Sections 7.1-7.3 of the textbook

Sampling Distributions and the Central Limit Theorem

Recall: A (univariate) random variable is defined to be a function from a sample space S into the real numbers, $\mathbb{R}$. We've talked a lot about distributions of random variables (including multivariate random variables), and we've used these distributions to find probabilities and other quantities like expected values that can be used to make decisions.

Often, we are not interested in the random variable(s) themselves, but instead in functions of the random variable(s), such as the **sample mean**.

Specifically, we can focus on functions of the variables $Y_1, Y_2, \dots, Y_n$ **observed in a random sample** selected from the population of interest.

In such scenarios, we need the probability distribution of these functions of random variables in order to make inferences about population parameters, and, in turn, make decisions.

Y_i can be Discrete or Continuous.

Definition: (Random Sample)
$Y_1, Y_2, \dots, Y_n$ are a **random sample** of size n from a population with distribution (CDF) $F(y)$ if $Y_1, Y_2, \dots, Y_n$ are:

- **independent** ⇒ Unrelated
- **identically distributed** (have the same probability distribution)

This is also called an **iid (independent and identically distributed)** sequence.

Example 1: Let Y be the number of calls received per hour at a customer service center. The observed values of Y in randomly selected 8 hours are given below:

$y_1 = 3, \quad y_2 = 1, \quad y_3 = 2, \quad y_4 = 1, \quad y_5 = 10, \quad y_6 = 3, \quad y_7 = 0, \quad y_8 = 12$

Here, Y_i is the number of calls received per hour and it can take any value from $\{0, 1, 2, \dots\}$. Y_i has a Poisson distribution with parameter λ which is the average number of calls per hour.
Since these 8 hours are randomly selected, Y_i are independent here.

Here, $n = 8$. And the Poisson distribution is the population distribution in this example.
Typically, the value(s) of the parameter(s) is unknown. That is, λ is unknown. Our goal is to use the random sample and estimate the unknown parameter(s).

Often, we are not interested in the random variables themselves, but instead in functions of the random variables.

Definition 7.1: (Statistic)
If $Y_1, Y_2, \ldots, Y_n$ are a **random sample**, then any function of $Y_1, Y_2, \ldots, Y_n$ and only of $Y_1, Y_2, \ldots, Y_n$ is a **statistic**.
(Note: A **statistic** cannot depend on any unknown population parameter(s) and can depend on known constants.) such as $\mu, \sigma, \alpha, \beta, \lambda$

Examples: Let $(Y_1, Y_2, \ldots, Y_n)$ (known) be a random sample.

① Maximum of random sample is a statistic because there are no unknowns and it can be computed from the sample.

② Mean of sample $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a statistic because there are no unknowns

③ Variance of Sample $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is a statistic because there are no unknown.

④ $\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2$ is NOT a statistic because σ is an unknown parameter.

Different samples will have different values (NOT exactly same but close) for the **same statistic**.

Therefore, a **statistic** is a random variable.

Definition: (Sampling Distribution)
A **statistic** is also a random variable and has its own distribution, which depends on

- the distribution of Y_i
- the value of n

The probability distribution of the statistic is called the **sampling distribution**.

To see the properties of sampling distribution of $\bar{Y}$ and S^2 using simulation, go to http://onlinestatbook.com/stat_sim/sampling_dist/ and click on **Begin** button. And then change the name of the population distribution and sample sizes.

independent and identically distribution Mean

Example 2: Let $Y_1, Y_2, \ldots, Y_n$ be an **iid** random sample from a population with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ ← variance

(Note that a specific population distribution has not been provided!).

⇒ Population distribution is unknown.

Two commonly used statistics are the sample mean, $\bar{Y}$, and the sample variance, S^2.

- The **sample mean**,

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$$ is Linear Function of $Y_1, Y_2, \ldots Y_n$

- The **sample variance**,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right]$$

Even though a specific population distribution wasn't provided, we can still find the following characteristics of the sampling distributions of $\bar{Y}$ and S^2.

Find the **mean and variance** of the sampling distribution of $\bar{Y}$. $E(\bar{Y}) = ?$ $Var(\bar{Y}) = ?$

Use Expected value and Variance of a Linear function.

$$\text{Mean} = E(\bar{Y}) = E\left[\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)\right]$$

$$= \frac{1}{n}\left[\underbrace{E(Y_1)}_{\mu} + \underbrace{E(Y_2)}_{\mu} + \cdots + \underbrace{E(Y_n)}_{\mu}\right] = \frac{1}{n}(n\mu) = \mu$$

(n terms)

$\mu_{\bar{Y}} = \text{Mean of } \bar{Y} = \mu$.

⇒ Sampling distribution of $\bar{Y}$ is centered at μ which is the mean of the population distribution

$$Var(\bar{Y}) = Var\left(\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)\right)$$

$$= \frac{1}{n^2} Var(Y_1 + Y_2 + \cdots + Y_n)$$

$$= \frac{1}{n^2}\left[\underbrace{Var(Y_1)}_{\sigma^2} + \underbrace{Var(Y_2)}_{\sigma^2} + \cdots + \underbrace{Var(Y_n)}_{\sigma^2}\right]$$

(n terms)

because Y_i are independent, covariances are zeros

$$= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

$\sigma_{\bar{Y}}^2$ = Variance of $\bar{Y}$ = $\frac{\sigma^2}{n}$ < σ^2

⇒ Variance of sampling distribution of $\bar{Y}$ is smaller than variance of population distribution and it goes down (decreases) when the sample size(n) goes up (increase).

(*) These properties of sample mean are always true regardless of population distribution.

Sampling from the Normal Distribution (Section 7.2)

Many phenomena observed in the real world (like sales prices of homes in a particular area, height, heart rate, corn yield) are closely approximated by a normal distribution.

In these situations, it is reasonable to assume $Y_1, Y_2, \ldots, Y_n$ are a random sample of normally distributed random variables. If this is true, then $\bar{X}$ and S^2 have some special properties.

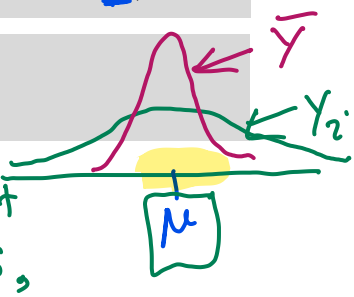
Theorem 7.1:

Let $Y_1, Y_2, \ldots, Y_n$ be a **random** sample of size n from a **normal** distribution with **mean μ** and **variance σ^2**. Then

"Y bar" → $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N\left(\text{Mean} = \mu_{\bar{Y}} = \mu, \quad \text{Variance} = \sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}\right)$ ← Normal

Standard deviation = $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

$z = \frac{\text{Variable} - \text{Mean}}{\text{Standard Deviation}} \Rightarrow Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ ← Standard Normal



Proof:

Since $\bar{Y}$ is a linear function of independent normally distributed random variables, from Theorem 6.3, this theorem is true.

So, in other words, if we have a **random sample of normally** distributed RVs, we know the sampling distribution of the **sample mean**, $\bar{Y}$, will also be **normally distributed**.

This result helps to make inferences about population mean μ based on sample mean $\bar{Y}$ and find probability about sample mean $\bar{Y}$ when population has a normal distribution.