Point Estimation. Hypothesis Testing. Confidence Intervals.

Statistics for Data Science

Bruno Damásio

✉ bdamasio@novaims.unl.pt

🐦 @bmpdamasio

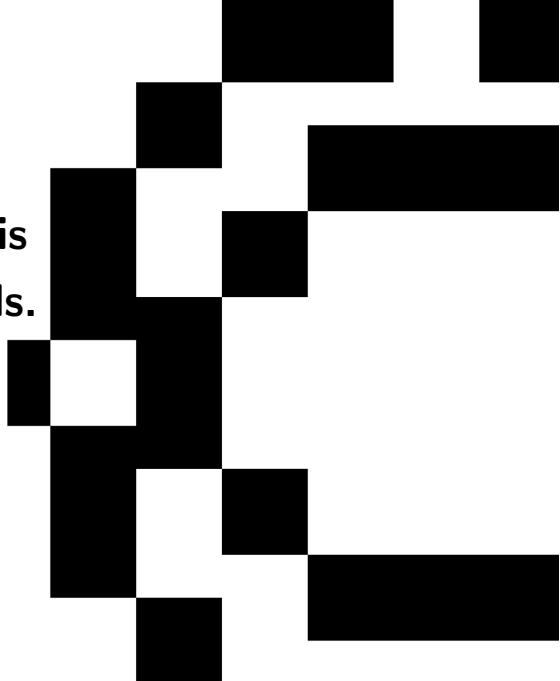2025/2026

Table of contents i

NOVA IMS INFORMATION MANAGEMENT SCHOOL

Point Estimation

Table of Contents

NOVAIMS INFORMATION MANAGEMENT SCHOOL

Point Estimation

Context

Suppose we have an unknown population parameter, such as a population mean μ or a proportion p, which we would like to estimate. For example:

- $p =$ the proportion of Portuguese university students, 18-24, that rent a room during the school year.

- $\mu =$ the mean number of hours that Portuguese university students spent on commute during the school year.

We cannot survey the entire population of Portuguese university students. So, we do what comes naturally and take a random sample from the population, and use the resulting data to estimate the value of the population parameter.

Point Estimation

We will focus on two methods:

- Method of Moments;
- Method of Maximum Likelihood.

Definitions

First, let's start with some formal definitions. Let $X_1, X_2, \ldots, X_n$ be n random variables arising from a random sample. The corresponding observed values of a specific random variable are denoted: $x_1, x_2, \ldots, x_n$.

Parameter Space

The range of possible values of the parameter θ is called the parameter space Ω.

For example, if μ denotes the mean grade of all university students, then the parameter space (assuming a 20-point grading scale) is:

$$\Omega = \{\mu : 0 \leq \mu \leq 20\}$$

Definitions

Point Estimator

The function $X_1, X_2, \ldots, X_n$, that is, the statistic $T(X_1, X_2, \ldots, X_n)$, used to estimate θ is called a point estimator of θ. For example, the function:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a point estimator of the population mean μ.

Definitions

Point Estimate

The function $T(x_1, x_2, \ldots, x_n)$ computed from a set of data is an observed point estimate of θ. For example, if x_i are the observed grade point averages of a sample of 88 students, then:

$$\bar{x} = \frac{1}{88} \sum_{i=1}^{88} x_i = 14.5$$

is a point estimate of μ.

Table of Contents

NOVAIMS INFORMATION MANAGEMENT SCHOOL

Method of Moments

The method of moments involves equating sample moments with theoretical moments.

Definitions

1. $E(X^k)$ is the kth theoretical moment of the distribution (about the origin), for $k = 1, 2, \ldots$
2. $m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$ is the kth sample moment, for $k = 1, 2, \ldots$

Method of Moments

Method of moments estimators are found by equating the first k sample moments to the corresponding k population moments, until we have as many equations as we have parameters. Then, we solve the resulting system of simultaneous equations. More precisely, define

$$m_1 = \frac{1}{n} \sum_{i=1}^{n} X_i^1, \quad \mu_1' = \mathsf{E}\left[X^1\right]$$

$$m_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2, \quad \mu_2' = \mathsf{E}\left[X^2\right]$$

$$\cdots$$

$$m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k, \quad \mu_k' = \mathsf{E}\left[X^k\right]$$

The method of moment estimator is obtained by solving the following system of equations:
$m_1 = \mu_1', \ldots, m_k = \mu_k'$

Method of Moments - Example Bernoulli Distribution

Suppose $X_1, \ldots, X_n$ are i.i.d. that follow a Bernoulli with parameter p. The first theoretical moment is:

$$E(X_i) = p \tag{1}$$

We have just one parameter for which we are trying to derive the method od moments estimator. Therefore, we need just one equation. Equating the first theoretical moment about the origin with the corresponding sample moment we get:

$$p = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2}$$

Method of Moments - Example Normal Distribution

Suppose now that $X_1, \ldots, X_n$ are i.i.d. that follow a Normal(μ, σ^2). We have:

$$
\begin{aligned}
m_1 &= \bar{X} \\
m_2 &= (1/n) \sum X_i^2
\end{aligned}
$$

$$
\begin{aligned}
\mu_1' &= E\left[X^1\right] = \mu \\
\mu_2' &= E\left[X^2\right] = \mu^2 + \sigma^2
\end{aligned}
$$

Thus, we must solve:

$$
\begin{aligned}
\frac{1}{n} \sum X_i &= \mu \\
\frac{1}{n} \sum X_i^2 &= \mu^2 + \sigma^2
\end{aligned}
$$

Method of Moments - Example Normal Distribution

Solving for μ and σ^2 yields the method of moments estimator

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum \left(X_i - \bar{X} \right)^2$$

S²

estimating the pop mean (μ) with the sample mean (X bar)

estimating the pop variance (σ²) with the sample variance (S²)

Table of Contents

NOVAIMS INFORMATION MANAGEMENT SCHOOL

Context

- Suppose we have a random sample $X_1, X_2, \ldots, X_n$, whose assumed probability distribution depends on some unknown parameter θ.

- Our primary goal here will be to find a point estimator, such that is produces a "good" point estimate of θ.

- For example, if we plan to take a random sample for which X_i are assumed to be normally distributed with mean μ and variance σ^2, then our goal will be to find a good estimate of μ, using the data $x_1, x_2, \ldots, x_n$, that we obtained from our specific random sample.

Maximum Likelihood Estimator

- It seems reasonable that a good estimate of the unknown parameter θ would be the value of θ that maximizes the probability (i.e., likelihood) of getting the data we observed.
- The method of maximum likelihood is, by far, the most popular technique for deriving an estimator.
- If $X_1, \ldots, X_n$ are an i.i.d. sample from a population with pdf $f(x|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k) \tag{3}$$

Maximum Likelihood Estimator

- Typically, we take the log of the likelihood and obtain the log-likelihood function:

$$logL\left(\theta|x\right) = \sum_{i=1}^{n} \log\left(f\left(x_i|\theta_1, \ldots, \theta_k\right)\right) \tag{4}$$

- Since the logarithm is a strictly increasing function, maximizing the log of the likelihood is the same as maximizing the likelihood.

- Then, possible candidates for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that solve

$$\frac{\partial}{\partial\theta_i}logL\left(\theta|\boldsymbol{X}\right) = 0, i = 1, \ldots, k \tag{5}$$

Putting our calculus hats on, it is easier to differentiate the log-likelihood function than the likelihood!

Maximum Likelihood Estimator

Is this sounding like too much abstract gibberish? Let's take a look at an example...

MLE - Bernoulli Distribution

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ where:

- $X_i = 0$ if a randomly selected student does not own a sports car, and
- $X_i = 1$ if a randomly selected student does own a sports car.

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i|p) = p^{x_i}(1-p)^{1-x_i} \tag{6}$$

MLE - Bernoulli Distribution

Then, the likelihood function is, by definition:

$$L\left(p|x_1,\ldots,x_n\right) = \prod_{i=1}^{n} f\left(x_i|p\right) =$$

$$p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n} \quad (7)$$

And the log-likelihood function is:

$$\log L\left(p|x_1,\ldots,x_n\right) = \left(\sum_{i=1}^{n} x_i\right) \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p) \quad (8)$$

MLE - Bernoulli Distribution

Now, taking the derivative of the log-likelihood, and setting it to 0, we get:

$$\frac{\partial logL(p)}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1 - p} = 0 \tag{9}$$

Multiplying through by $p(1 - p)$, we get:

$$\left(\sum_{i=1}^{n} x_i\right)(1 - p) - \left(n - \sum_{i=1}^{n} x_i\right)p = 0 \tag{10}$$

Solving for p will yield the MLE estimator:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{11}$$

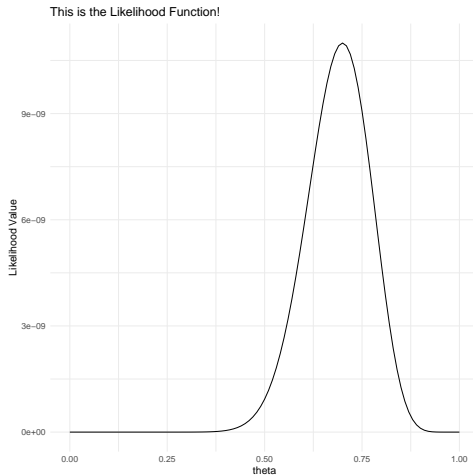**MLE - Bernoulli Distribution**

Still very abstract? Let's see this through a plot...

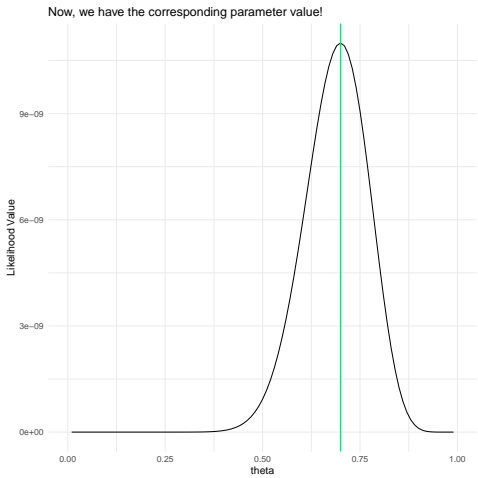## MLE - Bernoulli Distribution

Consider the data below:

```
data = c(1, 1, 1, 0, 1, 1, 1,
         0, 1, 0, 1, 0, 1, 1,
         1, 1, 0, 1, 1, 1, 1,
         0, 1, 0,1, 0, 0, 1, 1, 1)
```
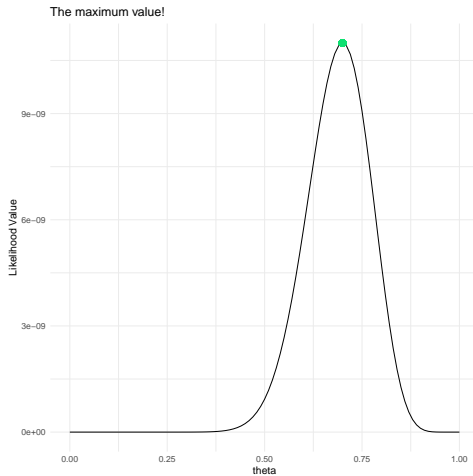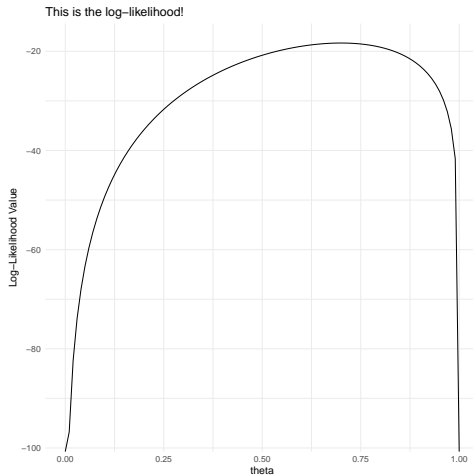
MLE - Bernoulli Distribution
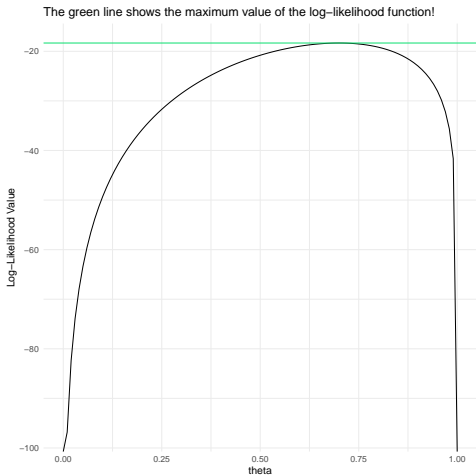
MLE - Bernoulli Distribution

MLE - Bernoulli Distribution

MLE - Bernoulli Distribution

MLE - Bernoulli Distribution

The green line shows the maximum value of the log–likelihood function!

NOVA IMS INFORMATION MANAGEMENT SCHOOL
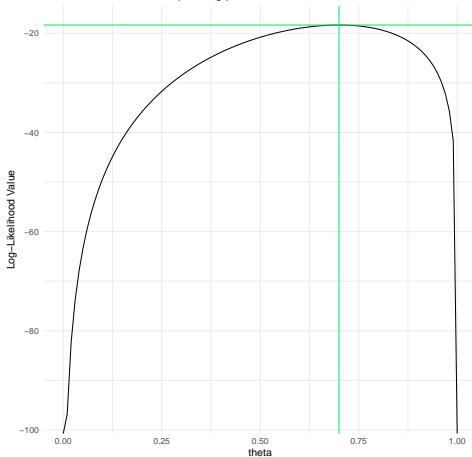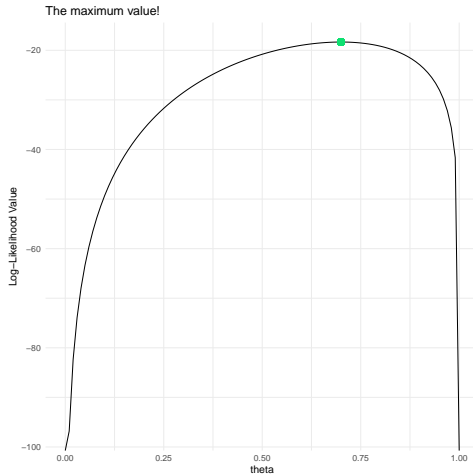
MLE - Bernoulli Distribution

Now, we have the corresponding parameter value!

MLE - Bernoulli Distribution

MLE - Bernoulli Distribution

The maximum value of the log-likelihood:

```
##               theta Log-likelihood Value
##            0.70000           -18.32593
```

MLE - Bernoulli Distribution

Maximizing the likelihood or the log-likelihood will always yield the same result!

Table of Contents

NOVAIMS INFORMATION MANAGEMENT SCHOOL

Point Estimator Properties

- The quality assessment of a given point estimator, should not be based on the particular estimate, but on the sampling distribution of that given estimator.
- Now, we will discuss some relevant properties of the point estimators.

Sampling Distribution

The sampling distribution of a point estimator is the probability distribution based on a large number of samples of size n, from a given population.

Unbiasedness

Unbiasedness

An estimator $T = T(X_1, X_2, \ldots, X_n)$ of the parameter θ is unbiased, if and only if

$$E(T) = \theta, \forall \theta \in \Omega \tag{12}$$

- What does this mean? The expected value of the estimator should be equal to the true value of the parameter we are estimating, no matter what the value of θ is, within the parameter space.

- Why is this relevant? Using an unbiased estimator allows a balance between the cases of over and underestimation, "on average" the value is accurate.

Unbiasedness

Let's see an example. We define the population parameters as:

```
popMean <- 100
popSD <- 15
```

These are the parameter's true values.

Unbiasedness

Then, let's generate a population, that follows a Normal Distribution, considering these parameters:

```
population <- rnorm(10000000, mean = popMean, sd = popSD)
```
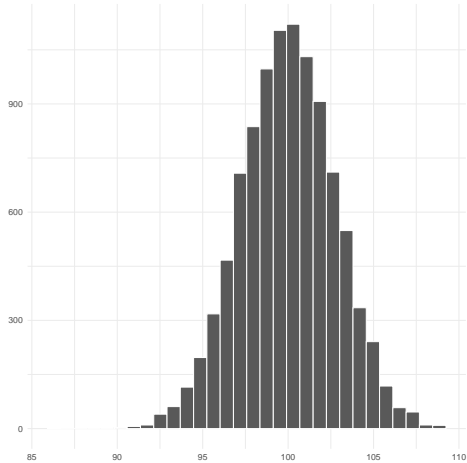
Unbiasedness

Finally, we will simulate, and extract several small samples (with a sample size equal to 30 observations):

```
sampDist <- replicate(10000,mean(sample(population, 30, replace=TRUE)))
```

Unbiasedness

The sampling distribution is given by:

Unbiasedness

Note that the mean of the sampling distribution is a pretty good match for the original
population mean.

```
mean(sampDist)

## [1] 99.9876
```

The sample mean is an unbiased estimator!

Efficiency

Efficiency

Let $T = T(X_1, X_2, \ldots, X_n)$ and $T' = T'(X_1, X_2, \ldots, X_n)$ be two unbiased estimators of the parameter θ. The estimator T is more efficient than T' if and only if

$$\text{Var}(T) \leq \text{Var}(T'), \forall \theta \in \Omega \qquad (13)$$

The estimator T is the most efficient if and only if the relationship (16) verifies for any other unbiased estimator T' of θ.

Note that the previous definition has two concepts: the first one is **relative efficiency** and the second one is **absolute efficiency**

Efficiency

How do we know we have absolute efficiency?

Efficiency

Fréchet-Cramér-Rao inequality

Let $(X_1, X_2, \ldots, X_n)$ be a sample of the population with pdf $f(x \mid \theta)$ and let $T = T(X_1, X_2, \ldots, X_n)$ be an unbiased estimator of θ, then

$$\text{Var}(T) \geq \frac{1}{n\mathcal{I}(\theta)} \tag{14}$$

where

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln\{f(X \mid \theta)\}}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ln\{f(X \mid \theta)\}}{\partial^2 \theta}\right] \tag{15}$$

where $\mathcal{I}(\theta)$ is the information matrix.

Efficiency

- The Fréchet-Cramér-Rao inequality allows to determine the inferior limit from which we can compare our variance with.
- The first step should be calculate the information matrix $(\mathcal{I}(\theta))$, to define the inferior limit.
- Then, we find the variance of our estimator and compare it with this inferior limit:
  - If the variance of our estimator is the same as our limit $\rightarrow$ we have absolute efficiency, i.e., we are certain that it does not exist another unbiased estimator with smaller variance.

Efficiency

What can we use to compare the relative efficiency?

Efficiency

The estimator with lower MSE is considered "better", because it contains estimates more concentrated around θ true value.

Consistency

The final useful property:

> **Consistency**
>
> An estimator $T_n = T(X_1, X_2, \ldots, X_n)$ is said to be consistent in quadratic mean if an only if
>
> $$\lim_{n \to +\infty} \mathsf{E}\{(T_n - \theta)^2\} = 0, \forall \theta \in \Omega \tag{17}$$
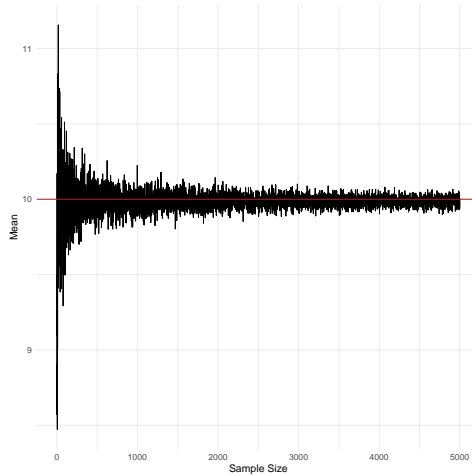
This property means that as the sample increases, the closer the estimator will be to the real population parameter.

Consistency

Let's see an example for the sample mean. We will draw several samples from a population, with a Normal distribution, mean equal to 10 and standard deviation equal to 2.
We will increase the sample size, starting from 1 until 5000 observations.

```
##The population
N = 5000
X = rnorm(N, mean = 10, sd = 2)
```

Consistency

Summary

In summary, the ideal properties of an estimator...

- Unbiasedness: the expected value of the estimator should be equal to the true value of the parameter we are estimating.
- Efficiency: the variance of our estimator is the smallest (could be in relative or absolute terms).
- Consistency: as the sample increases, the estimator will be closer to the true value of the parameter we are estimating.
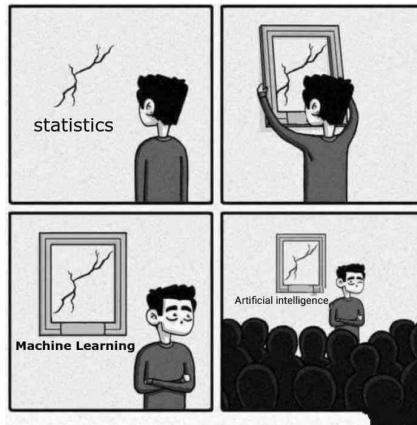
Summary

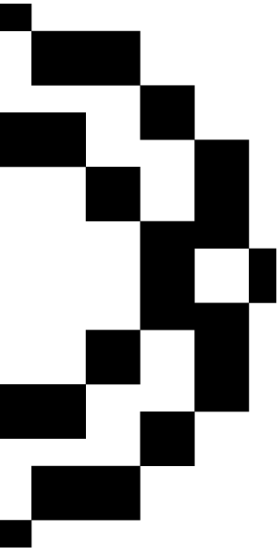This class was a bit theoretical, a sharp student could ask:

Aren't these estimation methods and point estimator properties already widely studied for the common statistical distributions?

Why is this relevant?

Summary

- Statistics is the basis of more complex methods in data science.
- Understand the basic concepts behind point estimation is crucial to applying and interpreting it in a mindfully way.

Hypothesis testing

Table of Contents

NOVA IMS INFORMATION MANAGEMENT SCHOOL

Hypothesis Testing

- In real life, we don't typically have access to the whole population. In these cases, we can use the sample data to estimate the population parameter.
- However, to conclude about possible values on the population parameter, we need to use statistical inference.
- Here, we are going to look at a more formal method for testing whether a given value is a reasonable value of a population parameter. To do this we need to have a hypothesized value of the population parameter.

Five Step Hypothesis Testing Procedure

1. Write hypothesis
2. Calculate the test statistic
3. Determine the p-value
4. Make a decision
5. State a conclusion

Write Hypothesis

- The first step in conducting a hypothesis test is to write the hypothesis statements that are going to be tested. For each test you will have a null hypothesis and an alternative hypothesis.
- **Null Hypothesis**: The statement that there is not a difference in the population, denoted as H_0.
- **Alternative Hypothesis**: The statement that there is some difference in the population(s), denoted as H_1.

Table of Contents

N✪VAIMS INFORMATION MANAGEMENT SCHOOL

One Sample Mean

Research question	Is the population mean different from μ_0	Is the population mean greater than μ_0	Is the population mean less than μ_0
Null Hypothesis, H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
Alternative Hypothesis, H_1	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Type of Hypothesis test	Two-tailed, non-directional	Right-tailed, directional	Left-tailed directional

μ_0 is the hypothesized population mean

Calculate test statistic

Consider the data `sleep`. This dataset contains the effect, i.e., difference in hours of sleep compared to a control group, of two treatments. Suppose we want test if this difference regardless the treatment, is different from 0. Then, the hypothesis are:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

Calculate test statistic

- We want to test if the mean is different from zero. To perform any inference on a given sample, we need the test statistic.
- In this case, the test statistic is given by:

$$t = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t(n-1)$$

where $\bar{X}$ is the sample mean, $\hat{\sigma}$ is the estimated standard deviation, μ_0 is the hypothesized population mean and n is the sample size.

Calculate test statistic

- The t distribution is used to estimate the sampling distribution when the sample size is large (at least 30) or when the population is known to be normally distributed (but σ is unknown).
- The normal distribution is used when the population is normal and the population standard deviation is know.

Calculate test statistic

Example

The test statistic for our example is given by:

$$t_{obs} = \frac{1.54 - 0}{0.45} = 3.413$$

Determine the p-value

- Now, through the test statistic calculated, we will compute the probability that a population with the specified parameter would produce a sample statistic as extreme or more extreme to the one we observed in our sample.

- This probability is known as the *p*-value and it is used to evaluate statistical significance.
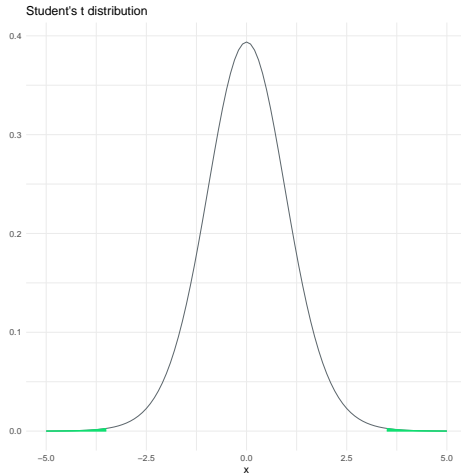
Determine the p-value

- *p-value*: Given that the null hypothesis is true, the probability of obtaining a sample statistic as extreme or more extreme than the one in the observed sample, in the direction of the alternative hypothesis.

- A test is considered to be statistically significant when the p-value is less than or equal to the level of significance, also known as the alpha (α) level. The most frequently used alpha level in many fields is 0.05.

Determine the p-value

```
2*(1-pt(3.413,19))

## [1] 0.002917389
```

Determine the p-value

Student's t distribution

Fast calculation in R

```
t.test(sleep$extra)

##
##  One Sample t-test
##
## data:  sleep$extra
## t = 3.413, df = 19, p-value = 0.002918
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5955845 2.4844155
## sample estimates:
## mean of x
##      1.54
```

Make a decision

- Since the *p*-value is lower than 0.05, we reject the null hypothesis.
- Another alternative, would be to verify if the test statistic is in the critical region of the test.
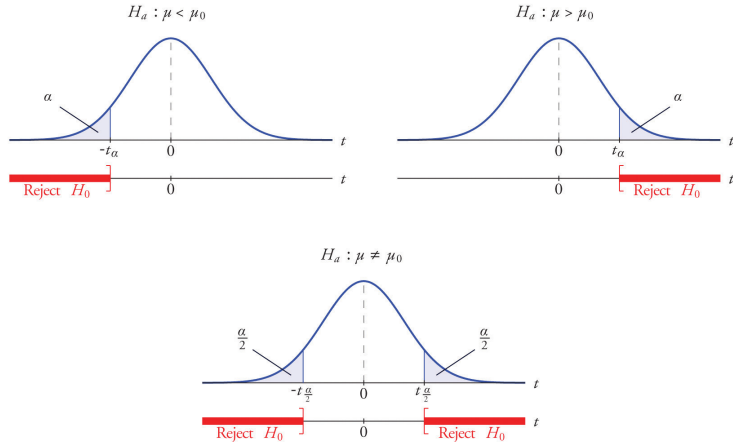- This region is given by

$$\mid t_{obs} \mid \geq t_{1-\alpha/2;n-k-1}$$

  where $t_{1-\alpha/2;n-k-1}$ corresponds to the quantile of the *t*-student distribution.
- Notice that we consider $\alpha/2$ because our test is bilateral (it has $\neq$ on H_1).

Critical Regions
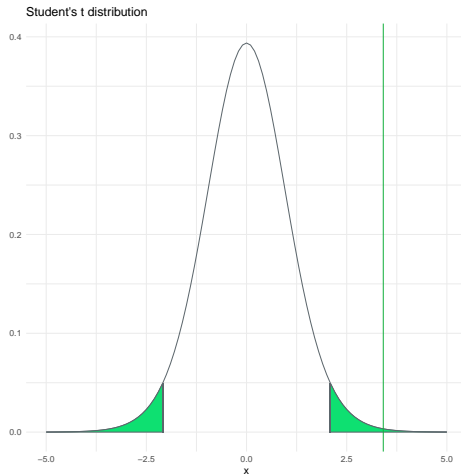
The critical regions for each test are:

Make a decision

To get the quantiles for the critical region:

```
qt(c(0.025,0.975),19)

## [1] -2.093024  2.093024
```

Make a decision

Student's t distribution

State a conclusion

- In conclusion, we can state with a 5% significance level that the difference in hours of sleep between the treated students and the control group is different from zero.

Table of Contents

NOVAIMS INFORMATION MANAGEMENT SCHOOL

Hypothesis Testing: One sample proportion

- Statistical inference can be applied to different parameters of the population.
- Besides the mean, the proportion is also a relevant statistic to study.

One sample proportion

Suppose we have gathered a sample of patients that had a stroke. We want to test if the proportion of patients with hypertension is greater than 0.4 (i.e., 40%). From the 136 patients that had a stroke, 129 had hypertension.

One sample proportion

Research question	Is the population proportion different from p_0	Is the population proportion greater than p_0	Is the population proportion less than p_0
Null Hypothesis, H_0	$p = p_0$	$p = p_0$	$p = p_0$
Alternative Hypothesis, H_1	$p \neq p_0$	$p > p_0$	$p < p_0$
Type of Hypothesis test	Two-tailed, non-directional	Right-tailed, directional	Left-tailed directional

p_0 is the hypothesized population proportion.

One sample proportion

- First, we should define the hypothesis:

$$\begin{cases} H_0 : p = 0.4 \\ H_1 : p > 0.4 \end{cases}$$

- The test statistic in this case is given by

$$Z_{obs} = \frac{\hat{p} - 0.4}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0, 1)$$

- The critical region for this test:

$$Z_{obs} \geq Z_{1-\alpha}$$

where $Z_{1-\alpha}$ is the $1 - \alpha$ percentile of the standard normal distribution

One sample proportion

- In **R**, we can use the function prop.test():

```r
prop.test(x=129,
          n=136,
          p=0.4,
          alternative = c('greater'))
```

where *x* corresponds to the number of 'sucesses' (here the number of patients that had a stroke), *n* corresponds to the size of the sample, *p* the probability value we want to test and the *alternative* specifies the alternative hypothesis.

One sample proportion

```
prop.test(x=129, n=136, p=0.4, alternative = c('greater'))

##
##  1-sample proportions test with continuity correction
##
## data:  129 out of 136, null probability 0.4
## X-squared = 168.22, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is greater than 0.4
## 95 percent confidence interval:
##  0.9031356 1.0000000
## sample estimates:
##         p
## 0.9485294
```

Table of Contents

NOVA IMS INFORMATION MANAGEMENT SCHOOL
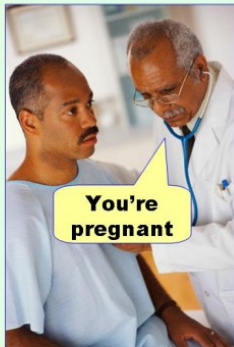
Type I and Type II Errors

- When conducting a hypothesis test there are two possible decisions: reject the null hypothesis or fail to reject the null hypothesis.
- You should remember though, hypothesis testing uses data from a sample to make an inference about a population.
- When conducting a hypothesis test we do not know the population parameters. In most cases, we don't know if our inference is correct or incorrect.

Type I and Type II Errors

- When we reject the null hypothesis there are two possibilities. There could really be a difference in the population, in which case we made a correct decision. Or, it is possible that there is not a difference in the population but our sample was different from the hypothesized value due to random sampling variation. In that case we made an error. This is known as a **Type I error**.

- When we fail to reject the null hypothesis there are also two possibilities. If the null hypothesis is really true, and there is not a difference in the population, then we made the correct decision. If there is a difference in the population, and we failed to reject the null hypothesis, then we made a **Type II error**.

Type I and Type II Errors

Type I and Type II Errors

Decision/Reality	H_0 is true	H_0 is false
Reject H_0	Type I error (α)	Correct decision
Fail to reject H_0	Correct decision	Type II error (β)

Table of Contents

NOVAIMS INFORMATION MANAGEMENT SCHOOL

Power of test

- The probability of rejecting the null hypothesis, given that the null hypothesis is false, is known as power. In other words, power is the probability of correctly rejecting H_0.

$$Power = 1 - \beta$$

where β = probability of committing a Type II Error.

Power of test

- The power of a test can be increased in a number of ways, for example increasing the sample size, decreasing the standard error, increasing the difference between the sample statistic and the hypothesized parameter, or increasing the alpha level.

Power of test

Power of test

Question

When we fail to reject the null hypothesis, can we accept the null hypothesis? For example, with a *p*-value of 0.12 we fail to reject the null hypothesis at 0.05 alpha level. Can we say that the data support the null hypothesis?

Confidence Intervals

Confidence Intervals

- Besides the *t*-test, another conventional way of performing statistical inference is through a confidence interval.
- **Confidence Interval** is a range computed using sample statistics to estimate an unknown population parameter with a stated level of confidence.

Applying confidence Intervals

The general form of a confidence interval is given by:

$$\bar{X} \pm t_{1-\alpha/2}(\frac{\hat{\sigma}}{\sqrt{n}})$$

when constructing a confidence interval for a population mean the point estimate is the sample mean, $\bar{x}$. The multiplier is the percentile $1 - \alpha/2$ from t distribution. And, the standard error is equal to $\frac{\hat{\sigma}}{\sqrt{n}}$.

Applying confidence Intervals in R

Consider the vector `nhtemp` that contains information regarding the mean annual
temperature in degrees Fahrenheit in New Haven, Connecticut, from 1912 to 1971. Calculate
a confidence interval for the mean temperature, between 1912 and 1971.

Applying confidence Intervals in R

```
t.test(nhtemp)

##
##  One Sample t-test
##
## data:  nhtemp
## t = 313.12, df = 59, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  50.83306 51.48694
## sample estimates:
## mean of x
##     51.16
```

Applying confidence Intervals in R

```
library(Rmisc)
CI(nhtemp)

##    upper     mean    lower
## 51.48694 51.16000 50.83306

library(DescTools)
MeanCI(nhtemp)

##     mean   lwr.ci   upr.ci
## 51.16000 50.83306 51.48694
```

Interpreting confidence Intervals

- Confidence intervals are often misinterpreted. The logic behind them may be a bit confusing. Remember that when we're constructing a confidence interval we are estimating a population parameter when we only have data from a sample.
- We don't know if our sample statistic is less than, greater than, or approximately equal to the population parameter. And, we don't know for sure if our confidence interval contains the population parameter or not.
- The correct interpretation of a 95% confidence interval is that 'we are 95% confident that the population parameter is between X and X.'
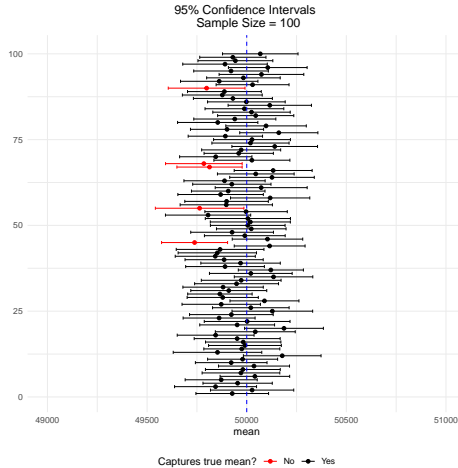
Interpreting confidence Intervals

Hence,

- A confidence interval contains a range of acceptable estimates of the population parameter.
- Values in a confidence interval are reasonable estimates for the true population value.
- Values not in the confidence interval are not reasonable estimates for the population value.

Interpreting confidence Intervals

Let's visualize an example, where we calculated 100 confidence intervals from 100 different samples, drawn from the same population...
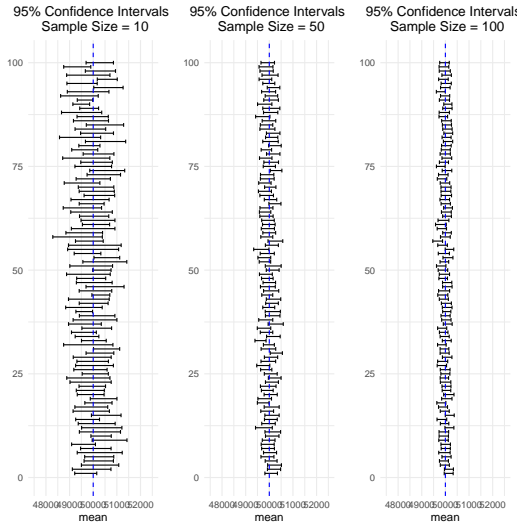
Interpreting confidence Intervals

Impact of Sample Size on Confidence Interval

- Earlier in this lesson we learned that the sampling distribution is impacted by sample size.
- As the sample size increases, the standard error decreases. With a larger sample size there is less variation between sample statistics.
- Let's see and example of this impact...

Impact of Sample Size on Confidence Interval

NOVAIMS INFORMATION MANAGEMENT SCHOOL

Inference for two samples

Two sample proportion test

Suppose we have gathered a sample of patients that had a stroke in two different hospitals. We want to test if the proportion of patients with hypertension in hospital A is greater than in hospital B. For hospital A, 140 patients that had a stroke, 119 had hypertension. From hospital B, 240 patients that had a stroke, 200 had hypertension.

Two sample proportion

- First, we should define the hypothesis:

$$\begin{cases} H_0 : p_A = p_B \\ H_1 : p_A > p_B \end{cases}$$

- The test statistic in this case is given by

$$Z_{obs} = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}} \sim N(0, 1)$$

- The critical region for this test:

$$Z_{obs} \geq Z_{1-\alpha}$$

Two sample proportion test

```
patients <- c(140, 240)
hypertension <- c(119, 200)
prop.test(hypertension, patients, alternative = c('greater'))


##
##  2-sample test for equality of proportions with continuity correction
##
## data:  hypertension out of patients
## X-squared = 0.079566, df = 1, p-value = 0.3889
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.0524679  1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.8500000 0.8333333
```

We fail to reject H0 so the interpretation is:
-> We dont have data to support the claim that Pa is bigger than Pb

Two sample mean test

Exercise

Suppose we have gathered two samples of students of public schools in Lisbon and Oporto. We want to study if there is a difference of the proportion of students that have a personal computer at home between Lisbon and Oporto. In Lisbon, in the sample collected of 300 students, 213 had a personal computer at home. In Oporto, in the sample collected of 250 students, 145 had a personal computer at home.

Two sample mean test

Example

A school that implemented a healthy food program wants to test if there is a difference in their students' weights, comparing to the school that didn't implement the same program. The collected sample from each school is:

```
schoolA = c(38.9, 61.2, 73.3, 51.8, 63.4, 64.6, 58.4, 58.8, 68.5)
schoolB = c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
```

Two sample mean test

- First, we should define the hypothesis:

$$\begin{cases} H_0 : \mu_A = \mu_B \\ H_1 : \mu_A < \mu_B \end{cases}$$

- If the samples have different variances, the test statistic in this case is given by

$$t_{obs} = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}} \sim t(n_A + n_B - 2)$$

- If the samples have the same variance, the test statistic is

$$t_{obs} = \frac{\hat{\mu}_A - \hat{\mu}_B}{\hat{\sigma}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t(n_A + n_B - 2)$$

- The critical region for this test:

$$t_{obs} \leq t_{\alpha; n_A + n_B - 2}$$

Test of Variances

- Before we test for the difference in means between the two samples, we must determine whether the variances are equal through a test. The hypothesis are:

$$\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \neq \sigma_B^2 \end{cases}$$

- The test statistic is

$$F_{obs} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} \sim F(n_A - 1, n_B - 1)$$

- And, the critical region for this test is:

$$F_{obs} \geq F_{\alpha; n_A - 1, n_B - 1}$$

Two of Variances

```r
schoolA = c(38.9, 61.2, 73.3, 51.8, 63.4, 64.6, 58.4, 58.8, 68.5)
schoolB = c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
var.test(schoolA, schoolB)
```

```
##
##  F test to compare two variances
##
## data:  schoolA and schoolB
## F = 1.1371, num df = 8, denom df = 8, p-value = 0.8602
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2564992 5.0411826
## sample estimates:
## ratio of variances
##            1.137128
```

Looking at the p-value, we don't reject the null hypothesis, thus we don't have statistical evidence that the variances among samples are different.

Two sample mean test

```
schoolA = c(38.9, 61.2, 73.3, 51.8, 63.4, 64.6, 58.4, 58.8, 68.5)
schoolB = c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
t.test(schoolA, schoolB, alternative = c('less'), var.equal=TRUE)


##
##  Two Sample t-test
##
## data:  schoolA and schoolB
## t = -1.9943, df = 16, p-value = 0.03172
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -1.134833
## sample estimates:
## mean of x mean of y
##  59.87778  68.98889
```

We define var.equal = TRUE, because we did not reject the null on the previous test.

Hypothesis Testing

Flow chart: which test statistic should you use?