Inference

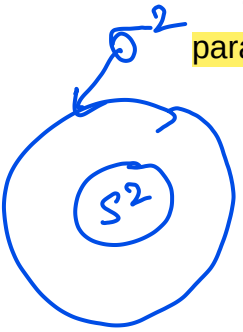
We make conclusions about the population based on a random sample.

Why do we care about $\bar{Y}$ and S^2 so much? Because we use those to make inferences about the population mean (μ) and population variance (σ^2).

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
  $$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

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**?



Population variance (σ^2) is unknown. It has to be estimated using the sample variance (s^2).

$s^2 \neq \sigma^2$ $\quad |s^2 - \sigma^2| \neq 0$

Derived Distributions

When the population variance (σ^2) is estimated using the sample variance (s^2), there is an error introduced because $s^2 \neq \sigma^2$ and s^2 is a random variable (like $\bar{Y}$) while σ^2 is a constant (like μ).
Therefore, the normal distribution cannot be used to make inferences about population mean (μ) when σ^2 is unknown.

$N(0,1) \Rightarrow$ Standard Normal

Sections 7.1-7.3 of the textbook

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

1. $Z \sim N(0,1)$ such as $Z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}}$ has $N(0,1)$
2. $W \sim \chi^2(v)$ such as $W = \frac{(n-1)s^2}{\sigma^2}$ has $\chi^2 (v = n-1)$
3. Z and W are **independent** such as $\bar{y}$ and s^2 are independent.

then the random variable

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

$$\Rightarrow T = \frac{\frac{\bar{y}-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}} = \frac{\bar{y}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

has a t distribution with $df = n-1$ when population has a normal distribution.

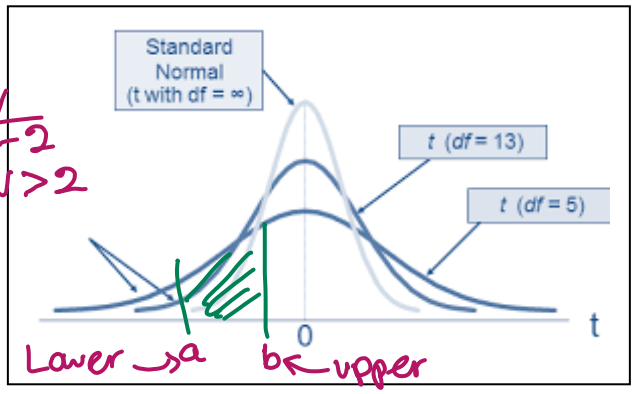
Properties of Student's t Distribution:

- Symmetric about 0, with shape similar to the standard normal (heavier tails ⇒ "flatter")
- As $v \longrightarrow \infty$, $T \sim N(0,1)$ That is, when v is larger, the t distribution is closer to the Standard Normal distribution.

TI 83 or 84 calculator.

$E(T) = 0$

$Var(T) = \frac{v}{v-2}$ where $v > 2$



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.**

If using tables, then use Table 5 at the end of the textbook.

In R `pt(a, v, lower.tail = FALSE)` to get $P(T > a)$.

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). ⇒ Population has 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?

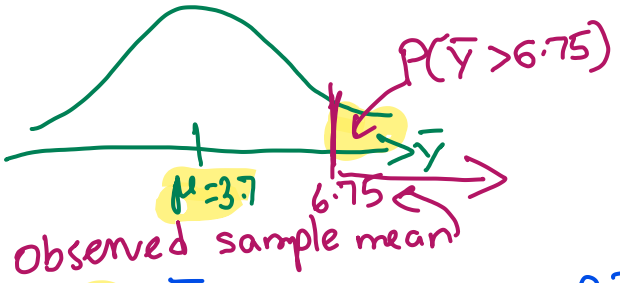
What probability do we want here? How could we find it?

Let $\bar{Y}$ be the mean of 8 random observations. Since we know the population has a normal distribution

and σ^2 is unknown, we use the t distribution.

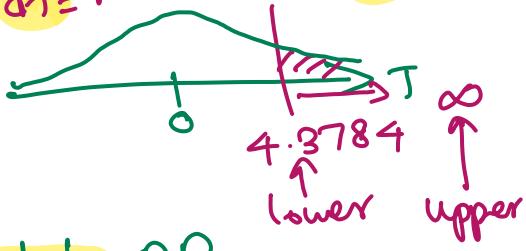
From Theorem 7.1, $\bar{Y}$ has a normal distribution with mean $= \mu$ and variance $= \frac{\sigma^2}{n}$.

If the company is correct, then $\mu = 3.7$. Since the observed sample mean of 6.75 is higher than $\mu = 3.7$, $\bar{Y} > 6.75$ indicates more extreme values for $\bar{Y}$. Find $P(\bar{Y} > 6.75)$



First, compute the T for $\bar{y} = 6.75$.

$$T = \frac{\bar{Y} - \mu}{\left(\frac{S}{\sqrt{n}}\right)} = \frac{6.75 - 3.7}{\left(\frac{\sqrt{3.882}}{\sqrt{8}}\right)} = 4.3784 \text{ with } df = n - 1 = 8 - 1 = 7$$



$\Rightarrow P(\bar{Y} > 6.75) = P(T > 4.3784) = 0.00162$

Using tcdf(4.3784, 10^{99}, 7) on Ti calculator OR

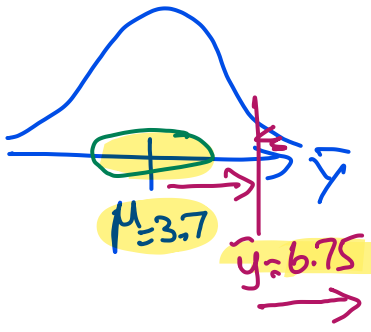
Using pt(4.3784, 7, lower.tail = FALSE) in R.

OR use Table 5 and get 0.005.

$\Rightarrow$ If the company is correct ($\mu = 3.7$), then it is rare that a random sample of 8 observations has a mean of 6.75 or more extreme value.

That is, the company is wrong.

If the company was correct, then the observed sample mean has to be closer to the population mean of 3.7. and cannot be rare.



(*) Most often use 0.05 for the highest probability of an unusual or rare occurrence.

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) CLT

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:

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

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

That is,

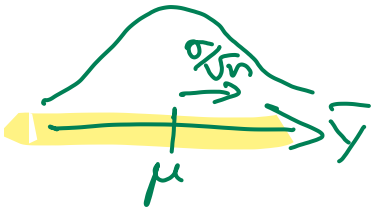
$$\lim_{n\to\infty} P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \quad \text{for all } z$$

The **practical application** of this is:

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

mean, variance, standard normal, approximately, normal

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?

When the population does not have a normal distribution and the sample size (n) is at least 30.

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. NOT normal

Population (NOT sample)

$\mu = 1$, $\sigma = 1$

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?

Let $\bar{Y}$ be the average maintenance time per unit for a random sample of 70 units. $n = 70$. Although the population does not have a normal distribution, the sample size ($n = 70$) is bigger than 30. Therefore, from CLT,