BMI 713 / GEN 212

Lecture 1: Probability & Distributions

- Introduction
- Statistical inference
- Binomial distribution
- Normal distribution
- Inference on a mean

September 9, 2010

General Information

- 1:02-2:30 lecture
- 2:45-4 hands-on exercises
- The TAs and I are here to help!
- Textbooks
- Contact info
  - emails: put '[stat]' in the subject line
  - feedback is welcomed and appreciated at any time

General Information

- A bit about my lab's research
  - Epigenetics: fly, human; dosage compensation, cancer, stem cells
  - Analysis of data from high-throughput sequencing platforms (ChIP-seq, RNA-seq, nucleosomes, CNV)
- Acknowledgments:

  Lecture notes are partially based on materials by
  - **Kim Gauvreau** (Bio201, HSPH)
  - **Marcello Pagano** (Bio200, HSPH)
  - **Rebecca Betensky, Yves Chretien**, HST 190

Introduction

- Survey of medical students in Britain (1980)

  Epidemiology: "dull, and neither useful nor difficult"

  Statistics: "neither interesting nor helpful and… difficult"
- Unfortunately you will not understand how important it is until you have to use it
- It is now easy to collect large amounts of data, both from your experiments and from public resources
- The challenge is to sort through all the data and find meaningful results
- Statistics is more important that ever!

Statistics

- … is not magic. A statistician cannot tell you which data point is correct and which data point is wrong.
- … is not torturing the data until it confesses!
- … can be misleading.
  - A p-value is an estimate based on many assumptions.
  - It is generally difficult to determine whether these assumptions are satisfied.
  - Theoretical results generally require a large sample size.
  - You may have a statistical significant but biologically irrelevant result, depending on what your null hypothesis is.
- … is not a black box (google is not the best way to do it)
- … is art, not science.

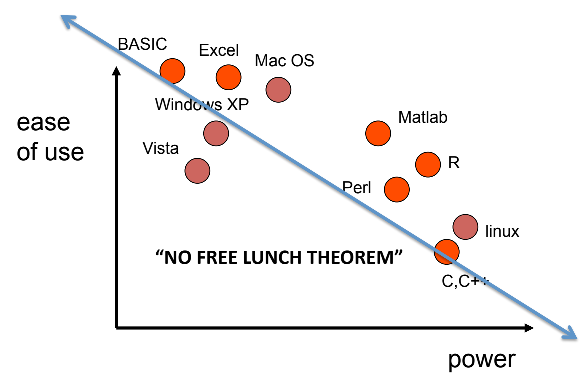
Setting proper expectations

- This class: emphasis on thinking critically about statistics; basic data analysis and visualization skills
- The difficult question is which test to apply, not how to apply it
- You will not be an expert statistician or a programmer in a semester!
- Different people will respond differently
- Trade-off between rigor and practicality
- This is an experiment!

Comments from last year

- Blog/webgroup/discussion board?
- "Peter is wonderful and extremely dedicated, just that he has over estimated the ability of the student over the programming"
- "Be prepared to spend a significant amount of time on the homework. Try to find a study group to work with on the homework."
- "If there are examples of how to use R in the class lecture slides, make sure you try to do them for yourself after class"
- "ASK QUESTIONS early and frequently. Make it a personal goal to ask at least one question (especially if you think it is dumb) per lecture."
- "Single people out to participate … it would have totally petrified me but would have helped"

What software to use?

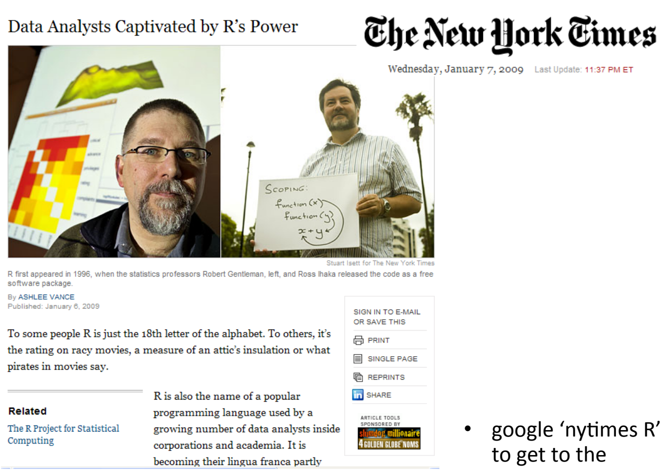


Excel

- Excel can be a useful tool
- Cannot handle very large data sets
- Probably not good for most substantial analysis
- What happens if you want to redo the entire analysis with one more sample?
- Watch out for automatic conversion errors:
  - At least 30 genes, e.g., DEC1 → 1-DEC
  - ~2000 Riken identifier, e.g., 2310009E13 → 2.31E+13
  - Ref: Zeeberg et al, *BMC Bioinformatics*, 2004

The R Language

- Data analysis involves repeated tasks that can be automated and shared
- Interactive views of the data let you “get to know” the data and can help guide analysis
- Should avoid re-inventing the wheel
- Should take advantage of state-of-the-art algorithms developed by others
- Reproducible research requires access to computational code (most results are NOT re-producible from the information given in the paper)
- R is free; many people actively working on it
- Not the easiest one to get started



Data Analysts Captivated by R's Power

The New York Times

Wednesday, January 7, 2009 Last Update: 11:37 PM ET

Stuart Isett for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly

Related
The R Project for Statistical Computing

SIGN IN TO E-MAIL OR SAVE THIS
PRINT
SINGLE PAGE
REPRINTS
SHARE
ARTICLE TOOLS SPONSORED BY
slumdog millionaire 4 GOLDEN GLOBE NOMS

- google ‘nytimes R’ to get to the article

Outline

- Introduction
- Statistical inference
- Binomial distribution
- Normal distribution
- Inference on a mean

Statistical Inference

- Statistical inference: drawing a conclusion about a population or a general phenomenon on the basis of a limited sample.
- **Deductive reasoning proceeds from general to specific.**
  - Example: Systolic blood pressure in a population is normally distributed with mean 140 and std. dev. 9. What fraction of the population has SBP ≥155?
- **Inductive reasoning proceeds from specific to general.**
  - Example: Out of 20 mice tested, 8 mice responded to the drug treatment I developed for diabetes. What is the best estimate of the proportion expected to response in the mouse population? How confident are you?

Types of Data

- **Nominal:** unordered categories
  - blood type, gender
- **Ordinal:** ordered categories
  - severity of condition: mild, moderate, severe
- **Discrete:** counts
  - number of hospital visits
- **Continuous:** spectrum of values
  - systolic blood pressure

Basic Probability

- **Random variable:** a variable that can assume a number of different values such that any particular outcome is determined by chance
- **Discrete random variable:** a finite or countable number of outcomes
  - e.g., number of infections
- **Continuous random variable:** can take on any value within a specified interval or continuum
  - e.g., height

How do we measure variability?

- Suppose my data points are $x_1, x_2, \ldots, x_n$
- How about $\sum_{i=1}^{n}(x_i - \bar{x})$?
- How about $\sum_{i=1}^{n}|x_i - \bar{x}|$?
- How about $\sum_{i=1}^{n}(x_i - \bar{x})^2$?
- Do we want it to depend on the number of samples?
- Divide by n? (n-1)? "degree of freedom"

Mean and Variance

- Mean $\bar{x} = \sum_{i=1}^{n} x_i$
- Variance $s^2 = Var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Standard deviation $s = \sqrt{Var(X)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- Standard notation: when known, μ, σ
  when estimated, $\hat{\mu}, \hat{\sigma}$ or $\bar{x}, s$
  (capital X for random variable, small x for data)

Random Variables

- **Random variable:** a variable that can assume a number of different values such that any particular outcome is determined by chance
- **Discrete random variable:** a finite or countable number of outcomes
  - e.g., number of infections
- **Continuous random variable:** can take on any value within a specified interval or continuum
  - e.g., height

Probability Distributions

- **Binomial** distribution: discrete-valued
- **Normal** distribution: continuous-valued
- Some of the statistical tests will have discrete and continuous versions
- The normal distribution ("Gaussian") can be used to approximate a binomial distribution

Binomial Distribution

- I flipped the coin 5 times and got all heads. Is this a biased coin?
- How unusual is it to have 3 children be all girls?
- Binomial distribution
  - two categories: "success" and "failure"
  - each trial is independent with probability *p*
  - a fixed number of trial
- This is a **discrete** probability distribution

Binomial Distribution

- Let *Y* be a random variable that represents a coin flip. *Y*=1 if a Head, *Y*=0 if a Tail. Let *p* be the probability of obtaining a Head
- Suppose we have two coin tosses
- Introduce a new variable *X* that represents the number of Heads in the two trials

Outcome	of Y	Probability	X
0	0	(1-p)(1-p)	0
1	0	p(1-p)	1
0	1	(1-p)p	1
1	1	pp	2

- P(X=1) = p(1-p) + (1-p)p
- P(X=0)+P(X=1)+P(X=2) = 1

Binomial Distribution

- What is the probability that a family with six children have exactly four boys?

- Probability of having a boy: p=½
- Probability of having four boys: p^4
- Probability of not having two boys is $(1-p)^2$
- How many ways are there to have 4 girls out of 6 children?
  - "6 choose 4": 6!/(4!2!) = 15

$$\binom{6}{4} p^4 (1-p)^2 = 15/64 = .234$$

Binomial Distribution

- What about having **at least** 4 girls?

$$\binom{6}{4} p^4 (1-p)^2 + \binom{6}{5} p^5 (1-p)^1 + \binom{6}{6} p^6 (1-p)^0$$

- If a random experiment has two possible outcomes and we do *n* independent repetitions with identical success probability *p*, then *X ~ Bin(n,p)* and

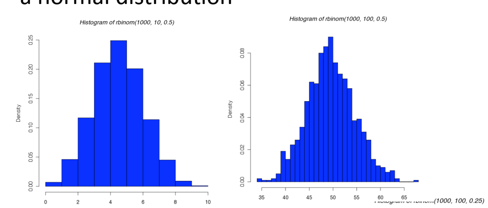
$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- The probability of obtaining at least *k* successes is

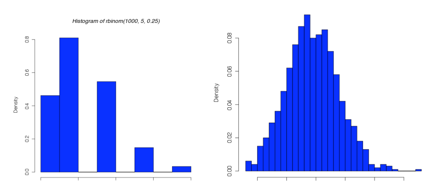
$$P(X \ge k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

Normal Approximation to the Binomial Distribution

- For large *n,* the binomial distribution can be approximated by a normal distribution



1000 samples
n=10, n=100
p=.5



1000 samples
n=10, n=100
p=.25

Normal Approximation to the Binomial Distribution

- What are the mean and variance of the normal approximation?

- ***E(X) = np***

  (E(X) is the mean or the "expected value" of X)

- ***Var(X) = npq*** where ***q=1-p***

Normal Distribution

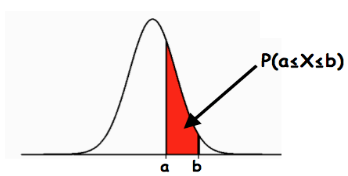
- 'Bell-shaped'; 'Gaussian'

$$X \sim N(\mu, \sigma^2)$$

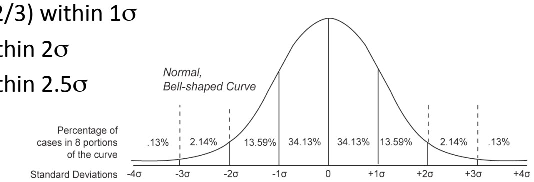
- "X is normally distributed with mean μ and variance σ^2"
- Normal random variables are continuous.
- "P(X=a)" does not make sense.
- We need to write intervals, e.g., "P(a≤X ≤b)" or "P(X ≤c)"
- Why is this ubiquitous?

Normal Distribution

- *P(a ≤ X ≤ b)* = probability that X falls between *a* and *b*



- 68% (~2/3) within 1σ
- 95% within 2σ
- 99% within 2.5σ



Standard Normal

- To figure out the probability that a normal random variable falls in a given range, first transform the variable into a standard normal variable.

"z-score"

- If $X \sim N(\mu, \sigma^2)$ and $Z = \dfrac{X - \mu}{\sigma}$, then $Z \sim N(0,1)$

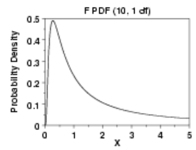
- In fact, this is the form of most statistical testing:

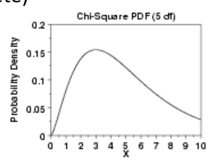
$$\frac{\text{Statistic - Hypothesized value}}{\text{Square root of the variance of the statistic}}$$

follows a known probability distribution

A Little Bit of Statistical Theory

- There are few well-known distributions in statistics
  - Normal distribution
  - t-distribution
  - F-distribution
  - χ^2(chi-square)-distribution
  - Binomial distribution (discrete)
  - Poisson distribution (discrete)





Example

- Systolic blood pressure in a population is normally distributed with mean 140 and std. dev. 9. What fraction of the population has SBP ≥155?”
- Solution: $X \sim N(140, 9^2)$

$$P(X \geq 155) = P\left(\frac{X-140}{9} \geq \frac{155-140}{9}\right)$$

$$= P\left(Z \geq \frac{155-140}{9}\right)$$

$$P(Z \geq 1.67) = 0.0475$$

Inference on a Mean

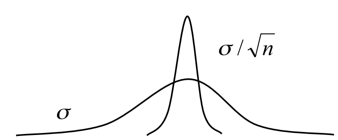
- Suppose height is normally distributed. I take a sample of 10 students and calculate the average. Is the deviation of this average from the population average is unusual?
- Let’s do a thought experiment
- Take a sample of 10, calculate its average $\bar{x}_1$
- Take another sample of 10, calculate its average $\bar{x}_2$
- Take another sample of 10, calculate its average $\bar{x}_3$
- What does the distribution of $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ look like?
- What is distribution of the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?

Distribution of Sample Mean

- $\bar{X}$ are approximately normally distributed when *n* is large

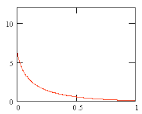
$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

- The larger the sample size, the less spread you should see
- Standard deviation of sample mean → ‘standard error’
- The denominator has square root of *n*, not *n*

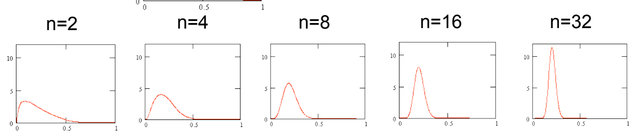


A Little Bit of Statistical Theory

- For $\overline{X}$ to be normally distributed, does X have to be normally distributed?
- The distribution of an average tends to be Normal, even when the distribution from which the average is computed is non-Normal (Central Limit Theorem)



Suppose my variable of interest has this distribution. What is the distribution of the averages?



Confidence Intervals

- Previously, we saw that for a normal variable, 95% of the data are contained in $(\mu - 2\sigma, \mu + 2\sigma)$
- So, we know that there is 95% probability that

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

contains the true mean

- After we draw the sample we cannot say, "The probability that μ is contained in the interval is 95%."
- μ is fixed, not random. Once we have calculated the interval, it simply either contains μ or it doesn't.

R programming

- Interpreted vs compiled languages
- When R is running, all variables, data, etc are stored in the memory as objects
- Users act on these objects with *operators* and *functions*
- Functions have default and optional arguments
- Functions need to be written with parentheses
- Various editors can be used to write scripts
- Users can read/write files (need to specify directory)
- Should learn to be comfortable with vector and matrix operations
- Many packages are available

Resources

- R project: http://www.R-project.org
- Downloading R: http://cran.r-project.org
- Graphics examples: http://addictedtor.free.fr/graphiques
- Bioinformatics packages: http://www.bioconductor.org
- To search mailing list archive http://tolstoy.newcastle.edu.au/R
- Manuals: http://cran.r-project.org/other-docs.html
- I would recommend:

http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf