Introduction

Outline for today:

- About the course
- Course objectives
- About the instructor
- Why we use R
- Dryad with Heather Piwowar
- Organizing data for analysis in R
- Totally non-threatening quiz
- Review of some basic concepts in statistics
- First Discussion paper

Course components:

- Lectures: Every Thursday afternoon 1 - 3 pm, Biodiv 224.
- Presentations and discussions of papers: Also on Thursday between 1 and 3
- R workshops: (aka “labs”) Tuesday afternoon 1 - 3 pm, Biodiv 224.
- Assignments: Occasional.
- There are no exams.
- First presentation and discussion is next Thursday – need two volunteers!
- First workshop is next Tuesday – Introduction to R.
- Bring your own laptops if you have one, with latest R (2.14) installed.
- See the course web site for recommended add-on packages to install.
- For those with no laptop: I will borrow a Biodiversity Centre laptop (Mac) for you.

Web site

- http://www.zoology.ubc.ca/biol548/
- Updated regularly – hit your refresh button.
- Lecture overheads will be placed there in pdf format before lecture time.
- Discussion papers will be placed there.
- Assignments will also be put there.
- The “R tips” help pages contain just about all the clues you will need to carry out workshops.
- Help is organized by topic, roughly corresponding to workshops.

Intended list of lecture topics

1. Introduction
2. Graphics
3. Experimental design
4. Linear models
5. Mixed-effects models
6. Likelihood
7. Generalized linear models
8. Model selection
9. Bayesian methods
10. Computer intensive
11. Meta analysis
12. Multivariate methods
13. Reproducibility project presentations

The course

- Developed by Dolph Schluter in response to needs identified by graduate students in the Biodiversity Research Centre. Please help to improve it.
- This is a “second” course in data analysis, to take you beyond the most basic, introductory level, which I’m assuming you have already (up to ANOVA and linear regression).
- Students who eavesdrop are welcome – bring your own laptop.

Grading of registered students based on

- Presentations of readings.
- Contribution to discussions of readings.
- Occasional assignments involving data and R.

Textbook

- No required textbook.
- Course web site lists useful books, many available online.
- Use Whitlock and Schluter (2009) or alternative as a basic stats reference.

About the instructor

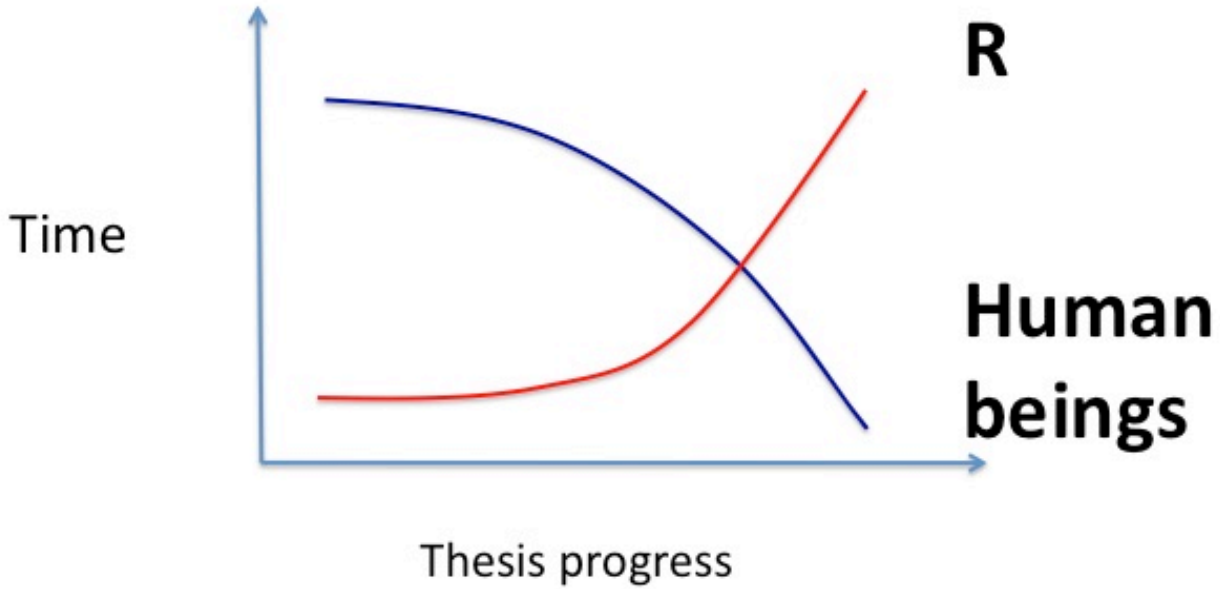
- I’m an evolutionary ecologist focusing on speciation.
- I’m not a statistician, but have basic training and learned from mistakes.
- I won’t be able to answer all your stats questions but will work with you to answer them.
- I started using R four years ago for statistical analyses. Used Statistica in the past: expensive and license renewal a drag. Now also use R for figures and programming.
- I am not an R expert. I have a head start on most of you.
- R often provides multiple ways to solve same problem: share them and new cool things.
- I have used R mainly on a Mac, but Windows version is virtually identical.
- My office hours: Tuesday 3-5 pm, Biodiv 242

Course aims:

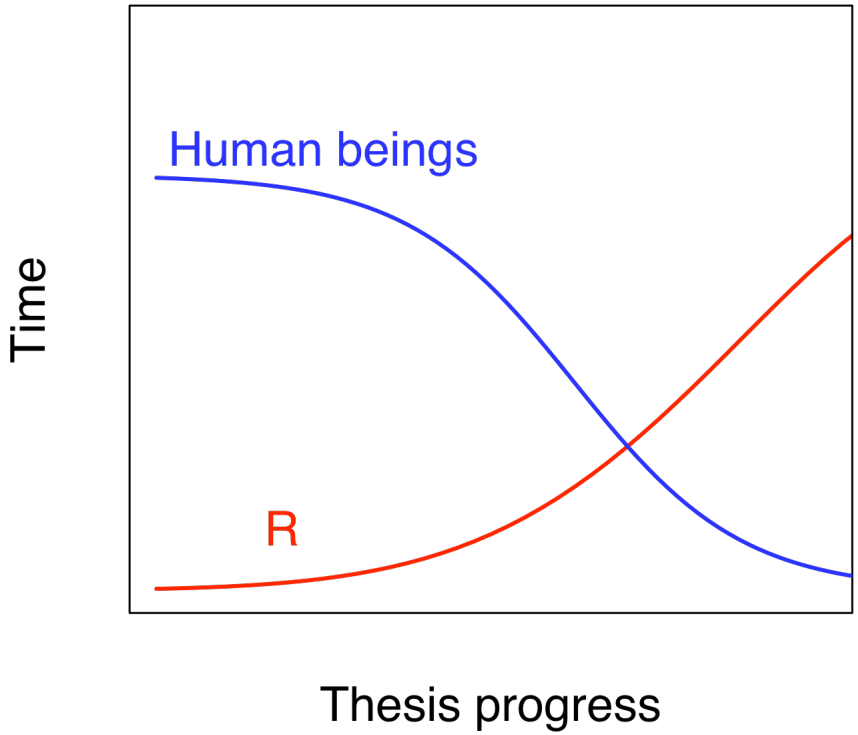
- To help prepare you for research by instruction in basic principles for designing good studies, gathering and organizing data, and properly analyzing those data.
- Introduce you to innovative new approaches increasingly used in biology to analyze data.
- Provide the computational tools to carry them out, namely, R.
- To provide broad coverage of current methods, rather than a deep foundation on few topics. It is expected that in your research you will need to study further those particular methods that turn out to be most appropriate.
- General linear models will be our framework.
- As far as possible this is a practical course: learn by doing.

Learning R has a notoriously steep learning curve.

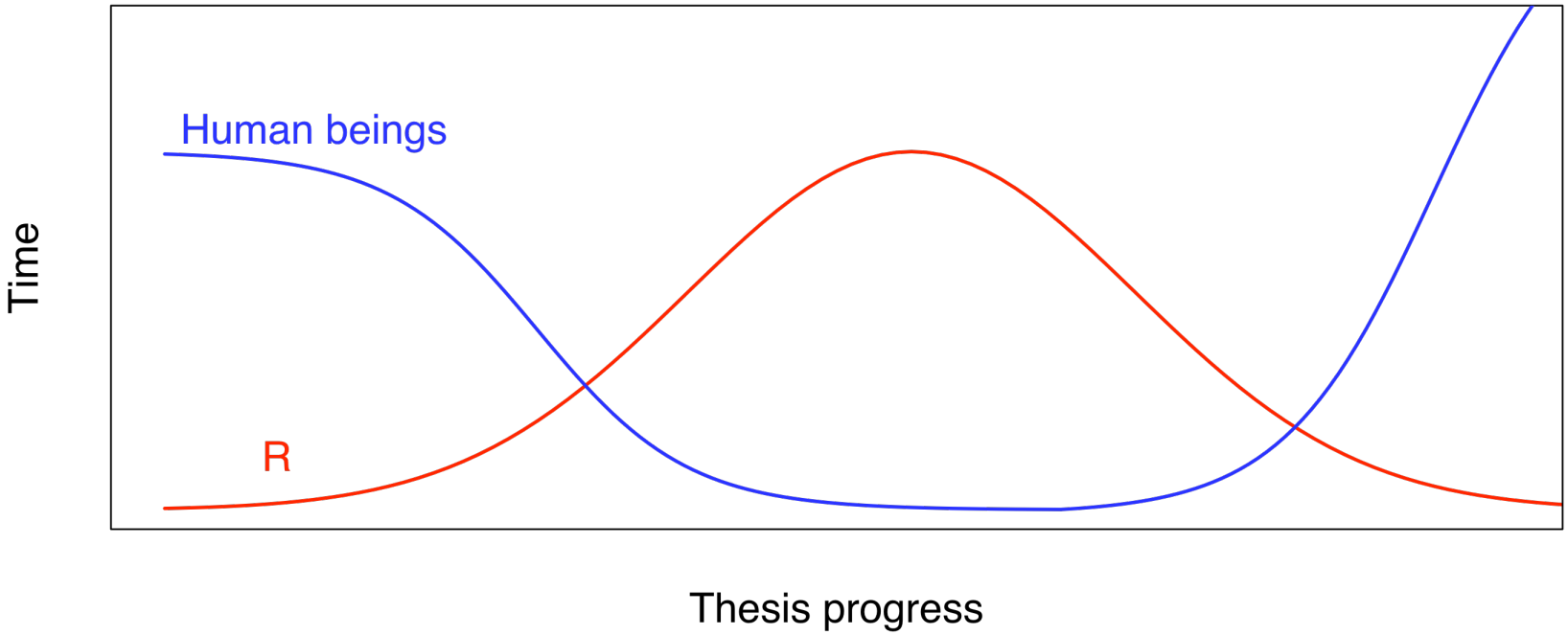
Ecology and Evolution 2011 retreat slide.



First of all: graph is not made in R.



Secondly, the time scale is wrong (increase `xlim`).



But, the great thing is that this grad student uses R and is on his way to heightened levels of social interactions while being R savvy.

Good things about R

1. Powerful and flexible.

2. Runs on all computer platforms.

3. Free!

4. New stuff always coming online – yet all in a common language
   gplots – graphics tools, including error bars
   mra – analysis of mark-recapture data
   nlme – linear mixed-effects models, generalized least squares
   qtl – QTL analysis
   shapes – geometric morphometrics
   RgoogleMaps – use static Google maps in R

5. Superb data management and manipulation capabilities (split columns, rearrange dates, transform data, analyze by subsets, and more).

Good things about R

6. Superb graphics capabilities. (Vectorized graphs are saved as pdf files, then can be imported to e.g. Illustrator or OpenOffice Draw for editing. Not PowerPoint!)
7. R uses scripts to execute commands rather than menus and a mouse. Keeping all steps in a text file allows you to go back years later and determine how you carried out the analysis, and how to repeat it.
8. In R it can be so easy to do otherwise difficult things (e.g., randomization tests)
9. You can write your own functions for specific needs.
10. Great programming tool.
11. Lots of help available. Someone has already solved your problem (e.g., Google search: error bars R)

http://www.google.ca/search?hl=en&q=error+bars+R&btnG=Google+Search&meta

Bad things about R (with partial solutions)

- R uses scripts to execute commands rather than menus and a mouse. Takes time to learn, a bit like a language. (This course will help get you started.)
- In R it can sometimes be so difficult to do otherwise simple things. (Improves with practice.)
- It is not a great spreadsheet. (Enter your data in a standard spreadsheet program and then save in a readable text format.)
- There are several kinds of data objects. (Stick with data frames whenever you can.)
- Some variation in command styles, e.g. older commands like `lsfit` compared with newer commands like `lm`.
- Quality control concerns? (Core programs are well-tested, whereas add-ons need checking.)
- Add-ons may disappear in future, e.g., xlsReadWrite. (for this particular example you should probably be storing your data in a non-proprietary format like .csv instead.)

[Dryad time with Heather Piwowar]

Data talk

Basic considerations for data entry, storage, organization. Valuable to consider as you are starting out on a career of data collection and interpretation.

Modified from:

Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer. 2009. Some simple guidelines for effective data management. Bulletin of the Ecological Society of America 90: 205-214.

1. Use a scripted program for analysis.
    a. E.g., R. (Others include MATLAB and SAS).
    b. Harder to start but reduces problems in future.
    c. Menu-based programs leave no record of analyses carried out, and you will forget.
    d. Scripts (commands) are written records of your analyses.
    e. Keep in a separate text file from your data.
    f. Surround commands with detailed comments on your choices and actions.

Data talk

2. Store data in a nonproprietary software format

    a. E.g., use comma delimited text files, .csv.

    b. Text files can always be read, whereas proprietary formats can become unavailable.

    c. You can still use spreadsheet programs to create the text files (Calc from www.openoffice.org, or Excel).

3. Store data in nonproprietary hardware formats

    a. The modern DVD format is just today's version of the 8-track tape.

    b. If possible, keep data on the internet, which probably won't die soon.

Data talk

4. Leave an uncorrected data file with all its bumps and warts

    a. Otherwise you might change something that you later discover was correct.

    b. Corrections directly to the data file go unrecorded.

    c. Make corrections instead within the scripted language (e.g., R) so you have a record, and can undo later if necessary.

    d. Keep comments in your script (command) file that explain reasons for corrections.

5. Use descriptive names for your data files

    a. Use names that are short but indicative of file contents.

    b. “SVancIsland_VegBiodiv_2007.csv” not “Veg_2007.csv”.

    c. Some applications have trouble importing file names with blank spaces.

Data talk

6. Include a “header” first line in data file with descriptive variable names
    a. Variable names should be terse but descriptive, without blank spaces or commas
    b. `read.csv` command in R assumes by default that first line is a header line.

Example:

Huey, R. B. and A. E. Dunham. 1987. Repeatability of locomotor performance in natural populations of the lizard *Sceloporus merriami*. Evolution 42: 1116-1120.

lizard	sprint_speed_1984
1	1.43
2	1.56
3	1.64
4	2.13
5	1.96

Data talk

7. Use plain ASCII text for names and data values

    a. Includes all letters of English alphabet (uppercase and lowercase), numbers, and many common punctuation marks (_ - * . are ok)

    b. Avoid commas because they separate fields in .csv format

    c. Avoid symbols (e.g., α © ﬁ)

8. When you add data to a database, add rows not columns

    a. Set up data files to maximize consistency of column content

    b. Use “long” format rather than “wide”

Data on sprint speed in *wide* format:

lizard	sprint_speed_1984	sprint_speed_1985
1	1.43	1.37
2	1.56	1.30
3	1.64	1.36
4	2.13	1.54
5	1.96	1.82
...		

Data on sprint speed in *long* format (preferred):

lizard	sprint_speed	year
1	1.43	1984
2	1.56	1984
3	1.64	1984
4	2.13	1984
5	1.96	1984
1	1.37	1985
2	1.30	1985
3	1.36	1985
4	1.54	1985
5	1.82	1985
...		

Data talk

9. A column of data should contain only one data type (i.e., either numerical or character, not both)
    a. R will interpret any column with even a single character as character or factor data

don't:

lizard	speed	year
1	1.43	1984
2	1.56	1984
3	1.64?	1984
4	2.13	1984

do:

lizard	speed	year	comment
1	1.43	1984	ok
2	1.56	1984	ok
3	1.64	1984	dubious
4	2.13	1984	ok

Data talk

10. Record full dates, using standardized formats
    a. For dates use YYYY-MM-DD (promoted by the International Organization of Standards). Other formats can be ambiguous.
    b. For datetime use YYYY-MM-DDThh:mm:ss
    c. Or use different columns for year, months, day, time.

11. Create a relational database
    a. Put separate information collected at different scales into different files.
    b. E.g., one file for SITE data (temperature, elevation). Another file for measurements of SPECIES collected within sites. Both files contain the SITE variable, allowing data to be merged as needed (using `match` command in R) .

Data talk

12. Maintain effective metadata (data about the data)

    a. Ten years from now you won't remember what the site looked like, which sample you dropped, or how you assigned a single value for "depth" of a pond.

    b. Record why you collected the data.

    c. Write down details of methods.

    d. Include names of all files associated with the study, definitions for data and treatment codes, missing value codes, definitions, unit of measurement for each variable.

    e. Consider using a metadata standard such as Ecological Metadata Language (EML).

Finding data (besides collecting your own)

- The lizard data was extracted from a scatter plot in the original article.
- There is no copyright on published data, which is useful when you need an example or are carrying out a meta-analysis.
- Graphics tool (shareware): http://www.datathief.org/
- Online data archives, e.g., Ecological Archives, http://esapubs.org/archive/, Genbank, Dryad (http://datadryad.org/repo)
- Permissions/conditions may be required to publish results from archives.

Quiz:

Take out a sheet of paper and state your understanding of the following terms (you may answer anonymously). Use just one sentence for each concept.

1. Sampling error of an estimate.
2. Bias of an estimate.
3. Random sample.
4. Probability.
5. Standard error.
6. 95% confidence interval.
7. *P*-value.

Brush up on some basic principles:

1. **Sampling error**

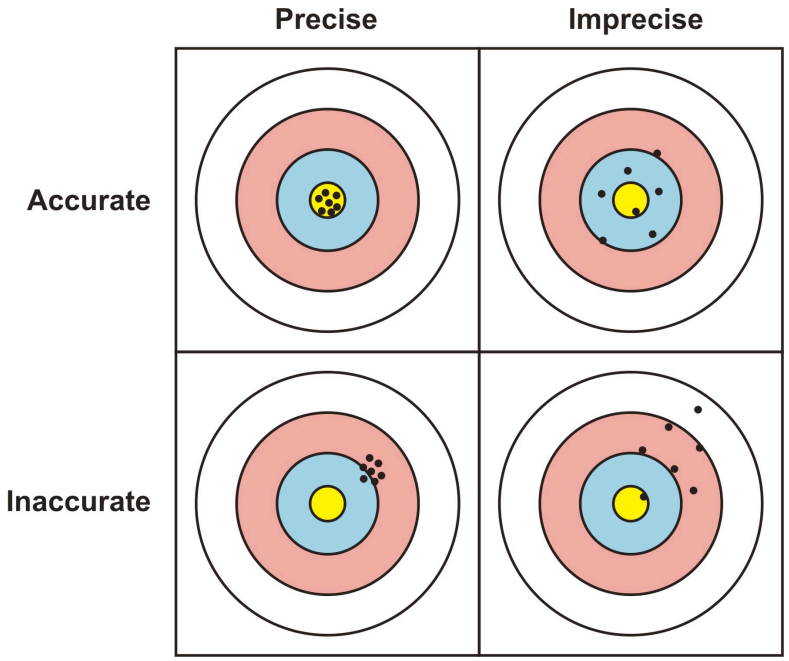
> **Sampling error** is the chance difference between an estimate and the population parameter being estimated.

2. **Bias**

> **Bias** is a systematic discrepancy between estimates and the true population characteristic.

Low sampling error = low spread = high precision

Low bias = centered on bull's eye = high accuracy



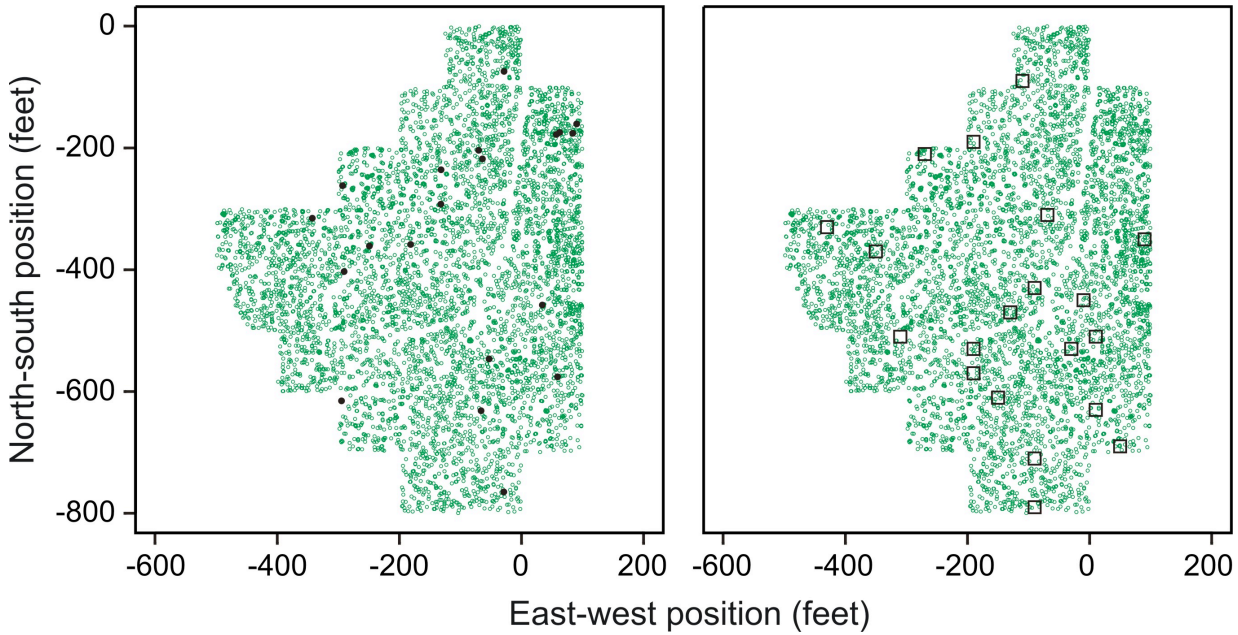
Analogy between estimation and target shooting. An accurate estimate is centered around the bull's-eye, whereas a precise estimate has low spread.

3. **Random sample**

In a **random sample**, each member of a population has an equal and independent chance of being selected.

Random sampling minimizes bias and makes it possible to measure the amount of sampling error.

Example of random sample



The locations of all 5699 trees present in the Prospect Hill Tract of Harvard Forest in 2001 (green circles). The black dots in the left panel are a random sample of 20 trees. The squares in the right panel are a random sample of 20 quadrats (each 20 feet on a side).

4. **Probability** (frequentist definition)

The **probability** of an event is the proportion of times the event would occur if we repeated a random trial over and over again under the same conditions.

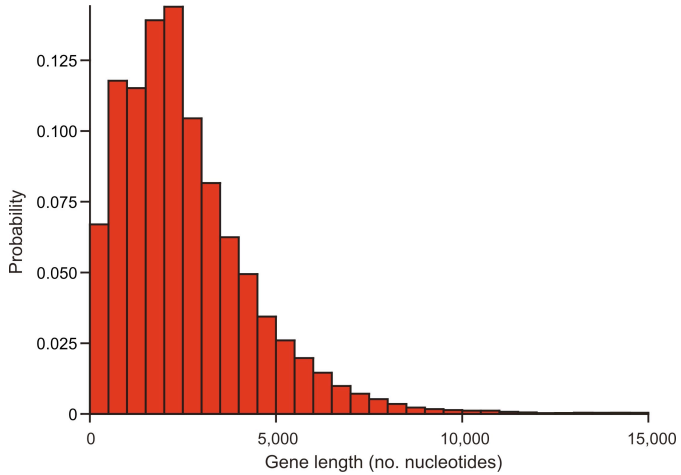
5. **Standard error**

The **standard error** of an estimate is the standard deviation of the estimate's sampling distribution.

The **sampling distribution** is the probability distribution of values for an estimate that we might obtain when we sample a population.

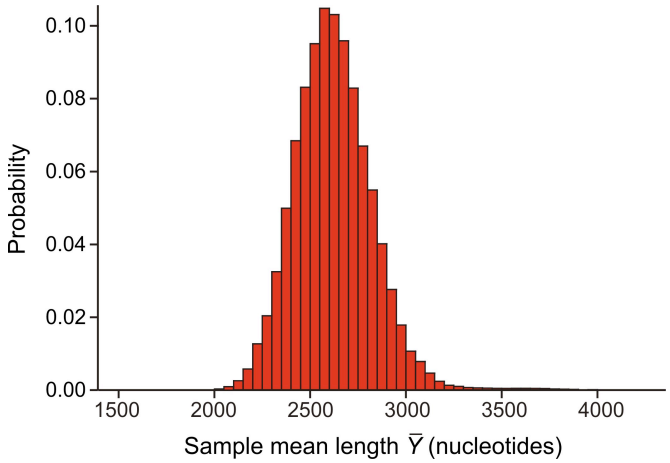
Population distribution: has a standard deviation measuring spread

Probability distribution of gene lengths
in the known human genome.



Sampling distribution: has a standard deviation (=standard error) measuring spread

The sampling distribution of mean gene length, $\overline{Y}$,
when *n* = 100. Note the change in scale from above.



Note: the mean of the sampling distribution is the true mean gene length (i.e., $\overline{Y}$ is unbiased)

Notice how bell-shaped the sampling distribution for $\overline{Y}$ is (illustration of central limit theorem).

6. **Confidence interval**

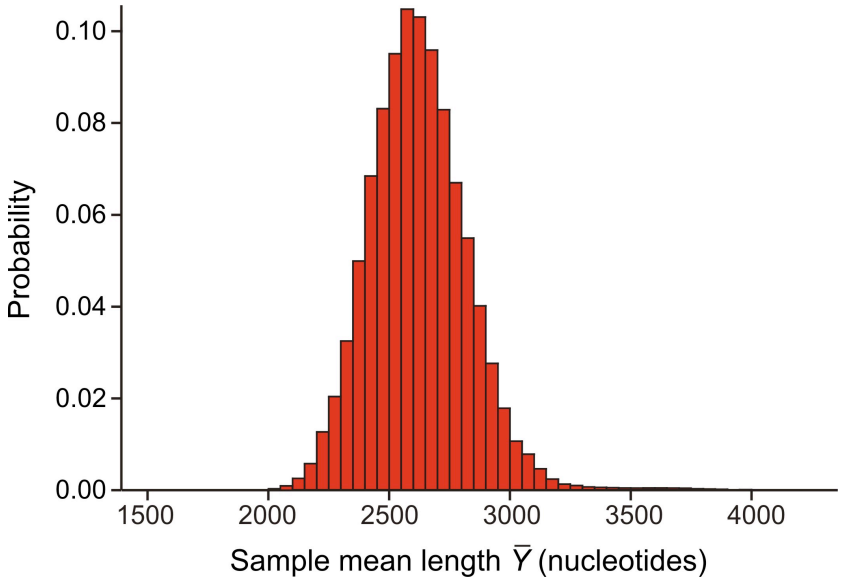
A **confidence interval** is a range of values surrounding the sample estimate that is likely to contain the population parameter.

95% CI's calculated from independent random samples will include the value of the parameter 19 times out of 20.

A rough approximation to the 95% confidence interval can be found from the sample estimate plus and minus two standard errors.

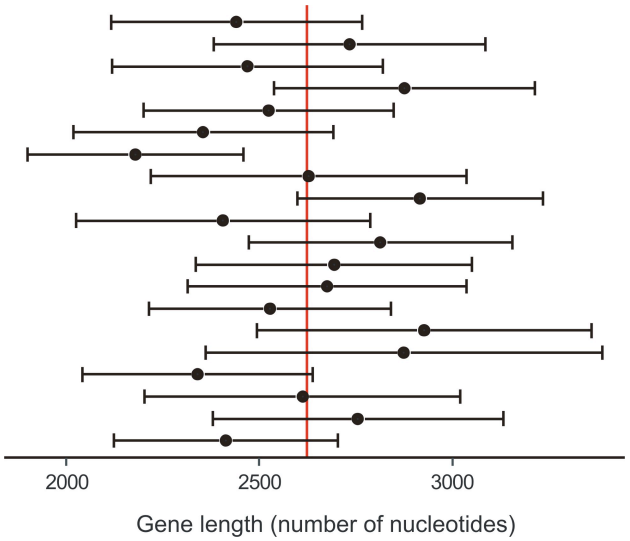
Why the 2SE rule works (approximately):

~ 95% of samples of $\overline{Y}$ lie within 2 standard deviations (SE's) of the mean



The sampling distribution of mean gene length, $\overline{Y}$, when n = 100.

Example of 95% confidence intervals



The 95% confidence intervals for the mean calculated from 20 separate random samples of n = 100 genes from the known human genome. Dots indicate the sample means. The vertical line (colored) represents the known population mean, μ = 2622.0. In this example, 19 of 20 intervals included the population mean, whereas one interval did not.

7. ***P*-value**

The ***P*-value** is the probability of obtaining the data (or data showing as great or greater difference from the null-hypothesis) if the null hypothesis were true.

The P-value is calculated from the tails of the NULL sampling distribution, the probability distribution of values for test statistic (e.g., *t*, *F*) that we might obtain when we sample a population when the null hypothesis is true.

Brush up on some basic principles:

If these statements are different from the ones that you wrote on your answers to the quiz, then consider reading Chapters 4 and 6 of Whitlock and Schluter (2009) to refresh.

Discussion paper:

Wainer (1984) How to display data badly

Placed online under the assignments tab at the Biol 548 web site.

Need two presenters and session leaders for next week: 15-20 minute presentation