M378K: October 16th, 2024.

Statistical Set-Up

Population. e.g., all the people in this class

$\boxed{\textbf{Sample.}}$ e.g., a committee of 4 students

We assume that we have a representative sample.

Use the same word for the results of measuring/polling from the same population with an unknown but common dist'n.

Def'n. A random sample of size n from distribution D is a random vector

$$(Y_1, Y_2, \dots, Y_n)$$

such that:

1. $Y_1, Y_2, \dots, Y_n$ are independent
2. every Y_i has the distribution D

Example. Consider 10 measurements $Y_1, Y_2, \dots, Y_{10}$. Care was taken so that the measurements are independent.

It is a standard model to assume that

Y_i are normally distributed w/ an unknown mean μ

Scenario #1. We know the standard deviation 0.1.

Then, $\boxed{Y_i \sim N(\mu, 0.1) \quad , i=1..10}$

Scenario #2. We don't know the standard deviation σ.

Then, $\boxed{Y_i \sim N(\mu, \sigma), \ i=1..10}$

Def'n. Any function of the random sample is called a **STATISTIC**.

A **point estimator** is any function (rule, procedure) of the sample $(Y_1, Y_2, \ldots, Y_n)$ (including known constants) w/ the purpose of estimating a model parameter.

IT CANNOT CONTAIN THE UNKNOWN PARAMETER THAT WE'RE ESTIMATING.

An **interval estimator** is a pair of point estimators.

e.g.,
- $\dfrac{\#\text{ of people of the 4 students who like ice cream}}{4}$, i.e.,

  a sample proportion

- In the normal example, we look @ the sample mean

$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

M378K Introduction to Mathematical Statistics
Problem Set #11
Order Statistics.

Problem 11.1. *An insurance company is handling claims from two categories of drivers: the good drivers and the bad drivers. The waiting time for the first claim from a* **good** *driver is modeled by an exponential random variable T_g with mean 6 (in years). The waiting time for the first claim from a* **bad** *driver is modeled by an exponential random variable T_b with mean 3 (in years). We assume that the random variables T_g and T_b are independent.*

What is the distribution of the waiting time T until the first claim occurs (regardless of the type of driver this claim was filed by)?

$\rightarrow$: $T = \min(T_g, T_b)$ $\quad S_T = [0, +\infty)$

$$F_T(t) = \mathbb{P}[T \le t] = \mathbb{P}[\min(T_g, T_b) \le t] = 1 - \mathbb{P}[\min(T_g, T_b) > t]$$

$$= 1 - \mathbb{P}[T_g > t, T_b > t] = 1 - \mathbb{P}[T_g > t] \cdot \mathbb{P}[T_b > t]$$

$$= 1 - e^{-t/\tau_g} \cdot e^{-t/\tau_b} = 1 - e^{-t(1/\tau_g + 1/\tau_b)}$$

$T \sim E(\tau)$ w/ $\tau = 1/(1/\tau_g + 1/\tau_b) = 1/\tfrac{1}{2} = 2$ $\square$

Definition 11.1. *Let $Y_1, \dots, Y_n$ be a* **random sample***. The random sample ordered in an increasing order is called an* order statistic *and denoted by*

$$Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}.$$

Question Write $Y_{(1)}$ as a function of $Y_1, Y_2, \dots, Y_n$.

Question Write $Y_{(n)}$ as a function of $Y_1, Y_2, \dots, Y_n$.