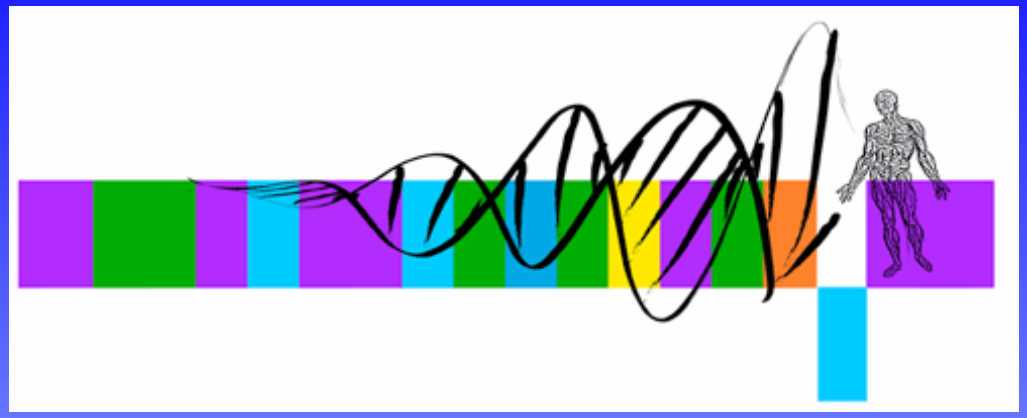


Statistics for Biologists

Lecture 3:

Selected biostatistical applications

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

Outline

- Lecture 2 review
- Analyzing the intersection of two sets
- Comparing proportions
- Analyzing survival data
- Regression and correlation analysis
- Determining over-representation

From lecture 2

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

Example: Analyzing the intersection between two sets

- A knockout mouse model has certain genes that are downregulated.
- A mouse treated with a drug has certain genes that are downregulated.
- What is the intersection of these gene lists?
- Is this intersection greater than we'd expect by chance?
- Biological goal: to determine if the gene knocked out and the drug act through a similar mechanism.

Sample data: Analyzing the intersection between two sets

- Using an array of 15,000 genes
  - the knockout model had 150 "down" genes
  - the drug-treated mice had 120 "down" genes
  - the intersection is 20 genes

	Knockout model		
Drug treatment		Yes	No
	Yes	20	100
	No	130	14,750

- Question: Is the proportion of "down" genes with the different genotypes influenced by the drug treatment?
- Need to find: Is the proportion of genes in both sets the product of the proportion of genes in each set * the total number of genes?

Tests of independence: analyzing an intersection

- Input data as a contingency table

20	100
130	14,750

- Statistical tests:
  - Fisher's exact test: best test for 2x2 matrix
  - Chi-squared test: best test for larger matrices; approximate for 2x2 matrix
- Excel command: CHITEST
  - but expected frequencies needed as input

- R commands:

```
# Create a matrix of 2 rows (and 2 columns)
ko.vs.drug = matrix(c(20, 100, 130, 14750), nrow=2)
fisher.test(ko.vs.drug)
chisq.test(ko.vs.drug)
```

Fisher's exact test output: Analyzing the intersection between two sets

```
> fisher.test(ko.vs.drug)

        Fisher's Exact Test for Count Data

data:  ko.vs.drug
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not
   equal to 1
95 percent confidence interval:
 12.87998 38.26746
sample estimates:
odds ratio
  22.67740
```

Conclusion: drug and KO are not independent

$$\frac{14750/130}{100/20}$$

```
> chisq.test(dat)$expected
          KO.yes     KO.no
Drug.yes     1.2     148.8
Drug.no    118.8   14731.2
```

Expected values, assuming indpendence

The Chi-squared goodness of fit test to check for Mendelian ratios

- The most traditional biological use of the chi-squared test
- Example: number of colors from a flower cross

red	pink	white
54	122	58

- Do these equal the expected ratio 1:2:1 ?
- R commands:

```
flowers = c(54, 122, 58)
chisq.test(flowers, p=c(0.25, 0.5, 0.25))
```

p-value = 0.75

Survival analysis

- Given data about patient or animal mortality, can we calculate the probability of being alive at a given time?
- Given data for >1 different groups, does one live longer than the other?
- Lifetime data are generally censored: exact time of death is not known:
  - left censoring: upper bound is known
  - right censoring: lower bound is known

Survival analysis goal 1: estimate the survival function

- The survival function *S(t)*: the probability of surviving more than t units of time

- The Kaplan Meier estimate: reduce survival probability by

  $\frac{R_t - 1}{R_t}$ if there is a death a time t

  and a population of R_t is still alive at that time

- Input data required for estimate:
  - age animal was last observed
  - if animal was alive or dead

Survival analysis goal 1: sample data and R commands

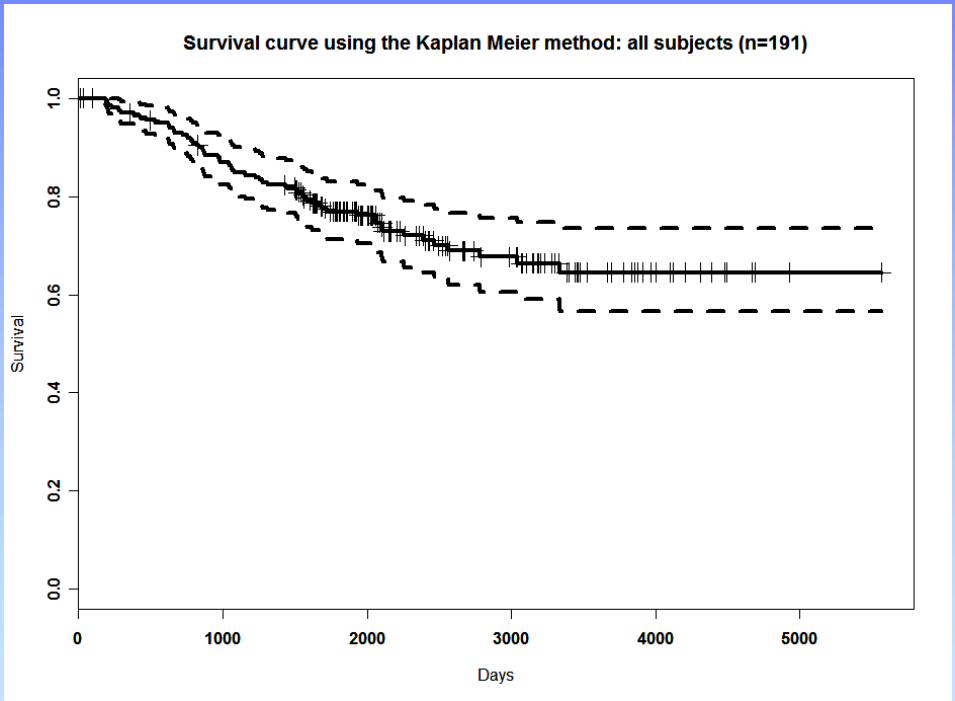
mouse ID	days	dead 0 = no; 1 = yes	group
A	350	0	Control
B	450	0	Control
C	103	1	KO
D	305	0	KO

```
library(survival)   # load the R package
dat = read.delim("my.data.txt", header=T)
# Compute an estimate of a survival curve
# Look at all data together
my.surv.all = survfit(Surv(days, dead), data=dat)
# Does group influence survival?  =>  '~' means "described by"
my.surv.by.group = survfit(Surv(days, dead) ~ group, data=dat)
```

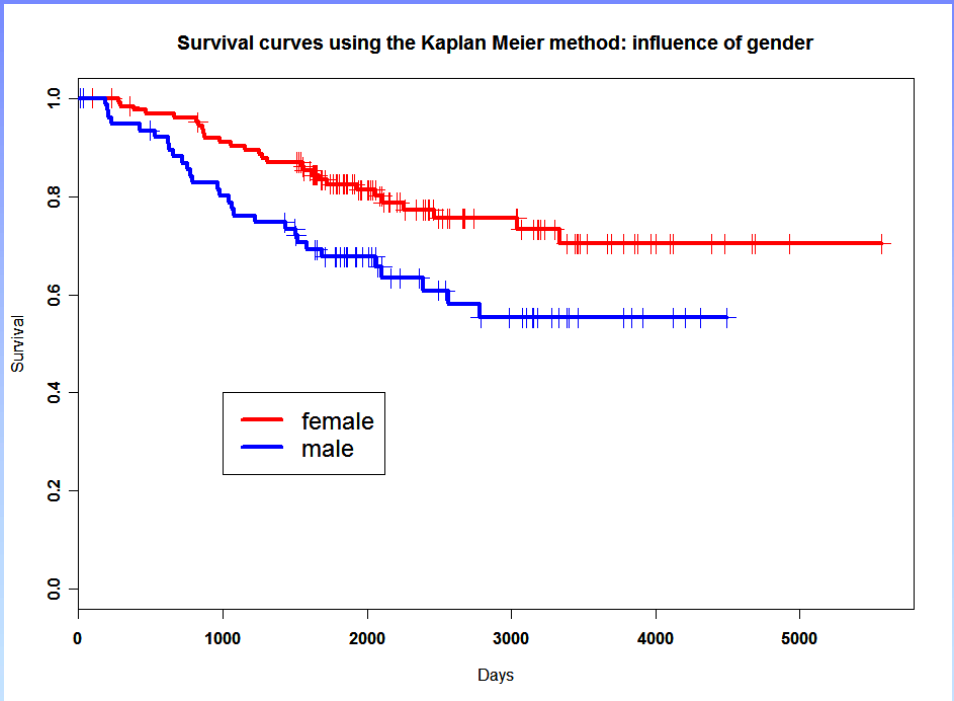
Survival analysis goal 1: sample R output

summary(my.surv) # Get text output

plot(my.surv.all)



plot(my.surv.group)



Confidence interval(s) (95%) can be shown

Survival analysis goal 2: compare two curves

- Use the log-rank test to compare the survival function for each group at each time point using nonparametric statistics

```
> survdiff(Surv(days, dead) ~ group, data=dat)

Call:
survdiff(formula = Surv(days, dead) ~ group, data = dat)

                N Observed Expected (O-E)^2/E (O-E)^2/V
group=control 119       28     37.6      2.47      7.27
gruop=KO       72       29     19.4      4.80      7.27

 Chisq= 7.3  on 1 degrees of freedom, p= 0.00699
```

Regression and correlation

- Both create a model for the relationship between two (or more) variables
- regression: technique for describing a relationship in which the magnitude of one variable is a function of another variable – *but not necessarily cause and effect*

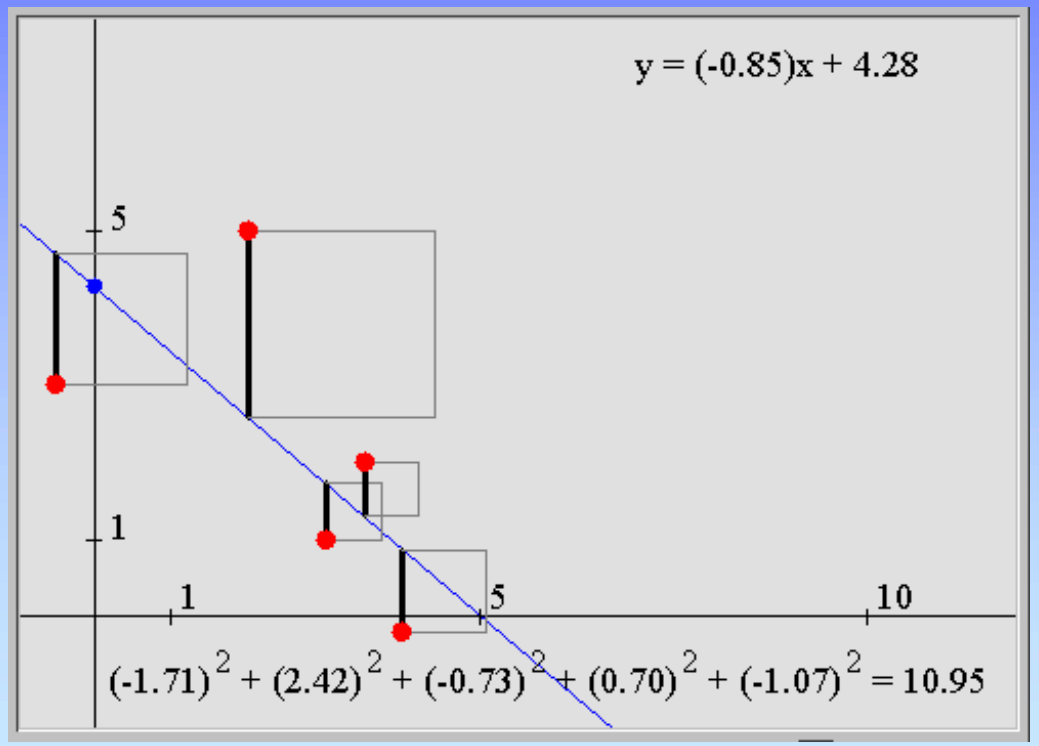
  ex: *height = f(age)* *cell size = f([glucose])*
- Simple linear regression: tries to explain one variable with another variable with an equation of the form

$$y = \beta x + \alpha + \varepsilon$$

  where
  - ε = error = residual α = y-intercept
  - β = slope = regression coefficient
- Does this make sense biologically? visually (from the plot)? Could data be transformed?

Visual example: linear regression

- How can we draw a "best fit line"?
  - the method of least squares: minimize $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$



where
y = actual value of y
$\hat{y}$ = predicted value of y

So to get slope(β) and intercept(α)

$$\beta = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\alpha = \bar{y} - \beta\bar{x}$$

figure from http://standards.nctm.org

Significance of a regression line

- How sure are you that
  - the slope of the line $\neq 0$?
  - the intercept of the line $\neq 0$? [or some other slope to test]
- For the slope, use the Student's t-test
  - H_0: slope = 0 H_a: slope $\neq$ 0

    and calculate t = (estimated slope – slope being tested) / s_b

    where s_b is the standard error of the slope (a messy equation)
- The intercept can be tested in a similar manner
  - but value of intercept may be of questionable biological meaning

Example: simple linear regression

Create a linear model (lm) of height as a function of age in R:

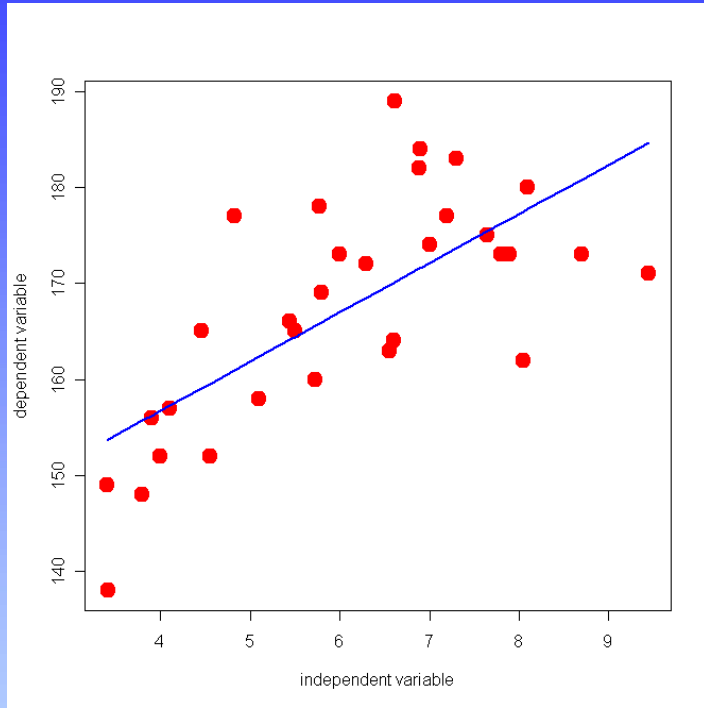
```
> summary(lm(height ~ age))

Call:
lm(formula = height ~ age)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7284  -5.7207  -0.2493   4.3736  18.8360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 136.2688     6.0748  22.432  < 2e-16 ***
age           5.1201     0.9652   5.305 9.85e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.725 on 30 degrees of freedom
Multiple R-Squared: 0.484,      Adjusted R-squared: 0.4668
F-statistic: 28.14 on 1 and 30 DF,  p-value: 9.853e-06
```

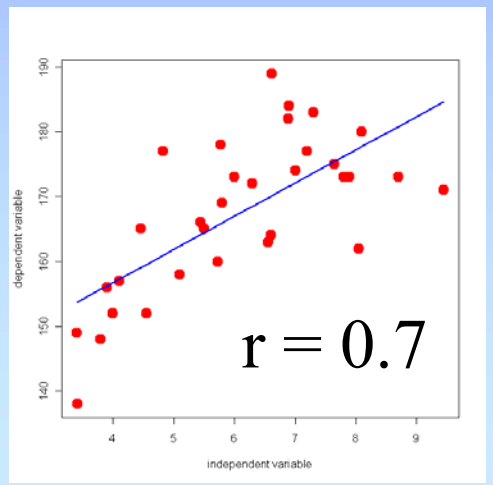


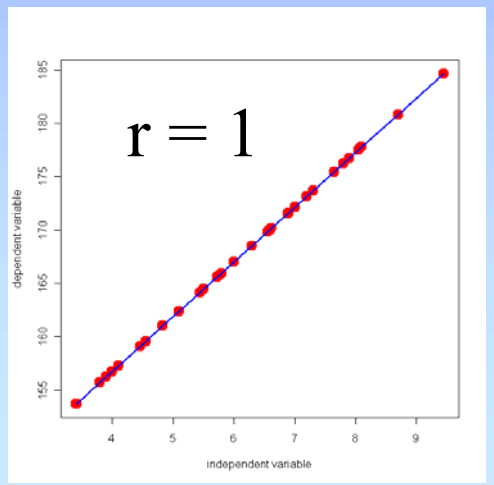
For Excel, see the Data Analysis tools

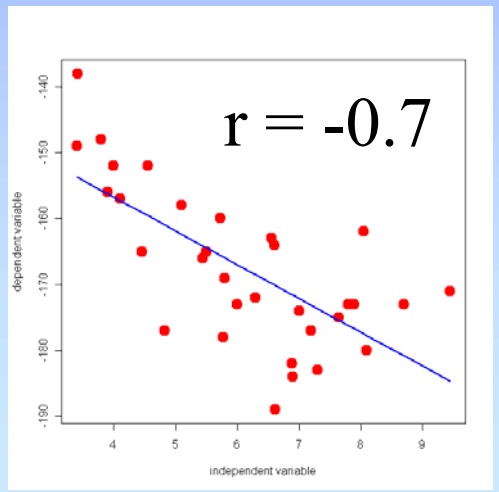
Correlation

- The correlation coefficient (r) describes the strength and the direction of the relationship between two variables $-1 \leq r \leq 1$
- coefficient of determination = r^2

  describes the percentage of the variation in y that is determined by x (e.g., the regression line)







Correlation commands

- Standard correlation analysis with assumes that the Y's at each X – and the X's at each Y -- are normal
- In Excel
  - PEARSON or CORREL
  - ex: PEARSON(A2:A6,B2:B6)
- In R
  - cor(a, b, method=c("pearson", "kendall", "spearman"))
  - "pearson" is the standard method (with assumptions)
  - other methods are nonparametric (without assumptions)
  - try cor.test(a, b) to test for correlation

Example: Determining over-representation

- Using Gene Ontology terms, 15,000 genes from a model species have been annotated
- A screen identifies 150 genes.
- We're interested in "development" (GO:0007275)
- Number of genes assigned this term:
  - In the whole genome: 900
  - In the genes from our screen: 15
- Are more genes from this screen involved in development than one would expect by chance?

Sample data: Determining over-representation

- Input data

	Development	Any GO term	Ratio
Genome	900	15,000	0.06
Gene set	15	150	0.10

- Is the gene set ratio significantly larger than the genome ratio?
- Biological interpretations:
  - Is the gene set enriched for developmental genes?
  - Is development one of the "themes" of this gene set?

More precise question: Determining over-representation

Given the prevalence P of a factor (ex: a GO term) in a population of size N and the prevalence of p in a sample (n),

- In a random sample of size n, what is the probability that this factor would be present more often than observed?
- What is the sum of the probabilities for (p+1)/n to n/n times?
- Is this P(more often than observed) close to zero?
- If so, is the observed prevalence significantly greater than the expected prevalence?

Hypergeometric distribution: Determining over-representation

- The hypergeometric distribution describes the probability of "sampling without replacement"
  - P(drawing cards) vs. P(rolling dice)
- To get the probability of an event occurring more often than p times in a set of size n, we have two choices:

1: $P(p+1) + P(p+2) + \ldots + P(n)$

2: $1 - (P(0) + P(1) + \ldots + P(p))$

Hypergeometric distribution: Determining over-representation

- Input:

	# of Hits	Size
Gene set	15	150
"Genome"	900	15,000

- In Excel: HYPGEOMDIST
  - calculates each separate probability
  - they must be summed
- In R: dhyper(0:h, H, N-H, n) for a list of separate probabilities

h = hits in sample = 15	n = size of sample = 150
H = Hits in population = 900	N = size of population = 15,000

Hypergeometric distribution: input and output from R

- Input data

h = hits in sample = 15	n = size of sample = 150
H = Hits in population = 900	N = size of population = 15,000

- Get the separate probabilities and find their cumulative sum:
  pvals = 1 - cumsum(dhyper(0:h, H, N-H, n))

```
> pvals = 1 - cumsum(dhyper(0:15, 900, 15000-900, 150) )
> pvals
 [1] 0.99991121 0.99905197 0.99492728 0.98183123 0.95089329 ...
[13] 0.11632698 0.06667349 0.03580922 0.01805521
```

- Get the last in the list and print it out

```
> pval = pvals[length(pvals)]
> cat("p = ", pval, "\n")
p =  0.01805521
```

Practical applications of GO enrichment analysis

- Tools like GO Term Finder (and many other GO tools) can analyze all GO terms for over-representation
- Multiple hypothesis testing is usually necessary
- Statistical vs. biological significance:
  - beware of too general or too specific terms

Summary

- Analyzing intersection between 2 sets
  - Fisher's exact test or chi-squared test
- Comparing proportions
  - Chi-squared test
- Analyzing survival data
  - Kaplan Meier curves; log-rank test
- Regression and correlation analysis
- Determining over-representation
  - Hypergeometric distribution
- 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 3 - To do

Using Excel and R:

- Analyzing intersection between 2 sets
  - Fisher's exact test or chi-squared test
- Comparing proportions
  - Chi-squared test
- Determining over-representation
  - Hypergeometric distribution
- Analyzing survival data
  - Kaplan Meier curves; log-rank test
- Regression and correlation analysis