Chapter 3 and 4 Summary

Paul Luo, Tina Wang, Jackson Cong

Investigation 3.1

- With two samples for a binary categorical variable, two-way table (2x2 table of counts) can be used.
- In R, we can input matrix(c(x, x, x, x), ncol = 2) code to construct two-way table.
- **Conditional proportion**: Comparing proportions of two groups instead of the overall proportion.
  - Eg. $\hat{p}_1 - \hat{p}_2$

    Two additional "rules for random variables" is, for two random variables X and Y,
    - $E(X \pm Y) = E(X) \pm E(Y)$
    - $Var(X \pm Y) = Var(X) + Var(Y)$, as long as X and Y are independent
  - According to the formula, Variability is easily calculated from $\hat{p}_1$ and $_2$2.
- **Central Limit Theorem for the difference in two sample proportions**
  - Two independent samples n_1 and n_2 from large populations, the distribution of difference of $\hat{p}_1 - \hat{p}_2$ is approximately normal. The mean equals to $\pi_1 - \pi_2$. Standard deviation equals to SD($\hat{p}_1 - \hat{p}_2$).
  - We consider the sample size is large enough to make a normal model if: $n_1\pi_1 \geq 5$, $n_1(1-\pi_1) \geq 5$, $n_2\pi_2 \geq 5$, $n_2(1-\pi_2) \geq 5$
- Two proportion t-test in r: iscamtwopropztest($observed_1$, n_1, $observed_2$, n_2, hypothesized difference, alt = ("less", "greater", or "two.sided"), conf.level = 0.95)
- Confidence interval formula of two proportion

  Approximate (100 × *C*)% Confidence interval for $\pi_1 - \pi_2$:

  1. *Two-sample z-interval*:

     An approximate (100 × *C*)% interval: $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

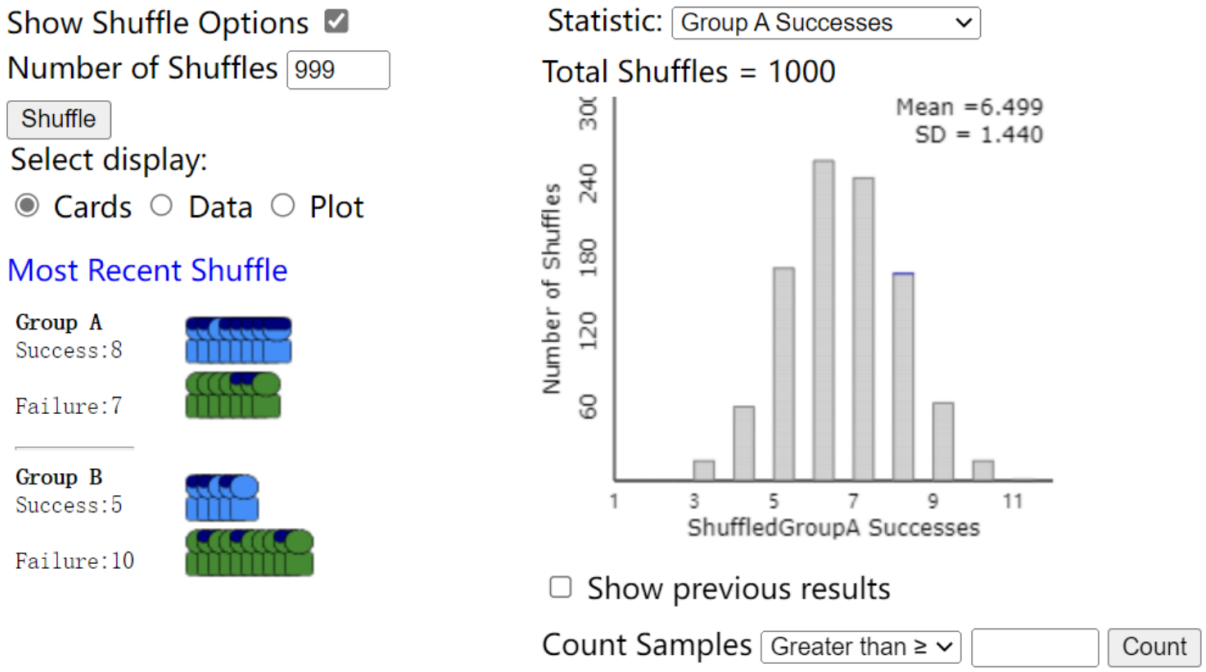
     where $-z^*$ is the $100 \times (1 - C)/2^{th}$ percentile from the standard normal distribution.

     This method is considered valid when $n_1\hat{p}_1 \geq 5$, $n_1(1-\hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$, and $n_2(1-\hat{p}_2) \geq 5$.
- **The table from The Statistical Sleuth**
  - Random sampling enhances the external validity or generalizability of your results, because it helps ensure that your sample is unbiased and representative of the whole population.
  - Random sampling is a way of selecting members of a population to be included in the study. Random assignment is a way of sorting the samples into control and experimental groups.
  - **Random sampling** + **Random assignment**: Random sample is selected in population, and units are distributed randomly into groups.
  - **Random sampling** + **Not Random assignment**: Random samples from populations, the inferences can be generalized to populations.
  - **Not Random sampling** + **Random assignment**: Units of sample are randomly assigned, which can draw cause and effect conclusions.
  - **Not Random sampling** + **Not Random assignment**: Available units and assigned into irregular groups, which results in confounding variables and potential sampling bias.

Investigation 3.5

- **P-value with random assignment (Dolphin therapy group and control group)**

- With Dolphin Study Applet, 1000 numbers of shuffles, 13 blue and 17 green. Mean = 6.499



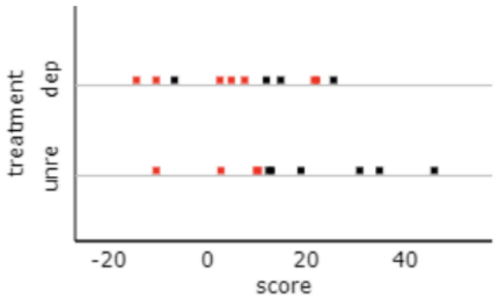
Investigation 4.2

- **Comparing two population means**: Choose an SRS of size n_1 from Population 1 with mean μ_1 and standard deviation σ_1 and an independent SRS of size n_2 from Population 2 with mean μ_2 and standard deviation σ_2
  - The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is **Normal** if both population distributions are Normal.
  - The mean of the sampling distribution of is $\bar{x}_1 - \bar{x}_2$ is $\mu_{\bar{x}_1-\bar{x}_2}$ = μ_1 - μ_2
  - The standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is $SD_{\bar{x}_1-\bar{x}_2}$ = $\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$
  - C% confidence interval for $\mu_1 - \mu_2$ is $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$
  - where t^* is the critical value with C% of the area between $-t^*$ and t^* for the t distribution with degrees of freedom (n-1)
  - Valid condition: the sample distributions are reasonably symmetric or the sample sizes are both at least 20
  - In R, two sample t-test: iscamtwosamplet(x_1, sd_1, n_1, x_2, sd_2, n_2, alt= “less”, “greater”, or “two.sided”, conf.level) # Investigation 4.4
- **Simulating a Randomization Test for a Quantitative Response**

- Comparing Groups (Quantitative) applet:

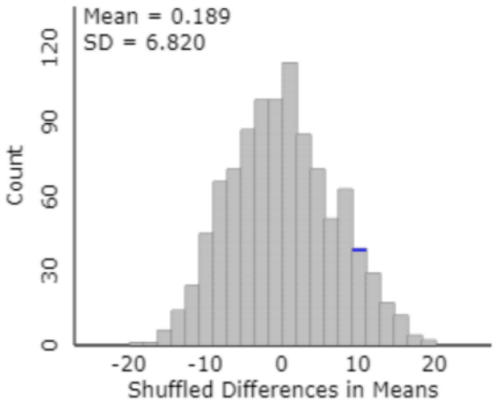
Show Shuffle Options: ☑
Number of Shuffles: 999
Hypothesized μ_2-μ_1: 0
Shuffle Responses
Select display:
○ Data ◉ Plot

Most Recent Shuffle



☑ Show Original Colors

Statistic: Difference in means
Total Shuffles = 1000



Count Samples Greater Than ≥ Count

Shuffled summary statistics:

	n	Mean	SD
dep	11	6.90	13.57
unre	10	16.52	16.45
pooled	21	11.48	15.00

Shuffled diff = 9.62

- The dotplots in picture shows how two groups of people distributed after 1000 shuffles. (Two groups represent deprived and unrestricted sleeping of people)
- For exact P-value, thinking about all possible random assignments of units distributed into groups, and determining difference in means or medians. Then, counting how many of the simulated statistics or as or more extreme as the observed.
- In R:

(with Sleep Deprivation file loaded and attached)
In the R Console, set the number of repetitions and initialize the vector that will store the differences.

```
> I = 10000                ! use upper case
> diff = 0                 ! initializes the vector
```

Then create a loop that will mix up the order of the response variable values and calculate a difference in means with each new random assignment:

```
>   for (i in 1:I){                    ! the loop for I iterations
+     rerandom = sample(improvement)   ! mixes up response values
+     boxplot(rerandom~condition,
+     horizontal=TRUE)                 match up the explanatory variable
+     diff[i]=mean(rerandom[1:10]) -   group sizes
      mean(rerandom[11:21])
>   }
>   Then you can examine a histogram of diff and compute the
>   empirical p-value as before.
    hist(diff); abline(v =15.92,     ! adds vertical line
    col=2)                           ! p-value
    sum(diff >= 15.92)/I
```