Statistical Inference

Dr. João Victor da Fonseca Pinto

November 18, 2022

Table of contents

Introduction

In real life, we work with data that are affected by randomness, and we need to extract information and draw conclusions from the data.

- Suppose that we would like to predict the outcome of an election. Since we cannot poll the entire population, we will choose a random sample from the population and ask them who they plan to vote for. In this experiment, the randomness comes from the sampling.
- Note also that if our poll is conducted one month before the election, another source of randomness is that people might change their opinions during the one month period.

Introduction

- **Frequentist (classical) Inference:** Based on the samples (random variable X) we can propose some distribution with unknown (fixed) parameter θ that best describe the population.
- **Bayesian Inference:** Inthe bayesian approch, θ is a random variable.

Introduction: Basic Statistical Concepts

Definition

A random variable X is a function from the sample space to the real numbers.

$$X : S \to \mathbb{R}$$

Toss a coin five times. This is a random experiment and the sample space can be written as

$$S = TTTTT, TTTTH, ..., HHHHHH$$

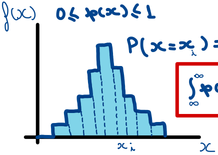
in this experiment, we are interested in the number of heads. We can define a random variable X whose value is the number of observed heads.

$$X = [0, 1, 2, 3, 4, 5]$$

depending on the outcome of the random experiment.

Introduction: Basic Statistical Concepts

x discrete:

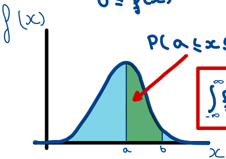


$0 \leq p(x) \leq 1$

$P(x=x_i) = p(x_i)$

$$\boxed{\int_{\infty}^{\infty} p(x)\,dx = 1}$$

x continuous:



$0 \leq f(x)$

$P(a \leq x \leq b) = \int_a^b f(x)\,dx$

$$\boxed{\int_{-\infty}^{\infty} f(x)\,dx = 1}$$

Expected Mean:

$$\mu_x = E[X] = \begin{cases} \int_{-\infty}^{\infty} x f(x)\,dx, & x \text{ continuous} \\ \sum_{\text{all } x} x\, p(x), & x \text{ discrete} \end{cases}$$

Expected variance:

$$\sigma^2 = V[x] = E[(x-\mu_x)^2] = \begin{cases} \int_{-\infty}^{\infty} (x-\mu_x)^2 f(x)\,dx, & x \text{ continuous} \\ \sum_{\text{all } x} (x-\mu_x)^2 p(x), & x \text{ discrete} \end{cases}$$

Introduction: Basic Statistical Concepts

let c is a constant:

- $\boxed{E[c] = c}$
- $\boxed{E[cx] = c\mu_x}$
- $\boxed{V[c] = 0}$
- $V[x] = E[(x-\mu_x)^2] =$
  $E[x^2 - 2x\mu_x + \mu_x^2] =$
  $E[x^2] - 2\mu_x \underset{\mu_x}{\cancel{E[x]}} + \underset{\mu_x^2}{\cancel{E[\mu_x^2]}}$

  $\boxed{V[x] = E[x^2] - \mu_x^2}$

- $V[\underbrace{cx}_{y}] = E[(y-\mu_y)^2] =$
  $E[y^2] - \mu_y^2$
  $E[(cx)^2] - (E[cx])^2$
  $c^2E[x^2] - c^2(E[x])^2$
  $c^2(E[x^2] - \mu^2)$

  $\boxed{V[cx] = c^2 V[x]}$

if we have two random variables $x \begin{cases} E[x] = \mu_x \\ V[x] = \sigma_x^2 \end{cases}$ $y = \begin{cases} E[y] = \mu_y \\ V[y] = \sigma_y^2 \end{cases}$

- $E[x+y] = E[x] + E[y] = \mu_x + \mu_y$
- $Cov(x,y) = E[(x-\mu_x)(y-\mu_y)]$
- $V[\underbrace{x+y}_{z}] = E[z^2] - z^2 = E[(x+y)^2] - (E[(x+y)])^2$
  $= E[(x^2 + 2xy + y^2)] - (\mu_x^2 + 2\mu_x\mu_y + \mu_y^2)$
  $= E[x^2] + 2E[xy] + E[y^2] - \mu_x^2 - 2\mu_x\mu_y - \mu_y^2$
  $= V[x] + V[y] + 2\,Cov(x,y)$
- If x and y are independent, $Cov(x,y) = 0$

Sampling from Pratical Example

Our goal is to investigate the height distribution of people in a well defined population We choose a random sample of size n **with replacement from the population** and let X_i be the height of the ith chosen person. More specifically,

- We chose a person uniformly at random from the population and let X_1 be the height of that person. Here, every person in the population has the same chance of being chosen.
- o determine the value of X_2, again we choose a person **uniformly** (and **independently from the first person**) at random and let X_2 be the height of that person. Again, every person in the population has the same chance of being chosen.
- In general, X_i is the height of the ith person that is **chosen uniformly and independently** from the population.

Sampling

- You might ask why do we do the sampling with replacement?
- In practice, we often do the sampling without replacement, that is, we do not allow one person to be chosen twice. If the population is large, then the probability of choosing one person twice is extremely low.
- The big advantage of sampling with replacement is that X_i's will be independent and this makes the analysis much simpler.

Sampling

Definition

The collection of random variables $X_1, X_2, ..., X_n$ is said to be a **random sample** of size n if they are independent and identically distributed (i.i.d.), i.e.,

- $X_1, X_2, ..., X_n$ are independent **random variables**, and
- they have the same distribution, i.e,

$$f_{X_1}(x) = f_{X_2}(x) = ... = f_{X_n}(x) \text{ for all } x \in \mathbb{R}$$

Sampling: Sample Mean and Sample Variance

Statistical inference makes considerable use of quantities computed from the observations in the sample.

- **Sample mean:** The mean **estimator** is defined as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Sample variance:** The variance **estimator** is defined as

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

Sometimes $S = \sqrt{S^2}$, called the **sample standard deviation**, is used as a measure of dispersion.

Point Estimation

A point estimator is a random variable.

- The **sample** mean ($\overline{x}$) is a **point estimation** of the **population** mean (μ).
- The **sample** variance (S^2) is a **point estimation** of the **population** variance (σ^2).

Several properties are required of **good point estimators**. Two of the most important are the following:

- **Unbiased:** The expected value of the point estimator should be equal to the parameter that is being estimated.
- **Minimum variance:** The variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter.

We may easily show that $\overline{x}$ and S^2 are unbiased estimators of μ and σ^2, respectively.

$\bar{x}$ is an unbiased estimator

$$E[\bar{x}] = E\left[\frac{1}{m}\sum_{i=1}^{m} x_i\right]$$

$$E[\bar{x}] = \frac{1}{m} E\left[\sum_{i=1}^{m} x_i\right]$$

$$E[\bar{x}] = \frac{1}{m}\sum_{i=1}^{m} E[x_i]$$

$$E[\bar{x}] = \frac{1}{m}\cdot m\mu = \mu$$

$$\boxed{E[\bar{x}] = \mu}$$

S^2 is a unbiased estimator

$$E[S^2] = E\left[\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{x})^2\right]$$

$$E[S^2] = \frac{1}{m-1}E\left[\sum_{i=1}^{m}(x_i-\bar{x})^2\right]$$

$$= \frac{1}{m-1}E\left[\sum_{i=1}^{m}(x_i^2 - 2\bar{x}x_i + \bar{x}_i^2)\right]$$

$$= \frac{1}{m-1}E\left[\sum_{i=1}^{m}x_i^2 - 2\bar{x}\cancelto{m\bar{x}}{\sum_{i=1}^{m}x_i} + \underset{m\bar{x}^2}{\cancel{\sum_{i=1}^{m}\bar{x}^2}}\right]$$

$$= \frac{1}{m-1}E\left[\sum_{i=1}^{m}x_i^2 - m\bar{x}^2\right]$$

$$= \frac{1}{m-1}\left(\sum E[\underset{\sigma^2+\mu^2}{\cancel{x_i^2}}] - E[m\bar{x}^2]\right) \quad \rightarrow E\left[\left(\frac{1}{m}\sum_{i=1}^{m}x_i\right)^2\right]$$

$$\frac{1}{m-1}\left(m(\sigma+\mu) - mE[\bar{x}^2]\right)$$

$$\boxed{E[S^2] = \sigma^2}$$

The Normal Distribution

- The probability distribution of a statistic is called a **sampling distribution**.
- the most important sampling distributions is the normal distribution. If x is a normal random variable, the probability distribution of x is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{x-\mu}{\sigma}\right)^2} \quad , \infty < x < \infty \Rightarrow \; N(\mu, \sigma^2)$$

- We can also map any random variable x described by the normal distribution to the standard normal distribution, using

$$z = \frac{x-\mu}{\sigma} \text{ where we map } \; N(\mu, \sigma^2) \rightarrow \; N(0, 1)$$

The Central Limit Theorem (CLT)

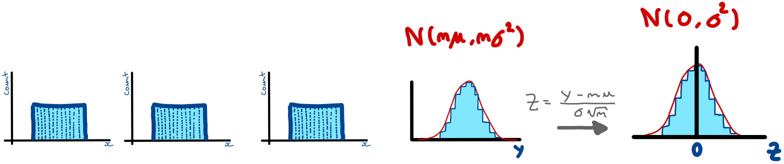
Definition

If $X_1, X_2, ..., X_N$ is a sequence of n independent and identically distributed random variables with $E[X_i] = \mu$ and $V[X_i] = \sigma^2$ (both finite) and $Y = X_1 + X_2 + ... + X_n$, then the limiting form of the distribution of

$$Z_n = \frac{Y - n\mu}{\sigma\sqrt{n}}$$

as $n \to \infty$, is the **standard normal distribution**.

The Central Limit Theorem (CLT)



$$\underbrace{\begin{bmatrix} * \\ * \\ * \\ \vdots \\ * \\ * \end{bmatrix}}_{X_1} + \underbrace{\begin{bmatrix} * \\ * \\ * \\ \vdots \\ * \\ * \end{bmatrix}}_{X_2} + \cdots + \underbrace{\begin{bmatrix} * \\ * \\ * \\ \vdots \\ * \\ * \end{bmatrix}}_{X_m} = \underbrace{\begin{bmatrix} * \\ * \\ * \\ \vdots \\ * \\ * \end{bmatrix}}_{y}$$

$$\boxed{z = \frac{y - m\mu}{\sigma\sqrt{m}}}$$

$E[X_1] = \mu$ $\quad$ $E[X_2] = \mu$ $\quad$ $E[X_m] = \mu$ $\quad$ $E[y] = m\mu$

$V[X_1] = \sigma^2$ $\quad$ $V[X_2] = \sigma^2$ $\quad$ $V[X_m] = \sigma^2$ $\quad$ $E[y] = m\sigma^2$

$$E[y] = E[x_1 + x_2 + \dots + x_m] = E[x_1] + E[x_2] + \dots + E[x_m] = m\mu$$

$$V[y] = V[x_1 + x_2 + \dots + x_m] = V[x_1] + V[x_2] + \dots + V[x_m] = m\sigma^2$$

Confidence Intervals

Definition (Interval estimation)

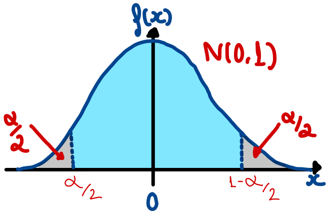
Let X_1, X_2, ..., X_n be a random sample from a distribution with parameters θ that is to be estimated. As **interval estimator** with **confidence level** $1-\alpha$ consists of two estimators $\hat{\theta}_h(X)$ and $\hat{\theta}_h(X)$ such that

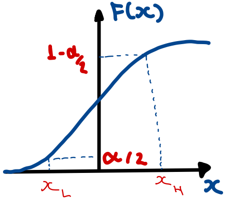
$$P(\hat{\theta}_l \leq \theta \leq \hat{\theta}_h) \geq 1-\alpha$$

for every possible value of θ.

Confidence Intervals for large samples: Example

Let $Z \to N(0,1)$, find x_l and x_h such that $P(x_l \leq Z \leq x_h) = 0.95$



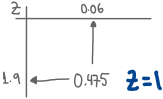


$$P(x_L \leq Z \leq x_H) = 0.95$$

$$\begin{cases} P(Z \leq x) = \alpha/2 \\ P(Z \leq x) = 1 - \alpha/2 \end{cases}$$

$$\begin{cases} F_X(x_L) = \alpha/2 \rightarrow x_L = F_X^{-1}(\alpha/2) \\ F_X(x_H) = 1 - \alpha/2 \rightarrow x_H = F_X^{-1}(1-\alpha/2) \end{cases}$$

CDF for standard normal distribution is given



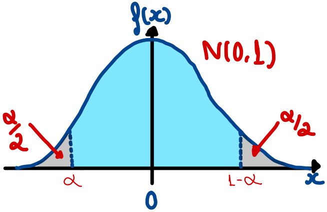
$z = 1.96$

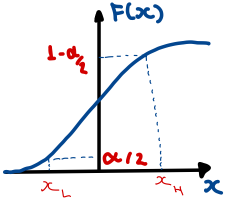
$$F_X^{-1}(0.025) = x_L \rightarrow \boxed{x_L = -1.96}$$

$$F_X^{-1}(0.975) = x_H \rightarrow \boxed{x_H = 1.96}$$

Confidence Intervals: Example

Let $x_1, x_2, ..., x_i$ be a random sample from a normal distribution $(N(\theta, 1))$. Find 95% confidence interval for θ





$$P(x_L \leq z \leq x_H) = 0.95$$

$$\begin{cases} P(z \leq x) = \alpha/2 \\ P(z \leq x) = 1 - \alpha/2 \end{cases}$$

$$\begin{cases} F_x(x_L) = \alpha/2 \rightarrow x_L = F_x^{-1}(\alpha/2) \\ F_x(x_H) = 1 - \alpha/2 \rightarrow x_H = F_x^{-1}(1 - \alpha/2) \end{cases}$$

CDF for standard normal distribution is given

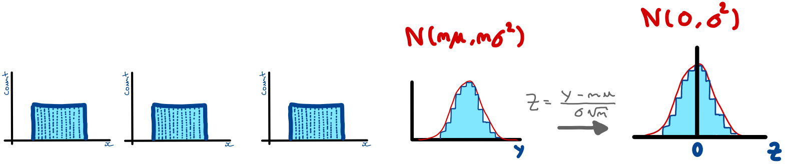
z	0.06
1.9	0.475

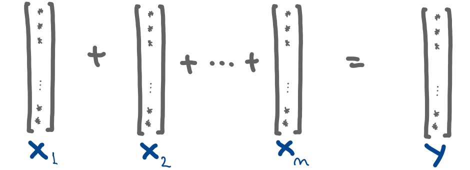
$z = 1.96$

$$F_x^{-1}(0.025) = x_L \rightarrow \boxed{x_L = -1.96}$$

$$F_x^{-1}(0.975) = x_H \rightarrow \boxed{x_H = 1.96}$$

Confidence Intervals





$$\boxed{z = \frac{y - m\mu}{\sigma\sqrt{m}}}$$

$E[X_1] = \mu$ $\quad E[X_2] = \mu$ $\quad E[X_m] = \mu$ $\quad E[y] = m\mu$

$V[X_1] = \sigma^2$ $\quad V[X_2] = \sigma^2$ $\quad V[X_m] = \sigma^2$ $\quad E[y] = m\sigma^2$

$$E[y] = E[x_1 + x_2 + \dots + x_m] = E[x_1] + E[x_2] + \dots + E[x_m] = m\mu$$

$$V[y] = V[x_1 + x_2 + \dots + x_m] = V[x_1] + V[x_2] + \dots + V[x_m] = m\sigma^2$$

Hypothesis Testing

A statistical hypothesis is a statement either about the parameters of a probability distribution or the parameters of a model.

- Given two experiments with mean μ_1 and μ_2. We can reflects some conjecture about the problem situation.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

- The statement $H_0 : \mu_1 = \mu_2$ is called the **null hypothesis** and should be always **equal**.
- otherwise, $H_1 : \mu_1 \mu_2$ is called the **alternative hypothesis**.

Hypothesis Testing

To test a hypothesis, we devise a procedure for

- Taking a random sample,
- computing an appropriate **test statistic**,
- and then **rejecting or failing to reject** the null hypothesis H_0 based on the computed value of the test statistic.

Rejecting the **null hypothesis**:

- The set of values for the test statistic that leads to **rejection of** H_0 is called the **critical region** or rejection region for the test.
- Two kinds of errors may be committed when testing hypotheses.

Hypothesis Testing: Error Types

- If the null hypothesis is rejected when it is true, a type I error has occurred.

  $$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$

  where α is called the **significance level** of the test,
- If the null hypothesis is not rejected when it is false, a type II error has been made.

  $$\beta = P(\text{type II error}) = P(\text{accept } H_0 | H_0 \text{ is false})$$
- Sometimes it is more convenient to work with the **power** of the test, where

  $$\text{Power} = 1 - \beta = P(\text{reject } H_0 | H_0 \text{ 0s false})$$

Hypothesis Testing: Error Types

truth

		H_0 false	H_0 true
test	Reject H_0	Power $1 - \beta$	type I error (α)
	Accept H_0	type II error (β)	$1 - \alpha$

Hypothesis Testing: P-Value

Definition

- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- In a significance test, the null hypothesis H_0 is rejected if the p-value is less than or equal to a predefined threshold value α, which is referred to as the significance level.

Intuitively, if the P-Value is small, it means that the observed data is very unlikely to have occurred under H_0

Hypothesis Testing: Chi-Square Goodness of Fit Test

- The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not.

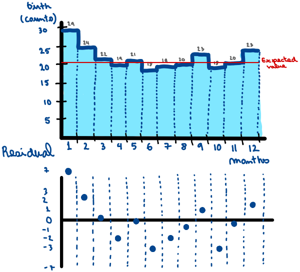
  $$\chi^2_{df} = \sum_{i=1}^{n} \frac{(x_i - E[x_i])^2}{E[x_i]}$$

  where *df* is the degrees of freedom and *n* is the number of observations.

- Data values that are a simple random sample from the full population.
- Categorical or nominal data. The Chi-square goodness of fit test is not appropriate for continuous data. Use histograms instead where each bin (range value) is a categorial class.
- A data set that is large enough so that at least 30 values are expected in each of the observed data categories.

Chi-Square Goodness of Fit Test: Birth Example

nº of bins: 12 total of observations: 256



month	Observed	Expected	Residual	$(\text{Obs.}-\text{Exp.})^2$	X^2
1	29	21.33	7.67	58.79	2.75
2	24	21.33	2.67	7.11	0.33
3	22	21.33	0.67	0.45	0.02
4	19	21.33	-2.33	5.44	0.25
5	21	21.33	-0.33	0.11	0.005
6	18	21.33	-3.33	11.11	0.52
7	19	21.33	-2.33	5.44	0.25
8	20	21.33	-1.33	1.77	0.08
9	23	21.33	1.67	2.77	0.13
10	18	21.33	-3.33	11.10	0.52
11	20	21.33	-1.33	1.77	0.08
12	23	21.33	1.67	2.77	0.13
					$X^2=5.09$

degrees of freedom: 11

Chi-Square statistic : 5.09

$$\boxed{X^2_{11} = \sum_{i=1}^{12} \frac{(O_i - E_i)^2}{E_i}}$$

Degrees of freedom for Chi-square goodness of fit test is equal to the number of groups (bins) minus one.

Chi-Square Goodness of Fit Test: Birth Example

- Now, let's test the hypothesis with 5% of significance level (or 95% of confidence)

$$H_0 : \text{The births are uniformly distributed}$$

$$H_1 : \text{The births are not uniformly distributed.}$$

- To reject the null hypothesis we need to find the p-value for the chi-square test.
- Using the chi-square calculator with significance level (α) of 0.05 we have:

$$\chi^2_{11} = 5.09 \Rightarrow \text{ p-value} = 0.92673$$

- Since the p-value is higher than α, we can **accept the null hypothesis**

Bayesian Statistical Inference

- The goal is to draw inferences about the **unknown** variable X by observing a related random variable Y. The unknown variable is modeled as a random variable X, with **prior distribution**

$$f_X(x), \text{ if } X \text{ is continuous,}$$

$$P_X(x), \text{ if } X \text{ us discrete}$$

- After observing the value of the random variable Y, we find the **posterior** distribution of X. This is the conditional pdf (or pmf) of X giben $Y = y$

$$f_{X|Y}(x|y) \;\text{ or }\; P_{X|Y}(x|y)$$

- The posterior distribution is usually fount using **Bayes' formula**. Using the posterior distribution, we can then find the point or interval estimates of X.

Bayesian Statistical Inference

- Let X be the random variable whose value **we try to estimate**. Let Y be the **observed random variable**.
- That is, we have observed $Y = y$, and we would like to estimate X. Assuming both X and Y are discrete, we can write

  $$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

  The above equation, as we have seen before, is just one way of writing Bayes' rule
- For the continuous case we can write:

  $$f_{\theta|Y}(x|y) = \frac{f_{X|\theta}(x|\theta)f_{\theta}(\theta)}{f_X(x)}$$