

# Introduction to Statistical Inference

Felipe José Bravo Márquez

April 20, 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 statistical 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 statistic** 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  $X_1, \dots, X_n \sim F$ , 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  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  are called **parametric models**.
- Example: if we assume that the data comes from a normal distribution  $N(\mu, \sigma^2)$ ,  $\mu$  and  $\sigma$  would be the parameters of the model.

# Frequentist Approaches

The statistical methods to be presented in 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 postulates, 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 sense, 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**  $\theta$ .
- By convention, we denote a point estimate of  $\theta$  by  $\hat{\theta}$  or  $\hat{\theta}_n$ .
- It is important to remark that while  $\theta$  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  $X_1, \dots, X_n$  be  $n$  IID data points from some distribution  $F$ .
- A point estimator  $\hat{\theta}_n$  of a parameter  $\theta$  is some function of  $X_1, \dots, X_n$ :

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

- 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  $n$ <sup>1</sup>.
- 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 situations we can't obtain all possible samples.
- Actually, in most cases we will only work with a single sample.

---

<sup>1</sup><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  $X_1, X_2, \dots, X_n$  be a random sample of a population of mean  $\mu$  and variance  $\sigma^2$ .
- Let's suppose that we are interested in estimating the **population mean**  $\mu$  (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  $\mu$ :

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \times \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}(n \times \mu) = \mu$$

# The Standard Error of the Sample Mean

- The standard error of the sample mean  $se(\bar{X}_n) = \sqrt{\mathbb{V}(\bar{X}_n)}$  can be calculated as:

$$\mathbb{V}(\bar{X}_n) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \mathbb{V}\left(\sum_{i=1}^n X_i\right) = \frac{n}{n^2} \mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Then,

$$se(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

- The formula for the standard error of the mean implies that the quality of our measurement involves two quantities: the population variability  $\sigma$ , 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  $\sigma$  of the population.
- In those cases we can estimate  $\sigma$  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 in 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  $\mu$  and the population standard deviation  $\sigma$ .

```
> 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
```

$\mu=4$  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

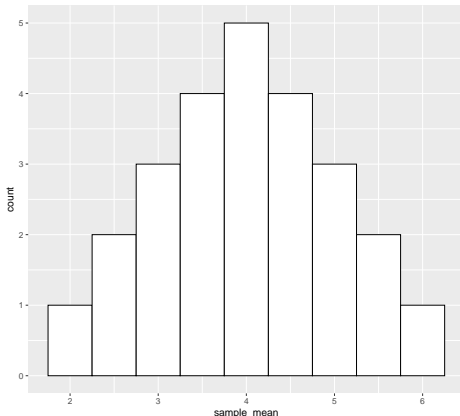
- We can calculate the sample mean of each sample using the command “mutate”:

```
> 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 distribution**.
- We can visualize its shape by plotting an histogram:

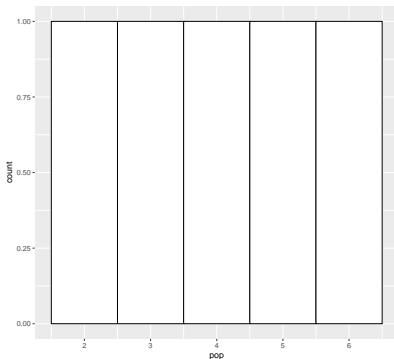
```
ggplot(samples, aes(x=sample_mean)) +  
  geom_histogram(bins = 10, color="black", fill="white")
```



# The Sampling Distribution of the Sample Mean

- You may noticed that the histogram 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  $\mu_{\bar{X}}$  equals the population mean  $\mu$ .
- 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 deviation 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  $\theta$  is **consistent** if it converges to the true value when the number of data in the sample tends to infinity.

# Consistency

- Theorem: If for an estimator  $\hat{\theta}_n$ , its *bias*  $\rightarrow 0$  and its *se*  $\rightarrow 0$  when  $n \rightarrow \infty$ ,  $\hat{\theta}_n$  is a consistent estimator of  $\theta$ .
- 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

- The estimators we have presented so far (e.g., the sample mean, the sample proportion) are intuitive, easy to compute, and consistent.
- Maximum Likelihood Estimation (MLE) is a more general framework for estimating the **parameters** of any **parametric model**.
- In MLE, we assume that the sample data is generated by a given probability distribution (continuous or discrete) parameterized by  $\theta$  and try to find the value of  $\theta$  that maximizes the joint probability of the data under that distribution.
- Idea: find the parameter values of the assumed statistical model that make the observed data most probable.
- For example, we can assume that each data point is generated by  $N(\mu, \sigma^2)$ , then we will compute the joint PDF of our data and find the parameter values for  $\mu$  and  $\sigma$  that maximize that joint density.

# Maximum Likelihood Estimation

- Let  $X_1, \dots, X_n$  be IID with PDF  $f(x; \theta)$ .
- Since we are assuming that our data samples are independent random variables, the joint density (PDF) would be the product of each PDF:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i; \theta)$$

- We refer to this joint density as the **likelihood function**.
- The likelihood function is just the joint density of the data, except that we treat it is a function of the parameter  $\theta$ .
- In MLE, we turn the estimation task into an optimization problem:

$$\max_{\theta} \mathcal{L}_n(\theta) \tag{1}$$

- The maximum likelihood estimator MLE, denoted by  $\hat{\theta}_n$ , is the value of  $\theta$  that maximizes  $\mathcal{L}_n(\theta)$ .

# Maximum Likelihood Estimation

- In many cases the log-likelihood is easier to optimize  $l_n(\theta)$ :

$$l_n(\theta) = \log(\mathcal{L}_n(\theta)) = \log\left(\prod_{i=1}^n f(X_i; \theta)\right) = \sum_{i=1}^n \log(f(X_i; \theta))$$

- Since the logarithm is a monotonic function, the maximum of occurs at the same value of  $\theta$ .
- If the  $l_n$  is differentiable we can find  $\hat{\theta}_n$  by setting the derivatives to zero:

$$\frac{\partial l_n}{\partial \theta} = 0$$

- The MLE has many good mathematical properties that go beyond the scope of this course to discuss.
- One property that is important to know is that the MLE is **consistent**.

# Maximum Likelihood Estimation

- Example 1: Suppose that  $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ .
- The probability mass function is  $f(x; p) = p^x(1 - p)^{1-x}$  for  $x = 0, 1$ . The unknown parameter is  $p$ .
- Then,

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i; p) = \prod_{i=1}^n p^{X_i}(1 - p)^{1-X_i} = p^S(1 - p)^{n-S}$$

where  $S = \sum_i X_i$ . Hence,

$$l_n(\theta) = S \log p + (n - S) \log(1 - p).$$

- Take the derivative of  $l_n(\theta)$ , set it equal to 0 to find that the MLE is  $\hat{p}_n = S/n$ , which is the same value as the sample proportion shown earlier.

# Maximum Likelihood Estimation

- Example 2: Let  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ .
- The parameter is  $\theta = (\mu, \sigma)$  and the likelihood function (ignoring some constants) is:

$$\mathcal{L}_n(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma} \exp \left( -\frac{1}{2\sigma^2} (X_i - \mu)^2 \right) \quad (2)$$

$$= \frac{1}{\sigma^n} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right) \quad (3)$$

$$= \frac{1}{\sigma^n} \exp \left( -\frac{nS^2}{2\sigma^2} \right) \exp \left( -\frac{n(\bar{X} - \mu)^2}{2\sigma^2} \right) \quad (4)$$

- Where  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is the sample mean and  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ .



# Maximum Likelihood Estimation

- The last equality above follows from the fact that

$$\sum_{i=1}^n (X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$$

which can be verified by writing

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2$$

and then expanding the square.

- The log-likelihood is:

$$l_n(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}$$

- Solving the equations  $\frac{\partial(\mu, \sigma)}{\partial \mu} = 0$  and  $\frac{\partial(\mu, \sigma)}{\partial \sigma} = 0$
- We conclude that  $\hat{\mu} = \bar{X}$  and  $\hat{\sigma} = S$ .
- It can be verified that these are indeed global maxima of the likelihood.

# 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 in the long run.
- 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  $\theta$  with a **confidence level**  $1 - \alpha$ , is an interval  $C_n = (a, b)$  where:

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

- In addition  $a = a(X_1, \dots, X_n)$  and  $b = b(X_1, \dots, X_n)$  are functions of the data.
- The  $\alpha$  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  $\theta$  since  $\theta$  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  $\theta_1$ . Then on day 2, I do the same for a parameter  $\theta_2$  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  $\theta$  as if it were a random variable and we do make probability statements about  $\theta$ .
- In particular, we will make statements like “the probability that  $\theta$  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  $X_1, \dots, X_n$  (IID) of distribution  $N(\mu, \sigma^2)$ .
- Suppose  $\mu$  is **unknown** but  $\sigma^2$  is **known**.
- We know that  $\overline{X}_n$  is an unbiased estimator of  $\mu$ .
- By the law of large numbers we know that the distribution of  $\overline{X}_n$  is concentrated around  $\mu$  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.

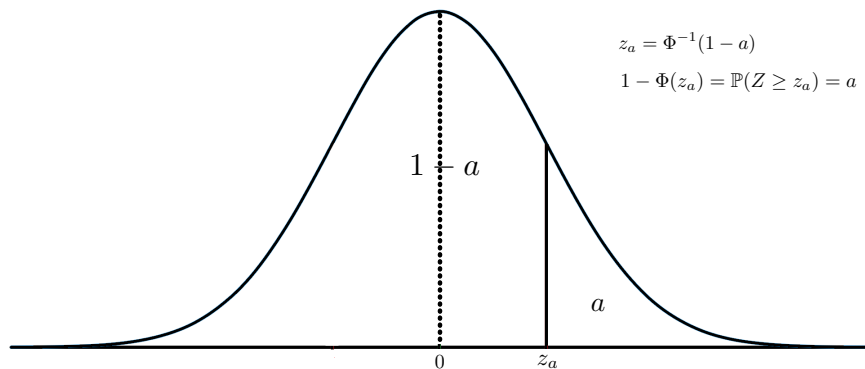
# Confidence Interval

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

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

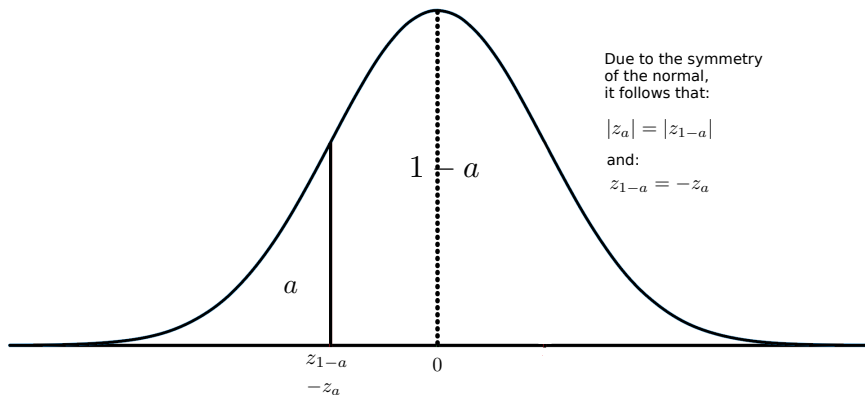
- Let  $z_a = \Phi^{-1}(1 - a)$ , with  $a \in [0, 1]$  where  $\Phi^{-1}$  is the quantile function of a standardized normal.
- This is equivalent to saying that  $z_a$  is the value such that  $1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$ .
- By symmetry of the normal distribution:  $z_{\alpha/2} =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

# Confidence Interval

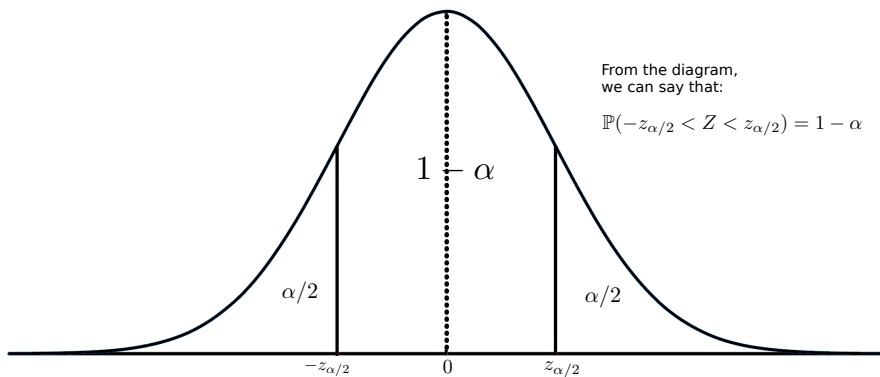




# Confidence Interval



# Confidence Interval



# Confidence Interval

- The confidence interval for  $\mu$  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  $\alpha$  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  $\mu$  we are unlikely to know  $\sigma$ .
- If we estimate  $\sigma$  using  $s$ , confidence intervals are better build using the distribution **T-student**, especially when the sample size is small.

## T Distribution

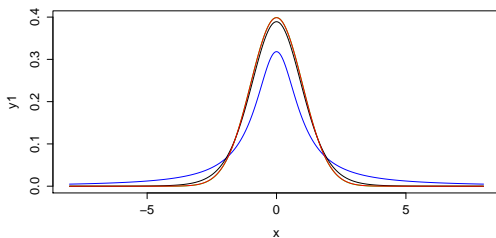
- An R.V. has distribution  $t$  with  $k$  degrees of freedom when it has the following PDF:

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

- When  $k = 1$  it is called **Cauchy** distribution.
- When  $k \rightarrow \infty$  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  $s^2 = \frac{1}{n-1} \sum_i^n (X_i - \bar{X}_n)^2$  we have:

$$T = \frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

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

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

- 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
```



# Confidence Interval for a Population Proportion

- Suppose we want to compute the proportion of subjects who will vote for a candidate and we also want a confidence interval for the estimated proportion.
- As showed earlier, the sampling distribution of a the sample proportion follows a Normal distribution.
- The confidence interval  $C_n$  for a proportion is:

$$C_n = \left( \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)$$

# Confidence Interval for a Population Proportion

- Example: 1,219 respondents indicated that they would vote for candidate A in a survey of 3,532 people.
- Compute a 95% confidence interval for the proportion of voters:

$$\hat{p} = \frac{1219}{3532} = 0.345$$

- and  $z_{\alpha/2} = 1.96$

```
> qnorm(1-0.025)
[1] 1.959964
```

$$C_n = 0.345 \pm 1.96 \sqrt{\frac{0.345(1 - 0.345)}{3532}} = (0.329, 0.361)$$

- In R:

```
> prop<-prop.test(1219, 3532, correct=FALSE)
> prop$conf.int
[1] 0.3296275 0.3609695
```

# The Bootstrap

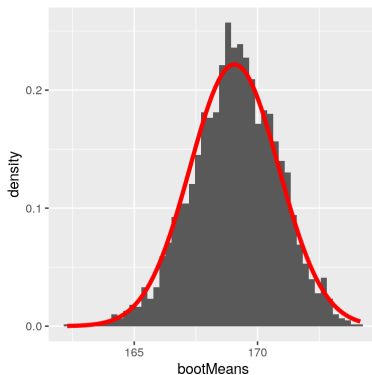
- We have used our knowledge of the sampling distribution of the mean to compute the standard error of the mean.
- But what if we can't assume that the estimates are normally distributed, or we don't know their distribution?
- The idea of the bootstrap is to use the data themselves to estimate an answer.
- The bootstrap method was conceived by Bradley Efron of the Stanford Department of Statistics, who is one of the world's most influential statisticians.
- The idea behind the bootstrap is that we repeatedly sample from the actual dataset.
- We sample with replacement, such that the same data point will often end up being represented multiple times within one of the samples.

# The Bootstrap

- We then compute our statistic of interest on each of the bootstrap samples, and use the distribution of those estimates as our sampling distribution.
- In a sense, we treat our particular sample as the entire population, and then repeatedly sample with replacement to generate our samples for analysis.
- This makes the assumption that our particular sample is an accurate reflection of the population, which is probably reasonable for larger samples but can break down when samples are smaller.
- This technique can be used to estimate the standard error of any statistic and to obtain a confidence interval (CI) for it.

# The Bootstrap

- Let's use the bootstrap to estimate the sampling distribution of the mean:



- The histogram shows the distribution of means across bootstrap samples, while the red line shows the normal distribution based on the sample mean and standard error.

# The Bootstrap

- From the figure, we can see that the distribution of means across bootstrap samples is fairly close to the theoretical estimate based on the assumption of normality.
- It doesn't make much sense to use bootstrap for the sample mean because we know the theoretical shape of the sampling distribution.
- The bootstrap would more often be used to generate standard errors and confidence intervals for estimates of statistics (e.g., the median, the maximum) where we know or suspect that the normal distribution is not appropriate.
- Bootstrap is especially useful when CI doesn't have a closed form, or it has a very complicated one.
- Tutorials:  
<https://www.datacamp.com/community/tutorials/bootstrap-r>,  
<https://github.com/statsthinking21/statsthinking21-R/blob/master/07-ResamplingAndSimulation.Rmd>.

# 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.