Statistics for Differential Expression in Sequencing Studies

Naomi Altman
naomi@stat.psu.edu

Outline

- Preliminaries
  - what you need to do before the DE analysis
- Stat Background
  - what you need to know to understand the analysis
- The Analysis

Altman, 2014

Preliminaries

- Before this point:
  - convert sequences to mapped reads
  - reduce data to counts per feature per sample

- At this point we assume that the data are in the form of a expression matrix of counts
- row=feature (gene or exon or ...)
  - column=sample.

- If an RNA sample was split across several lanes, we sum the counts across the lanes.

Altman, 2014

Stat Background

We assume that in sample i, some percentage π_{ij} of the reads come from feature j.

We want to test whether π_{ij} varies with treatment.

Differential expression analysis compares the differences in π_{ij} between treatments to what is expected from random noise introduced by biological, sample and measurement variation.

Altman, 2014

Stat Background

We assume that in sample i, some percentage π_{ij} of the reads come from feature j.

We want to test whether π_{ij} varies with treatment.

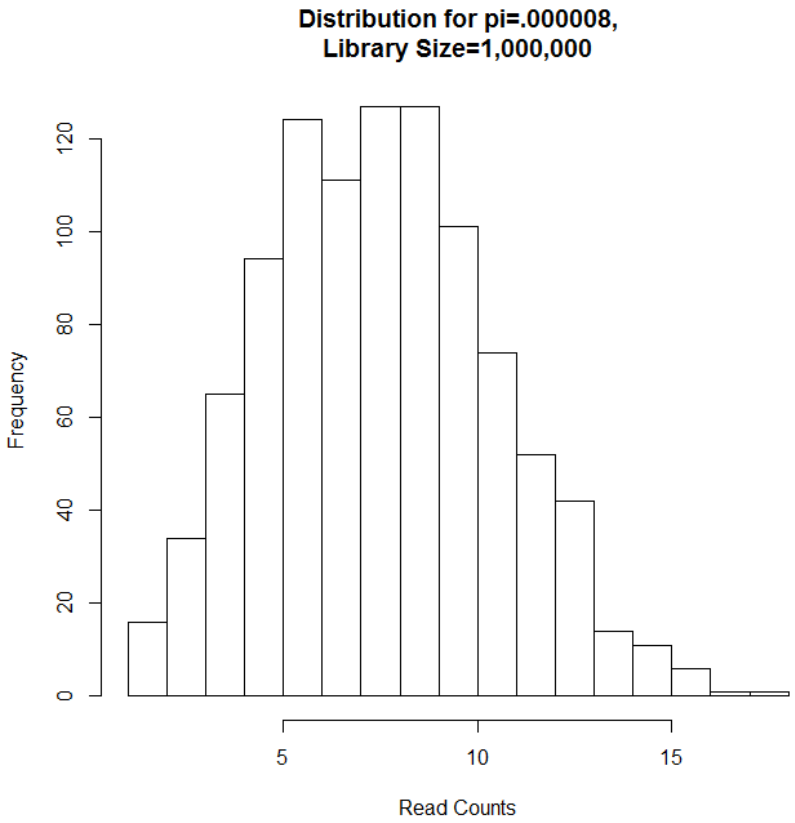
Differential expression analysis compares the differences in π_{ij} between treatments to what is expected from random noise introduced by biological, sample and measurement variation.

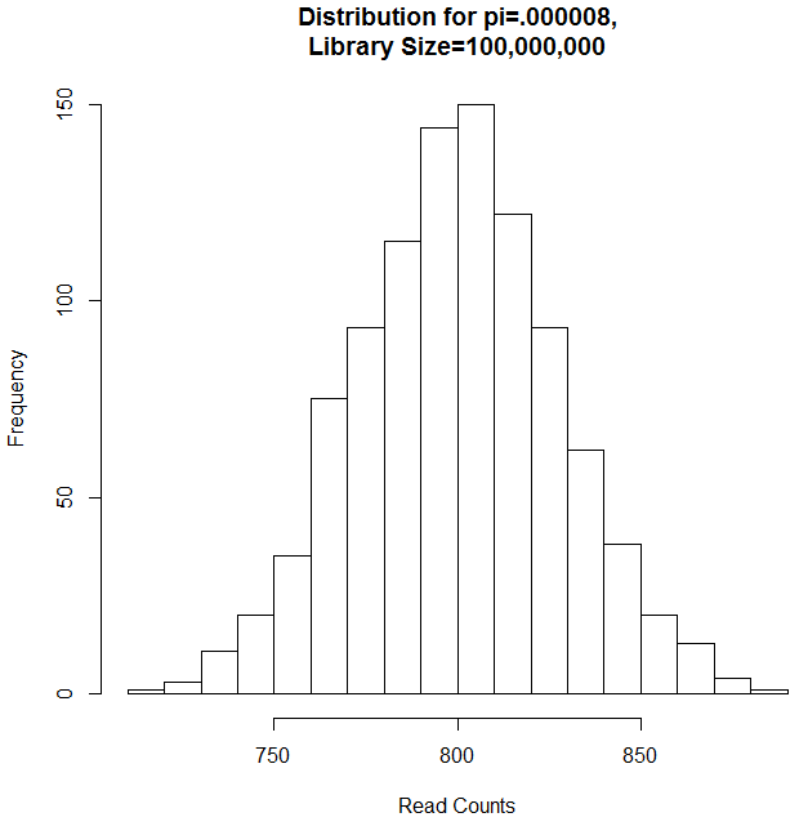
So we are going to need to estimate the random noise.

Altman, 2014

Stat Background

So we are going to need to estimate the random noise.





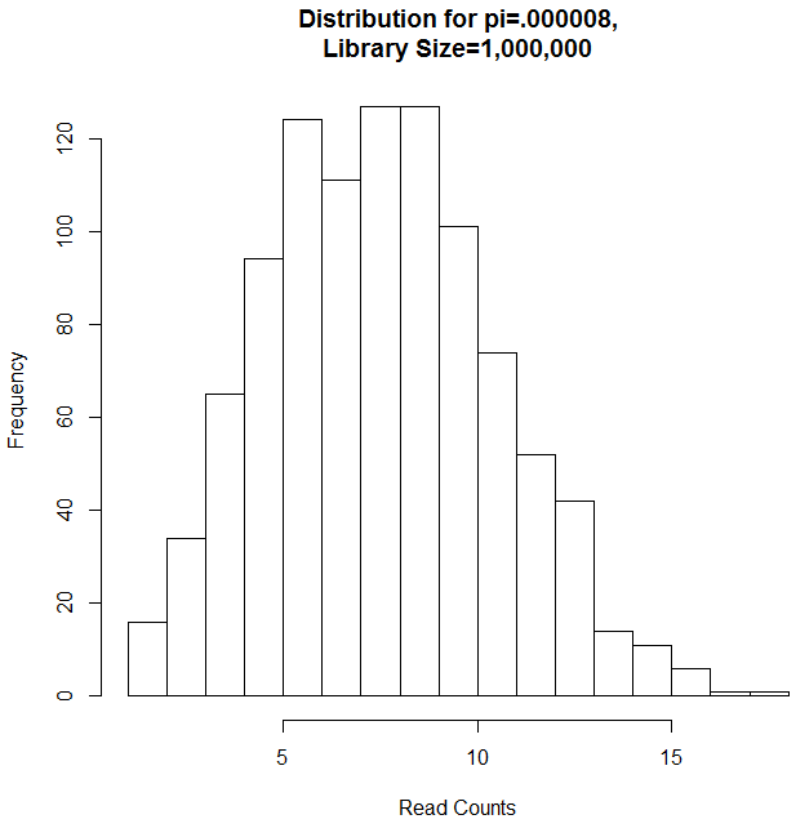
SD=sqrt(8)

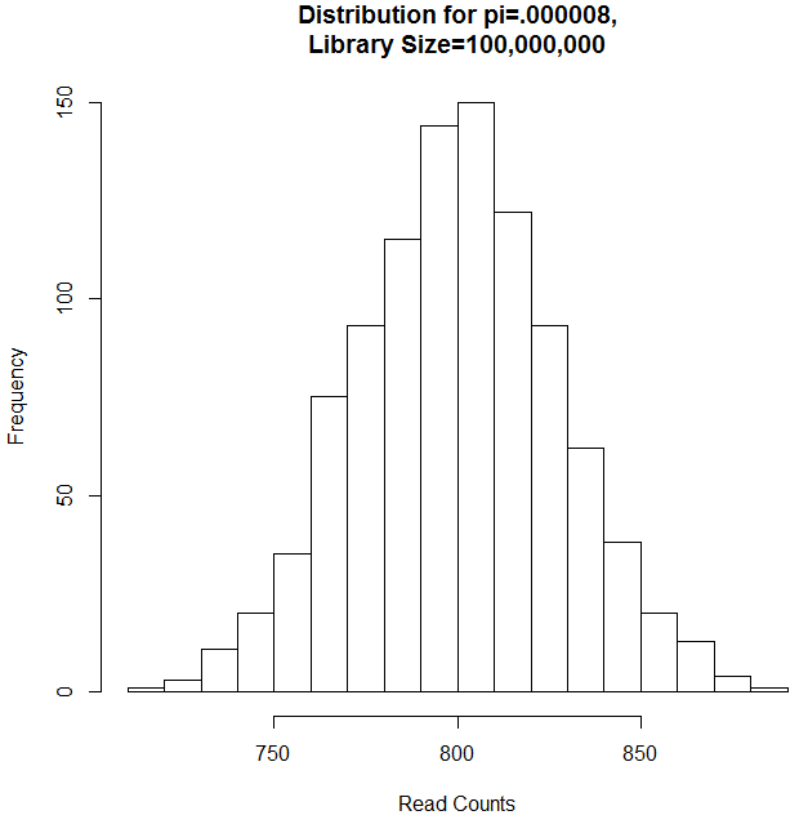
SD=10 sqrt(8)

Altman, 2014

Stat Background

So we are going to need to estimate the random noise.





6 and 10 are "typical"

600 and 1000 are not typical

Altman, 2014

Stat Background

Sequence data are counts.

Some mapping software returns FPKM but this is not suitable because the noise depends on the counts, not the percentage.

Larger features often do produce more reads than smaller ones, but most comparisons are between the same features in different samples, so that does not matter.

Altman, 2014

Stat Background

Sequence data differs in distribution from microarray data.

We will do the equivalent of t-tests or ANOVA, taking into account that the data are counts.

Count data have (at least) 3 sources of variability:

1. Poisson variability

2. Biological variability

3. Systematic variability

Altman, 2014

Stat Background

Sequence data differs in distribution from microarray data.

We will do the equivalent of t-tests or ANOVA, taking into account that the data are counts.

Count data has (at least) 3 sources of variability:

1. Poisson variability due to the fact that each tag either is or is not captured.
2. Biological variability due to variable numbers of tags of each type in each sample
3. Systematic variability due to the "treatments" (genotype, exposure, tissue type ...)

Altman, 2014

Stat Background

The goal of differential expression analysis is to identify

- Systematic variability due to the "treatments" (genotype, exposure, tissue type ...)

- This is usually measured as the ratio of the estimated difference in treatment percentages over the expected size of difference based on the Poisson and biological variation

Altman, 2014

Agenda

- Models for count data
- Filtering low counts
- Exploratory analysis (Quality assessment)
- Data Preparation
- Normalization
- Modeling and Moderating Dispersion
- Linear Models for Differential Expression
- Differential Expression analysis
  - the lab

Altman, 2014

Models for Count Data

Let n_i be the number of mapped reads in sample i and π_{ij} be the percentage of reads from feature j in sample i.

Statistical theory tells us that when $n_i\pi_{ij}$ is small, the observed number of reads from feature j in sample i should come from a Poisson distribution with mean $n_i\pi_{ij}$.

Technical replication confirms this.

But due to biological variability, π_{ij} varies among

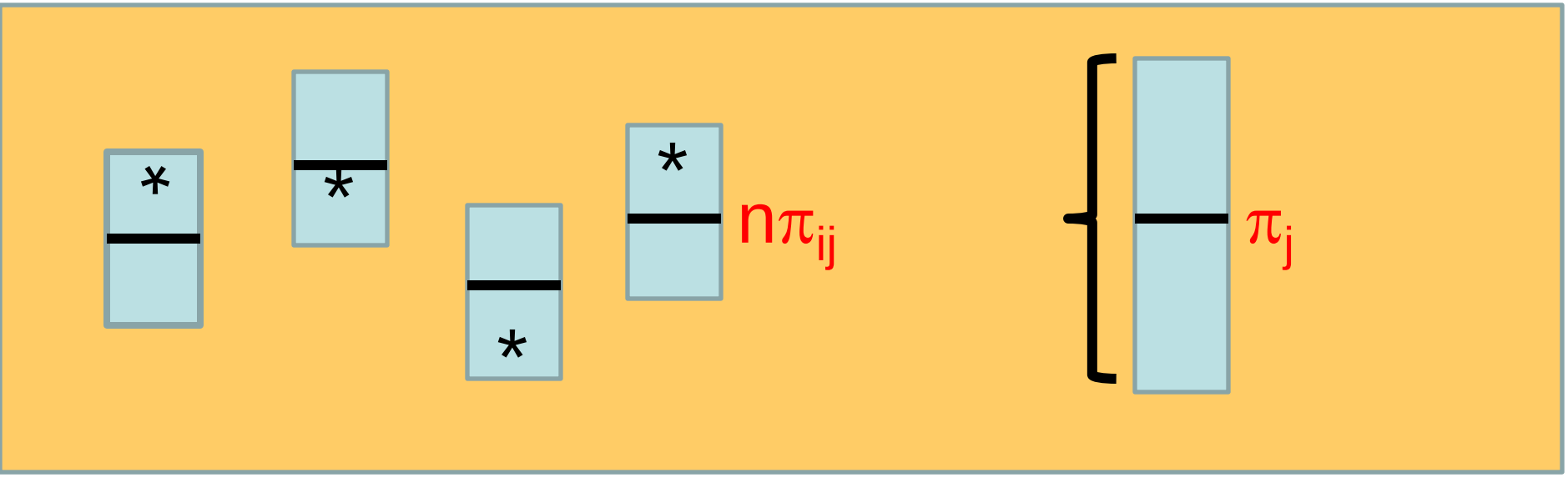
Altman, 2014

Models for Count Data

The expected number of reads for feature j in sample i is $n_i\pi_j$.

The differences in the percentages is accounted for in the variability which is $n_i\pi_j(1+n_i\pi_j\phi_j)$

ϕ_j is called the dispersion.

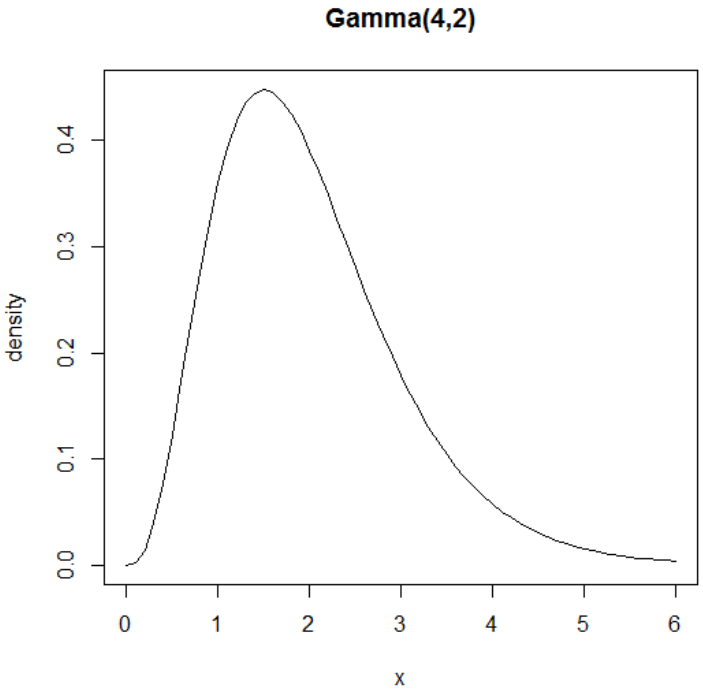


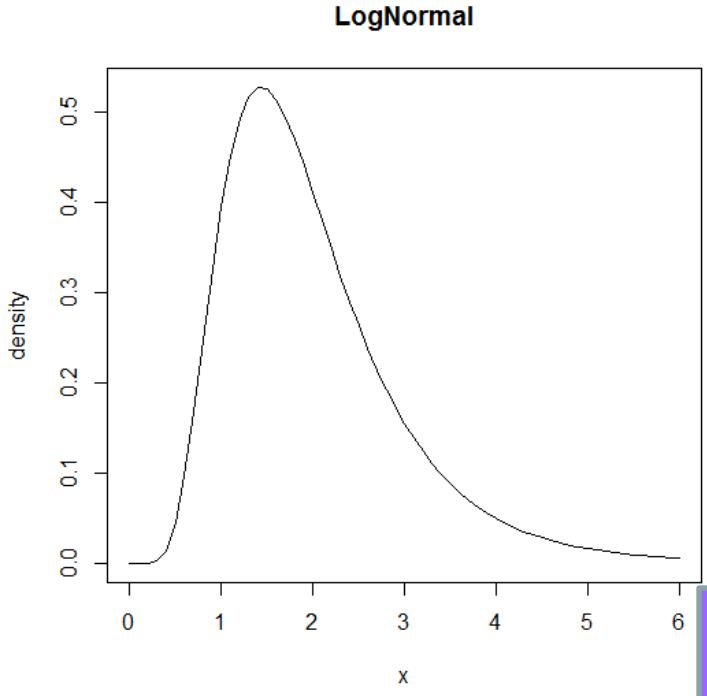
Samples are Poisson($n\pi_{ij}$). Aggregate is more variable.

Models for Count Data

There are 2 commonly used models for $n_i\pi_{ij}$ when $n_i=n$ is constant.

1) $n_i\pi_{ij} \sim \text{Gamma}(a,b)$ → Negative Binomial
2) $\log(n_i\pi_{ij}) \sim \text{Normal}(\mu,\sigma^2)$ → Poisson-LogNormal





Altman, 2014

Models for Count Data

Most packages for RNA-seq differential expression analysis use a log-linear model for the mean $\mu = n_i \pi_{ij}$

$$\log(\mu) = \beta_0 + \beta_1 x = \log(n_i) + \log(\pi_{ij})$$

where x is the treatment indicator.

Note that n_i is the known library size and only $\log(\pi_{ij})$ actually depends on the unknown parameters.

Altman, 2014

Models for Count Data

Most packages for RNA-seq differential expression analysis also model the variance.

When the data are Poisson (no biological variance) then the variance is the SAME as the mean.

The packages vary in how they model the variance but most use:

$$Var(Y)=\mu + f(\mu,\phi)$$

where f is a function that needs to be estimated and ϕ is a parameter controlling dispersion – $\phi=0 \rightarrow$ Poisson

Models for Count Data

Most packages for RNA-seq differential expression analysis also model the variance.

$$Var(Y)=\mu + f(\mu,\phi)$$

For sequencing data, if the gene expresses at different levels at different treatment combinations, the within variance also changes.

Altman, 2014

Models for Count Data

We will look at 2 software packages both from the Smyth lab:

edgeR uses the Negative Binomial model with dispersion moderation, and uses a normalization factor to account for library size.

voom is part of LIMMA. It uses CPM counts per million reads on the log-scale and weighted least squares. This is called “quasi-likelihood” because it uses the relationship between the mean and variance but NOT a probability model for the data.

Altman, 2014

Filtering Low Expressing Genes

- The power of statistical tests to detect real expression differences for count data depends on the mean counts

- For this reason, we may prefilter the data to eliminate features with very low total read counts before doing differential expression analysis.

- If we have more samples or more reads per sample or aggregate features we can increase the power to detect differential expression.

Altman, 2014

Multiple Testing

Work done by my students suggests that filtering the low expressing features is the best fix for FDR estimation.

There are several heuristics for filtering low expressing features.

We will filter out features with fewer than 10 reads across all our samples.

These features can be viewed separately to determine the distribution of reads.

Altman, 2014

Example

Altman, 2014

GEO dataset GSE17274

Blekhman, et al, (2010) liver samples in male and female human, chimpanzee and rhesus monkey.

Samples were mapped to human homologs.

There were 3 biological replicates of each treatment, each divided in 2 sequencing lanes. 20689 features were tabulated in each of the 36 lanes

71 million mappable 35 bp reads.

Processed counts per feature

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17274

Data Preparation

Step 3a: Compute the library sizes for each sample.

HSM1	PTF1	RMM1	HSF1	PTM1
2844253	**5345035**	5163215	3385112	3176224

RMF1	RMF2	HSM2	PTF2	RMM2
4511466	3872339	4385229	3791833	3916798

HSF2	PTM2	RMM3	RMF3	HSM3
3421198	3554650	4531203	5220250	3799601

PTF3	PTM3	HSF3
3652193	3548743	2844471

The library sizes differ

Altman, 2014

Data Preparation

Step 3b: Understand the highly expressed genes.

HSM1	PTF1	RMM1	HSF1	PTM1	RMF1
6.8%	8.9%	11.5%	16.7%	**4.5%**	9.4%

RMF2	HSM2	PTF2	RMM2	HSF2	PTM2
16.2%	14.6%	5.9%	10.5%	14.8%	10.9%

RMM3	RMF3	HSM3	PTF3	PTM3	HSF3
6.4%	5.4%	10.8%	8.7%	5.4%	8.1%

The most highly expressed gene is a large percentage of each sample

Altman, 2014

Preliminaries

Step 3b: Understand the highly expressed genes.

In these samples, the 9 most abundant transcripts account for 25% of the reads, but the percentage of the most abundant transcript varies from 4.5% to 16.2% affecting the relative abundance of every other transcript.

(The most abundant transcript is ALBUMIN in most of these liver samples.)

Altman, 2014

Preliminaries

Note: when a few genes dominate the reads, the relative expression of ALL other genes in the same appears to be depressed.

You need to understand the biology – is the massive over-expression of a few genes depressing expression or just using up "sequencing space"?

In single cell processing we can actually determine this!

Altman, 2014

Normalization

The effective read totals should adjust BOTH for the actual differences in library sizes and the differing numbers of reads coming from a few aberrant highly expressed genes.

This is done by normalization.

The simplest method is to compute the mean (median...) of the 75th quartile of reads in the sample, and then compute the ratio for each sample to create an effective library size.

Altman, 2014

Normalization

RNA-seq normalization methods compute an effective library size.

Methods based on counts use the effective library size as an offset replacing the actual library size.

i.e. $\log(\mu)=\log(n_i) + \log(\pi_{ij})$

Altman, 2014

Models for Treatment Comparisons

Recall:

Models for treatment comparisons need to take into account:

1. Biological variation between replicates.
2. Total read count differences among samples.
3. "Extra-Poisson" variation (dispersion)

Altman, 2014

Models for Treatment Comparisons

Assume that in
treatment t
sample i
gene j

we observe n_{tij} reads

The true percentage of reads from gene j in sample t,i is π_{tij}.

Averaging over the response of the population to treatment t, the true mean percentage of reads from gene j is $\pi_{t\bullet j}$.

Models for Treatment Comparisons

For treatment 1 we have

$n_{11j}\, n_{12j}\, n_{13j} \bullet \bullet \bullet$

For treatment 2 we have

$n_{21j}\, n_{22j}\, n_{23j} \bullet \bullet \bullet$

Our test is: If $\pi_{1\bullet j} = \pi_{2\bullet j}$ and the observed library sizes, what is the probability that the counts we could observe in different samples are as different or more different than actually observed.

Models for Treatment Comparisons

Our test is: If $\pi_{1\bullet j} = \pi_{2\bullet j}$ and given the observed library sizes, what is the probability that the counts we could observe in different samples are as different or more different than actually observed?

This probability is the p-value.

To estimate the p-value, we need:

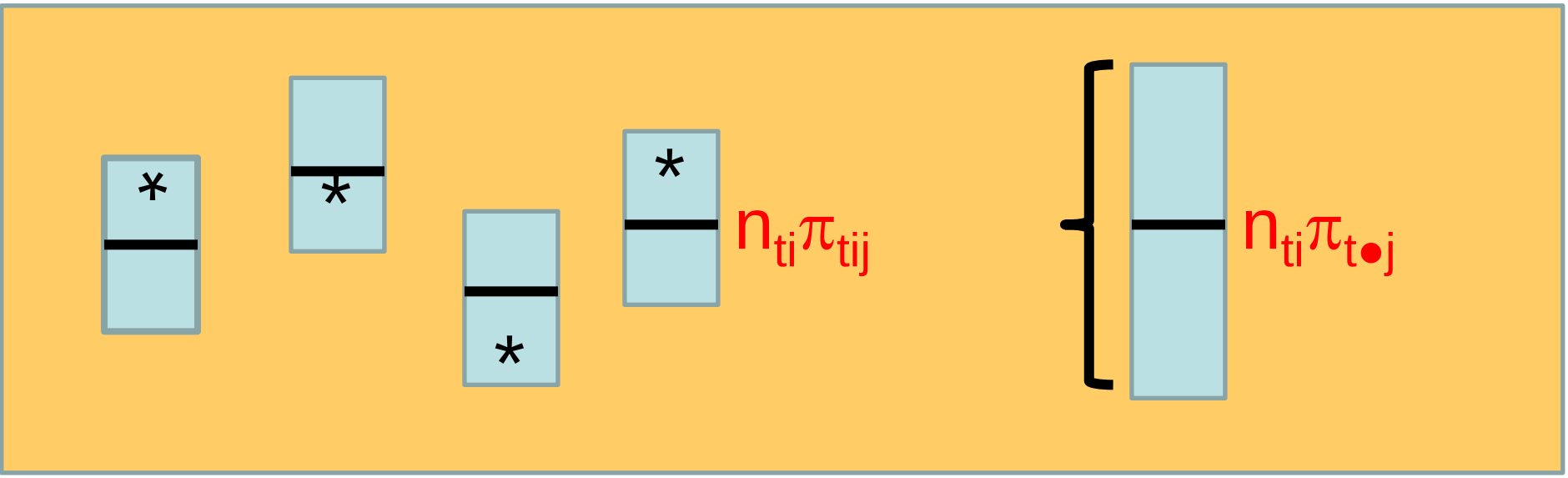
- the read counts
- an estimate of the variability of read counts from sample to sample when $\pi_{1\bullet j} = \pi_{2\bullet j}$

Altman, 2014

one-way design

Models for Treatment Comparisons

The expected number of reads for feature j in sample i is $n_{ti}\pi_{t\bullet j}$. where n_{ti} is the library size.



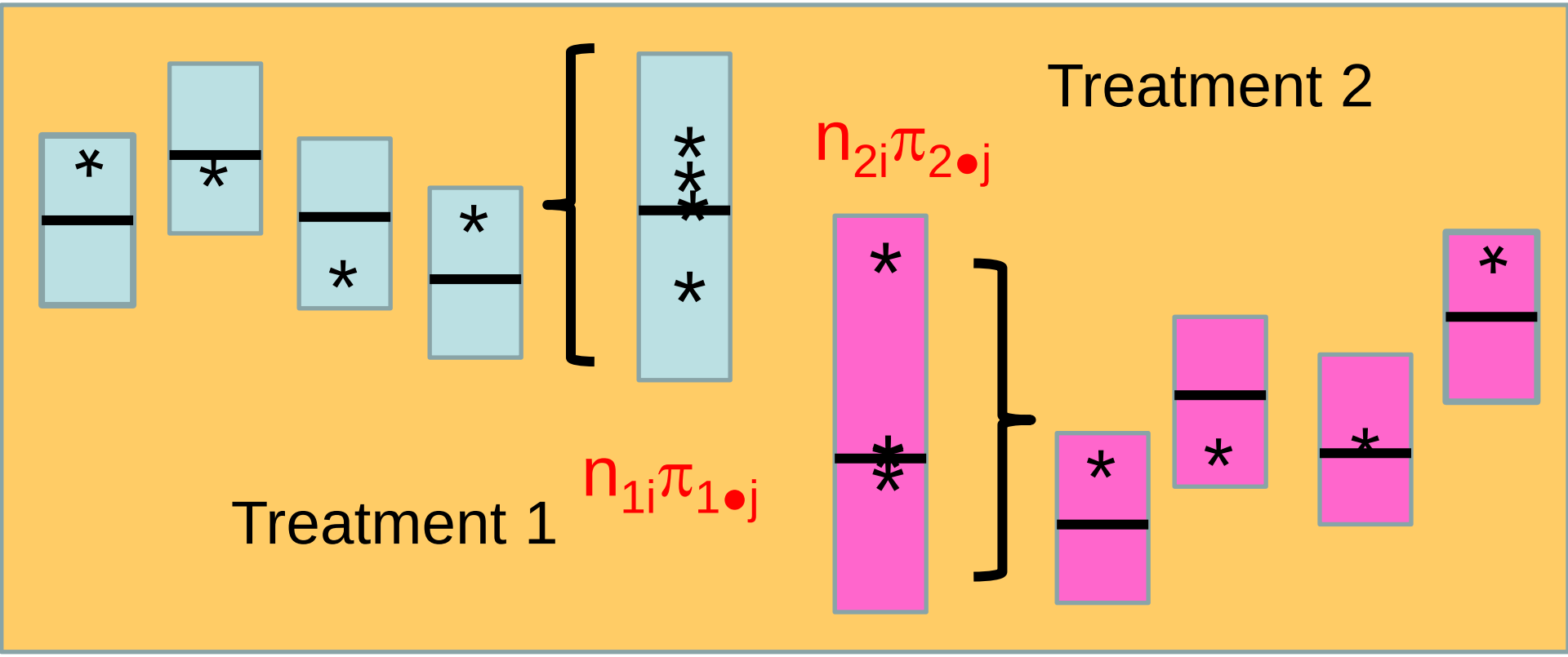
The differences in the percentages is accounted for in the variability which is $n_{ti}\pi_{t\bullet j}(1+n_{ti}\pi_{t\bullet j}\phi_{t\bullet j})$

$\phi_{t\bullet j}$ is called the dispersion and is estimated from all the counts for that feature.

Altman, 2014

one-way design

Models for Treatment Comparisons



$\text{sqrt}(n_{ti}\pi_{t\bullet j}(1+n_{ti}\pi_{t\bullet j}\phi_{t\bullet j}))$ is a measure of height of the big boxes.

If the heights are big compared to the distance between the central bars, this is likely to occur by chance.

Altman, 2014

Dispersion

Dispersion is difficult to estimate so most popular software e.g. edgeR, DESeq use estimates which shrink the dispersion towards some type of “typical” value.

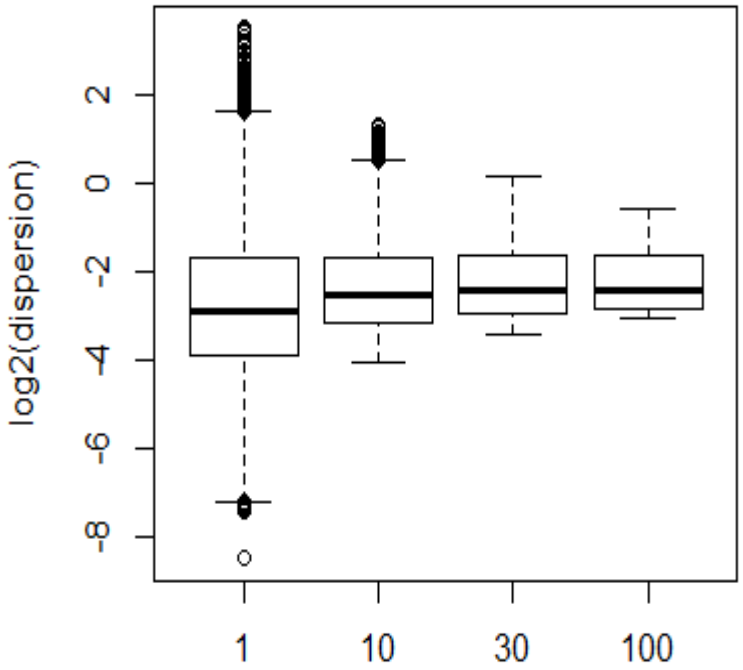
Altman, 2014

Dispersion

Dispersion is computed for each gene (1).
The labels(10, 30,100) indicate the amount of shrinkage.

This is from edgeR.

DESeq pulls the low dispersions up towards the common value, but leaves the high dispersions, which is more conservative.



Altman, 2014

Dispersion

ENSG00000105549 0 0 0 0 0 0 0 0 0 0 0 0 11 0 0
ENSG00000132446 0 0 0 0 0 0 1 0 0 0 0 0 9 0 0
ENSG00000134438 0 0 0 0 0 0 0 0 0 11 0 0 0 0 0
ENSG00000149968 1 0 0 0 0 3 0 0 0 11 0 0 0 0 0
ENSG00000152822 0 0 0 0 1 2 0 0 0 0 0 0 12 0 0
ENSG00000165566 0 0 0 0 0 18 0 0 0 0 0 0 0 0 0
ENSG00000169393 0 0 0 0 0 0 0 0 0 7 0 0 0 5 0
ENSG00000184909 0 0 0 0 0 0 0 0 0 11 0 0 0 0 0

The genes with individual dispersion >10 are mostly 0 but have a few samples with high counts.

Altman, 2014

Models for Treatment Comparisons

After all this preparation, all we need to do is tell our software which groups we want to test.

We get a p-value for each feature for each comparison of equality of expression.

We infer that the features differentially express if the p-value is small.

We can also do more complicated comparisons

Altman, 2014

Multiple Testing

The p-value is the probability of seeing differences this big or bigger when the percentages of reads are on average the same in the 2 treatments.

Suppose P=1%

Is that small?

Altman, 2014

Multiple Testing

The p-value is the probability of seeing differences this big or bigger when the percentages of reads are on average the same in the 2 treatments.

Suppose P=1%

Is that small?

Altman, 2014

Multiple Testing

Suppose I roll 2 dice. What is the probability of 2 sixes?

What if I roll 36000 times?

Altman, 2014

Multiple Testing

In our example, there are 14981 genes.

5% of 14981= 749
1% of 14981= 150

Comparison	**p<0.05**	**p<0.01**
Human Male vs Female	373	82
Male Human vs Chimp	3032	1588

Estimated False Discovery Rate for Male Human vs Chimp:
P<0.01

150/1588=9.4%

Altman, 2014

Stat 555

Statistical Analysis of High Throughput Genomics Data

learn to do your own analysis

(Stat 500 preferred but not required)

Spring semester only.

See you there!

Altman, 2014