Statistical Distributions.

Statistics for Data Science

Bruno Damásio

bdamasio@novaims.unl.pt

@bmpdamasio

2025/2026

Table of contents

Motivation and important concepts

Motivation

- A **random variable** is a variable that can take on certain numerical values with certain probabilities.
- The collection of these probabilities is called the **probability distribution** for the random variable.
- A **probability distribution** specifies how the total probability (which is always 1) is distributed among the various possible outcomes.

Motivation

The building blocks of statistical models are probability distributions. Specifying a model implies finding the probability distributions that describe the process of data generation. A statistical distribution describes how values are distributed. This allows to estimate the probability of any specific observation in a sample space.

Important Concepts

With every random variable X, we associate a function called the cumulative distribution of X.

Cumulative Distribution Function

The cumulative distribution function or cdf of a random variable X, denoted by $F(x)$, is defined by

$$F(x) = P(X \leq x), \text{for all } x$$

Example

Consider the experiment of tossing three fair coins, and let $X =$ number of heads observed. The cdf of X is

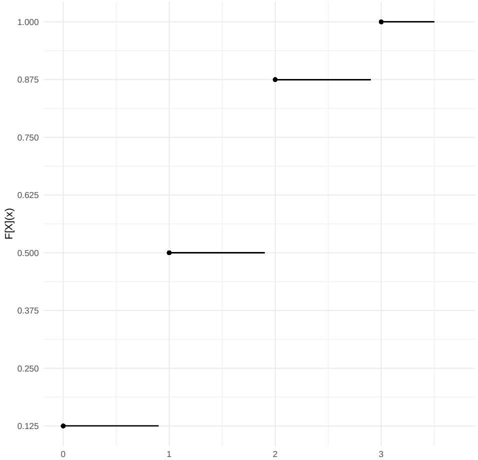
$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ \frac{1}{8}, & 0 \leq x < 1 \\ \frac{1}{2}, & 1 \leq x < 2 \\ \frac{7}{8}, & 2 \leq x < 3 \\ 1, & 3 \leq x < \infty \end{cases} \tag{1}$$

Thus,

$$F(2.5) = P(X \leq 2.5) = P(X = 0, 1, \text{or } 2) = \frac{7}{8}$$

Example

Visually, the previous example is given by:



Important Concepts

Associated with a random variable X and its cdf F is another function, called either the probability density function (pdf) or probability mass function (pmf). The terms pdf and pmf refer, respectively, to the continuous and discrete cases.

Probability mass function

The probability mass function (pmf) of a discrete random variable X is given by

$$f(x) = P(X = x), \text{for all } x$$

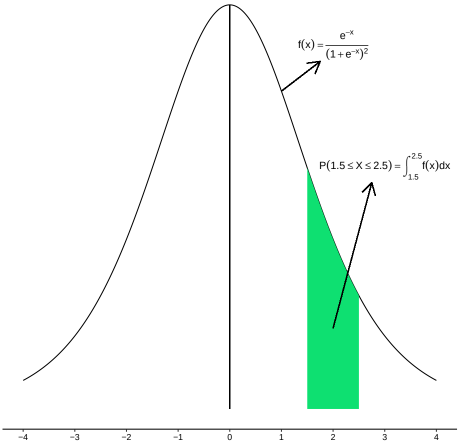
Probability density function

The probability density function (pdf) of a continuous random variable X is the function that satisfies

$$F(x) = \int_{-\infty}^{x} f(t)dt, \text{for all } x$$

Example

An example with the logistic curve:



Discrete Distributions

Discrete Distributions

- A random variable X is said to have a discrete distribution if the range of X, the sample space, is countable.
- In most situations, the random variable has integer-valued outcomes.
- In this section, we will study three discrete distributions: Bernoulli, Binomial and Poisson.

Bernoulli distribution

- The Bernoulli distribution is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$
- Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes-no question.

Bernoulli distribution

- Such questions lead to outcomes that are boolean-valued: a single bit whose value is success/yes/true/one with probability p and failure/no/false/zero with probability q.
- It can be used to represent a (possibly biased) coin toss where 1 and 0 would represent "heads" and "tails" (or vice versa), respectively, and p would be the probability of the coin landing on heads or tails, respectively.
- In particular, unfair coins would have $p \neq 0.5$.

Bernoulli distribution properties

- If X is a random variable with this distribution, $X \sim Ber(p)$, then:

$$P(X = 1) = p = 1 - P(X = 0) = 1 - q.$$

- The probability mass function f of this distribution, over possible outcomes x, is

$$f_{(x) = p^x (1-p)^{1-x} \quad \text{for } x \in \{0,1\}}$$

Bernoulli distribution properties

- Mean: $E\left(X\right) = p$ because

$$E[X] = P(X = 1) \cdot 1 + P(X = 0) \cdot 0 = p \cdot 1 + q \cdot 0 = p.$$

- Variance: $Var[X] = pq = p(1 - p)$ We first find

$$E[X^2] = P(X = 1) \cdot 1^2 + P(X = 0) \cdot 0^2 = p \cdot 1^2 + q \cdot 0^2 = p$$

From this follows

$$Var[X] = E[X^2] - E[X]^2 = p - p^2 = p(1 - p) = pq$$

Binomial distribution

- The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes-no question, and each with its own boolean-valued outcome: success/yes/true/one (with probability p) or failure/no/false/zero (with probability $q = 1 - p$).
- A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment and a sequence of outcomes is called a Bernoulli process
- For a single trial, i.e., $n = 1$, the binomial distribution is a Bernoulli distribution.

Binomial distribution

- The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N.

Binomial distribution properties

- In general, if the random variable X follows the binomial distribution with parameters $n \in \mathbf{N}$ and $p \in (0,1)$, we write $X \sim Bin(n,p)$.
- The probability of getting exactly x successes in n independent Bernoulli trials is given by the probability mass function:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, ..., n, \text{ where}$$

$\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ is the binomial coefficient, hence the name of the distribution.

Binomial distribution properties

- The formula can be understood as follows: x successes occur with probability p^x and $n - x$ failures occur with probability $(1 - p)^{n-x}$.
- However, the x successes can occur anywhere among the n trials, and there are $\binom{n}{x}$ different ways of distributing x successes in a sequence of n trials.

Binomial distribution properties

- If $X \sim Bin(n, p)$, that is, X is a binomially distributed random variable, n being the total number of experiments and p the probability of each experiment yielding a successful result, then the expected value of X is:

$$E[X] = np$$

  This follows from the linearity of the expected value along with fact that X is the sum of n identical Bernoulli random variables, each with expected value p. In other words, if $X_1, \ldots, X_n$ are identical (and independent) Bernoulli random variables with parameter p, then

$$X = X_1 + \cdots + X_n \text{ and}$$

$$E[X] = E[X_1 + \cdots + X_n] = p + \cdots + p = np.$$

Binomial distribution properties

- The variance is:

$$Var[X] = np(1-p).$$

  This similarly follows from the fact that the variance of a sum of independent random variables is the sum of the variances.

Problem

- Suppose there are twelve multiple choice questions in an English class quiz.
- Each question has five possible answers, and only one of them is correct.
- Find the probability of having four or less correct answers if a student attempts to answer every question at random.

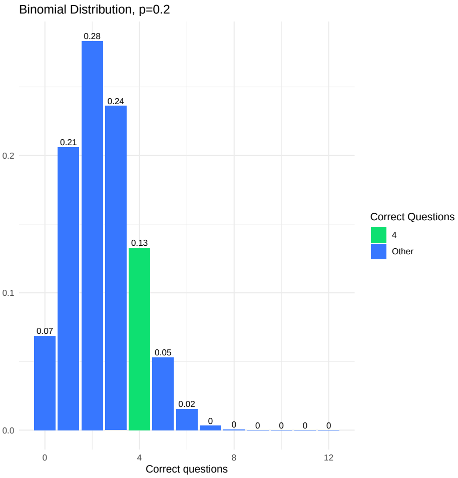
Solution

- Since only one out of five possible answers is correct, the probability of answering a question correctly by random is $1/5 = 0.2$.
- We can find the probability of having exactly 4 correct answers by random attempts as follows.

```
dbinom(4, size=12, prob=0.2)

## [1] 0.1328756
```

Solution



Binomial Distribution, p=0.2
0.28
0.24
0.21
0.13
0.07
0.05
0.02
0
0
0
0
0
0
0.2
0.1
0.0
0
4
8
12
Correct questions
Correct Questions
4
Other

Solution

To find the probability of having four or less correct answers by random attempts, we apply the function `dbinom()` with $x = 0, 1, 2, 3, 4$.

```
dbinom(0, size=12, prob=0.2) +
dbinom(1, size=12, prob=0.2) +
dbinom(2, size=12, prob=0.2) +
dbinom(3, size=12, prob=0.2) +
dbinom(4, size=12, prob=0.2)

## [1] 0.9274445
```

Solution

Alternatively, we can use the cumulative probability function for binomial distribution `pbinom()`.

```
pbinom(4, size=12, prob=0.2)

## [1] 0.9274445
```

The probability of four or less questions answered correctly by random in a twelve question multiple choice quiz is 92.7%.

Poisson distribution

- The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.
- The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.
- The Poisson distribution is popular for modeling the number of times an event occurs in an interval of time or space.

Poisson distribution

Some examples of variables that could follow a Poisson distribution:

- number of phone calls received by a call center per hour;
- number of decay events per second from a radioactive source;
- amount of mail a company receives each day;
- number of visits to the doctor office;
- number of visits to a museum.

Poisson distribution

- A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if, for $x = 0, 1, 2, ...$, the probability mass function of X is given by:

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

where

  - e is Euler's number ($e = 2.71828...$)
  - $x!$ is the factorial of x.
  - The positive real number λ is equal to the expected value of X and also to its variance.

$$E(X) = Var(X) = \lambda$$

Problem

If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

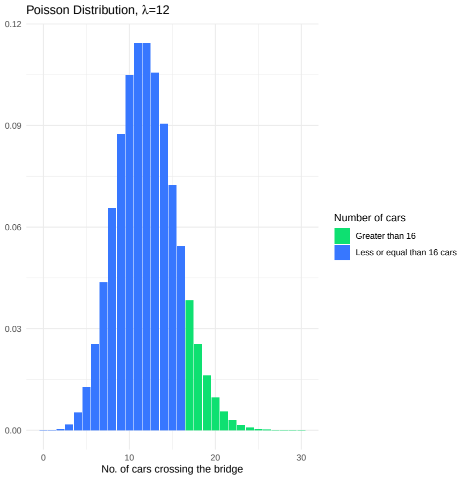
Solution

The probability of having sixteen or less cars crossing the bridge in a particular minute is given by the function `ppois()`.

```
ppois(16, lambda=12) # lower tail

## [1] 0.898709
```

Solution



Poisson Distribution, λ=12
0.12
0.09
0.06
0.03
0.00
0
10
20
30
No. of cars crossing the bridge
Number of cars
Greater than 16
Less or equal than 16 cars

Solution

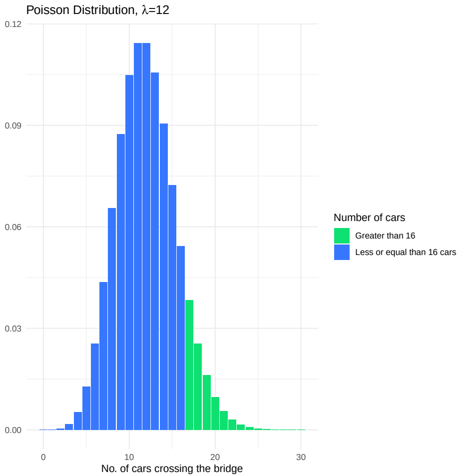
Hence the probability of having seventeen or more cars crossing the bridge in a minute is in the upper tail of the probability density function.

```
ppois(16, lambda=12, lower=FALSE) # upper tail

## [1] 0.101291
```

If there are twelve cars crossing a bridge per minute on average, the probability of having seventeen or more cars crossing the bridge in a particular minute is 10.1%.

Solution



Poisson Distribution, λ=12
0.12
0.09
0.06
0.03
0.00
0
10
20
30
No. of cars crossing the bridge
Number of cars
Greater than 16
Less or equal than 16 cars

Discrete probability distributions in R

In summary...

Distribution	pmf	cdf
Binomial	`dbinom`	`pbinom`
Poisson	`dpois`	`ppois`

Continuous Distributions

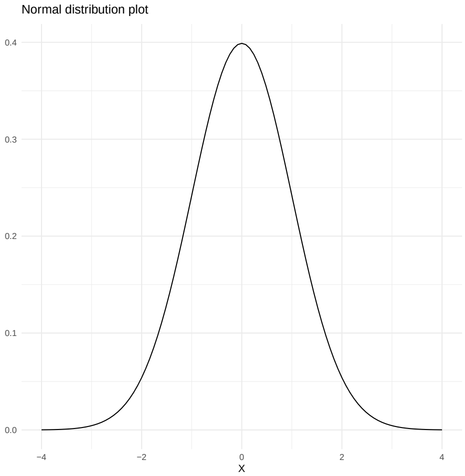
Normal distribution

- A random variable X is normally distributed with mean μ and variance σ^2 if it has the probability density function of X as:

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

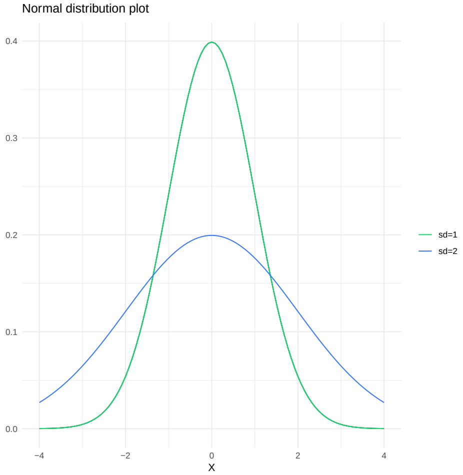
- This results in the usual bell-shaped curve.
- As shorthand notation we may use the expression $X \sim N(\mu, \sigma^2)$, indicating that X is distributed according to a normal distribution (denoted by N), with mean μ and variance σ^2.
- A standard normal distribution has a mean of 0 and standard deviation of 1. This is also known as the z distribution.

Normal distribution



Normal distribution

If σ^2 is large then the spread is going to be large, otherwise if the σ^2 value is small then the



spread will be small.

Finding normal distributions proportions - “less than” in R

- **Scenario**: vehicle speeds at a highway location have a normal distribution with a mean of 102 kmh and a standard deviation of 8 kmh.
- **Question**: what is the probability that a randomly selected vehicle will be going 120 kmh or slower?

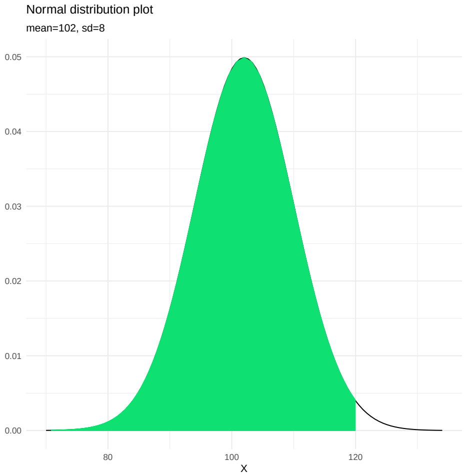
Finding normal distributions proportions - “less than” in R

Let’s construct a normal distribution with a mean of 102 and standard deviation of 8 to find the area less than 120.

```
pnorm(120,mean=102,sd=8, lower.tail=TRUE)

## [1] 0.9877755
```

Finding normal distributions proportions - “less than” in R



Finding normal distributions proportions - “greater than” in R

- **Scenario**: vehicle speeds at a highway location have a normal distribution with a mean of 102 kmh and a standard deviation of 8 kmh.
- **Question**: what is the probability that a randomly selected vehicle will be going more than 120 kmh?

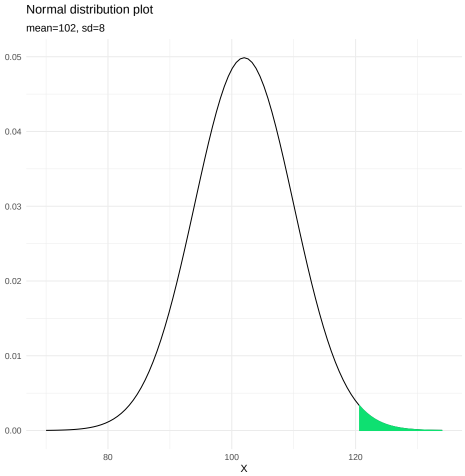
Finding normal distributions proportions - “greater than” in R

Let’s construct a normal distribution with a mean of 102 and standard deviation of 8 to find the area greater than 120.

```
pnorm(120,mean=102,sd=8, lower.tail=FALSE)

## [1] 0.01222447
```

Finding normal distributions proportions - “greater than” in R



Finding normal distributions proportions - "in between" in R

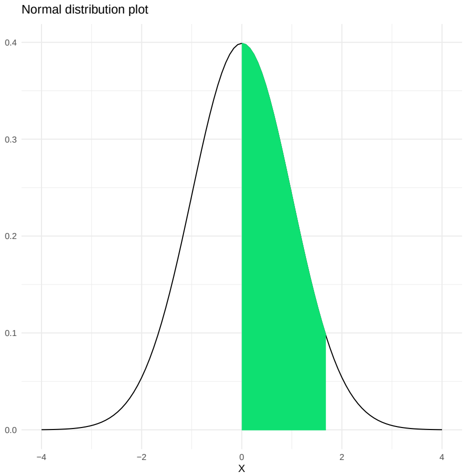
Question: what proportion of the standard normal distribution is between a z score of 0 and a z score of 1.75?

Recall that the **standard normal distribution** (i.e., distribution) has a mean of 0 and standard deviation of 1.

```
pnorm(1.75,mean=0,sd=1, lower.tail=TRUE)-
        pnorm(0,mean=0,sd=1, lower.tail=TRUE)

## [1] 0.4599408
```

Finding normal distributions proportions - “in between” in R



Finding normal distributions proportions - "more extreme than" in R

Question: what proportion of the standard normal distribution is more extreme than a z value of $\pm$ 2?

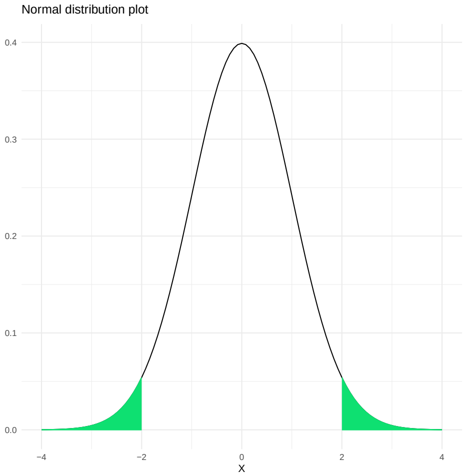
Recall that the **standard normal distribution** (i.e., distribution) has a mean of 0 and standard deviation of 1.

```
pnorm(2,mean=0,sd=1, lower.tail=FALSE)*2

## [1] 0.04550026
```

NOVA IMS INFORMATION MANAGEMENT SCHOOL

Finding normal distributions proportions - “more extreme than” in R



Finding quantiles given proportions in R

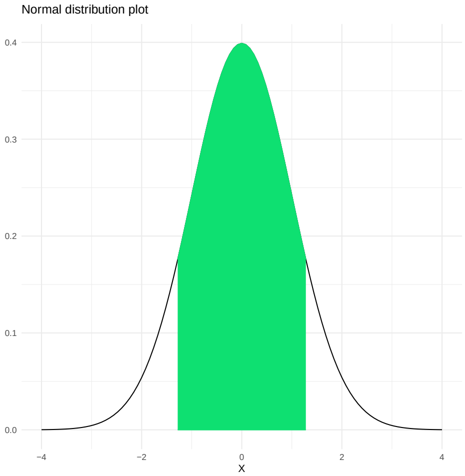
Question: what z-scores separate the middle 90% of the standard normal distribution from the outer 10%?

Recall that the **standard normal distribution** (i.e., distribution) has a mean of 0 and standard deviation of 1.

```
qnorm(0.95,mean=0,sd=1, lower.tail=TRUE)

## [1] 1.644854
```

Finding quantiles given proportions in R



Finding quantiles given proportions in R

Scenario: vehicle speeds at a highway location have a normal distribution with a mean of 102 kmh and a standard deviation of 8 kmh.

Question: What speed separates the top 10% of vehicles?

Finding quantiles given proportions in R

We will construct a normal distribution with a mean of 102 and standard deviation of 8. We will find the point that separates the to 0.10 from the bottom 0.90:

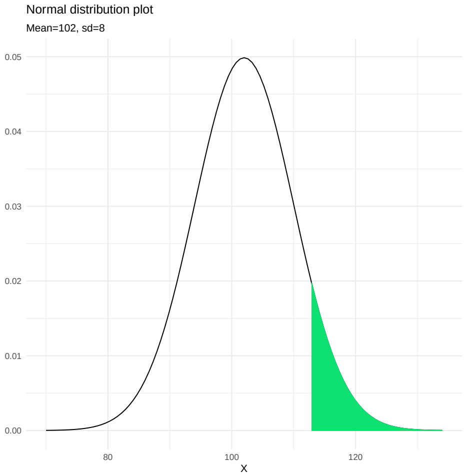
```
qnorm(0.9,mean=102,sd=8, lower.tail=TRUE)

## [1] 112.2524

qnorm(0.1,mean=102,sd=8, lower.tail=FALSE)

## [1] 112.2524
```

Finding quantiles given proportions in R



χ^2 distribution

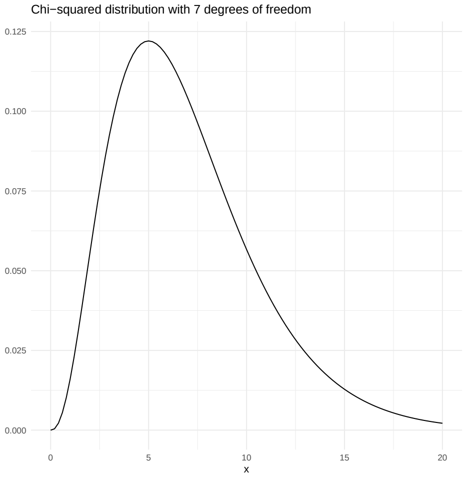
- If $X_1, X_2, ..., X_n$ are n independent random variables having the standard normal distribution, then the following quantity follows a Chi-squared distribution with n degrees of freedom.
- Its mean is n, and its variance is $2n$.

$$V = X_1^2 + X_2^2 + ... + X_n^2 \sim \chi^2_{(n)}, \; n > 0$$

- Here is a graph of the χ^2-squared distribution 7 degrees of freedom.

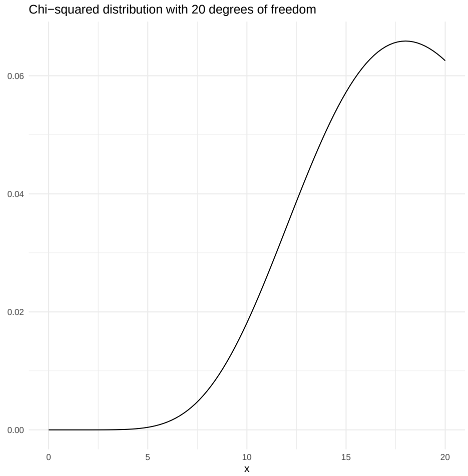
NOVA IMS INFORMATION MANAGEMENT SCHOOL

χ^2 distribution



χ^2 distribution

Here is a graph of the χ^2-squared distribution 20 degrees of freedom.



Problem

Find the 95th percentile of the chi-squared distribution with 7 degrees of freedom.

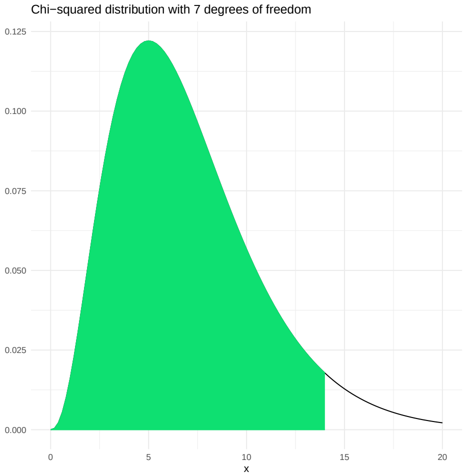
Solution

We apply the quantile function `qchisq()` of the Chi-Squared distribution against the decimal values 0.95.

```
qchisq(.95, df=7)

## [1] 14.06714
```

Solution

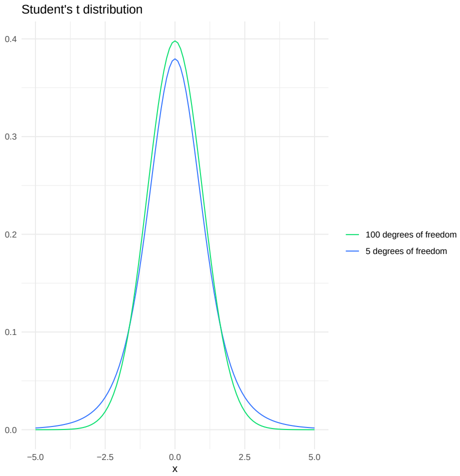


Chi−squared distribution with 7 degrees of freedom
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
x

Student's t distribution

- The t-distribution plays a role in a number of widely used statistical analyses, including Student's t-test for assessing the statistical significance of the difference between two sample means, the construction of confidence intervals for the difference between two population means, and in linear regression analysis.
- The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean.

Student's t distribution



Problem

Find the 2.5th and 97.5th percentiles of the Student's t distribution with 5 degrees of freedom.

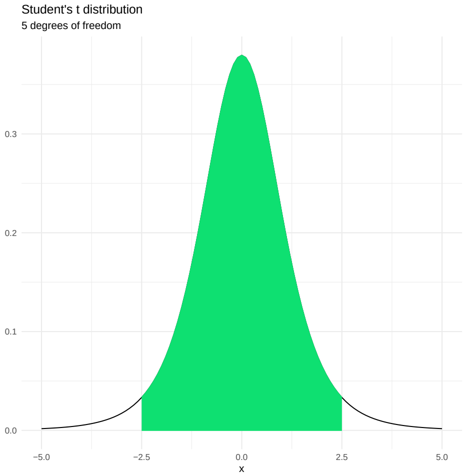
Solution

- We apply the quantile function `qt()` of the Student's t distribution against the decimal values 0.025 and 0.975.

```
qt(c(.025, .975), df=5) # 5 degrees of freedom
## [1] -2.570582  2.570582
```

- The 2.5th and 97.5th percentiles of the Student t distribution with 5 degrees of freedom are -2.5706 and 2.5706 respectively.

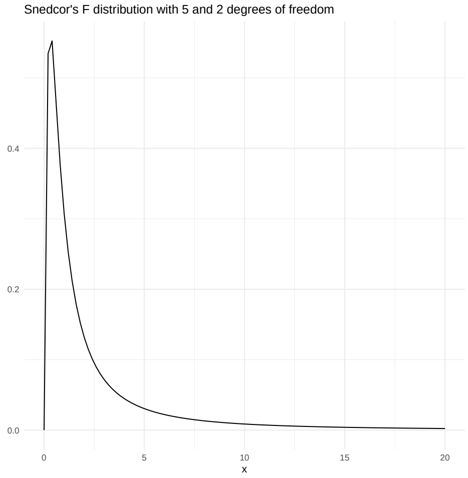
Solution


Student's t distribution
5 degrees of freedom
0.3
0.2
0.1
0.0
−5.0
−2.5
0.0
2.5
5.0
x

Snedcor's F distribution

- In probability theory and statistics, the F-distribution, also known as Snedecor's F distribution or the Fisher-Snedecor distribution (in honor to Ronald Fisher and George W. Snedecor) is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA), e.g., F-test.
- Here is a graph of the F distribution with $(5, 2)$ degrees of freedom.

Snedcor's F distribution



Problem

Find the 95th percentile of the F distribution with $(12, 5)$ degrees of freedom.

Solution

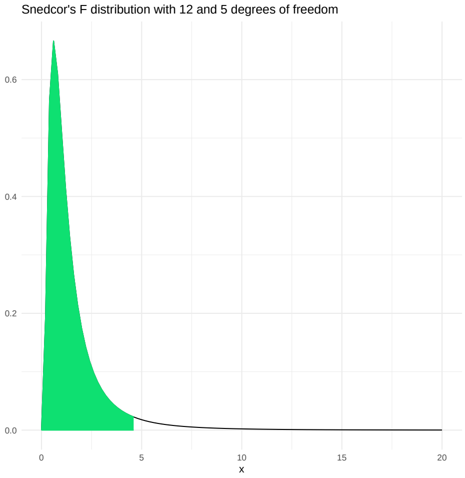
- We apply the quantile function `qf()` of the F distribution against the decimal value 0.95.

```
qf(.95, df1=12, df2=5)

## [1] 4.677704
```

- The 95th percentile of the F distribution with $(12, 5)$ degrees of freedom is 4.677704.

Solution



Snedcor's F distribution with 12 and 5 degrees of freedom
0.6
0.4
0.2
0.0
0
5
10
15
20
x

Continuous probability distributions in R

In summary...

Distribution	pdf	cdf	quantile
Normal	`dnorm`	`pnorm`	`qnorm`
Chi-squared	`dchisq`	`pchisq`	`qchisq`
t-student	`dt`	`pt`	`qt`
F	`df`	`pf`	`qf`

How do these distributions relate?

Figure 1: Continuous Distributions Relationships