbers x_i $\quad$ **Var(x)** $= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

It's the mean squared deviation (from the mean)

Covariance of n numbers x_i $\quad$ **Cov(x,y)** $= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Standard deviation of n numbers x_i $\quad$ **s** $= \sqrt{Var(x)}$ (also known as $\mathbf{s_n}$)

Unbiased estimate of population variance $\mathbf{s_{n-1}^{2}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Unbiased estimate of population std. dev. is $\mathbf{s_{n-1}}$ (i.e. the square root of the above)

Pearson's product-moment correlation $\quad$ **R** $= \frac{Cov(x,)}{\sqrt{Var(x).Var(y)}}$

Regression line is $\quad y = a + bx$

a is the constant term (also known as the intercept)
b is the gradient of the line (also known as the regression coefficient for x)

$$\mathbf{b} = \frac{Cov(x,y)}{Var(x)}$$

It is known that the OLS line of best fit goes through the point $(\bar{x}, \bar{y})$
Therefore, *a* can be obtained from the equation: $\quad \bar{y} = a + b\,\bar{x}$

In a set of numbers x_i sorted in ascending order $x_{(i)}$:

Median (M) is $x_{\left(\frac{n+1}{2}\right)}$ $\quad$ the number at rank (n+1)/2

Lower quartile (Q1) is $x_{\left(\frac{n+1}{4}\right)}$ $\quad$ the number at rank (n+1)/4

Upper quartile (Q3) is the number found at rank 3(n+1)/4

Range = max – min = $x_{(n)} - x_{(1)}$

Inter-quartile range = Q3 - Q1

Bowley's (Interquartile) coefficient of skew is (Q3 + Q1 - 2M) / (Q3 - Q1)

Pearson's coefficient of skew is 3 (mean - median) / (standard deviation)

Moments: $\mathbf{m_k} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$

Moment skewness = $\frac{m_3}{m_2^{1.5}}$

Excess kurtosis = $\frac{m_4}{m_2^{2}} - 3$

Spearman's correlation coefficient ('Rho') for n pairs of ranks (a_i, b_i) is given by:

$$\rho = 1 - \frac{6\ \Sigma\ D_i^2}{(n+1)n(n-1)}$$

where D_i is the difference in rank between a_i and b_i. May also be found as the Pearson's correlation coefficient computed on ranked data.

A value x from a population with mean μ and standard deviation σ can be expressed as a **normal deviate z** (how many standard deviations it lies from the mean) by the following transformation:

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

The standard deviation of the *sample mean* is: $s_{\bar{x}} = \frac{s_{n-1}}{\sqrt{n}}$ and is also known as the standard error.

Where the distribution of the sample mean can be approximated by the normal distribution (usually when n ≥ 30), the **95% confidence interval** for the estimate of the population mean is:

$$\mu = \bar{x} \pm 1.96 s_{\bar{x}}$$

Replace 1.96 by 2.58 for 99% confidence intervals.
Replace 1.96 by the appropriate value in the t-table (d.f. = n-1) when n < 30.

The **finite population correction** to the standard error is:

$$\sqrt{\frac{(N-n)}{(N-1)}}$$

(Apply if n > N / 20 where n is the sample size, and N the population size).

t-tests

Comparing the difference between means of two independent groups,

$\overline{X} = (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$.   Usually, $\mu_1 = \mu_2$

$s^2 = \frac{s_1^{\ 2}}{n_1} + \frac{s_2^{\ 2}}{n_2}$ where s is the s.d. for the difference in means X

When n_1 and n_2 are not big enough to use the normal approximation, and so t must be used, the degrees of freedom are derived from a complicated formula, but are usually close to

$(n_1 + n_2 - 2)$.

The probability of x occurrences for a **Poisson's distribution** with population mean of μ is given by:

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}$$

where e ≈ 2.71828... (a special π-like mathematical number), known as the exponential number, or usually just 'e'; and x! = x (x-1) (x-2)....2 1, is known as factorial x. [N.B. 1! = 1, and 0! = 1 also!!!!]. The standard deviation of a Poisson distribution is always $\sigma = \sqrt{\mu}$.

The **Binomial distribution** for x occurrences from n *independent* events, each of probability p (with q = 1 – p), is given by:

$$p(x) = {}^nC_x \, p^x q^{n-x}$$

where **${}^n\mathbf{C}_x$** = $\frac{n!}{(n-x)!x!}$ is the number of combinations of x from n.

For large n (≥ 20), np ≥ 5, and nq ≥ 5, use normal approximation: $\mu = np$; $\sigma = \sqrt{npq}$
Perhaps also use *continuity correction* of 0.5 in numerator, to reduce size of z.

Sign test

Exact tables for small n. Otherwise use normal approximation to binomial, as above.

Wilcoxon's Signed Rank test (for matched pairs)

Exact tables for small n. For n ≥ 10, use normal approximation: $z = \frac{\sum R_i}{\sqrt{\sum R_i^{\ 2}}}$

Where R_i are signed ranks, with biggest unsigned difference getting largest numerical rank.

Mann-Whitney U-test (Wilcoxon's Rank Sum test)

Exact tables for small n1 & n2. For larger n (either n1 or n2 > 20) use normal approximation:
T_j is the total sum of the ranks for j.

U = Smaller of { (T_1 – (n_1 (n_1+1) / 2), and T2 – (T_1 – (n_2 (n_2+1) / 2) }
$\mu = n_1 n_2 / 2$
$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

Then $z = \dfrac{U + 0.5 - \mu}{\sigma}$

Runs test

0001110011111 would be 4 runs. For small numbers of events, exact tables exist. For larger numbers, use normal approximation. If there are a (e.g. a = 5 zeros) of one kind and b (e.g. = 8 ones) of another in the sample, then:

$\mu = 1 + \dfrac{2ab}{(a+b)}$ the mean number of runs expected from a & b

$\sigma^2 = \dfrac{2ab(2ab - a - b)}{(a+b)^2(a+b-1)}$ the variance

$z = (\text{runs} - \mu + C) / \sigma$ where C (continuity correction) is +0.5 if runs < μ
and -0.5 if runs > μ

N.B. this test detects unexpectedly many runs (runs > μ) as well as unexpectedly few (runs < μ).

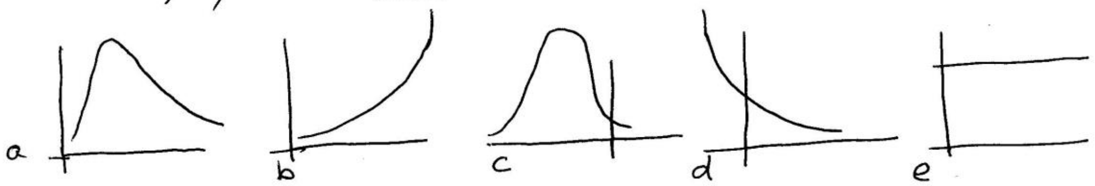
Chi-square test for r x c contingency table

$\chi^2 = \sum_{i=1}^{n} \dfrac{(O_i - E_i)^2}{E_i}$ with (rows-1)(cols-1) degrees of freedom. N.B. n = rows x cols.

Required sample size $n = \dfrac{z^2\, s^2}{e^2}$ where e is the desired margin of error

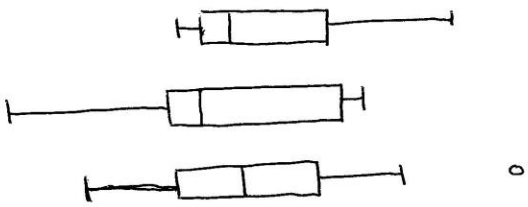
Stats Quiz

1. Below are 5 distributions. Write against each whether it has +, −, or zero skew



2. Place a to e in order (− to +) of how skewed they are.

3. Consider the three boxplots. Estimate whether you think each of Bowley's, Pearsons, and the moment measures of skew will give +, −, or zero skew. Ans. in table.



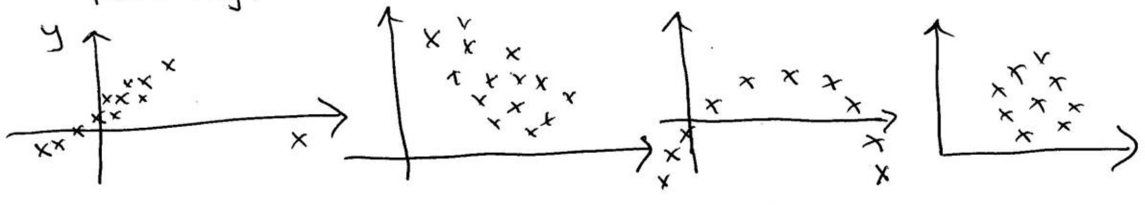
Bowley's	Pearson's	Moment	
			(a)
			(b)
			(c)

4. (In 3) Which of a, b, c has the largest Q_1, Q_3, median? which has the largest IQR?

5. Which of the following would you perform on the distributions in (1) to try to remove skew? Please tick in boxes.

	a	b	c	d	e
log $\ln(x)$					
sq. root $\sqrt{x}$					
none x					
square x^2					
exponentiate e^x					

6. ~~Guess~~ Estimate what the correlation is between x and y in the following:



7. You wish to use regression to estimate the number of Googled websites that mention a new brand "J-Stats", from its launch to one month later (readings taken daily). The idea is to estimate the number of mentions in the future. What is a suitable y to use? What is a suitable x? In the standard regression equation $y = a + bx$, what are the units of measurement for a, & b? Do we expect a, b to be $-$, 0, or $+$?

8. Estimating fuel consumption (m.p.g.) from the weight of the car (Kg), what are the answers to Q7 here? Assume you have a number of cars in your survey.

9. Can R and b have opposite signs?

10. $x_{(i)}$ are a set of n numbers that have been ordered from smallest $x_{(1)}$ to largest $x_{(n)}$. Write down formulae for:

a) median
b) Q_1
c) Q_3
d) IQR
e) Range
f) mean
g) Mean Absolute Deviation (from Mean)
h) standard deviation

(The un-ordered numbers are noted by x_i)

11. Without using a calculator, if you take logs of the following numbers, is the result $+$, $-$, or 0?

a) 5
b) ·5
c) 1
d) −1
e) 0

12. The ratio of men to women in the ~~five~~ six departments of a company are: 1·1, 1·2, 0·9, 0·4, 2·0, 2·4. The company wishes to maintain as close to a 1-to-1 ratio of men to women throughout its departments. Which is the department that is most imbalanced?

13. Take logs of the ratios in (12)

14. Hum quietly.

Hypotheses

Please tick all the following statements that are true, and put a cross besides those that are false.

1. You can never KNOW that you have made a Type I error.
2. A Type I error is being misled by an unusual (fluky) sample of data into believing in some relationship where none really exists.
3. Type I errors can usually be eliminate by good research design.
4. Type II errors occur when your data does not show a pattern that really exists in the population.
5. To determine the probability of a 2-tailed test, simply halve the probability of the 1-tailed test.
6. When some values x_i have been transformed to <u>standardized</u> z_i scores, which of the following are true?

   $z = 1$; $z = 0$; $z = x$; $s_z = s_x$; $s_z = 1$; $s_z = 0$.
7. About 5% of a normal distribution lies at or above 2 standard deviations from the mean.
8. (7) could be written more algebraically as: $p\ (z_i \geq 2) = .05$.
9. The hypothesis: “Female PhD students complete their theses faster than Males” should be tested using a 2-tailed test.
10. Write a null hypothesis (H_0) for the hypothesis in (9):

    ____
11. If someone’s hypothesis is tested and the result is “significant at the 1% level”, then:
    (a) The probability that they are making a Type I error is 1%.
    (b) The probability that they are making a Type II error is 1%.
    (c) If the research were repeated very many times but on new samples, on 1% of occasions the researcher would find an equally extreme pattern of results.
    (d) If the research were repeated very many times but on new samples, on 1% of occasions the researcher would find an equally extreme pattern of results, if H_0 were true.
    (e) If the research were repeated very many times but on new samples, on 1% of occasions the researcher would find an equally extreme pattern of results, if H_0 were false.
    (f) The researcher can be 99% confident that his/her hypothesis is correct.
    (g) If the researcher repeats his/her study many times, he/she can expect that 99% of them will also give significant results.
12. Rejecting H_0 is the same as saying “my results are unlikely to have occurred by chance.”
13. We never directly test a hypothesis, instead we test the null hypothesis.
14. Testing hypotheses establishes facts, not theories.

Questions about t-tests and comparing means

1. A survey of 49 people chosen at random in a particular city shows that on average people travel 2.3 miles to work, with a standard deviation of 1.4 miles[6]. Try to sketch a plausible histogram, showing frequencies for different distances travelled.
2. What are the 90% confidence limits on this average? What does this 90% confidence limit mean? What are the 95% limits, and the 99% limits? What assumptions about the distribution of the average (=mean).
3. If 4900 people were surveyed, what would happen to each of these (90%, 95%, 99%) limits?
4. If I wanted to be 95% sure that the *margin of error* on my estimate of the average distance travelled was no more than 0.1 miles, how many people would I need to survey? How many would I need to survey in order to be 99% confident?
5. A newsagent takes a sample of 36 of its 2000 customers, and determines that *on average* they spent £0.20p more in this week than in the same week last year. The standard deviation of the increased amount spent is £0.90. Calculate the 95% confidence interval for the average increase in spend. Can the newsagent conclude from this that average spend has gone up?
6. Suppose it had been Sarah who had sampled her own paper-round of 60 customers. (Otherwise same figures as above). Calculate the 95% confidence interval for the average increase in spend. What can she conclude?
7. Danny's paper-round has increased its average spend in the same week by £0.10p. This figure has a standard deviation of £0.70p. He has sampled 30 of his 72 customers to reach this figure. Calculate the 95% confidence interval for the average increase in spend.
8. Repeat the analysis in (2), presuming that Sarah had only sampled 16 of her 60 customers. What are you assuming about the distribution of increased spend.
9. Biff-Bang Boots has been testing out its revolutionary new boots on a randomly selected group of 12 amateur footballers in the local area. Each player wears the boot for half the games he plays in. Over the course of a season Biff-Bang Boots observes the following number of goals scored to the 12 players: (3,1), (3,5), (5,2), (6,4), (4,5), (1,1), (7,2), (12,10), (2,1), (6,2), (3,0), (9,6), where (n,m) indicates n goals scored with, and m scored without the magic boot, respectively. What can BBB conclude?
10. With what confidence could we say that average spend has gone up more on Sarah's round than on Danny's?
11. Dr Pinkerton-Smythe is well pleased with herself. Her remarkable new method of treating disease X has resulted in an average stay in hospital of only 10.5 days, whereas Professor Tablet-Gulper's method has resulted in an average stay of 12.3 days. What other (statistical) information should Dr Pinkerton-Smythe know before putting it about that her method is better than Professor Tablet-Gulper's? Invent the missing data for yourself, and decide whether she should "go public".
12. Repeat 9 as if the two groups were independent. Note the loss of power.

[6] (using the 1/(n-1) rather than the 1/n method to calculate variance).

Normal Distribution

z-score	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.5	0.000233	0.000224	0.000216	0.000208	0.000200	0.000193	0.000185	0.000178	0.000172	0.000165
-3.4	0.000337	0.000325	0.000313	0.000302	0.000291	0.000280	0.000270	0.000260	0.000251	0.000242
-3.3	0.000483	0.000466	0.000450	0.000434	0.000419	0.000404	0.000390	0.000376	0.000362	0.000349
-3.2	0.000687	0.000664	0.000641	0.000619	0.000598	0.000577	0.000557	0.000538	0.000519	0.000501
-3.1	0.000968	0.000935	0.000904	0.000874	0.000845	0.000816	0.000789	0.000762	0.000736	0.000711
-3	0.001350	0.001306	0.001264	0.001223	0.001183	0.001144	0.001107	0.001070	0.001035	0.001001
-2.9	0.001866	0.001807	0.001750	0.001695	0.001641	0.001589	0.001538	0.001489	0.001441	0.001395
-2.8	0.002555	0.002477	0.002401	0.002327	0.002256	0.002186	0.002118	0.002052	0.001988	0.001926
-2.7	0.003467	0.003364	0.003264	0.003167	0.003072	0.002980	0.002890	0.002803	0.002718	0.002635
-2.6	0.004661	0.004527	0.004396	0.004269	0.004145	0.004025	0.003907	0.003793	0.003681	0.003573
-2.5	0.006210	0.006037	0.005868	0.005703	0.005543	0.005386	0.005234	0.005085	0.004940	0.004799
-2.4	0.008198	0.007976	0.007760	0.007549	0.007344	0.007143	0.006947	0.006756	0.006569	0.006387
-2.3	0.010724	0.010444	0.010170	0.009903	0.009642	0.009387	0.009137	0.008894	0.008656	0.008424
-2.2	0.013903	0.013553	0.013209	0.012874	0.012545	0.012224	0.011911	0.011604	0.011304	0.011011
-2.1	0.017864	0.017429	0.017003	0.016586	0.016177	0.015778	0.015386	0.015003	0.014629	0.014262
-2	0.022750	0.022216	0.021692	0.021178	0.020675	0.020182	0.019699	0.019226	0.018763	0.018309
-1.9	0.028717	0.028067	0.027429	0.026803	0.026190	0.025588	0.024998	0.024419	0.023852	0.023295
-1.8	0.035930	0.035148	0.034380	0.033625	0.032884	0.032157	0.031443	0.030742	0.030054	0.029379
-1.7	0.044565	0.043633	0.042716	0.041815	0.040930	0.040059	0.039204	0.038364	0.037538	0.036727
-1.6	0.054799	0.053699	0.052616	0.051551	0.050503	0.049471	0.048457	0.047460	0.046479	0.045514
-1.5	0.066807	0.065522	0.064255	0.063008	0.061780	0.060571	0.059380	0.058208	0.057053	0.055917
-1.4	0.080757	0.079270	0.077804	0.076359	0.074934	0.073529	0.072145	0.070781	0.069437	0.068112
-1.3	0.096800	0.095098	0.093418	0.091759	0.090123	0.088508	0.086915	0.085343	0.083793	0.082264
-1.2	0.115070	0.113139	0.111232	0.109349	0.107488	0.105650	0.103835	0.102042	0.100273	0.098525
-1.1	0.135666	0.133500	0.131357	0.129238	0.127143	0.125072	0.123024	0.121000	0.119000	0.117023
-1	0.158655	0.156248	0.153864	0.151505	0.149170	0.146859	0.144572	0.142310	0.140071	0.137857
-0.9	0.184060	0.181411	0.178786	0.176186	0.173609	0.171056	0.168528	0.166023	0.163543	0.161087
-0.8	0.211855	0.208970	0.206108	0.203269	0.200454	0.197663	0.194895	0.192150	0.189430	0.186733
-0.7	0.241964	0.238852	0.235762	0.232695	0.229650	0.226627	0.223627	0.220650	0.217695	0.214764
-0.6	0.274253	0.270931	0.267629	0.264347	0.261086	0.257846	0.254627	0.251429	0.248252	0.245097
-0.5	0.308538	0.305026	0.301532	0.298056	0.294599	0.291160	0.287740	0.284339	0.280957	0.277595
-0.4	0.344578	0.340903	0.337243	0.333598	0.329969	0.326355	0.322758	0.319178	0.315614	0.312067
-0.3	0.382089	0.378280	0.374484	0.370700	0.366928	0.363169	0.359424	0.355691	0.351973	0.348268
-0.2	0.420740	0.416834	0.412936	0.409046	0.405165	0.401294	0.397432	0.393580	0.389739	0.385908
-0.1	0.460172	0.456205	0.452242	0.448283	0.444330	0.440382	0.436441	0.432505	0.428576	0.424655
0	0.500000	0.496011	0.492022	0.488034	0.484047	0.480061	0.476078	0.472097	0.468119	0.464144
0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965620	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999534	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999651
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999822	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967
4	0.999968	0.999970	0.999971	0.999972	0.999973	0.999974	0.999975	0.999976	0.999977	0.999978

t-distribution	=tinv(t,df,tails)		Probability level	(2-tailed test)				
	0.2	0.1	0.05	0.02	0.01	0.005	0.001	0.0001
d.f.								
1	3.08	6.31	12.71	31.82	63.66	127.32	636.62	6366.2
2	1.89	2.92	4.30	6.96	9.92	14.09	31.60	99.99
3	1.64	2.35	3.18	4.54	5.84	7.45	12.92	28.00
4	1.53	2.13	2.78	3.75	4.60	5.60	8.61	15.54
5	1.48	2.02	2.57	3.36	4.03	4.77	6.87	11.18
6	1.44	1.94	2.45	3.14	3.71	4.32	5.96	9.08
7	1.41	1.89	2.36	3.00	3.50	4.03	5.41	7.88
8	1.40	1.86	2.31	2.90	3.36	3.83	5.04	7.12
9	1.38	1.83	2.26	2.82	3.25	3.69	4.78	6.59
10	1.37	1.81	2.23	2.76	3.17	3.58	4.59	6.21
11	1.36	1.80	2.20	2.72	3.11	3.50	4.44	5.92
12	1.36	1.78	2.18	2.68	3.05	3.43	4.32	5.69
13	1.35	1.77	2.16	2.65	3.01	3.37	4.22	5.51
14	1.35	1.76	2.14	2.62	2.98	3.33	4.14	5.36
15	1.34	1.75	2.13	2.60	2.95	3.29	4.07	5.24
16	1.34	1.75	2.12	2.58	2.92	3.25	4.01	5.13
17	1.33	1.74	2.11	2.57	2.90	3.22	3.97	5.04
18	1.33	1.73	2.10	2.55	2.88	3.20	3.92	4.97
19	1.33	1.73	2.09	2.54	2.86	3.17	3.88	4.90
20	1.33	1.72	2.09	2.53	2.85	3.15	3.85	4.84
21	1.32	1.72	2.08	2.52	2.83	3.14	3.82	4.78
22	1.32	1.72	2.07	2.51	2.82	3.12	3.79	4.74
23	1.32	1.71	2.07	2.50	2.81	3.10	3.77	4.69
24	1.32	1.71	2.06	2.49	2.80	3.09	3.75	4.65
25	1.32	1.71	2.06	2.49	2.79	3.08	3.73	4.62
26	1.31	1.71	2.06	2.48	2.78	3.07	3.71	4.59
27	1.31	1.70	2.05	2.47	2.77	3.06	3.69	4.56
28	1.31	1.70	2.05	2.47	2.76	3.05	3.67	4.53
29	1.31	1.70	2.05	2.46	2.76	3.04	3.66	4.51
30	1.31	1.70	2.04	2.46	2.75	3.03	3.65	4.48
35	1.31	1.69	2.03	2.44	2.72	3.00	3.59	4.39
40	1.30	1.68	2.02	2.42	2.70	2.97	3.55	4.32
45	1.30	1.68	2.01	2.41	2.69	2.95	3.52	4.27
50	1.30	1.68	2.01	2.40	2.68	2.94	3.50	4.23
60	1.30	1.67	2.00	2.39	2.66	2.91	3.46	4.17
70	1.29	1.67	1.99	2.38	2.65	2.90	3.44	4.13
80	1.29	1.66	1.99	2.37	2.64	2.89	3.42	4.10
90	1.29	1.66	1.99	2.37	2.63	2.88	3.40	4.07
100	1.29	1.66	1.98	2.36	2.63	2.87	3.39	4.05
200	1.29	1.65	1.97	2.35	2.60	2.84	3.34	3.97
300	1.28	1.65	1.97	2.34	2.59	2.83	3.32	3.94
400	1.28	1.65	1.97	2.34	2.59	2.82	3.32	3.93
500	1.28	1.65	1.96	2.33	2.59	2.82	3.31	3.92
1000	1.28	1.65	1.96	2.33	2.58	2.81	3.30	3.91
Normal distr.	1.28	1.64	1.96	2.33	2.58	2.81	3.29	3.89

Pearson's R (2-tailed)		
N	0.05	0.01
3	0.9969	0.99988
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561
21	0.433	0.549
22	0.423	0.537
23	0.413	0.526
24	0.404	0.515
25	0.396	0.505
26	0.388	0.496
27	0.381	0.487
28	0.374	0.479
29	0.367	0.471
30	0.361	0.463
35	0.335	0.433
40	0.314	0.407
45	0.296	0.384
50	0.280	0.364
60	0.256	0.333
70	0.237	0.309
80	0.221	0.288
90	0.208	0.272
100	0.197	0.258
200	0.140	0.184
300	0.114	0.150
400	0.098	0.129
500	0.088	0.115
1000	0.0621	0.0817

Chi-squared =chiinv(X,df)								
	0.99	0.95	0.9	0.5	0.1	0.05	0.01	0.001
d.f.								
1	0.00016	0.0039	0.0158	0.45	2.71	3.84	6.63	10.83
2	0.0201	0.1026	0.2107	1.39	4.61	5.99	9.21	13.82
3	0.115	0.352	0.584	2.37	6.25	7.81	11.34	16.27
4	0.297	0.711	1.06	3.36	7.78	9.49	13.28	18.47
5	0.554	1.15	1.61	4.35	9.24	11.07	15.09	20.52
6	0.872	1.64	2.20	5.35	10.64	12.59	16.81	22.46
7	1.24	2.17	2.83	6.35	12.02	14.07	18.48	24.32
8	1.65	2.73	3.49	7.34	13.36	15.51	20.09	26.12
9	2.09	3.33	4.17	8.34	14.68	16.92	21.67	27.88
10	2.56	3.94	4.87	9.34	15.99	18.31	23.21	29.59

Skew and Kurtosis

Unlike measure of location and dispersion, measures of Skew and Kurtosis have no units of measurement. In this sense they are like correlation coefficients. Skew and kurtosis summarise something about the general shape of the curve, as Figures 1 and 2 illustrate. Actually, it is quite hard to spot kurtosis by eye. If a distribution has positive skew, the positive tail of the right-hand tail of the distribution is extended or stretched a long way into or towards the positive numbers. If it has negative skew, the left-hand tail of the distribution is extended towards the negative numbers. If it has no skew, the distribution is symmetrical.

If a distribution is symmetrical, the distribution may yet extend further into *both* tails than a normal distribution would. Such a distribution is described as having fat tails, or positive kurtosis. If the distribution does not extend as far as a normal distribution would into the tails, then it will have negative kurtosis. A *uniform* distribution has negative kurtosis. Having positive kurtosis usually causes more problems for analysis than having negative kurtosis.

Here are three measures of skew, and one for kurtosis.

Pearson's Coefficient of Skew.

Consider the mean, median, and mode of the distribution in Figure 1. If we move through the distribution from low to high numbers (left to right), the first measure of location we expect to meet is the mode, then the median, and finally the mean, which is dragged to the right by large positive numbers. If the distribution is skewed the other way, we meet the mean first, then the median, and lastly, the mode. A rule of thumb is that the more skewed a set of data, the greater the separation between the mean and the mode. In theory, then, (mean – mode) / standard deviation could be used as a measure of skew. Dividing by the standard deviation (s) is a way of scaling the data so that the difference in dispersion from sample to sample has been "divided out". Also, it has no units of measurement. Can you see why?

However, we already know that the mode is not much practical use because of its instability. What Pearson knew was that for a wide variety of distributions:

mean – mode ≈ 3 (mean – median) [≈ means approximately equal]

So, instead Pearson used:

3 (mean – median) / s

as his measure of skew. Unfortunately, the greater the skew, the greater the separation between mean and median (GOOD) - but also the greater the standard deviation s (BAD). The increasing bottom of the equation acts to mask the increasing top of the equation. [The data in Table 1 have a Pearson's skew = -1.061.]

Quartile Measure of Skew (Bowley's)

An alternative measure of skew and be understood by inspecting the box of a boxplot. If there is positive skew (Q3 – Median) will be greater than (Median – Q1). How much greater it is can be expressed as the IQR. So, we get:

Quartile Measure of Skew = [(Q3 – Med) – (Med – Q1)] / (Q3 – Q1)

= (Q3 + Q1 – 2 Med) / (Q3 – Q1)

It varies between +1 for maximum positive skew (when the Median and Q3 are at the same point), to –1 for maximum negative skew (when the Median and Q1 are at the same point). Note that the quartile measure of skew does not use the upper 25% of data, nor the lower 25% of data.

Moment measure of skew

The kth moment of a distribution is defined to be:

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

In words, we take the deviations of each point from the mean, and then square them (k=2), or cube them (k=3), etc. Finally we average all these numbers. So the second moment (m_2) is what we have already met as the variance, and $s^2 = m_2$. Another measure of skew, closes to the one calculated by SPSS, and Excel is defined as:

$$\text{skew} = \frac{m_3}{(m_2)^{3/2}} = \frac{m_3}{s^3}$$

This measure works because if the numbers stretch out towards the negative end, then the deviations from the mean will be larger at the negative end. When these deviations get cubed, they retain the negative sign, but get even larger than the more numerous small positive deviations (see the final column in Table 1, where the -125 outweighs all the positive numbers).

Table 1. Calculating the moment measure of skew for five observations.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$
	1	-5	25	-125
	5	-1	1	-1
	7	1	1	1
	8	2	4	8
	9	3	9	27
Means	6	0	8	-18
			$(=m_2 = s^2)$	$(=m_3)$
			s^3	22.62742
			Skew =	-0.7955

Comparing the three measures of skew, the quartile measure of skew pays no attention to extreme values, because it is calculated only on the quartiles and median. The moment measure of skew pays most attention to the extremes of the distribution, as the above example illustrates, because deviations get cubed.

Kurtosis

Again, using moments, the moment measure of kurtosis is:

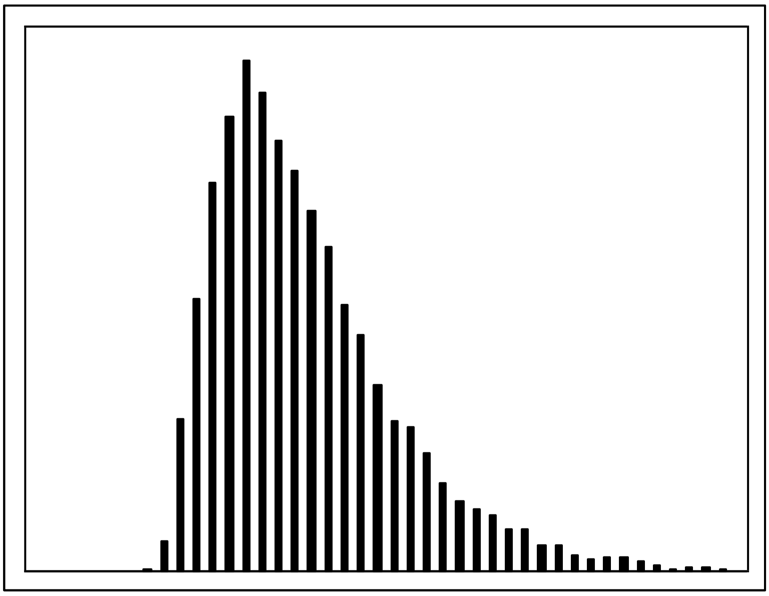
$$\frac{m_4}{(m_2)^2} = \frac{m_4}{s^4}$$

Since the normal distribution has a kurtosis of 3 measured in this way, many statistics packages routinely subtract 3 from the calculation to give:

$$\text{kurtosis} = \frac{m_4}{s^4} - 3$$

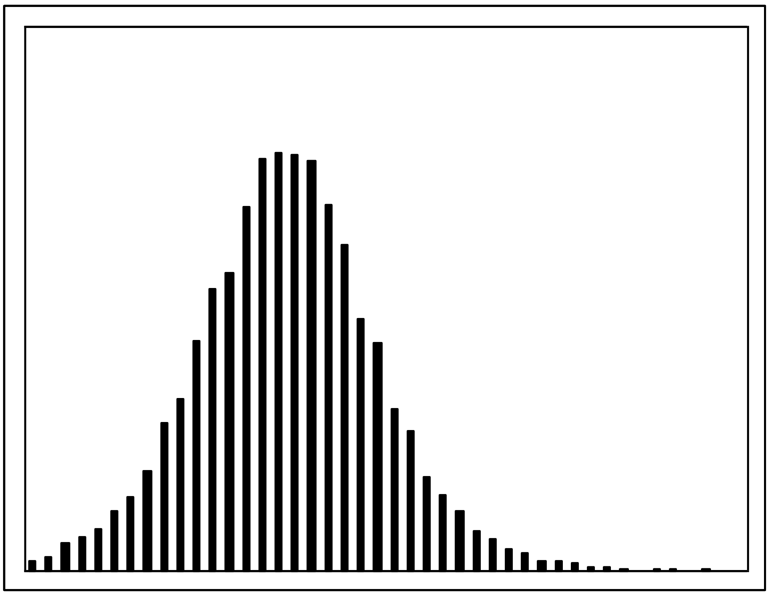
Subtracting the 3 leads to what is sometimes called "excess kurtosis." Be aware of which packages do (SPSS, SAS, Excel, Minitab, BMD), and which do not (Stata) subtract the 3. Furthermore, for various reasons, many statistical packages differ slightly from this simple version in their calculations of kurtosis.

Figure 1. Example of a distribution with positive skew.



Pearson's	0.654799
Quartile	
Skew	0.169934
Moments	1.750036
Kurtosis	6.010709

Figure 2. Example of a distribution with no skew, but positive kurtosis (fat tails).



Pearson's	-0.01614
Quartile	
Skew	-0.0199
Moments	-0.05111
Kurtosis	1.672773

Boxplot, and Stem-and-Leaf Diagrams

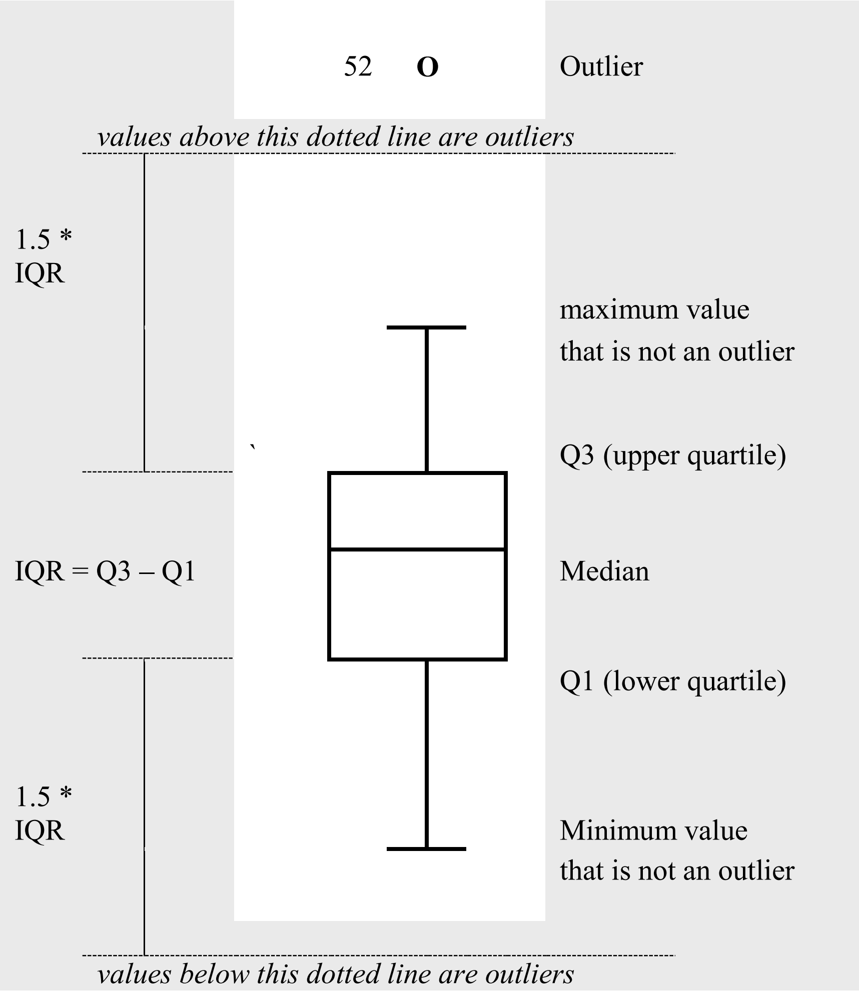
(The Figures may look better when printed, or using Print Preview)

Constructing a boxplot.

Knowing how to calculate a median and the quartiles, we can now use this information to construct a boxplot. It is a simple but useful diagram that tells us quite a bit about the distribution of our data. *In the diagram below (Figure 1), all comments that appear in the grey shaded background are my annotations for your benefit, and would not normally appear in a boxplot.*

First we draw the box in the middle. The top of the box represents the upper quartile (Q3), the bottom represents the lower quartile (Q1), and the line in the middle is the median. The box is supposed to represent the bulky middle part of a normal-like distribution, in contrast to the less dense tails. Recall that the IQR is a stable measure of dispersion, which is why we use it to help define how far away from the middle region is going to count as outlier country.

Figure 1. An annotated boxplot, showing that observation #52 is an outlier.


52 O
Outlier
values above this dotted line are outliers
1.5 * IQR
maximum value that is not an outlier
Q3 (upper quartile)
IQR = Q3 – Q1
Median
Q1 (lower quartile)
1.5 * IQR
Minimum value that is not an outlier
values below this dotted line are outliers

We define outliers as anything that lies *above* Q3 + (1.5 * IQR), or anything that lies *below* Q1 - *1.5 * IQR). These boundaries (the long dotted lines) are sometimes known as the upper and lower fences. Values in this region are noted with an **o** for outlier. SPSS also identifies which observation it is with a number, e.g. 52 for the fifty second number as it appears in your data (52 is not its value!). Now we know what an outlier is, we can calculate the maximum and minimum values that are not outliers. We draw so-called *whiskers*, extending from Q1 to the minimum that is not an outlier. Similarly we draw a line (whisker) from Q3 to the maximum that is not an outlier. The original name for a boxplot was a box-and-whiskers plot.

The multiplying factor of 1.5 is slightly arbitrary, but has been found by practice to strike the right balance between identifying too many and too few far-away values as outliers. SPSS also identifies extreme outliers with an **x** instead of an **o.** This is the usual convention. Extreme outliers are defined to be 3 * IQR beyond the upper and lower quartiles. If you ever find one of these in your data take a good long look at it before you proceed. In my opinion, being an *extreme outlier* would be reason enough to remove it from my analysis.

What do Normal boxplots look like?

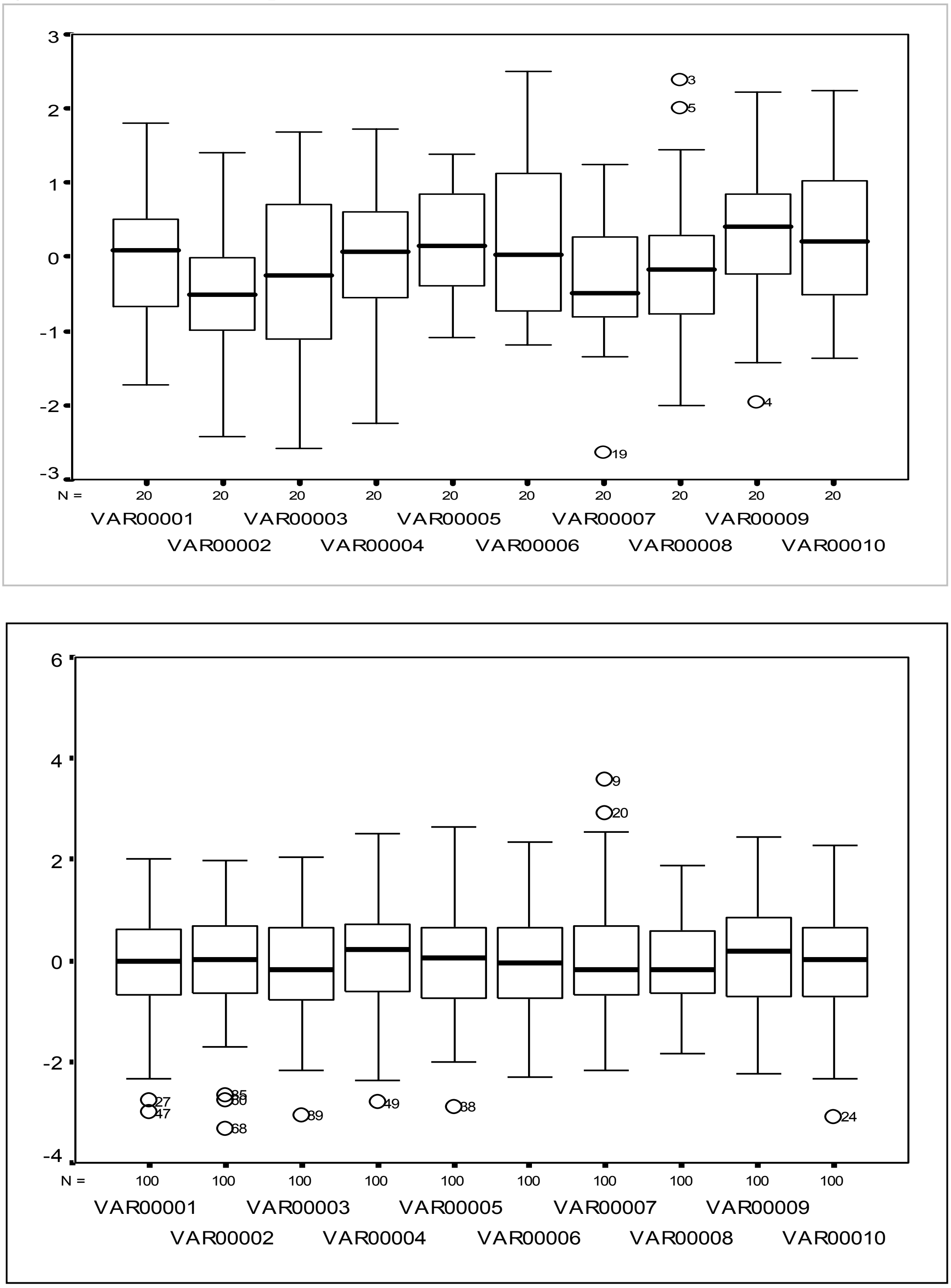
It is helpful to get an idea of what a usual boxplot looks like so that you can identify when you find one that is unusual. Eventually you may see enough of them so you develop your own feel for it, but for now, it is forced learning. In Figure 2 I have used a computer to generate random samples from a population that is Gaussian / normally distributed, having a mean of 0 and a standard deviation of 1. In the upper panel I have generated 10 samples, each sample consisting of 20 numbers. I have got SPSS to draw boxplots for each of these samples. In the lower panel I have done the same, but my samples each have 100 numbers in them. As you will see, there is quite a variety in the boxplots, more so when the samples are smaller (n=20). However, very approximately, we see that each of the whiskers is about one IQR in length. So the box and whiskers divide the distribution up into 3 approximately equal lengths.

We also see that outliers are not a once-in-a-lifetime occurrence. As we might expect, the more numbers there are in the sample (100 versus 20), the more outliers we unearth. But even with small samples, some values do get labelled as outliers, and even when we know they have been generated by a Gaussian / normal distribution. In fact, approximately 1 in 140 observations from a Gaussian distributions will be identified as an outlier by the criterion now standardly adopted by researchers (1.5 * IQR). This is a small price to pay.

Also, notice that the median of the samples does move around from the zero point, which is the population mean. This is more apparent for the smaller samples, as we would expect, because the larger the sample, the more like the entire population it should look. Thus, the n=100 samples not only look more like each other than the n=20 samples do, they also look more like what the population as a whole would look like.

Finally, boxplots are not much use for data that have been measured on a 5-point scale, say. For instance, if the majority reply “4” to a particular question, we may end up with Q1 = Median = Q3 = 4, in which case IQR = 0, and the upper and lower fences are both at 4, so everything that is not a “4” is classified as an extreme outlier. The boxplot was not intended for that kind of data.

Figure 2. BoxPlots of samples drawn from a normal distribution



In all cases the populations have a normal distribution, with mean = 0, and standard deviation = 1.
In the first picture, 10 different samples of size 20 (N=20) have been drawn from this population.
Similarly, in the second picture 10 samples have been drawn of size 100 (N=100).
By our rule of thumb OUTLIERS BEYOND 1.5 * IQR, about 1 in 140 observations would be outliers if the distribution were normal*. In picture 1 we have 200 observations and 4 outliers; in picture 2 we have 11 outliers from 1000 observations. So this isn't too far out of line with expectations.

*(Statistical theory predicts this as we shall see later)

Stem-and-leaf diagrams

Even simpler than the boxplot is the stem-and-leaf diagram. It is just a kind of histogram that allows you to retrieve the original data, which a histogram doesn't. It lets you see more of the detail of a distribution than a boxplot does, an so is a useful complement. It is available on SPSS.

Given the following numbers: 43, 46, 46, 47, 48, 48, 49, 51, 57, 58, 60, 62, 63, 74, they are re-written in a way that is easier to see than describe. The underlined 1 in version A is the number 47, and comes from the fact that we need to add on 1 to the stem (46) to get 47. The choices to make in a stem-and-leaf are: (i) which number to start with, and (ii) the category or bin width. These are the same issues you must confront when making a histogram. The B version of the same data starts with 42, rather than 43, but has the same category-width of 3. In version B, 47 is represented as 45 + 2. In version C, the category width is 6. Here 47 is represented as 42 + 5. Both versions A and B are at about the right level of detail to reveal the possible bimodality of the distribution. Version C, on the other hand, does not bring this out, so the bin-widths are too wide. Bin-widths can also be too small. It is a matter of judgement.

The stem-and-leaf diagram conveys the same information as a histogram only if the numbers in the "leaves" have the same width. If this were not the case, we could cram many more 1s into the same space as 0s, giving a misleading impression, because the eye pays attention to the ragged edge of the distribution.

Finally, two groups of numbers can be put back-to-back, as in D, to help compare them.

Figure 3. Stem-and-leaf diagrams of the same data (versions A, B, and C). Also shown is a back-to-back stem-and-leaf diagram (D).

A

43	0				
46	0	0	**1**	2	2
49	0	2			
52					
55	2				
58	0	2			
61	1	2			
64					
67					
70					
73	1				

bin-width, 3
starting value, 43

B

42	1		
45	1	1	**2**
48	0	0	1
51	0		
54			
57	0	1	
60	0	2	
63	0		
66			
69			
72	2		

bin-width, 3
starting value, 42

C

42	1	4	4	**5**
48	0	0	1	3
54	3	4		
60	0	2	3	
66				
72	2			

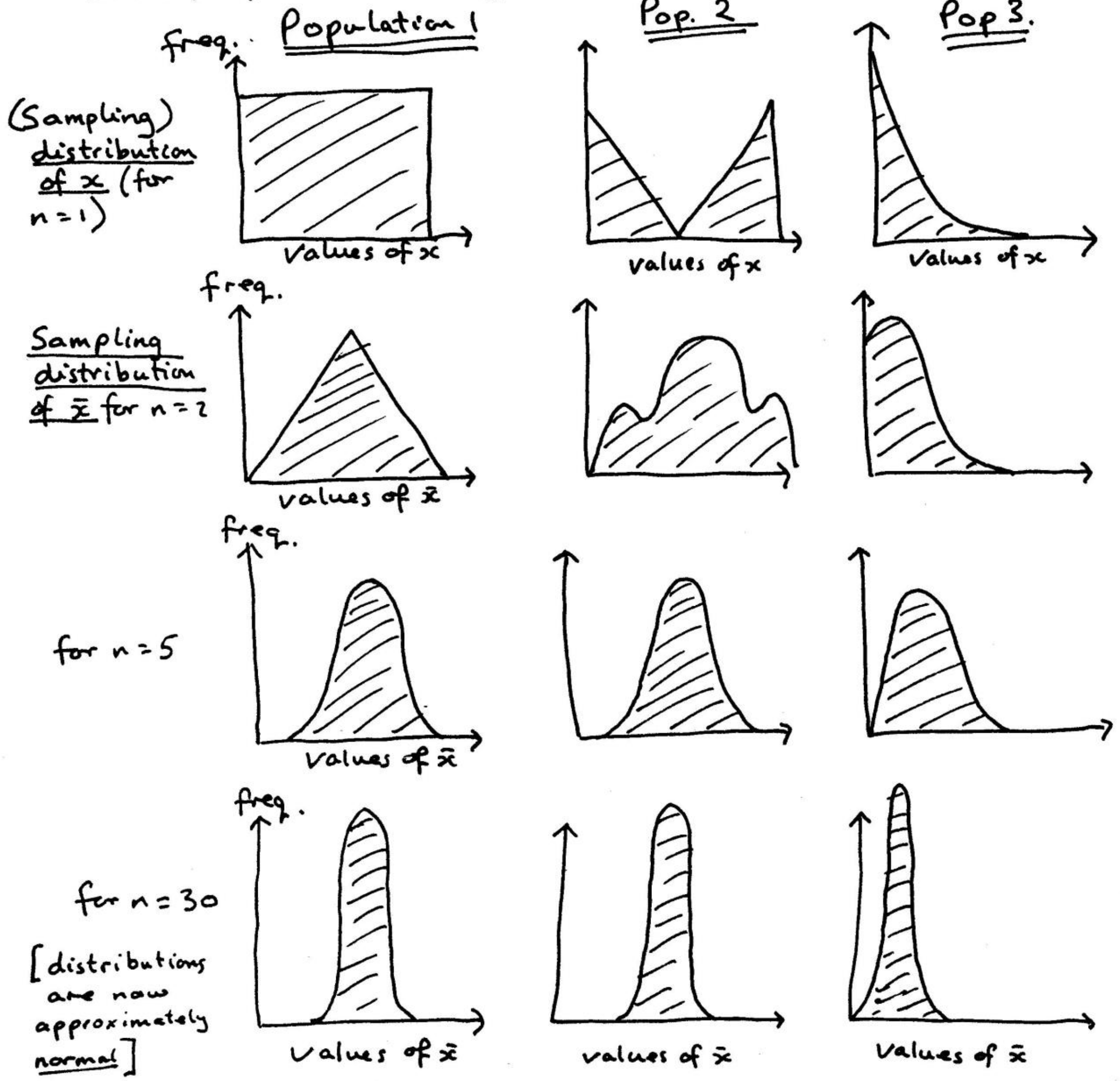
bin-width, 6
starting value, 42

D

					43	0				
			2	1	46	0	0	1	2	2
2	1	0	0	0	49	0	2			
	2	2	2	0	52					
		1	1	0	55	2				
				2	58	0	2			
					61	1	2			
					64					
					67					
					70					
					73	1				

Central Limit Theorem

If random samples of size n from a population, the sampling distribution of the sample mean, $\bar{x}$, can be approximated by a normal probability distribution (the bell-shaped curve), as the sample size, n, becomes large. Also, $S_{\bar{x}} = S_x/\sqrt{n}$.



What does the equation: ln(y) = a + b.ln(x) look like in terms of x and y?

First, remembering that: $10^{\log_{10}(m)} = m$ (e.g. $10^{\log_{10}(100)} = 10^2 = 100$);
and similarly, $e^{\ln(m)} = m$, where "e" is the special mathematical number 2.71828.

We do the following

$$e^{\ln(y)} = e^{(a + b.\ln(x))}$$

$$y = e^a . e^{b.\ln(x)}$$

but since $b.\ln(x) = \ln(x^b)$

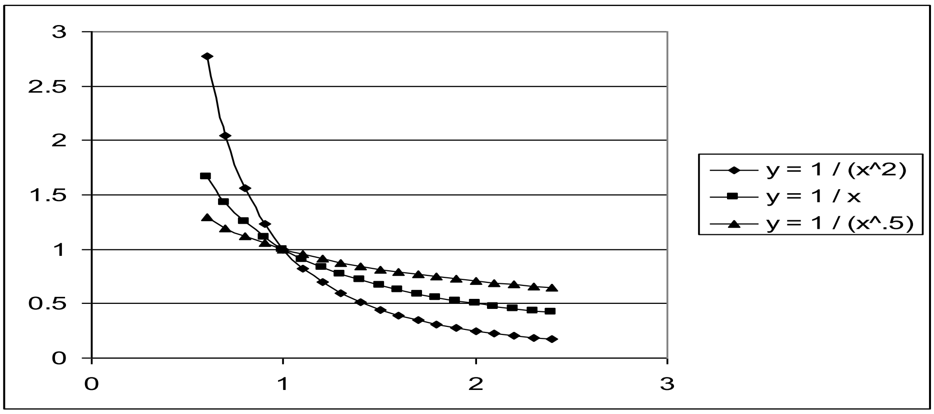
$$y = e^a . e^{\ln(x^b)}$$

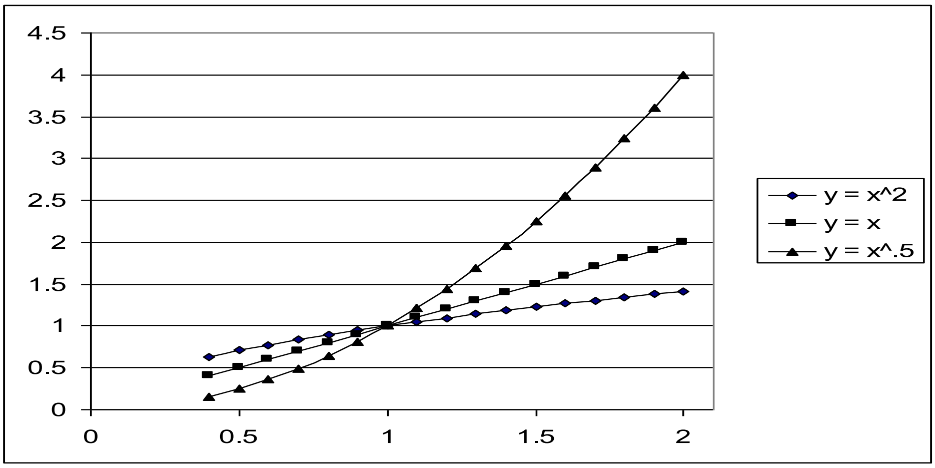
$$= e^a . x^b$$

$$= k . x^b$$ because e^a is a constant, which we shall call k.

So, taking logs of x and y we are trying to predict k and b, to give us a family of curves of best fit. If b turns out to be 1, then we just have a line: y = kx. If b = -1, then we get y = k/x. The first figure give the idea for b = -2, -1 and -.5; the data we have in Q3 looks to conform to one of these. The second figure shows b = .5, 1, and 2.

(Note: x^2 means x^2)





The Rank-transformation approach

I questioned respondents over the phone and asked how many times in the last month (4 weeks) they had shopped at supermarket X. I noted down whether it was 0, 1, 2, or 3 times. For larger numbers than this, I simply recorded it as “frequently” (*i.e.*, ≥ 4). There are a few points to be made here about the nature of the “scale” thus produced:

1. This scale (or level of measurement) is clearly at least *ordinal*, but the first four points on it are also *ratio* scaled.
2. The net effect of coding all numbers at or above 4 is to “Winsorise” the data (see: http://en.wikipedia.org/wiki/Winsorising). That is, all values above a certain criterion value (here 4) are reduced to that value. This is considered good practice when the data may contain outliers or be highly positively skewed. It protects against the analyses being dominated by a small number of people who give very much more frequently than the majority.
3. When calculating means with Winsorised data, we can interpret the statistic as a lower bound on the mean, rather than the actual mean itself.
4. Nonetheless, even had the scale been straightforwardly ordinal (e.g., *never, occasionally, sometimes, often, frequently*), standard practice found in every journal has been to code these as 1, 2, 3, 4, and 5 and treat the data as if it was *interval* scaled. This allows the researcher to analyse the data by using means, t-tests, OLS regression, etc. Sometimes commentators decry this practice and stipulate the use of non-parametric tests instead. Non-parametric tests are generally taught and understood by appeal to the theory of rank permutations to derive p-values. However, what is less well known is that most of the standard non-parametric tests are themselves mathematically / statistically equivalent to performing interval-scaled tests on ranked versions of the raw data. For example, a Pearson’s correlation taken on ranked data *is* a Spearman’s correlation. A Pearson’s correlation on the data that constitutes a 2 x 2 contingency table *is* the phi coefficient of association. Conover and Iman (1981, p.124) describe it this way:

“Simply replace the data with their ranks, then apply the usual parametric t test, F test, and so forth, to the ranks. We call this the rank transformation (RT) approach. This approach results in a class of nonparametric methds that includes the Wilcoxon-Mann-Whitney test, the Kruskal-Wallis test, the Wilcoxon signed rank test, the Friedman test, Spearman’s rho, and others.”

Therefore in following the frowned-upon practice of using parametric statistics on ordinal data whose levels have been recoded as ranks, researchers are actually carrying out the equivalent of non-parametric tests without even realising it. Conversely, in performing standard non-parametric tests by-the-book, researchers are actually themselves engaging in the frowned-upon practice!

The authors continue:

“The rank transformation approach also furnished useful methods in multiple regression, discriminant analysis, cluster analysis, analysis of experimental designs, and multiple comparisons.”

It can now be seen that the analyses in this thesis simply follow Conover and Iman’s rank transformation approach by applying parametric tests to ranked data.

Conover, W. J. & R. L. Iman (1981). Rank transformation as a bridge between parametric and nonparametric statistics. *The American Statistician, 35(3).* 124-129.

FD.R (some R commands)

Read in the data. It is separated by tabs, hence sep="\t" which does it.

```
> pqr=read.table("D:/FilmDistributors.txt",header=T,sep="\t")
```

You will probably need to change D:/ to wherever it is that you have put the file FilmDistributors.txt. Also, be careful when copying either single or double quotes from MS-Word. you will need to re-type them by hand in R.

What column names are there?

```
> colnames(pqr)
```

Do the shorthand so we don't need to write pqr$ every time

```
> attach(pqr)
```

Examine these pairs:

```
> mean(IAgrossm)       # Mean of all the data for gross revenue
> by(IAgrossm,Distributor,mean)   # mean of gross broken down by distributor

> median(IAgrossm)      # median of all the data for gross revenue
> by(IAgrossm,Distributor,median)  # median of gross by distributor

> IQR(IAgrossm)      # IQR of all the data for gross revenue
> by(IAgrossm,Distributor,IQR)   # IQR of gross broken down by distributor

> summary(IAgrossm)      # summary of all gross revenue (min Q1, Md, Q3, max)
> by(IAgrossm,Distributor,summary)   # summary of gross broken by distributor

> table(Distributor)      # count how many films for each distributor
> by(Distributor,Genre,table)   # ditto, but broken down by Genre too
> by(Distributor,MPAARating,table)   # similarly for MPAARating
> by(Distributor,Year,table)   # similarly for Year

> boxplot(IAgrossm)
> boxplot(log(IAgrossm+1))
> boxplot((IAgrossm+1)~Distributor)
> boxplot(log(IAgrossm+1)~Genre,main=”Inflation Adjusted Gross”)

> windows() # do not replace graphs but keep all of them

> qqnorm(IAgrossm)
> qqnorm(log(IAgrossm+1))    # add one to avoid log of zero
> qqnorm(log(1+IAgrossm[Distributor==1]))  # qq plot for Distributor 1 only.
                                            # Note it is == NOT =.
> t.test(IAgrossm[Distributor==1],IAgrossm[Distributor==2])
> t.test(IAgrossm[Distributor==5],IAgrossm[Distributor==3])   # etc.

# t-test on logged version of data (logged because of + skew)
> t.test(log(1+IAgrossm[Distributor==1]),log(1+IAgrossm[Distributor==2]))

# non-parametric alternative to t-test:
> wilcox.test(IAgrossm[Distributor==1],IAgrossm[Distributor==2])

> wilcox.test(IAgrossm[MPAARating=="PG"],IAgrossm[MPAARating=="G"])
> wilcox.test(IAgrossm[MPAARating=="PG"],IAgrossm[MPAARating=="R"])

> stem(log(1+IAgrossm[Distributor==5]),scale=2)
```

***Finally, FILE / SAVE HISTORY* will save all the R commands you typed in.**

Chi-squared tests

H_0: The launch of films is uniformly distributed across months of the year
(more accurately “across the year” - see below).
H_a: Some months are more popular than others for launching films.

..or..

```
H0: O = E     (Observed numbers do not differ from Expected)
Ha: O ≠ E
```

```
> mnO = table(Month)     # the Observed number in each month
> chisq.test(mnO)
```

This is a ”quick-and-dirty” χ^2 test on the data: it assumes there are the same number of days in each month.

What we need to do is to taking into account that months vary in #days. Create a variable containing the number of days in each month.

```
> mn = c(31, 28.25, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
> sum(mn)    # will be 365.25, as we would expect
> sum(mnO)  # this will be 2881 - the total number of films (no filtering).
```

The next step more accurately apportions the total films by size of month.
These are the *Expected* values (#films for each month according to H_0)

```
> mnE = sum(mnO*mn/sum(mn)  # e.g. for January it is 2881 * 31 / 365.25
                            # For February it is 2881 * 28.25 / 365.25, etc.
```

Now we calculate χ^2 by hand:

```
X2 = sum( (mnO-mnE)^2 / mnE )   # white spaces are to help the eye
```

In more conventional maths, it this calculation is:

$$\chi^2(\mathrm{df}=11) = \sum \frac{(O-E)^2}{E}$$ where we are summing over all 12 months.

X2 is 77.489, where previously we had 82.858. In other words, the variability of #days in each month was artifically inflating our value of χ^2, and hence giving ourselves a slightly too-liberal test. However, we still reject H_0 and conclude that there is variability across the year in the number of films launched. We turn our χ^2(df=11) into a p-value as follows:

```
> 1-pchisq(X2,11)  # gives p = 4.504e-12
```

That is, p = 0.000000000004504, which is clearly less than .05. In rejecting H_0, we do so accepting that there is a small probability that we might be wrong. More formally, the probability of making a Type I error is 4.504e-12, which is less than five in a trillion.

Exploratory hypothesis – Towards Content Analysis

Do films with long titles gross more than those with short titles, or is it the other way round (H_a is 2-tailed)? Or is title length unimportant (H_0)?

Ha: Title length is correlated with film gross. r(length, gross) ≠ 0.
Ho: Title length is not correlated with film gross. r(length, gross) = 0

```
> TLen = nchar(as.character(Movie))  # Is this OK? Check one or two films.
> cor(TLen,IAgrossm, method=”pearson”)  # But is it significnt?
> cor.test(TLen,IAgrossm,”two.sided”,method='pearson')
```

Also possible are Spearman's and Kendall's non-parametric correlations:

```
> cor.test(TLen,IAgrossm,”two.sided”,method=”spearman”)
> cor.test(TLen,IAgrossm, ”two.sided”,method=”kendall”)
```

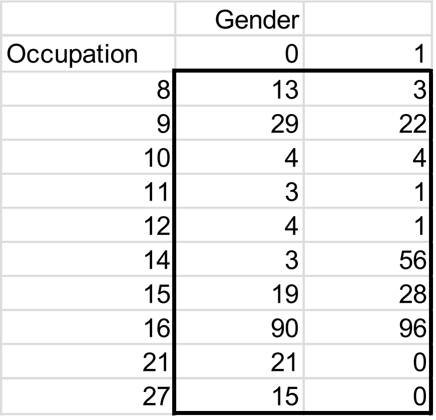
Subsetting in R using [,]

```
>xxx=read.table(“C:/personclean.txt”, header=T,delim=”\t”)  # or wherever you have put the file
# alternatively, rather than reading directly from a named file, we could copy contents of that
# file to the clipboard and copy from there.
>xxx = read.table(“clipboard”, header=T)
>attach(xxx)
>genocc=table(Occupation,Gender)
```

# genocc is now a table that has the counts ready for χ^2 testing . If we do:

```
>genocc
```

It looks something like:



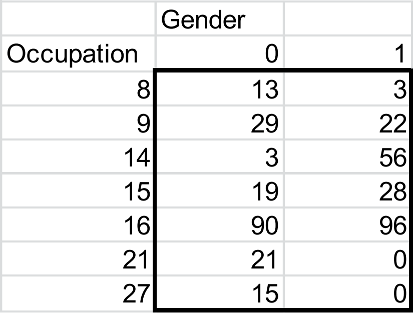
	Gender	
Occupation	0	1
8	13	3
9	29	22
10	4	4
11	3	1
12	4	1
14	3	56
15	19	28
16	90	96
21	21	0
27	15	0

What we want to do (erroneously) is:

```
>chiqs.test(genocc)
```

Although R does compute χ^2(df=9) as 94.34, p <.001, it also warns that the result may not be correct. What is wrong here? Notice that Occupations 10, 11, and 12 all have Expected frequencies less than 5. This makes the χ^2 unreliable. Perhaps we should remove Occupations 10, 11, and 12, and redo our analyses. Best not use subset(), but use [,].This is how:

```
>genoccX=genocc[c(1, 2, 6, 7, 8, 9, 10), ]  # picks out rows 1, 2, 6, 7, 8, 9, 10.
>genoccX
```



	Gender	
Occupation	0	1
8	13	3
9	29	22
14	3	56
15	19	28
16	90	96
21	21	0
27	15	0

```
>chisq.test(genoccX)
```

Now R computes χ^2(df=6) as 92.30, p<.001, but doesn't display a warning. Note in passing that the probability quoted by R in both the 10-row version and the 7-row version is 2.2e-16. This is not an amazing co-incidence, but just the limit on the probability that R is willing to compute for us. In both cases they're tiny tiny.

```
>genocc[6,2]    # the sixth row and column two (i.e., 56)
>genocc[6,]      # the sixth row (all of it, because column unrestricted)
>genocc[,]     # prints all of genocc; therefore same as: >genocc
>genocc[c(4, 2, 3),]  # prints rows 4, 2, and 3, in that order.
>genocc[c(“16”, “27”), ]  # print rows for Occupations 16 and 27. Word quotes don’t work in R. Don’t copy, but type
them!
```

Therefore, genocc[“8”,] prints the first row, while genocc[8,] prints the eight row.