

Classical statistics Issues in their application to ma data

1

Overview of parametric statistics

2

Sampling and testing

Discrete observations

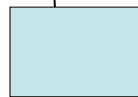
#red = 3



When do I think that I am not sampling from this box anymore?

How many reds could I expect to get just by chance alone!

Random sample of 10 balls from the box



10% red balls and 90% blue balls

3

Sample

Discrete observations

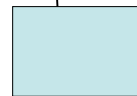
#red = 3

Test statistic



Rejection criteria (based on your observed sample, do you have evidence to reject the hypothesis that you sampled from the null population)

Random sample of 10 balls from the box



10% red balls and 90% blue balls

Null hypothesis (about the population that is being sampled)

4

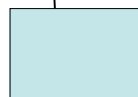
Sample

Continuous observations

mean = 3, sd = 0.6

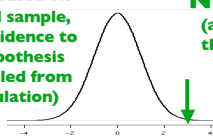
Test statistic

4, 2.3, 5.2, 4.7, 2.1, 3.5,



Rejection criteria (based on your observed sample, do you have evidence to reject the hypothesis that you sampled from the null population)

Null hypothesis (about the population that is being sampled)



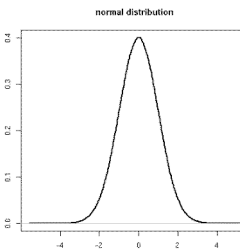
Lets formalise the test!

5

- Distribution from null population is normal, defined by a mean and variance
 - $N(\mu_0, \sigma_0)$
- Take a sample size n , observe
 - $N(\mu_1, \sigma_1)$
- Null hypothesis : random sample comes from population with $N(\mu_0, \sigma_0)$

6

- If $x=c(x_1, \dots, x_n) \in N(\mu_0, \sigma_0)$

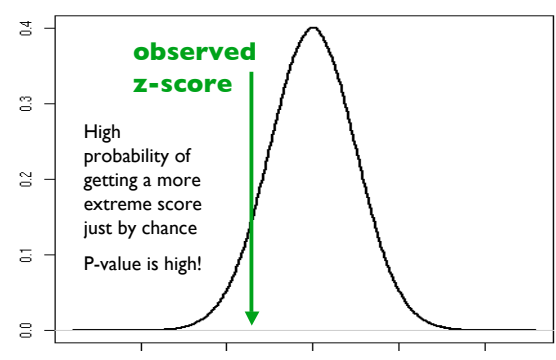
$$\frac{\text{mean}(x) - \mu_0}{\sqrt{\sigma_0/n}} \sim N(0, 1)$$


This standardised statistic is called the z-statistic

Most of the data are between +/- 3 sd's

If you observed a z-score greater than +/- 3 this is evidence that you did not sample from the hypothesized distribution.

normal distribution

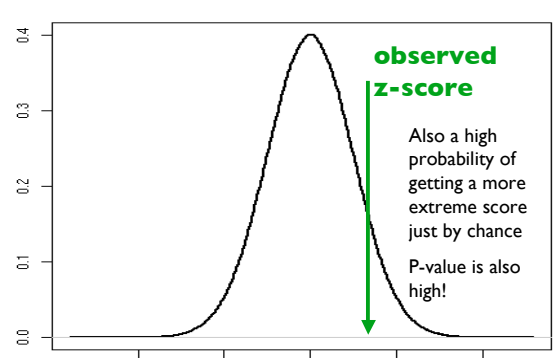


observed z-score

High probability of getting a more extreme score just by chance

P-value is high!

normal distribution

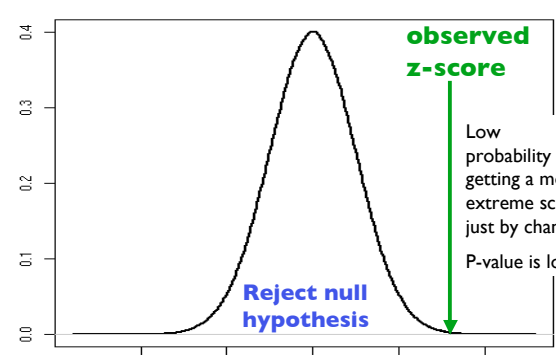


observed z-score

Also a high probability of getting a more extreme score just by chance

P-value is also high!

normal distribution



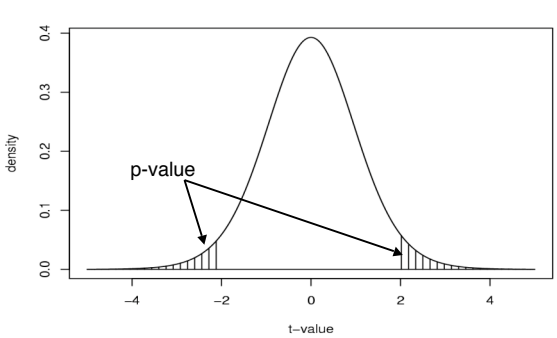
observed z-score

Low probability of getting a more extreme score just by chance

P-value is low

Reject null hypothesis

The p-values for two-sided tests



density

t-value

p-value

The **z-score** is based on the assumption that you are sampling from a normal distribution.

The normality assumption doesn't matter so much when you sample large numbers of observations.

The law of large numbers takes care of things for us. In the long run, the mean of a large number of observations tends to be distributed normally.

The z-score assumes that you **know the variance** of the null distribution - in practise this is rarely the case!!

To account for the increased uncertainty in our sample mean and variance estimates we use a **t-distribution** (not a normal distribution), thus we perform what is called a **t-test**

The distribution of the t-test is similar to the normal distribution, except **the tails of the bell curve are wider**.

This makes sense because we are only estimating the variance, from a small number of observations, so we'll get more extreme scores by chance.

As n gets really large, our sample variance estimate gets better and the t-distribution gets closer to the normal $(0,1)$.

But what do we do, if we don't think we are sampling from a normal distribution, or if we have outliers.

We cannot assume a distribution for the mean - we assume no distribution - we use **non-parametric statistics**.

13

non-parametric statistics in brief

14

Non-parametric tests usually rely on statistics based on the rank of observations.

Wilcoxon signed rank test : 1 sample test

Wilcoxon rank sum test : 2 sample test

(tests if two samples are equivalent by checking that the ranks of all observations are equally distributed between the two samples)

Kruskal Wallis test : n samples test

(tests if many samples are equal)

All of these tests require a reasonable sample size (not usual with microarray data)

15

Issues with classical statistics for microarray data

16

First issue is : we cannot assume that relative gene expression measurements (observed M-values) are emitted from a normal distribution.

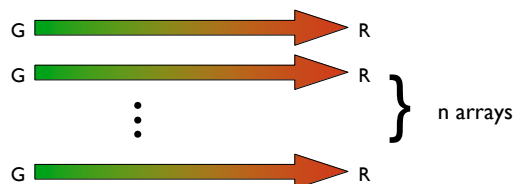
In fact M-values have a very heavy tailed distribution (i.e. we get a lot more observations in the tails of the bell curve). If we only had one gene, a t-test would be fine, but because we have thousands of t-tests, and each gene comes from a different distribution (different variance) then a t-test isn't so helpful anymore.

Lets understand why!

17

Sample size from replicate arrays

The sample size for each gene is dependent on the amount of replication for each type of array. We must have replication to do statistical analysis of differential expression. We will consider first, the basic problem of all arrays comparing the same two samples (RNA sources).



18

Gene-wise summaries

- Each gene give a series of log-ratios
- Summarize log-ratios by the average and standard deviation for each gene

$$M_1, \dots, M_n$$

$$M = \text{ave}M \quad s = \text{st.dev } M$$

19

Summarising replicates to determine differential expression

Obvious thing : average M's

$$\text{av}M$$

But averages can be driven by outliers

Better than that : account for variability

$$t = \text{av}M / \text{SE}$$

But with 10,000 or so genes, some will have very small SE

Better still : use smoothed SE's

$$t^* = \text{av}M / \text{SE}^*$$

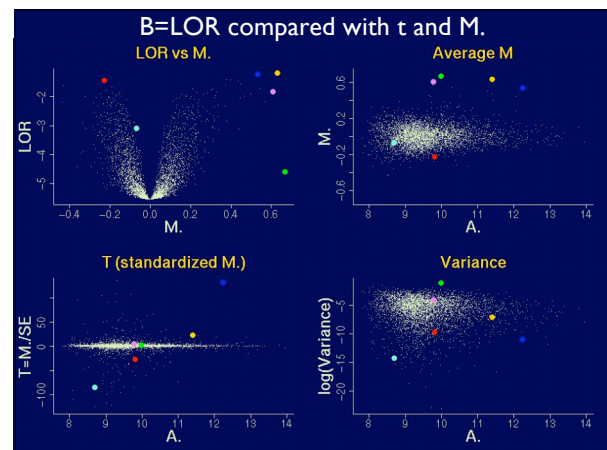
This is a modified t-statistic (also referred to as a moderated t). A similar statistic often used is the B statistic (it is the log odds ratio of differential expression, i.e. LODS, or LOR). The B statistic smoothes the standard errors it uses in its calculation. We call this Empirical Bayes smoothing!

20

Borrowing information from genes

- When there are thousands of genes we can get a better idea of the variability than from just the individual gene variance estimates
- We can't borrow information when there are only a few genes, but when there are tens of thousands of genes we can.
- We want a compromise between individual gene variance estimates and a single variance estimate for all genes.
- The compromise is achieved by empirical Bayes methods which give a weighted combination...

21



Summary

- Microarray experiments typically have thousands of genes, but only few (1-10) replicates for each gene.
- Averages can be driven by outliers.
- t-statistics can be driven by tiny variances.
- B (or moderated t-statistic)
 - use information from all the genes
 - combine the best of M. and t
 - avoid the problems of M. and t

Ranking on B could be helpful.

23

What we want to do is...

- Analyse data all at once
- Use standard deviances not just fold changes
- Use ensemble information to shrink variances
- Assess differential expression for all comparisons together (because microarray experiments will rarely be just a simple comparison between two samples)

24

Ranking is easier

- How many genes are differentially expressed?
- If there was only one gene, a t-test would give a reliable P-value for judging whether the true log-ratio was zero
- With so many genes, computing absolute P-values on the basis of probability models is problematic
- Much easier to simply rank the genes in order of evidence for differential expression

25

Why judging significance is hard

- Log-ratios aren't normally distributed, hard to check log-ratios for different genes are correlated in unknown way
- High level of multiple testing means that very small p-values are required – distributional assumptions must hold in extreme tail

26

Choosing a cut-off

- Could choose a threshold for differential expression if there were known DE and non-DE genes
- Print artificial genes on microarray, then spike corresponding RNA into target RNA before labelling and hybridization
- Choose a cut-off that seem sensible!! Careful and thorough graphical exploration and the choice of ranking statistic are probably the most important aspect to choosing DE. Follow up experimentation that the biologist intends to perform will also play an important role.

27

Multiple samples

Finding differentially expressed genes when there are more than one kind of array comparison

In classical statistics,
t-test for 1 sample
t-test for 2 samples
regression
anova and
linear models...

28

Extensions include dealing with

- Replicates within and between slides
- Several effects: use a linear model
- ANOVA: are the effects equal?
- Time series: selecting genes for trends

29