Course Introduction, Descriptive Statistics and Data Visualization

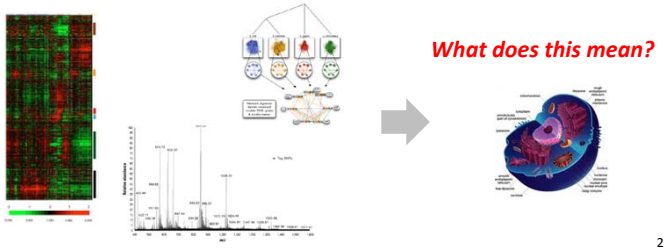
GENOME 560, Spring 2012

Su-In Lee, CSE & GS
suinlee@uw.edu

1

Why Taking This Course?

- ***Data*** are interesting because they help us understand the world
- *Genomics*: Massive Amounts of Data Data …
- Statistics is fundamental in genomics because it is integral in the **design, analysis** and **interpretation** of experimental data



2

Why Taking This Course?

- ***Data*** are interesting because they help us understand the world
- *Genomics*: Massive Amounts of Data Data …
- Statistics is fundamental in genomics because it is integral in the **design, analysis** and **interpretation** of experimental data
- This course covers the key statistical concepts and methods necessary for extracting biological insights from experimental data

3

Learning Goals

- 5 weeks is too short to cover every specific topic that might arise in the course of your research…
- It is not a good strategy to treat what we learn in this course as "recipes" to follow
- Instead, we should focus on
  - rigorous understanding of fundamental concepts that will provide you with the tools necessary to address routine statistical analyses
  - foundation to understand and learn mode specific topics

4

Course Schedule

- **Syllabus:**

Date	Topic
May 1	Descriptive Statistics and Data Visualization
May 3	Random Variables and Probability Theories
May 8	Probability Distributions
May 10	Parameter Estimation
May 15	Regression Methods
May 17	Hypothesis testing I – t-test, confidence interval
May 22	Hypothesis testing II – ANOVA
May 24	Hypothesis testing III – Analysis of Categorical Data
May 29	Bootstrapping, cross validation and permutation tests
May 31	Assessing significance in high dimensional experiments

  - Special topics that may be discussed in class include Bayesian networks, Expectation Maximization (EM) algorithm, principal component analysis

- **Grading:** 5 problem sets (20% each)

5

Books and Resources

- Course website
  - http://www.cs.washington.edu/homes/suinlee/genome560/
- No required text
- Good on-line resources
  - http://www.math.wm.edu/~trosset/Courses/351/book.pdf
  - http://www.statsoft.com/textbook/stathome.html
  - http://www.stat.berkeley.edu/~stark/SticiGui/Text/toc.htm
- Some good books if you ever have some extra $$$:
  - Probability and Statistics for Engineering and the Scientists 6th Ed. Jay L. Devore (2004). Duxbury press, Thompson-Brooks/Cole.
  - Statistical Inference. Casella, G. and Berger, R. L. (1990). Wadsworth, Belmont, CA.
  - Probabilistic Graphical Models: Principles and Techniques. Koller, D. and Friedman, N. (2009). MIT Press.

6

Class Meetings

- Class meets twice a week
  - Tue/Thu 9-10:20am @ Foege S110
- Each class will last for 80 minutes and be primarily lecture based
- Other forms of learning and interactions will be included
- We will often interrupt lectures to work on problems in small groups as well as work through statistical analyses using R (please bring a laptop with R installed!)

7

What is R?

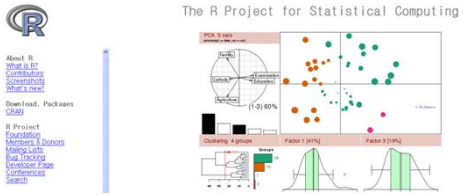
- The R statistical programming language is a free open source package based on the S language developed by Bell Labs
- May statistical functions are already built in
- Contributed packages expand the functionality to cutting edge research
- Amazing graphics
- Widely used in genetics, genomics, computational biology

8

R Resources

- Windows, Mac and Linux binaries available at

  http://www.r-project.org



- Extensive resources at the above web-site, in particular see:

  http://cran.r-project.org/other-docs.html

9

Lecture 1: Descriptive Statistics and Data Visualization

10

Outline

- What is descriptive statistics and exploratory data analysis?
- Basic numeral summaries of data
- Basic graphical summaries of data
- Basic operations in R
- (If time permits) How to use R for calculating descriptive statistics and making graphs

11

Why Descriptive/Graphical Summary?

- Before making inferences from data, it is essential to examine all your variables
- Why?
- To listen to the data:
  - to catch mistakes
  - to see patterns in the data
  - to find violations of statistical assumptions
  - to generate hypotheses
  - … and because if you don't, you will have trouble later

12

Types of Data

- Categorical
  - Binary: 2 categories
  - Nominal: more categories
  - Ordinal: order matters
  - E.g. gender, ethnicity, disease state, genotypes, etc
- Continuous (or Quantitative)
  - Numeric values that can be ordered sequentially, and that do not naturally fall into discrete ranges.
  - E.g. weight, number of seconds it takes to perform a task, gene expression levels, etc

13

Dimensionality of Data Sets

- **Univariate:** Measurement made on one variable per subject
- **Bivariate:** Measurement made on two variables per subject
- **Multivariate:** Measurement made on many variables per subject

14

Numerical Summaries of Data

- ***Central tendency measures.*** They are computed to give a "center" around which the measurements in the data are distributed.
- ***Variation or variability measures.*** They describe "data spread" or how far away the measurements are from the center.
- ***Relative standing measures.*** They describe the relative position of specific measurements in the data

15

Central Tendency Measures: Mean

- To calculate the **mean** of a set of observations, add their value and divide by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

16

Central Tendency Measures: Median

- **Median:** the exact middle value
- **Calculation:**
  - If there are an odd number of observations, find the middle value
  - If there are an even number of observations, find the middle two values and average them
- **Example:**

  Some data:

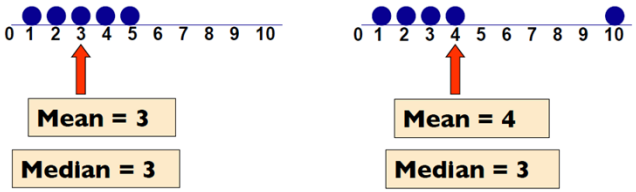
  Age of participants: 17 19 21 22 23 23 23 38

  Median = (22+23)/2 = 22.5

17

Which Measure Is Best?

- **Mean** is best for symmetric distributions without outliers
- **Median** is useful for skewed distributions or data with outliers



18

Scale: Variance

- Average of squared deviation of values from the mean

$$\hat{\sigma}^2 = \frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}$$

19

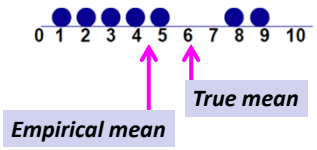
Why Squared Deviations?

- Squares eliminate the negatives
- Absolute values do not have nice mathematical properties
- **Result:**
  - Increasing contribution to the variance as you go farther from the mean

20

Why Divide By (n-1), not n ?

$$\hat{\sigma}^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$



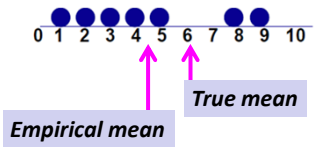
- You compute the difference between each observation and the mean of all *n* observations.
- You don't know the true mean of the population; all you know is the mean of your samples (*empirical* mean)
- Except for the rare cases where the sample mean happens to equal the population mean, the data will be closer to the sample mean than it will be to the true population mean.
- So the numerator will probably be a bit smaller (and can't be larger) than what it would be if you used the true mean.

Biased estimator of the population variance

21

Why Divide By (n-1), not n ?

$$\hat{\sigma}^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$



- To make up for this divide by (n-1) rather than n.

Unbiased estimator of the population variance

- If you knew the sample mean, and all but one of the values, you could calculate what that last value must be. Statisticians say there are n-1 degrees of freedom.

22

Scale: Standard Deviation

- Variance is somewhat arbitrary
- What does it mean to have a variance of 10.8? Or 2.2? Or 1459.092? Or 0.000001?
- Nothing. But if you could "standardize" that value, you could talk about any variance (i.e. deviation) in equivalent terms
- Standard deviations are simply the square root of the variance

23

Scale: Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

$$\hat{\sigma} = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}}$$

24

Interesting Theoretical Result

- Regardless of how the data are distributed, a certain percentage of values must fall within *k* standard deviations from the mean

Note use of μ (mu) to represent "mean".

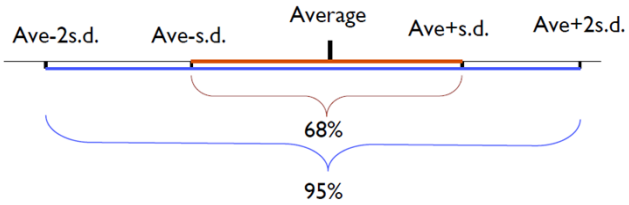
Note use of σ (sigma) to represent "standard deviation."

At least	within
$(1 - 1/1^2) = 0\%$	$k=1$ $(\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$	$k=2$ $(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	$k=3$ $(\mu \pm 3\sigma)$

25

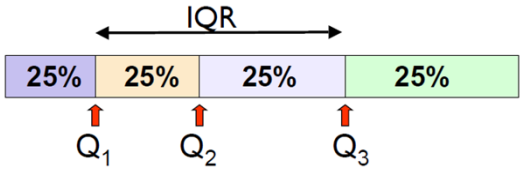
Often We Can Do Better

- For many lists of observations, especially if their histogram is bell-shaped
  - Roughly 68% of the observations in the list lie within 1 σ (standard deviation) of the average
  - 95% of the observations lie within 2σ of the average



26

Scale: Quartiles and IQR

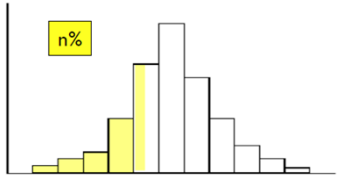


- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the Q_3

27

Percentiles (aka Quantiles)

- In general the n^{th} percentile is a value such that n% of the observations fall at or below of it



$Q_1 = 25^{th}$ percentile

Median = 50^{th} percentile

$Q_2 = 75^{th}$ percentile

28

Graphical Summaries of Data

- Dimensionality of data matters …
  - **Univariate:** Measurement made on one variable per subject
  - **Multivariate:** Measurement made on many variables per subject

29

Univariate Data

- Histograms and bar plots
- What is the difference between a histogram and bar plot?

Bar plot:

- Used for categorical variables to show frequency or proportion in each category
- Translate the data from frequency tables into a pictorial presentation…

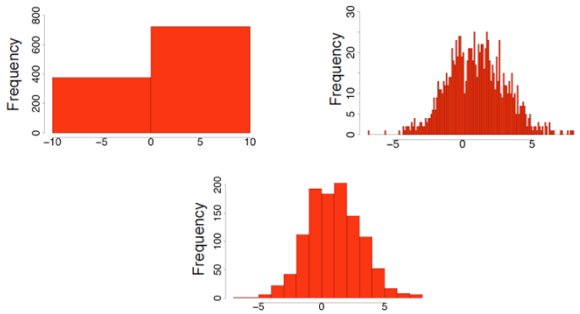
Histogram:

- Used to visualize distribution (shape, center, range, variation) of continuous variables
- "Bin size" is important

30

Effect of Bin Size on Histogram

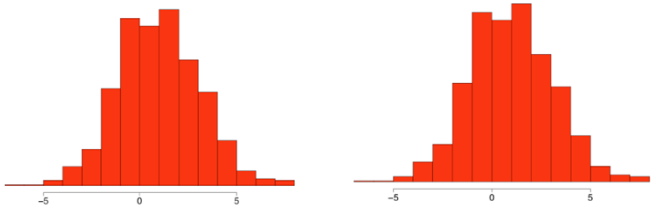
- Simulated 1,000 N(0,1) and 500 N(1,1)…



31

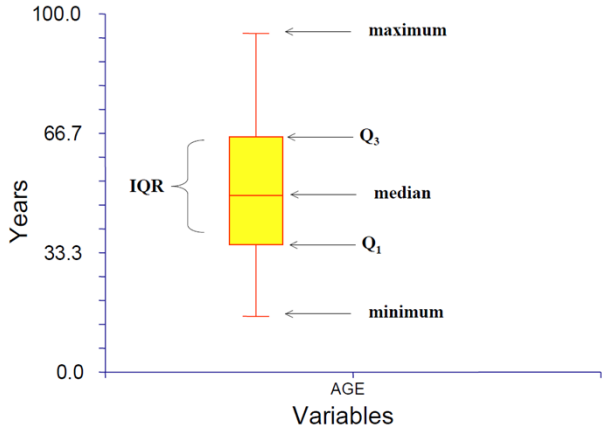
More on Histograms

- What's the difference between a frequency histogram and a density histogram?



32

Box Plots



33

Multivariate Data

- **Clustering**
  - Organize variables into clusters
  - Descriptive, not inferential
  - Many approaches
  - "Clusters" always produced
- Data reduction approaches
  - Reduce n-dimensional dataset into much smaller number
  - Finds a new (smaller) set of variables that retains most of the information in the total sample
  - Effective way to visualize multivariate data

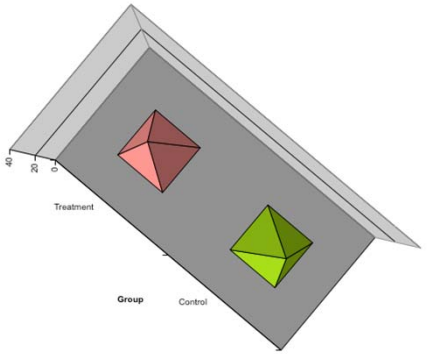
34

How to Make a Bad Graph

- The aim of **good** data graphics:
  - Display data accurately and clearly
- Some rules for displaying data **badly**:
  - Display as little information as possible
  - Obscure what you do show (with chart junk)
  - Use pseudo-3d and color gratuitously
  - Make a pie chart (preferably in color and 3d)
  - Use a poorly chosen scale

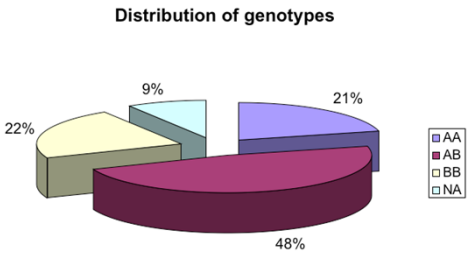
35

Example 1



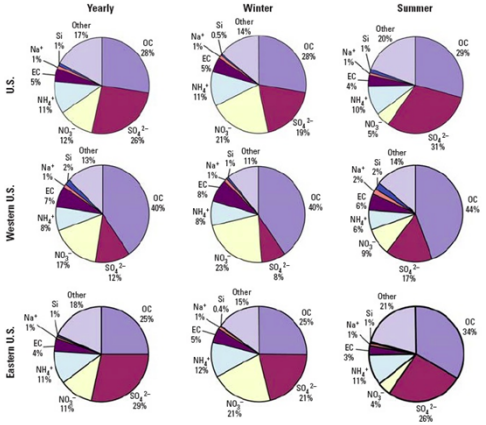
36

Example 2



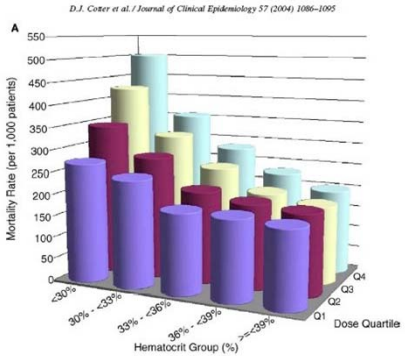
37

Example 3



38

Example 4



39

Goals of Our R Tutorial Today

- Installing R
- Using R as a fancy calculator
- Data structures: scalars, vectors, data frames, matrices
- Reading in data from a file
- Subsetting and extracting data
- Writing and executing simple R scripts

40

Probabilities

- Event
- Variable $\mathbf{X}XP(S)P(S)\ P(S)\log\log$
- Probability
- Statistical independence
- Joint probability

41

Probabilities

- Probabilities of mutually independent events are summed

  Pr(a die comes up 2 or 3)

  = Pr(comes up 2) + Pr(comes up 3)

  = 1/6 + 1/6 = 1/3
- Probabilities of independent events are multiplied to get the joint probability

  Pr(one die comes up 2 and the other one comes up 3)

  = Pr(first one comes up 2) + Pr(second one comes up 3)

  = 1/6 + 1/6 = 1/3
- Conditional probabilities are the joint probability divided by the probability of the event that they are conditioned on:

$P(D|S) = \int P(D|S,\theta)P(\theta|S)d\theta$

42

$$P(D|S) = \int P(D|S,\theta)P(\theta|S)\,d\theta$$

43

Goals of Our R Tutorial Today

- Installing R
- Using R as a fancy calculator
- Data structures: scalars, vectors, data frames, matrices
- Reading in data from a file
- Subsetting and extracting data
- Writing and executing simple R scripts

44

Some Jargon

- **Units:** the basic objects on which the experiment is done
- **Variable:** a measured characteristic of a unit
- **Treatment:** any specific experimental condition applied to the units. A treatment can be a combination of specific values (called levels) of each experimental factor

45