Statistics for Biologists
Outline

• estimation and hypothesis testing
• two sample comparisons
• linear models
• non-linear models
• application to genome scale data
Warning

• while the quantities often seem simple
• NEVER IMPLEMENT THEM YOURSELF

• use good software that already exists (R, SAS, MatLab)
  – numerical/scientific computing has many pitfalls for the unwary
Warning

• what went wrong:
  > x = sqrt(2)
  > x
  [1] 1.414214
  > x * x == 2
  [1] FALSE

• R FAQ 7.31, Why doesn’t R think these numbers are equal?
Estimation

• given some set of data one might want to estimate some parameters of that data
  – mean, variance, mean

• **point estimates**
  – mean=122.2

• **interval estimates**
  – the mean is between 101 and 133

• in **general** we make assumptions about the underlying probability model (randomness) and choose estimates with specific properties
  – unbiased, minimum variance
  – we can be frequentist or Bayesian
  – confidence intervals (have a frequentist interpretation)
Hypothesis Testing

• a hypothesis is a statement about the real world
  – I think the mean is 100 ($H_0 : \mu = 100$)

• the null hypothesis should typically represent the status quo, or a presumption of no effect

• we use the data, plus our chosen inference paradigm to compute quantities that help us determine whether the null hypothesis is likely to be true, or not
Two-types of mistake

• there are two kinds of mistakes that can be made
  – reject the null hypothesis when it is true
  – accept the null hypothesis when it is false
• the size of a test is the probability that we reject the null when true
• the power of a test is the probability of rejecting the null hypothesis when it is false
  – this generally requires us to specify how it is false
• in general we use the size of the test to control the first type of mistake at some fixed level
• for a given size there are many tests, we attempt to choose ones that are more powerful for likely alternatives
p-values

• are quantities that relate to the null hypothesis
  – you cannot have a p-value without a null hypothesis
  – the p-value measures how likely it is to see evidence as extreme or more extreme as that observed assuming the null hypothesis is true
  – small p-values are evidence against the null hypothesis; they are not the probability it is true!
  – Bayesian’s use a different approach and typically end up with quantities that do have probabilistic interpretations
Equivalence

• there is a very direct relationship between confidence intervals and hypothesis tests

\[ H_0 : \theta = X \]

- if the value, X, lies inside of a 95% CI then the null hypothesis would not be rejected at the 5% level
- if X, lies outside the 95% CI, then the null hypothesis would be rejected at the 5% level

- do not reject \( H_0 \)

- reject \( H_0 \)
Significance

• statistical significance should never be confused with scientific significance

• statistical significance tells us the surprise factor:
  – if all my assumptions are correct, and the null hypothesis is true, how surprised should I be by my data
  – at some level of surprise we choose to decide that our null hypothesis is unlikely to be true (usually we check to be sure our assumptions are reasonable)

• scientific significance is concerned with whether what we found is likely to have any relevance to our understanding of nature
Significance

• statistical significance is affected by sample size
• scientific significance is not
• getting more data often ensures statistical significance
  – new data technologies give us too much data
  – eg flow cytometry, sequencing
  – many things are scientifically uninteresting, but statistically significant
Two Concepts

• **variance**: when we estimate a quantity using data, we generally get both a point estimate and some estimate of the variability of that estimate
  – as sample sizes increase this variance tends to decrease

• **bias**: this is the difference between what we intended to measure and what we did measure
  – we estimate RPKMs incorrectly due to mapping issues
  – bias is never improved by sampling more, it usually requires changes in technology to reduce
Two Important Theorems

- a **central limit theorem** basically says that the **average** (mean) of a set of numbers (assumed to come from some distribution) will behave approximately like a Normal random variable as the set grows.

- the **law of large numbers** says that the mean of a set of numbers (assumed to come from some distribution) will get arbitrarily close to the mean (expected value) of the distribution.
Two Sample Comparisons

• paired vs non-paired comparisons
  – eg. before/after, or two related measurements
  – a paired comparison usually increases power

• non-parametric tests vs parametric tests
  – parametric tests tend to be more powerful, for a given sample size, but they often achieve that at the expense of making assumptions

• t-test, Wilcoxon, Mann-Whitney are favorites
t-test

• test is for equality of the means
  \[ H_0 : \mu_1 = \mu_2 \]

• various versions can address different underlying assumptions
  – paired vs independent

• assumptions:
  – no strong ones, the CLT provides rationale for reasonable samples
  – this is a parametric test (\( \mu \) is the parameter)
Non-parametric two-sample tests

- Mann-Whitney (two independent samples)
- Wilcoxon (paired samples)
- they have a different null hypothesis
  \[ H_0 : F_1 = F_2 \]
- equality of the two underlying distributions
- while this includes equality of the means, it is more restrictive
- in particular we do not expect correspondence between these tests and the t-test
When to use tests

• non-parametric tests are often used when one does not want to make specific assumptions about the data
  – but they are less powerful, so if you don’t have much data they won’t work very well
• when you have lots of data and the assumptions are reasonable both parametric and non-parametric methods have similar behavior
• so I would use the non-parametric tests when I want to test $H_0 : F_1 = F_2$
• and the parametric tests when I want to test $H_0 : \mu_1 = \mu_2$
Limitations

• the two sample tests can be extended in a number of ways
  – inclusion of covariates; linear and non-linear regression
  – multiple groups; ANOVA (and friends)
Linear Models

• a linear model

\[ y = a + bx + e \]

• where \( y \) represents the independent variable
• \( a \) is the intercept (value for \( y \) when \( b=0 \))
• \( b \) is the slope of the relationship
• \( x \) are the known covariates
• \( e \) are the errors
Ancscombe’s Quartet

• four data sets for which most summary statistics and indeed, a, b and $\sigma^2$, are identical
• but regression is appropriate for only one
Anscombe’s Quartet
Linear Models

• often the model is fit and parameters estimated using least squares
  – this gives estimates of $a$, $b$ and from them the residuals can be obtained

\[
\hat{e} = y - \hat{a} - \hat{b}x
\]

• the residuals can be used to determine whether the model is reasonable

• hypothesis tests generally focus on questions about $b$
The t-test as a linear model

- if we let $x$ be 0 or 1, depending on whether the observation is Treated or Not Treated,
- then for every observation in the treated group our model is $y = a + e$

- and for every observation in the untreated group the model is $y = a + b + e$

- so we can interpret $a$ as the mean in the treated group, and $a+b$ as the mean in the untreated group
- the test of $b=0$, is identical to the t-test, for unpaired samples
Linear model

• but the advantage of this formulation is that we can add other variables
  – eg sex, tissue, complex treatments
  – these are then adjusted for in our comparisons
• the residuals should always be examined, since they tell you about whether or not your model is appropriate
• testing $b=0$ makes the strong assumption that the model is correct
  – it is important that you learn to assess whether model assumptions are reasonable
Non-linear models

• while there is only one kind of linear model, there are lots of different non-linear models
• we will discuss generalized linear models
• this class of models includes logistic regression Poisson regression and Negative Binomial regression models
• logistic regression is used to model 0/1 data
• Poisson and Neg Binomial are suitable for modeling count data
  – the latter is more general and is being used for much of the DE of next gen sequencing data
Non-linear models

• good software exists for fitting these
  – Modern Applied Statistics in S (MASS), Venables and Ripley
  – Julian Faraway’s books, Linear Models in R, and one on non-linear models
Application to Genome scale data

• several problems/issues became apparent
  – the test statistics seemed to often associate with other variables
    • for microarrays DE genes were those with high intensity
    • for RNA-seq, GC content seems to matter in some cases
  – these indicate the need for normalization
Genome Scale

• the test statistics could be large due to variability in the estimate of the variance
  – led to moderated t-tests, and other approaches

• how do we assess significance when doing many tests
  – p-value correction methods
Moderated t-tests

\[
\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2 / n}}
\]

- the t-test can be large if
  1. the means are different
  2. \(n\) is large
  3. our estimate of SE is small
- 1. is mostly what we are interested in
  - so we sometimes include a fold-change requirement
- 2. is a problem with flow cytometry and for some RNA-seq problems
- 3. is common in microarray experiments and limma and others use some form of moderated estimate of the SE
Moderated tests

• they are effective for small sample sizes, the advantages of moderation drop off as the sample size increases

• there is nothing special about t-tests and limma fits more general models
  – most other methods can be similarly adapted
p-value Adjustments

• p-values are really interpreted for a single test
• when you do many some more careful thinking is required to ensure that error rates are controlled
• the false discovery rate is the expected value of the proportion of all tests for which $H_0$ is rejected where it is actually true
• this turns out to be a relatively easy quantity to estimate and it is of reasonable importance
p-value Adjustments

• we can often live with quite high FDR values
  – in some discovery projects FDR=0.5 is considered pretty good

• as with all cut-offs/approaches the FDR does not tell the whole story
  – it is attempting to control false discoveries
  – it says nothing about missing true discoveries
  – indeed, if one takes those tests just below the cut-off, they are enriched for true discoveries