\user\jobb\stat.98
980504
Torbjörn Ledin, Dept ENT, University Hospital, Linköping
email Torbjorn.Ledin@inr.liu.se

Adopted for the Basic Statistics Course for the PhD students. The examples shall be seen as computation illustrations to the use of statistical packages, therefore the calculations are not complete, only principally sketched. The possibility of errors can not be ruled out at this stage. The survey does not claim to fully cover the area. Comments are most welcome.

HITCHHIKER'S GUIDE TO ELEMENTARY STATISTICS

BRIEF INTRODUCTION

Statistics, why?

The question is not that of a first year lazy student.

Statistics could be said to be the art of science that tries to compute things that cannot be computed. It has built a system of rules, methods and assumption strategies that allow us to conclude things from what our senses can read or measure. Sometimes our measurements produce data that after processing by the laws of statistics tell us to decide something, but it could very well be so that this decision is entirely wrong. Statistics must therefore not only supply us with a value that allows us to decide something, but it must also deliver an estimation of how large the risk of a false decision is. But, do not forget, all decisions and risk estimations are based on that we follow the prerequisites of our procedures. If those are violated, we can say exactly nothing.

Looks like we are at step one again.

Inferential statistics

In medical profession research, we want to do experiments, evaluate the outcome, and draw a conclusion based on the data. Most often we want to prove that a treatment is more effective than another, or that our group of patients differ from another group. The term statistical inference describes how to draw conclusions from statistical computations. Another branch of statistics deal with describing sets of data, among the most simple analyses is how to compute the mean value, or tell how many persons have a certain characteristic.

What is probability?

Back to statistical inference, the term probability is essential. If we make a decision that the male patients in general are taller than female, a proper statistical analysis would contain (at least) the following information: the men had a certain length in average (eg 178 cm), this value had a certain accuracy (eg 'SD=7 cm'), the women had some corresponding data, the group sizes are known, and statistics perhaps tells us that the difference is a number of centimeters, and that this 'probably is larger than zero'. The last fact must rely on a risk estimate of being wrong, eg 'the probability that this is wrong is less than 5 %'. This 'wrong decision risk' is often termed the alpha-error or probability (of 'null hypothesis truth'). In many introductory texts on statistics this is illustrated by a normal distribution curve and a small area towards the right-most part of it showing that the small area is 5% of the total. However, this 'wrong decision risk' can be defined for any type of data distribution (not only the normal = Gaussian distribution).

Why these p's and stars?

The probability estimates are generally denoted 'p' in scientific literature (and many authors seem to expect this to be self evident despite it sure is not). A statement $p<0.05$ means that the probablity estimate is less than 5%. Often the level 5% is denoted with one star (*), 1% with two stars (**), and 0.1% with three stars (***). However, this is certainly not a universal law, and only star notations (without definitions) in a scientific text should not be accepted. The choice of 5%, 1%, and 0.1% as 'golden standards' for probabilities is also entirely arbitrary, you may define any limits. There is in practice no difference between $p<0.049$ and $p<0.051$, but using a cut-off decision rule at 5%, the first is significant but the second not. From this standpoint, another idea is of course to give the exact probability values you have computed (which also enables you to compute exact confidence intervals for the studied variable see special section on this).

One or two tails in the risk estimate?

To make things even more complicated, the alpha error can be based on both a one-tailed test and a two-tailed test. One-tailed means that you compute only the risk that the true value is, say, less than the observed one. Thus it is useful for experiments such as are men taller than women, or is the income of lawyers larger than that of statisticians. However, in all experiments where data can go in any direction it is wiser to use a two-tailed test, which answers to the question if the value of one group differs from that of another group, ie we are not interested in which direction it differs, but if it does. In 'symmetric' data populations, such as the normal distribution, the two-tailed probabilities are twice as large as the one-tailed, ie the risk of being wrong in your conclusions is doubled. As a general suggestion, be very careful in using one-tailed risk estimations. Two-tailed tests make you sleep much better at night.

DATA QUALITY

What types of data are there?

1. Nominal data: Categories such as men/women, party sympathies. These data are of fairly low quality and permit only limited statistics. But, try it!

2. Ordinal data: Can be ranked in a non-disputable fashion, such as T in the TNM system. Everybody would agree that T1 is better than T2, or T3, or T4. The possible values of the variable T are in a clear way possible to rank.

3. Interval data: The numeric difference between two locations on the scale has some meaning, however, the quotient between locations is not trustworthy. Example (one of the few) temperatures Celsius centigrades. From 10 to 12 is fully comparable to the distance from 20 to 22, but 20 is not twice as much as 10.

4. Ratio data: Two locations on the scale can be divided by each other and a valuable number is obtained. Examples are tumor size, body length. Two centimeters is twice as much as one cm.

As clearly can be seen, these data categories are in an ascending order. Ratio is better than interval in terms of data processing, and interval is better than ordinals. For practical purposes however, all reasonable procedures can be applied to both ratio and interval data. In the following they are treated as one group.

Normally distributed data and the Kolmogorov Smirnov test

The most efficient statistical procedures require data to be normally distributed (Gaussian distribution). If data are not, one should beware of using the parametric procedures below. One way of establishing that data are not significantly different from the normal distribution is to apply the Kolmogorov-Smirnov (vodka) test. Simply do as follows.

Construct a frequency histogram of your observed data:

*** fig

From this construct the cumulative histogram by "walking" from left to right successively adding the number of subjects corresponding to the location in the frequency histogram. In this way the value at one location is identical to the total number of subjects to the left of that location in the frequency histogram:

*** fig

On top of this, draw the cumulative histogram of the normal distribution using the computed mean and standard deviation of your sample of data:

*** fig

The Kolmogorov Smirnov test looks for the largest distance between the observed cumulative histogram of your material and the same histogram of the normal distribution. For a given maximum distance, the test will say 'yes' or 'no' to the statement that your data sample is normally distributed. In the shown example it is very well corresponding to the normal distribution.

DIFFERENT TYPES OF STATISTICAL DESIGNS

What is the design of the study?

Having established what type of data is available, a statistical test is to be applied. The investigation design is critical to this choice.

1. One group test. One set of patients is investigated. The task is to decide what the properties of this group are.

2. Two groups test. The two different groups shall in some way be compared, and a decision on the presence or absence of group differences is the aim.

3. Three or more groups are investigated. Do these groups differ in some way, and in that case, which group(s) are in some way "outstanding"?

Paired tests, a special case, or....?

A special problem is how to treat paired designs of investigations. For most elementary statistics this is quite often not critical. Simply, replace the two values for each measured subject with that person's difference! This requires the analysis to be "linear" with respect to the measurement scale, a condition which is fulfilled for most tests beneath.

CHOOSING THE ANALYSIS STRATEGY

A novel strategy for statistical test selection

Here is a simple rule how to choose type of statistical test:

1. Nominal data: Choose "Category procedures"

2. Normally distributed data not of ordinal scale type: Choose "Parametric tests"

3. Other data: Choose "Nonparametric tests"

Table of test possibilities

Having observed the above questions, it should be quite easy to decide which test to use, just have a glance at the following:

DESIGN	CATEGORY PROCEDURES	PARAMETRICS	NONPARAMETRICS
One group	Binomial test Sign test Poisson test McNemar test	Student's t-test	Wilcoxon rank sum test
Two groups	Fisher's exact test Chi-2 test (Yates correction)	Student's t-test	Mann-Whitney U test
Many groups	Chi2 test	ANOVA + 2nd stage Bonferroni methods	e.g. Kruskall Wallis test
Correlation	-	Pearson's r	Spearman rank correlation
Available descriptors	Frequency counts	mean, SD	median, percentiles

GLOSSARY OF TESTS

CATEGORY PROCEDURES

Binomial test

In the studied group, did a reasonable number of subjects have a certain property?

In a group of ten subjects we expect 20% to have a disease. However, four subjects have it. Is this by chance alone?

The binomial test with N=10, prob=0.2, observed=4 gives the answer: p<0.12.

This is obviously not a very challenging finding!

Sign test

This is actually more of a paired test. Measure subjects before and after a treatment. Some got better, some got worse, some were unchanged. Is this a random outcome?

Suppose 20 patients were operated, 12 got better, 3 unchanged, 5 worse.

Sign test gives us p<0.15.

Dept ENT, University Hospital, S-581 85 Linköping, Sweden

We can not prove that we did something useful.

Poisson test

This has nothing to do with hazardous chemicals or french fish.

It answers: Out of all those patients we had last year, hundreds, we expected to have only 2 of that very special rare one 'on the average' as computed from statistics over the years. However, we saw 5! Is this something remarkable or just by chance?

Poisson test for expected = 2, observed = 5 gives us p<0.053.

We can almost classify ourselves as a lucky clinic regarding education of the youngsters of the profession.

The test shall only be used with rare events that actually occur a few times over a certain period, at least not those which occur many dozen times.

Fisher's exact test

If we have two groups and two possible states, do the observed frequencies in our study reflect chance only, or is there a manifest deviation from the random outcome?

Consider eg 14 cancer patients out of which 4 have brown shoes, and as control group 12 CEOs out of which none have brown shoes. Is it dangerous to have brown shoes?

A two-by-two table would be

	Brown	Other
Cancer	4	10
CEO	0	12

In this case we expect quite few (approximately two) brown shoes 'by random' in each group, and the Chi2 test (see below) is not allowed to use. Hopefully your computer gives you a warning if an attempt is performed!

Our friend Fisher will tell the verdict: $p<0.10$.

We can safely go on using brown shoes, despite CEOs don't.

The Fisher test is always allowed to be used, Chi2 only when numbers grow.

Chi2 test ('chi-square')

The question is as for Fisher's exact test, however, the test can also be used with more than two groups, and/or more than two states. In the two-by-two situation, however, the test is possible to 'sharpen' using the Yate's correction (for the continuity approximation of integer numbers).

So, as before, we study cancer and executives.

Preference:	Bourbon	Smirnoff
Cancer	5	47
CEO	31	11

Chi2 says $p<0.001$ (or even better...) using Yate's correction.

Confounding factors may invalidate the conclusion that bourbon protects against cancer.

Using more groups a table could look like:

Shovel when?	Morning	Evening	Holiday
Group A	12	9	6
Group B	8	10	12
Group C	10	4	6

This looks to deviate a little, but as $p<0.36$ it has probably arisen by chance alone.

Fisher's exact test is not available in this case.

McNemar's test

McNemars test evaluates if it is reasonable that a number of individuals in a group improve betwen the measurement times, and that simultaneously a number of individuals deteriorate. The subjects that remain in their original states do not influence the test at all. The procedure is a Chi2 version using two cells, the improvers (N1) and the deteriorators (N2). The tested quantity is (square of (N1-N2-1))/(N1+N2), with 1 df. We here assume that N1 is the bigger number, otherwise change them. A Yate's correction has also been used. The test appears to give identical results to the sign test.

PARAMETRIC TESTS FOR MAXIMUM TWO GROUPS

Student's t-test

This is well known but has nothing to do with any student at all. The inventor (whose name was William Gossett) was working in some kind of brewery quality control function, and was afraid his ideas wouldn't be accepted if he wrote his correct name and institution on the paper.

It can be done as a paired test, in this case compute the difference before-after on each patient and use the procedure below for one-group test.

If two groups are investigated before and after a treatment, simply compute the differences in each group and enter 'difference' data into the procedures for two groups below.

First, the mean and standard deviation SD of a data set is given.

mean = (sum of all observations) / N(umber of observations)

SD = square root of ((sum of squared deviations from mean)/(N-1))

the latter is however identical to

SD = square root of ((sum of squared observations)-squared sum of observations/N)/N-1)

The last line allows you to conduct meta-analysis of several other papers giving mean, SD and N of their patient groups, and you can construct the total mean, SD and N without asking for raw data from the authors.

Student's t-test is now quite straightforward.

1. Compute your observed difference (before vs after, vs the Swedish well known average, between groups, etc)
2. Construct a measure of the SD.
3. Establish a relevant divisor relative to your sample size.
4. Compute t=difference/(SD*divisor)
5. Establish the number of freedom degrees (one group=N-1, two groups = N1+N2-2)

Dept ENT, University Hospital, S-581 85 Linköping, Sweden

6. Check a t-table for the significance level.

Three examples show the details:

1. One group compared to the Swedish mean. Ten men have length mean=183, SD 8, Swedish grand mean value = 178 cm is considered a fixed value (as it is based on a very large study population, compare below).
In this case the SD estimate is identical to observed SD 8 cm.
The divisor is identical to the square root of 1/N in this case.

We have
t=(183-178)/(8*square root of (1/10))=1.98.
Degrees of freedom is N-1 =9.
t-table says not significant, p<0.08 (two tailed).

2. A paired study measured something before and after a treatment. The investigator calculated the differences for all subjects and found mean=5, SD=7 and N=20. If we want to see if there was any effect of the treatment we should compare this difference to zero, ie the no effect value. We do exactly as above and find
t= 5/(7*square root of (1/20))=3.19.
Degrees of freedom is 19.
t-table says p<0.01, clearly significant (again two tailed).

3. Two groups were investigated in a study. In one group the results were m1=10, SD1=10, N1=10, in the other m2=20, SD2=20 and N=20 were found.
In this case the SD estimate of the difference is identical to the long expression
SD=square root of ((SD1square*(N1-1)+SD2square*(N2-1))/(N1+N2-2))

thus in our example

SD=square root of (10*10*(10-1)+20*20*(20-1))/(10+20-2)) = square root of 303.6=17.4

The divisor in these cases is square root of (1/N1+1/N2) thus in this case square root of (1/10+1/20)=square root of 0.15 = 0.39

So we get
t= (10-20) / (17.4 * 0.39) = 1.47 drop the minus sign!
Degrees of freedom is N1+N2-2 = 28

The t-table says not significant, p<0.15 two-tailed.

NONPARAMETRIC TESTS FOR MAXIMUM TWO GROUPS

Wilcoxon rank sum test

This is the first non-parametric test in our series. The term "non-parametric" is based on that no characteristic measure of the data is compared in the test (like means and SDs are compared in the different versions of Student's t-test), rather the calculations are based on creating the observations in some kind of logic order and then considering their relative positions.

The basic concept in this field is the well known median of the data. To find the median, just order the numbers from smallest to largest and take the one in the middle. If an even number of data are at hand, take the average of the two in the middle. An example scatters some light on this:

Observed data:	1	3	9532	2	99	21
In ascending order:	0	1	2	3	21	99

The middle one is the median, ie 3. Those large ones at the end didn't play an important role despite they are huge compared to the others!

Further on you may define what is called percentiles. The median is actually the middle, ie the 50th percentile. The value 25% distance "from the bottom" is the 25th percentile, or the first quartile. In an analogue fashion it is easy to understand what the 75th percentile = the third quartile means.

One measure of the dispersion of these data is called the interquartile range, it is the distance between the first and the third quartile.

Now, the Wilcoxon rank sum test is the way to evaluate if it is likely that the median of our data is equal to some number, or not. The most useful procedure is to use it in a paired design investigation. If no effect is at hand, the median should be zero. Thus, we test if the differences of all individuals investigated is zero or not.

Here is a set of data:

Person no	1	2	3	4	5	6	7
Before value	4	3	2	1	4	5	2
After value	3	4	2	4	2	3	5
Difference	1	-1	0	-3	2	2	-3

The next step is to disregard the sign of the differences and order them from smallest to largest. The zero value does not affect the calculations and is disregarded. The rank of each value above will thus be:

Rank	1.5	1.5	-	5.5	3.5	3.5	5.5

The sign of each difference should be disregarded at this stage.

Note that the 1 and -1 occupy places 1 and 2 and therefore will have score 1.5 each, and a similar argument applies for the other halfnumbers.

Finally, compute the rank sums of all positive differences, and the corresponding sum for all negative values:

Positives = 1.5 + 3.5 + 3.5 = 8.5
Negatives = 1.5 + 5.5 + 5.5 = 12.5

The zero value is omitted, no difference was observed in that case.

The Wilcoxon test now gives us that this is not significant, $p<0.60$. The conclusion is that no significant difference has occurred between the two measurements.

Mann-Whitney U-test

This test compares two sets of data in a fashion similar to the Wilcoxon test.

Here is a sample of data from an investigation:

Group 1	3	1	99	12	15	
Group 2	5	6	7	8	9	10

Arrange them in ascending order regardless of group:

Group 1	1	3					12	15	99	
Group 2			5	6	7	8	9	10		

Rank them regardless of group:

Group 1	1	2					9	10	11	
Group 2			3	4	5	6	7	8		

The sum of ranks in the different groups are

Group 1 : 1 + 2 + 9 + 10 + 11 = 33
Group 2 : 3 + 4 + 5 + 6 + 7 + 8 = 33

Mann Whitney tells us that this is not significant, $p<0.65$. The conclusion is that the values observed in the groups are not significantly different.

COMPARING SEVERAL GROUPS

The multi-comparison 'fishing expedition' problem

If the means or medians of more than two groups shall be compared one could feel tempted to repeatedly compare pairs of groups using t-test or the Mann-Whitney U-test. However this results in multiple comparisons and thus it can be expected that a 5% significance to turn up just by random approximately in every twentieth comparison. The model to use in these cases to estimate an 'overall' significance level can be based on the expression (assuming all pairs to be independent):

'overall-p'=1-product of all (1-p's)

As an example, having five 5% significances results in a total significance level of

p=1-(1-0.05)*(1-0.05)*(1-0.05)*(1-0.05)*(1-0.05)
=1-0.95*0.95*0.95*0.95*0.95 = 0.226

ie the significance level is this way approaching 23%, nearly five times as much as the individual tests.

The Bonferroni correction

The above gives a hint how to address the problem 'the rough way'; if we divide a requested significance level of eg 5% down to a lower value, perhaps 1%, we have a small protection against this 'multiple comparisons problem'. A simple way is to divide the 'normal level' of significance by the total number of comparisons. Of course other methods can be applied as well. This idea is termed the Bonferroni method.

Parametric ANOVA procedures

A more general way to take care of these problems in the parametric case is to use analysis of variance (ANOVA). It represents a variety of methods where the variability of the observed data is modelled as a sum of variability due to the group to which the subjects belongs, other measured factors, interaction between the factors and finally the (hopefully) random variability that remains unexplained by the model. ANOVA answers if the variability due to the model of the data is large enough to deduct that it has an 'explanatory value' in the experiment. In practice, the question answered is if the groups and / or factors studied have a significant impact on the measured variable (or if they are completely unrelated). If they have, you are allowed to look for where the differences are (using so called secondary statistics; Tukey, Duncan, Dunett and many more methods). If they do not differ, it is not recommended to look for possible differences related to groups or studied factors.

Non-parametric 'ANOVA' versions

If your data do not fulfill the criteria for use with the parametric tests, there are also statistical procedures available if more than two groups are studied. They can analogously be termed 'non-parametric ANOVAs'. The most well known is the Kruskall Wallis test. To consult a statistician in these matters (as well as in all but the most simple versions of ANOVA) is probably not a bad idea.

MISCELLANEOUS

Until the next version
- regression
- correlation t=r*sqrt ((N-2)/(1-r2)), df=N-2, Pearson's r and Spearman rank correlation
- multivariate methods
- non-linear procedures
- confidence interval

and many more have been left out.