Point Estimates and Confidence Intervals

Statistical inference is the process of analyzing sample data to gain insight into the population from which the data was collected and to investigate differences between data samples. In data analysis, we are often interested in the characteristics of some large population, but collecting data on the entire population may be infeasible. For example, leading up to U.S. presidential elections it could be very useful to know the political leanings of every single eligible voter, but surveying every voter is not feasible. Instead, we could poll some subset of the population, such as a thousand registered voters, and use that data to make inferences about the population as a whole.

1. Point Estimates

Point estimates are estimates of population parameters based on sample data. For instance, if we wanted to know the average age of registered voters in the U.S., we could take a survey of registered voters and then use the average age of the respondents as a point estimate of the average age of the population as a whole. The average of a sample is known as the sample mean.
The sample mean is usually not exactly the same as the population mean. This difference can be caused by many factors including poor survey design, biased sampling methods and the randomness inherent to drawing a sample from a population. Let's investigate point estimates by generating a population of random age data and then drawing a sample from it to estimate the mean:

In [1]:
```
set.seed(12)
population_ages <- c(rexp(1000000,0.015)+18,   # Generate a population
                     rpois(500000,20)+18,
                     rpois(500000,32.5)+18,
                     rpois(500000,45)+18)

population_ages <- ifelse(population_ages<100, population_ages, population_ages%%100+18)


true_mean <- mean(population_ages)          # Check the population mean

true_mean
```

Out[1]:
```
51.2188371860945
```

In [2]:
```
set.seed(10)
sample_ages <- sample(population_ages, size=1000) # Take a sample of 1000 ages
```

```
sample_mean <- mean(sample_ages)          # Make a point estimate of the mean

sample_mean

sample_mean-true_mean #Check difference between estimate and population parameter
```

Out[2]:
52.1636089223386
Out[2]:
0.944771736244064

Our point estimate based on a sample of 1000 individuals overestimates the true population mean by almost a year, but it is pretty close. This illustrates an important point: we can get a fairly accurate estimate of a large population by sampling a relatively small subset of individuals.
Another point estimate that may be of interest is the proportion of the population that belongs to some category or subgroup. For example, we might like to know the race of each voter we poll, to get a sense of the overall demographics of the voter base. You can make a point estimate of this sort of proportion by taking a sample and then checking the ratio in the sample:

In [3]:

```
set.seed(12)
population_races <- c(rep("white",1000000), # Generate some dummy demographic data
                      rep("hispanic",500000),
                      rep("black",500000),
                      rep("asian",250000),
                      rep("other",250000))

demographic_sample <- sample(population_races, size=1000)     # Take a sample

for (race in unique(demographic_sample)){          # Loop over each race*
   print(paste(race," proportion estimate:"))
   print(sum(demographic_sample==race)/1000)       # Print the estimated proportion
}
```

```
[1] "white  proportion estimate:"
[1] 0.4
[1] "asian  proportion estimate:"
[1] 0.096
[1] "other  proportion estimate:"
[1] 0.093
[1] "black  proportion estimate:"
[1] 0.207
[1] "hispanic  proportion estimate:"
[1] 0.204
```

*Note: The function unique() takes a vector and returns a new vector with duplicate elements removed.

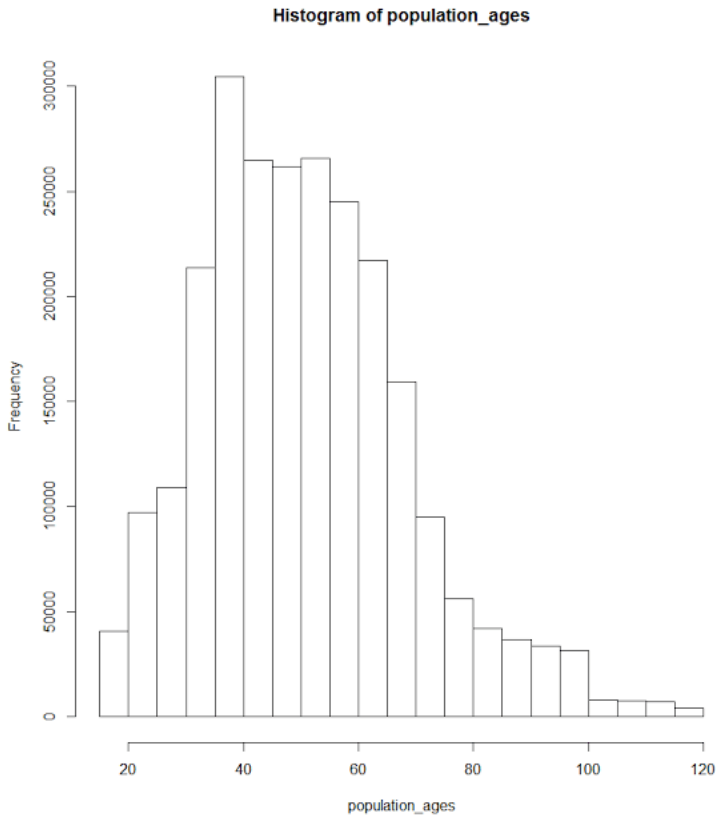
Sampling Distributions and The Central Limit Theorem

Many statistical procedures assume that data follows a normal distribution, because the normal distribution has nice properties like symmetricity and having the majority of the data clustered within a few standard deviations of the mean. Unfortunately, real world data is often not normally distributed and the distribution of a sample tends to mirror the distribution of the population. This means a sample taken from a population with a skewed distribution will also tend to be skewed. Let's investigate by plotting the data and sample we created earlier and by checking the skew:

In [4]:
```
library(e1071)
```

In [5]:
```
hist(population_ages, breaks=20)  # Create histogram of population

skewness(population_ages)        # Check the skewness
```



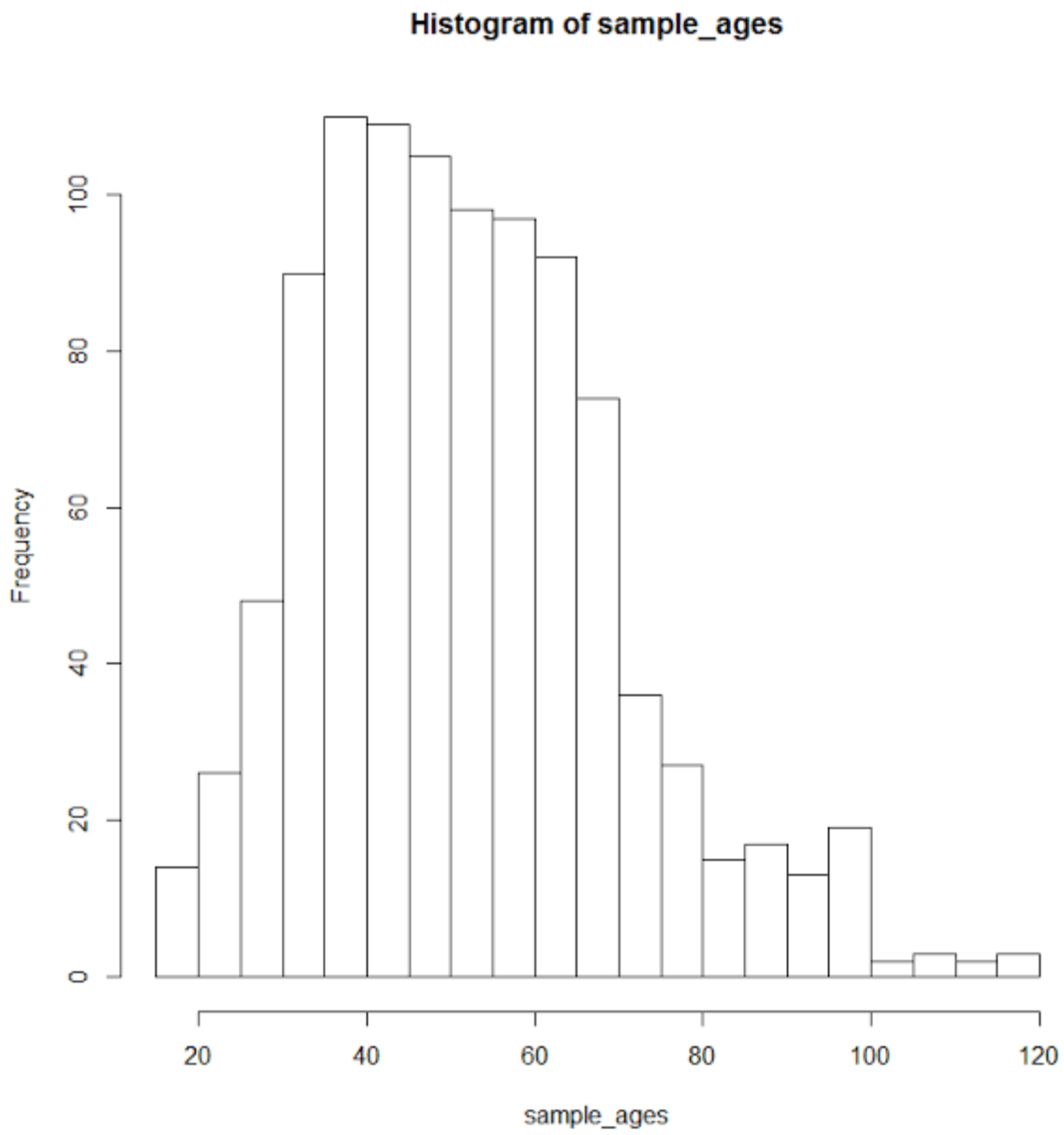
Out[5]:
```
0.655602782446446
```

The histogram shows a distribution with right skew, which is confirmed by the skewness measurement of 0.6556. The sample we drew should have roughly the same shape and skewness:

In [6]:

```
hist(sample_ages, breaks=20)   # Create histogram of the sample

skewness(sample_ages) # Check the skewness (point estimate of skewness)
```



Out[6]:

0.670960675526386

The sample has roughly the same skew as the underlying population. This suggests that we can't apply techniques that assume a normal distribution to this data set. In reality, we can, thanks the central limit theorem.

The central limit theorem is one of the most important results of probability theory and serves as the foundation of many methods of statistical analysis. At a high level, the theorem states the distribution of many sample means, known as a sampling distribution, will be normally distributed. This rule holds even if the underlying distribution itself is not normally distributed. As a result we can treat our a sample mean as if it were drawn normal distribution.

To illustrate, let's create a sampling distribution by taking 200 samples from our population and then making 200 point estimates of the mean:

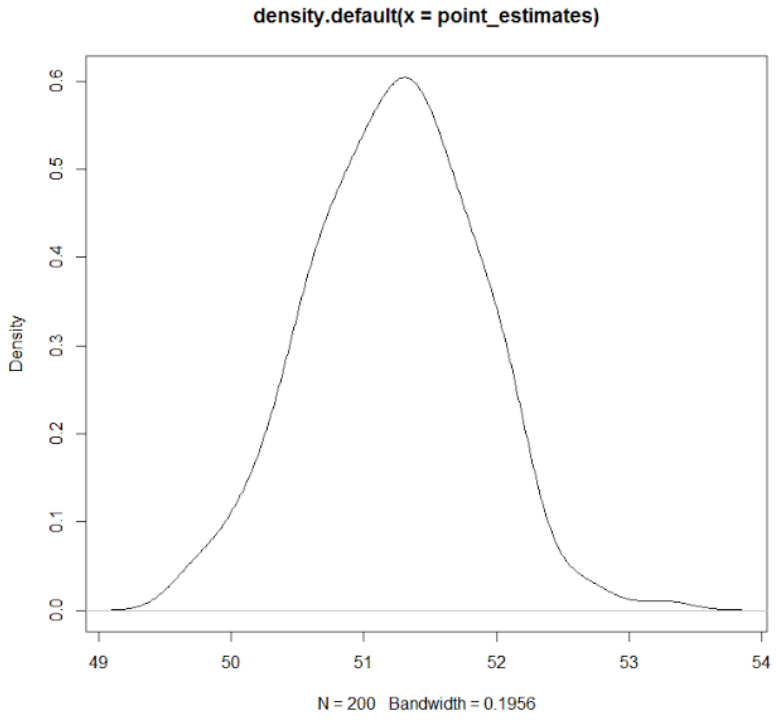
In [7]:

```
set.seed(12)
point_estimates <- c()    # Create an empty vector to hold results

num_samples <- 200        # Initialize number of samples to take

for (x in 1:num_samples){      # Draw 200 samples and make 200 point estimates
   sample <- sample(population_ages, size=1000)
   point_estimates <- c(point_estimates, mean(sample))
}

plot(density(point_estimates))  # Plot the sampling distribution
```



The sampling distribution appears to be roughly normal, having significantly less skew than the original distribution:

In [8]:

```
skewness(point_estimates)
```

Out[8]:

```
-0.0130738593557108
```

In addition, the mean of the sampling distribution approaches the true population mean:

In [9]:

```
mean(point_estimates)

mean(point_estimates)-true_mean# Difference between true mean and sample means
```

Out[9]:

```
51.2249750096931
```

Out[9]:

```
0.00613782359863535
```

The more samples we take, the better our estimate of the population parameter is likely to be.

2. Confidence Intervals for One Mean

In this section, we'll learn how to calculate a confidence interval for a population mean. As we'll soon see, a confidence interval is an interval (or range) of values that we can be really confident contains the true unknown population mean. We'll get our feet wet by first learning how to calculate a confidence interval for a population mean (called a ***Z*-interval**) by making the unrealistic assumption that we know the population variance. (Why would we know the population variance but not the population mean?!) Then, we'll derive a formula for a confidence interval for a population mean (called a ***t*-interval**) for the more realistic situation that we don't know the population variance. We'll also spend some time working on understanding the "confidence part" of an interval, as well as learning what factors affect the length of an interval.

Objectives

- To learn how to calculate a confidence interval for a population mean.
- To understand the statistical interpretation of confidence.
- To learn what factors affect the length of an interval.
- To understand the steps involved in each of the proofs in the lesson.
- To be able to apply the methods learned in the lesson to new problems.

The Situation

Point estimates, such as the sample proportion ($\hat{p}$), the sample mean ($\bar{x}$), and the sample variance (s^2) depend on the particular sample selected. For example:

(1) We might know that $\hat{p}$, the proportion of a sample of 88 students who use the city bus daily to get to campus, is 0.38. But, the bus company doesn't want to know the sample proportion. The bus company wants to know population proportion *p*, the proportion of *all* of the students in town who use the city bus daily.

(2) We might know that $\bar{x}$, the average number of credit cards of 32 randomly selected American college students is 2.2. But, we want to know μ, the average number of credit cards of *all* American college students.

The Problem

(1) When we use the sample mean $\bar{x}$ to estimate the population mean μ, can we be confident that $\bar{x}$ is close to μ? And, when we use the sample proportion $\hat{p}$ to estimate the population proportion *p*, can we be confident that $\hat{p}$ is close to *p*?

(2) Do we have any idea as to how close the sample statistic is to the population parameter?

A Solution

Rather than using just a point estimate, we could find an interval (or range) of values that we can be really confident contains the actual unknown population parameter. For example, we could find lower (*L*) and upper (*U*) values between which we can be really confident the population mean falls:

$$L < \mu < U$$

And, we could find lower (*L*) and upper (*U*) values between which we can be really confident the population proportion falls:

$$L < p < U$$

An interval of such values is called a **confidence interval**. Each interval has a **confidence coefficient** (reported as a proportion):

$$1 - \alpha$$

or a **confidence level** (reported as a percentage):

$$(1 - \alpha)100\%$$

Typical confidence coefficients are 0.90, 0.95, and 0.99, with corresponding confidence levels 90%, 95%, and 99%. For example, upon calculating a confidence interval for a mean with a confidence level of, say 95%, we can say:

"We can be 95% confident that the population mean falls between *L* and *U*."

As should agree with our intuition, the greater the confidence level, the more confident we can be that the confidence interval contains the actual population parameter.

2.1 A Z-Interval for a Mean

Now that we have a general idea of what a confidence interval is, we'll now turn our attention to deriving a particular confidence interval, namely that of a population mean μ. We'll jump right ahead to the punch line and then back off and prove the result. But, before stating the result, we need to remind ourselves of a bit of notation.

The value:

$$z_{\alpha/2}$$

is the *Z*-value (obtained from a standard normal table or calculated by the qnorm() function) such that the area to the right of it under the standard normal curve is $\alpha/2$. That is:

$$P(Z \geq z\alpha/2) = \alpha/2$$

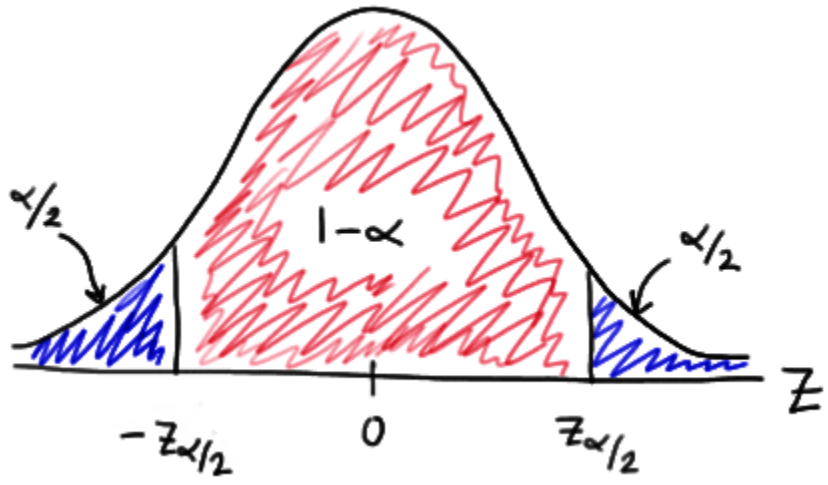
Likewise:

$$-z\alpha/2$$

is the *Z*-value (obtained from a standard normal table) such that the area to the left of it under the standard normal curve is $\alpha/2$. That is:

$$P(Z \leq -z\alpha/2) = \alpha/2$$

This notation can be illustrated with the following diagram of a standard normal curve:



With the notation now recalled, let's state the formula for a confidence interval for the population mean.

Theorem. Assume:

(1) $X_1, X_2, ..., X_n$ is a random sample from a normal population with mean μ and variance σ^2. So that:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \text{ and } Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

(2) The population variance σ^2 is known.

Then, a **$(1 - \alpha)100\%$ confidence interval for the mean μ** is:

$$\bar{X} \pm z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$$

Proof. From the above diagram of the standard normal curve, we can see that the following probability statement is true:

$$P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1 - \alpha$$

Then, simply replacing Z, we get:

$$P[-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}] = 1 - \alpha$$

Now, let's focus only on manipulating the inequality inside the brackets for a bit. Because we manipulate each of the three sides of the inequality equally, each of the following statements are equivalent:

$$\begin{aligned}
-z_{\alpha/2} &\leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \\
-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq \bar{X} - \mu \leq +z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \\
-\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq -\mu \leq -\bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \\
\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq \mu \leq \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)
\end{aligned}$$

So, in summary, by manipulating the inequality, we have shown that the following probability statement is true:

$$P\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1 - \alpha$$

In reality, we'll learn on the next page why we shouldn't (and therefore don't!) write the formula for the Z-interval for the mean quite like that. Instead, we write that we can be $(1 - \alpha)100\%$ confident that the mean μ is in the interval:

$$\left[\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

The interval, because it depends on Z, is often referred to as the ***Z*-interval for a mean**.

More generally, we can find the values for any confidence level. This is usually denoted in reverse by calling it a (1-α)100% confidence level. Where for any α in (0,1) we can find a z^* with

$P(-z^* < z < z^*) = 1-\alpha$

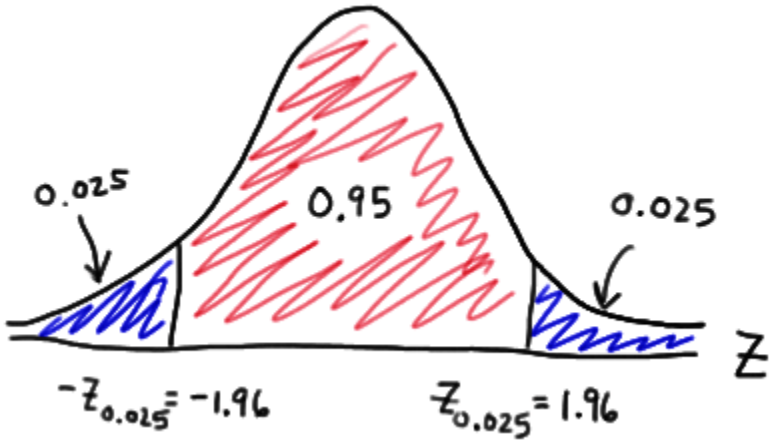
Often such a z^* is called $z_{1-\alpha/2}$ from how it is found. For R this can be found with the qnorm function

```
> alpha = c(0.2,0.1,0.05,0.001)
> zstar = qnorm(1 - alpha/2)
> zstar
[1] 1.281552 1.644854 1.959964 3.290527
```

Example

A random sample of 126 police officers subjected to constant inhalation of automobile exhaust fumes in downtown Cairo had an average blood lead level concentration of 29.2 $\mu g/dl$. Assume X, the blood lead level of a randomly selected policeman, is normally distributed with a standard deviation of $\sigma = 7.5$ $\mu g/dl$. Historically, it is known that the average blood lead level concentration of humans with no exposure to automobile exhaust is 18.2 $\mu g/dl$. Is there convincing evidence that policemen exposed to constant auto exhaust have elevated blood lead level concentrations? (Data source: Kamal, Eldamaty, and Faris, "Blood lead level of Cairo traffic policemen," Science of the Total Environment, 105(1991): 165-170.)

Solution. Let's try to answer the question by calculating a 95% confidence interval for the population mean. For a 95% confidence interval, $1-\alpha = 0.95$, so that $\alpha = 0.05$ and $\alpha/2 = 0.025$. Therefore, as the following diagram illustrates the situation, $z_{0.025} = 1.96$:



Now, substituting in what we know ($\bar{x}$ = 29.2, n = 126, σ = 7.5, and $z_{0.025}$ = 1.96) into the formula for a *Z*-interval for a mean, we get:

$$[29.2 - 1.96(7.5/\sqrt{126}), 29.2 + 1.96(7.5/\sqrt{126})]$$

Simplifying, we get a 95% confidence interval for the mean blood lead level concentration of all policemen exposed to constant auto exhaust:

$$[27.89, 30.51]$$

That is, we can be 95% confident that the mean blood lead level concentration of all policemen exposed to constant auto exhaust is between 27.9 $\mu g/dl$ and 30.5 $\mu g/dl$. Note that the interval does not contain the value 18.2, the average blood lead level concentration of humans with no exposure to automobile exhaust. In fact, all of the values in the confidence interval are much greater than 18.2. Therefore, there is convincing evidence that policemen exposed to constant auto exhaust have elevated blood lead level concentrations.

2.2 A t-Interval for a Mean

So far, we have shown that the formula:

$$\bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

is appropriate for finding a confidence interval for a population mean if two conditions are met:

(1) The population standard deviation σ is known, and

(2) $X_1, X_2, ..., X_n$ are normally distributed. (The truth is that $X_1, X_2, ..., X_n$ need not be normally distributed as long as the sample size n is large enough for the Central Limit Theorem to apply. In this case, the confidence interval is an *approximate* confidence interval.)

Now, as suggested earlier in this section, it is unrealistic to think that we'd ever be in a situation where condition (1) would be met. That is, when would we ever know the population standard deviation σ, but not the population mean μ? Let's entertain, then, the realistic situation in which not only the population mean μ is unknown, but also the population standard deviation σ is unknown.

What if σ is unknown?

Yes, the reasonable thing to do is to estimate the population standard deviation σ with the sample standard deviation:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Then, in deriving the confidence interval, we'd start out with:

$$\frac{\bar{X} - \mu}{S / \sqrt{n}}$$

instead of:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Then, to derive the confidence interval, in this case, we just need to know how:

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

is distributed!

How is $T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$ distributed?

Given that the ratio is typically denoted by the capital letter *T*, we probably shouldn't be surprised that the ratio follows a *T* distribution!

Theorem. If X_1, X_2, ..., X_n are normally distributed with mean μ and variance σ^2, then:

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

follows a *T* distribution with $n - 1$ degrees of freedom.

Now that we have the distribution of $T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$ behind us, we can derive the confidence interval for a population mean in the realistic situation that σ is unknown.

Proof. The proof is as simple as recalling a few distributional results from our work in Stat 414. Recall the definition of a T random variable, namely if $Z \sim N(0,1)$ and $U \sim \chi^2_{(r)}$ are independent, then:

$$T = \frac{Z}{\sqrt{U/r}}$$

follows the T distribution with r degrees of freedom. Furthermore, recall that if $X_1, X_2, ..., X_n$ are normally distributed with mean μ and variance σ^2, then:

(1) $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

(2) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

(3) $\bar{X}$ and S^2 are independent

Now, we just have to put all that we've remembered together:

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

The first equality simply defines a T random variable using (1), (2), and (3) above. The second equality comes from canceling out the $(n-1)$ terms in the denominator. The third equality comes from canceling out the σ terms, leaving us with:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

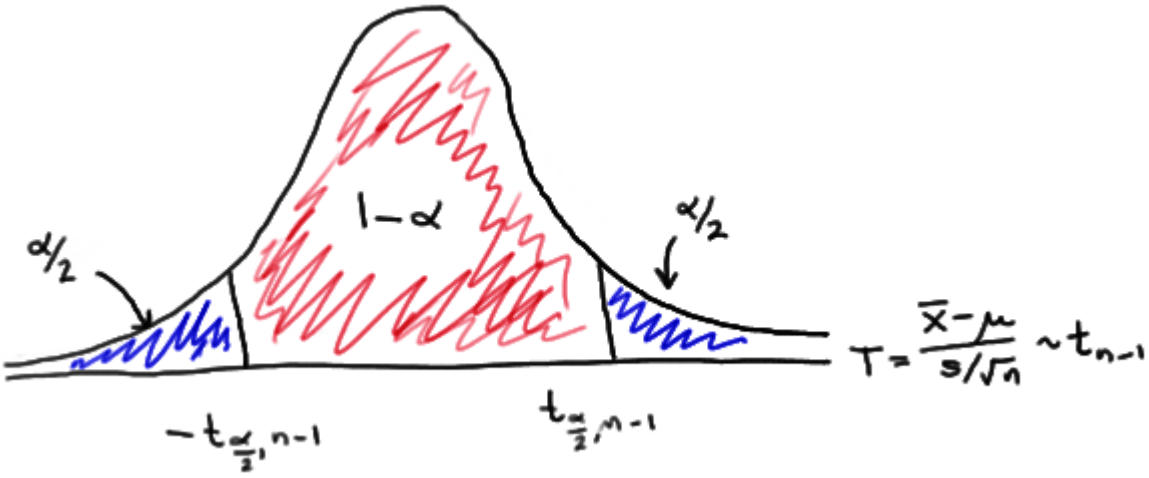
following a T distribution with $n - 1$ degrees of freedom, as was to be proved!

Theorem. If $X_1, X_2, ..., X_n$ are normally distributed random variables with mean μ and variance σ^2, then a $(1-\alpha)100\%$ confidence interval for the population mean μ is:

$$\bar{X} \pm t_{\alpha/2}(S\sqrt{n})$$

This interval is often referred to as the "***t*-interval for the mean**."

The proof is very similar to that for the Z-interval for the mean. We start by drawing a picture of a T-distribution with $n - 1$ degrees of freedom:



From the diagram, we can see that the following probability statement is true:

$$P[-t_{\alpha/2,n-1} \leq T \leq t_{\alpha/2,n-1}] = 1-\alpha$$

Then, simply replacing T, we get:

$$P\left[-t_{\alpha/2,n-1} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}\right] = 1-\alpha$$

Let's again focus only on the inequality inside the brackets for a bit. Because we manipulate each of the three sides of the inequality equally, each of the following statements are equivalent:

$$\begin{aligned}
-t_{\alpha/2,n-1} &\leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1} \\
-t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) &\leq \bar{X}-\mu \leq +t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \\
-\bar{X}-t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) &\leq -\mu \leq -\bar{X}+t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \\
\bar{X}-t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) &\leq \mu \leq \bar{X}+t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)
\end{aligned}$$

That is, we have shown that a $(1-\alpha)100\%$ confidence interval for the mean μ is:

$$\left[\bar{X}-t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right), \bar{X}+t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)\right]$$

as was to be proved.

Definition. With the formula for the *t*-interval:

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

in mind, we say that:

(1) $\bar{x}$ is a "**point estimate**" of μ

(2) $\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$ is an "**interval estimate**" of μ

(3) $\frac{s}{\sqrt{n}}$ is the "**standard error of the mean**"

(4) $t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$ is the "**margin of error**"

Now, let's take a look at an example!

Example

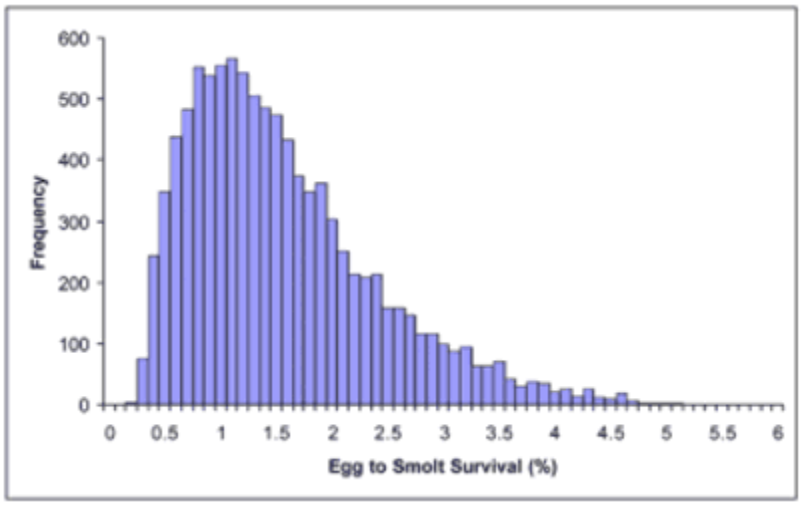
A random sample of 16 Americans yielded the following data on the number of pounds of beef consumed per year:

118 115 125 110 112 130 117 112
115 120 113 118 119 122 123 126

What is the average number of pounds of beef consumed each year per person in the United States?

Exercise 1. a) Show that the data follow normal distribution, since according to the above theorem states, in order for the *t*-interval for the mean to be appropriate, the data must follow a normal distribution, and b) calculate a 95% confidence interval for the mean.

Non-normal Data



So far, all of our discussion has been on finding a confidence interval for the population mean μ when the data are normally distributed. That is, the *t*-interval for μ (and *Z*-interval, for that matter) is derived assuming that the data X_1, X_2, ..., X_n are normally distributed. What happens if our data are skewed, and therefore clearly not normally distributed?

Well, it is helpful to note that as the sample size *n* increases, the *T* ratio:

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

approaches an approximate normal distribution regardless of the distribution of the original data. The implication, therefore, is that the *t*-interval for μ:

$$\bar{x} \pm t_{a/2,n-1}(\frac{s}{\sqrt{n}})$$

and the *Z*-interval for μ:

$$\bar{x} \pm t_{a/2}(\frac{s}{\sqrt{n}})$$

(with the sample standard deviation *s* replacing the unknown population standard deviation σ!) yield similar results for large samples. This result suggests that we should adhere to the following guidelines in practice.

In practice!

(1) Use $\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$ if the data are normally distributed.

(2) If you have reason to believe that the data are not normally distributed, then make sure you have a large enough sample ($n \geq 30$ generally suffices, but recall that it depends on the skewness of the distribution.) Then:

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \text{ and } \bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

will give similar results.

(3) If the data are not normally distributed *and* you have a small sample, use:

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Exercise 2

A random sample of 64 guinea pigs yielded the following survival times (in days):

36	18	91	89	87	86	52	50	149	120
119	118	115	114	114	108	102	189	178	173
167	167	166	165	160	216	212	209	292	279
278	273	341	382	380	367	355	446	432	421
421	474	463	455	546	545	505	590	576	569
641	638	637	634	621	608	607	603	688	685
663	650	735	725						

What is the mean survival time (in days) of the population of guinea pigs? (Data from K. Doksum, *Annals of Statistics*, 2(1974): 267-277.). Find the confidence interval that contains the mean survival time for the population of guinea pigs with 95% confident.

Exercise 3

For the population generated in section 1 of this document, a) create the sample of size 1000, b) compute the z-critical value using the qnorm R function, and compute the confidence interval using the z-value.

The following code generates 25 confidence intervals and plots them. Discuss what you observed.

```
set.seed(12) sample_size <- 1000
```

```
intervals <- c()  # Create and store 25 intervals   for (sample in 1:25){ sample_ages <-
sample(population_ages, size=sample_size)  # Take a sample of 1000 ages

sample_mean <- mean(sample_ages)  # Get the sample mean

z_critical <- qnorm(0.975)       # Get the z-critical value*

pop_stdev <- sd(population_ages)  # Get the population standard deviation

margin_of_error <- z_critical * (pop_stdev / sqrt(sample_size)) # Calculate margin of er ror

confidence_interval  <- c(sample_mean - margin_of_error,  # Calculate the the interval
sample_mean + margin_of_error)

intervals <- c(intervals, confidence_interval)    }

interval_df <- data.frame(t(matrix(intervals,2,25)))  # Store intervals as data frame

library(ggplot2)

# Plot confidence intervals and show the true mean

my_plot <- ggplot(data=interval_df, aes(x=1:nrow(interval_df))) +

geom_errorbar(aes(ymax = X2, ymin = X1)) +              geom_point(aes(y=rowMeans(interval_df)),
shape=1, size=3) +              geom_abline(intercept=true_mean, slope=0,color="red",lwd=1) +

          ylab("Interval Range (Red Line=True Mean)") +    xlab("Interval Number")
```

Exercise 4

Take a new, smaller sample from the population used in exercise 3 and then create a confidence interval without the population standard deviation, using the t-distribution. Note: when using the t-distribution, you have to supply the degrees of freedom (df). For this type of one test, the degrees of freedom is equal to the sample size minus 1. If you have a large sample size, the t-distribution approaches the normal distribution. Compare the t-interval with z-interval. Use the t.test(sample) function to calculate the t-interval.

Exercise 5

We can also make a confidence interval for a point estimate of a population proportion. In this case, the margin of error equals:

$$z * \sqrt{\frac{p(1-p)}{n}}$$

Where z is the z-critical value for our confidence level, p is the point estimate of the population proportion and n is the sample size. Calculate a 95% confidence interval for Hispanics according to the sample proportion 0.204.

Note: As with the confidence interval for the mean, you can use a built in R function to get a confidence interval:

```
prop.test(x=204,     # Number of observations        n=1000)    # Total number of samples
```

Apply this function and report the results.