Statistics for Biologists

Lecture 2:

Inferential statistics:

Pairwise comparisons, power and multiple testing

George Bell, Ph.D.
Senior Bioinformatics Scientist
Bioinformatics and Research Computing
Whitehead Institute

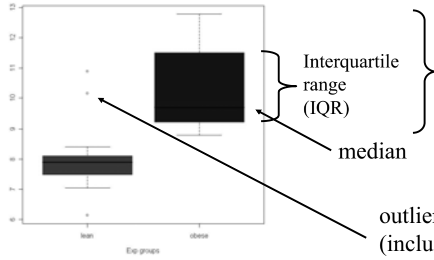
Outline

- Lecture 1 review
- Intro to pairwise comparisons
- The t-test and other tests
- Statistical power
- Intro to multiple hypothesis testing
- Methods to correct for multiple hypothesis testing
- Using the R Commander
- Exercises

From lecture 1

- Why use statistics?
- Descriptive statistics
  - – central tendency + variability
- Visualization of quantitative data
  - – What are you trying to show?
- Inferential statistics: H_0, H_a, α, β
- False positives and false negatives
- Software for statistics
- Exercises

Exercise 1 question: boxplot



Interquartile range (IQR)

Range of values <= 1.5 * IQR from box borders

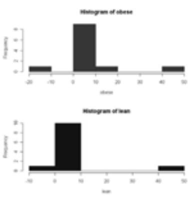
median

outlier (included in calculation of median, etc.)

Exercise 1 questions: R

- Place multiple figures on the same plot:

```
par(mfrow=c(2,1))
hist(obese, col="red")
hist (lean, col="blue")
par()   # reset
```



- Save a plot to a file: right click or

```
savePlot(filename="myPlot.pdf",
   device=2, type="pdf")
```

- Open another window for a new figure

```
windows()
```

Intro to pairwise comparisons

- A common research question is,
  - – "Is one measurement equal to another measurement?"
- More specifically, Q: "Is the mean of values in set X equal to the mean in set Y?
  - – sample answers: no; yes; X > Y; X:Y = 1.5
- To get an answer that includes some measure of confidence, we need to do a statistical test:
  - – T-test (parametric: uses actual values)
  - – Wilcoxon Rank Sum (Mann-Whitney) test (nonparametric: uses ranks; less powerful: greater β)

Student’s t-test

- History: Gossett (“Student”) created this as a new technique to analyze Guinness beer using a small sample size.
- Assumptions:
  - measurements are independent
  - input data are normally distributed
  - Samples are from populations with equal variances (but test can be modified (“Welch’s test”) when this isn’t true)

  Fortunately the test is robust, but…
- Two-tailed (testing for ≠) or one-tailed (testing for > or <)
- T statistic = $\frac{\text{mean}_1 - \text{mean}_2}{\text{SE}}$ df = $n_1 + n_2 - 2$

Performing the t-test

Data sets: a = { 2 4 6 } b = { 7 8 9 }

- Excel command: =TTEST(array1, array2, tails, type)
  example: =TTEST(A1:A3, B1:B3, 2, 3)
- R command: t.test(a, b, alternative="two.sided", var.equal=FALSE)

Output:

```
   Welch Two Sample t-test

data:  a and b
t = -3.0984, df = 2.941, p-value = 0.05479
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 -8.1553995  0.1553995
sample estimates:
mean of x mean of y
        4         8
```

Prob (H_a is false)

for mean(a) – mean(b)

Comparing variances

- Use the F test
  - Excel: =FTEST(array1, array2)
  - R: var.test(a, b) # a=c(2, 4, 6); b=c(7, 8, 9)
- H_0: True ratio of variances is 1

```
        F test to compare two variances

data:  a and b
F = 4, num df = 2, denom df = 2, p-value = 0.4
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
   0.1025641 156.0000000
sample estimates:
ratio of variances
                 4              # To calculate variance:  =VAR  or  ‘var’
```

Prob (H_a is false)

for var(a) / var(b)

The paired t-test

- Used when one measurement in each group comes from the same “experimental unit”
- example:
  - mouse assay after treatment by injection
  - one leg is treated; other leg is control
- Method (done by paired test):
  - get difference between each measurement pair
  - Are these differences ≠ 0?
  - (Do a one-sample t-test)
- Use when experimental design warrants it
- R command: t.test(a, b, paired=TRUE)

Set A	Set B
100	105
50	55
20	23
250	260
0.94 vs. 0.03	

T-test: summary

- Tests for a difference between two means
- Assumes data are normally distributed
- Test comes in multiple flavors:
  - 1 vs. 2 tails
  - equality of variances?
- Use the paired test when appropriate
- Excel just gives the take-home message
  - Or see Tools >> Data Analysis

Moderated t-test

- Generic t statistic: $t = \frac{\bar{x}_1 - \bar{x}_2}{s}$
- Moderated t statistic $t = \frac{\bar{x}_1 - \bar{x}_2}{s + s_0}$
  - Best known in microarray analysis
  - Corrects gene expression standard deviations toward a pooled estimate
  - Modify all standard deviations used in the t-test
  - s_0 = 90th percentile of all s (`limma` R package)
  - s_0 = exchangeability constant (SAM)
  - gains power from sharing variation data across genes

Power and sample size for the t-test

- Power reflects the probability of finding a true difference.
- Determinants of power:

$$n \ge \frac{2s_p^2}{\delta^2}(t_{\alpha,\upsilon} + t_{\beta(1),\upsilon})^2$$

- Calculating power in R using $\log_2$-transformed data:
  - How many measurements are needed to be 80% sure that a 2-fold difference [$\log_2(2)=1$] can be determined when using $\alpha=0.05$ on data with a sd of 0.45 (standard deviation of log2-transformed measurements)?

power.t.test(n=NULL, delta=1, sd=0.45, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")

Power and sample size for the t-test

R command: power.t.test(n=NULL, delta=1, sd=0.45, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")

- Output:

```
     Two-sample t test power calculation

              n = 4.3824
          delta = 1
             sd = 0.45
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

Wilcoxon Rank Sum test

- A non-parametric test:
  - An alternative to a t-test
  - Ranks of data (rather than actual values) are used.
- Also known as the Mann-Whitney test
- No assumptions about normally distributed data are required.
- Data sets a = { 0.2 0.4 0.6 } b = { 70 80 90 }
  become a = { 6 5 4 } b = { 3 2 1 }
- R command:
  wilcox.test(a, b, alternative="two.sided") => p-value = 0.1
- But this test doesn't have much power:
  - same data in t-test yield a p-value of 0.005

Intro to multiple hypothesis testing

- A p-value reflects the probability of a false positive call in a statistical test.
- When multiple hypotheses are tested on the same data, the rate of false positives greatly increases

Example for $\alpha = 0.05$:

Number of genes tested (n)	Expected number of FPs = $\alpha * n$	Probability of at least one FP = $100(1 - (1 - \alpha)^n)$
1	0.05	5%
2	0.1	9.75%
100	5	99.4%

- So what to do?

Choices for multiple hypothesis testing

- How many true positives do you expect?
- How costly is a false positive call?
- How costly is a false negative call?
- Major choices:
  - Family-wise error rate (ex: Bonferroni)
  - False discovery rate (ex: Benjamini and Hochberg)

Corrections with the FWER

- Family-wise error rate = the probability of at least one false positive in the "family" of positive tests.
- "Bonferroni correction"
- A good choice if you predict that there are very few – if any – true positive tests
- Large correction means that
  - the false positive rate is very low
  - the false negative rate can be very high
- Practical implication: p-value must be really low to make it past a FWER correction.

Performing a Bonferroni correction

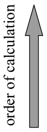
- Corrected p-value = raw p-value * n
  - n is the number of tests
  - if corrected p-value > 1, set to 1
- Example:
  - a microarray assays 10,000 genes
  - All are tested for differential expression with the t-test
  - If the raw p-value for one gene = 10^{-5}
    - The corrected p-value:
      = 10^{-5} * 10,000 = 0.1
    - If α = 0.05, then this gene cannot be described as differentially expressed.

Corrections with the FDR

- False discovery rate = the frequency of false positives among the positive tests
- Introduced by Benjamini and Hochberg, 1995
- Tolerates a chosen proportion of false positives
- Much less conservative than Bonferroni:
  - false positive rate is higher
  - false negative rate is lower
- A good choice if you predict that there are many truly positive tests
- Practical implications:
  - more tests stay below α compared to FWER
  - appropriate correction for many tests of differential expression in microarray experiments

Performing a FDR correction

- Sort list of p-values in increasing order and starting at the bottom row
- Corrected p-value = the minimum between
  1: raw p-value * (n/rank)
  2: corrected p-value below
  - n is the number of tests
  - rank is the position in the sorted list
- Example: a microarray assays 5 genes for differential expression



Gene	Rank	Raw p-value	Formula	Corrected p-value
C	1	0.001	min (0.001 * (5/1), 0.0125)	0.005
A	2	0.005	min (0.005 * (5/2), 0.017)	0.0125
B	3	0.01	min (0.01 * (5/3), 0.063)	0.017
E	4	0.05	min (0.05 * (5/4), 0.1)	0.063
D	5	0.1	0.1 * (5/5)	0.1

Performing multiple hypothesis corrections in R

```
# Read the data (tab-delimited text, with header fields; p-values in second field)
data.raw.pvals = read.delim("pvals_raw.txt", h=T)

# Correct the p-values using one of 7 methods ('?p.adjust' to see them)
# [,2] ➔ use the data in column 2

FDR.p.vals = p.adjust(data.raw.pvals[,2], "fdr")

# Combine original file with corrected p-value output
data.adjp = cbind(data.raw.pvals, FDR.p.vals)

# Print the output to a tab-delimited file
write.table(data.adjp, file =
   "data_pvals_corrected.txt", sep="\t", quote=F)
```

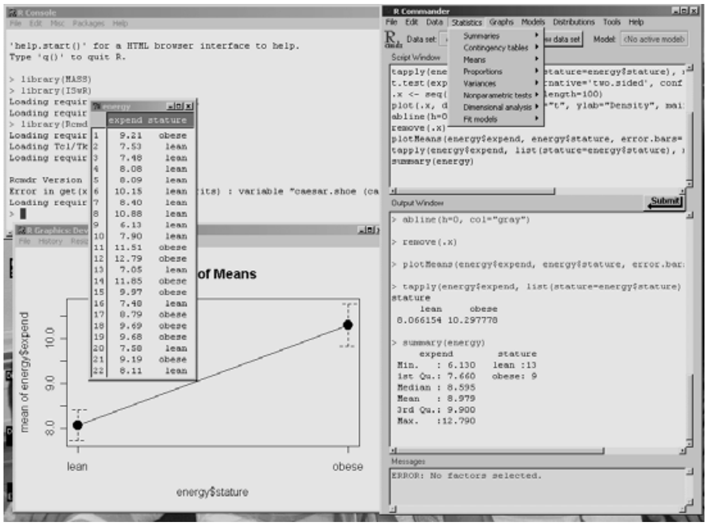
	id	Raw.p	bonferroni	BY	holm	hochberg	BH	fdr
1	A	0.0400	0.1600	0.1389	0.1200	0.1000	0.0667	0.0667
2	B	0.0010	0.0040	0.0083	0.0040	0.0040	0.0040	0.0040
3	C	0.2000	0.8000	0.4167	0.2000	0.2000	0.2000	0.2000
4	D	0.0500	0.2000	0.1389	0.1200	0.1000	0.0667	0.0667

Summary: multiple hypothesis corrections

- Beware that with raw α = 0.05, 5% of tests will be positive just by chance
- Whenever performing >1 statistical test together, corrections should be done
- Select method based on desired FP and FN error rates
- Use R, Excel or BaRC web tool

The R Commander

- Graphical interface designed to facilitate learning R
- Pull down menus and multiple-choice interface for common statistics and graphics
- Created by John Fox from McMaster University
- Installed with R on barra
- To start => begin R => library(Rcmdr)
- For Macs (but installation may be difficult), Windows, and Unix/Linux
- Other R graphical tools: affylmGUI, limmaGUI



```
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

> library(MASS)
> library(ISwR)
Loading requir
Loading requir
> library(Rcmd
Loading requir
Loading Tcl/Tk
Loading requir

Rcmdr Version
Error in get(x                 its) : variable "caesar.shoe (ca
Loading requir
>
```

```
   expend stature
1    9.21   obese
2    7.53    lean
3    7.48    lean
4    8.08    lean
5    8.09    lean
6   10.15    lean
7    8.40    lean
8   10.88    lean
9    6.13    lean
10   7.90    lean
11  11.51   obese
12  12.79   obese
13   7.05    lean
14  11.85   obese
15   9.97   obese
16   7.48    lean
17   8.79   obese
18   9.69   obese
19   9.68   obese
20   7.58    lean
21   9.19   obese
22   8.11    lean
```

```
tapply(ene                    stature=energy$stature), :
t.test(exp                    rnative='two.sided', conf
.x <- seq(                    length=100)
plot(.x, d                    ="t", ylab="Density", mai
abline(h=0
remove(.x)
plotMeans(energy$expend, energy$stature, error.bars=
tapply(energy$expend, list(stature=energy$stature), :
summary(energy)
```

```
> abline(h=0, col="gray")

> remove(.x)

> plotMeans(energy$expend, energy$stature, error.bar

> tapply(energy$expend, list(stature=energy$stature)
stature
     lean     obese
 8.066154 10.297778

> summary(energy)
     expend        stature
 Min.   : 6.130   lean :13
 1st Qu.: 7.660   obese: 9
 Median : 8.595
 Mean   : 8.979
 3rd Qu.: 9.900
 Max.   :12.790
```

```
ERROR: No factors selected.
```

Summary

- Intro to pairwise comparisons
- The t-test and other pairwise tests
- Intro to multiple hypothesis testing
- Methods to correct for multiple hypothesis testing
- Using the R Commander
- Exercises

References

- Zar JH. *Biostatistical Analysis.* Prentice Hall, 1998. [or any general biostatistics textbook]
- Dalgaard P. *Introductory Statistics with R.* Springer, 2002.
- Venables W.N. and Ripley B.D. Modern Applied Statistics with S. Springer, 2002.
- Tufte E. *The Visual Display of Quantitative Information.* Graphics Press, 1992.
- Lots of web sites
- R documentation

Exercise 2 – To do

Using Excel and R:

- Transform data to get a more normal distribution
- Perform different t-test flavors for several different types of data
- Compute power for some t-tests
- Perform a Wilcoxon rank sum test
- Given a series p-values, perform multiple hypothesis testing
  - Bonferroni
  - FDR

Helpful R commands

```
# HTML help page
help.start()
```

```
# Help for a command
?t.test

# Search help pages
help.search("t-test")

# Read a tab-delimited data table
energy = read.delim("energy.txt", header=T)
dat = read.delim(file("clipboard"), h=T)
rats = read.delim("http://.../rats.txt", h=T)
```

Google mailing lists, etc. => r-project t-test

Exercise 2 functions

Excel

- TTEST
- RANK
- SORT
- LOG

R

- t.test
- wilcox.test
- p.adjust
- log
- power.t.test