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.

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

- **independent**
- **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:

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

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.)

Examples:

Different samples will have _____ for the **same statistic**.

Therefore, a **statistic** is a _____

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.

Example 2: Let $Y_1, Y_2, \ldots, Y_n$ be an **iid** random sample from a population with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$.
(Note that a specific population distribution has not been provided!).

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

- The **sample mean**,
  $$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$
- 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 $\bar{Y}$.

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 $Y_1, Y_2, \dots, Y_n$ 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 $Y_1, Y_2, \dots, Y_n$ be a **random** sample of size n from a **normal** distribution with **mean** $\boldsymbol{\mu}$ and **variance** $\boldsymbol{\sigma^2}$. Then

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

$$\Longrightarrow \quad Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \quad \sim \quad N(0,1)$$

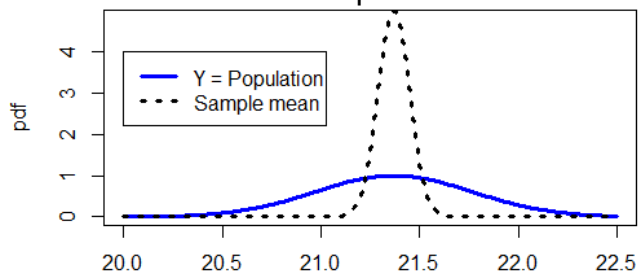
Proof:

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.

Example 3: A candy maker produces mints that have a label weight of 20.4 grams. Assume that the distribution of the weights of these mints has a normal distribution with mean = 21.37 and variance = 0.16.

(a) Let Y be the weight (in grams) of a **single** mint selected at random from the production line. Find the probability that $Y > 21.6$.



(b) Find the probability that the **average** weight (in grams) of a random sample of 25 mints from the production line will be within 0.5 grams of the population mean.
Let $\bar{Y}$ be the **average** weight (in grams) of a random sample of 25 mints from the production line.

Example 4: A company manufactures electoral resistors and assume that the distribution of resistance is a normal distribution with variance of 0.01.

If a random sample of 25 resistors are selected for inspection, then find the probability that sample mean resistance will **differ from the population mean** by **no more than 2.5.**

What about the sampling distribution of other statistics?

Theorem 7.3:

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

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}[Y_i - \bar{Y}]^2 \sim \chi^2\,(v = \mathrm{n} - 1)$$

where χ^2 is a **Chi-square** distribution.

Further, $\bar{Y}$ and S^2 are **independent** random variables.

Facts about the χ^2 Distribution:

- **Definition 4.10**: The $\chi^2(v)$ pdf is a special case of the gamma with $\alpha = v/2$ and $\beta = 2$. If $Y \sim \chi^2\ (v)$**,** then the pdf of Y is

$$f(y) = \frac{1}{\Gamma(v/2)\ 2^{v/2}} y^{(v/2)-1}\ e^{-y/2} \qquad ; \qquad y > 0$$

where v is **a positive integer** that represents the **degrees of freedom (df).** It follows that $E(Y)\ = v$ and $Var(Y)\ = 2v$**.**

- If $Z \sim N(0,1)$, then $Z^2\ \sim\ \chi^2\ (v = 1)$. (You did prove this in Homework 3)
- **Theorem 7.2**: If $X_1, X_2, \ldots, X_n$ are **<u>independent</u>** and $X_i \sim \chi^2\ (v_i)$, then

$$\sum_{i=1}^{n} X_i\ \sim \chi^2 \left(\sum_{i=1}^{n} v_i \right)$$

<u>Ti 83 o4 84 Calculator instructions</u>

<u>If $Y \sim \chi^2(v)$, to get probability that Y is between a and b (that is, find $P(a < Y < b)$):</u>

1. Press [2nd] button and then [VARS] button.
2. Press arrow down to χ^2**cdf(** and press [ENTER].
3. Enter the values for a, b, and v with a comma between each. Close parenthesis and press [ENTER]. That is, a is **lower**, b is **upper,** and v is **df.**
4. If lower bound is $-\infty$, then enter -10^{99} and if upper bound is ∞, then enter 10^{99}

Example 5: A candy maker produces mints that have a label weight of 20.4 grams. Assume that the distribution of the weights of these mints has a normal distribution with mean = 21.37 and variance = 0.16.
Let S^2 be the **variance** of weight (in grams) of a **random sample of 25** mints from the production line. Find the probability that sample variance is within 0.05 of 0.16.
That is, find

Inference

Why do we care about $\bar{Y}$ and S^2 so much?

Certain functions (notably $\bar{Y}$ and S^2) of samples from normal distributions are very important in statistical analyses.

- **S^2 is used to estimate the parameter σ^2 and to make inferences about σ^2**
  For example, if $Y_1, Y_2, \ldots, Y_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution, then we know
  $$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2\,(v = n-1)$$
- **$\bar{Y}$ is used to estimate the parameter μ and to make inferences about μ**
  For example, if $Y_1, Y_2, \ldots, Y_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution, then we know
  $$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N\,(0,1)$$

Why might using $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ or $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N\,(0,1)$ as a basis for inference about the parameter μ present **problems**?

Derived Distributions

Definition 7.2: (Student's t Distribution) If . . .

1. $Z \sim N(0,1)$

2. $W \sim \chi^2\ (v)$

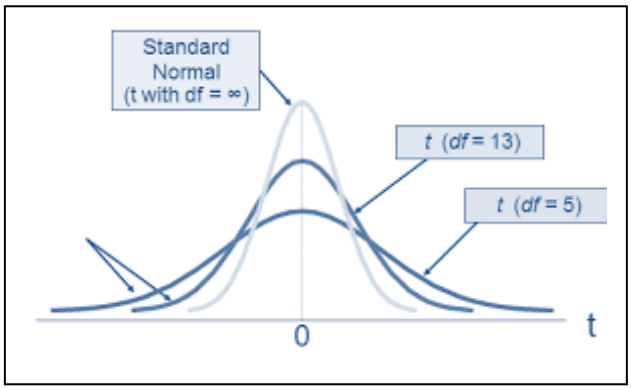
3. Z and W are **independent**

then the random variable

$$T = \frac{Z}{\sqrt{W/v}} \quad \sim \quad \text{the } \mathbf{student's\ t\ distribution} \text{ with } \mathbf{df} = v$$

Properties of Student's t Distribution:

- Symmetric about 0, with shape similar to the standard normal (heavier tails ⇒ "flatter")
- As $\boldsymbol{v} \longrightarrow \infty$, $\boldsymbol{T} \sim \boldsymbol{N(0,1)}$



If $Y \sim T(v)$, to get probability that Y is between a and b (that is find $P(a < Y < b)$):

1. Press [2nd] button and then [VARS] button. This will get you a menu of probability distributions.
2. Press arrow down to **tcdf(** and press [ENTER]. This puts **tcdf(** on the home screen.
3. Enter the values for a, b, and v with a comma between each. Close parenthesis and press [ENTER]. That is, a is **lower**, b is **upper,** and v is **df.**

Example 6: The Nielsen Company reported that U.S. mobile phone subscribers average 3.7 hours per month watching videos on their phone.
You think this number is too low, and so you collect 8 random observations from a cell phone company (you have every reason to believe time spent watching videos follows a normal distribution).
The sample mean and variance are $\bar{y}$ = 6.75 and s^2 = 3.882, respectively.
If the Nielsen Company is correct, how likely is it you find a sample of size 8 with a mean of 6.75 hours, or something more extreme?
What probability do we want here? How could we find it?

Central Limit Theorem (Sections 7.3)

When the distribution of a statistic is unknown, it is often necessary to approximate the distribution of the statistic. Let's look at a (the most famous!) theorem to approximate the sampling distribution of the sample mean.

Theorem 7.4: (Central Limit Theorem)
Let $Y_1, Y_2, \ldots, Y_n$ be **independently and identically distributed** random variables with expected value $E(Y_i) = \mu$ and variance $Var(Y_i) = \sigma^2 < \infty$.
Define the following:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \mathrm{Z} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sqrt{n}\,\sigma}$$

Then, distribution function of Z converges to the **standard normal distribution** as $n \to \infty$.
That is,

$$\lim_{\mathrm{n}\to\infty} P(\mathrm{Z} \le \mathrm{z}) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt \quad \text{for all z}$$

The **practical application** of this is:

$$\textbf{When } n \textbf{ is large}, \quad \bar{Y} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$$

In other words, the **Central Limit Theorem (CLT)** tells us that if we have a random sample, $Y_1, Y_2, \ldots, Y_n$, from a **population distribution that is not necessary normal**, then the sampling distribution of the sample mean, $\bar{Y}$, is an **approximately normal** distribution with the same mean, μ, as the original population distribution and variance σ^2/n .

Note that this theorem applies only to **sample means** !!!

- When is the CLT useful?

Example 7: The time (in hours) that a technician requires to perform preventive maintenance on an air-conditioning unit is governed by a left skewed distribution with mean 1 hour and standard deviation 1 hour.
Your company has a contract to maintain 70 of these units in an apartment building. You must schedule technicians' time for a visit to this building.
Is it safe to budget an average of 1.1 hour for each unit?
Or should you budget an average of 1.25 hours?

Sections 7.1-7.3 of the textbook