

Introduction to Statistical Tests (part 1)

Emmanuel Jeannot

INRIA
LaBRI

January, 2024

- 1 Introduction
- 2 Probability 101
- 3 Law of large numbers and CLT
- 4 Distribution of the sum of a random variable

Objectives

Welcome!

- I have run 20 experiments. In 15 cases algorithm A is better than B and in 5 cases it is the opposite.

Objectives

Welcome!

- I have run 20 experiments. In 15 cases algorithm A is better than B and in 5 cases it is the opposite.
- What can I conclude? With which certainty?

Objectives

Welcome!

- I have run 20 experiments. In 15 cases algorithm A is better than B and in 5 cases it is the opposite.
- What can I conclude? With which certainty?
- 6 benchmarks were used to compare two systems. The observations are:
 $\{(5.4, 19.1), (16.6, 3.5), (0.6, 3.4), (7.3, 1.7), (1.4, 2.5), (0.6, 3.6)\}$.

Objectives

Welcome!

- I have run 20 experiments. In 15 cases algorithm A is better than B and in 5 cases it is the opposite.
- What can I conclude? With which certainty?
- 6 benchmarks were used to compare two systems. The observations are:
 $\{(5.4, 19.1), (16.6, 3.5), (0.6, 3.4), (7.3, 1.7), (1.4, 2.5), (0.6, 3.6)\}$.
- Is one system better than the other?

Objectives

These are relevant questions!

- Statistics can help us to make insightful conclusions
- Statistics can help us to bound the confidence of these conclusions

Goal

- Know that a theory exists to answer these questions
- Make statistics tests and interpret them
- Know where to find more insightful resources

This lesson will be

- 1/3 of theory
- 1/3 of cookbook
- 1/3 of practice

Disclaimer

Be advise that:

- I am not a statistician...
- I am not even a mathematician...
- I did not learn this during my studies (but I would have love to)...
- I am a researcher in algorithms and I have found the technics I will develop here very useful to do a better science!

Other Resources

I used the following resources

- The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling, April 1991 of Raj Jain
- MOOC *Statistical Inference* on Coursera.
- <https://leanpub.com/LittleInferenceBook> that comes with the above course

Web page:

<https://www.labri.fr/perso/ejeannot/teaching>

Outline

- 1 Introduction
- 2 Probability 101
- 3 Law of large numbers and CLT
- 4 Distribution of the sum of a random variable

Fundamentals of Probability

- Ω : sample space (all possible outcome of an experiments).
- \mathcal{A} : set of events (containing zero or more outcomes)
- \mathbb{P} : assignment of probabilities of events in \mathcal{A} .

Axiom/Properties

- $\forall A \in \mathcal{A} : 0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}(\Omega) = 1$
- $A_1 \in \mathcal{A}, A_2 \in \mathcal{A}, A_1 \cap A_2 = \emptyset : \mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$
- $A \in \mathcal{A}, \bar{A} = \mathcal{A} \setminus A : \mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1 \Leftrightarrow \mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 0$
- A_1 et A_2 are independent then $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$.

Example

- I throw n dice
- What is the probability that at least one of them shows a “1”?

Solution

- $A =$ “the dice is 1”

Example

- I throw n dice
- What is the probability than at least one of them show a “1”?

Solution

- $A =$ “the dice is 1”
- $\mathbb{P}(A) = p$ ($p = 1/6$ for a 6-faces dice)

Example

- I throw n dice
- What is the probability that at least one of them shows a “1”?

Solution

- $A =$ “the dice is 1”
- $\mathbb{P}(A) = p$ ($p = 1/6$ for a 6-faces dice)
- $1 - p$: probability that the dice is not “1”.

Example

- I throw n dice
- What is the probability that at least one of them shows a “1”?

Solution

- $A =$ “the dice is 1”
- $\mathbb{P}(A) = p$ ($p = 1/6$ for a 6-faces dice)
- $1 - p$: probability that the dice is not “1”.
- $(1 - p)^n$: probability that no dice among n show a “1” (throws are assumed independent)

Example

- I throw n dice
- What is the probability than at least one of them show a “1”?

Solution

- $A =$ “the dice is 1”
- $\mathbb{P}(A) = p$ ($p = 1/6$ for a 6-faces dice)
- $1 - p$: probability that the dice is not “1”.
- $(1 - p)^n$: probability that no dice among n show a “1” (throws are assumed independent)
- $1 - (1 - p)^n$: probability than at least one of them show a “1”.

Random Variables

- A way to think about numeric outcome of experiments (e.g. dice roll).
- Random variable: **numerical outcome** of an experiment.
- Can be discrete or continuous.
- X a random variable: $X : \Omega \rightarrow E$. Ex. if $E = \mathbb{N}$, X is discrete; if $E = \mathbb{R}$, X is continuous (real).

Random Variables

- A way to think about numeric outcome of experiments (e.g. dice roll).
- Random variable: **numerical outcome** of an experiment.
- Can be discrete or continuous.
- X a random variable: $X : \Omega \rightarrow E$. Ex. if $E = \mathbb{N}$, X is discrete; if $E = \mathbb{R}$, X is continuous (real).

Examples

- Dice roll: $E = \{1, 2, 3, 4, 5, 6\}$
- Coin: $E = \{0, 1\}$
- Time of bus arrival at a station: $E = [0, 24[$.
- Number of requests per second arriving to a web server: $E = \mathbb{R}^+$.

Probability Mass Function (pmf)

- Gives the probability that discrete random variable takes a given value.
- $X : \Omega \rightarrow E$, a discrete random variable. Then, the probability mass function is $p : E \rightarrow [0, 1]$, with:
 - $p(x) = \mathbb{P}(X = x)$.
 - $\sum_{x \in E} p(x) = 1$.

Example

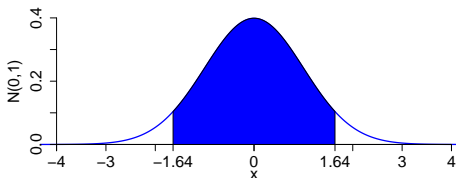
- Balanced coin flip. $X(\text{head}) = 0$. $X(\text{tail}) = 1$.
 $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$
- $p(x) = (1/2)^x(1/2)^{1-x}$, for $x = 0, 1$.
- Unbalanced coin. θ probability of tail, $p(x) = \theta^x(1 - \theta)^{1-x}$, for $x = 0, 1$.

Probability Density Function (pdf)

- Gives the probability that a continuous variable is between two values
- The **area under a pdf is a probability**.
- $X : \Omega \rightarrow E$, a continuous random variable. Then, the probability density function is a **non negative integrable function** f_X , such that:
 - $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx, a, b \in E.$
 - $\int_E f(x) dx = 1.$
- In general, $\mathbb{P}(X = a) = 0.$

Probability Density Function (pdf)

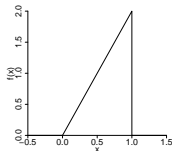
- Gives the probability that a continuous variable is between two values
- The **area under a pdf is a probability**.
- $X : \Omega \rightarrow E$, a continuous random variable. Then, the probability density function is a **non negative integrable function** f_X , such that:
 - $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx, a, b \in E.$
 - $\int_E f(x) dx = 1.$
- In general, $\mathbb{P}(X = a) = 0.$



$$X \sim \mathcal{N}(0, 1), \mathbb{P}(-1.645 \leq X \leq 1.645) \simeq 0.9$$

Example

Suppose that the proportion of answered calls in an help center on a random day is $f(x) = 2x$, $x \in [0, 1]$.

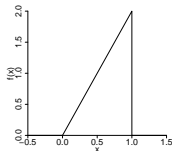


Questions

- Is this a valid PDF?

Example

Suppose that the proportion of answered calls in an help center on a random day is $f(x) = 2x$, $x \in [0, 1]$.

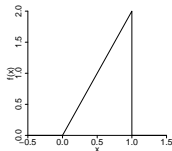


Questions

- Is this a valid PDF?
- Hints: check non negativity and area.

Example

Suppose that the proportion of answered calls in an help center on a random day is $f(x) = 2x$, $x \in [0, 1]$.

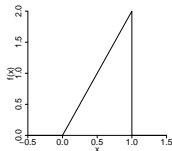


Questions

- Is this a valid PDF?
- Hints: check non negativity and area.
- What is the probability that 75% or fewer of the calls get addressed?

Example

Suppose that the proportion of answered calls in an help center on a random day is $f(x) = 2x$, $x \in [0, 1]$.

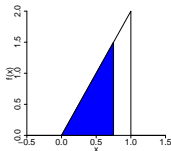


Questions

- Is this a valid PDF?
- Hints: check non negativity and area.
- What is the probability that 75% or fewer of the calls get addressed?
- Hint: find the right area.

Example

Suppose that the proportion of answered calls in an help center on a random day is $f(x) = 2x$, $x \in [0, 1]$.



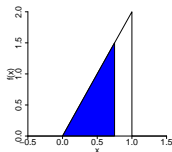
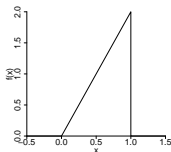
Questions

- Is this a valid PDF?
- Hints: check non negativity and area.
- What is the probability that 75% or fewer of the calls get addressed?
- Hint: find the right area.
- $p = \frac{0.75 \times 1.5}{2} = 0.5625$

Cumulative Distribution Function (cdf)

- Gives the probability that a random variable is less or equal to a given value
- $X : \Omega \rightarrow E$, a continuous random variable. Then, the cumulative distribution function is a non negative integrable function F_X , such that $F_X(x) = \mathbb{P}(X \leq x)$.
- F_X is increasing.
- $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$, $a, b \in E$.
- F_X is the primitive of f_X
- f_X is the derivative of F_X

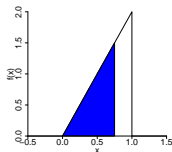
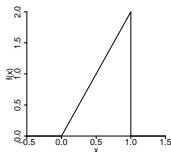
Example



Questions

- $f_X(x) = 2x, x \in [0, 1]$. What is the CDF?

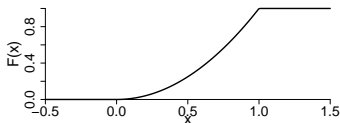
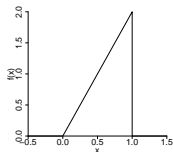
Example



Questions

- $f_X(x) = 2x, x \in [0, 1]$. What is the CDF?
- $F_X(x) = \mathbb{P}(X \leq x) = \frac{x \times 2x}{2} = x^2, x \in [0, 1]$.

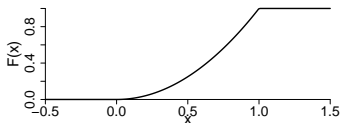
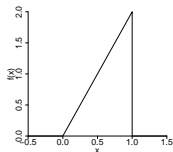
Example



Questions

- $f_X(x) = 2x, x \in [0, 1]$. What is the CDF?
- $F_X(x) = \mathbb{P}(X \leq x) = \frac{x \times 2x}{2} = x^2, x \in [0, 1]$.
- $F_X(x) = 0, x \leq 0$.

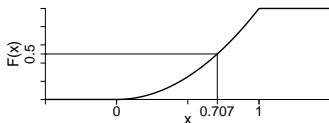
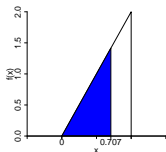
Example



Questions

- $f_X(x) = 2x, x \in [0, 1]$. What is the CDF?
- $F_X(x) = \mathbb{P}(X \leq x) = \frac{x \times 2x}{2} = x^2, x \in [0, 1]$.
- $F_X(x) = 0, x \leq 0$.
- $F_X(x) = 1, x \geq 1$.

Example



Questions

- $f_X(x) = 2x$, $x \in [0, 1]$. What is the CDF?
- $F_X(x) = \mathbb{P}(X \leq x) = \frac{x \times 2x}{2} = x^2$, $x \in [0, 1]$.
- $F_X(x) = 0$, $x \leq 0$.
- $F_X(x) = 1$, $x \geq 1$.
- x such that $F_X(x) = 0.5 \implies x = \sqrt{0.5} \approx 0.707$ is the median.

Quantile

- At which value x we have a fraction of a population above a given threshold α ($\alpha \in [0, 1]$).
- The α^{th} quantile of a distribution with distribution function F is the point x_α such that $F(x_\alpha) = \alpha$
- $\mathbb{P}(X < x_\alpha) = \alpha$.

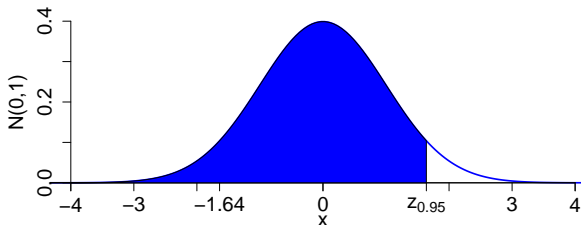
Examples

- Median : $\alpha = 0.5$
- First quartile $\alpha = 0.25$
- Third quartile $\alpha = 0.75$
- Percentile is a quantile when α is expressed in percent. The median is the 50th percentile.

CDF vs Quantile

Each function is the inverse of the other

- Let X a random variable. You need to solve $\mathbb{P}(X \leq C) = \alpha$.
- If you know C then α is given by the CDF : $\alpha = F_X(C)$.
- If you know α then C is given by the quantile function: $C = q_\alpha$ where q_α the α^{th} quantile.
- Remark: for the standard normal $-\mathcal{N}(0, 1)$ – the α^{th} quantile is noted z_α .



The 0.95 quantile ($z_{0.95}$) of $\mathcal{N}(0, 1)$ is 1.645.

Mean vs mediane

- Mean (*esperance*) : center of density of the distribution
- Discrete variable with PMF $p(x)$: $\mathbb{E}(X) = \sum_x xp(x)$
- Continuous variable with PDF $f(x)$: $\mathbb{E}(X) = \int_E xf(x)dx$

Mean vs mediane

- Mean (*esperance*) : center of density of the distribution
- Discrete variable with PMF $p(x)$: $\mathbb{E}(X) = \sum_x xp(x)$
- Continuous variable with PDF $f(x)$: $\mathbb{E}(X) = \int_E xf(x)dx$

Examples

- 6-faces dice:

$$\mathbb{E} = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = \frac{1}{6} \times 21 = 3.5$$

- Help center example: $\int_0^1 x \times 2x dx = \int_0^1 2x^2 dx = \left[\frac{2x^3}{3} \right]_0^1 = \frac{2}{3}$

Mean vs mediane

- Mean (*esperance*) : center of density of the distribution
- Discrete variable with PMF $p(x)$: $\mathbb{E}(X) = \sum_x xp(x)$
- Continuous variable with PDF $f(x)$: $\mathbb{E}(X) = \int_E xf(x)dx$

Examples

- 6-faces dice:

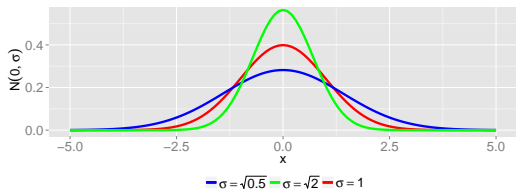
$$\mathbb{E} = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = \frac{1}{6} \times 21 = 3.5$$

- Help center example: $\int_0^1 x \times 2x dx = \int_0^1 2x^2 dx = \left[\frac{2x^3}{3} \right]_0^1 = \frac{2}{3}$

Help center

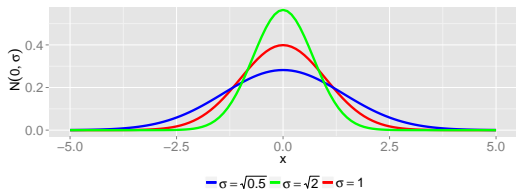
- Mean = $\frac{2}{3}$: on average $\frac{2}{3}$ of the calls are answered.
- Median = $\sqrt{0.5}$: half of the days answer $\sqrt{0.5}$ calls or more.

Variance and Standard deviation



- The variance is a measure of the spread of a random variable.
- μ mean of X , then : $\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- σ the standard deviation is the square root of the variance (same unit as the data): $\sigma^2 = \text{Var}(X)$.

Variance and Standard deviation



- The variance is a measure of the spread of a random variable.
- μ mean of X , then : $\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- σ the standard deviation is the square root of the variance (same unit as the data): $\sigma^2 = \text{Var}(X)$.

Example

- 6-faces dice:

$$\sigma^2 = \frac{1}{6} \times 1^2 + \frac{1}{6} \times 2^2 + \frac{1}{6} \times 3^2 + \frac{1}{6} \times 4^2 + \frac{1}{6} \times 5^2 + \frac{1}{6} \times 6^2 - 3.5^2 \approx 15.17 - 12.25 = 2.92$$

Random distribution with R

Examples

- `xunif`: continuous uniform distribution; `xbinom`: binomial distribution; `xnorm`: normal distribution; `xbeta`: beta distribution, etc.
- $x \in \{d, p, q, r\}$ for the density, the distribution function, the quantile function and for generating random deviates.
- mnemonic: `d` is for distribution. `p` for probability, `q` for quantile, `r` for random.

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?
- `dunif(1, min = 0, max = 2)`

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?
- `dunif(1, min = 0, max = 2)`
- What is the probability that X is lower than 1.5?

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?
- `dunif(1, min = 0, max = 2)`
- What is the probability that X is lower than 1.5?
- `punif(1.5, min = 0, max = 2)`

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?
- `dunif(1, min = 0, max = 2)`
- What is the probability that X is lower than 1.5?
- `punif(1.5, min = 0, max = 2)`
- What is the median of X ?

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?
`dunif(1, min = 0, max = 2)`
- What is the probability that X is lower than 1.5?
`punif(1.5, min = 0, max = 2)`
- What is the median of X ?
`qunif(0.5, min = 0, max = 2)`

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?
- `dunif(1, min = 0, max = 2)`
- What is the probability that X is lower than 1.5?
- `punif(1.5, min = 0, max = 2)`
- What is the median of X ?
- `qunif(0.5, min = 0, max = 2)`
- Generate 10 numbers following X .

Random distribution with R

Uniform distribution

Consider X a random variable following a uniform distribution between 0 and 2

- What is the value of the density function in 1?
- `dunif(1, min = 0, max = 2)`
- What is the probability that X is lower than 1.5?
- `punif(1.5, min = 0, max = 2)`
- What is the median of X ?
- `qunif(0.5, min = 0, max = 2)`
- Generate 10 numbers following X .
- `runif(10, min = 0, max = 2)`

Exercises

Binomial distribution

- `pbinom(q, size, prob)`: probability that q or less events among $size$ ones have happen when each event is of probability $prob$.

Exercises

Binomial distribution

- `pbinom(q, size, prob)`: probability that q or less events among $size$ ones have happen when each event is of probability $prob$.
- What is the probability that 100 coin flips result in only heads?

Exercises

Binomial distribution

- `pbinom(q, size, prob)`: probability that `q` or less events among `size` ones have happen when each event is of probability `prob`.
- What is the probability that 100 coin flips result in only heads?
- Answer: `pbinom(0, 100, 0.5) = 7.888609e-31`

Exercises

Binomial distribution

- `pbinom(q, size, prob)`: probability that q or less events among `size` ones have happen when each event is of probability `prob`.
- What is the probability that 100 coin flips result in only heads?
- Answer: `pbinom(0, 100, 0.5) = 7.888609e-31`
- And 100 coin flips result in exactly 50 heads?

Exercises

Binomial distribution

- `pbinom(q, size, prob)`: probability that q or less events among `size` ones have happen when each event is of probability `prob`.
- What is the probability that 100 coin flips result in only heads?
- Answer: `pbinom(0, 100, 0.5) = 7.888609e-31`
- And 100 coin flips result in exactly 50 heads?
- Answer: `pbinom(50, 100, 0.5) - pbinom(49, 100, 0.5)`
`0.07958924 = dbinom(50, 100, 0.5)`

Exercises

Binomial distribution

- `pbinom(q, size, prob)`: probability that q or less events among `size` ones have happen when each event is of probability `prob`.
- What is the probability that 100 coin flips result in only heads?
- Answer: `pbinom(0, 100, 0.5) = 7.888609e-31`
- And 100 coin flips result in exactly 50 heads?
- Answer: `pbinom(50, 100, 0.5) - pbinom(49, 100, 0.5)`
`0.07958924 = dbinom(50, 100, 0.5)`
- Your friend claims that changing the font to comic will result in more ad revenue on your web sites. When presented in random order, 9 pages out of 10 had more revenue when the font was set to comic. If it was really a coin flip for these 10 sites, what's the probability of getting 9 or 10 out of 10 with more revenue for the new font?

Exercises

Binomial distribution

- `pbinom(q, size, prob)`: probability that q or less events among `size` ones have happen when each event is of probability `prob`.
- What is the probability that 100 coin flips result in only heads?
- Answer: `pbinom(0, 100, 0.5) = 7.888609e-31`
- And 100 coin flips result in exactly 50 heads?
- Answer: `pbinom(50, 100, 0.5) - pbinom(49, 100, 0.5)`
`0.07958924 = dbinom(50, 100, 0.5)`
- Your friend claims that changing the font to comic will result in more ad revenue on your web sites. When presented in random order, 9 pages out of 10 had more revenue when the font was set to comic. If it was really a coin flip for these 10 sites, what's the probability of getting 9 or 10 out of 10 with more revenue for the new font?
- Answer: `1 - pbinom(8, 10, 0.5) = 0.01074219`

Exercises

Normal distribution

- `pnorm(q, mean, sd)`: probability that a random variable following a normal distribution with mean `mean` and standard deviation `sd` is less or equal than `q`.

Exercises

Normal distribution

- $\text{pnorm}(q, \text{mean}, \text{sd})$: probability that a random variable following a normal distribution with mean mean and standard deviation sd is less or equal than q .
- A software company is doing an analysis of documentation errors of their products. They sampled their very large codebase in chunks and found that the number of errors per chunk was approximately normally distributed with a mean of 11 errors and a standard deviation of 2. When randomly selecting a chunk from their codebase, what's the probability of fewer than 5 documentation errors?

Exercises

Normal distribution

- $\text{pnorm}(q, \text{mean}, \text{sd})$: probability that a random variable following a normal distribution with mean mean and standard deviation sd is less or equal than q .
- A software company is doing an analysis of documentation errors of their products. They sampled their very large codebase in chunks and found that the number of errors per chunk was approximately normally distributed with a mean of 11 errors and a standard deviation of 2. When randomly selecting a chunk from their codebase, what's the probability of fewer than 5 documentation errors?
- Answer: $\text{pnorm}(5, 11, 2) = 0.001349898$

Exercises

Normal distribution

- $\text{pnorm}(q, \text{mean}, \text{sd})$: probability that a random variable following a normal distribution with mean mean and standard deviation sd is less or equal than q .
- A software company is doing an analysis of documentation errors of their products. They sampled their very large codebase in chunks and found that the number of errors per chunk was approximately normally distributed with a mean of 11 errors and a standard deviation of 2. When randomly selecting a chunk from their codebase, what's the probability of fewer than 5 documentation errors?
- Answer: $\text{pnorm}(5, 11, 2) = 0.001349898$
- Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What's the probability that a given day has fewer than 93 hits per day?

Exercises

Normal distribution

- $\text{pnorm}(q, \text{mean}, \text{sd})$: probability that a random variable following a normal distribution with mean mean and standard deviation sd is less or equal than q .
- A software company is doing an analysis of documentation errors of their products. They sampled their very large codebase in chunks and found that the number of errors per chunk was approximately normally distributed with a mean of 11 errors and a standard deviation of 2. When randomly selecting a chunk from their codebase, what's the probability of fewer than 5 documentation errors?
- Answer: $\text{pnorm}(5, 11, 2) = 0.001349898$
- Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What's the probability that a given day has fewer than 93 hits per day?
- Answer: $\text{pnorm}(93, 100, 10) = 0.2419637$

Normal distribution

- The population mean BMI for men is reported as $29 \text{ kg}/\text{m}^2$ with a standard deviation of 4.73. Assuming normality of BMI, what is the population 95th percentile? (hint use `qnorm`)

Exercises

Normal distribution

- The population mean BMI for men is reported as $29 \text{ kg}/\text{m}^2$ with a standard deviation of 4.73. Assuming normality of BMI, what is the population 95th percentile? (hint use `qnorm`)
- Answer: `qnorm(.95, 29, 4.73) = 36.78016`

Exercises

Normal distribution

- The population mean BMI for men is reported as $29 \text{ kg}/\text{m}^2$ with a standard deviation of 4.73. Assuming normality of BMI, what is the population 95th percentile? (hint use `qnorm`)
- Answer: `qnorm(.95, 29, 4.73) = 36.78016`
- Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What number of web hits per day represents the number so that only 5% of days have more hits?

Exercises

Normal distribution

- The population mean BMI for men is reported as $29 \text{ kg}/\text{m}^2$ with a standard deviation of 4.73. Assuming normality of BMI, what is the population 95th percentile? (hint use `qnorm`)
- Answer: `qnorm(.95, 29, 4.73) = 36.78016`
- Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What number of web hits per day represents the number so that only 5% of days have more hits?
- Answer: `qnorm(0.95, 100, 10) = 116.4485`

Exercises R

The Standard Normal Distribution

- Use `rnorm` to draw 1000 samples normally distributed with mean $m = 24$ and standard deviation $s = 7$. Put these sample in vector `v`.

Exercises R

The Standard Normal Distribution

- Use `rnorm` to draw 1000 samples normally distributed with mean $m = 24$ and standard deviation $s = 7$. Put these sample in vector `v`.
- Print the mean of `v` and the standard deviation of `v`

Exercises R

The Standard Normal Distribution

- Use `rnorm` to draw 1000 samples normally distributed with mean $m = 24$ and standard deviation $s = 7$. Put these sample in vector v .
- Print the mean of v and the standard deviation of v
- Compute the standard deviation and the mean of v/k , for $k = 7$.

Exercises R

The Standard Normal Distribution

- Use `rnorm` to draw 1000 samples normally distributed with mean $m = 24$ and standard deviation $s = 7$. Put these sample in vector v .
- Print the mean of v and the standard deviation of v
- Compute the standard deviation and the mean of v/k , for $k = 7$.
- Compute the standard deviation and the mean of $(v-m) / s$.

Exercises R

The Standard Normal Distribution

- Use `rnorm` to draw 1000 samples normally distributed with mean $m = 24$ and standard deviation $s = 7$. Put these sample in vector v .
- Print the mean of v and the standard deviation of v
- Compute the standard deviation and the mean of v/k , for $k = 7$.
- Compute the standard deviation and the mean of $(v-m) / s$.
- Do it with other values of m , s and k . What can you conclude?

Exercises R

The Standard Normal Distribution

- Use `rnorm` to draw 1000 samples normally distributed with mean $m = 24$ and standard deviation $s = 7$. Put these sample in vector `v`.
- Print the mean of `v` and the standard deviation of `v`
- Compute the standard deviation and the mean of `v/k`, for $k = 7$.
- Compute the standard deviation and the mean of `(v-m) / s`.
- Do it with other values of m , s and k . What can you conclude?

Important properties

- if $X \sim \mathcal{N}(\mu, \sigma)$ then $\frac{X}{k} \sim \mathcal{N}(\mu/k, \sigma/k)$, for $k \neq 0$.
- if $X \sim \mathcal{N}(\mu, \sigma)$ then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$

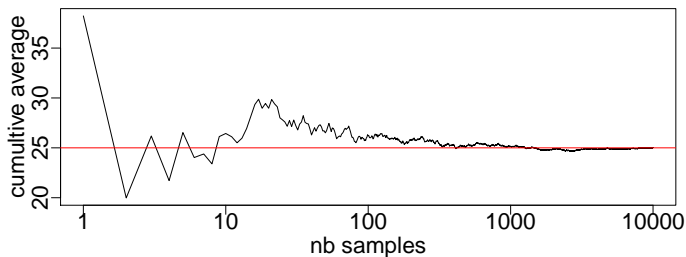
Law of large numbers

Try this code

```
library(dplyr)

max <- 50
N <- 10000
cum_avg <- cummean(runif(N, 0, max))
pdf("law_of_large_numbers.pdf", width=8, height=4)
plot(cum_avg, log="x", type="l",
      xlab="nb samples", ylab="cumulative average")
abline(h = max/2, col = "red")
dev.off()
```

Law of large numbers



Law of large numbers

The sample mean of an independent and identically distributed (iid) sample asymptotically converges to the expected value (the population mean)

Distribution of the sum of a random variable

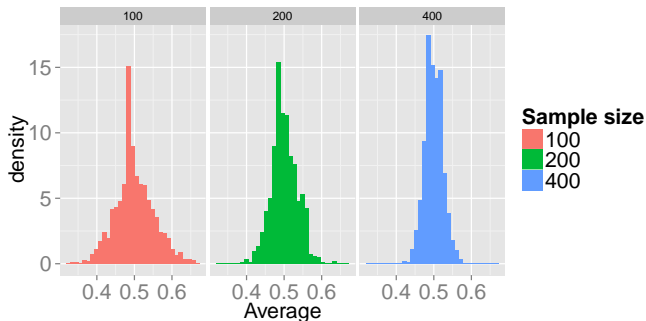
The binomial case

- We flip a coin N times and count 1 for head and 0 for tail.
- We do this M times and obtain M samples of N flips.
- For each sample we compute the average (of the N flips).
- Draw the histogram of the M s with M (number of samples) = 100 and N (sample size) = 100, 200, 400.

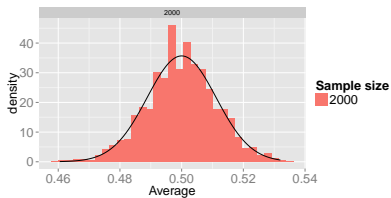
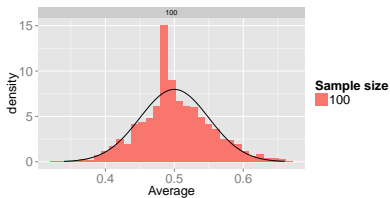
Distribution of the sum of a random variable

The binomial case

- We flip a coin N times and count 1 for head and 0 for tail.
- We do this M times and obtain M samples of N flips.
- For each sample we compute the average (of the N flips).
- Draw the histogram of the M s with M (number of samples) = 100 and N (sample size) = 100, 200, 400.

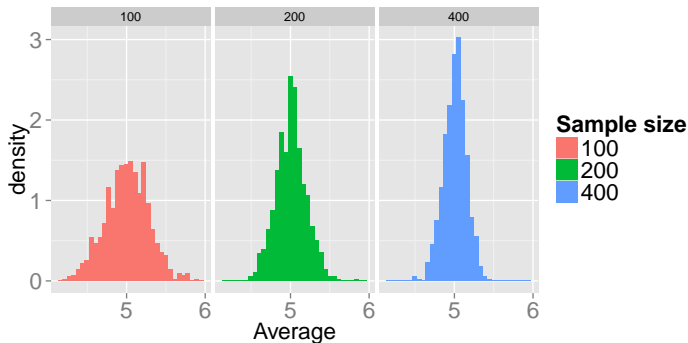


Distribution of the sum of a random variable



The variance decreases with the sample size

Distribution of the sum of a random variable



This also work for any other distribution.

Here uniform distribution in $[0, 10]$ with 100, 200 and 400 samples

Central Limit Theorem

Theorem

- Let X be a random variable of mean μ and variance σ^2 .
- Let $X_n = \sum_n X$.
- Then $X_n \sim \mathcal{N}(n\mu, \sigma\sqrt{n})$ when n is large.
- *"The sum of iid samples tends to be normally distributed."*

Central Limit Theorem

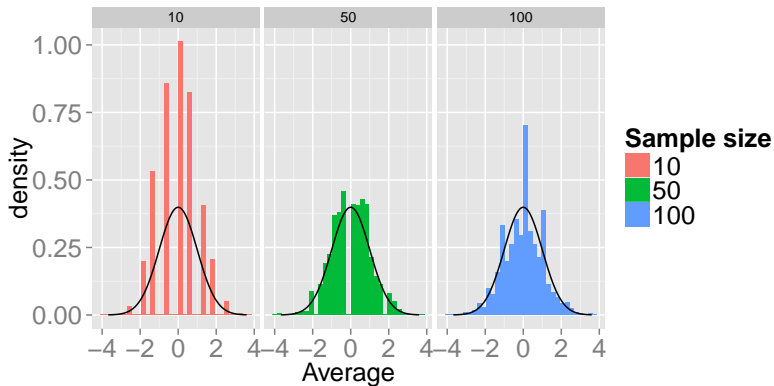
Theorem

- Let X be a random variable of mean μ and variance σ^2 .
- Let $X_n = \sum_n X$.
- Then $X_n \sim \mathcal{N}(n\mu, \sigma\sqrt{n})$ when n is large.
- "The sum of iid samples tends to be normally distributed."

Corollary

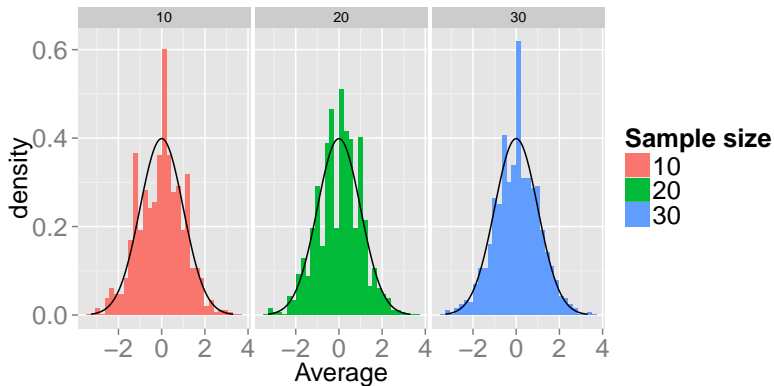
- Recall: if $X \sim \mathcal{N}(\mu, \sigma)$ then :
 - $\frac{X}{k} \sim \mathcal{N}(\mu/k, \sigma/k)$ and,
 - $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$
- $X_n \sim \mathcal{N}(n\mu, \sigma\sqrt{n}) \iff \bar{X}_n = \frac{X_n}{n} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
- $\iff \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$
- σ/\sqrt{n} is called the *standard error*.

Central Limit Theorem: speed of convergence



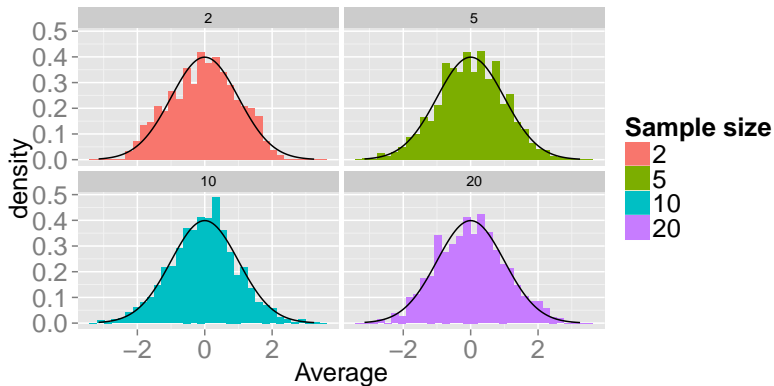
Flip coin (normalized histograms)

Central Limit Theorem: speed of convergence



6 faces dice (normalized histograms)

Central Limit Theorem: speed of convergence

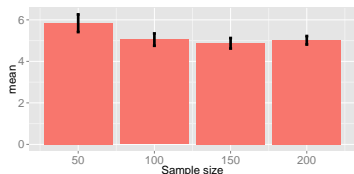


Continuous Uniform distribution in $[0, 10]$ (normalized histograms)

The Standard Error

If n is large enough

- $\bar{X}_n = \frac{X_n}{n} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$.
- σ/\sqrt{n} is called the standard error and is also the standard deviation of \bar{X}_n .
- Hence, the standard error is a good way for representing error bars when plotting averages (it decreases with number of samples accounting for a better certainty for larger samples size.).



Mean of continuous uniform distribution in $[0, 10]$
with different sample size.