Lecture 7: Hypothesis Testing

Jing Ma, Statistics, TAMU

21 October, 2019

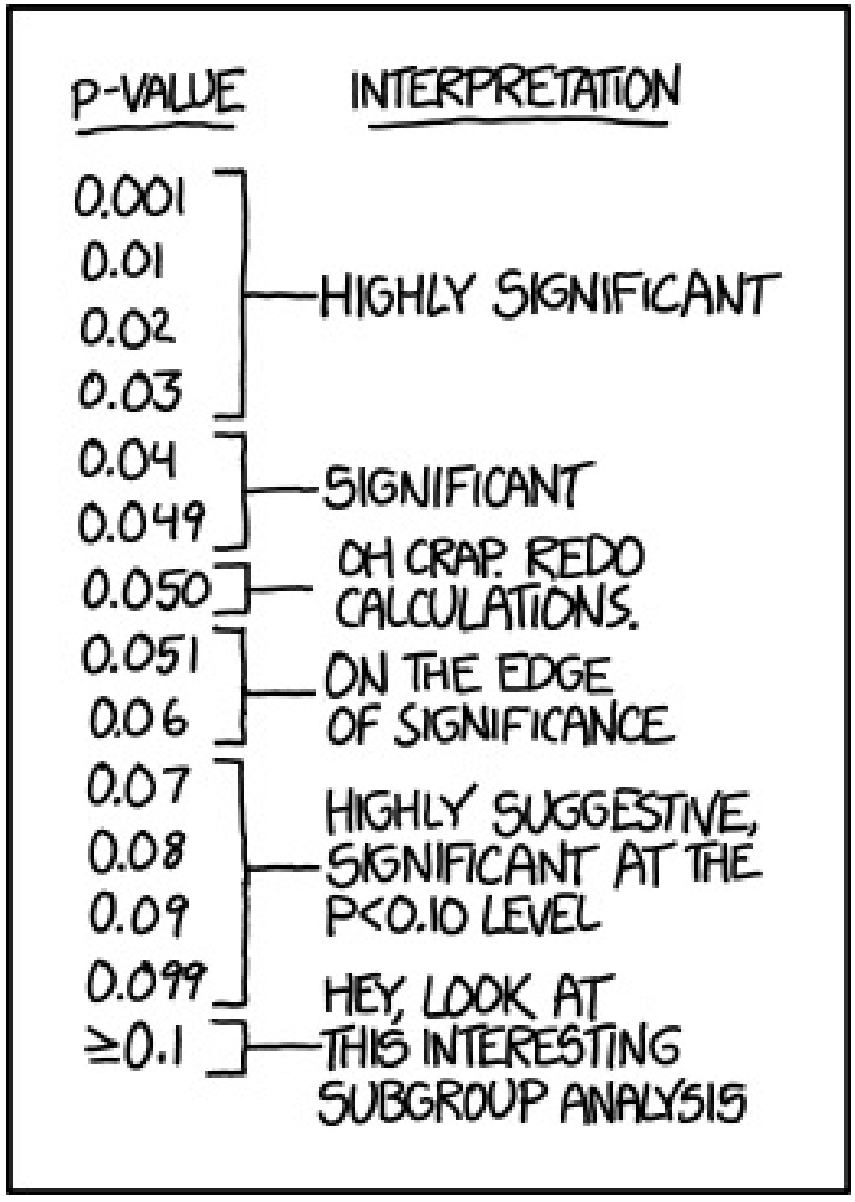
Recap

- Different notions of **distances**
- k-means and hierarchical clustering
- How to determine the number of clusters

This lecture

- What is a p-value?
- Understand the basic principles of hypothesis testing, its pitfalls, strengths, use cases and limitations.
- What changes when we go from single to multiple testing?
- False discovery rates and p-value 'adjustments'.

Hypothesis testing



Statistical hypothesis testing

Questions we may ask:

- Does the mean gene expression over ALL patients differ from that over AML patients?
- Is the mean gene expression different from zero?
- To what extent are gene expression values normally distributed?
- Are there outliers among a sample of gene expression values?
- How can it be tested whether the frequencies of nucleotide sequences of two genes are different?

Statistical hypothesis testing

- Population parameters were used to define theoretical distributions.
- In any research setting, the specific values of such parameters are unknown so they must be estimated.
- Once estimates are available it becomes possible to statistically test biologically important hypotheses.

Statistical hypothesis testing

- Let μ_0 be a number representing the hypothesized population mean by a researcher on the basis of experience and knowledge from the field.

Statistical hypothesis testing

- Let μ_0 be a number representing the hypothesized population mean by a researcher on the basis of experience and knowledge from the field.
- With respect to the population mean, the null hypothesis can be formulated as $H_0 : \mu = \mu_0$ and the alternative hypothesis as $H_1 : \mu \neq \mu_0$.

Statistical hypothesis testing

- Let μ_0 be a number representing the hypothesized population mean by a researcher on the basis of experience and knowledge from the field.
- With respect to the population mean, the null hypothesis can be formulated as $H_0 : \mu = \mu_0$ and the alternative hypothesis as $H_1 : \mu \neq \mu_0$.
- The alternative hypothesis is true if $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$ holds true. This type of hypothesis is called "two-sided".

Statistical hypothesis testing

- Let μ_0 be a number representing the hypothesized population mean by a researcher on the basis of experience and knowledge from the field.
- With respect to the population mean, the null hypothesis can be formulated as $H_0 : \mu = \mu_0$ and the alternative hypothesis as $H_1 : \mu \neq \mu_0$.
- The alternative hypothesis is true if $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$ holds true. This type of hypothesis is called "two-sided".
- In case $H_1 : \mu > \mu_0$, it is called "one-sided".

Statistical hypothesis testing

- Such a null hypothesis will be statistically tested against the alternative using a suitable distribution of a statistic (e.g. standardized mean).

Statistical hypothesis testing

- Such a null hypothesis will be statistically tested against the alternative using a suitable distribution of a statistic (e.g. standardized mean).
- After conducting the experiment, the value of the statistic can be computed from the data.

Statistical hypothesis testing

- Such a null hypothesis will be statistically tested against the alternative using a suitable distribution of a statistic (e.g. standardized mean).
- After conducting the experiment, the value of the statistic can be computed from the data.
- By comparing the value of the statistic with its distribution, the researcher draws a conclusion with respect to the null hypothesis: H_0 is rejected or it is not.

Statistical hypothesis testing

- Such a null hypothesis will be statistically tested against the alternative using a suitable distribution of a statistic (e.g. standardized mean).
- After conducting the experiment, the value of the statistic can be computed from the data.
- By comparing the value of the statistic with its distribution, the researcher draws a conclusion with respect to the null hypothesis: H_0 is rejected or it is not.
- The probability to reject H_0, given the truth of H_0, is called the significance level which is generally denoted by α. We shall follow the habit in statistics to use $\alpha = 0.05$, but it will be completely clear how to adapt the procedure in case other significance levels are desired.

The Z-test

The Z-test applies to the situation where we want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ and the standard deviation σ is known.

- Assume that gene expression values $x_1, \ldots, x_n$ are from a normal distribution.
- We compute the standardized value $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma$.
- The p-value equals

$$P(Z \leq -|z|) + P(Z \geq |z|) = 2 \cdot P(Z \leq -|z|).$$

- The conclusion: If the p-value is larger than the significance level α, then H_0 is not rejected; if it is smaller than the significance level, then H_0 is rejected.

Example on the Z-test

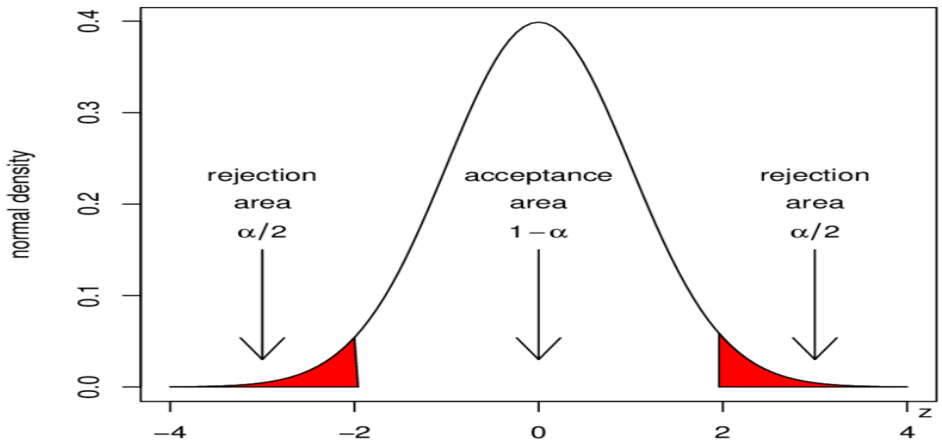
- We use the gene expression value from row 2058 of the **Golub** data set.
- A quick search through the NCBI site suggests that this gene may not be directly related to leukemia. Hence, we may hypothesize that the population mean of the ALL gene expression values equals zero; or $H_0 : \mu = 0$
- For the sake of illustration, we shall pretend that we know the standard deviation $\sigma = 0.25$.
- How to compute the z-value and test H_0?

Example on the Z-test

```
data(golub, package = "multtest")
gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
sigma <- 0.25;
n <- 27;
mu0 <- 0
x <- golub[2058,gol.fac=="ALL"]
z.value <- sqrt(n)*(mean(x) - mu0)/sigma
2*pnorm(-abs(z.value),0,1)
```

Since it is clearly larger than 0.05, we conclude that the null hypothesis of mean equal to zero is not rejected. There is not enough evidence to reject the null.

Acceptance and rejection regions of the Z-test



- If z falls in the interval $(z_{0.025}, z_{0.975})$, then H_0 is not rejected and consequently this region is called "acceptance region".

Confidence interval

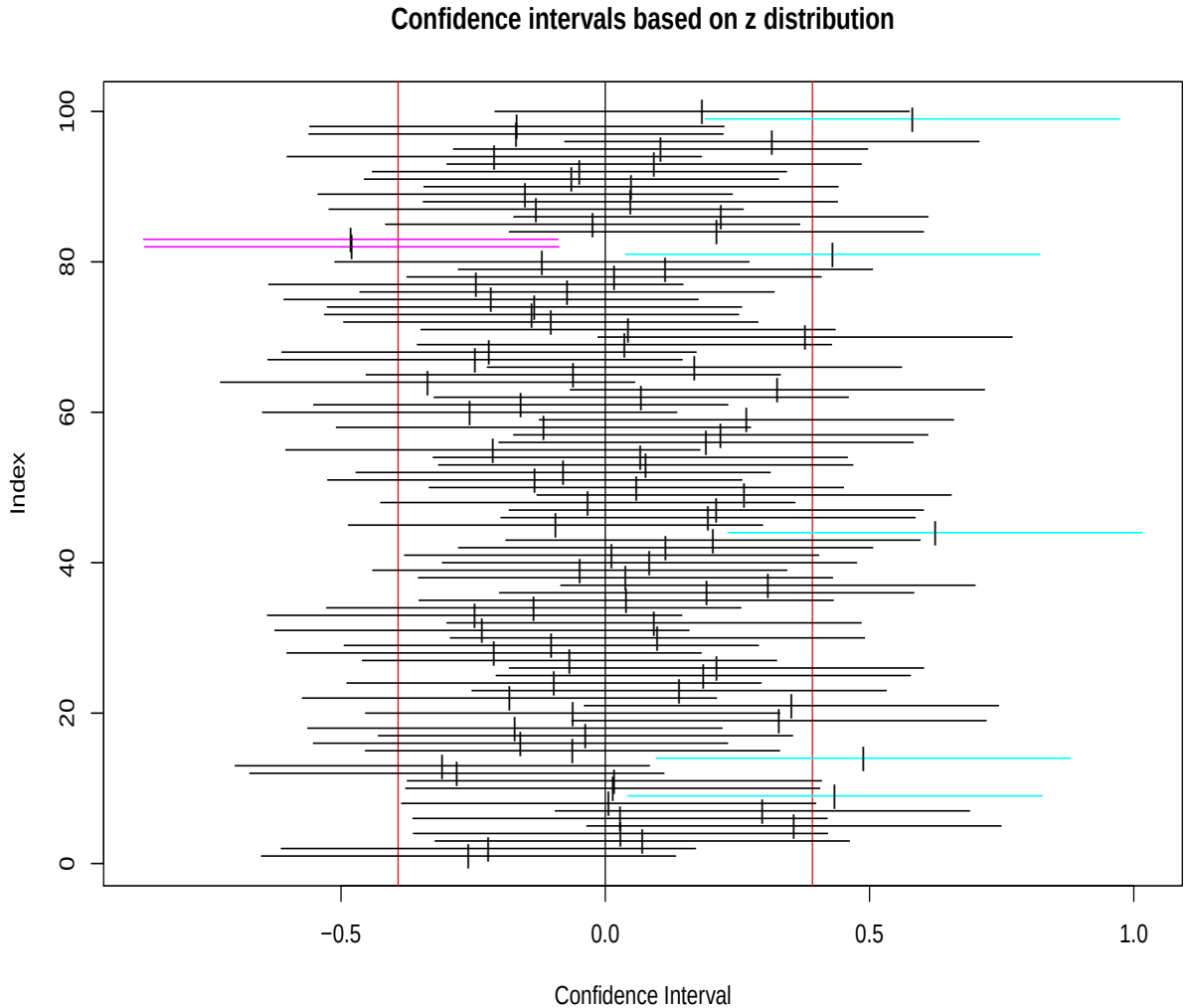
- The interval $(z_{0.025}, z_{0.975})$ is often called the confidence interval, because if the null hypothesis is true, then we are 95% confident that the observed z-value falls in it.
- It is custom to rework the confidence interval into an interval with respect to μ. In particular, the 95% confidence interval for the population mean μ is

$$\left(\bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}}\right).$$

- `z.test` in **TeachingDemos** package

Confidence interval (cont.)

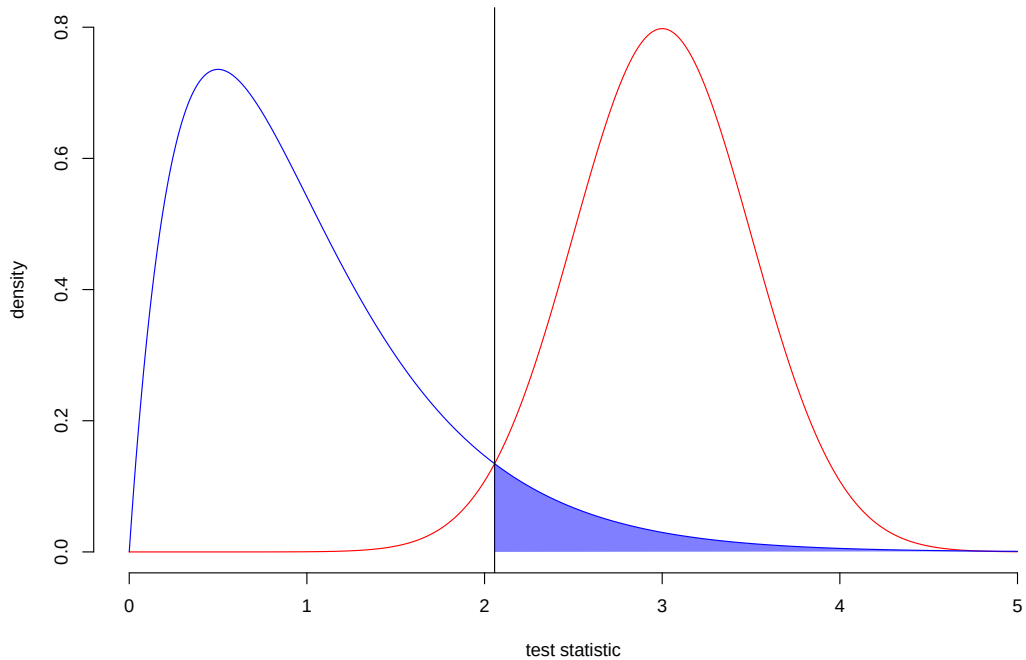
```
ci.examp(mean.sim =0, sd = 1, n = 25, reps = 100,
   method = "z", lower.conf=0.025, upper.conf=0.975)
```



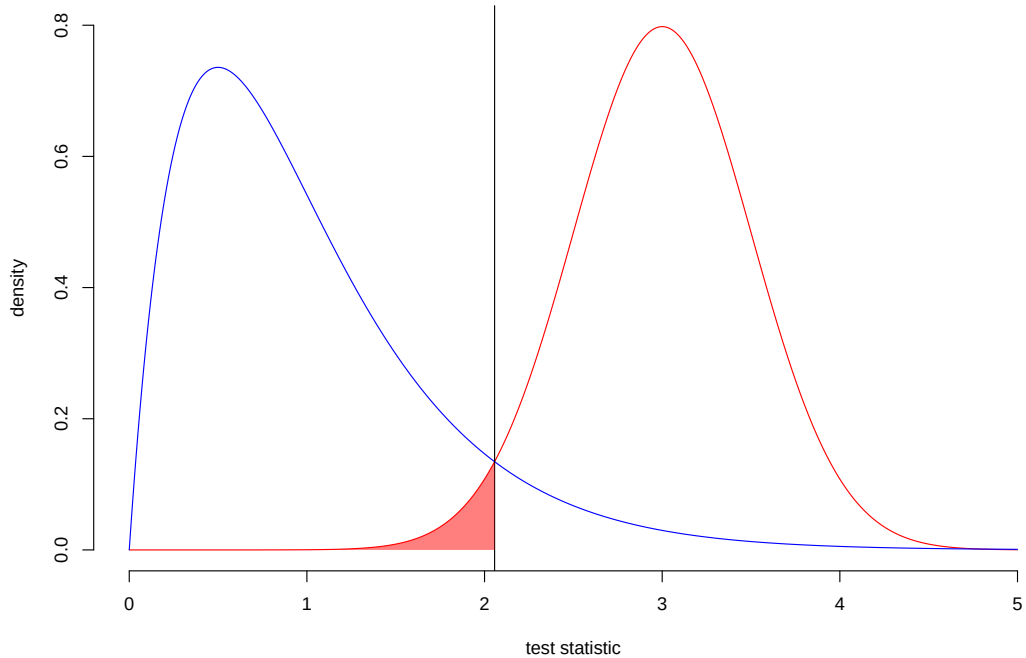
Confidence interval

- 100 samples of size 25 from the $N(0, 1)$ distribution are drawn and for each of these the confidence interval for the population mean is computed and represented as a line segment.
- Apart from sampling fluctuations, the confidence level corresponds to the percentage of intervals containing the true mean (colored in black) and that the significance level corresponds to intervals not containing it (colored in red or blue).

False positive rate (type I error)

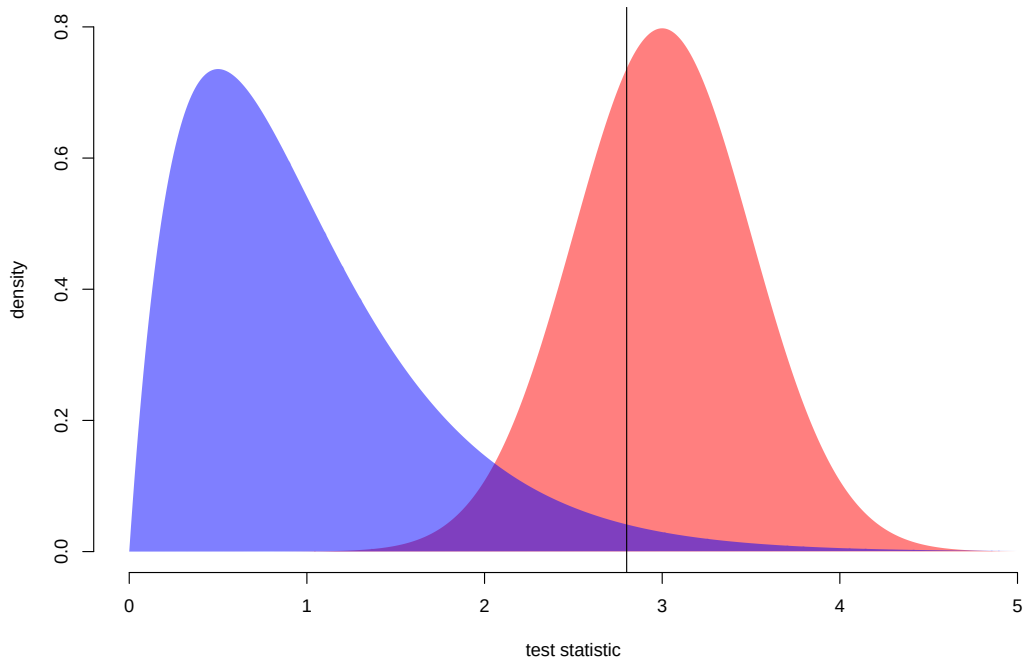


False negative rate (type II error)



Question

What if we move the decision boundary as follows? Do you think it is a better or worse decision rule?



Types of error in testing

Test vs reality	**Null hypothesis is true**	**. . . is false**
Reject null hypothesis	Type I error (false positive)	True positive
Do not reject	True negative	Type II error (false negative)

One-sample t-test

- In almost all research situations with respect to gene expression values, the population standard deviation σ is unknown so that the Z-test is not applicable.
- In such cases *t*-tests are very useful for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
- Test statistic: $t = \sqrt{n}(\bar{x} - \mu_0)/s$.
- p-value: $2 \cdot P(T_{n-1} \leq -|t|)$.
- 95% Confidence interval for the population mean

$$\left(\bar{x} + t_{0.025,n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{0.975,n-1} \frac{s}{\sqrt{n}} \right).$$

Example on one-sample t-test

Let's test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ for the ALL population mean of the Gdf5 gene expressions. The latter is collected in row 2058 of the golub data.

```
x <- golub[2058,gol.fac=="ALL"]; mu0 <- 0; n <- 27
t.value<-sqrt(n)*(mean(x) - mu0)/sd(x)
t.value

2*pt(0.0010,26)

# Now compare with t.test
t.test(x,mu=0)
t.test(x,mu=0,alternative = 'greater')
```

Two-sample t-test with unequal variances

- Suppose that gene expression data from two groups of patients (experimental conditions) are available and that the hypothesis is about the difference between the population means μ_1 and μ_2.
- $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$
- Reformulate hypothesis $H_0 : \mu_1 - \mu_2 = 0$ and $H_1 : \mu_1 - \mu_2 \neq 0$.
- Given data $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$, let $\bar{x}$ and $\bar{y}$ be the mean of the respective group, and s_1^2 and s_2^2 be the variance of the respective group. Then the t-statitic is

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

Example on two-sample t-test

Golub et al. (1999) argue that gene CCND3 Cyclin D3 plays an important role with respect to discriminating ALL from AML patients, which translates to this particular gene has different expressions among ALL compared to the AML patients.

The null hypothesis of equal means can be tested by the function `t.test` and the appropriate factor and specification `var.equal=FALSE`.

```
t.test(golub[1042,] ~ gol.fac, var.equal=FALSE)
```

Since the p-value is extremely small, the conclusion is to reject the null-hypothesis of equal means. The data provide strong evidence that the population means differ.

Two sample t-test with equal variances

- Consider the same setting as before, but now the variances σ_1^2 and σ_2^2 for the two groups are known to be equal.
- $H_0 : \mu_1 = \mu_2$
- Test statistic

$$t = \frac{\bar{x} - \bar{y} - \mu_1 - \mu_2}{s_p \sqrt{1/n_1 + 1/n_2}},$$

where the pooled sample variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Example on two-sample t-test

We test the same null hypothese for gene CCND3 Cyclin D3 that the mean of the ALL differs from that of AML patients.

```
t.test(golub[1042,] ~ gol.fac, var.equal=TRUE)
```

F-test on equal variances

In case of any uncertainty about the validity of the assumption of equal population variances, one may want to test this.

- Null hypothesis: the two population variances are equal, or $H_0 : \sigma_1^2 = \sigma_2^2$.
- This can be done using the F-test.
- From the sample variances s_1^2 and s_2^2, the test statistic $f = s_1^2/s_2^2$ follows F_{n_1-1,n_2-1} distribution.
- H_0 is not rejected if

$$P(F_{n_1-1,n_2-1} < f) \geq \alpha/2, f < 1$$

or

$$P(F_{n_1-1,n_2-1} > f) \geq \alpha/2, f > 1$$

```
var.test(golub[1042,] ~ gol.fac)
```

Chi-squared test

One may want to test multiple probabilities where
$H_0 : (\pi_1, \ldots, \pi_m) = (p_1, \ldots, p_m)$ against
$H_1 : (\pi_1, \ldots, \pi_m) \neq (p_1, \ldots, p_m)$.

- Expected number of observations $e_i = n \cdot p_i$ where n is the total number of observations.
- Observed number o_i
- Test statistic

$$q = \sum_{i=1}^{m} (o_i - e_i)^2 / e_i$$

  follows a Chi-squared (χ^2_{m-1}) distribution with $m - 1$ degrees of freedom.
- The p-value is $P(\chi^2_{m-1} \geq q)$.

Rejection region of Chi-squared test

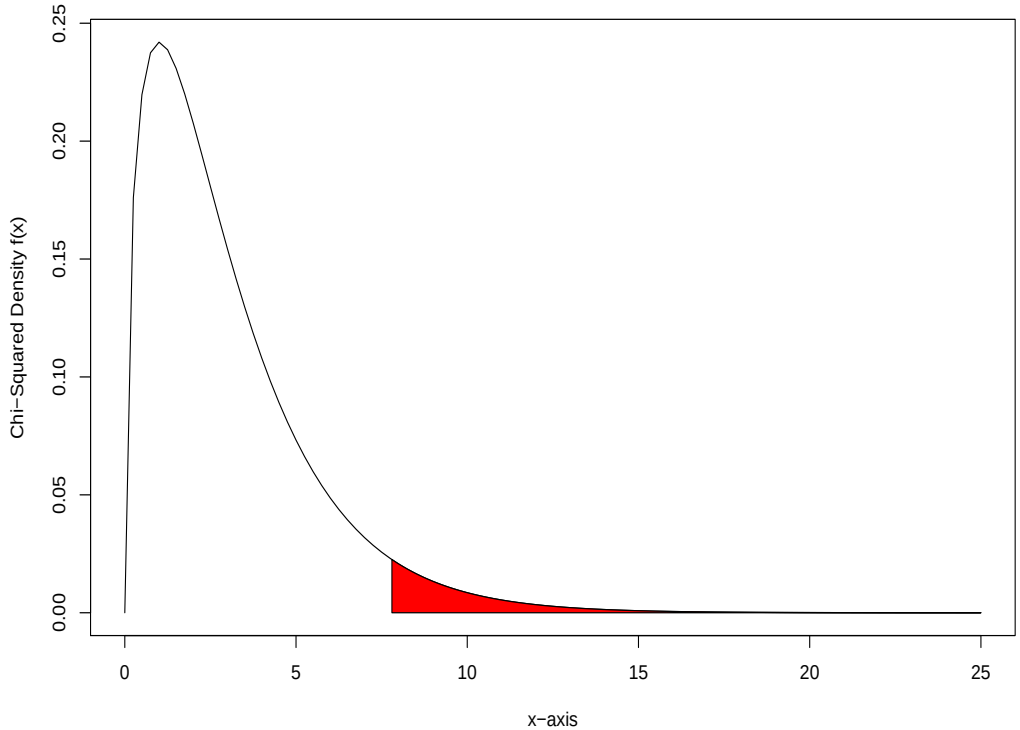


Fig. 1: Rejection region of Chi-squared test with df=3 and q=7.8.

Example on Chi-squared test

- We want to test the hypothesis that the nucleotides of Zyxin have equal probability. Let the probability of $\{A, C, G, T\}$ to occur in the sequence be $(\pi_1, \pi_2, \pi_3, \pi_4)$.
- Null hypothesis $H_0 : (\pi_1, \pi_2, \pi_3, \pi_4) = (1/4, 1/4, 1/4, 1/4)$.
- The observed frequencies are

```
##
##   a   c   g   t
## 410 789 573 394
```

- We can use `chisq.test` in R to test the null hypothesis.

```
chisq.test(zyxinfreq)
```

Normality tests

We can use the Shapiro-Wilk test to test the hypothesis that a data set is normally distributed.

```
shapiro.test(golub[1042, gol.fac=="ALL"])
```

Since the p-value is greater than 0.05, the conclusion is not to reject the null hypothesis that CCND3 Cyclin D3 expression values follow from a normal distribution.

Wilcoxon rank test

If data are not normally distributed, then the t-test is not "good" for testing $H_0 : \mu_1 = \mu_2$.

- The type I error may not be controlled.

The alternative is Wilcoxon rank test.

- The null hypothesis is $H_0 : F = G$, i.e. whether the distribution F in the first group is equal to the distribution of the second group G.

```
wilcox.test(golub[1042,] ~ gol.fac)
```

Avoid fallacy

The p-value is not the probability that the null hypothesis is true.
Absence of evidence $\neq$ evidence of absence.

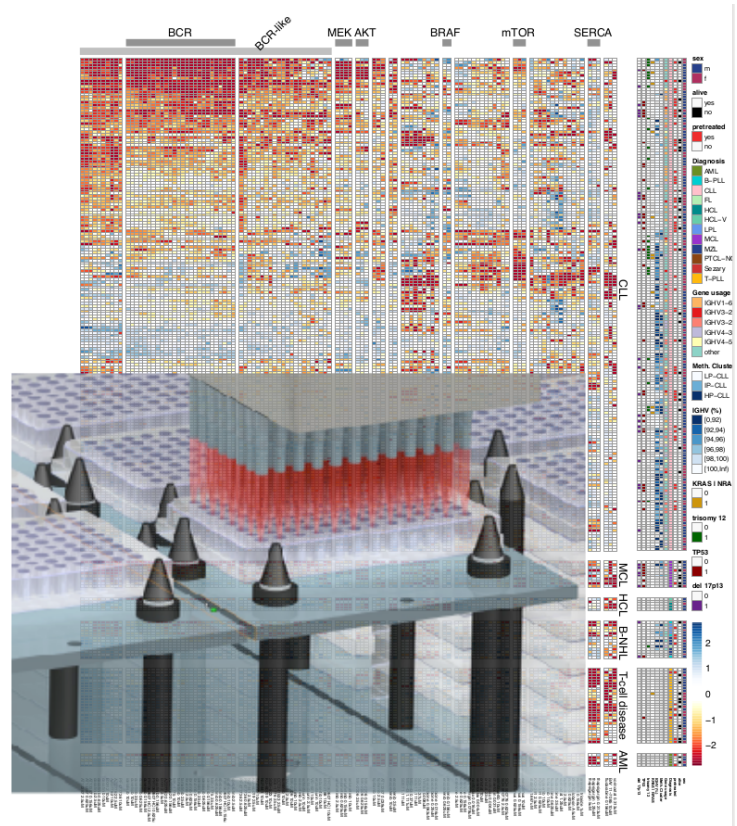
Summary: single hypothesis testing

- P-values are random variables: uniformly distributed if the null hypothesis is true – and should be close to zero if the alternative holds. Note: We only observe one draw.
- We prove something by disproving ('rejecting') the opposite (the null hypothesis). Reject = Discover.
- Not rejecting does not prove the null hypothesis.
- Repeating the experiment (under the null): around 5% of the times the p-value will be less than 0.05 by chance.
- All this reasoning is probabilistic. Testing and p-values are for rational decision making in uncertain contexts.

Multiple testing

Many data analysis approaches in genomics employ item-by-item testing.

- Differential expressed genes
- Genome-wide association studies
- Variant calling



Multiple testing

xkcd cartoon

- Why didn't the newspaper report the results for the other colors?

Multiple testing

xkcd cartoon

- Why didn't the newspaper report the results for the other colors?
- Maybe we really care about each jelly bean color. Or maybe we just wanted to publish something.

What is p-value hacking?

- On the same data, try different tests until one is significant.

What is p-value hacking?

- On the same data, try different tests until one is significant.
- On the same data, try different hypotheses until one is significant (hypothesizing after results are known).

What is p-value hacking?

- On the same data, try different tests until one is significant.
- On the same data, try different hypotheses until one is significant (hypothesizing after results are known).
- Moreover, retrospective data picking, “outlier”" removal, the 5% threshold and publication bias.

What is p-value hacking?

- On the same data, try different tests until one is significant.
- On the same data, try different hypotheses until one is significant (hypothesizing after results are known).
- Moreover, retrospective data picking, "outlier"" removal, the 5% threshold and publication bias.

What can we do about this?

The multiple testing burden

When performing multiple tests, type I error goes up: for $\alpha = 0.05$ and n independent tests, probability of no false positive result is

$$\underbrace{0.95 \times 0.95 \times \cdots \times 0.95}_{n \text{ times}} \ll 0.95.$$

Family wise error rate (FWER)

- Bonferroni correction controls FWER.
- For m hypothesis tests, multiply each p-value by m. Then see if anyone still remains below 0.05.

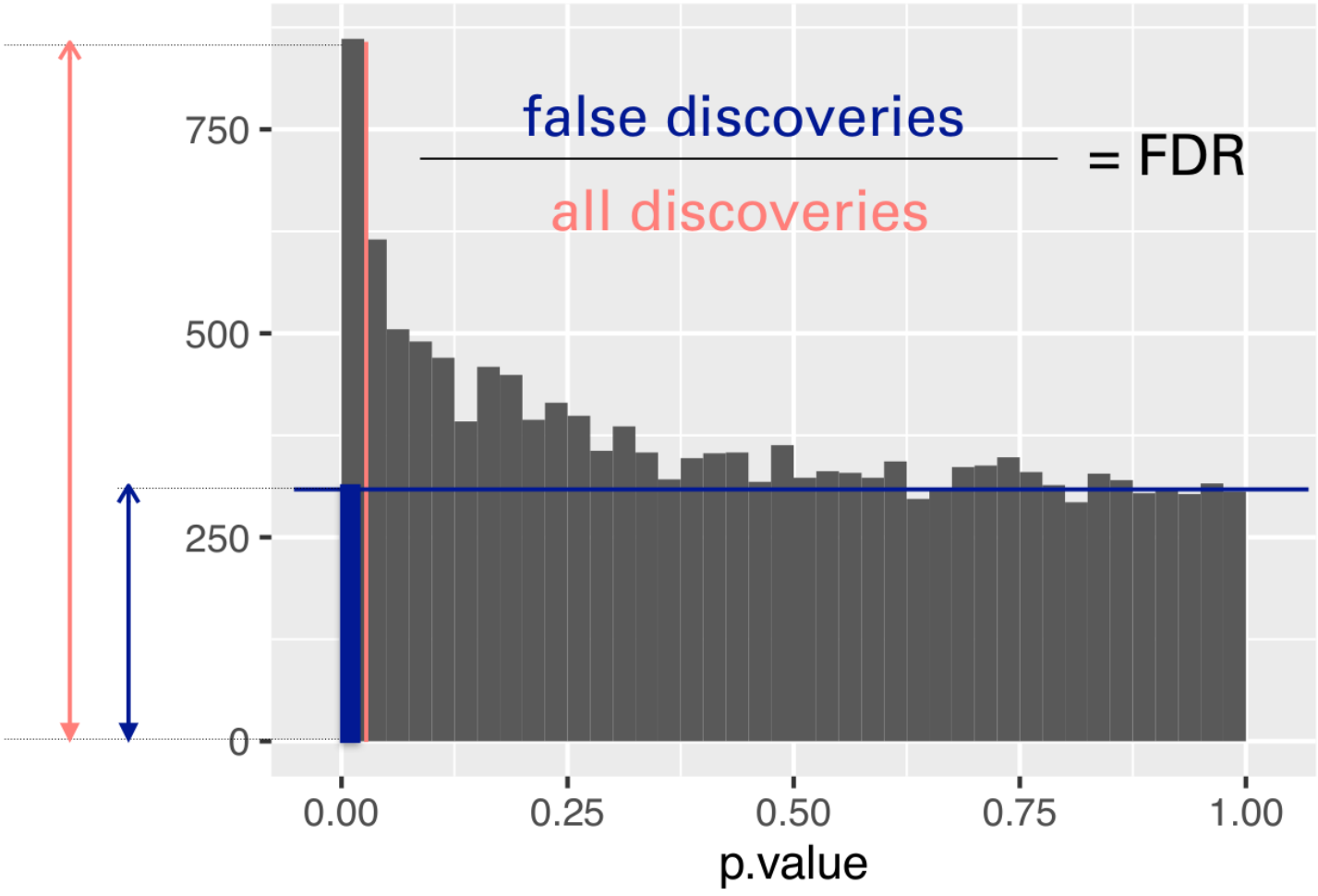
False discovery rate (FDR)

Alternatively, we can control for the false discovery rate (FDR), where

$$\text{FDR} \approx \frac{\text{false discoveries}}{\text{all discoveries}}.$$

(FDR $= 0$ if we make no discoveries.)

Method of Benjamini and Hochberg



Method of Benjamini and Hochberg

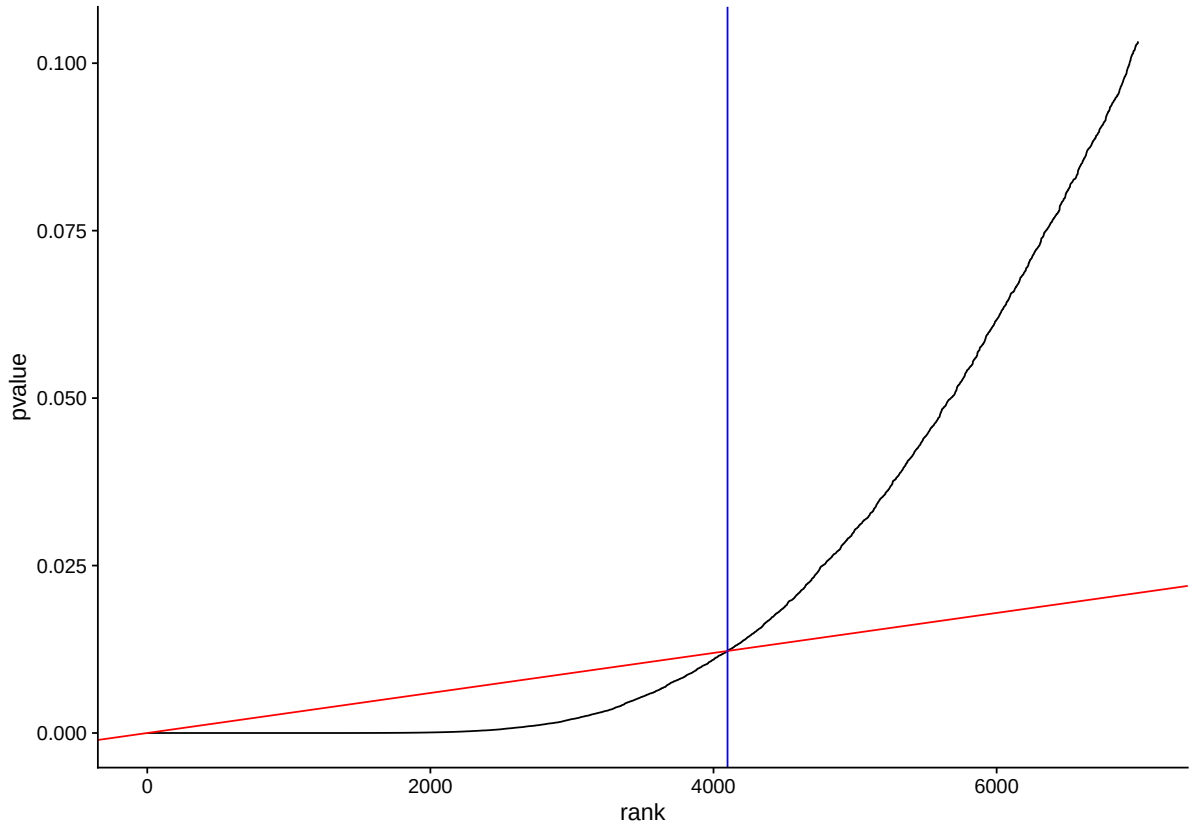


Fig. 2: Visualization of the Benjamini-Hochberg procedure. Shown is a zoom-in to the 7000 lowest p-values. Red line has the slope alpha/m.

Method of Benjamini and Hochberg

In R, we use the function `p.adjust` to adjust p-values obtained from testing multiple hypotheses.

```
p.adjust(p, method = "BH")
```

Experiment-wide type I error rates

Test vs Reality	**Null Hypothesis is true**	**...is false**	**Total**
Rejected	V	S	R
Not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m: total number of hypotheses
- m_0: number of null hypotheses
- V: number of false positives (a measure of type I error)

- **Family-wise error rate (FWER)** refers to the probability of one or more false positives, $P(V > 0)$.
- For large m_0, this is difficult to keep small.

Experiment-wide type I error rates

Test vs Reality	**Null Hypothesis is true**	**...is false**	**Total**
Rejected	V	S	R
Not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m: total number of hypotheses
- m_0: number of null hypotheses
- V: number of false positives (a measure of type I error)

- **False discovery rate (FDR):** the expected fraction of false positives among all discoveries,

$$E[\frac{V}{\max\{R, 1\}}].$$

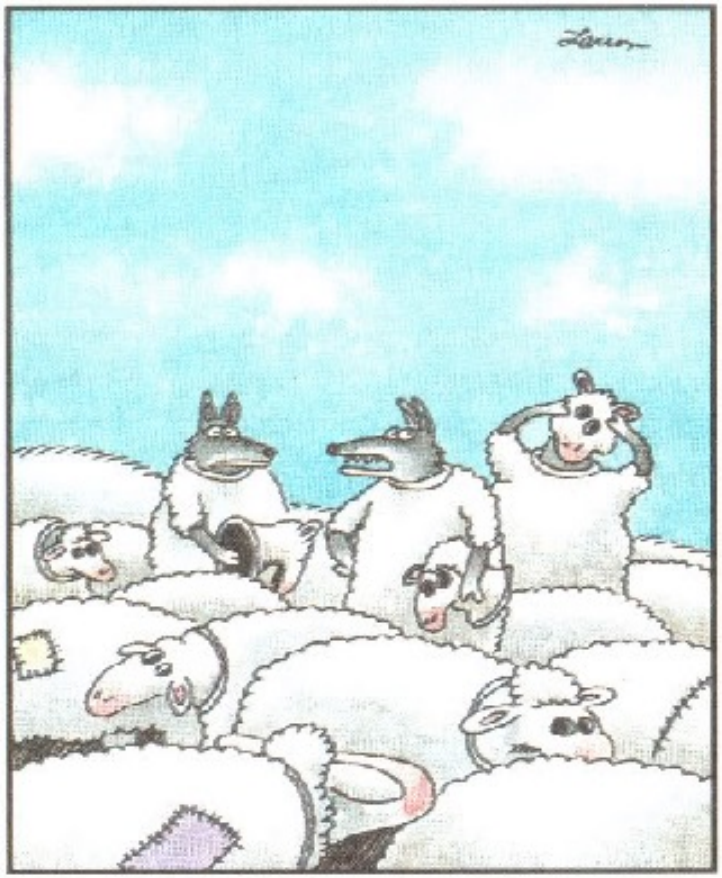
- If $m_0 = m$, then FDR = FWER.

False positive rate vs false discovery rate

- **FPR**: fraction of FP among all negatives.
- **FDR**: fraction of FP among discoveries called.

Example: 20,000 genes, 500 differentially expressed, 100 discoveries called, 10 of them wrong.

- **FPR**: $10/19{,}500 \approx 0.05$
- **FDR**: $10/100 = 10$



"Wait a minute! Isn't anyone here a real sheep?"

An RNA-seq dataset

- Data: gene expression measurements (gene-level counts) of four primary human airway smooth muscle cell lines with and without treatment with dexamethasone, a synthetic glucocorticoid.
- For each gene, we perform a test for differential expression. Conceptually, the tested null hypothesis is very similar to that of the t-test (details in later lecture).
- Reference: Himes et al. (2014). "RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells." PLoS ONE, 9(6), e99625

R code for DESeq analysis (run yourself)

```
library("DESeq2")
library("airway")
data("airway")
aw   = DESeqDataSet(se = airway, design = ~ cell + dex)
aw   = DESeq(aw)
awde = as.data.frame(results(aw)) %>% dplyr::filter(!is.na(pvalue))
head(awde)

ggplot(awde, aes(x = pvalue)) +
  geom_histogram(binwidth = 0.025, boundary = 0)

p.BH = p.adjust(awde$pvalue, method="BH")
sum(p.BH<0.1)
```

Multiple testing with `multtest` package

We will illustrate some functionality of `multtest` with gene expression data from the leukemia ALL/AML study of Golub et al. (1999). Load the leukemia dataset:

```
require(multtest)
rm(list=ls())
data(golub)
class(golub)
dim(golub)
head(golub[,1:5])
```

Leukemia gene expression data

- Note that each column is a sample, and `golub[j, i]` is the expression level for gene j in tumor mRNA sample i.
- There are also gene identifiers and tumor class labels (0 for ALL, 1 for AML).

```
dim(golub.gnames)
golub.gnames[1:4, ]
golub.cl
```

Computing simple test statistics

- The `mt.teststat` and `mt.teststat.num.denum` functions provide a convenient way to compute test statistics for each row of a data frame, e.g., two-sample Welch t-statistics, Wilcoxon statistics, F-statistics, paired t-statistics, and block F-statistics.

```
teststat = mt.teststat(golub, golub.cl)
require(ggplot2)
plt = ggplot(data.frame(teststat), aes(sample = teststat)) + stat_qq() +
  theme_bw()
plt
```

Question 1

(Multiple Choice) What can we say about those points on the plot that look like outliers?

a. They could correspond to genes whose expression levels differ between the ALL and AML groups.

b. They definitely correspond to genes whose expression levels differ between the ALL and AML groups.

c. We cannot say anything

Computing the numerators and denominators of the test statistics

```
parts = mt.teststat.num.denum(golub, golub.cl)
names(parts)

head(parts$teststat.num)
head(parts$teststat.denum)
teststatNew = parts$teststat.num/parts$teststat.denum
plt = ggplot(data.frame(teststat, teststatNew),
             aes(x = teststat, y = teststatNew)) +
    geom_line() + geom_abline(colour = "red", linetype = "dotted", size = 2) +
    theme_bw()
plt
```

Question 2

(Multiple Response) Which of the following commands can be used to check that two numeric vectors or matrices are identical?

a) `identical(A, B)`

b) `max(abs(A-B)) == 0`

Question 3

(Multiple Response) The commands in Question 2 work for the vast majority of cases. But there are exceptions. Run the following code:

```
identical(2^100, 2^100 - 1)
identical(2^20, 2^20 - 1)
```

What should we be mindful of when comparing two numbers?

a) Their relative sizes

b) Their absolute sizes

Adjusting p-values

- The `mt.rawp2adjp` function computes adjusted p-values for simple multiple testing procedures from a vector of raw (unadjusted) p-values.
- The procedures include the Bonferroni, Holm (1979), Hochberg (1988), and Sidak procedures for strong control of the family-wise Type I error rate (FWER), and the Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) procedures for (strong) control of the false discovery rate (FDR).
- First we will compute raw nominal two-sided p-values. For this data, we'll assume that it's safe to use a standard normal distribution for the 3,051 test statistics.

Adjusting p-values

```
rawp = 2 * (1 - pnorm(abs(teststat)))
procedures = c("Bonferroni", "Holm", "Hochberg", "SidakSS", "SidakSD",
               "BH", "BY")
adjusted = mt.rawp2adjp(rawp, procedures)
adjusted$adjp[1:10, ]
```

The results are stored in increasing order of the raw p-values.

To display them based on the original data order, use

```
adjusted$adj[order(adjusted$index)[1:10], ]
```

The mt.plot function

```
res = mt.rawp2adjp(rawp, c("Bonferroni", "BH", "BY"))
allp = res$adjp[order(res$index), ]
procs = dimnames(allp)[[2]]
cols = c(1, 2, 5, 6)
ltypes = c(1, 2, 3, 3)
mt.plot(allp, teststat, plottype = "pvst", logscale = TRUE, proc = procs,
    leg = c(7.5, 4), pch = ltypes, col = cols)
```

The mt.reject function

```
ord = order(res$index)
rawp = res$adjp[ord,1]
BH = res$adjp[ord,3]
mt.reject(cbind(rawp, BH), seq(0, 1, 0.1))$r
```

Summary

- t-tests are common!
- Multiple testing is not a problem but an opportunity.
- Bonferroni and FDR correction for multiple testing.

Further reading

- Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science 286.5439 (1999): 531-537.
- Himes, Blanca E., et al. "RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells." PloS One 9.6 (2014): e99625.