Statistics with R for Biologists

James H. Bullard
Kasper Daniel Hansen
Margaret Taub

Berkeley, California
July 7-11, 2008

Testing

What can we do with statistics

Common goals with a statistical analysis

- Estimation
- Hypothesis testing
- Understanding
- Prediction
- Get low p-values and publish

These goals might look similar, but in actual use cases they can lead to surprisingly different strategies for analyzing data.
In this lecture we will focus on hypothesis testing.

The two-sample location problem

We will start by considering a classical problem: the two-sample location problem. We will attack this problem using a variety of methods, and essentially touch many important concepts in modern statistics.
We are observing observations from two groups – this could be gene expression for two different organisms, blood pressure from males/females, etc.
We assume that the observations are iid. (independent, identically distributed) within each group. We also assume that the distributions of observations from the two groups are equal, *except* a location shift (mean shift).
The objective is to decide whether there is any difference between the two groups.

Testing: setup

The basic ingredients in the testing framework are

- A model (assumptions)
- A hypothesis
- A test statistic

Contrary to what some people seem to suggest, there is always a model. Sometimes the model is really obscured (especially if you read a paper).

test statistic

A test statistic is some univariate function of the data. It should be small when the hypothesis is true and big when the hypothesis is false

(A better description for big/small is perhaps "extreme", "less extreme")

Example: t-test

The t-test is a solution to the two-sample problem

(Classic) model the two groups are both normal distributed with the same variance. They differ in their mean (location).

Hypothesis the location of the two groups are the same.

Test statistic the difference in mean between the two groups divided by an estimate of the standard deviation of the difference. Also known as the t-statistic.

It is clear that the t-statistic is "extreme" (either very positive or very negative) if the location is different and close to zero if the location is the same.

Example

Let us simulate some data and compute the t-statistic. Usually the t-statistic is the difference between the means in the two groups, divided by an estimate of the standard deviation of the difference in means. The are different ways to get the standard deviation estimate, we will be using the standard one (see later though). How this estimate is derived should be discussed in any decent intro level statistics book (left for some serious self-study).

Example, cont'd

```
> group1data <- rnorm(10, mean = 4,
+     sd = 1)
> group2data <- rnorm(13, mean = 7,
+     sd = 1)
> t.stat <- function(x, y) {
+     n1 <- length(x)
+     n2 <- length(y)
+     variance.estimate <- ((n1 -
+         1) * var(x) + (n2 - 1) *
+         var(y))/(n1 + n2 - 2)
+     (mean(x) - mean(y))/sqrt(variance.estimate *
+         (1/n1 + 1/n2))
+ }
> t.stat(group1data, group2data)
> t.test(group1data, group2data,
+     var.equal = TRUE)$statistic
```

When is a test statistic big?

So now we have a test statistic t and we can compute the value on our data $t(\text{data})$. We know that a big value indicates that the hypothesis might be false.
But how big is "big"?
We need an appropriate scale to measure the test statistic on in order to decide how big it is.
And we need a cutoff on that scale in order to decide whether it is too big.

The distribution of the test statistic

The standard approach to this question is to answer the question: "Assuming the hypothesis is true, how likely is our observation". This is an *interpretable probability* scale.
This consists of two interlinked steps

- Find the distribution of the test statistic when the hypothesis is true.
- Compute the probability of a more extreme event happening and choose a cutoff.

p-value

The p-value is the probability of observing a more extreme test statistic when the hypothesis is true.

distribution of a t-statistic

So we want to compute the distribution of the t-statistic when the model is true. To do so we simulate repeatedly from a situation where the means are the same:

```
> group1 <- replicate(1000, rnorm(10,
+     mean = 4, sd = 1), simplify = FALSE)
> group2 <- replicate(1000, rnorm(13,
+     mean = 4, sd = 1), simplify = FALSE)
> tdist <- mapply(t.stat, group1,
+     group2)
```

Example

What does this distribution represent? Is our observed t-statistic extreme? What is the p-value. Try different values for the mean and the standard deviation.

Questions

There are a few things we did, that might seem wrong: We used some specific values of the mean and the variance in the simulation. How do we get these values for "real" data where we do not know the truth? Would we get another result if these numbers change? Why did we use 4 for the mean and not 7 (the two true means were 4 and 7) and why not a 3rd number?
Here a bit of theory comes in great handy. It is not at all obvious, but the distribution we get is actually independent of what value we used for the mean.
That is not always the case, so watch out when you create your own test statistics.
What we did is ok, but it required some values (the truth) we cannot know in a real situation.

Exact / Asymptotic distribution

In some cases it is possible to compute the exact distribution of the test statistics under the hypothesis. In (more often) other cases we resort to an asymptotic argument, where we can find the distribution of the test statistic if the number of observations is big enough.
For the t-statistic, the exact distribution is a t-distribution with $n-2$ degrees of freedom, when the observations are normal distributed.

Example

How does this compare to our simulations? Does our observations support the hypothesis?

Interlude: Unequal variance

It is straightforward to extend the t-test to the situation where the variances of the two groups are different. You simply replace the estimate of the standard deviation of the difference in means in the denominator with a slightly different estimate. This should be used in most cases. It is called a Welsch t-test. However, since we estimate the variance in each group separately, you need to consider whether you have enough observations in each group to do this reliably.

Rejecting the hypothesis

When we reject the hypothesis we say "the is sufficient evidence in the data to support the conclusion, that the hypothesis is wrong".
When we accept the hypothesis we say "our data seems to be in good correspondence with the hypothesis." Or "there is not sufficient evidence for concluding that the hypothesis is wrong".
These two statements are not symmetric: rejecting the hypothesis is "stronger" than accepting it.
This is a basic fact of science, we don't prove theories, we fail to disprove them.

Normality

You often see stated that the data has to be normal distributed in order for the t-test to work. Is this really true? Somehow it would make sense to compute the statistic no matter what – right?
We have seen that the mean of a large number of iid. observations is approximately normal (CLT) so perhaps we can use the t-statistic anyway. How well this approximation works depends on the number of observations and what the "true" distribution is. We still assume the data are iid. except for the location shift.

Example

Examine the t-test in the case that we are sampling from a gamma distribution with shape 1. Try group sizes of (50,60) and (10,15). Try with a shape parameter of 10. What about other distributions?

P-values and power

We have computed some p-values. Often we accept/reject the hypothesis bases on whether the p-value is greater or smaller than some cutoff – typically 5% (mainly historical reasons). Having the right distribution for the test-statistic makes it certain that we have the right *interpretation* of the p-value. This is essentially tied to how often we accept the hypothesis if the hypothesis is true.
But in real life we are also interested in how good we are at rejecting the hypothesis, if the hypothesis is false. This is called power, and is an important way to compare tests.
But power is much harder to calculate, because what does it mean that the hypothesis is false? What *alternative* are we considering?

A permutation test

Permutation tests can be very powerful. They are based on the same ideas we have just seen.
Let us consider the t-test example. We will still use the t-statistic, but we will use a different distribution (scale) to measure it with.
The basic idea is that if the two groups have the same location, all data are iid. – *the group label has no influence on the outcome*. So we could generate samples from the hypothesis just by re-shuffling the group labels.
This has to be done many times in order to get a distribution.
This is the main insight behind permutation tests.

Example

Permute the group labels 1000 (10000) times and compute the permutation distribution. Below is one permutation:

```
> alldata <- c(group1data, group2data)
> idx <- sample(1:23, size = 23,
+     replace = FALSE)
> group1new <- alldata[idx[1:10]]
> group2new <- alldata[idx[11:23]]
> t.stat(group1new, group2new)

[1] -0.6371764
```

See replicate. What is the conclusion this time?

A step back

Let us be clear here. We have now seen

- One statistic – the t-statistic (two if you count the Welsch t-statistic)
- 3 ways of generating a reference distribution of this statistic – using a t-distribution, using a permutation test and straight simulation.

A statistical *test* is a combination of a test statistic and a reference distribution. How effective a test statistic is depends on both.
The reference distribution needs to be "right" in order for the p-value to be interpretable. Without an interpretable p-value, a cutoff is meaningless.

Permutations test, comments

We just computed a (large) number of permutations.
The number of permutations is finite. If the number of observations is even remotely big, the number of possible permutations is astronomical (but it can be "tiny" if there a few observations).
Sometimes, clever arguments (often based on symmetry arguments) and good approximations makes it possible to indeed compute a p-value as if all possible permutations have been considered. This is tough stuff.

Permutation tests, comments cont'd

Permutation tests (sometimes called exact tests) can be very useful. Some further comments

- They can be very computer intensive, especially if you need really small p-values.
- They do not provide confidence intervals.
- They (can) operate under somewhat different assumptions (one point of view is that the grouping is randomly assigned and the observations are fixed).
- There are examples of complicated models where it is not clear that this methodology can be used.
- They require a decent number of observations.

In general permutation tests are probably under utilized, although that depends on whom you talk to.

Non-parametric tests

There is also a whole suite of tests that are called *non-parametric* tests. The idea is to be a bit more flexible regarding the distribution of the data.
Or perhaps rather construct test statistics that are more "robust".
For example the mean is quite affected by outliers, whereas the median is not.
Non-parametric tests often use functions like median and ranks.

The Mann-Whitney test

If we say that the two distributions are the same, but just differ in their location (center), we could construct a test statistic based on how often one groups observations are greater than the other. Since there is no natural pairing, we set

$$Z_{i,j} = \begin{cases} 1 & \text{if } X_i < Y_j \\ 0 & \text{otherwise} \end{cases}$$

(i, j denotes the observations in the two groups)
We then compute

$$U = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} Z_{i,j}$$

if the two locations are about the same, U should be close to $1/2$, otherwise it should be close to 1 or 0.

Wilcoxon rank-sum

There is another statistic which is very similar:

$$W = \sum_{i=1}^{I} R_i \text{sign}(X_i)$$

with R_i being the rank of X_i.
When there are no ties in the data, $U = W + \text{const}$, so the two test statistics essentially have the same value (and hence the same distribution).
The reference distribution is obtained either by large sample theory (asymptotic arguments) or by permutation.

Do it!

Example

The `R` function for a Mann-Whitney and Wilcoxon rank sum test is wilcox.text. Use it on our data. Compute p-values based both on a permutation test approach and based on an asymptotic approach (see the `exact` argument). What is your conclusion?

Summary so far

So far we have seen the following solutions to the two-sample problem

- t-statistic with an (asymptotic) reference distribution.
- t-statistic with permutation based reference distribution.
- Mann-Whitney with asymptotic reference distribution.
- Mann-Whitney with permutation based reference distribution. .

All of them gave similar answers to our problem.
There are differences though (later).

Interlude: test names

There are many statistics with similar (or exactly the same) names. There are at least two Wilcoxon test statistics (rank-sum and signed-rank) and (at least) two ways of obtaining p-values.
What does the statement "I made a Wilcoxon test" cover?
This becomes way more bewildering when one talks about a "chi-square test".

Interlude: one-sample location problem

There is a similar one-sample location problem. Here we have one group and we wish to assess whether the group has a specific location.
Paired two-sample problems reduce naturally to a one-sample problem. This is often a very powerful approach.
(In paired two-sample problems, every observation from one group is naturally matched to an observation from the other group – for example blood pressure before and after treatment).

Parametric / Non-parametric – what is the big deal?

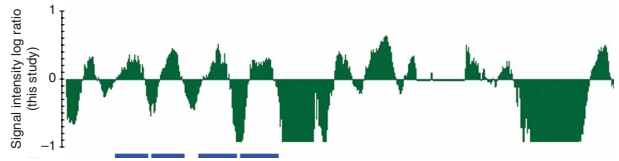
Much has been written about whether you should prefer one over the other.
In this series of examples, the main assumption that all methods make is the assumption of *independence* (and identical distribution, but that can be "fixed"). And that is very hard to verify and can be tough to really believe.
How you get the p-value (the reference distribution) is important.
In this specific case (the two-sample location problem), there is ample evidence (theoretical arguments as well as simulation studies) that show you should always do a Mann-Whitney.

Power simulation

Besides handling outliers better, the reason for preferring the Mann-Whitney test is power – it is better at rejecting a false hypothesis (more precise: at worst, it is a little worse than the t-test, at best it is a lot better).
We will do a simulation experiment to show this, using the t-distribution as error model.
First we examine the level of the test: we simulate data where the hypothesis is true, and examine the p-values. If the interpretation is correct, the p-values should be uniformly distributed.
Then we examine the power of the test when the alternative is a shift of 1 and the significance cutoff is 5%

Back to reality...

We consider data from Lee et. al. 2007, Nature Genetics, *A high-resolution atlas of nucleosome occupancy in yeast*.



They use a tiling microarray to probe the yeast genome. As a result they get data where for each 4bp they have an intensity measure related to the nucleosome occupancy at that position.

Their Question

For each "region" in the genome, they compute a nucleosome occupancy score, which is essentially an average of the probe measurements inside that region.
This allows them to compute a nucleosome occupancy score for each "gene", "50 kb upstream of each transcription start site", etc.
They now want to conclude something like "the nucleosome occupancy is (high) in (genes)" where () could be changed depending on the question of interest.
In order to make this a bit easier they divide their regions into 2-3 classes. Examples are "within gene" vs. "upstream of gene" vs. "downstream of gene". Another example is "highly expressed gene" vs "medium expressed genes" vs "low expressed genes" (where the gene expression is estimated using a different experiment).

Their figure

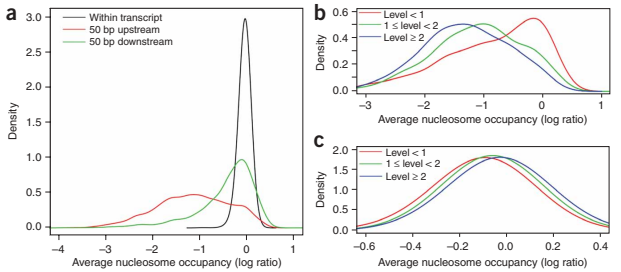


Figure 2 TSSs are demarcated by NDRs. (**a**) Kernel density plot showing the distribution of nucleosome occupancy for regions surrounding verified transcription segments that overlap ≥50% of a verified gene on the 5′ end, as defined in ref. 12 ($n = 5{,}015$). Red shows the distribution of average nucleosome occupancy within a region 50-bp upstream of a verified transcript, green shows occupancy within a region 50-bp downstream, and black shows occupancy within the transcript. (**b**) Nucleosome occupancy within a region 50-bp upstream of verified transcription segments as separated by transcription level. Red shows the distribution of average nucleosome occupancy for promoters of segments with expression level < 1 ($n = 759$), green shows that for segments with expression level between 1 and 2 ($n = 1{,}859$), and blue shows the most highly expressed genes with level ≥ 2 ($n = 2{,}397$). (**c**) Same as in **b**, but showing average nucleosome occupancy within verified transcripts.

Their test

Our data also show that nucleosome occupancy within coding regions correlates with transcription level, but in the opposite manner. Specifically, highly expressed genes are significantly more occupied by nucleosomes than are genes expressed in small amounts or not at all (**Fig. 2c**). This observation also holds true when genes are separated by transcriptional frequency rather than by steady-state mRNA levels (ref. 18 and data not shown). When the nucleosome occupancies of genes expressed in different amounts are compared, the distinctions are statistically significant (t-test, $P < 1 \times 10^{-15}$ for all three expression levels). A possible explanation for the observed patterns of occupancy is that the act of transcription promotes or requires form-

t-test in microarrays

We said that Mann-Whitney tests are prefferable to t-tests for two-sample problems. That is not always true.
Remember that in the t-test we divided by a standard deviation. What if we had additional data to estimate this?
This is something that is not possible to formulate easily in the Mann-Whitney framework.
It is something that is crucial in microanalysis: the moderated t-statistic.

Microarray data

The standard microarray experiment compares two experimental conditions. For each gene on the array we have a number of observations in the two groups (conditions), and we are interested in the genes which are "differentially expressed" This is usually done by a t-test.
When you do an (ordinary) t-test you will estimate a new variance for each gene: σ^2_{gene}.
You could also estimate one variance which a shared across all genes on the array: σ^2.

Microarray data, variance estimation

The problems here are

- The shared variance σ^2 is unlikely to be a good fit. The genes probably will have individual variances.
- But we usually don't have many arrays: σ^2_{gene} is poorly estimated.

Perhaps the truth is somewhere between the two extremes:

$$\sigma^2_{\text{mod, gene}} = \alpha_1 \sigma^2_{\text{gene}} + \alpha_2 \sigma^2$$

(we will ignore how α_1, α_2 are computed).
This is the moderated t-statistic, with good reason widely celebrated in microarray analysis. You should always use this!

Multiple testing, the problem

Recall the definition of a *p*-value

p-value

The *p*-value is the probability of observing a more extreme test statistic when the hypothesis is true.

If we choose *p*-value cutoff of 5% we will make the wrong conclusion (reject the hypothesis) 5% of the time, if the hypothesis of no difference is true.
This has an impact when we do many tests, for example on a microarray.

Multiple testing, be precise

We will need to be specific. Now that we are doing several tests, are we trying to control

- Whether we make 1 (or k) or more error(s) amongst our accepted hypotheses (if we assume all hypotheses are true)
- Whether we make ?% errors amongst our accepted hypotheses (if we assume all hypotheses are true)

This is different types of *Type 1 error rates* (statistical lingo).
The first is called the familiwise error rate, the second is called the false discovery rate (FDR).
We need to choose what type of error rate we control. This concept does not really make sense when we just did one hypothesis test.

After choosing a type 1 error rate, we need to choose a *method* for controlling the error rate. Popular choices are "Bonferroni", "Benjamini-Hochberg", "Holm" etc.
Most of the popular methods in the literature are "marginal methods".
Without going into great detail, marginal methods keep the ordering of the different statistics, they only (essentially) change the cutoff – how many tests are called significant.
When we do correction for multiple tessting we should in general expect to see fewer rejected hypotheses.

Bonferroni correction

Bonferroni is easy to understand. We will control the family-wise error rate.
We assume that all the tests are independent, and that we have a total of n tests. We want to have "5% chance for making one or more error if all hypothesis are true".
We want to choose a cut off p such that if we reject at the rate p for the individual tests, we control the error rate.
But

$$p_{\text{new}} = P(\mathbb{U}_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) = np_{\text{old}}$$

based on a well known formula for probabilities. Hence, if we have a value for p_{new} we are aiming at, we can dominate it by using $p_{\text{old}} = p_{\text{new}}/n$.

Multiple testing: discussion

A common stated point of view is the following: " multiple testing is crap, I just pick my top 10 genes, they are great."
Another statement is "multiple testing means that the interpretation of a single test depends on what else I did".
Discuss! Do we need p-values? Do they give insight? Does it depend on what kind of answer we are interested in?

Types of data

Statisticians typically divide data into

Categorical a distinct number of different categories. The number of categories is finite. Binary data is an important special case.

Ordinal categorical data that has an underlying order, eg. "bad", "mediocre", "good"

Discrete integer data

Continous continous data.

What type of data is "gene expression data", "high-throughput sequencing data", "genotype data", "copy-number variation data".

Categorical data

We have a distinct finite number of different categories. The common model is the multinomial distribution, with the binomial distribution an important special case when there are only two categories.
The different categories are represented through a probability vector. The number of items in each category is envisioned to have arisen through a series of iid. assignments.
Said differently: we "draw" a number of items independently. Each item is assigned to a category according to the probability vector.
The number of draws is always assumed to be known in advance (otherwise you need different methods than what we have outlined here).
Example: Coin flip. Geno type.

Classic tests

There are a number of classic tests for categorical data.
The first test is called "goodness of fit". We have a probability vector p_0 and we want to examine if the data could have been generated using this probability vector.
Sometimes the probability vector p_0 is a mix of estimated quantities and of know things (this will be clearer later).
The other test is known as the "chisquare test for independence" or "the chisquare test for no association". In this case we observe two categorical variables and we want to assess if there is any dependence between them. An important special case is known as "Fisher's exact test".

The test statistic

We will be using the well known Pearson test statistic, that has the form

$$X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

m is the number of categories, O_i is the observed frequency of category i and E_i is the expected frequency of category i under the hypothesis.
This statistic is easy to explain, and it works well in practice.
There is also a likelihood ratio test statistic. The LR test statistic is also not that hard to understand, but it does require a couple of hours to truely explan. So most often people use the Pearson test statistic above. Often the two statistics are equal to each other.
There is no real reason for prefering the Pearson test statistic, when you use a computer program.

Example: Hardy-Weinberg

We assume that we know that the population frequency of allele A is 0.7. I want to know if my locus is in Hardy-Weinberg equilibrium. I observe the following number of genotypes

AA	226
Aa	165
aa	109
	500

Under the assumption of HW, I get the following expected frequencies

			E_i
AA	np^2	$500 * 0.7^2$	245
Aa	$2np(1-p)$	$2 * 500 * 0.7 * 0.3$	210
aa	$n(1-p)^2$	$500 * 0.3^2$	45

My value of X^2 is 106.26.

The reference distribution

Using asymptotic arguments one can show that the distribution of X^2 approximately is a Chisquare distribution with the degrees of freedom equal to the number of categories - 1 - the number of parameters under the hypothesis (For the HW the degrees of freedom is 1).
For the aproximation to work you need a decent number of expected frequencies in each category.
But (as always): we can also find the distribution through Monte-Carlo simulation. This always works and should be implemented in any decent statistics program (how to do this can be hard, if you want the procedure to be fast).
For our data, the approximation should work fine.

Test for independence

The chisquare test for no association works in a similar fashion. If we let c_j be the marginal probabilities for one variable (column) and we let r_k be the marginal probabilities for the other variable (row), independence says that $p_{j,k} = c_j r_k$.
Watch out: before we used i to index the categories. Now we have a two-dimensional table that we index by (j, k).

$p_{j,k}$	r_k
c_j	

We estimate c_j, r_k be the empirical column/row proportions and plug them into the Pearson test statistic.

Example: genotype vs. disease

We want to investigate whether there is any association between genotype and disease status. Our data are as follows

Disease status	*AA*	*Aa*	*aa*
well	200	234	182
ill	40	100	146

Also: is the locus in Hardy-Weinberg equilibrium? (Estimate the population proportion of A and plug it into the test statistic).
Use the R function chisq.test (it can handle both cases).

Fisher's exact test

Fisher's exact test is just the chisquare test for independence, but in a 2x2 table. In this special case it is possible to compute an exact distribution of the test statistic without using simulation or approximation. This was really exciting 70 years ago, before we could use simulation.
The main comment worth making about 2x2 tables is the fact that there is one – and only one – reasonable measure of association between the two variables (think correlation). This is the odds ratio. This was established at least 60 years ago in the statistics literature,
Unfortunately, we sometimes see people "discovering" "new" ways of measuring dependency in 2x2 tables. Such nonsense is better ignored.