Statistics for Biologists

Lecture 1:

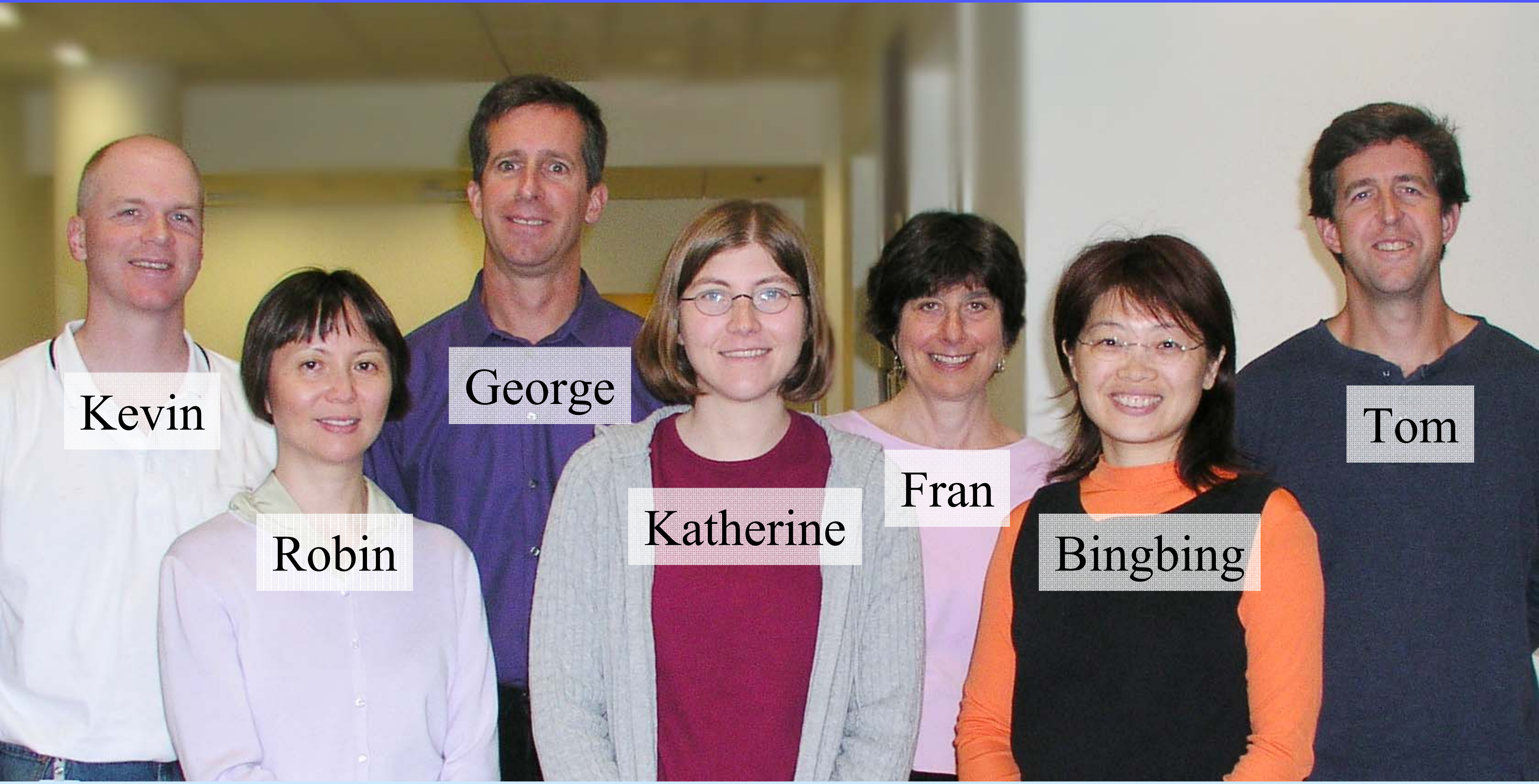
Descriptive statistics and introduction to inferential statistics

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

Course goals

- Review concepts in statistics most relevant to Whitehead researchers
- Discuss general biological questions and examples of statistical applications
- Perform methods using Excel and/or R
- Encourage the use of statistics before, during, and after experimentation

Bioinformatics and Research Computing



Outline

- Why use statistics?
- Descriptive statistics
- Visualization of quantitative data
- Intro to inferential statistics
- False positives and false negatives
- Statistics software
- Exercises

Why use statistics?

- To reduce data to a manageable amount with an effective summary
- To determine if data are worth getting excited about
- To separate interesting variability from uninteresting variability
- To measure the reliability and confidence of your or others' conclusions
- To plan more effective experiments

Descriptive statistics

- Measures of central tendency + variability
- Generally try to describe a population using sample data
- No hypotheses; no p-values; no comparisons
- Reduce the quantity of data
- Should data be summarized?
- What measures should be used?

The mean

- Other names: average; arithmetic mean
- Sample mean = $\overline{X}$
- Population mean = μ
- The center of gravity of a histogram
- All measurements contribute
- Not robust to outliers
- More robust modification: the trimmed mean
- Geometric mean = $\sqrt[n]{a_1 \bullet a_2 \bullet \ldots \bullet a_n}$

  <= arithmetic mean; only works with positive numbers

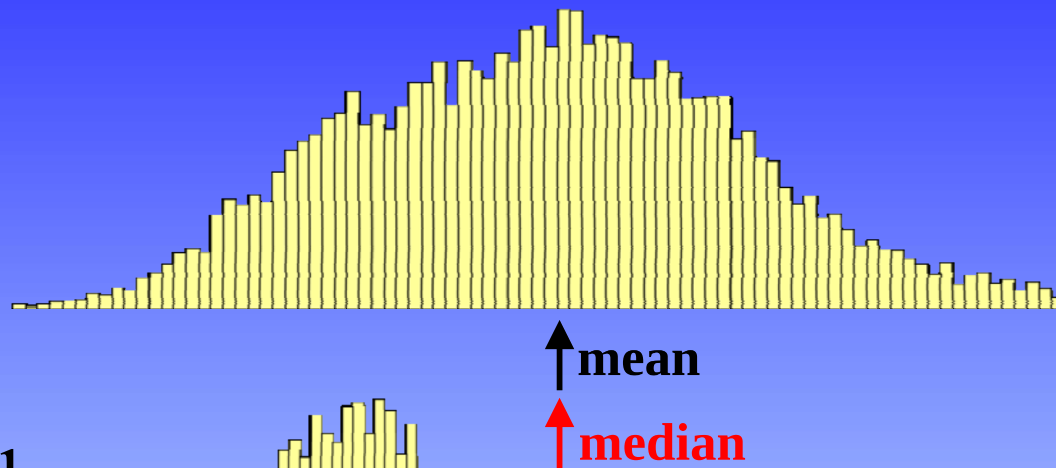
  antilog of the (arithmetic) mean of the logs of values

The median

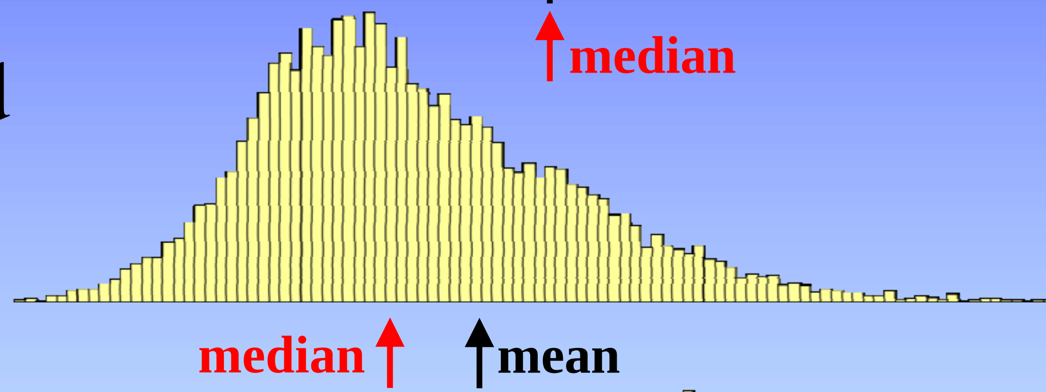
- The middle measurement in an ordered set of data
- With an even number of points, use the mean of the two center points
- Divides a histogram into two equal areas
- Most measurements don't contribute
- Contains less information than the mean
- Robust (resistant) to outliers

Frequency distributions

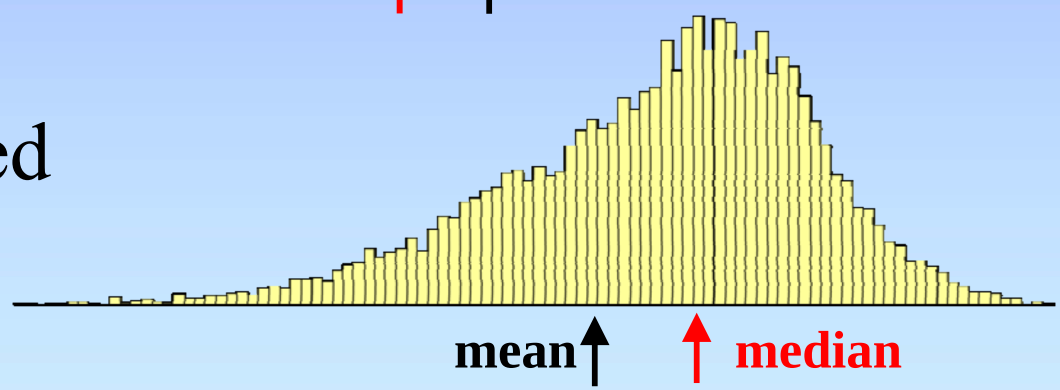
- Symmetrical



- Positively skewed (skewed right)



- Negatively skewed (skewed left)



Measures of variability

- Range (minimum – maximum)
- Interquartile range (25th – 75th percentiles)
- Standard deviation
  - describes variability in a population
  - STDEV in Excel
  - Has same units as original measurements
  - positive square root of variance

$$= \sqrt{\frac{\sum (x_i - \text{mean})^2}{N - 1}}$$

- Standard error [of the mean] (SE; SEM)
  - describes the stdev of sample means

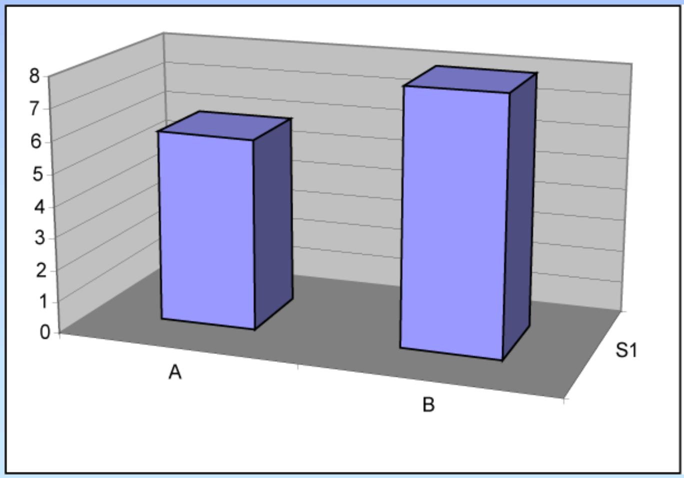
$$= \text{stdev} / \sqrt{N}$$

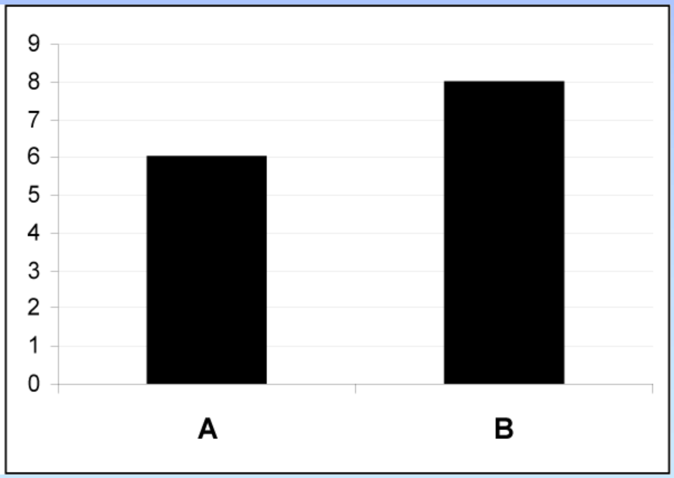
N = number of measurements

- coefficient of variation CV = 100 * stdev/mean

Visualizing data

- What are you trying to show?
- Is the figure understandable on its own?
- Is the main idea(s) clear?
- Have you removed unnecessary "junk"?





Visualizing data

- All data, summaries, or both?
- Some types of figures:
  - scatterplot
  - bar plot (mean ± stdev)
  - boxplot (median, IQR, outliers)
  - histogram
  - volcano plot (fold change vs. p-value)
  - pie chart

Hypothesis testing - part 1

- In statistics, it's a formal way of asking a question.
- Differentiates between two hypotheses:
  - null hypothesis (H_0): "there's no difference"
  - alternative hypothesis (H_a): $a \neq b$; $a < b$; $a > b$
- The magnitude of the difference is not part of the hypothesis.
- Statistical hypotheses are stated <u>before</u> data collection and examination.

Hypothesis testing - part 2

- If there's enough evidence, we can reject the null hypothesis.
- If there's not enough evidence, we can't say that there is no difference – just that there is not enough evidence to support a difference.
- Alternative hypotheses:

  $a \neq b$ => 2-tailed tests

  $a < b$ or $a > b$ => 1-tailed tests

Hypothesis testing and decision making

1. Calculate test statistic
2. Compare test statistic to distribution of values obtained if there was no difference between data sets (e.g., if H_0 was true)
3. Get p-value

   = probability of getting a result at least this extreme if H_0 were true
4. Compare p-value to selected cutoff (α; "significance level")
5. Accept H_0 ("there is no difference") or reject H_0 ("there is a difference")

Types of errors with hypothesis testing

	Conclusion from statistical test		
Reality		Accept H_0 (means are the same)	Reject H_0 (means differ)
	Means are the same	Everybody's happy	Type I error (False positive) probability = α
	Means differ	Type II error (False negative) probability = β	Everybody's happy

Types of errors with hypothesis testing

- The p-value from a statistical test reflects the false positive error rate.
- The p-value indicates nothing about your confidence at identifying a difference that exists in reality.
- To get an idea of the false negative error rate, calculate the power of the test:

  power = 1 - β ex: if power = 0.95, ...

  inputs: n; stdev; true difference; α

Selecting a significance level (α)

- The choice of α should be made before looking at the data.
- What error rate can you and others tolerate?
- Choosing a significance level of 0.05 is based more on convention than on statistical reasoning.
- Increasing n (sample size) reduces the probability of false positives and false negatives.
- If the p-value for a statistical test is close to α, increasing n may help to determine which hypothesis is supported.
- Substituting another statistical test (or variation) just to achieve α invalidates your statistics.
- α and β are inversely related.

Receiver Operating Characteristic (ROC) curves

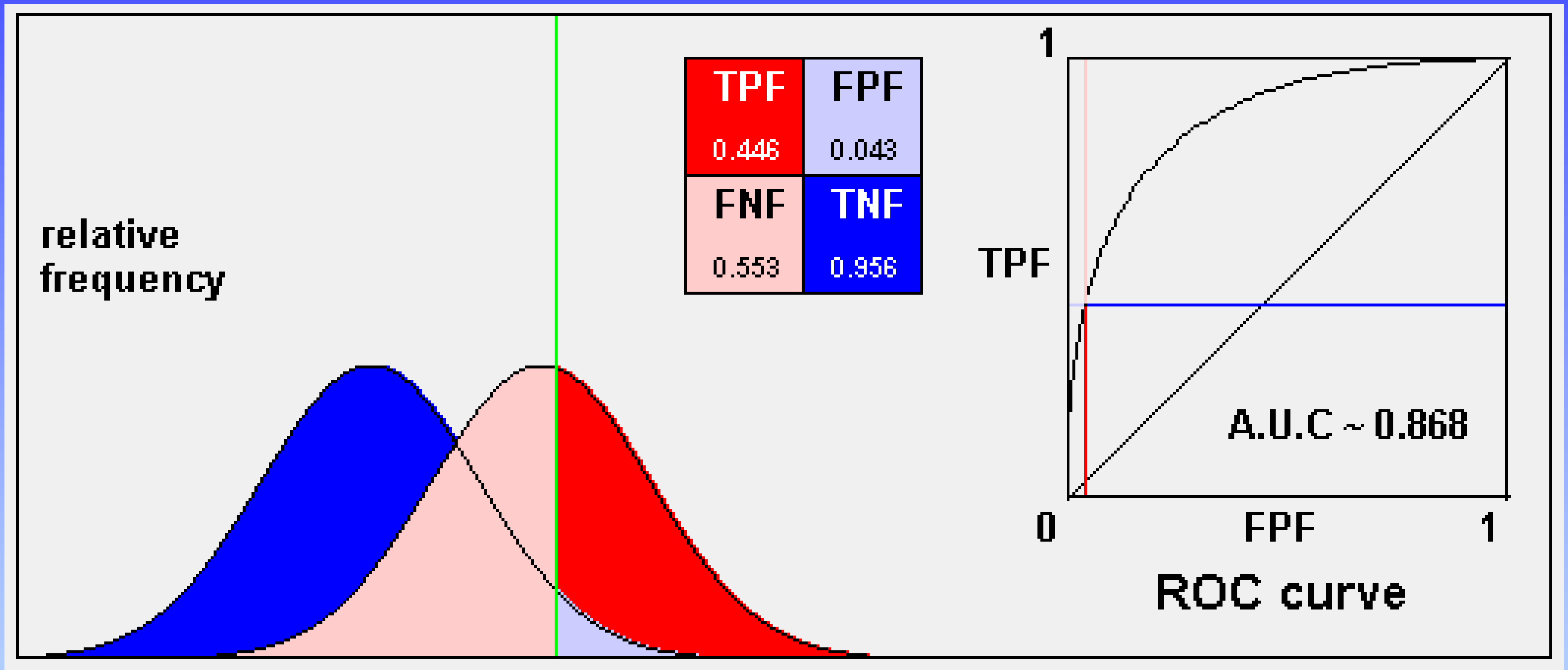
$p < 0.043$

[TF][PN]F =

True/False

Positive/ Negative

Fraction



- red curve: distribution of positive data
- blue curve: distribution of negative data

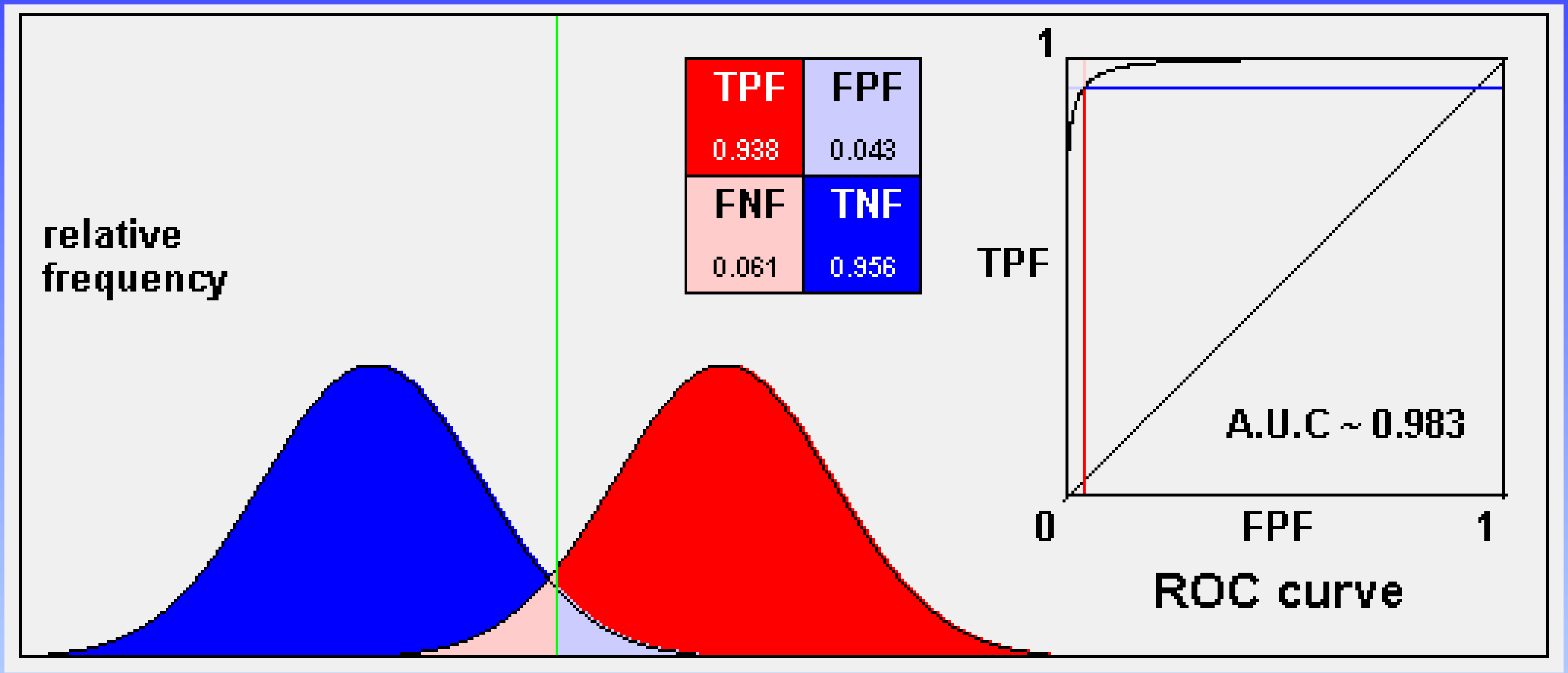
A.U.C. = area under curve

sensitivity = TP/(TP+FN)

specificity = TN/(FP+TN)

ROC curves: good scenario

p < 0.043

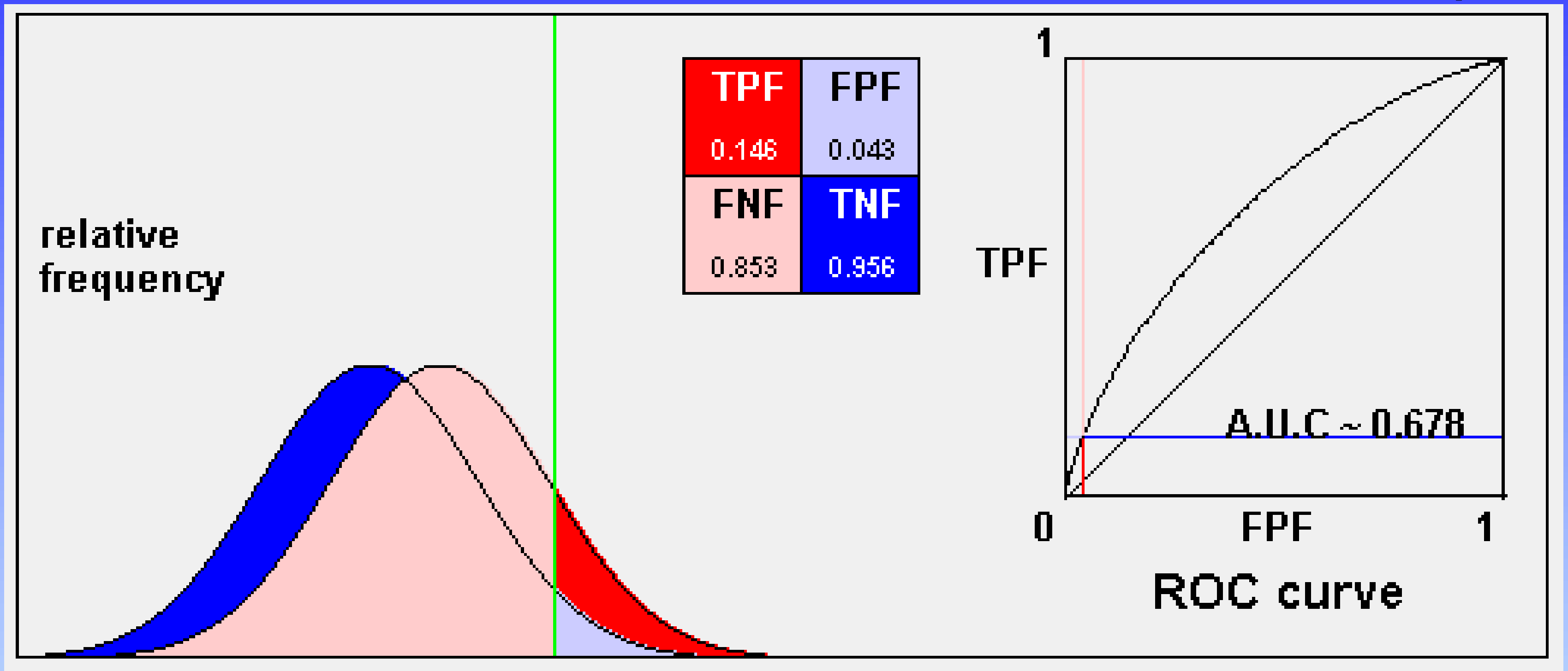


- red curve: distribution of positive data
- blue curve: distribution of negative data

A.U.C. = area under curve
sensitivity = TP/(TP+FN)
specificity = TN/(FP+TN)

ROC curves: bad scenario

p < 0.043



- red curve: distribution of positive data
- blue curve: distribution of negative data

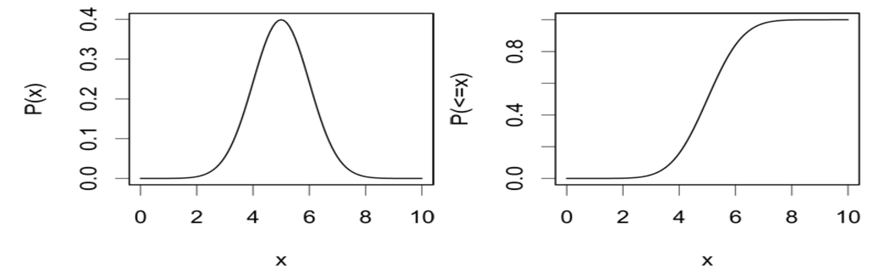
A.U.C. = area under curve
sensitivity = TP/(TP+FN)
specificity = TN/(FP+TN)

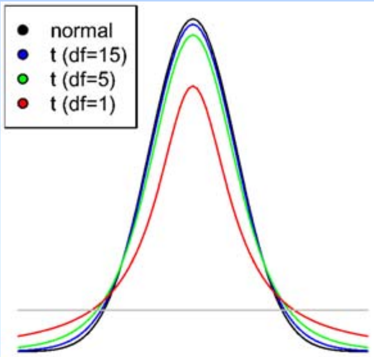
Probability distributions

- Functions describing the probability that a variable will have a given value.
- discrete or continuous
- Examples of distributions
  - normal (Gaussian)
  - binomial: successes in P/F experiments
  - Student's t: a family of distributions for small sample sizes
    - approaches the normal when N (or df) approaches infinity

density

CDF = cumulative distribution function



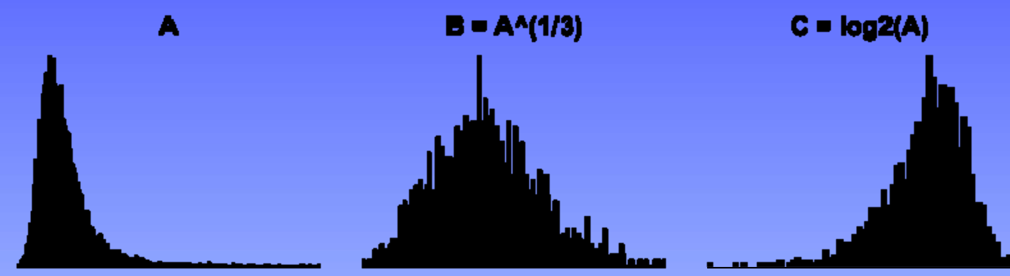


How normal is your distribution?

- Plot a histogram.
  - Is it bell-shaped?
- Compare mean and median.
  - Are they the same?
- Try verifying the empirical rule
  - What percent of measurements are $\pm\ \sigma$, etc.?
- Plot a quantile-quantile (q-q) plot
  - Does it make a straight line?

Transformations to create a more normal distribution

- For positively (right) skewed data:
  - Square root
  - Logarithm
  - Inverse (1/x)



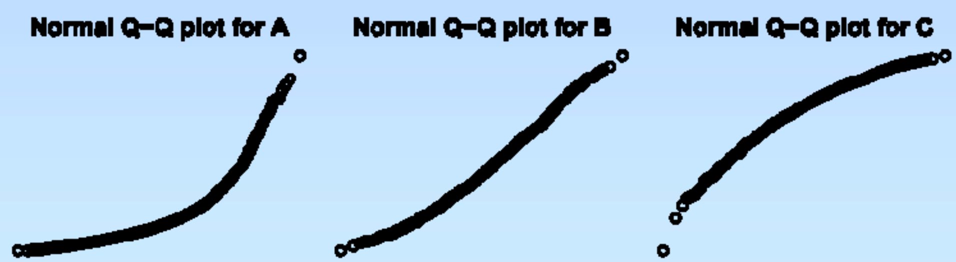
- For negatively skewed data:
  - Reflect data ➔ add constant ➔

    ➔ perform above method ➔ reflect again

- Check results with a quantile-quantile plot:

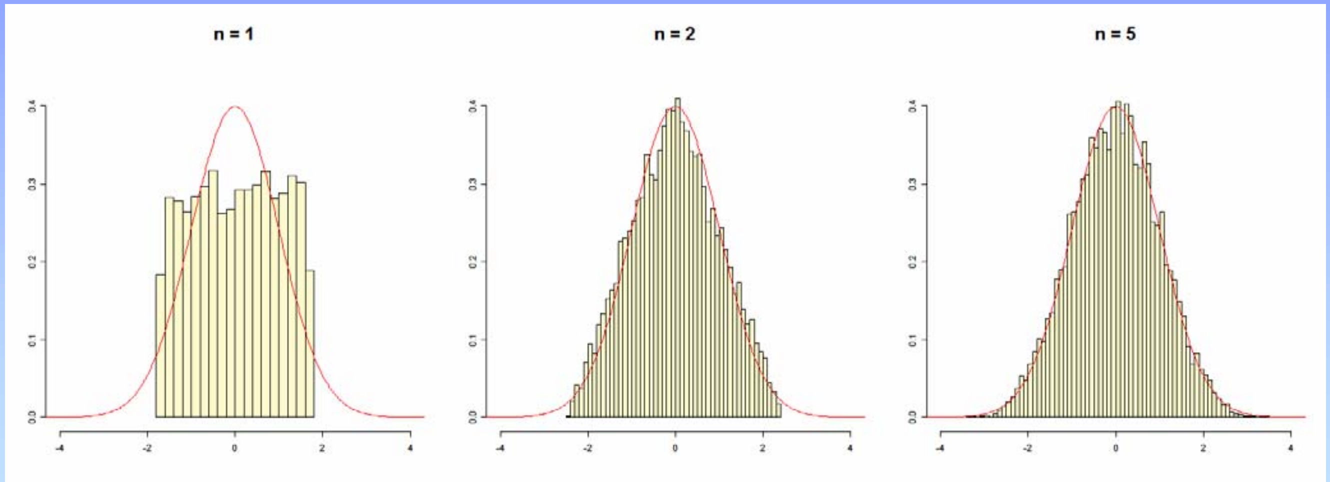
  sample quantiles vs.

  theoretical quantiles



Central limit theorem

- Even with data that comes from a distribution that is far from normal, the distribution of averages tends to be normal.
- This distribution of averages can be made arbitrarily close to normal by increasing the sample size.



- The theorem permits inferences about a population when we only have data about a sample

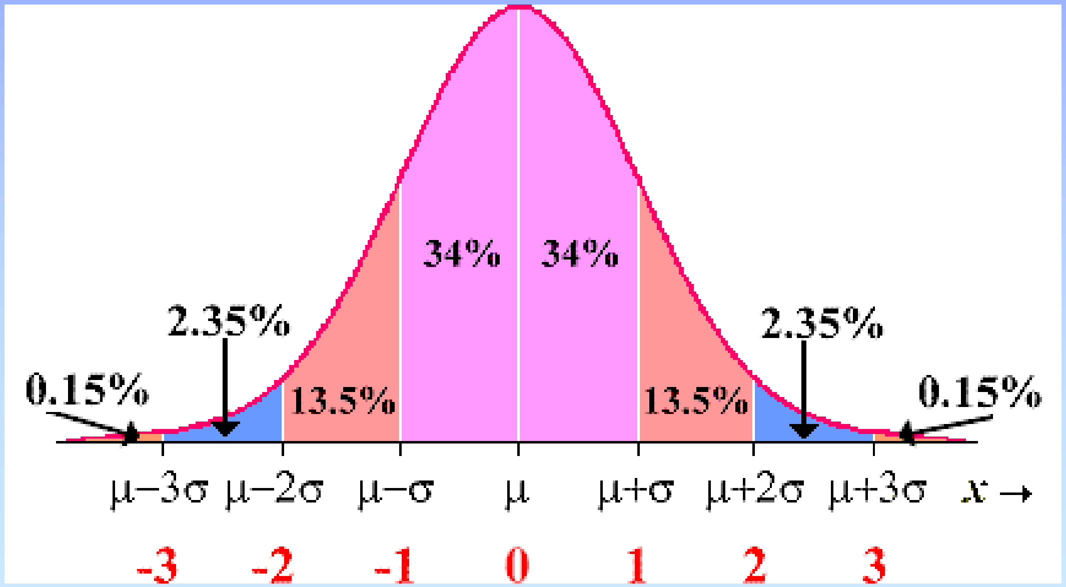
The Empirical Rule

For a normal distribution:

~68% of the data fall in the interval mean ± 1 stdev

~95% of the data: mean ± 2 stdevs

~99.7% of the data: mean ± 3 stdevs



Describing data relative to a normal distribution

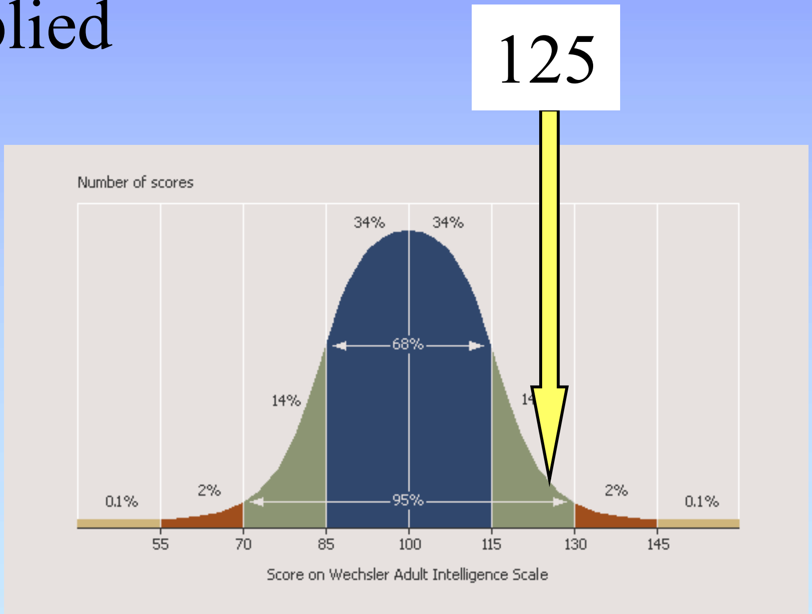
- Z-score = number of standard deviations from a population mean in a normal distribution

$$z = \frac{\text{measurement} - \text{mean}}{\text{stdev}}$$

- The Empirical Rule can then be applied
- Ex: sample score = 125
  - population mean = 100
  - stdev = 15

    z = (125 - 100) / 15 = 1.67

    `pnorm(125, mean=100, sd=15)`
  - 95.2% measurements are below



Statistics tools

- Excel
- Office Calc (www.openoffice.org)
- The R Project for Statistical Computing
  - http://www.r-project.org/
- Bioconductor (microarray packages for R)
  - http://www.bioconductor.org/
- BaRC analysis tools:
  - http://iona.wi.mit.edu/bio/tools/bioc_tools.html

Excel or Office Calc for statistics

- Good:
  - familiar
  - friendly and flexible interface
- Bad
  - not good for inferential statistics
  - not robust on older computers
  - Limited number of rows: 2^{16}

R for statistics

- Good:
  - powerful and flexible
  - created and used by Ph.D. statisticians
  - commands can be saved as scripts
  - comes with microarray analysis routines
- Bad:
  - command-line interface takes a while to figure out

Exercise 1 - Excel syntax

A2	Cell reference
A2:A100	Series of cells
=B5	Formula
=B5	Absolute link ('$')
=data!B4	Reference other sheet
=[otherFile.xls]data!B4	Reference other file

Exercise 1: Excel functions

- AVERAGE
- MEDIAN
- STDEV
- TRIMMEAN
- PERCENTILE
- CONFIDENCE
- VLOOKUP
- Tools >> Data Analysis

The R Project for Statistical Computing - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.r-project.org/

George's Genome Br... WIBR genome regio... Camargo Lab - Lolig... Eorsa TargetScan Jura TargetScan Nriti's orthology tools 2can Bioinformatics ...

The R Project for Statistical Computing

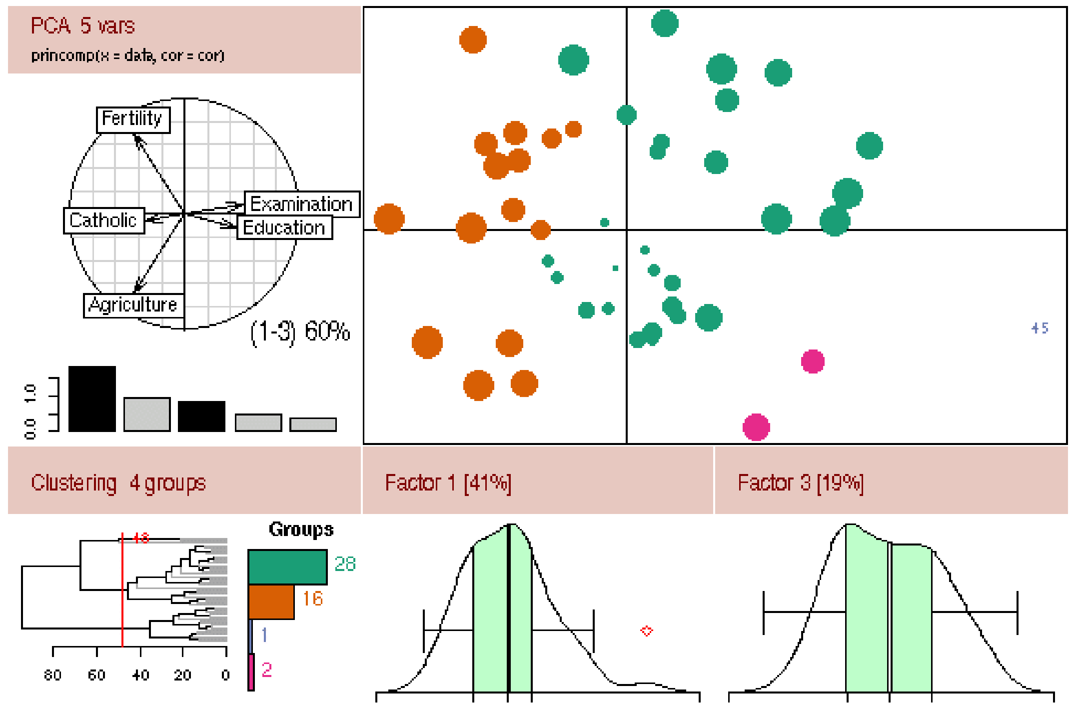
About R
[What is R?](#)
[Contributors](#)
[Screenshots](#)
[What's new?](#)

Download
[CRAN](#)

R Project
[Foundation](#)
[Members & Donors](#)
[Mailing Lists](#)
[Bug Tracking](#)
[Developer Page](#)
[Conferences](#)
[Search](#)

Documentation
[Manuals](#)
[FAQs](#)
[Newsletter](#)
[Wiki](#)
[Books](#)
[Other](#)

Misc
[Bioconductor](#)
[Related Projects](#)
[Links](#)



Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News:

- **R version 2.4.0** has been released on 2006-10-03.
- **DSC 2007**, the 5th workshop on Directions in Statistical Computing, February 15-16, 2007, Auckland, New Zealand.
- **R News 6/4** has been published on 2006-10-31.
- **The R Wiki** provides an online forum where useRs can help other useRs.

Introduction to R

```
# Read a data file
dat = read.delim("Data1.txt", header=T)
dim(dat)        # Get dimension of matrix
summary(dat)        # Get data summary
colnames(dat)        # Get names of columns
mean(dat[,"my.col.1"])
# Draw a boxplot of first 2 columns of matrix
boxplot(dat[,1], dat[,2])
q()     # quit [or use pull-down menu]
```

Summary

- 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

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 1 - To do

Using Excel and R:

- Calculate some descriptive statistics
  - mean, median, stdev, IQR, CI
- Draw some figures
  - histogram, scatterplot, boxplot, Q-Q plot