APS606 - Lecture Note (Week 3)

Norman Lo, M.S. in Quantitative Economics

9/30/2020

Introduction to Sampling Distribution

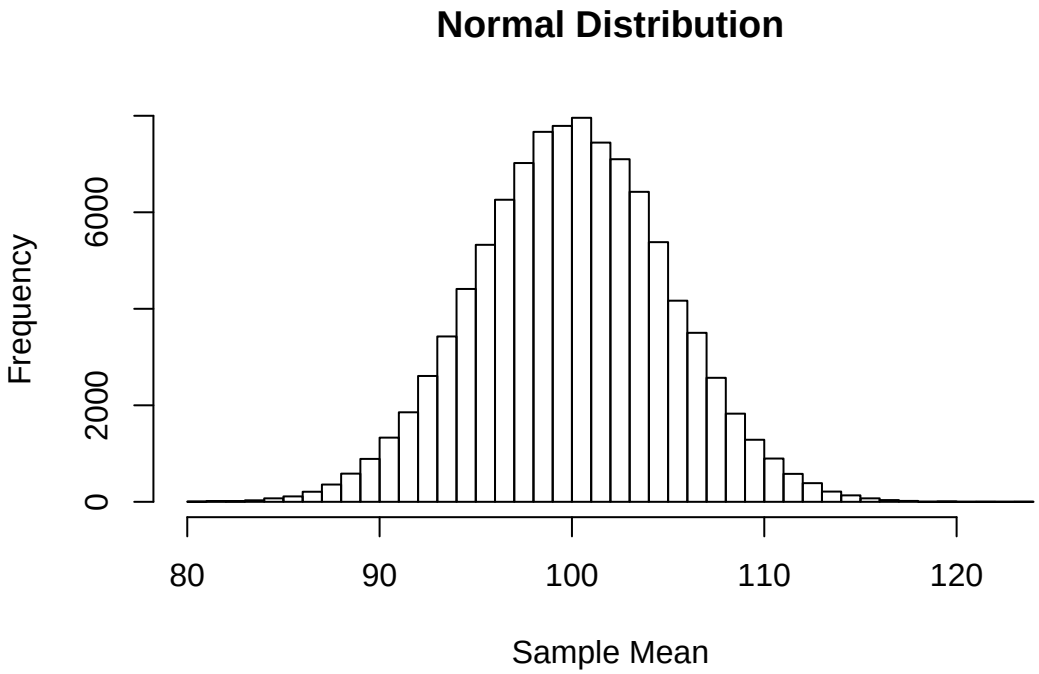
I. Central Limit Theorem
II. Sampling Distribution (Student's t Distribution)
III. t Statistcis and p-value
IV. Confidence Interval

I. Central Limite Theorem:

When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of $\bar{x}$. In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution as the sample size becomes large. The standard deviation of the sampling distribution is referred to the **standard error**. Here are the statistics for sampling distribution:

Finite Population	Infinite Population
$E(\bar{x}) = \mu$	$E(\bar{x}) = \mu$
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}(\frac{\sigma}{\sqrt{n}})$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
$E(\bar{p}) = p$	$E(\bar{p}) = p$
$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}$	$\sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}}$

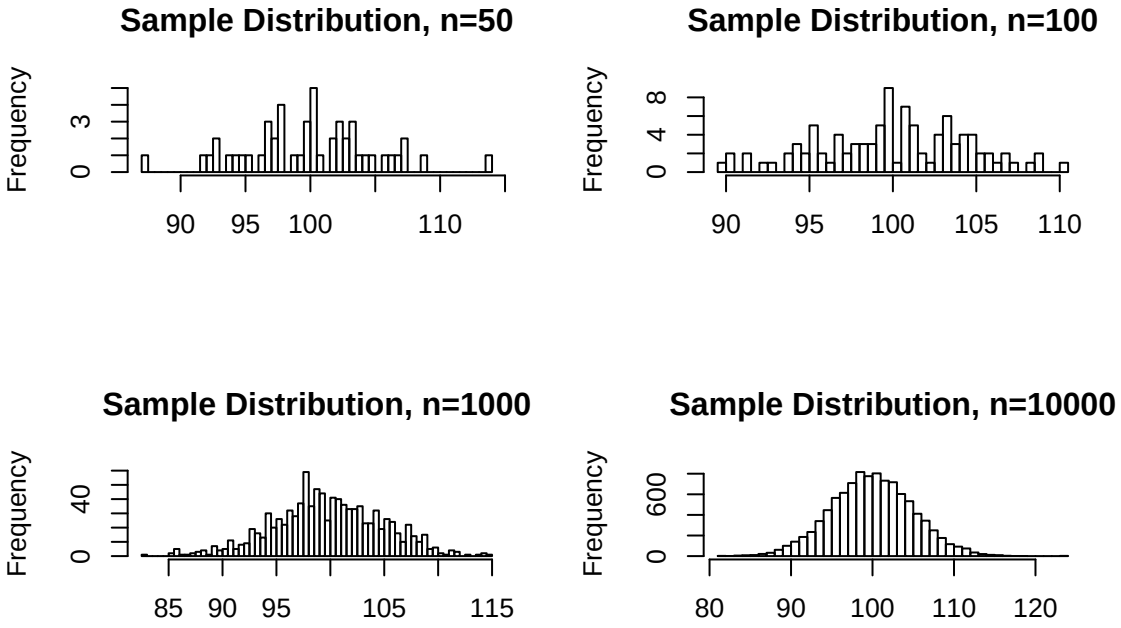
To demonstrates the **Central Limit Theorem**, we can simulate a normally distributed data with mean 100 and sd 5 as the population data for our study. Then, draw different size of sample from the population and see how the distribution shape getting closer to closer to normal distribution.



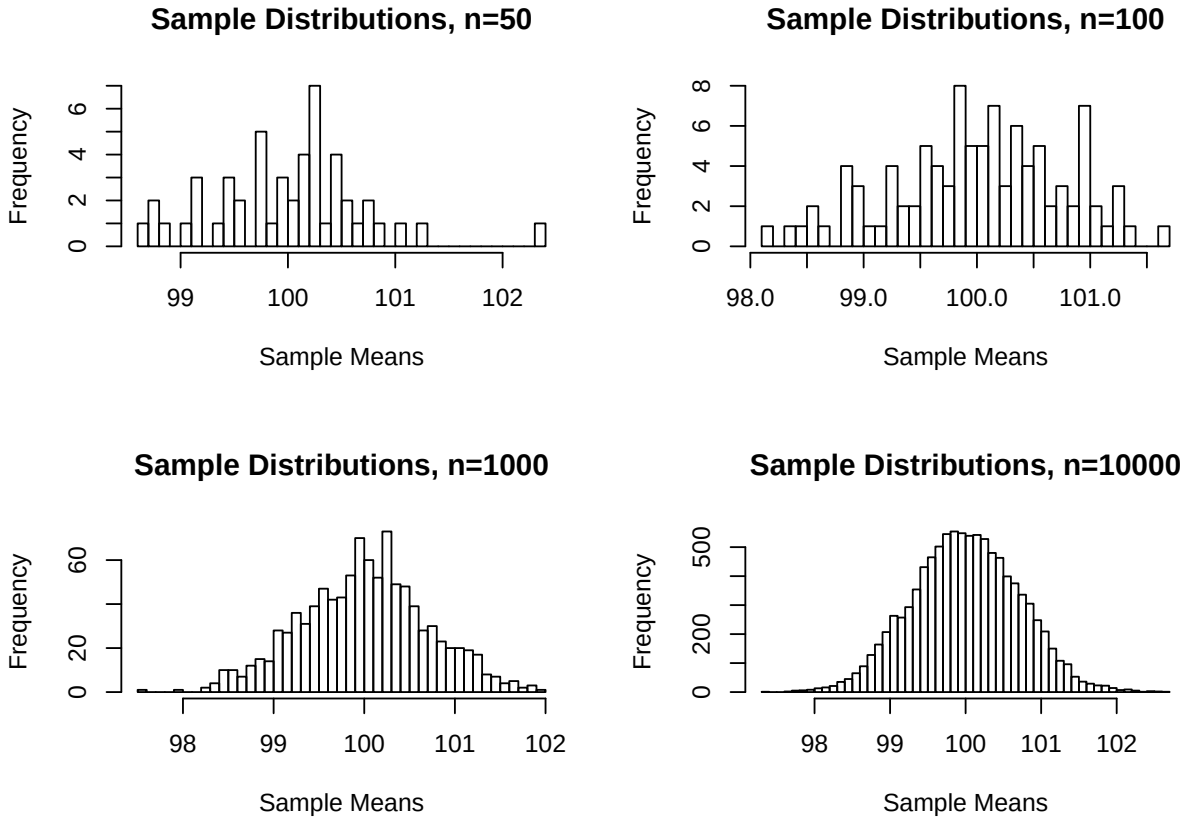
If we select 50 random sample from the population data, we can can calculate the mean of the samples.

```
## [1] "The sample mean from 50 randomly selected samples is  99.717"
```

If we increase the size of sample, the distributions of the sample data are going to get closer and closer to normal.



To be more accurate, we should randomly select sample sets with size of 50 observations multiple times from the population to check the distribution of the mean of the sample means to confirm the central limit theorem.



Overall, the mean of our samples, if we take enough, will equal the mean of the population. As we keep taking samples from the population data, the distribution of the means of those samples will stack up close to 100, forming a normal distribution.

Why do we want to know the sample distribution approximates the normal distribution?

One of the key objective for learning the central limit theorem is to know that the sampling distribution is normally distribute with mean $\bar{x}$ and standard deviation s, so that we can apply the probabiity distribution properties that we learned from the previous section with the replacement of σ by $\sigma_{\bar{x}}$.

II. Sampling Distribution

Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about "making inferences" the way statisticians think about it, we need to be a bit more explicit about what it is that we're drawing inferences from (the sample) and what it is that we're drawing inferences about (the population).

In almost every situation of interest, what we have available to us as researchers is a sample of data. We might have run experiment with some number of participants; a polling company might have phoned some number of people to ask questions about voting intentions; etc. Regardless: the data set available to us is finite, and incomplete. We can't possibly get every person in the world to do our experiment; a polling company doesn't have the time or the money to ring up every voter in the country etc. Our only goal was to find ways of describing, summarising and graphing that sample.

Unbias Estimation is a form of statistical inference, which we use the data from the sample to compute a value of a sample statistics that serves as an estimate of a population parameter. We refer $\bar{x}$ as the unbias estimator of the population mean μ. s is the unbias estimator of the population standard deviation σ.

Why do we want to know the sample distribution approximates the normal distribution?

One of the key objective for learning the central limit theorem is to know that the sampling distribution is normally distribute with mean $\bar{x}$ and standard deviation s, so that we can apply the probabiity distribution properties that we learned from the previous section with the replacement of σ by $\sigma_{\bar{x}}$.

Sample Means Normal Distribution Statistics:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Example:

A census data indicates that the population, on average, spend $\mu = 13$ minutes exercising every week and is normally distributed with standard deviation $\sigma = 4.6$ minutes. What is the probability to randomly select 10 samples from the population that spend at least 15 minutes $\bar{x} > 15$ on average to exercise weekly?

Solution:

$$P(\bar{x} > 15) = ?$$

Step 1: Convert $\bar{x}$ to the standard normal distribution

$$z = \frac{15 - 13}{\frac{4.6}{\sqrt{10}}} = \frac{15 - 13}{1.455} = 1.375$$

Step 2: Find the area under the standard normal curve to the right of z = 2.433 (using the pnorm() function in r)

```
# Calculate the probability with the z score
round(1 - pnorm(1.375), 3)
```

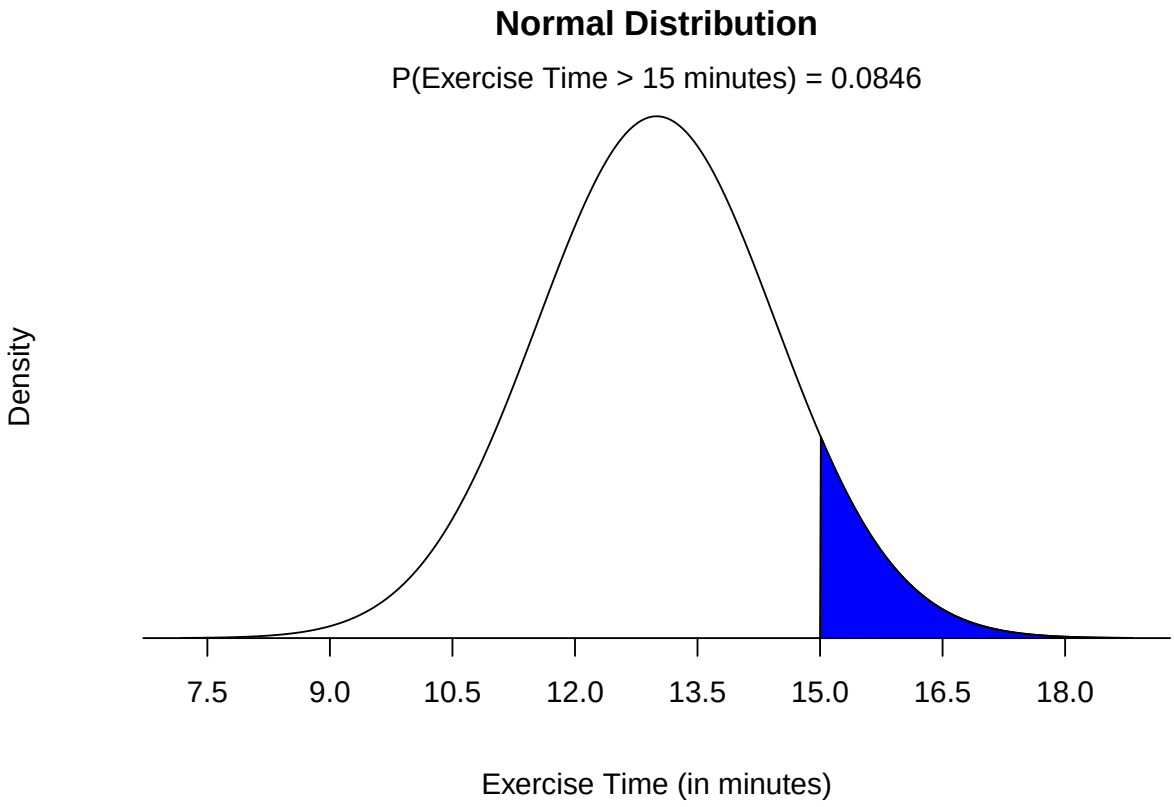
```
## [1] 0.085
```

```
# Alternative, calculate the probability with the average time
round(1 - pnorm(15, mean=13, sd=1.455), 3)
```

```
## [1] 0.085
```

$$P(z > 1.375) = 0.0846 \; or \; 8.46\%$$

Graphical Solution:



III. Student's t Distribution

Sample Distribution without a Known Population Standard Deviation

In the last example, we assume that the population standard deviation σ is given. However, it is difficult to imagine we always know about the population parameters, especially σ. So, the problem is what if we only know the population mean μ, but not the standard deviation σ, can we still construct a distribution similar to the normal distribution for the probability problem?

Fortunately, the answer is "YES". A particular distribution we are going to apply is called the "**t-distribution**" or sometime referred to the "**Student's t-distribution**". The t-distribution is designed to approximate the normal distribution by estimating the unknown population standard deviation σ. The only tiny tweak to tranform the unbaised estimator is divide the sum of squares $\sum (x_i - \bar{x})^2$ by n minus 1, $(n-1)$, degree of freedom, to estimate the popluation variance.

Estimated Population Standard Deviation:

$$\hat{\sigma} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Estimated Standard Error:

$$\hat{\sigma x} = \frac{\hat{\sigma}}{\sqrt{n}}$$

Let's use an example to demonstrate the different between the two situations.

Case 1: σ is known

Given the population data with $\mu = 10$ and $\sigma = 2.5$, calculate the standard error $\sigma_{\bar{x}}$ for the normal distribution with sample size n = 25.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{25}} = 0.5$$

Case 2: σ is unknown

Given the population data with $\mu = 10$, calculate the estimated standard error $\hat{\sigma x}$ for the normal distribution with sample size n = 25.

Note: In this case, we need to calculate the estimated population standard deviation by the sample data.

Step 1: Calculate the estimated stardard deviation of the population.

$$\hat{\sigma} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Step 2: Calculate the estimated standard error of the t-distribution.

$$\hat{\sigma x} = \frac{\hat{\sigma}}{\sqrt{n}}$$

```r
# Generate the population data
p <- rnorm(100, mean=10, sd=2.5)

# Extract 25 samples from the population data
x <- sample(p, 25)

# Calculate the mean of the sample
xBar <- mean(x)

# Calculate the estimated population standard deviation
sdHat <- round(sqrt((sum((x - xBar)^2)/(25 - 1))), 3)

print(paste("The estimated population standard deviation is ", sdHat))
```

```
## [1] "The estimated population standard deviation is  2.266"
```

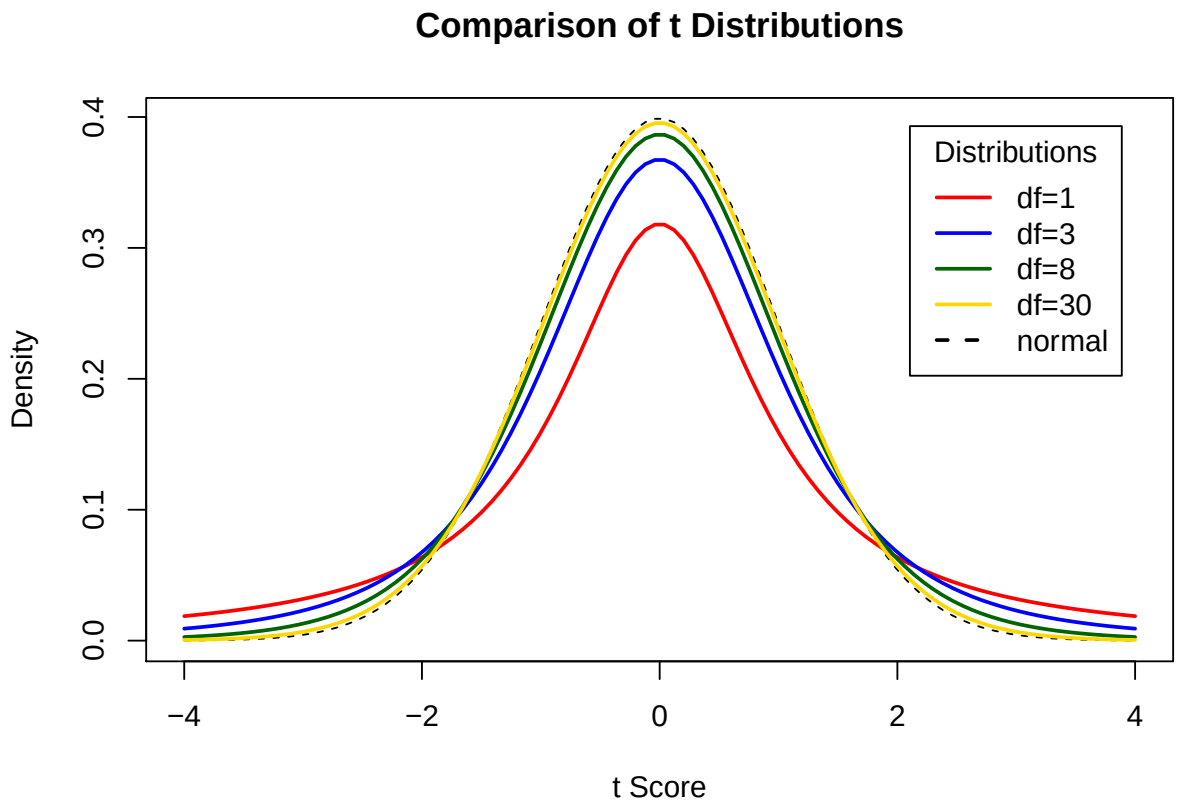
```r
# Calculate the estimated standard error for the t-distribution
se <- round(sdHat/sqrt(25), 3)

print(paste("The estimated standard erorr of the t-distribution is ", se))
```

```
## [1] "The estimated standard erorr of the t-distribution is  0.453"
```

Thanks to the t-distribution, now we can solve the probability problem with the new defined distribution. Since the shape of the t-distribution is mainly depends on the sample size (n), the probability distribution table and r function will be slightly different to the standardized normal distribution (z-distibution). The t-distribution has the standardized scores called "t statistic":

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

It is also worth to note that one of the properties of the t distribution is that the shape of the distribution is mainly determined by the sample size. Here we have a visualization to help to understand.



Example:

A survey data indicates that the population, on average, spend $\mu = 13$ minutes exercising every week and sample sum of squares $\sum(x_i - \bar{x})^2 = 91.4$ minutes. What is the probability to randomly select 10 samples from the population that spend at least 15 minutes $\bar{x} > 15$ on average to exercise weekly?

Solution:

$$P(\bar{x} > 15) = ?$$

Step 1: Calculate the estimated population standard deviation $\hat{\sigma}$

$$\hat{\sigma} = \sqrt{\frac{91.4}{(10-1)}} = 3.187$$

Step 2: Calculate the estimated standard error of the t-distribution

$$\hat{\sigma}_{\bar{x}} = \frac{3.187}{\sqrt{10}} = 1.008$$

Step 3: Convert $\bar{x}$ to the t statistic score

$$t = \frac{15 - 13}{1.008} = 1.984$$

Step 4: Find the area under the t distribution to the right of t = 1.984 (using the pt() function in r)

```
# Calculate the probability with the z score
round(1-pt(1.984, df=9), 3)
```

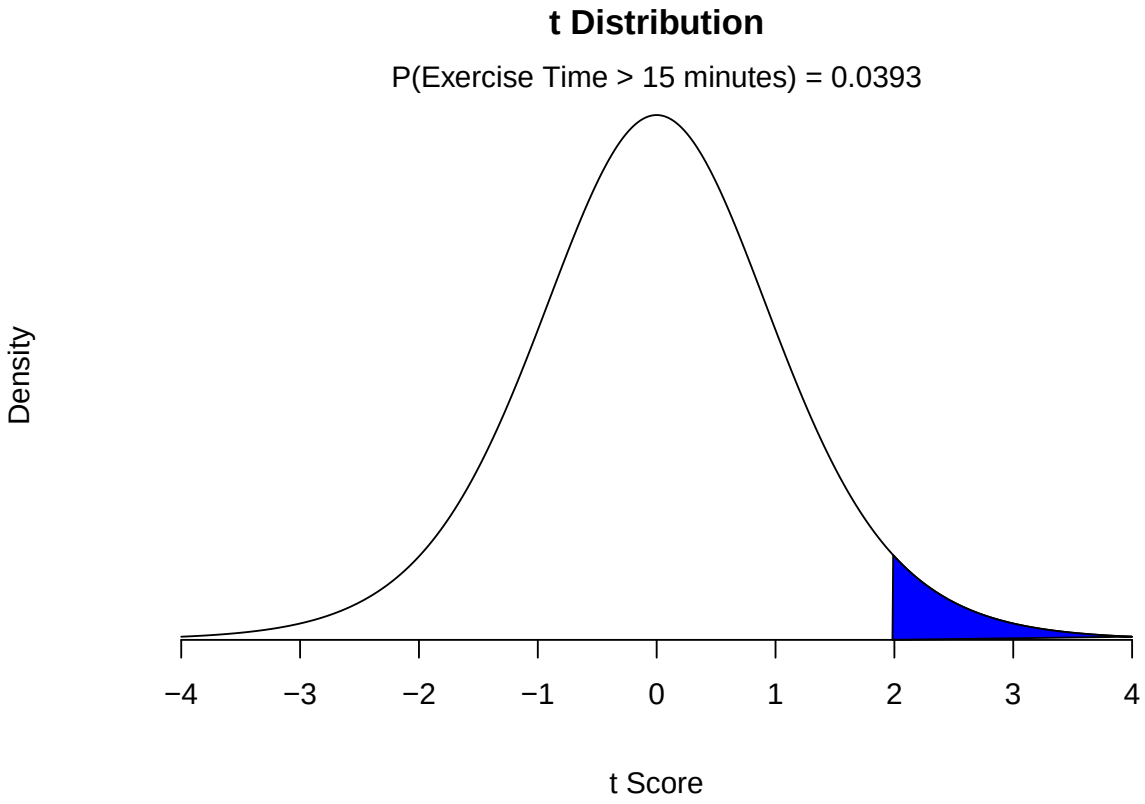
```
## [1] 0.039
```

```
# Alternative, use the parameter lower.tail=FALSE
round(pt(1.984, df=9, lower.tail=FALSE), 3)
```

```
## [1] 0.039
```

$$P(t > 1.984) = 0.0393 \; or \; 3.93\%$$

Graphical Solution:



IV. Confidence Interval

Interval Estimate of a Population Mean: (σ is unknown)

If an estimate of the population standard deviation σ cannot be developed prior to sampling, we use the sample standard deviation s to estimate σ. In this case, the interval estimate for μ is based on the t distribution. A specific t distribution depends on a parameter known as the degrees of freedom. Degrees of freedom refer to the number of independent pieces of information that go into the computation of **s**.

A t distribution with more degrees of freedom has less dispersion. As the degrees of freedom increases, the difference between the t distribution and the standard normal probability distribution becomes smaller and smaller. For more than 100 degrees of freedom, the standard normal z value provides a good approximation to the t value. Usually, a sample size of n $\geq$ 30 is adequate when using the expression $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ to develop an interval estimate of a population mean. If the population distribution is highly skewed or contains outliers, a sample size of 50 or more is recommended. If the population is not normally distributed but is roughly symmetric, a sample size as small as 15 will suffice. If the population is believed to be at least approximately normal, a sample size of less than 15 can be used.

Example:

A reporter for a student newspaper is writing an article on the cost of off-campus housing. A sample of 16 one-bedroom apartments within a half-mile of campus resulted in a sample mean of $750 per month and a sample standard deviation of $55.

Let us provide a 95% confidence interval estimate of the mean rent per month for the population of one-bedroom apartments within a half-mile of campus. We will assume this population to be normally distributed.

Solution:

At 95% confidence, $\alpha = 0.05$, and $\alpha/2 = 0.025$

$t_{0.025}$ is based on n - 1 => 16 - 1 = 15 degrees of freedom

Interval Estimate:

$$\bar{x} \pm t_{0.025}\frac{s}{\sqrt{n}}$$

$$750 \pm 2.131\frac{55}{\sqrt{16}} = 750 \pm 29.30$$

We are 95% confident that the mean rent per month for the population of one-bedroom apartments within a half-mile of campus is between $720.70 and $779.30.

Confidence Interval in R

```
t <- qt(0.025, df = 15, lower.tail = FALSE)
margin_of_error <- t *  55 / sqrt(16)
Upper <- 750 + margin_of_error
Lower <- 750 - margin_of_error
print(paste("95% C.I. of the average monthly cost for one-bedroom units is $", round(Lower, digits = 2)
```

```
## [1] "95% C.I. of the average monthly cost for one-bedroom units is $ 720.69 and $ 779.31 ."
```

In some cases, we may want to find out if a sample data significantly different (higher or lower) to the other sample given a different conditions. The confidence intervals is useful to define how significantly different between the two samples.

Example:

Imagine that we consider a town with a mean household income of greater than $100,000 to be high-income.

For Town A we sample some households, and calculate the mean household income and the 95% confidence interval for this statistic. The mean is $125,000, but the data is somehow variable, and the 95% confidence interval is from $75,000 to $175,000. In this case, we don't have much confidence that Town A is actually a high-income town. The point estimate for the population mean is greater than $100,000, but the confidence interval extends considerably lower than this threshold.

For Town B, we also get a mean of $125,000, so the point estimate is the same as for Town A. But the 95% confidence interval is from $105,000 to $145,000. Here, we have some confidence that Town B is actually a high-income town, because the whole 95% confidence interval lies higher than the $100,000 threshold.