

# Simulation

- Probability distributions
- Repetition, functions
- Bootstrap, permutation

## Basics of simulation

- Want to:
  - generate random numbers from known distribution
  - want to repeat the simulation multiple times

# Generating samples from different distributions

- **r**unif (uniform), **r**pois (poisson), **r**norm (normal), **r**binom (binomial), **r**gamma (gamma), **r**beta (beta)
- First argument for all is  $n$ , number of samples to generate
- Then parameters of the distribution (always check that the distribution is parameterized the way you expect)

## Your turn

### Generate

- 100 draws from  $N(0, 1)$
- 50 draws from  $N(10, 5)$
- 1000 draws from  $\text{Poisson}(50)$
- 100 draws from  $\text{Beta}(0.1, 0.1)$
- 30 draws from  $\text{Uniform}(0, 10)$

and PLOT them!

# Repetition

- Use the **replicate** function
- `replicate(n, expression)`

## Your turn

- Plot histogram of:
  - 1000 x mean of 2 Unif(0, 10)
  - 1000 x mean of 10 Unif(0, 10)
  - 1000 x mean of 100 Unif(0, 10)
- What do these examples show?

# Functions

- Let us avoid repetition

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

# Functions

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

# Exploring tests

- Two sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s/\sqrt{n}}$$

$$H_o : \mu_1 = \mu_2 \quad \text{vs} \quad H_a : \mu_1 \neq \mu_2$$

- How can we check how often it gets it wrong?

## Your turn

- Figure out how to do a two sample **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

# Exploring tests

- Compare samples simulated from
  - the same mean
  - different means
  - different, but close means
  - different distributions, same means
- How can we check how often it gets it wrong?

# Exploring tests

- Two ways to get it wrong:
  - Reject the null hypothesis, when its true - significance level
  - Fail to reject the null hypothesis, when its not true - power, hard to control

# Types of tests

t.test, chisq.test, cor.test, fisher.test,  
binom.test, wilcox.test, kruskal.test,  
prop.test, ...

# Bootstrap

- Take  $m$  samples with replacement, calculate statistic
- Understand the variability of statistic about population parameter, but treating the sample as the population

# Bootstrap

- Galaxies data
  - what's the distribution of the median?
  - If we knew the density function then the asymptotic distribution of the median of the sample would be normal centered at the true median, with a variance depending on the density function.

## Your turn

- Generate 1000 bootstrap samples of the diamond price, from the diamonds data
- Compute the standard deviation in the samples
- What does the distribution look like?

# Permutation

- Another way to test hypotheses
- Under the assumption that there is nothing of interest happening, what would we expect to see
- Explore the t-test, again.....

# Permutation

- For two sample test
  - shuffle the labels of each value
  - calculate p-value

# Permutation

- Graphically exploring association
  - Plot  $y$  vs  $x$
  - Shuffle  $y$ , replot
- Marginal distributions are unchanged, only the dependence between variables is changed.

## Your turn

- Use your subsetting skills to extract only the ideal and premium diamonds from the diamonds data.
- Use permutation to determine the p-value for the test that mean price is the same for ideal and premium diamonds.