Fundamentals of Inference

Lecture 5

Francesco Dotto
francesco.dotto@uniroma3.it

4/9/2021

Inference vs Probability

from population to data and back

- **Probability** starts from the population, which is described by the means of a probability distribution function, and predicts what happens in a sample extracted from it.

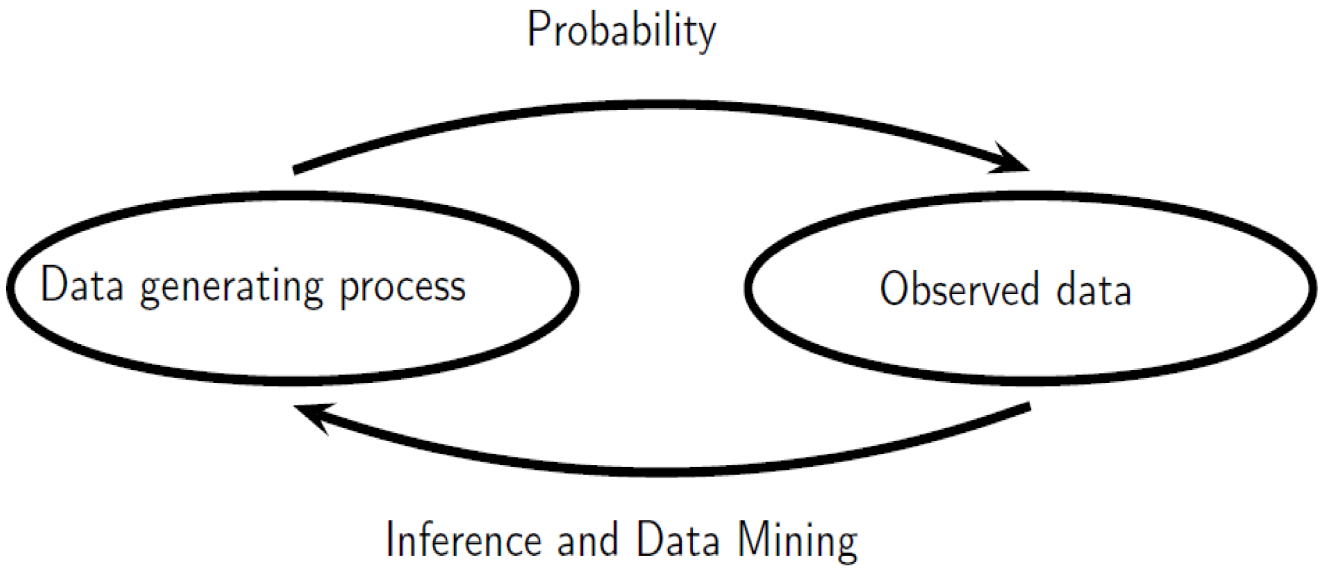
Given a probability law, which is the probability of a given event?

- **Inference** starts from a sample and describes the observed data with the aim of inferring relevant information on the population.

Given a sample, which are the parameters of the probability law that generated my sample?

What is Inference?

In a figure



What is Inference?

a general introduction

- **Estimate**: recover some parameter explaining the phenomenon that generates the data
  - **point estimate**: a *single number* that is our best guess for the parameter.
  - **interval estimate**: an *interval of numbers* that is believed to contain the actual value of the parameter.

- **Hypothesis testing**: using data to validate certain statements or predictions

Random sample: A practical example

For making inference we assume that our data (collected in a sample) come from a **probability** distribution. The probability distribution is assumed to be known but its **parameters are unknow**

- We want to estimate the **true proportion (p)** of Americans who favor doctor-assisted suicide.
- We want to estimate the **true mean (μ)** of Americans who favor doctor-assisted suicide.

Inference

Starting from the examples

- We want to estimate the **true proportion (p)** of Americans who favor doctor-assisted suicide.

$$p = \frac{\text{Americans who favor doctor assisted suicide}}{\text{Americans}}$$

* The population of interest is too large: we observe a **random** sample and estimate such proportion on a random sample

$$\hat{p} = \frac{\text{Americans who favor doctor assisted suicide IN THE SAMPLE}}{\text{Americans IN THE SAMPLE}}$$

Inference...formally speaking

Sampling distribution

- Suppose that, in a given sample of $n = 50$ we find 35 Americans that favour doctor assisted suicide. If, within the following days, we collect another sample, we expect to obtain a different result. **The sample proportion has its own variability**
- Formally speaking, when we record, for each element of the sample, the corresponding opinion we are making a Bernoulli experiment for each observation $i = 1, 2 \dots, n$

$$X_i = \begin{cases} 0 & \text{with probability } 1 - p \\ 1 & \text{with probability } p \end{cases}$$

Central Limit Theorem for the proportion

- When we compute the proportion in a sample we are computing a sample mean on variables that can only assume the values $\{0, 1\}$.

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

- For the Central Limit Theorem we can conclude that

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Central Limit Theorem for the proportion

why?

- The Central Limit Theorem states that

CLT

Given n radom variables $X_1, X_2, \ldots, X_n$ that are independent and that have the same distribution with mean $\mathbb{E}[X] = \mu$ and variance $\mathbb{E}[X] = \sigma^2$, the following holds

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is distributed as } \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Hang on: why are we doing all of this???

- Due the result provided above (and many others that will follow) we have a probability distribution associated to the results we obtained in a given sample.
- Is it really useful?
- Example: In a survey it was reported that 33 percent of women believe in the existence of aliens. I 100 women are selected at random, what is the probability that more than 45 percent will say that they believe in aliens?

Example: resolved

- We need to compute $P(\hat{p} > 0.45)$ and we know that $\hat{p} \sim N\left(0.33, 0.0022\right)$ since $p = .33$ and $p(1-p)/n = 0.0022$.

$$P(\hat{p} > 0.45) = P\left(\frac{\hat{p} - 0.3}{\sqrt{0.33(1-0.33)/100}} > \frac{0.45 - 0.33}{\sqrt{0.33(1-0.33)/100}}\right) =$$

$$P(Z > 2.55) = 0.0054 \text{ (where } Z \sim N(0,1))$$

Sampling distribution

The case of the sample mean

- We want to estimate the **true** daily mean time spent driving their motor vehicles by the Americans.

$$\mu = \frac{\text{total amount of time spent by American males}}{\text{total number of Americans who drive}}$$

* We use an estimate obtained from a sample

$$\bar{x} = \frac{\text{total amount of time spent by American males IN THE SAMPLE}}{\text{total number of Americans who drive IN THE SAMPLE}}$$

Central Limit Theorem heps us once again

- To derive the sampling distibution of the sample mean we rely to CLT which states that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- **Example:** A tire manufacturier claims that its tires last an average 60,000 miles with a standard deviation of 3,000 miles. 64 tires are placed on test. What is the probability that their failure miles will be more than 59,500 miles?

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{59,500 - 60,000}{30,000/\sqrt{64}} \text{ thus } P(\bar{x} > 59,500) =$$

$$= P(Z > -1.33) = 0.4082$$

Confidence intervals

Why?

- So far we evaluated the sampling distribution of quantitites taylored at estimating population parameters ($\hat{p}$ for estimating population proportion p and sample mean $\bar{x}$ for estimating population mean μ). These are **point estimates**
- The sampling distribution allows us to construct intervals of plausible values associated to the estimate of a population parameter. These are called **confidence intervals**

Confidence Interval

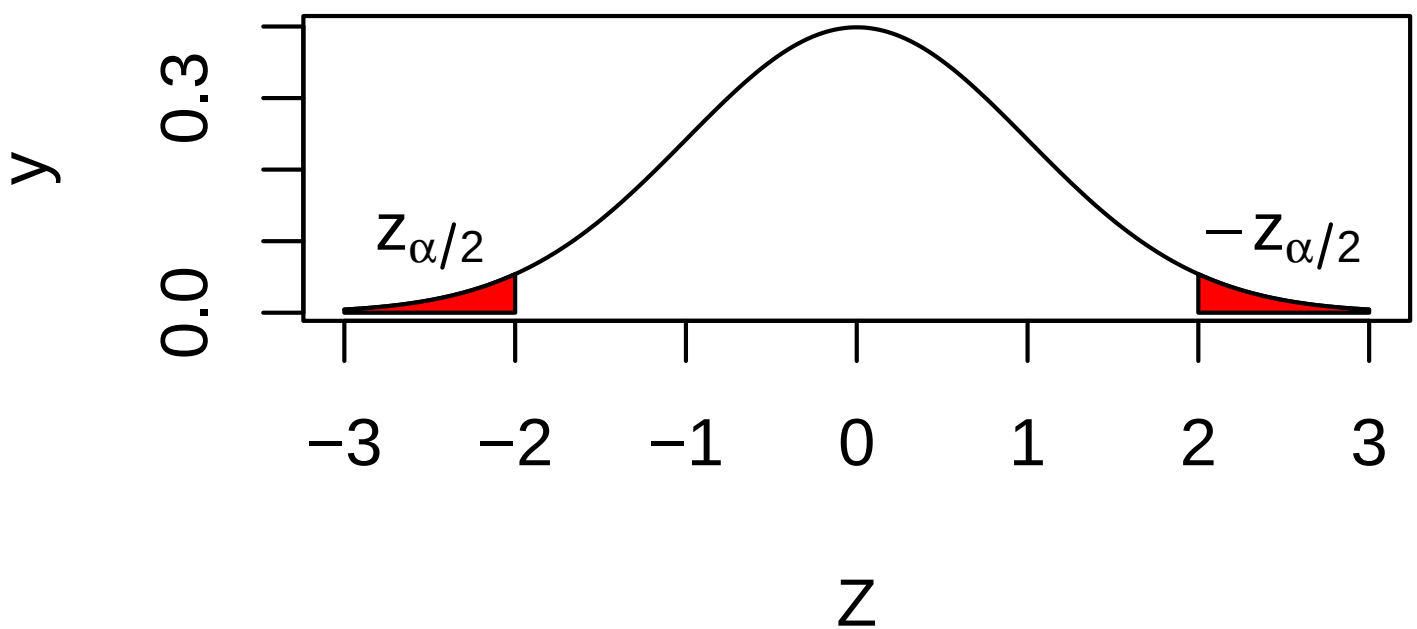
Given n radom variables $X_1, X_2, \dots, X_n$ and a parameter of interest θ, and interval $[L_1(X_1, X_2, \dots, X_n), L_1(X_1, X_2, \dots, X_n)]$ is Confidence Interval at $1 - \alpha$ confidence if it contains with probability $1 - \alpha$ the unkwnow θ parameter

Confidence Interval: understanding the definition

Build it for the sample proportion $\hat{p}$

- CLT states that

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} = Z \sim N(0,1) \text{ thus } 1-\alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \tag{1}$$



Confidence Interval: understanding the definition

Build it for the sample proportion $\hat{p}$

- Knowing that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

- Doing some calculations we obtain:

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

Confidence interval for the proportion

An exercise

- In a random sample composed by $n = 100$ persons, 77 percent of them declared that they regularly pray. Determine a 90 percent confidence interval for the true proportion of people that pray.
- $\widehat{p} = .77$, $\sqrt{\widehat{p}(p - \widehat{p})/n}$, $z_{\alpha/} = 1.645$ thus the interval

$$[0.77 - 1.645 \times 0.042;\ 0.77 - 1.645 \times 0.042]$$

contains the true proportion of persons that pray with probability 0.9.

Confidence interval for the sample mean

Same as for the proportion

- From CLT we know that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z \sim N(0,1) \text{ thus } 1-\alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

- Thus the following holds

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

which implies that the interval

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}};\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \text{ contains the true } \mu \text{ with probability } 1-\alpha$$

Confidence interval for the sample mean

Same as for the proportion

- Exercise: A random sample composed of $n = 100$ public school teachers has a mean salary of $\$31,578$ with a Standard deviation of $\$4,415$. Construct a 99% for the true mean salary
- Solution z_{α_2} is, in this case, equal to 2.575, $\bar{x} = 31,578$, $n = 100$ and $\sigma = 4,415$. y doing some calculations we obtain that $\sigma/\sqrt{n} = 441.5$ and $\sigma/\sqrt{n} \times z_{\alpha_2}$ and thus we are 99% **confident** that the true average salary lies within the following interval

$$[31,578 - 1,136;\ 31,578 + 1,136]$$

Hypothesis Testing

Conceptually

Hypothesis Testing

Hypothesis Testing is an inferential procedure that allow us to quantify how close are things IN THE POPULATION to our expectations or theories.

Hypothesis Testing

- A **statistical** hypothesis is an opinion about a population parameter.
- There are two types of hypothesis: Null hypothesis (H_0) and Alternative Hypothesis (H_1)

Hypothesis Statement

Your teacher claims that 60 percent of American males are married. You feel that such proportion is higher. In a random sample of $n = 100$ American males, 65 of them were marreid. Test the teacher's claim at 5 percent of significance.

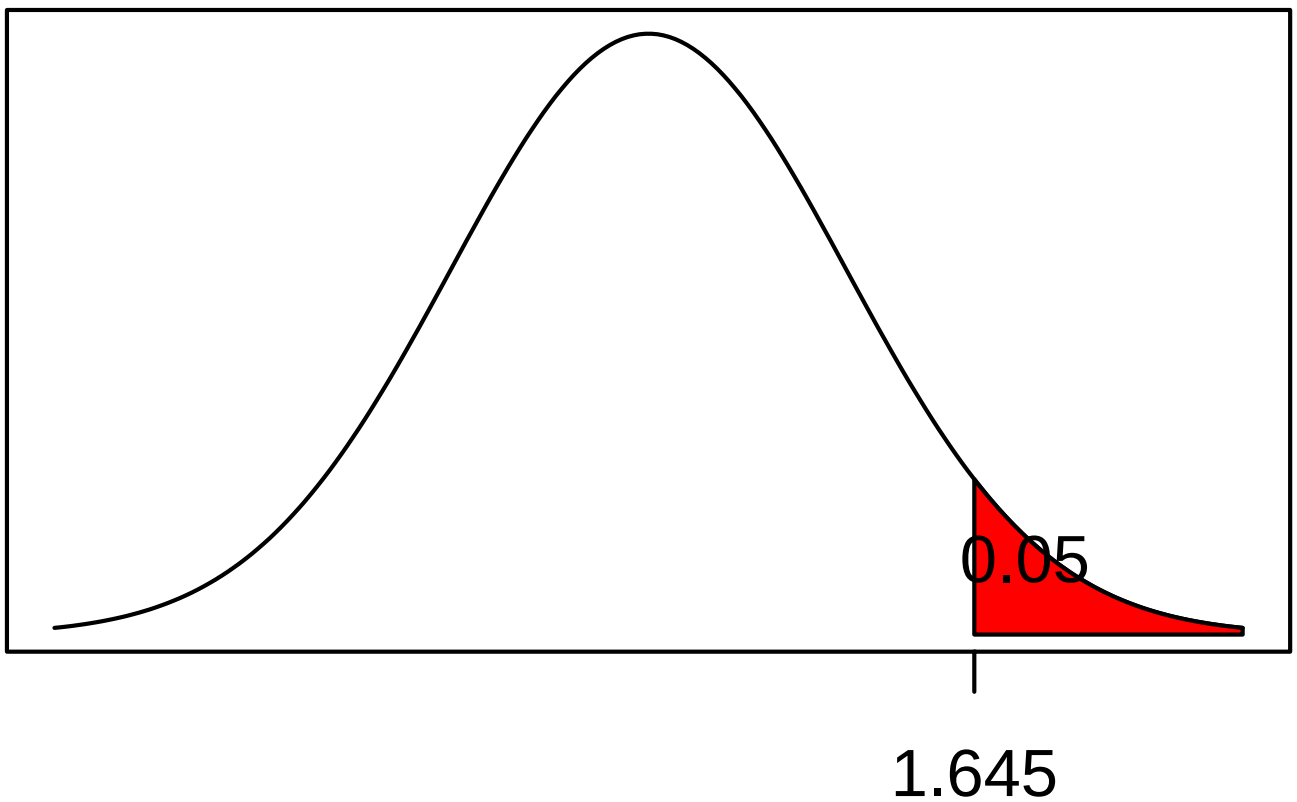
$$\begin{cases} \mathbf{H}_0 : p \leq 0.6 \\ \mathbf{H}_1 : p > 0.6 \end{cases}$$

Hypothesis Testing

Testing teacher claim

- Given that In this case $\hat{p} = 65/100 = .65$, p_0, the values that I am testing is 0.6, $\alpha = 0.05$, $z_\alpha = 1.645$ and from CLT we know that

$$Z_p = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$



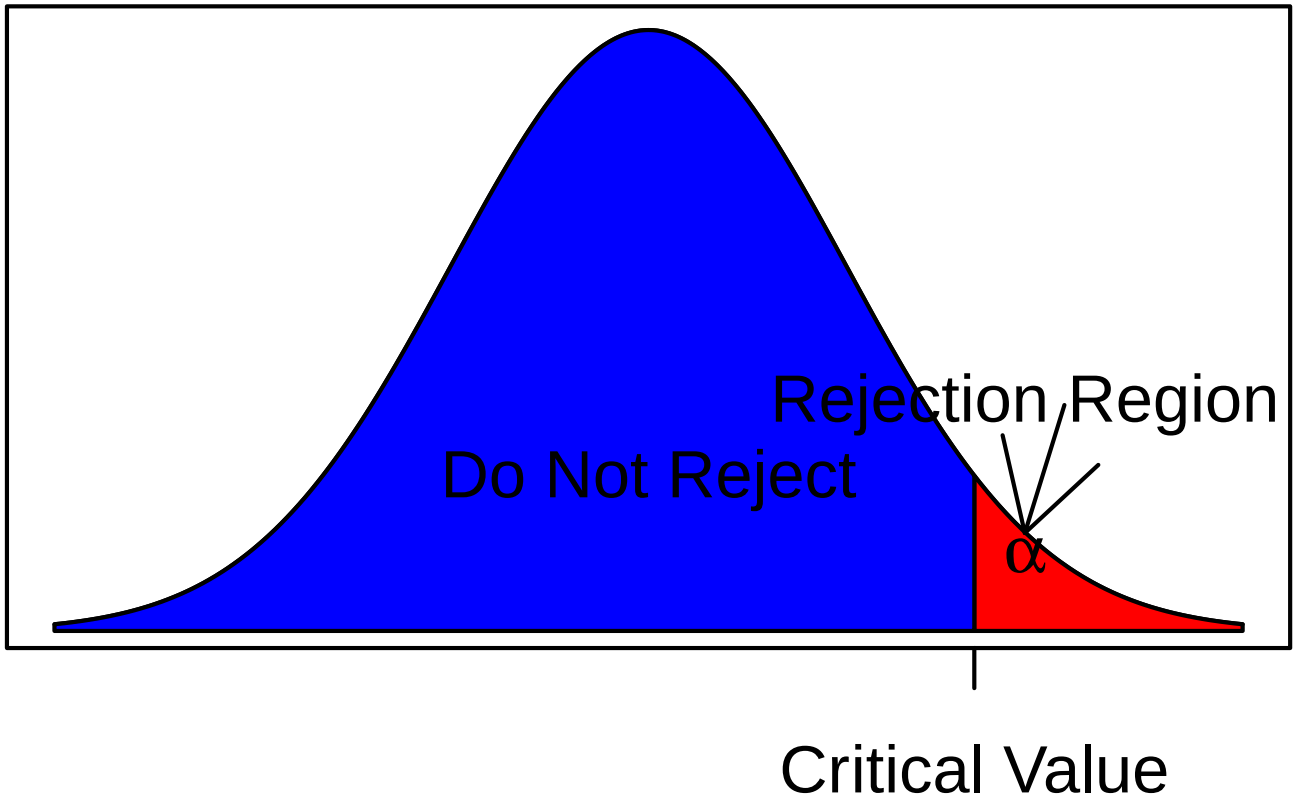
Hypothesis testing

Final Step

- We **reject** the null hypothesis H_0 if the computed value of test statistic $z > z_\alpha$.
- Since the observed z is equal to $1.0201 < 1.645$ we do not reject H_0
- Conceptually speaking,There is no sufficient sample evidence to calim that more than 60 percent of Americans are married at **5 percent of significance**. Any difference between the sample proportion and the postulated proportioon 0.6 may be due to the chance

Hypothesis testing

Graphically Speaking



#Hypothesis Testing

Ingredients of hypothesis testing

1. Two hypotheses:
   - NUll Hypothesis - A claim about the population parameter
   - Alternative Hypothesis - states that there is a difference between a parameter and a specific value
2. Test Statistic. A function of the data whose distribution **is known**
3. A critical value
4. A decision rule

In the forthcoming slides we comment each ingredient **One by one**

Hypothesis testing

Ingrendient 1: Hypotheses

- Hypothesis Statement. This is quite important and common within the experiental context
  1. I experiment a new medicine. I make a clinical trial and observe the results. The starting point may be: the average effect μ is equal to 0. The alternative hypothesis will be: the new medicine HAS an effect. Thus H_1 states: $\mu > 0$.

Hypothesis Explained

The Null Hypothesis usually refers to the status quo the thing we are trying to find evidence against

Hypothesis testing

Ingrendient 2: Test Statistic

1. The Test Statistics is a Numerical Summary of a dataset.
2. It is used because its sampling distribution is known (it can be calculated)
3. The sampling distribution is the corner stone of the test.

$$\text{Test Statistic} \rightarrow \text{ Sampling Distrution}$$

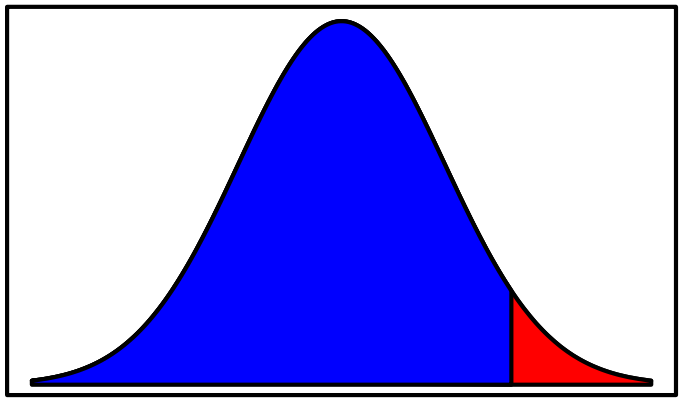
4. The sampling distribution allows us to evaluate if the difference among the Values observed within the sample and the value stated within the Null hypothesis is statistically significant or is a consequence of the variability among the different samples

Hypothesis testing

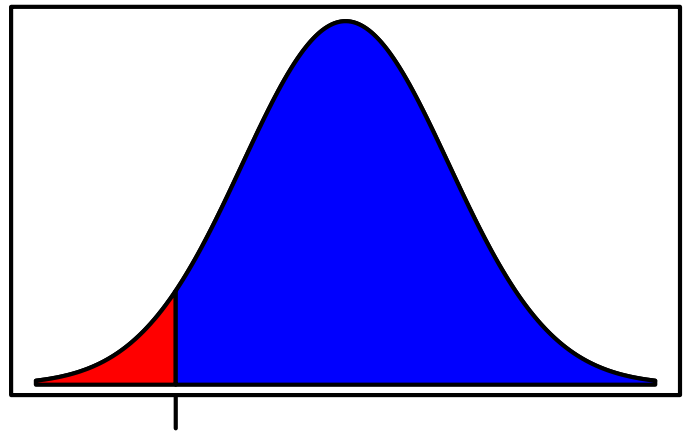
Ingredient 3: The critical value

1. The critical value separates Rejection Region from Non Rejection Region
2. How do I choose it? Whenever you do a test, you specify the level of significance (typical values are 1% or 5%).

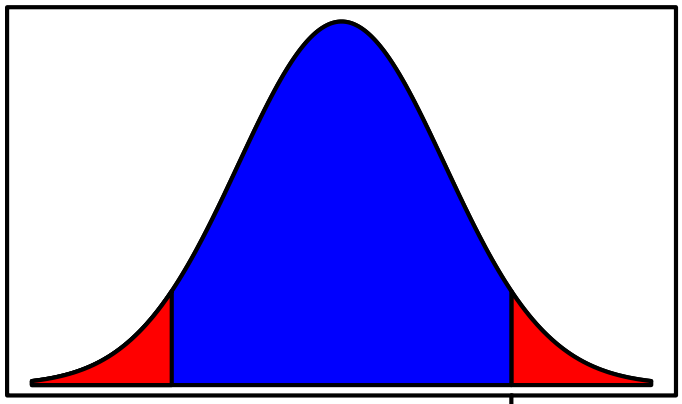
$H_1: \mu > \mu_0$



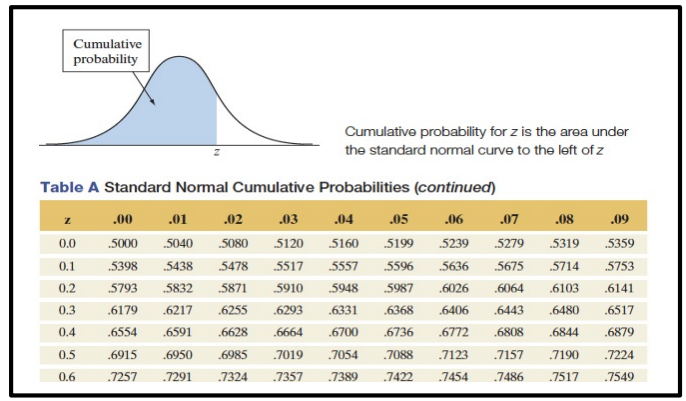
$H_1: \mu < \mu_0$



$H_1: \mu \neq \mu_0$



Critical Value



Cumulative probability for z is the area under the standard normal curve to the left of z

Table A Standard Normal Cumulative Probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549

Inference...formally speaking

Statistical Inference...formally speaking

The slides that are now displayed deal with inference in a formal way. Those are left to the interested reader and can be used as the corses of master goes by

Random sample: Formulas

A **random sample** is a collection of random variables $X_1, \ldots, X_n \sim f_{X_1,\ldots,X_n}$, that are:

- *independent*

$$f_{X_1,\ldots,X_n} = \prod_{i=1}^{n} f_{X_i}(x_i)$$

- *identically distributed*

$$f_{X_i}(x_i) = f_X(x_i) \quad \forall i$$

As a consequence

$$f_{X_1,\ldots,X_n} = \prod_{i=1}^{n} f_X(x_i)$$

An **observed sample** $(x_1, \ldots, x_n)$ is a realization of the random sample. Thus, using the previous example, every patient is a Binomial random variable, n is the number of trials, and p, which is **unknown** is the probability that the treatment works.

Toy Example

how to compute the sample distribution

Let $X_1, \dots, X_n$ i.i.d. (**i**ndependent **i**dentically **d**istributed) from a Poisson(λ).

The **sampling distribution** $f_{X_1,\dots,X_n}$ can be derived as follows:

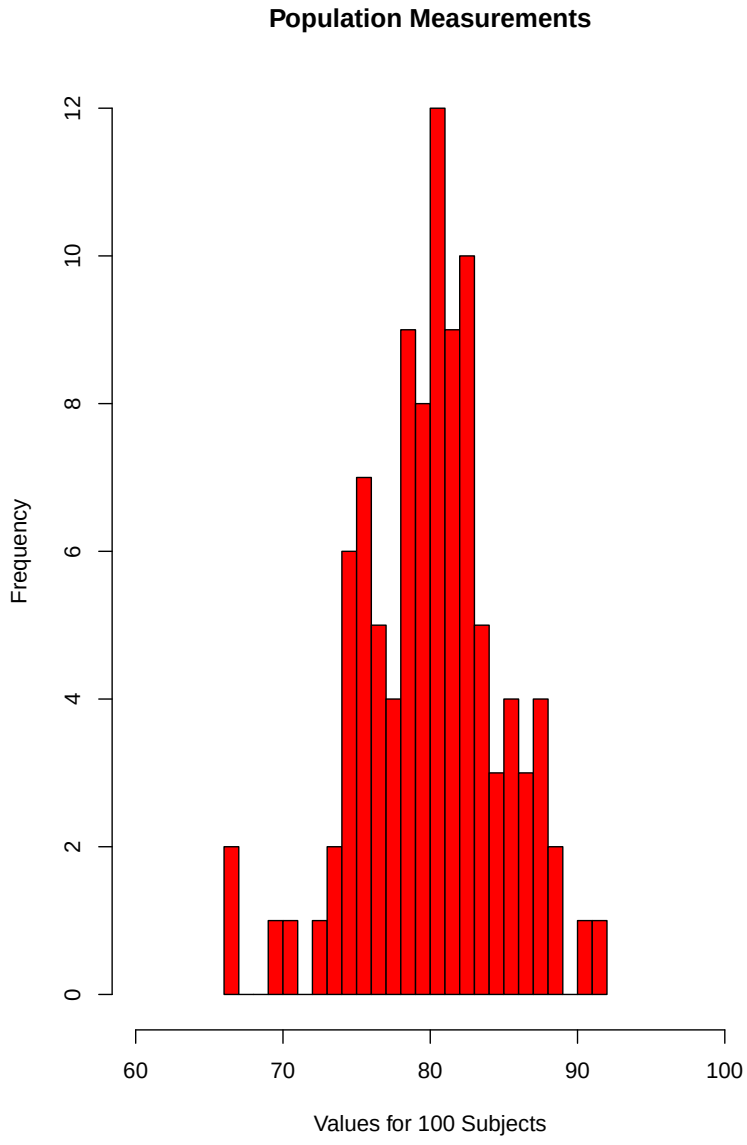
$$
\begin{aligned}
f_{X_1,\dots,X_n}(x_1, \dots, x_n) &= \prod_{i=1}^{n} f_X(x_i) \\
&= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
&= \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}
\end{aligned}
$$

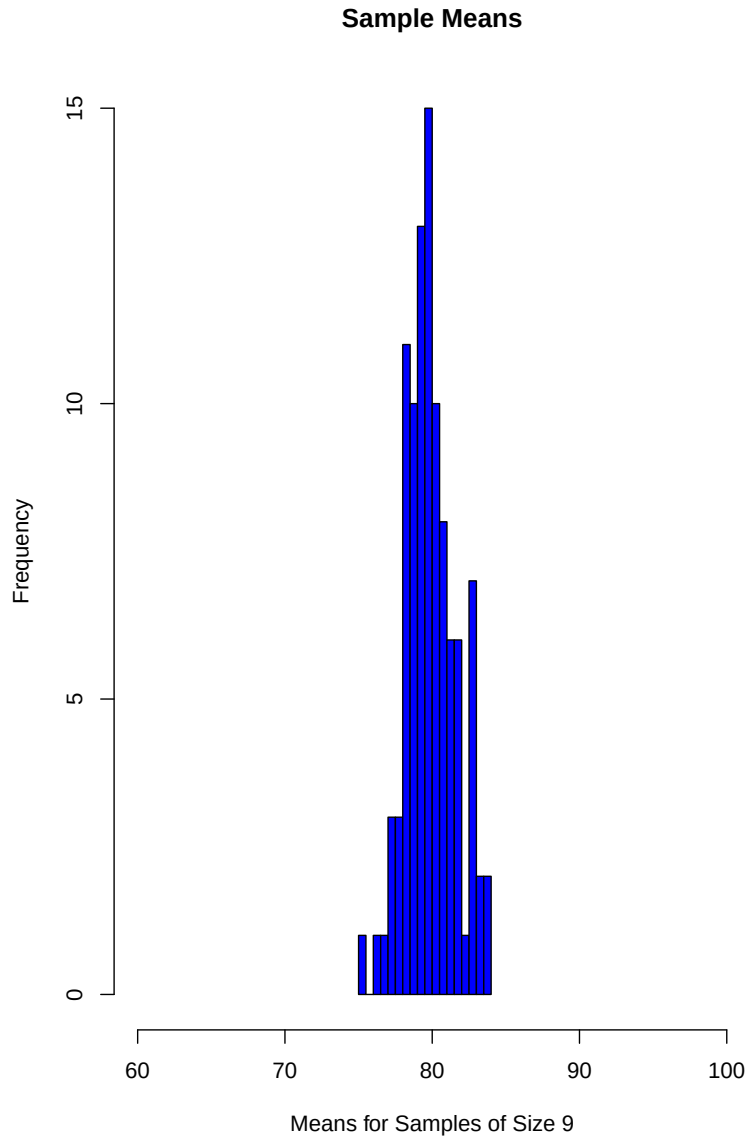
Basic Concepts

short glossary of estimation tools

- **Parameter**: numerical characteristic of the population that we are trying to recover (hence typically unknown)
  - Examples: λ in a Poisson,μ in a Gaussian and so on
- **Statistics**: numerical function of the sample that does not directly depend on any unknown parameter
  - Example: $S(X_1, \dots, X_n) = X_{(n)} - X_{(1)}$
- **Estimator**: a statistic used to estimate the population parameter
  - Example: $T(X_1, \dots, X_n) = \bar{X}$ is an estimator for μ
- **Estimate**: the value of an estimator corresponding to an *observed* sample:
  - Example: $T(x_1, \dots, x_n) = \bar{x}$ is an estimate corresponding to $\bar{X}$

Toy example


Population Measurements
Frequency
0
2
4
6
8
10
12
60
70
80
90
100
Values for 100 Subjects


Sample Means
Frequency
0
5
10
15
60
70
80
90
100
Means for Samples of Size 9

How do we define an estimator

The aim of the estimator is to try to recover the distribution that generated the data.

The are several *automatic* ways to derive an estimator, depending on how to use the data to recover the generating distribution.

ample at hand

- **Methods of Moments**:
  - find a distribution that has some features of the observed sample

- **Least squares**:

The Likelihood Function

the basic intuition

Let $X \sim \text{Binomial}(n, p)$, the probability mass function gives us the probability of observing a value x, once we know p and n.

Assume that $n = 10$ and we observe $x = 8$

- if $p = 0.5$, $P(X = 8) = \binom{10}{8}(0.5)^8(0.5)^2 = 0.043$
- if $p = 0.7$, $P(X = 8) = \binom{10}{8}(0.7)^8(0.3)^2 = 0.233$

For $x = 8$, the parameter $p = 0.7$ seems to be more likely than $p = 0.5$.

When we fix the realization x and we consider it a function of the parameter p, the p.m.f gives us a measure of **how compatible** x is with the value p.

This tells us how **plausible** a value of the parameter is, but it does not measure its **probability**.

The Likelihood Function

a little more formally

- $X_1, \ldots, X_n$ i.i.d random variables from a discrete distribution with parameter θ, and let $x_1, \ldots, x_n$ be an observed drawn from it. The **Likelihood function** $L(\theta; x_1, \ldots, x_n)$ corresponds to the **Probability Mass function** when taken to be a function of the parameter θ for a fixed value of $x_1, \ldots, x_n$.

- $X_1, \ldots, X_n$ i.i.d random variables from a continuous distribution with parameter θ, and let $x_1, \ldots, x_n$ be an observed drawn from it. The **Likelihood function** $L(\theta; x_1, \ldots, x_n)$ corresponds to the **Probability Density function** when taken to be a function of the parameter θ for a fixed value of $x_1, \ldots, x_n$.

The **log-likelihood function**, denoted by $l(\theta; x_1, \ldots, x_n)$ is the *logarithm* of the Likelihood function.

In conclusion: what does the likelihood do?

$$L(\theta; data) = P(data; \theta)$$

The equation above says that the probability density of the data given the parameters is equal to the likelihood of the parameters given the data. But despite these two things being equal, the likelihood and the probability density are fundamentally asking different questions — one is asking about the data and the other is asking about the parameter values

Maximum Likelihood Estimator

The **Maximum Likelihood Estimator (MLE)** is the value of the parameter that maximizes the Likelihood:

$$\hat{\theta}_{MLE} = \arg\max L(\theta; x_1, \dots, x_n) = \arg\max l(\theta; x_1, \dots, x_n)$$

Operationally the steps to find the **MLE** are:

1. **Compute the derivative** of the log-likelihood and equate it to 0: $dl(\theta; x_1 \dots, x_n)/d\theta = 0$
2. **Isolate** θ to find the candidate for the **MLE** (i.e. the critical point)
3. **Check the sign** of $d^2 l(\theta; x_1 \dots, x_n)/d\theta^2$ in the candidate θ to verify that this is not a min or a saddle

Example

Maximum Likelihood for the parameter λ of a Poisson:

Remember that if $X_1, \dots, X_n$ random sample, with $X_i \sim \text{Poisson}(\lambda)$ then:

- joint distribution

$$p_{X_1,\dots,X_n}(x_1, \dots, x_n; \lambda) = \prod_{i=1}^{n} \left[\frac{e^{\lambda} \lambda^{x_i}}{x_i!} \right] = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

- Likelihood

$$L(\lambda; x_1, \dots, x_n) = \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}$$

- log-Likelihood

$$l(\lambda; x_1, \dots, x_n) = \log \left(\frac{1}{\prod_{i=1}^{n} x_i!} \right) - n\lambda + \sum_{i=1}^{n} x_i \log(\lambda)$$

Example

Maximum Likelihood for the parameter λ of a Poisson:

1. Compute the derivative of $l(\lambda; x_1, \dots, x_n)$ and equate it to 0:

$$\frac{dl(\lambda; x_1, \dots, x_n)}{d\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

2. Isolate λ to get the MLE estimate:

$$-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \quad \iff \quad \hat{\lambda}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}_n$$

CAVEAT Even if $p_{X_1,\dots,X_n}(x_1, \dots, x_n; \lambda)$ denotes a discrete distribution, it is **a continuous function in** λ, hence we can compute derivatives to find the max.

Core of the Likelihood

The multiplicative factor **depending on the data** but **not on the parameter** $\frac{1}{\prod_{i=1}^{n} x_i!}$ disappeared when we computed the derivative. This is always true:

- if $L(\lambda; x) = h(x)g(x, \theta)$, then $l(\lambda; x) = \log(h(x)) + \log(g(x, \theta))$
- the derivative of $\log(h(x))$ does not depend on θ

$$\frac{dl(\theta; x)}{d\theta} = \frac{d\log(h(x))}{d\theta} + \frac{d\log(g(x, \theta))}{d\theta} = \frac{d\log(g(x, \theta))}{d\theta}$$

The function $g(x, \theta)$ is called the **core** of the likelihood and it contains all the information we need from the data.

Since **we can replace L with g without loss of information**, when we talk about *Likelihood* we actually talk about its *core*.

Maximum Likelihood: A brief recap

Likelihood based inference is *probably* the most widely used. In the example we maximized the likelihood function of a Poisson random variable, the same could be done for all the famous probability distributions analyzed so far. For the sake of brevity we report here a table containing, for each probability distribution, the corresponding MLE.

Table 1: MLE Estimators

Distribution	MLE
$Unif(0, \theta)$	$\max\{X_1, ..., X_n\}$
$Bin(n, p)$	$\sum_{i=1}^{n} X_i/n$
$Pois(\lambda)$	$\sum_{i=1}^{n} X_i/n$
$N(\mu, \sigma)$	$\sum_{i=1}^{n} X_i/n$
$Exp(\lambda)$	$\sum_{i=1}^{n} X_i/n$

Methods of Moments

for point estimation

The core idea is to *equate sample moments to population moments*, i.e.

$$\begin{cases} \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \mathbb{E}[X^2] = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ \mathbb{E}[X^3] = \frac{1}{n}\sum_{i=1}^{n} X_i^3 \\ \dots \end{cases}$$

Example:

Consider a random sample $X_1, \dots, X_n \sim \text{Unif}(0, \theta)$, for which $\mathbb{E}[X] = \theta/2$.

The MOM estimator is found by equating $\mathbb{E}[X] = \theta/2$ with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\theta/2 = \bar{X} \qquad \Rightarrow \hat{\theta}_{MOM} = 2\bar{X}$$

Exercise:

Let $X_1, \ldots, X_n \sim \text{Unif}(a, b)$

- compute the MOM estimator for a and b.
- find the maximum likelihood estimator
- What can we conclude about those estimators?

Remember that

$$X \sim \text{Unif}(a, b) \qquad \Rightarrow \mathbb{E}[X] = \frac{b+a}{2} \qquad \mathbb{V}[X] = \frac{(b-a)^2}{12}$$

Evaluating Point estimators

- An estimator T for a parameter θ, is said to be **unbiased** if $\mathbb{E}[T] = \theta$.
  - a "good" estimator is on average close to the real value of the parameter of interest

- An estimator T is **precise** if its variance $\mathbb{V}(T)$ is small.
  - a "good" estimator is *always* on target

The **Mean Squared Error** (MSE) evaluates the performance of the estimator combining these two desiderata:

$$MSE(T) = \mathbb{V}(T) + \text{Bias}(T)^2$$

MSE

- if $\mathbb{E}[T] = \theta$ we say that the estimator is **unbiased** and the MSE reduces to its variance

Consistency

- the MSE can be alternatively defined as

$$MSE(T) = \mathbb{E}[(T - \theta)^2]$$

- when

$$\lim_{n \to \infty} MSE(T) = 0$$

we have that as n grows T becomes closer and closer to real value of the parameter θ. This important property is called **consistency**, and reassures us that adding more observations improves the performances of the estimator

Bias and MSE: An example(I)

Let $X_1, X_x, ..., X_n \sim Bin(p)$. Consider the estimator given by $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- Compute its expected value.

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \times np = p$$

  The estimator is **unbiased**

- Compute the MSE (rember that MSE ie equal to variance if the estimator is unbiased:

$$\mathbb{V}[\bar{X}] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$$

Bias and MSE: An example(II)

Let $X_1, \ldots, X_n \sim (N, \sigma^2)$. Consider the two estimators:

- $T_1(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $T_2(X) = \frac{X_{(1)}+X_{(n)}}{2}$

Let us compute the bias:

- $\mathbb{E}[T_1(X)] = \frac{\mathbb{E}[X_1]+\mathbb{E}[X_2]}{2} = \frac{\mu+\mu}{2} = \mu$
- $\mathbb{E}[T_2(X)] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{n\mu}{n} = \mu$

The estimators are both unbiased (what does it mean?)

Bias and MSE: An example(III)

Let us compute the MSE which is equal to the variance

- $\mathbb{V}[T_1(X)] = \frac{1}{n^2}\mathbb{V}[X_i] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$
- $\mathbb{V}[T_2(X)] = \frac{1}{n^2}\mathbb{V}[X_1] + \mathbb{V}[X_n] = \frac{1}{2}(\sigma^2 + \sigma^2)$

T_1 has a smaller MSE; so it is better than T_2 (Did you expect that?)

Interval Estimates: General Idea.

With a point estimator we provide an estimate of the unknown parameter. We also to provide a set of plausible interval of values for the unknown parameter. To do that let us provide this results which is related to the Central Limit Theorem and states that, for large n

$$\frac{\sum_{i=1}^{n} X_i/n - \mu}{S/\sqrt{n}} \sim N(0,1)$$

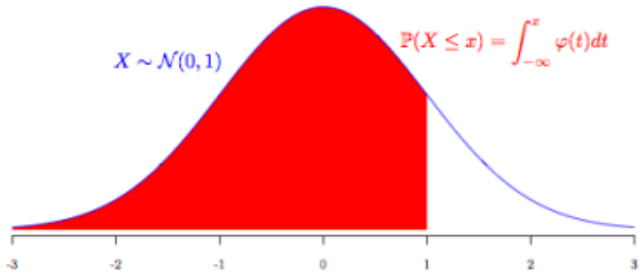
where S is the square root of the sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Interval estimate

Let us now observe that:

$$P\left(-1.96 < \frac{\sum_{i=1}^{n} X_i/n - \mu}{S/\sqrt{n}} < 1.96\right) = .95$$



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

Interval Estimates: How to build them?

$$P\left(-1.96 < \frac{\sum_{i=1}^{n} X_i/n - \mu}{S/\sqrt{n}} < 1.96\right) = .95$$

means that the the true unkown mean μ is contained within the interval with following probability

$$P\left(\sum_{i=1}^{n} X_i/n - 1.96S/\sqrt{n} < \mu < \sum_{i=1}^{n} X_i/n - 1.96S/\sqrt{n}\right) = .95$$

And thus the confidence interval for the mean at a condifence level of 95% is given by:

$$\left[\sum_{i=1}^{n} \frac{X_i}{n} - 1.96S/\sqrt{n}, \sum_{i=1}^{n} \frac{X_i}{n} + 1.96S/\sqrt{n}\right]$$

Interval Estimates

A Brief recap:

- The quantity $S/\sqrt{n}$ is called **standard error**
- 1.96 comes from the table of the standard normal distribution: the area under the standard normal density curve between -1.96 and 1.96 is 0.95
- If we want to increase the level of confidence of the interval we have to choose another value from the table of the standard Normal distribution. If we want to increase the level of confidency up 99% we ha have to choose the value such that the area under the curve is equal to 0.99

Interval Estimates (formally)

A **interval estimator** for a parameter θ is a random interval $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$, containing the most believable values for the parameter.

Intuitively, it is very difficult to predict the **exact** value of the unknown parameter (if T is a continuous random variable, this is even impossible, as by definition $P(T = \theta) = 0$), hence is more reasonable to ask for a range of possible parameters.

In addition a set of plausible values is more informative on the phenomenon than just a single guess.

The ingredients

A **confidence interval of level** $1-\alpha$ is a random interval $[L, U]$, where L and U are two *statistics*, such that

$$P(\theta \in [L, U]) = 1 - \alpha$$

The **confidence level** $(1-\alpha)$ is probability that the interval contains the true value of the parameter θ, *before the sample is observed*. Typically this value is chosen to be high (0.95 or 0.99).

Typically a confidence interval is built using the formula

$$T \pm err$$

where T is the point estimator for θ and err measures how accurate the point estimate is and depends on the level of confidence as well as $\mathbb{V}[T]$.

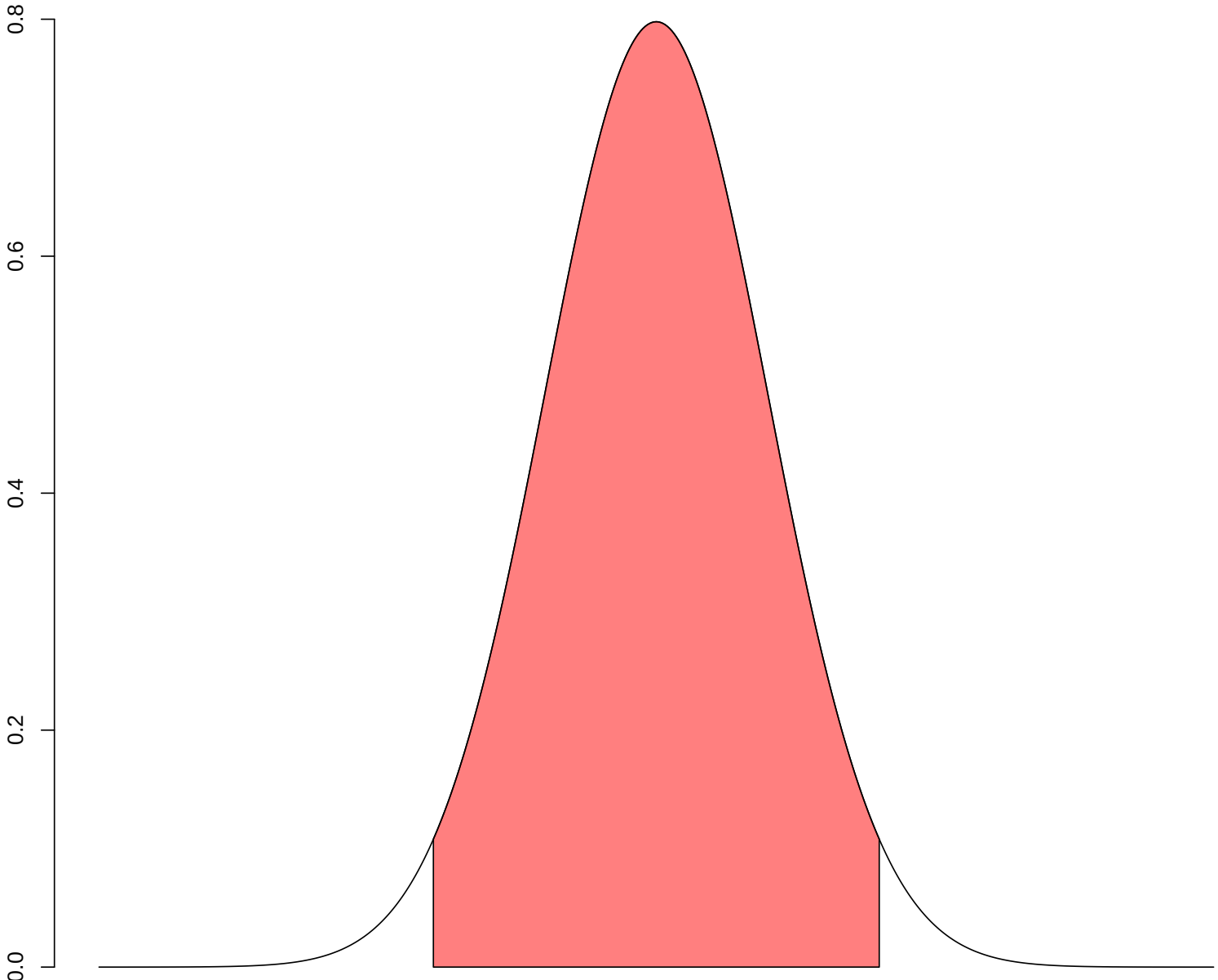
Confidence

a word of caution

BE CAREFUL: once we observe the sample, and we have an *estimate* of the confidence interval $[l, u]$, the probability that the parameter lies in this interval is either 0 or 1.

However, remembering the definition of probability as the limit of the relative frequency of an event, we can be **confident** that if we build a large number of confidence intervals, the parameter will be contained in the 95% of them.

Toy example



Exercise

- When a General Social Survey asked 1326 subjects, "Do you believe in science?", the proportion who answered yes was 0.82.

1. Construct the 95% confidence interval. Interpret it in context. **Here we are dealing with a proportion. A proportion is exactely like a sample mean. The values that we are averaging are either 0 or 1. Assuming that p is the **true** proportion in the population and $\hat{p}$ estimated proportion, the following holds:

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

Thus the confidence interval for p at a confidence level of 95%, as for the sample mean, is given by:

$$\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Exercise

1. How does the result in (2) change if you construct a 99% confidence interval?
2. Another source claims, "75% of people believe in science." Does the confidence intervals support this claim?
3. Describe the effect of the sample size on the confidence interval.

Hypothesis Testing

The main goal of **statistical testing** is to check whether the data support certain statements (**hypothesis**), usually expressed in terms of population parameters for variables measured in the study.

Usually, an *hypothesis* on the parameter θ is formalized as follows:

- $\theta = \theta_0$ *punctual* hypothesis
- $\theta \geq \theta_0$ or $\theta \leq \theta_0$ *one-sided* hypothesis
- $\theta \neq \theta_0$ *two-sided* hypothesis

Hypothesis

In a **hypothesis test** we compare two alternative hypothesis H_0 and H_1:

- The **Null Hypothesis** (H_0) is the hypothesis that is held to be true unless sufficient evidence to the contrary is obtained.
- The **Alternative Hypothesis** (H_1) represent the new theory we would like to test.

Example: We want to test whether an astrologer can correctly predict which of 3 personalities charts applies to a person.

- $H_0 : p = 1/3$
  - the astrologer doesn't have any predictive power (the probability of guessing the personality is $1/3$)
- $H_1 : p \geq 1/3$
  - the astrologer does have predictive power

Test logic

Innocent until proven guilty

	H_0 is true	H_0 is false
Accept H_0		Type II Error
Reject H_0	Type I Error	

- If we want to completely avoid Type II Error we should **always Reject** H_0
- If we want to completely avoid Type I Error we should **always Accept** H_0

It is impossible to simultaneously avoid both: which one is more important?

As H_0 represent the current condition, we would like to subvert it only when the data provide strong evidence against it

Testing procedure:

How to solve a test $H_0 = \theta \leq \theta_0$ versus $H_1 = \theta > \theta_0$:

1. Choose a level α of significance (i.e. the probability of Type I Error), typically $\alpha = 0.05$
2. Choose a test statistic T, i.e. a statistic that describes how far that point estimate falls from the parameter value given in the null hypothesis
3. Given an observed sample $(x_1, \dots, x_n)$, compute the $t = T(x_1, \dots, x_n)$
4. Compute the p-value, $P(T > t|H_0) = p$, a measure of how compatibles the data are with H_0
5. If $p \leq \alpha$, reject H_0, otherwise do not reject it

Toy Example

- A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.
  - **Step 1**: State the Null hypothesis. The accepted fact is that the population mean is 100, so: $H_0 : \mu = 100$
  - **Step 2**:State the alternative hypothesis. The claim is that the students have above average IQ scores, so: $H_1 : \mu > 100$
  - **Step 3**: Find the rejection region area (given by your α level equal to 0.05) from the z-table. An area of 0.05 is equal to a z-score of 1.645.
  - **Step 4**:Find the test statistic using this formula: $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = 4.56$
  - **Step 5**:The value of Z is greater than z_α (4.56 $>$ 1.645), so you can reject the null.

Exercise

In a sample of 402 TorVergata first-year students, 174 are enrolled into Statistics course.

1. Find the sample proportion.
2. Is the proportion of students enrolled into Statistics course in the population of all Tor Vergata first-year students different from 0.50 at the significance level $\alpha = 0.05$?