Introduction to Statistical Inference

Felipe José Bravo Márquez

April 19, 2021

Populations and Samples

- A **population** is the entire group of individuals that we are interested in studying.
- This could be anything from all humans to a specific type of cell.
- The main goal of statistical inference is investigate properties about a target **population**.
- Example: What is the average height of all people in Chile? Here the population is all the inhabitants of Chile.
- In order to draw conclusions about a **population**, it is generally not feasible to gather all the data about it.
- The special case where you collect data on the entire population is a **census**.

Populations and Samples

- In statisical inference we try to make reasonable conclusions about a population based on the evidence provided by **sample data**.
- We do this primarily to save time and effort.
- A **sample staticic** or simply **statistic** is a quantitative measure calculated from a sample. Examples: the mean, the standard deviation, the minimum, the maximum.
- Our goal in sampling is to determine the value of a statistic for an entire population of interest, using just a small subset of the population.

Samples and Surveys

- Random samples
- Stratified samples
- Biases

Statistical Inference

- The process of drawing conclusions about a population from sample data is known as **statistical inference**.
- From a general point of view, the goal of inference is to **infer** the distribution that generates the observed data.
- Example: Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?
- However, in most cases we are only interested in inferring some property of F (e.g., its **mean** value).
- Statistical models that assume that the distribution can be modeled with a finite set of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ are called **parametric models**.
- Example: if we assume that the data comes from a normal distribution $N(\mu, \sigma^2)$, μ and σ would be the parameters of the model.

Frequentist Aproaches

The satistical methods to be presented is this class are known as **frequentist (or classical)** methods. They are based on the following postulates [Wasserman, 2013]:

- Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.
- Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
- Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

There is another approach to inference called **Bayesian inference**, which is based on different posulates, to be discussed later in the course.

Point Estimation

- Point estimation is the process of finding the **best guess** for some quantity of interest from a **statistical sample**.
- In a general sence, this quantity of interest could be a parameter in a parametric model, a CDF F, a probability density function f, a regression function r, or a prediction for a future value Y of some random variable.
- In this class we will consider this quantity of interest as a **population parameter** θ.
- By convention, we denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$.
- It is important to remark that while θ is an unknown fixed value, $\hat{\theta}$ depends on the sample data and is therefore a random variable.
- We need to bear in mind that the process of sampling is by definition a **random experiment**.

Point Estimation

Formal Definition

- Let $X_1, \ldots, X_n$ be n IID data points from some distribution F.
- A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

- The **bias** of an estimator is defined as:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

- An estimator is unbiased if $\mathbb{E}(\hat{\theta}_n) = \theta$ or $\text{bias}(\hat{\theta}_n) = 0$

Sampling Distribution

- If we take multiple samples, the value of our statistical estimate $\hat{\theta}_n$ will also vary from sample to sample.
- We refer to this distribution of our estimator across samples as the **sampling distribution** [Poldrack, 2019].
- The sampling distribution may be considered as the distribution of $\hat{\theta}_n$ for all possible samples from the same population of size n[1].
- The sampling distribution describes the variability of the point estimate around the true population parameter from sample to sample.
- We need to bear in mind this is an imaginary concept, since in real sitations we can't obtain all possible samples.
- Actually, in most cases we will only work with a single sample.

[1] https://courses.lumenlearning.com/boundless-statistics/chapter/sampling-distributions/

Standard Error

- The standard deviation of $\hat{\theta}_n$ is called the **standard error** *se*:

$$se(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}$$

- The standard error tells us about the variability of the estimator between all possible samples of the same size.
- It can be think of as the standard deviation of the sampling distribution.
- It is a measure of the uncertainty of the point estimate.

The Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be a random sample of a population of mean μ and variance σ^2.
- Let's suppose that we are interested in estimating the **population mean** μ (e.g., the mean height of Chilean people).
- A sample statistic we can derive from the data is the **sample mean** $\overline{X_n}$

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The sample mean is a **point estimator** of the mean $\overline{X_n} = \hat{\mu}$.
- We can show that the sample mean is an unbiased estimator of μ:

$$\mathbb{E}(\overline{X_n}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n} \times \mathbb{E}(\sum_{i=1}^{n} X_i) = \frac{1}{n}(n \times \mu) = \mu$$

The Standard Error of the Sample Mean

- The standard error of the sample mean $se(\overline{X_n}) = \sqrt{\mathbb{V}(\overline{X_n})}$ can be calulated as:

$$\mathbb{V}(\overline{X_n}) = \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathbb{V}(\sum_{i=1}^{n} X_i) = \frac{n}{n^2}\mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Then,

$$se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$$

- The formula for the standard error of the mean implies that the quality of our measurement involves two quantities: the population variability σ, and the size of our sample n.

The Standard Error of the Sample Mean

- We have no control over the population variability, but we do have control over the sample size.
- Thus, if we wish to improve our sample statistics (by reducing their sampling variability) then we should use larger samples.
- However, the formula also tells us something very fundamental about statistical sampling.
- That the utility of larger samples diminishes with the square root of the sample size.
- This means that doubling the sample size will not double the quality of the statistics; rather, it will improve it by a factor of $\sqrt{2}$. [Poldrack, 2019]

Sample Variance

- A common problem when calculating $se(\overline{X_n})$ is that, in general, we do not know σ of the population.
- In those cases we can estimate σ using the **sample variance** s:

$$s^2 = \frac{1}{n-1}\sum_{i}^{n}(X_i - \overline{X_n})^2$$

- This is an unbiased estimator of the variance.
- The standard error of the sample mean when the population variance is unknown can be estimated as follows:

$$\hat{se}(\overline{X_n}) = \frac{s}{\sqrt{n}}$$

Population Variance

- There is also the population variance, defined as follows:

$$\sigma^2 = \frac{1}{N}\sum_{i}^{n}(X_i - \overline{X_N})^2$$

- The population variance should only be calculated from population data (all the individuals).
- Note that we are using *N* instead of *n* to denote the entire population rather than a sample.
- If is calculated from a sample, it would be a **biased** estimator of the population variance.

The Sampling Distribution of the Sample Mean

- We discussed earlier that the sampling distribution is an imaginary concept.
- Let's imagine the sampling distribution of the sample mean.
- Imagine drawing (with replacement) all possible samples of size n from a population.
- Then for each sample, calculate the sample statistic, which is this case is the sample mean.
- The frequency distribution of those sample means would be the sampling distribution of the mean (for samples of size n drawn from that particular population).
- In the next example we will calculate the sampling distribution for a toy example in which the population is known.

The Sampling Distribution of the Sample Mean

- Suppose our entire population is a family of 5 siblings and our property of interest is age measured in years.
- Our population consists of the following 5 values: 2, 3, 4, 5, and 6.
- Let's calculate the population mean μ and the population standard deviation σ.

```
> pop <-c(2,3,4,5,6)
> mean(pop)
[1] 4
> sd.p=function(x){sd(x)*sqrt((length(x)-1)/length(x))}
> sd.p(pop)
[1] 1.414214
```

μ=4 and $\sigma = 1.414214$

The Sampling Distribution of the Sample Mean

- Now, we will use the R library “gtools” to draw all 25 possible samples (with replacement) of size 2.

```
> library(gtools)
> library(tidyverse)
> samp_size <- 2
> samples<-as_tibble(permutations(length(pop), samp_size,
+                                 pop, repeats.allowed=TRUE))
> samples
# A tibble: 25 x 2
      V1    V2
   <dbl> <dbl>
 1     2     2
 2     2     3
 3     2     4
 4     2     5
 5     2     6
 6     3     2
 7     3     3
 8     3     4
 9     3     5
10     3     6
# ... with 15 more rows
```

The Sampling Distribution of the Sample Mean

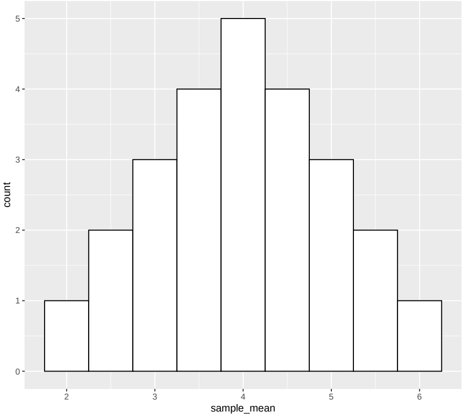
- We can calculate the sample mean of each sample using the command “mutate”:

```
> samples <- samples %>% rowwise() %>%
+   mutate(sample_mean=mean(c(V1,V2)))
> samples
# A tibble: 25 x 3
# Rowwise:
      V1    V2 sample_mean
   <dbl> <dbl>       <dbl>
 1     2     2         2
 2     2     3         2.5
 3     2     4         3
 4     2     5         3.5
 5     2     6         4
 6     3     2         2.5
 7     3     3         3
 8     3     4         3.5
 9     3     5         4
10     3     6         4.5
# ... with 15 more rows
```

The Sampling Distribution of the Sample Mean

- The distribution of these sample means is the **sampling distributiion**.
- We can visualize its shape by plotting an histogram:

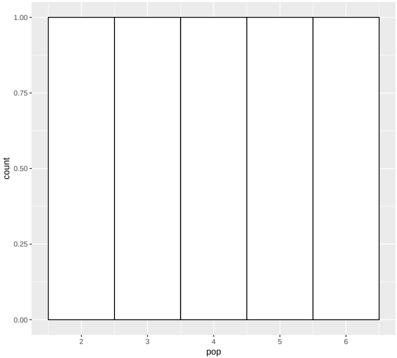
```
ggplot(samples, aes(x=sample_mean)) +
  geom_histogram(bins = 10, color="black", fill="white")
```



The Sampling Distribution of the Sample Mean

- You may noticed that the historgram is peaked in the middle, and symmetrical.
- This is a consequence of the Central Limit Theorem!!!
- We can see that the population distribution is very different from the sampling distribution:

```
ggplot(data.frame(pop), aes(x=pop)) +
  geom_histogram(bins = 5, color="black", fill="white")
```



The Sampling Distribution of the Sample Mean

- Let's calculate the mean and the standard deviation of the sample means:

  ```
  > mean(samples$sample_mean)
  [1] 4
  > sd.p(samples$sample_mean)
  [1] 1
  ```

- We can see that mean of the sampling distribution of the mean $\mu_{\overline{X}}$ equals the population mean μ.
- We can also calculate the theoretical standard error $se = \sigma/\sqrt{n}$

  ```
  > sd.p(pop)/sqrt(samp_size)
  [1] 1
  ```

  which is the same as the standard distribution of the sampling distribution of the sample mean.
- We have validated empirically that the sample mean is a good estimator of the population mean and that its standard error can be calculated from the population standard deviation and the sample size.

The Sampling Distribution of the Sample Mean

- The central limit theorem tell us the conditions under which the sampling distribution of the mean is normally distributed or at least approximately normal.
- If the population from which you sample is itself normally distributed, then the sampling distribution of the mean will be normal, regardless of sample size.
- If the population from which you sample is non-normal, the sampling distribution of the mean will still be approximately normal given a large enough sample size.
- What size is sufficient? Some authors say 30 or 40. But if the population distribution is extremely non-normal (i.e. very skewed) you will need more.

Point Estimation of a Proportion

- Suppose we want to estimate the fraction of people who will vote for a certain candidate.
- Our population parameter p corresponds to the true fraction of voters for this candidate.
- We can model a sample of independent voters $X_1, \dots, X_n$, as Bernoulli distributed random variables with parameter p.
- We interpret $X_i = 0$ as a negative vote and $X_i = 1$ as a positive vote.
- The sample proportion $\hat{p}_n = \frac{1}{n}\sum_i X_i$ is our estimator of p.

Point Estimation of a Proportion

- Then $\mathbb{E}(\hat{p}_n) = \frac{1}{n}\sum_i \mathbb{E}(X_i) = p$, and $\hat{p}_n$ is unbiased.
- The standard error *se* would be

$$se = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{p(1-p)/n}$$

- The estimated standard error $\hat{se}$:

$$\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$$

- By the Central Limit Theorem the sampling distribution of the sample proportion converges to a Normal distribution: $\hat{p}_n \approx N(p, \hat{se}^2)$.
- This is because the sample proportion is actually the sample mean of a binary population.

Consistency

- A good estimator is expected to be unbiased and of minimum standard error.
- Unbiasedness used to receive much attention but these days is considered less important
- Many of the estimators we will use are biased.
- A reasonable requirement for an estimator is that it should converge to the true parameter value as we collect more and more data.
- A point estimator $\hat{\theta}_n$ of a parameter θ is **consistent** if it converges to the true value when the number of data in the sample tends to infinity and its standard error converges to zero.

Consistency

- If for an estimator $\hat{\theta}_n$, its *bias* $\rightarrow 0$ and its *se* $\rightarrow 0$ when $n \rightarrow \infty$, $\hat{\theta}_n$, it is a consistent estimator of θ.
- For example, for the sample mean $\mathbb{E}(\overline{X_n}) = \mu$, which implies that the *bias* $= 0$.
- Then $se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$ converges to zero when $n \rightarrow \infty$.
- $\overline{X_n}$ is a consistent estimator of the mean.
- For the case of the Bernoulli experiment one has that $\mathbb{E}(\hat{p}) = p \Rightarrow$ *bias* $= 0$ and $se = \sqrt{p(1-p)/n} \rightarrow 0$ when $n \rightarrow \infty$.
- Then $\hat{p}$ is a consistent estimator of p.

Maximum Likelihood Estimation

Confidence Interval

- We know that the value of a point estimator **varies** from sample to sample.
- It is more reasonable to find an interval that is likely to trap the real value of the parameter with a certain probability.
- The general form of a confidence interval in the following:

$$\text{Confidence Interval} = \text{Sample Statistic} \ \pm \ \text{Margin Error}$$

- The wider the interval the more uncertainty there is about the value of the parameter.

Confidence Interval

Definition

- A **confidence interval** for an unknown population parameter θ with a **confidence level** $1 - \alpha$, is an interval $C_n = (a, b)$ where:

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha$$

- In addition $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data.
- The α value is known as the **significance** level, generally taken as 0.05, which is equivalent to working with a confidence level of 95%.
- Significance can be interpreted as the probability of being wrong.

Confidence Interval

- There is a lot of **confusion** about how to interpret a confidence interval.
- A confidence interval is not a probability statement about θ since θ is a fixed quantity in Frequentist inference setting, not a random variable
- One way to interpret them is to say that if we repeat the **same experiment** many times, the interval will contain the value of the parameter $(1 - \alpha)\%$ of the times.
- This interpretation is correct, but we rarely repeat the same experiment several times.
- A better interpretation: one day I collect data I create a 95% confidence interval for a parameter θ_1. Then on day 2, I do the same for a parameter θ_2 and so repeatedly *n* times. The 95% of my intervals will contain the actual values of the parameters.

Confidence Interval

- Later in the course, we will discuss Bayesian methods in which we treat θ as if it were a random variable and we do make probability statements about θ.
- In particular, we will make statements like “the probability that θ is in C_n, given the data, is 95 percent.”
- However, these Bayesian intervals refer to degree-of-belief probabilities.
- These Bayesian intervals will not, in general, trap the parameter 95 percent of the time.

Confidence Interval of the Mean

- We have n independent observations $X_1, \ldots, X_n$ (IID) of distribution $N(\mu, \sigma^2)$.
- Suppose μ is **unknown** but σ^2 is **known**.
- We know that $\overline{X_n}$ is an unbiased estimator of μ.
- By the law of large numbers we know that the distribution of $\overline{X_n}$ is concentrated around μ when n is large.
- By the CLT we know that

$$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

when n is large.

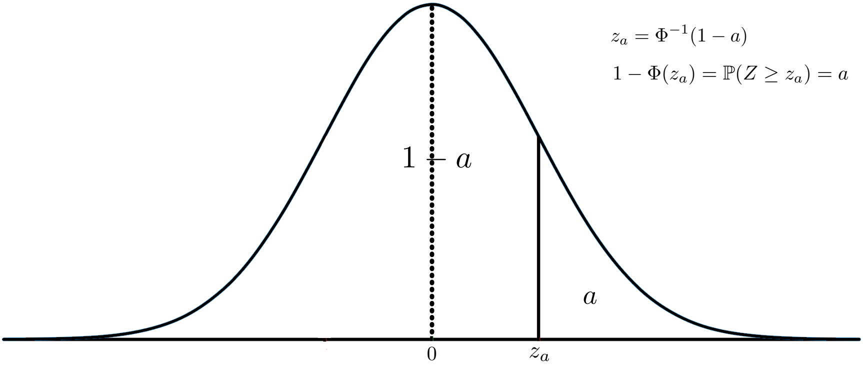
Confidence Interval

- We want to find an interval $C_n = (\mu_1, \mu_2)$ with confidence level $1 - \alpha$:

$$\mathbb{P}(\mu_1 \leq \mu \leq \mu_2) = 1 - \alpha$$

- Let $z_a = \Phi^{-1}(1 - a)$, with $a \in [0, 1]$ where Φ^{-1} is the quantile function of a standardized normal.
- This is equivalent to saying that z_a is the value such that $1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$.
- By symmetry of the normal distribution: $z_{\alpha/2} = -z_{(1-\alpha/2)}$.

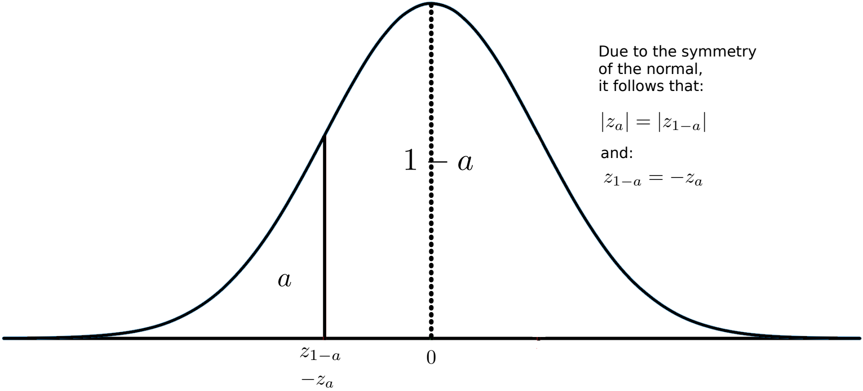
Confidence Interval



$z_a = \Phi^{-1}(1 - a)$

$1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$

Confidence Interval



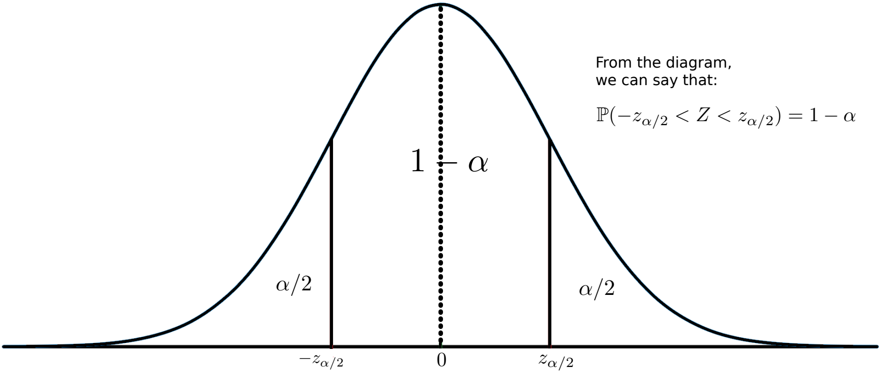
Due to the symmetry
of the normal,
it follows that:

$$|z_a| = |z_{1-a}|$$

and:

$$z_{1-a} = -z_a$$

Confidence Interval



From the diagram,
we can say that:

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$$

Confidence Interval

- The confidence interval for μ is:

$$C_n = (\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- Then $z_{\alpha/2}$ tells us how many times we have to multiply the **standard error** to build the interval.
- The smaller the value of α the larger the value of $z_{\alpha/2}$ and hence the wider the interval.
- Proof:

$$\begin{aligned}
\mathbb{P}(\mu \in C_n) &= \mathbb{P}(\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\
&= \mathbb{P}(-z_{\alpha/2} < \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) \\
&= \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= 1 - \alpha
\end{aligned}$$

Confidence Interval

- Since $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ we can use the quantile function of the normal to calculate confidence intervals in R.

```
> alpha <- 0.05
> xbar <- 5
> sigma <- 2
> n <- 20
> se <-sigma/sqrt(n)
> error <- qnorm(1-alpha/2)*se
> left <- xbar-error
> right <- xbar+error
> left
[1] 4.123477
> right
[1] 5.876523
>
```

T Distribution

- In practice, if we do not know μ we are unlikely to know σ.
- If we estimate σ using s, confidence intervals are better build using the distribution **T-student**, especially when the sample size is small.

T Distribution

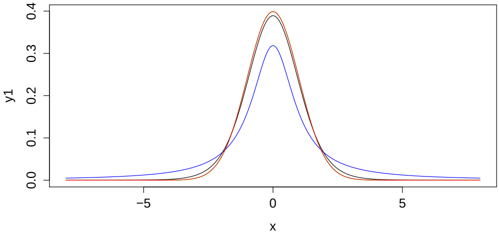
- An R.V. has distribution t with k degrees of freedom when it has the following PDF:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1+\frac{t^2}{k})^{(k+1)/2}}$$

- When $k = 1$ it is called **Cauchy** distribution.
- When $k \to \infty$ it converges to a standardized normal distribution.
- The t-distribution has wider tails than the normal distribution when it has few degrees of freedom.

T Distribution

```
x<-seq(-8,8,length=400)
y1<-dnorm(x)
y2<-dt(x=x,df=1)
y3<-dt(x=x,df=10)
y4<-dt(x=x,df=350)
plot(y1~x,type="l",col="green")
lines(y2~x,type="l",col="blue")
lines(y3~x,type="l",col="black")
lines(y4~x,type="l",col="red")
```



T-Distribution Confidence Interval

- Let $s^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X_n})^2$ we have:

$$T = \frac{\overline{X_n} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- Let $t_{n-1,a} = \mathbb{P}(T > a)$, equivalent to the quantile function *qt* evaluated at $(1 - a)$.
- The resulting confidence interval is:

$$C_n = (\overline{X_n} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \overline{X_n} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}})$$

- Since the tails of the *t* distribution are wider when *n* is small, the resulting confidence intervals are wider.

T-Distribution Confidence Interval

- Let's calculate a confidence interval for the mean of `Petal.Length` of the **Iris** data with 95% confidence.

```
>data(iris)
>alpha<-0.05
>n<-length(iris$Petal.Length)
>xbar<-mean(iris$Petal.Length)
>xbar
[1] 3.758
>s<-sd(iris$Petal.Length)
>se<-s/sqrt(n)
>error<-qt(p=1-alpha/2,df=n-1)*se
>left<-xbar-error
>left
[1] 3.473185
>right<-xbar+error
>right
[1] 4.042815
```

- Another way:

```
>test<-t.test(iris$Petal.Length,conf.level=0.95)
>test$conf.int
[1] 3.473185 4.042815
```

The Boostrap

References I

Poldrack, R. A. (2019).
Statistical Thinking for the 21st Century.

Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.