Math Revision Session

Statistics (4): Representative continuous random variables

Jukina HATAKEYAMA

The University of Osaka, Department of Economics

May 19, 2025

1. Normal Distribution

2. Standard Normal Distribution

3. Chi-Squared Distribution

4. t-Distribution

5. Degrees of Freedom

1. Normal Distribution
2. Standard Normal Distribution
3. Chi-Squared Distribution
4. t-Distribution
5. Degrees of Freedom

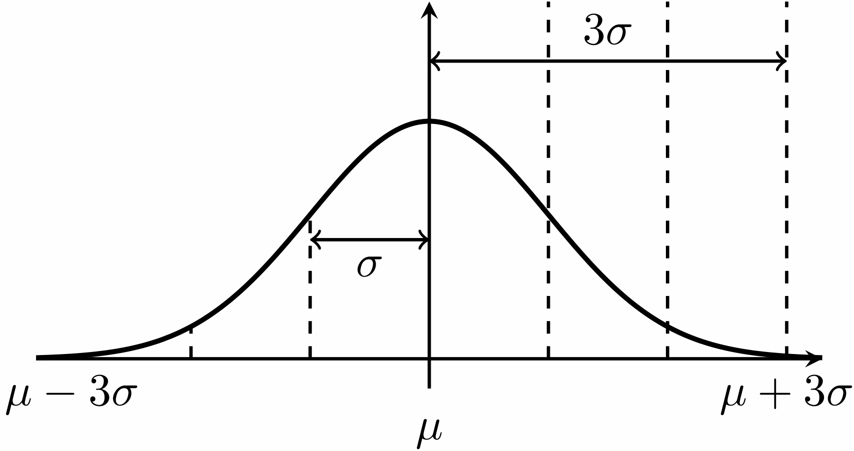
Normal Distribution (Gaussian Distribution)

- The normal distribution, also called the **Gaussian distribution**, is a continuous probability distribution.
- It is characterized by two parameters: the **mean** (μ) and the **standard deviation** (σ).
- The probability density function (PDF) of a normal distribution is given by:

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

- The graph of the normal distribution is bell-shaped and symmetric around the mean μ.
- The area under the curve represents probabilities. The total area under the curve is 1.
- The normal distribution is widely used in statistics due to the Central Limit Theorem.

PDF of normal distribution:



Coverage of the Probability Density Function

- The interval $\mu - 3\sigma$ to $\mu + 3\sigma$ covers almost all of the probability density function.
- In a normal distribution, approximately 99.7% of the data falls within this range.
- This range is commonly used to define practical limits in statistics and probability.

- The notation $X \sim N_{\mathbb{R}}(\mu, \sigma^2)$ represents that the random variable X follows a normal (Gaussian) distribution on $\mathbb{R}$.
- μ is the mean (expected value) of the distribution.
- σ^2 is the variance, which measures the spread of the distribution.
- The probability density function (PDF) of X is given by:

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

1. Normal Distribution
2. Standard Normal Distribution
3. Chi-Squared Distribution
4. t-Distribution
5. Degrees of Freedom

Standard Normal Distribution

- The standard normal distribution is a special case of the normal distribution where the mean is 0 and the variance is 1:

$$Z \sim N_{\mathbb{R}}(0, 1).$$

- The probability density function (PDF) is given by:

$$\phi(z) = \frac{1}{\sqrt{2\phi}} \exp\left(-\frac{z^2}{2}\right).$$

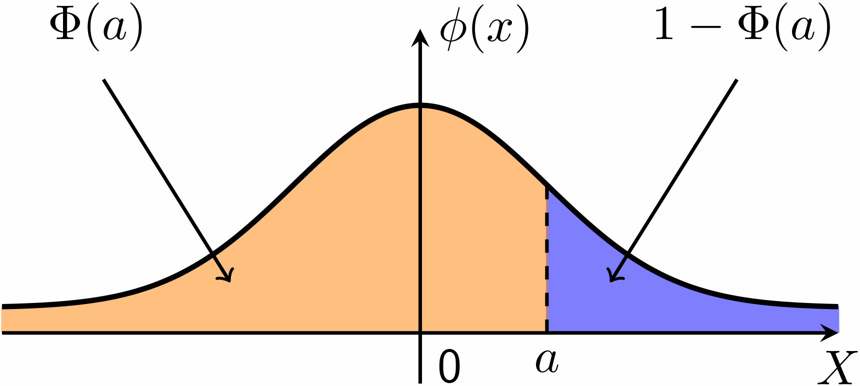
- The cumulative distribution function (CDF) is:

$$\Phi(z) = \int_{-\infty}^{z} \phi(t) dt.$$

- The standard normal distribution is symmetric about $Z = 0$, meaning:

$$\phi(-z) = \phi(z), \quad \Phi(-z) = 1 - \Phi(z).$$

PDF of standard normal distribution:



Linear Transformation of a Normal Variable

- If $X \sim N(\mu, \sigma^2)$, then for any constants a and b,
$$Y = aX + b$$
also follows a normal distribution.
- The mean and variance of Y are given by:
$$\mathbb{E}[Y] = a\mathbb{E}[X] + b = a\mu + b,$$
$$\mathsf{Var}(Y) = a^2\mathsf{Var}(X) = a^2\sigma^2.$$
- That is, the transformed variable follows:
$$Y \sim N(a\mu + b, a^2\sigma^2).$$
- In particular, if $X \sim N(0, 1)$, then $Y = \sigma X + \mu$ follows:
$$Y \sim N(\mu, \sigma^2).$$

Standardization and Normalization of a Normal Distribution

- **Standardization** (Z-score transformation):
  - Given $X \sim N(\mu, \sigma^2)$, we define the standardized variable:

    $$Z = \frac{X - \mu}{\sigma}.$$

  - This transformation results in:

    $$Z \sim N(0, 1),$$

    meaning Z follows the standard normal distribution.

1. Normal Distribution
2. Standard Normal Distribution
3. Chi-Squared Distribution
4. t-Distribution
5. Degrees of Freedom

Chi-Squared Distribution

The chi-squared distribution is a special case of the Gamma distribution. It is used in statistical tests, such as the chi-squared test, to assess the goodness of fit or to test for independence in contingency tables.

- The chi-squared distribution is defined by its degrees of freedom (df), which usually corresponds to the number of independent standard normal variables squared and summed.
- The distribution is positively skewed, and as the degrees of freedom increase, it becomes more symmetric, approaching a normal distribution.
- It is used in hypothesis testing, particularly for categorical data.

Let $X_1, X_2, \ldots, X_n$ be independent random variables, each following a normal distribution $N(\mu, \sigma^2)$.
Define the random variable W as:

$$W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

Where X_i are independent standard normal variables. Then, the random variable W follows a chi-squared distribution with n degrees of freedom:

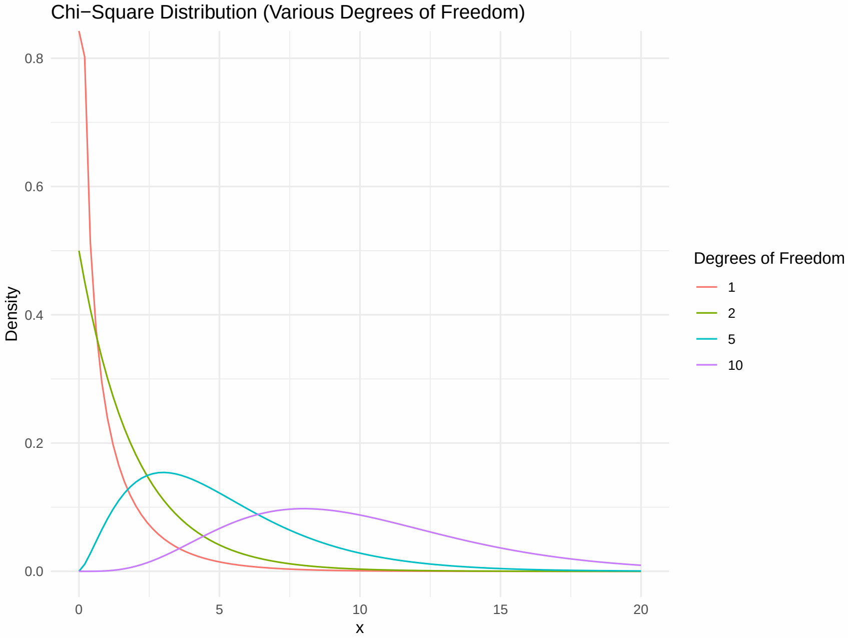
$$W \sim \chi^2(n)$$

This means that the sum of the squares of independent standard normal variables follows a chi-squared distribution.

For a chi-squared distribution with n degrees of freedom ($W \sim \chi^2(n)$), the mean and variance are as follows:

- **Mean:** The mean of a chi-squared distribution is equal to its degrees of freedom:
$$\mathbb{E}[W] = n$$
- **Variance:** The variance of a chi-squared distribution is twice the degrees of freedom:
$$\text{Var}(W) = 2n$$

This means that as the degrees of freedom increase, the distribution becomes more spread out, and the shape becomes closer to a normal distribution.


Chi−Square Distribution (Various Degrees of Freedom)
Density
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
x
Degrees of Freedom
1
2
5
10

1. Normal Distribution
2. Standard Normal Distribution
3. Chi-Squared Distribution
4. t-Distribution
5. Degrees of Freedom

t-Distribution

The t-distribution is defined as the ratio of a standard normal random variable and the square root of a chi-squared distributed random variable divided by its degrees of freedom. Specifically, let $Z \sim N(0,1)$ (standard normal) and $W \sim \chi^2(n)$ (chi-squared distribution with n degrees of freedom). The t-distributed random variable t is given by:

$$t = \frac{Z}{\sqrt{W/n}} \sim t(n)$$

Where:

- Z is a standard normal random variable ($Z \sim N(0,1)$),
- W is a chi-squared random variable with n degrees of freedom ($W \sim \chi^2(n)$),
- The denominator involves $\sqrt{W/n}$, which scales the chi-squared variable by its degrees of freedom.

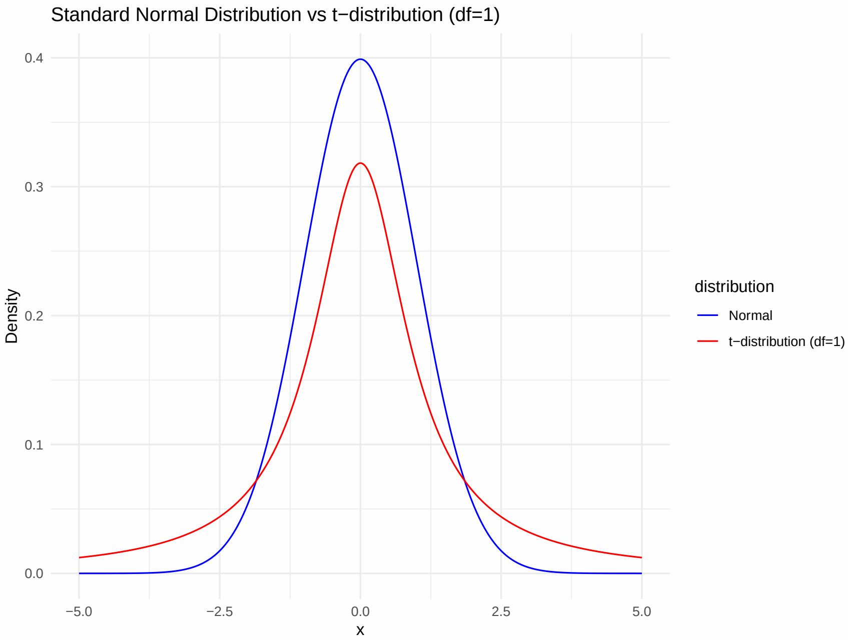
The t-distribution is used for hypothesis testing when the sample size is small and the population variance is unknown.

Summary of t-Distribution Characteristics

The t-distribution has the following key characteristics:

- It is defined as the ratio of a standard normal random variable Z to the square root of a chi-squared distributed variable W divided by its degrees of freedom n:
$$t = \frac{Z}{\sqrt{W/n}}$$
- The shape of the t-distribution is similar to the standard normal distribution, but it has heavier tails.
- As the degrees of freedom n increase, the t-distribution approaches the standard normal distribution.
- The t-distribution is used in hypothesis testing, particularly when the sample size is small or the population variance is unknown.

The t-distribution is essential for confidence intervals and significance testing in cases of small samples.


Standard Normal Distribution vs t−distribution (df=1)
0.4
0.3
0.2
0.1
0.0
Density
−5.0
−2.5
0.0
2.5
5.0
x
distribution
Normal
t−distribution (df=1)

Expectation and Variance of t-Distribution

The t-distribution has the following properties:

- **Expectation (Mean)**: The expected value (mean) of the t-distribution is given by:

  $$E(t) = 0$$

  This holds for all degrees of freedom n, as the t-distribution is symmetric around 0.

- **Variance**: The variance of the t-distribution depends on the degrees of freedom n and is given by:

  $$\text{Var}(t) = \frac{n}{n-2}, \quad \text{for} \quad n > 2$$

  For degrees of freedom $n \leq 2$, the variance does not exist. Specifically:

  - When $n = 1$, the variance is infinite.
  - When $n = 2$, the variance is undefined (it diverges).

Thus, while the expectation is always 0, the variance exists and is finite only when the degrees of freedom are greater than 2. For small n, the t-distribution has heavier tails, which is why its variance becomes large or undefined for very small degrees of freedom.

Exponential Distribution

The exponential distribution is a continuous probability distribution that is commonly used to model the time between events in a Poisson process.

- The probability density function (PDF) of the exponential distribution is given by:

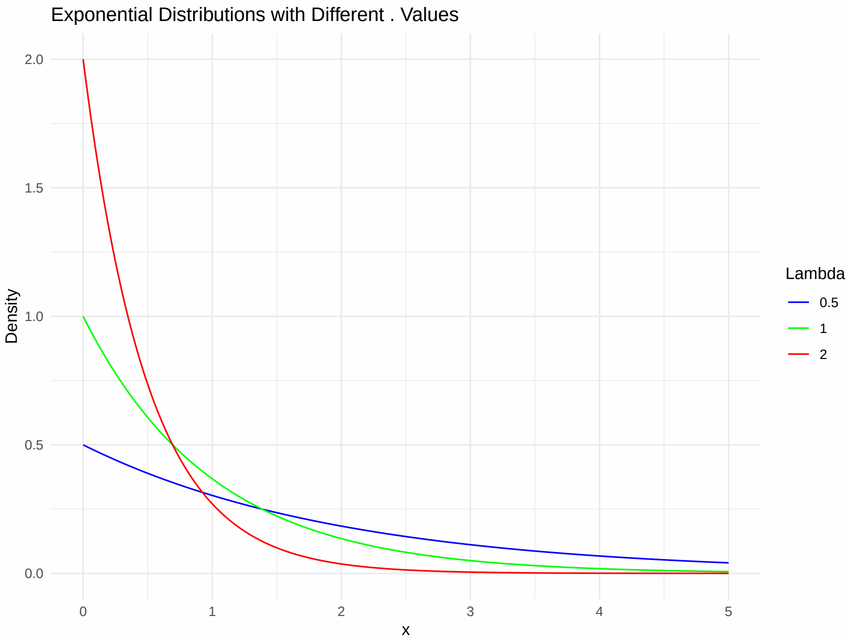
  $$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

  where $\lambda > 0$ is the rate parameter (also called the inverse of the mean), which controls the "rate" of events.
- **Key properties:**
  - **Mean**: $E(X) = \frac{1}{\lambda}$
  - **Variance**: $\mathrm{Var}(X) = \frac{1}{\lambda^2}$

- **Memoryless Property**: The exponential distribution is memoryless, meaning that the probability of an event occurring in the next time interval is independent of how much time has already passed:

$$P(X > x + y \mid X > x) = P(X > y)$$

- **Applications**: The exponential distribution is widely used in various fields, such as:
  - Modeling the time between arrivals of events (e.g., customers arriving at a service counter).
  - Describing the lifespan of certain types of machinery or devices.


Exponential Distributions with Different . Values
Density
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
x
Lambda
0.5
1
2

Expectation and Variance of Exponential Distribution

For an exponentially distributed random variable $X \sim \mathrm{Exp}(\lambda)$, where λ is the rate parameter, the following properties hold:

- **Expectation (Mean)**: The expected value of X, denoted $E(X)$, is given by:

  $$E(X) = \frac{1}{\lambda}$$

  This means the average time between events is $\frac{1}{\lambda}$.
- **Variance**: The variance of X, denoted $\mathrm{Var}(X)$, is given by:

  $$\mathrm{Var}(X) = \frac{1}{\lambda^2}$$

  This means the spread of the times between events is inversely proportional to λ^2.

These properties are key to understanding the exponential distribution. The mean $\frac{1}{\lambda}$ represents the average waiting time for the next event, and the variance $\frac{1}{\lambda^2}$ represents the variability in the waiting time.

Interpretation of Exponential Distribution as Waiting Time

The exponential distribution is often used to model the time between events in a Poisson process, where events occur independently at a constant rate.

- **Memoryless Property**: This property makes it suitable for modeling waiting times because the past does not affect the future.
- **Poisson Process**: In a Poisson process, events occur independently and at a constant average rate λ. The time between events follows an exponential distribution, making it a natural model for waiting times or intervals between occurrences.
- **Expectation and Variance as Waiting Time Interpretation**:
  - The expected value $E(X) = \frac{1}{\lambda}$ represents the average waiting time until the next event.
  - The variance $\text{Var}(X) = \frac{1}{\lambda^2}$ indicates the variability in the waiting time.
- **Example Applications**:
  - Time between customer arrivals at a service counter.
  - Time until failure of a machine or system.

1. Normal Distribution
2. Standard Normal Distribution
3. Chi-Squared Distribution
4. t-Distribution
5. Degrees of Freedom

Degrees of Freedom

Degrees of freedom (DF) is a concept in statistics that refers to the number of independent pieces of information available to estimate a parameter.

- **Definition**: Degrees of freedom represent the number of independent values or quantities that can vary in an analysis, after accounting for any constraints or parameters that have been fixed.
- When calculating a statistic, the degrees of freedom represent the number of independent values remaining after applying constraints or estimating parameters.

Degrees of freedom are essential for evaluating the precision and reliability of statistical estimates.

- **Common Examples**:
  - In calculating variance using the sample mean, the degrees of freedom for a sample of size n is $n-1$. This is because the mean is a fixed parameter, leaving $n-1$ independent values.
  - The degrees of freedom for the sample variance are $n-1$, since the sample mean has already been fixed.
- **Importance**: Degrees of freedom play a key role in statistical estimation and hypothesis testing. For example, the shape of the t-distribution, chi-squared distribution, and F-distribution depends on the degrees of freedom.
- **Degrees of Freedom and Distributions**:
  - In distributions such as the t-distribution and chi-squared distribution, the degrees of freedom affect the shape of the distribution.
  - As the degrees of freedom increase, the t-distribution approaches the standard normal distribution.

Example of Degrees of Freedom Using the Mean

Consider a set of 10 random variables $X_1, X_2, \ldots, X_{10}$ with known realizations. Let's assume that the true mean of these variables is 10.

- The true mean μ is fixed at 10. In this case, the sum of all the values must equal $10 \times 10 = 100$.
- For 9 of the values $X_1, X_2, \ldots, X_9$, the values can be chosen freely; each of them can take any value.
- However, once these 9 values are chosen, the 10th value X_{10} is no longer independent. It must be such that the sum of all values equals 100, so it is determined by the equation:

$$X_{10} = 100 - \sum_{i=1}^{9} X_i$$

This means that once 9 values are chosen, the 10th value is constrained to ensure the sum equals the true mean of 10.

- **Degrees of Freedom Interpretation**: Since we can freely choose 9 values but the 10th value is determined by the others, the degrees of freedom for the dataset is 9. In other words, there are 9 independent pieces of information, and the 10th is dependent on them.
- **Why This Matters**: This explains why the sample variance formula uses $n - 1$ as the degrees of freedom. When estimating the variance from a sample, one parameter (the mean) is estimated from the data, so we lose 1 degree of freedom.

Thus, the degrees of freedom reflect the number of independent pieces of data that can vary when calculating statistics.

Consider two variables x and y that satisfy the equation:

$$x + y = 1$$

In this case, the freedom we have in choosing values for x and y is limited by the constraint.

- **Free Choice for One Variable**: Since $x + y = 1$, if we choose a value for x, the value of y is automatically determined as $y = 1 - x$. Therefore, we only have one independent variable that can be freely chosen.
- **Degrees of Freedom**: Despite there being two variables x and y, because they are constrained by the equation $x + y = 1$, there is only one independent piece of information. Thus, the degrees of freedom for this system is 1.
- **Interpretation in Terms of Constraints**: The constraint $x + y = 1$ reduces the number of free variables from 2 to 1. For each possible value of x, there is a unique corresponding value of y, so only one of the variables is free to vary.