RCDS Statistics I – Imperial College London

Introduction to sampling and hypothesis testing

Jesús Urtasun Elizari

December 2024

Random sampling and hypothesis testing

Chapter 1. Random events and probability theory

i) Definition of random events and probability.

ii) Discrete probability; "mass" distributions.

iii) Continuous probability; "density" distributions.

Chapter 2. Expected values and the central limit theorem

i) Mean, variance as expected values.

ii) Parameter estimation, prediction vs inference.

iii) the law of large numbers (LLN).

iv) the central limit theorem (CLT).

Chapter 3. Hypothesis testing

i) Prediction vs inference. Formulate hypothesis.

ii) General approach. Significance and p-values.

iii) Real example { Compare mean to expected value (1 sample T)
                  { Compare two means (2 sample T)

1.1 Random events and probability

* Random events ("stochastic", from στοχαστικός "to guess")
  Something whose output we don't know

* Probability: number $\in [0,1]$ quantifying certainty / "surprise".
  Example: tossing coins (H, T)
  $P(H) = 0 \rightarrow$ Certain I will never get H
  $P(H) = 1 \rightarrow$ " always get H
  $0 < P(H) < 1 \rightarrow$ $\exists$ level of uncertainty / "surprise"

* Unitarity: the sum of probabilities for all possible outcomes x_i must add up to 1.

$$\sum_{\forall x_i} P(x_i) = 1$$

Example: tossing coins

$$P(H) + P(T) = \frac{1}{2} + \frac{1}{2} = 1 \quad \checkmark$$

Example: rolling dice

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = \frac{1}{6} + \frac{1}{6} + \dots + \frac{1}{6} = 1 \quad \checkmark$$

* A bit of history

i) Girolamo Cardano (1501 - 1576)

Probability in games and dice, Chance.

ii) Pierre de Fermat (1601 - 1665)

Blaise Pascal (1623 - 1662)

Foundations of mathematical probability theory.

iii) Pierre S. Laplace (1749 - 1827)

Frequentist definitions of probability.

P as fraction of favorable / possible outcomes.

iv) Andrey Kolmogorov (1903 - 1987)

Axiomatic foundations (1933)

"Foundations of theory of probability"

$P \in [0,1]$; finiteness. (*)

1.2 Discrete probability distributions

* Discrete: number of possible outcomes is a finite number. { coins, dice, counting }

i) Binomial distribution

Jacob Bernoulli; "Ars conjectandi" (1713)

"How many times X I get a specific result in n trials, if the probability of each success is p"

$$B(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

x: number of successes
n: number of trials
p: probability each success

Example: Probability of 5 times # tossing 10 times a coin

$$B\left(5; 10, \frac{1}{2}\right) = \binom{10}{5}\left(\frac{1}{2}\right)^5 \left(1-\frac{1}{2}\right)^{(10-5)} = 0.246$$

Exercise

Example: Probability of 3 times a 6 in 10 dice

$$B\left(3; 10, \frac{1}{6}\right) = \binom{10}{3}\left(\frac{1}{6}\right)^3 \left(1-\frac{1}{6}\right)^{(10-3)} = 0.155$$

Exercise

Example: Probability of passing exam (A, B, C) of 10 quest.

$$B\left(5; 10, \frac{1}{3}\right) = \binom{10}{5}\left(\frac{1}{3}\right)^5 \left(1-\frac{1}{3}\right)^{(10-5)} = 0.137$$

Exercise

* What if $x = n$?

Example : probability of tossing 10 heads out of 10

$$B(10; 10, \tfrac{1}{2}) = \binom{10}{10} \left(\tfrac{1}{2}\right)^{10} \left(1-\tfrac{1}{2}\right)^{0}$$

$$= 1 \cdot \left(\tfrac{1}{2}\right)^{10} \cdot 1 = \left(\tfrac{1}{2}\right)^{10} \checkmark$$

* General : recover the individual probability, raised to the number of attempts.

$$B(n; n, p) = \underbrace{\binom{n}{n}}_{1} p^{n} \underbrace{(1-p)^{0}}_{1} = p^{n}$$

iii) Poisson distribution

Simeon D. Poisson; "Research on probability" (1837)

"How many times I observe an event in an interval, know λ"

$$P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

x: number of times I observe

λ: observed average.

Example: Probability of observing 3 same patients, with $\lambda = 5$.

$$P(3;5) = \frac{e^{-5}5^{3}}{3!} = 0.140$$

Exercise

Example: Probability of observing 5 or less patients in same λ.

$$P(x \leq 5;5) = P(1) + P(2) + \ldots + P(5)$$

$$= \sum_{x=1}^{5} P(x;\lambda=5) = 0.616$$

Exercise

Cumulative distribution function cdf

Example: Probability of more than 5

$$P(x > 5;5) = 1 - cdf(5) = 0.394$$

↑ Unitarity

Exercise

1.3 Continuous distributions

* Continuous : amount of possible outcomes is infinite / uncountable

Coin: 2 outcomes $(H, T) \rightarrow p = \frac{1}{2}$

Dice: 6 outcomes $(1,2,3,4,5,6) \rightarrow p = \frac{1}{6}$

} Discrete cases "Mass dist." Frequentist definition ✓

Continuous variable $(T, h, \text{conc.})$: ∞ outcomes $\rightarrow p = \frac{1}{\infty} = 0$ WTF

↳ Frequentist approach does not work

Need a new mathematical object "Density distribution"

i) Discrete case

Probability $P \in [0,1]$

Unitarity $\sum_{\forall x_i} P(x_i) = 1$

ii) Continuous case

Density $f(x)$

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

* Gaussian

i) Gaussian distribution

Carl Friedrich Gauss; "theory of movement of celestial bodies" (1809)

"Random variable x centered at some μ and spreaded σ"

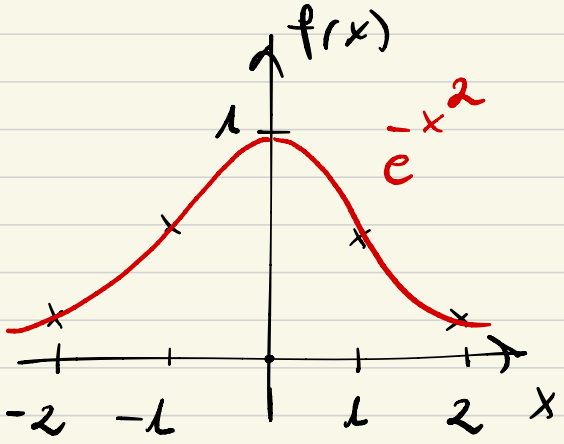
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

μ : mean value

σ : standard deviation

* Consider simplest case $\mu = 0$; $\sigma = 1$; ignore normalization factor

$$f(x) = e^{-x^2}$$; plot some values



$$x = 0;\ f(0) = e^{-0} = 1$$

$$x = 1;\ f(1) = e^{-1} = \frac{1}{e} = 0.37$$

$$x = -1;\ f(-1) = e^{-(-1)^2} = \frac{1}{e} = 0.37$$

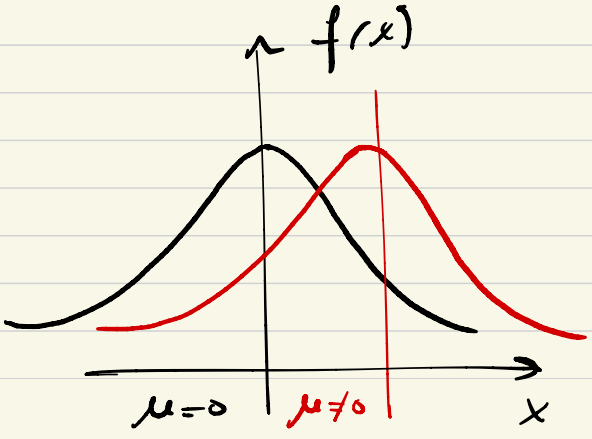
$$x = \pm 2;\ f(\pm 2) = e^{-4} = \frac{1}{e^4} = 0.018$$

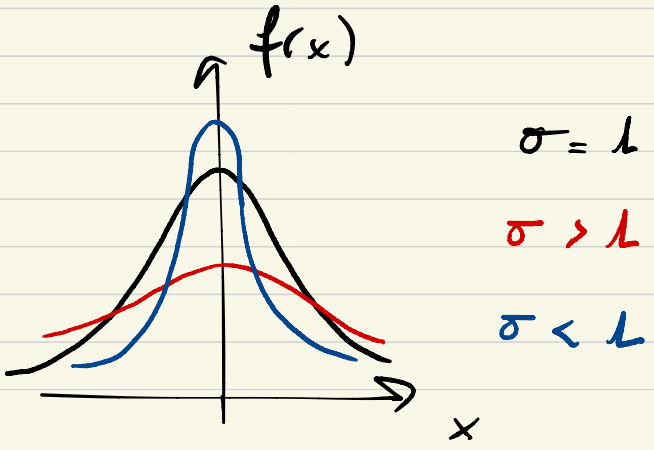
* Switch back on μ and σ

$$f(x) = e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

μ will displace the central value

σ will drive sharper/wider shape





* Add normalization factor

i) Remember discrete probability: Necessity $\sum_{\forall x_i} P(x_i) = 1$

ii) Continuous case. Density distribution $\int_{-\infty}^{+\infty} dx\, f(x) = 1$

$$\int_{-\infty}^{\infty} dx\; e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \sigma\sqrt{2\pi}$$

(*) Exercise

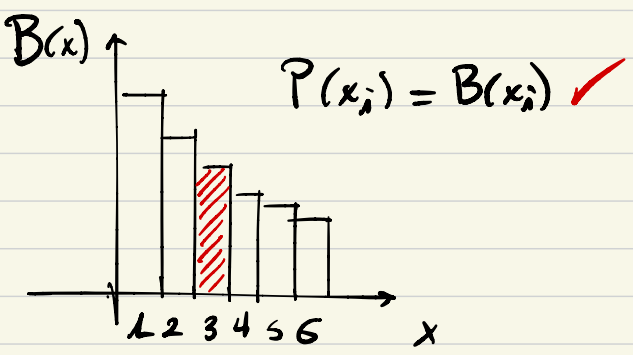
* for $f(x)$ to behave as a probability, we add a normalization factor

$$\left\{ f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \right\} \quad \left\{ \begin{array}{l} \text{Gaussian distribution;} \int_{-\infty}^{\infty} dx\, f(x) = 1 \\ \text{(Normal distribution)} \end{array} \right.$$

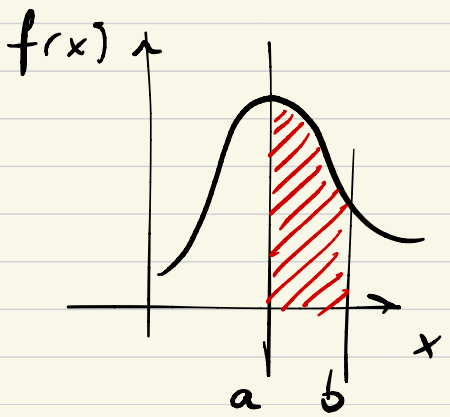
* Particular case $\mu = 0$; $\sigma = 1$

$$\left\{ f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2} \right\} \quad \begin{array}{l} \text{Standard Normal distribution } N(0,1) \\ (z \text{ distribution; } z(0,1)) \end{array}$$

* In continuous distributions, we can only compute probability in a given range. Never a single value.



$P(x_i) = B(x_i)$ ✓



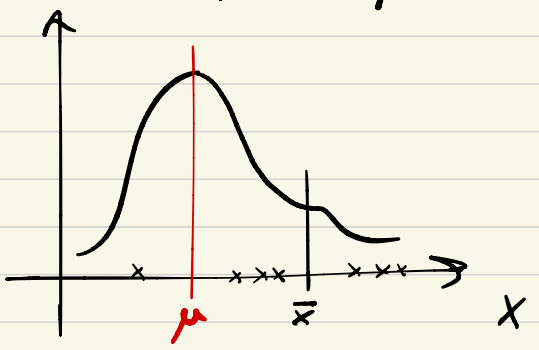
$P(a < x < b) = cdf(b) - cdf(a)$

(2.1) Parameter estimation

We have a set of observations $X = \{x_1, x_2, \ldots, x_n\}$

We assume they came from a certain distribution with some true μ and some true σ^2

$f(x)$: $B(x)$, $P(x)$, $Gauss(x)$ ...



I don't know the true μ, σ^2

I only have a discrete set of observations. I can build some "informative quantities" from data (*)

* observed average ("sample mean")

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

* observed variance ("sample variance")

$$s^2 = \frac{1}{n-1}\left\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2\right\} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

↳ Bessel's factor (*)

{ Sample mean $\bar{x}$ is an estimator of μ
Sample variance s^2 is an estimator of s^2 }

(2.2) the laws of large numbers.

Bernoulli, Chebyshev, Borel (1713)

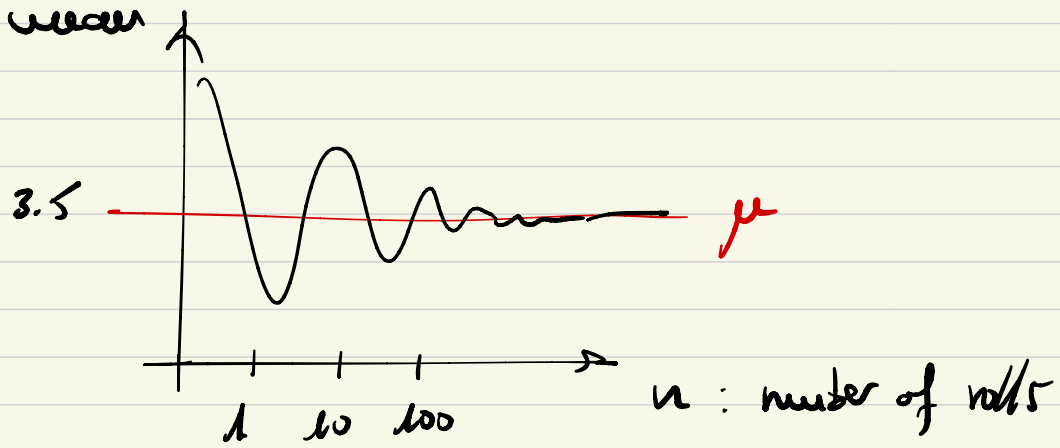
"Let $\{X_1, X_2, \ldots, X_n\}$ independent and identically distributed (iid) random variables, with same true mean μ (exp. value), the sample mean $\bar{X}$ converges to μ as n increases"

$$\left\{ \lim_{n \to \infty} \bar{X} = \mu \; ; \text{being } \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \right\}$$

Example: rolling dice.

i) true / expected mean $\mu = \frac{1}{6}(1+2+3+4+5+6) = 3.5$

ii) As we increase n rolls, sample mean $\bar{X}$ will converge to the true / expected value μ.

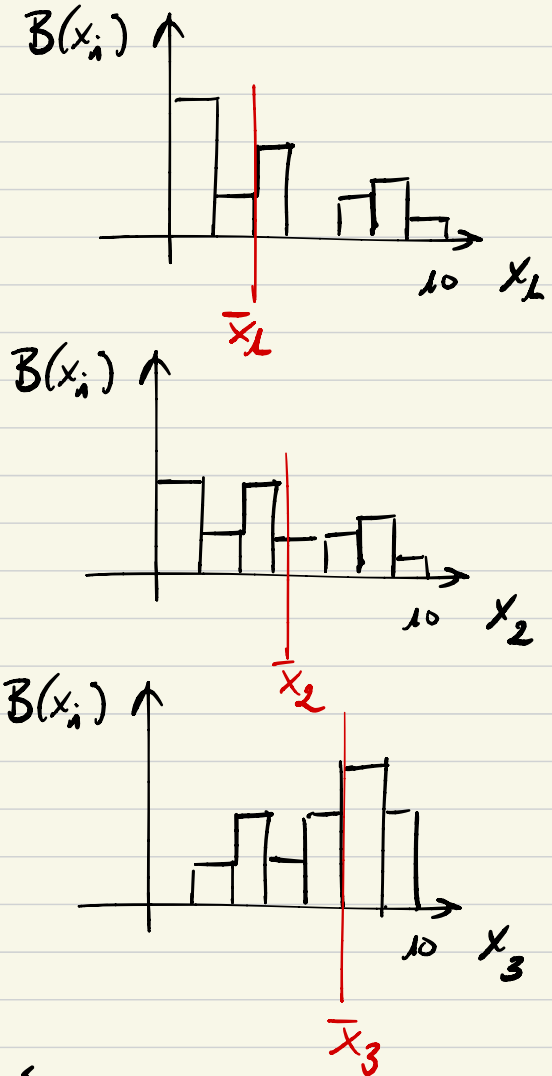


(2.3) the Central limit theorem

A. de Moivre, Laplace, Gauss (1733)

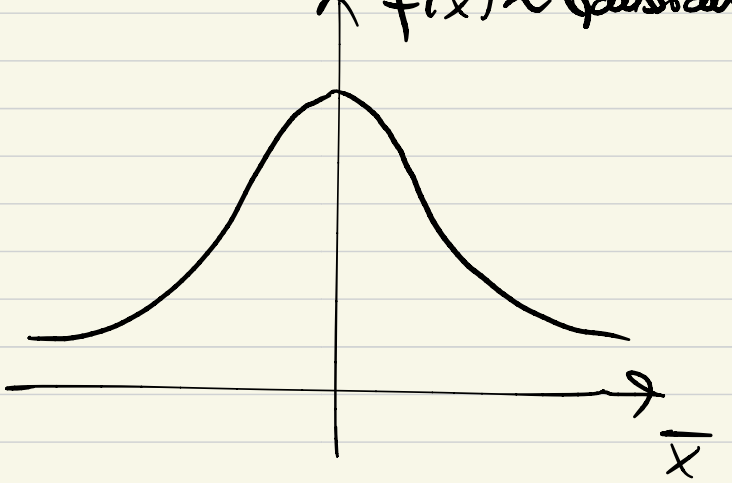
"For a set $\{x_1, \dots, x_n\}$ of n independent and identically distributed (iid) random variables, the distribution of sample means tends to gaussian as n increases, regardless of how are x_i distributed themselves"

i) Consider 10 rolls of a dice. How many times I observe a 6.
Binomial distributed $B(x; 10, 1/6)$. Repeat N times.



* As the sample size n increases, the distribution of means $\bar{x}_i$ tends to a gaussian distribution

$f(\bar{x}) \sim \text{Gaussian}(\bar{x})$



* Sample mean

$$\bar{x} \to N(\mu, \sigma^2/n)$$

* Standard error (SE)

$$SE = \frac{\sigma_i}{\sqrt{n}}$$; as I increase n, $\downarrow$ SE

ii) Build new random variable

$\bar{x} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N\}$

2.2 Confidence Intervals.

* See how many standard deviations an observation is away from the mean. Z-score (*)

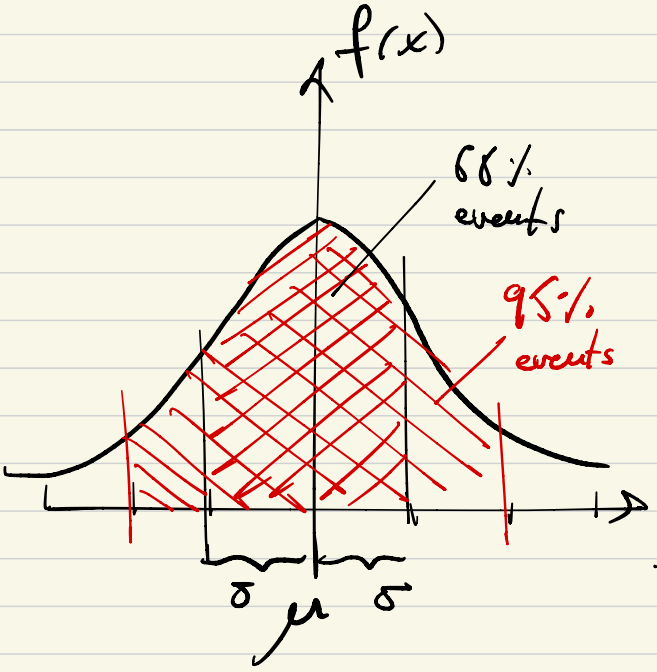
$$\left\{ z = \frac{x-\mu}{\sigma} \;;\; \text{if dealing with samples } z = \frac{x-\bar{x}}{s} \right\}$$

* When comparing sample means themselves, use SE instead of σ of individual observations.

$$\left\{ z = \frac{x-\mu}{\sigma/\sqrt{n}} \;;\; \text{dealing with samples } z = \frac{x-\bar{x}}{s/\sqrt{n}} \right\}$$

* Confidence interval $P_{critical} = (x+z, x-z)$

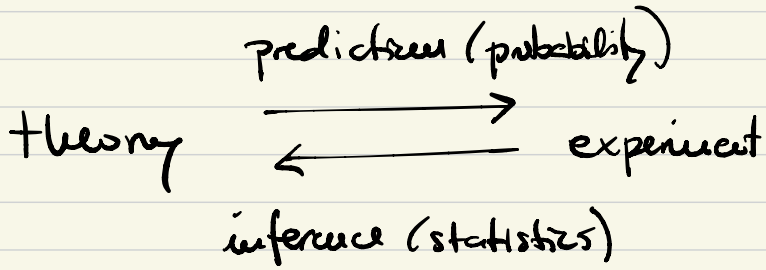
i) Consider a Gaussian distribution $f(x; \mu, \sigma)$



* 68% of events are contained in the $(\mu-\sigma, \mu+\sigma)$ interval. 1σ conf. int.
* 95% of events are contained in the $(\mu-2\sigma, \mu+2\sigma)$ interval. 2σ conf. int.
* 99% of events are contained in the $(\mu-3\sigma, \mu+3\sigma)$ interval. 3σ conf. int.

③ Hypothesis testing

* Prediction vs inference



i) Check if an observation is compatible with a given hypothesis H_0

* statistic test : a function / number computed out of our data

(t-test, χ^2 test, Fisher test, Wald test, ...)

* p-value : probability of obtaining a result at least as extreme as the one actually observed, under the assumption that our null hypothesis H_0 was correct.

(how likely / unlikely was to observe this result)

i) ↑ p-value : it was likely to obtain this result. Accept H_0.

ii) ↓ p-value : // unlikely // . Reject H_0.

ii) General approach

1. Formulate null hypothesis H_0 and significance level α.
2. Make measurement / observation.
3. Compute statistic test
4. Compute p-value
5. If p-value $< \alpha$, reject H_0. Otherwise, accept H_0.

* A bit of history

i) Thomas Bayes (1701-1761)

Bayes rule; probability as "updated degree of belief"

ii) Karl Pearson (1857-1936)

Correlation, regression, P-values

iii) Ronald Fisher (1890-1962)

Null hypothesis, analysis of variance

iv) Jerzy Neyman (1890-1891)

Egon Pearson (1895-1980)

False discovery rate / type I/II errors.

(3.1) One-sided t-test.

* Compare mean of observed die rolls to the expected mean of a fair die.

* For a series of observations $x = \{x_1, x_2, \ldots, x_n\}$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$\bar{x}$: observed average ("sample mean")

μ : expected average under H_0 ("population mean")

s : observed standard deviation ("sample standard dev.")

n : length of collected data ("sample size")

Example: Check if a die is biased. (1)

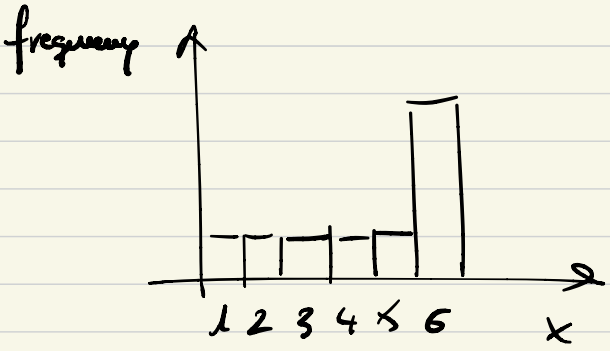
* You suspect your opponent is using a loaded die

i) Formulate null hypothesis $H_0 : p = \frac{1}{6}$

" alternative " $H_1 : p > \frac{1}{6}$

Choose significance level $\alpha = 0.01$

ii) Collect data. Roll die 15 times.



10 times out of 100 we get a 6. Suspicious!

iii) Perform 1-sample t-test.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{5 - 3.5}{1.6/\sqrt{15}} \simeq 3.6$$

μ : expected average in a fair die (all faces with same probability)

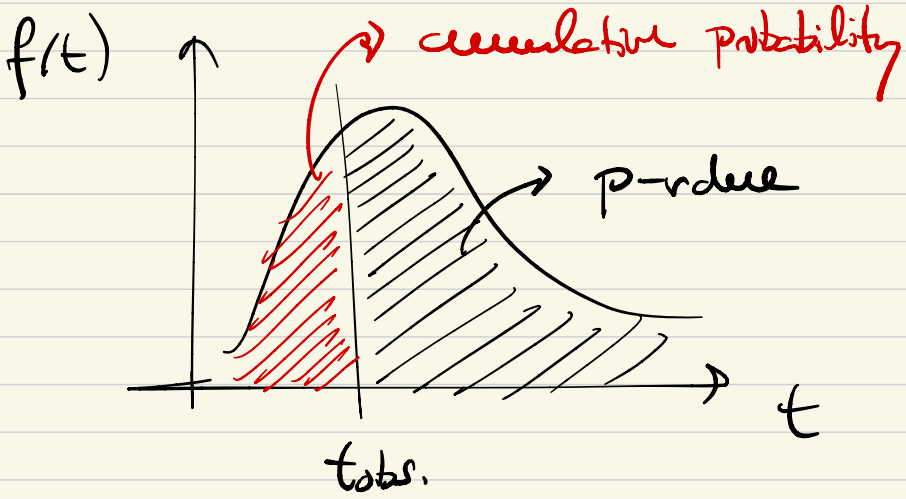
$$\mu = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

$\bar{x} = 5.$; $s = 1.6$; $n = 15$

(*) (*)

iv) P. value. Probability of obtaining a value at least as extreme as the one we obtained for our t-variable, given H_0 was true.

If H_0 is true, then t variable follows a "Student's t distribution"



$$\text{p-value} = P(t \geq t_{obs}) = 1 - \text{cdf}(t_{obs}) \quad (*)$$

{ Student's t tables
Computer simulations }

Example: Check if a die is biased. (II)

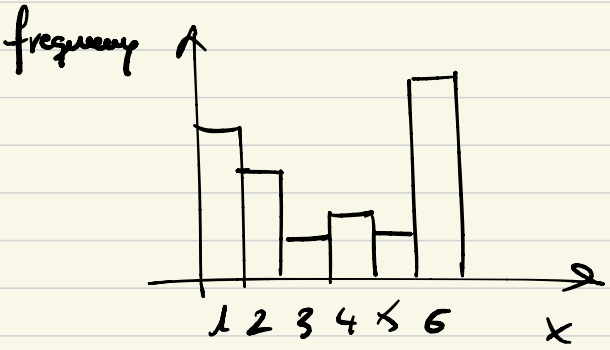
* You suspect your opponent is using a loaded die

i) Formulate null hypothesis $H_0 : p = \frac{1}{6}$

// alternative // $H_1 : p > \frac{1}{6}$

Choose significance level $\alpha = 0.01$

ii) Collect data. Roll die 100 times



43 times out of 100 we get a 6.

iii) Compute probability of 43 or less 6s, given H_0.

Given by Binomial distribution with $p = 1/6$

$B(43; 100, \frac{1}{6}) \sim 1$
cdf

Exercise

iv) Compute p-value. Probability of obtaining at least a value as extreme as 43.

$$\text{p-value} = 1 - B_{cdf}(43; 100, \tfrac{1}{6}) = 10^{-10}$$

↳ p-value $< \alpha$. ∃ evidence to reject H_0.

Example: check if coin is fair

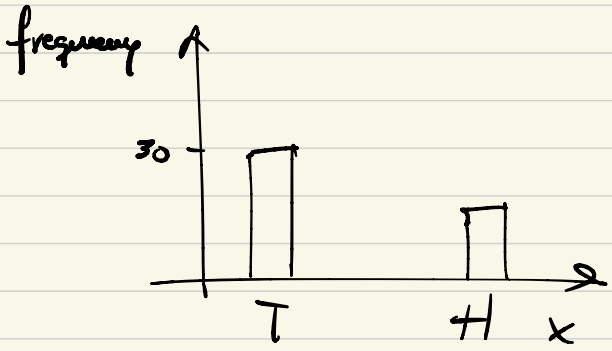
* You suspect your opponent is using a biased coin

i) Formulate null hypothesis $H_0: p = \frac{1}{2}$

// alternative // $H_1: p \neq \frac{1}{2}$

Choose significance level $\alpha = 0.05$

ii) Collect data. Roll die 100 times



20 heads in 50 flips of a coin

iii) Compute probability of 20 or less H, given H_0.

Given by Binomial distribution with $p = 1/2$

$$P(x = 20) = B_{cdf}\left(20; 50, \frac{1}{2}\right) = 0.1$$ (*)

iv) Compute p-value.

Now H_1 is just $p \neq \frac{1}{2}$, either larger or smaller.

p-value $= 2 \cdot cdf(20) = 0.2 < \alpha$ # evidence to reject H_0.

(3.3) Comparing means. t-test

* Compare if two samples have significantly different means

Given two sets of observations x_1 and x_2, of sample size n_1 and n_2, and sample means $\bar{x}_1$ and $\bar{x}_2$.

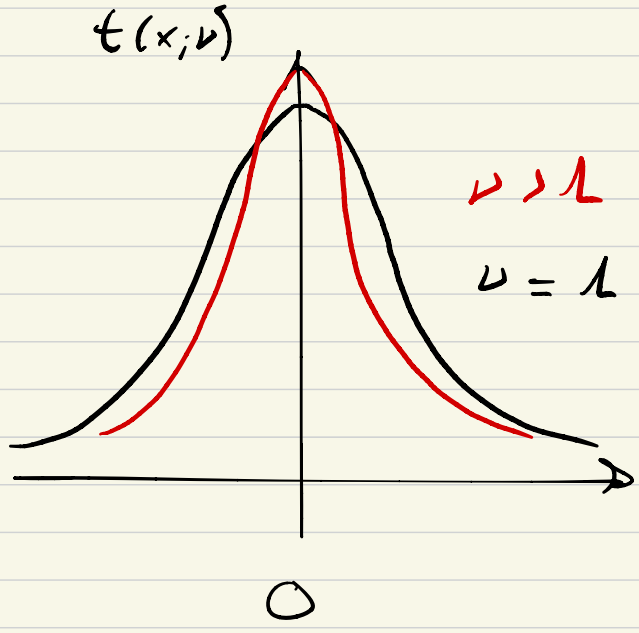
* Compute t variable (t "statistic")

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

being s_p the pooled / summed standard deviation

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

* Under null hypothesis of equal means, the t statistic follows a Student's t distribution with (n_1+n_2-2) dof.



i) For $\nu = 1$; $t(x;\nu)$ is referred to as the Cauchy distribution

$$f(x; x_0, \gamma) = \frac{1}{\pi} \frac{\gamma}{(x-x_0)^2 + \gamma^2}$$

ii) For ν large; $t(x;\nu)$ tends to a normal distribution

* A bit of history

1876. t-distribution derived first by Helmert and Lüroth

1895. General form from Pearson, as Pearson type IV distribution

1908. William Sealy Gosset publishes under name "Student"

1925. Popularized by Fisher as "Student's distribution"

Example: Comparing means of 2 populations

* Check if the means of birth weights are significantly different between smokers and non smokers.

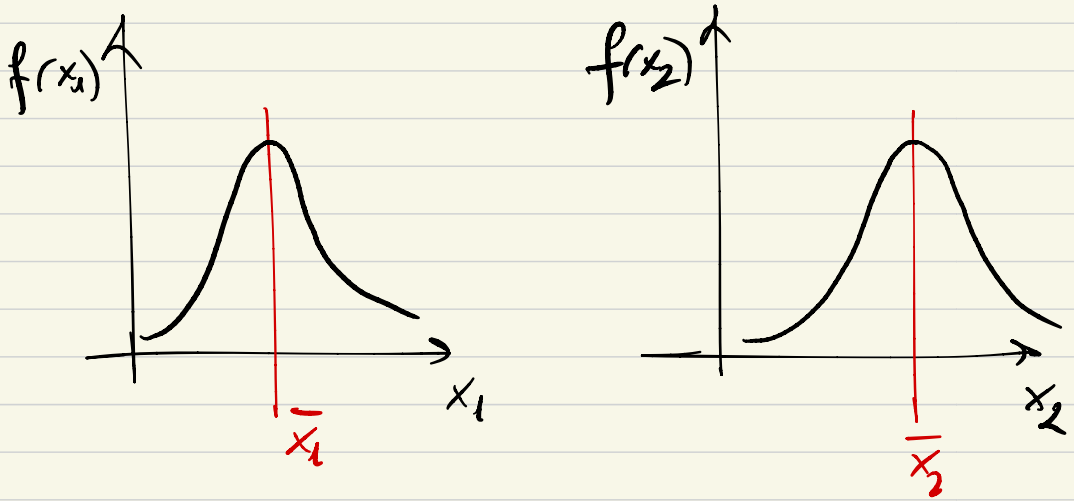
i) Formulate null hypothesis

H_0. $\nexists$ difference btw the means. $d = \bar{x}_1 - \bar{x}_2 = 0$

H_1. $\exists$ // . $d \neq 0$

Choose significance level α

ii) Collect data. Smokers and non smokers.



n_1 non smokers — n_2 heavy smokers

Sample mean $\bar{x}_1$ — Sample mean $\bar{x}_2$

standard deviation s_1 — standard deviation s_2

iii) Compute t statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

iv) Compute d.o.f

$$\nu = n_1 + n_2 - 2$$

v) Check if t falls inside the 95% confidence interval.

Example: Comparing means of two populations.

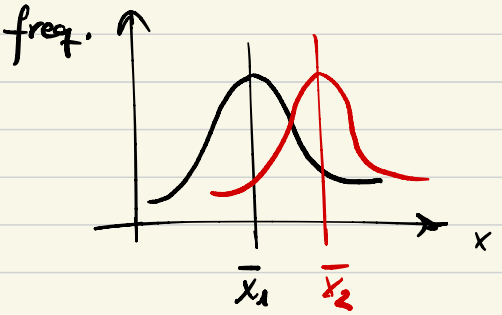
i) Formulate H_0, H_1

H_0: both samples have the same mean $\bar{x}_1 = \bar{x}_2 = \mu$

H_1: the samples have different means $\bar{x}_1 \neq \bar{x}_2$

ii) Collect data

Example: luminosity of 2 different types of stars.



$\{x_1\}$: type IV stars ("supergiant")

$\{x_2\}$: type V stars ("hypergiant")

iii) Compute t "statistic" (t variable)

$$\left\{ t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right\} \quad \text{if } \bar{x}_1 = \bar{x}_2 \Rightarrow t = 0 \checkmark$$

$$s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$ "pooled" standard deviation.

iv) Compute p-value.

"What was the probability of obtaining a value at least as extreme as the one we got for t."

* If $\bar{x}_1 = \bar{x}_2$, the t variable follows a Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. (*)