Stat 145, Wed 10-Mar-2021 -- Wed 10-Mar-2021
Biostatistics
Spring 2021

Wednesday, March 10th 2021

Wk 6, We
Topic:: P-values
Topic:: Randomization distributions

From Hesterberg & Chihara:
  This is the core idea of statistical significance or classical hypothesis
  testing--to calculate how often pure random chance would give an effect
  as large as that observed in the data, in the absence of any real effect.
  If that probability is small enough, we conclude that the data provide
  convincing evidence of a real effect.

Randomization distributions: first look

Various distributions:

- A **sampling distribution** gives the values a sample statistic can take, and their (relative) frequencies. It depends on
    - the population (from which samples are taken)
    - the statistic in question (just what we are computing from the sample: perhaps $\bar{x}, \hat{p}$, etc.)
    - the size n of samples

- A **null distribution** arises in the context of hypothesis testing. It is
    - like a sampling distribution: it relies on computing a particular sample statistic, from random samples of a particular size n.
    - even more difficult to pin down than a sampling distribution, as it must obey the null hypothesis.

- A **randomization distribution** uses simulated draws from the original sample to *approximate* the null distribution.

Some scenarios

Board overview is like that for generating a bootstrap distribution. We

1. draw a new sample, called a **randomization sample**, using data in the original sample, and

2. compute a **randomization statistic** from the randomization sample.

Repeating Steps 1 and 2 many (several thousand?) times, and plotting the randomization statistics that arise from Step 2, produces a **randomization distribution**.

- **Case**: univariate binary categorical data

    Our null hypothesis: $\mathbf{H}_0: p = p_0$, (p_0 is some number).

$H_0: \quad p = 1/3 \qquad$ (rock paper scissors)

See rock: Success
Don't see rock: failure

To look at null dist., decide
what sample size?
what sample statistic?

From StatKey saw in randomization dist numbers as high or higher than 25 occur 1% = .01

2-sided
P-value: $2(0.01) = 0.02$

$\underline{\qquad\qquad\qquad\qquad}$
$\qquad\qquad 2$

The relative freq. of obtaining a result as extreme (or more so)

as in our sample when null hypothesis is true.

- **Case**: bivariate data from independent samples, one binary categorical variable (explanatory; indicates `group`), other variable is quantitative (response)

  Our null hypothesis: $\mathbf{H}_0: \mu_1 - \mu_2 = 0$, (described as "no effect").

  Example[1] Six mice are divided randomly into two groups, with 3 given a drug and the other 3 acting as a control group. Afterward, all 6 mice are timed as they navigate a maze. Let μ_D, μ_C represent the mean finishing time for two populations, mice under the influence of the drug, and mice not so influenced. Suppose the 3 "drug" mice finish in 30, 25 and 18 seconds, while the 3 "control" mice finish in 20, 21, and 22 seconds.

```r
mice <- data.frame(time=c(30,25,18,20,21,22), grp = c(rep("Drug",3),rep("Control",3)))
mice
```

```
  time     grp
1   30    Drug
2   25    Drug
3   18    Drug
4   20 Control
5   21 Control
6   22 Control
```

$$\bar{x}_D = \frac{1}{3}(30 + 25 + 18) = 24.\overline{3}$$

$$\bar{x}_C = \frac{1}{3}(20 + 21 + 22) = 21$$

parameters

μ_D

μ_C

$H_0: \mu_D - \mu_C = 0$

We look at sample means and their difference:

```r
mean(time ~ grp, data=mice)    # shows the sample means for the 2 groups
```

```
 Control      Drug
21.00000 24.33333
```

```r
diffmean(time ~ grp, data=mice)    # finds the difference in sample means
```

```
diffmean
3.333333
```

Q: What is this difference of means a point estimate for?

In the picture that follows, the left side shows dot plots of `time` by treatment group as in the actual sampled data. The right side shows dot plots of `time` by group for a randomization sample.

---

[1]Taken from Chapter 3 of "Mathematical Statistics With Resampling and R", by Chihara and Hesterberg.

Original Sample

Control

Drug

20   24   28

time

Randomization Sample

Control

Drug

20   24   28

time
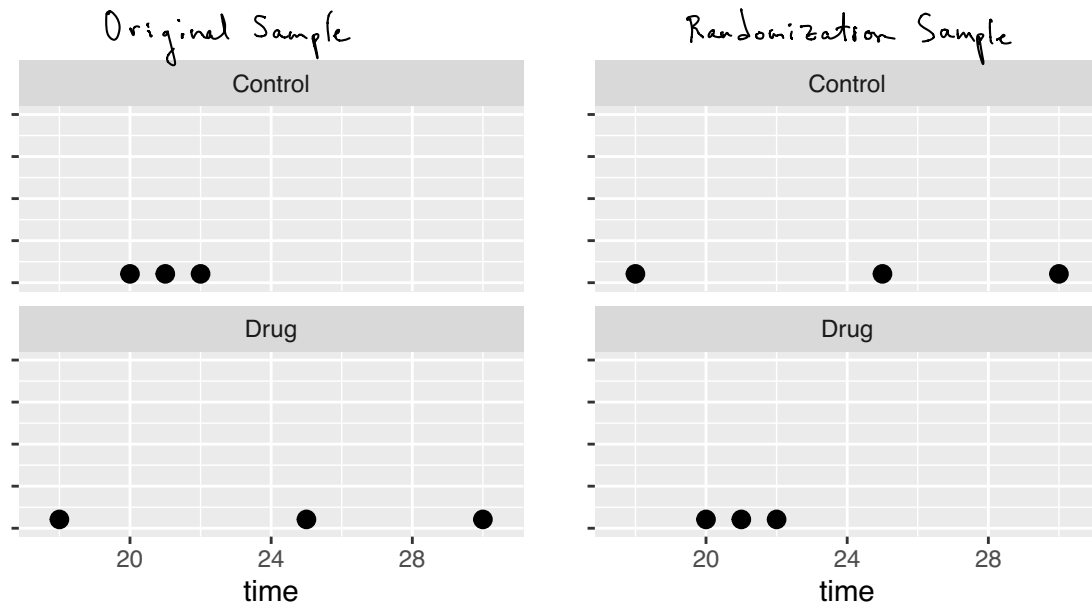
- **Case**: bivariate data, both (explanatory and response) variables are quantitative, and a linear assocation appears plausible

  Our null hypothesis: $\mathbf{H}_0\colon \rho = 0$.