STAT 339: Statistical Theory

Introduction to Statistical Theory

Anthony Scotina

Some "Fun" Stats Games

Game 1: (name withheld)

> I'm going to private message each of you some code to run in R!
>
> Run the code and **private message** me your *output*.

Game 1: (Spies versus Agents)

Surprise! I choose *N* of us at random to be **spies**!

- The remaining *10 - N* of us are **agents**.
- Spies and Agents had different success probabilities (i.e., probability of "1")!

Agent success probability: **2/3**

Spy success probability: **1/3**

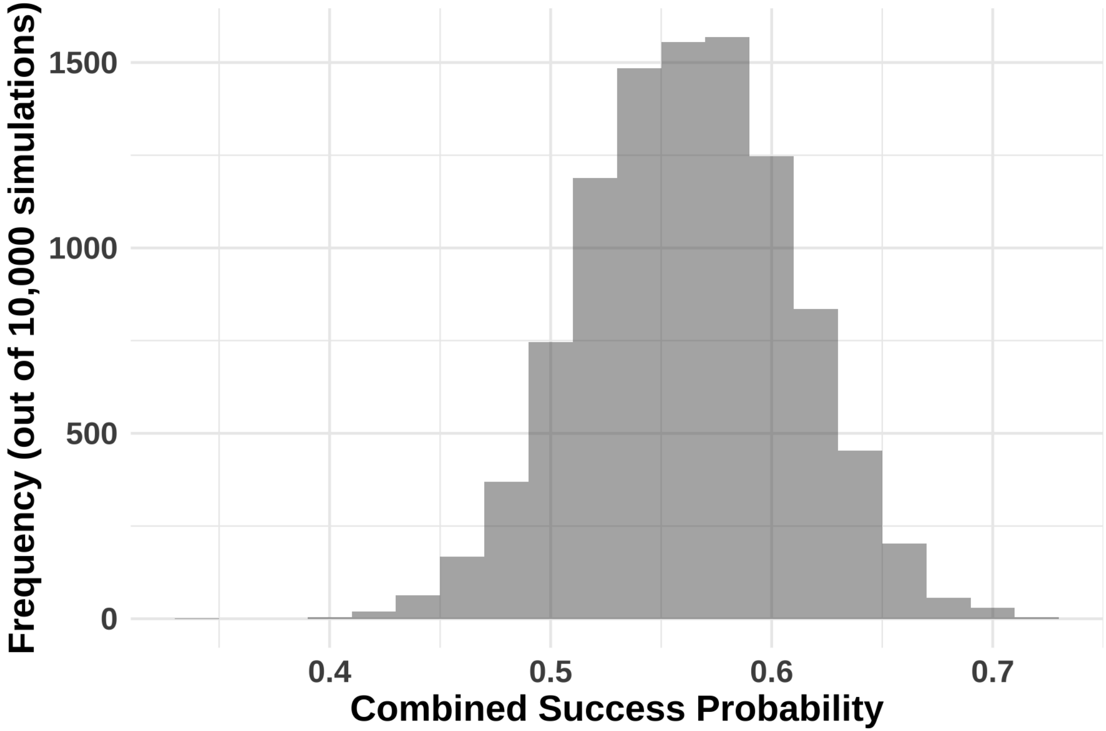
If we *pool* all 10 *results*, will the *combined* success probability be...

- 2/3? 🤔
- 1/3? 🧐

Game 1: (Spies versus Agents)

It turns out that the *combined* success probability is $(2-p)/3$, where p is the **proportion of spies**.

- How can we use our **DATA** to find p?



Game 2: Cell Phone Battery Life

Suppose we have a random sample of $n = 10$ cell phones, and we record their **battery life** (in minutes), $Y_1, Y_2, \ldots, Y_{10}$.

- We assume that the sample comes from an **Exponential** distribution with density function

  $$f(y \mid \theta) = \frac{1}{\theta} e^{-y/\theta}, \quad y > 0,$$

  where θ is **unknown**.

- **Note**: If $Y \sim Exp(\theta)$, then the *expected value* $E(Y) = \theta$.

Using the information provided to your group (I'll message you), try to **estimate** θ.

- **Group 1**: *The Raw Data* $\{393, 21, 211, 514, 73, 108, 116, 708, 387, 241\}$
- **Group 2**: *Sample Minimum* $Y_{(1)} = 21$
- **Group 3**: *Sample Mean* $\bar{Y} = 277.2$

Random Variables and Statistics

(Some Probability Review)

1. Random Variables

A **random variable (RV)** is a function from the sample space S to the real numbers, $\mathbb{R}$.

- Random variables are typically denoted by *capital letters*, for example, Y.
- Observed values of random variables are typically denoted by *lower-case letters*, for example, y.

Discrete RVs: Numerical variables that can take *whole, non-negative numbers*

- *Number* of calls to a call center (0, 1, 2, …)

Continuous RVs: Numerical variables that can take an *infinite range of numbers*

- *Lengths* of calls to a call center: $c \in [0, \infty)$

2. Probability Functions

Probability functions are *theoretical models for some frequency distribution of a population.*

- For example, we might choose to model cell phone battery life times, $Y_1, Y_2, \ldots, Y_n$ with an *Exponential* distribution that has *scale parameter,* θ:
  - $Y_i \sim Exponential(\theta)$
- Under this model, Y_i has probability *density* function (PDF)

$$f(y_i \mid \theta) = \frac{1}{\theta} e^{-y_i/\theta}, \quad y > 0$$

A **valid** probability function has the following properties:

Continuous RVs

1. $f(y \mid \theta) \geq 0$ for all y
2. $\int_{-\infty}^{\infty} f(y)\, dy = 1$

Discrete RVs

1. $0 \leq p(Y = y \mid \theta) \leq 1$ for all y
2. $\sum_y p(Y = y \mid \theta) = 1$

3. Linear Combinations of RVs

Let $Y_1, Y_2, \dots, Y_n$ denote a *random sample* of **independent and identically distributed** observations with finite mean $E(Y_i) = \mu$ and variance $Var(Y_i) = \sigma^2$.

Then for a **linear combination**

$$U = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n,$$

- $E(U) = a_1\mu + a_2\mu + \cdots + a_n\mu = \sum_{i=1}^{n} a_i\mu$
- $Var(U) = a_1^2\sigma^2 + a_2^2\sigma^2 + \cdots + a_n^2\sigma^2 = \sum_{i=1}^{n} a_i^2\sigma^2$

> What does this say about the **sample mean**, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$?

4. Order Statistics

For random variables $Y_1, Y_2, \ldots, Y_n$, the **order statistics** are the random variables $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$, where:

- $Y_{(1)} = \min(Y_1, Y_2, \ldots, Y_n)$
- $Y_{(2)} =$ the second-smallest of $Y_1, Y_2, \ldots, Y_n$
- $\cdots$
- $Y_{(n-1)} =$ the second-largest of $Y_1, Y_2, \ldots, Y_n$
- $Y_{(n)} = \max(Y_1, Y_2, \ldots, Y_n)$

🚨 For now, we'll assume that the Y_i are *iid* and **continuous** RVs with:

- *Distribution function* $F(y) = P(Y \leq y)$
- *Density function* $f(y) = F'(Y)$

4. Order Statistics

PDF for Minimum

In STAT 338, we derived the PDF for

$$Y_{(1)} = \min(Y_1, Y_2, \ldots, Y_n)$$

by *first* finding the distribution function, $P(Y_{(1)} \leq y)$.

- Because $Y_{(1)}$ is the **minimum** of $Y_1, Y_2, \ldots, Y_n$, the event $(Y_{(1)} > y)$ occurs *if and only if* each of the $(Y_i > y)$ events occur for $i = 1, 2, \ldots, n$:

$$\begin{aligned} P(Y_{(1)} > y) &= P(Y_1 > y, Y_2 > y, \ldots, Y_n > y) \\ &= P(Y_1 > y)P(Y_2 > y) \cdots P(Y_n > y) \end{aligned}$$

It turns out that the PDF for $Y_{(1)}$ is given by

$$f(1)(y) = n[1 - F(y)]^{n-1} f(y)$$

4. Order Statistics

Exponential Minimum Order Statistic

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample of cell phone battery lifetimes from an $Exponential(\theta)$ distribution with PDF

$$f(y_i \mid \theta) = \frac{1}{\theta} e^{y_i/\theta}, \quad y_i > 0.$$

> Let's show that $Y_{(1)} \sim Exponential(\theta/n)$.

A Note on Notation

In STAT 338, we would often write probability functions as follows:

$$f(y_i) = \frac{1}{\theta} e^{-y_i/\theta}, \quad y > 0,$$

rather than using $f(y_i \mid \theta)$.

- In STAT 339, we'll often add the

  $$|\theta$$

  to the $f(y_i)$ to emphasize that the probability function depends explicitly on the value of the **parameter** θ.
  - A goal in this class will be to gather *insight* on the parameter, θ.

- Each *named* probability distribution (e.g., Exponential, Binomial, Normal, ...) has a different probability function with different parameter(s).
  - See the **probability distribution cheatsheet**!