Simulation

Hadley Wickham

Aims

- Learn how to simulate data to:
  - Test your statistical intuition
  - Perform power calculation
  - Experiment with a new technique on known data
- Learn how to use functions to reduce duplication

Functions

- Let us avoid repetition

```
functionname <- function(argument1,...) {
   # do stuff here
}
```

Building up a function

- Start simple
- Do it outside of the function
- Test as you go
- Give it a good name

Next task

- We know (hopefully) that a t test works best on normally distributed data
- How can we test that?

Your turn

- Figure out how to do a **t.test** in R
- Figure out how to extract the p-value from that object (use **str** and your subsetting skills)
- Write a function to generate two vectors of n random normals, compare them with a t.test and return the p-value

Your turn

- Repeat several thousand times and draw a histogram for various values of n
- Try varying the parameters of the two normals. What happens when you vary the mean? What happens when you vary the standard deviation?
- What happens if you use non-normal data? Eg. uniform, or poisson data

Another exploration

- How does our sample estimate compare to the true unknown
- eg., when calculating the mean of a sample of random normals, how many do we need to draw to be reasonably certain we got the right value?

What do we want to see?

- A plot of the different estimates, vs. number of sample points?
- So we need a data.frame with columns n, and sample mean (and sample sd. as well)
- How can we do this?
- Can’t just use replicate

New function

- **sapply**
- Takes first argument, and calls second argument one at a time
- sapply(1:10, sum) vs sum(1:10)
- sapply(1:10, function(n) mean(rnorm(n)))

Create the data

- n <- rep(seq(1, 1000, by=10), each=10)
- mean <- sapply(n, function(x) mean(rnorm(x)))
- qplot(n, mean)

Your turn

- Explore what happens when you change the standard deviations
- What about when you estimate the standard deviation?
- What about other distributions? eg. poisson
- Try adding smoothed lines to the data (see qplot chapter)

Homework

- Write up an exploration of the sampling distribution of an estimate of a distribution (eg. mean or sd of normal) in the style of the cheatsheet (but with more graphics)