Statistics and Probability

Bas Machielsen

2025-06-13

Agenda

This presentation provides a refresher on the fundamental statistical concepts that form the bedrock of econometrics.

1. Basic Probability Theory
2. Conditional Probability
3. Random Variables
4. Discrete Probability Distributions
5. Continuous Probability Distributions
6. The Normal Distribution
7. Covariance and Correlation
8. Sampling
9. Sampling Distributions
10. The Central Limit Theorem
11. Estimation
12. Hypothesis Testing

1. Basic Probability Theory

Experiments, Outcomes, and Sample Spaces

- Experiment: A process or action whose result is uncertain.
  - *Example:* Rolling a six-sided die.
  - *Example:* Surveying a household to ask about their income.
  - *Example:* Observing next year's GDP growth rate.
- Outcome: A single possible result of an experiment.
  - *Example:* The die shows a 4.
  - *Example:* The household's income is $52,000$.
  - *Example:* GDP growth is 2.3.
- Sample Space (S): The set of *all possible* outcomes of an experiment.
  - *Example (Die Roll):* $S = 1, 2, 3, 4, 5, 6$
  - *Example (Household Income)*: $S \in [0, \infty]$

Events

- Event: A subset of the sample space; a collection of one or more outcomes. We can calculate probabilities for events.

Using the die roll example where $S = 1, 2, 3, 4, 5, 6$:

- **Event A:** The outcome is an even number.
  - $A = 2, 4, 6$
  - The probability of Event A is $P(A) = 3/6 = 0.5$
- **Event B:** The outcome is greater than 4.
  - $B = 5, 6$
  - The probability of Event B is $P(B) = 2/6 \approx 0.33$
- **The intersection of A and B ($A \cap B$):** The outcome is even AND greater than 4.
  - $A \cap B = 6$
  - $P(A \cap B) = 1/6$

2. Conditional Probability

The Probability of A, Given B

Conditional Probability is the probability of an event occurring, given that another event has already occurred.

The probability of event A occurring given that event B has occurred is written as $P(A|B)$.

Definition:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Intuition: We are restricting our sample space. We know B happened, so the "universe" of possible outcomes is now just B. Within that new universe, we want to know the chance that A also happens.

Conditional Probability: Example

Let's use our die roll example again: $S = 1, 2, 3, 4, 5, 6$,
$A = \{\text{Outcome is an even number}\} = \{2, 4, 6\}$,
$B = \{\text{Outcome is greater than } 4\} = \{5, 6\}$

Question: What is the probability that the number is even, *given* that we know it is greater than 4? We want to find $P(A|B)$.

1. **Find $P(B)$:** The probability of rolling a number greater than 4 is $P(B) = 2/6$.
2. **Find $P(A \cap B)$:** The probability of rolling a number that is even AND greater than 4 is $P(6) = 1/6$.
3. **Apply the formula:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{2/6} = \frac{1}{2} = 0.5$$

Intuition Check: If we know the outcome is in $B = 5, 6$, there are only two possibilities. Of these, only one (6) is even. So the probability is $1/2$. It matches!

3. Random Variables

Definition and Types

Random Variable (RV): A variable whose value is a numerical outcome of a random phenomenon. We use capital letters (e.g., X, Y) to denote a random variable. There are two main types of random variables:

Discrete Random Variable: A variable that can only take on a finite or countably infinite number of distinct values.

- *Example:* The number of heads in three coin flips (X can be 0, 1, 2, 3).
- *Example:* The number of defaults in a portfolio of 100 loans (X can be 0, 1, ..., 100).

Continuous Random Variable: A variable that can take on any value within a given range.

- *Example:* The exact price of a stock tomorrow.
- *Example:* The annual percentage growth in GDP (Y could be 2.1%, 2.11%, 2.113%...).

4. Distributions for Discrete RVs

Probability Mass Function (PMF)

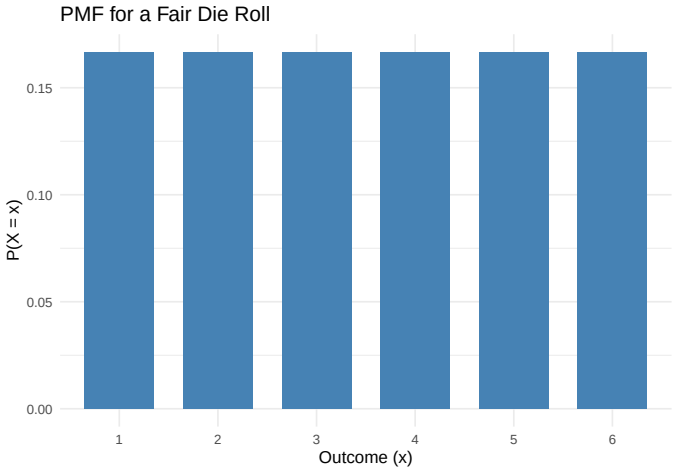
For a discrete random variable X, the **Probability Mass Function (PMF)** gives the probability that X is exactly equal to some value x.

$$f(x) = P(X = x)$$

A PMF has two key properties: 1. $0 \leq f(x) \leq 1$ for all x. 2. $\sum f(x) = 1$ (The sum of probabilities over all possible values is 1).

PMF Example

Example: Let X be the outcome of a fair die roll. The PMF is: $f(1) = 1/6$, $f(2) = 1/6$, $\ldots$, $f(6) = 1/6$.



Expected Value

The **Expected Value** of a discrete random variable X, denoted $E[X]$ or μ, is the long-run average value of the variable. It's a weighted average of the possible outcomes, where the weights are the probabilities.

Definition:

$$E[X] = \mu = \sum_x x \cdot P(X = x)$$

Example: Expected value of a fair die roll.

$E[X] = (1 \times 1/6) + (2 \times 1/6) + (3 \times 1/6) + (4 \times 1/6) + (5 \times 1/6) + (6 \times 1/6)$

$E[X] = (1 + 2 + 3 + 4 + 5 + 6)/6 = 21/6 = 3.5$

Note: The expected value doesn't have to be a possible outcome!

Variance and Standard Deviation

Variance, denoted $Var(X)$ or σ^2, measures the spread or dispersion of a random variable around its mean. A larger variance means the outcomes are more spread out.

Definition:

$$Var(X) = \sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 \cdot P(X = x)$$

Standard Deviation, $SD(X)$ or σ, is the square root of the variance. It's often easier to interpret because it's in the same units as the random variable itself.

$$SD(X) = \sigma = \sqrt{Var(X)}$$

Example: The Bernoulli Distribution

The **Bernoulli distribution** is a fundamental discrete distribution for any experiment with two outcomes, typically labeled "success" (1) and "failure" (0).

Let X be a Bernoulli random variable where $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

- **Econometric relevance:** Models binary outcomes like employed/unemployed, default/no-default, buy/don't-buy.
- **Expected Value:** $E[X] = (1 \times p) + (0 \times (1 - p)) = p$
- **Variance:**

$$\begin{aligned} Var(X) &= (1-p)^2 \times p + (0-p)^2 \times (1-p) \\ &= (1-p)^2 \times p + p^2 \times (1-p) \\ &= p(1-p) \times [(1-p) + p] = p(1-p) \end{aligned}$$

5. Distributions for Continuous RVs

Probability Density Function (PDF)

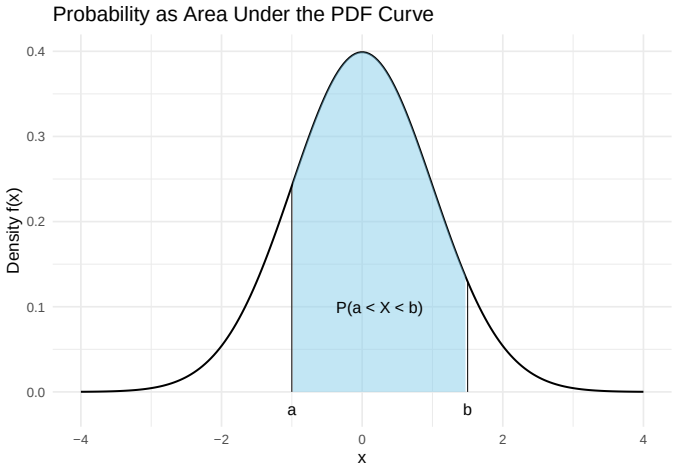
For a **continuous** random variable, the probability of it taking on any *single* specific value is zero! $P(X = x) = 0$. Why? Because there are infinitely many possible values.

Instead, we use a **Probability Density Function (PDF)**, $f(x)$.

Key Idea: Probability is represented by the **area under the curve** of the PDF.

$P(a \leq X \leq b) =$ Area under $f(x)$ between a and b.

Example PDF


Probability as Area Under the PDF Curve
Density f(x)
0.4
0.3
0.2
0.1
0.0
P(a < X < b)
a
b
−4
−2
0
2
4
x

Cumulative Distribution Function (CDF)

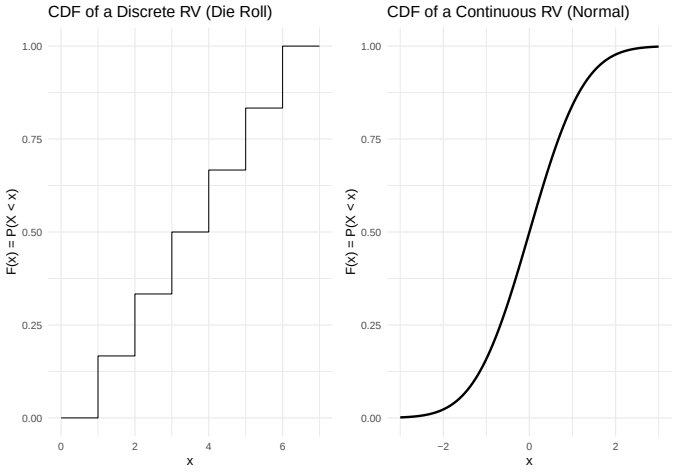
The **Cumulative Distribution Function (CDF)**, $F(x)$, gives the probability that a random variable X is *less than or equal to* a certain value x. It's a unifying concept for both discrete and continuous variables.

$$F(x) = P(X \leq x)$$

Properties:

- $F(x)$ is non-decreasing.
- $F(x)$ ranges from 0 to 1.
- For continuous RVs, $P(a \leq X \leq b) = F(b) - F(a)$.

Example CDF


CDF of a Discrete RV (Die Roll)
F(x) = P(X < x)
1.00
0.75
0.50
0.25
0.00
0
2
4
6
x
CDF of a Continuous RV (Normal)
F(x) = P(X < x)
1.00
0.75
0.50
0.25
0.00
−2
0
2
x

6. The Normal Distribution

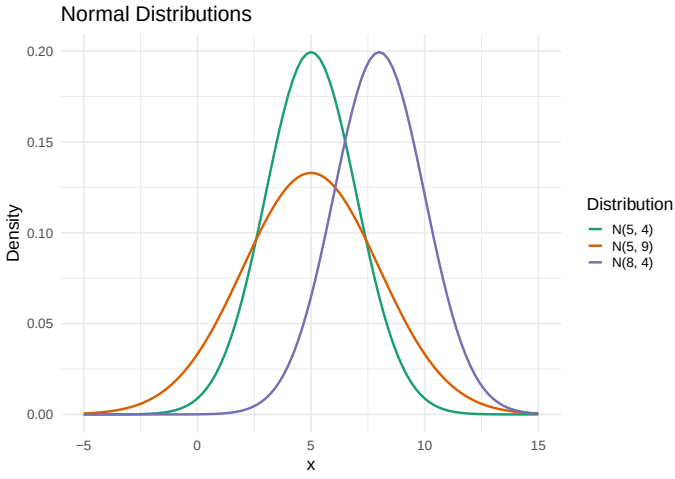
Properties of the Normal Distribution

The **Normal Distribution** is the most important probability distribution in statistics and econometrics. It is defined by its mean μ and its variance σ^2. We write $X \sim N(\mu, \sigma^2)$.

Properties:

- **Bell-shaped** and symmetric around its mean μ.
- Mean $=$ Median $=$ Mode.
- The curve is completely determined by μ (center) and σ (spread).

Example Normal Distribution



Linear Combinations of Normal Variables

An important property of the normal distribution is that linear combinations of independent normal variables are also normally distributed.

Rule 1: Scaling and Shifting

If $X \sim N(\mu, \sigma^2)$, then the new variable $Y = aX + b$ is also normally distributed:

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

Note that the new standard deviation is $|a|\sigma$.

Linear Combinations of Normal Variables (Cont.)

Rule 2: Sum/Difference of Independent Variables

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are **independent**, then their sum and difference are also normally distributed:

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Key point: Variances always add, even when subtracting the random variables.

Example

Let the time to complete Task A be $T_A \sim N(20, 3^2)$ minutes and Task B be $T_B \sim N(15, 4^2)$ minutes. What is the distribution of the total time $T_{total} = T_A + T_B$?

- New Mean: $\mu_{total} = \mu_A + \mu_B = 20 + 15 = 35$
- New Variance: $\sigma^2_{total} = \sigma^2_A + \sigma^2_B = 3^2 + 4^2 = 9 + 16 = 25$
- New Standard Deviation: $\sigma_{total} = \sqrt{25} = 5$

So, $T_{total} \sim N(35, 5^2)$. We can now calculate probabilities for the total time, e.g., the probability the total time is less than 45 minutes:

```
## [1] 0.9772499
```

The Standard Normal Distribution (Z)

The **Standard Normal Distribution** is a special case of the normal distribution with a mean of 0 and a variance of 1. $Z \sim N(0, 1)$.

Standardization (creating a Z-score): We can convert any normally distributed random variable $X \sim N(\mu, \sigma^2)$ into a standard normal variable Z using the formula:

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

Why is this useful? It allows us to use a single table (or software function) to find probabilities for *any* normal distribution. The Z-score tells us how many standard deviations an observation X is away from its mean μ.

Finding Probabilities

Historically, probabilities for the standard normal distribution were found using **Z-tables**, which provide $P(Z \leq z)$.

Today, we use software like R, Stata, or Python.

Example: Suppose annual returns on a mutual fund are normally distributed with a mean of 8% and a standard deviation of 10%. $X \sim N(0.08, 0.01)$. What's the probability of a negative return, $P(X < 0)$?

1. **Standardize the value:** $Z = (0 - 0.08)/0.10 = -0.8$
2. **Find the probability:** We need to find $P(X < 0) = P(\frac{X-0.08}{0.01} < \frac{0-0.08}{0.01}) = P(Z < -0.8)$.

Finding Probabilities (Cont.)

3. **Using R, Python or Stata:** The *pnorm*() and *norm.cdf* functions give the area to the left (the CDF).

```
pnorm(-0.8, mean = 0, sd = 1)
```

```
## [1] 0.2118554
```

```
from scipy.stats import norm
norm.cdf(-0.8)
```

```
## np.float64(0.2118553985833967)
```

So, there is about a **21.2%** chance of experiencing a negative return.

Conditional Density

Conditional Probability

Conditional probability is about how the probability of an event **A** changes when we know an event **B** has occurred.

- **Discrete Case:** We update probabilities for specific outcomes. $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- **Continuous Case:** What if we want to condition on a continuous random variable Y taking a specific value y? Since $P(Y = y) = 0$ for any continuous variable, the formula above is undefined.

We need to shift our thinking from the probability of *events* to the *probability density functions (PDFs)* of random variables.

Conditional Probability Density Functions (PDFs)

Definition

The conditional PDF of a random variable `X` given that `Y=y` is defined as the ratio of the joint PDF to the marginal PDF of Y.

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

Provided that the marginal density $f_Y(y) > 0$.

- $f_{X,Y}(x,y)$: The **joint PDF**, describing the probability density of `(X, Y)` occurring together.
- $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$: The **marginal PDF** of `Y`, found by "integrating out" `X`. It represents the distribution of `Y` on its own.

Intuition: Slicing the Joint Distribution

Think of the joint PDF $f_{X,Y}(x, y)$ as a 3D surface.

1. **Fix a value** `y` for the variable Y. This is like taking a 2D "slice" of the 3D surface at that `y`.
2. The shape of this slice gives the relative likelihood of `X`'s values, *given* that `Y=y`.
3. **Normalize the slice:** The area under this slice might not be 1. Dividing by $f_Y(y)$ (the area of the slice) scales it to become a valid probability density function.

This normalized slice *is* the conditional PDF, $f_{X|Y}(x|y)$.

Tbd: Visualization of this

hi

Conditional Expectation

The Average of the Conditional Distribution

Conditional expectation, $E[X|Y = y]$, is simply the expected value (the mean) of `X` calculated using its **conditional distribution**. It's the "center of mass" of that conditional slice we just discussed.

Formal Definitions

- Discrete case: if `X` and `Y` are discrete random variables, the expectation of `X` given `Y=y` is a weighted average using conditional probabilities:

$$E[X|Y = y] = \sum_x x \cdot P(X = x|Y = y)$$

- Continuous case: if `X` and `Y` are continuous random variables, the expectation is an integral using the conditional PDF:

$$E[X|Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx$$

A Crucial Distinction: $E[X|Y = y]$ vs. $E[X|Y]$

$E[X|Y = y]$ is a Function

The expression $E[X|Y = y]$ produces a **value** that depends on the specific, fixed `y` we conditioned on. We can think of this as a function of `y`.

Let's define a function $g(y)$:

$$g(y) = E[X|Y = y]$$

- If we know `Y=2`, we calculate $g(2) = E[X|Y = 2]$.
- If we know `Y=5`, we calculate $g(5) = E[X|Y = 5]$.

$E[X|Y]$ is a Random Variable

The notation $E[X|Y]$ (without specifying a value for Y) represents a **new random variable**.

- It is the random variable you get by taking the function $g(y)$ and plugging in the *random variable* Y itself.

$$E[X|Y] = g(Y)$$

- The value of this new random variable is not known until the value of Y is revealed.

Example: Suppose we find that the expected height of a child X given the mother's height Y is $E[X|Y = y] = 40 + 0.8y$. * $g(y) = 40 + 0.8y$ is the function. * $E[X|Y] = 40 + 0.8Y$ is a random variable whose outcome depends on the randomly selected mother's height Y.

Slide 5: The Law of Total Expectation (The Tower Property)

Tying It All Together

This powerful law connects unconditional and conditional expectations by using the distinction we just made. It states that the overall expected value of `X` is the expected value of the *random variable* $E[X|Y]$.

$$E[X] = E[E[X|Y]]$$

Intuition: "Averaging the Averages"

This means we can find the overall average of `X` in two steps: 1. First, find the conditional average of `X` for each possible value of `Y`. This gives you the function $g(y) = E[X|Y = y]$. 2. Then, find the expected value of that function with respect to the distribution of `Y`. This is what $E[g(Y)]$ means.

Example: Suppose a student's score `X` depends on their hours of study `Y`. * If they study `Y=0` hours, their expected score is 50. * If they study `Y=10` hours, their expected score is 90. * To find the overall average score, we would need to know the distribution of study hours (the distribution of `Y`) and then average these conditional expectations over that distribution. This is the essence of the Law of Total Expectation.

7. Covariance and Correlation

Covariance

Covariance measures the joint variability of two random variables. It tells us the *direction* of the linear relationship.

Definition:

$$Cov(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

- $Cov(X, Y) > 0$: X and Y tend to move in the same direction. When X is above its mean, Y tends to be above its mean.
- $Cov(X, Y) < 0$: X and Y tend to move in opposite directions.
- $Cov(X, Y) = 0$: No linear relationship between X and Y.

Drawback: The magnitude of covariance is hard to interpret because it depends on the units of X and Y. (e.g., Cov(GDP, Consumption) will be a huge number).

Correlation

The **Correlation Coefficient (ρ or r)** is a standardized version of covariance that measures both the *strength* and *direction* of the linear relationship between two variables.

Definition:

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Properties:

- Ranges from **-1 to +1**.
- $\rho = +1$: Perfect positive linear relationship.
- $\rho = -1$: Perfect negative linear relationship.
- $\rho = 0$: No linear relationship.
- It is unitless, making it easy to interpret and compare.

Correlation $\neq$ Causation

This is the most important lesson in all of econometrics.

A strong correlation between two variables does not mean that one *causes* the other. There could be:

1. **Reverse Causality:** Y causes X.
2. **Omitted Variable Bias (Lurking Variable):** A third variable Z causes both X and Y.

Classic Example:

- Ice cream sales (X) and drowning deaths (Y) are highly positively correlated.
- Does eating ice cream cause drowning? No.
- The omitted variable is **hot weather (Z)**, which causes people to both buy more ice cream and swim more (leading to more drownings).

8. Rules for Sums of Random Variables In General

1. Expected Value

Let X and Y be two random variables with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2. Let a and b be constants. These rules are fundamental and apply to **all** random variables, discrete or continuous.

The expectation of a linear combination is the linear combination of the expectations. This property is called **Linearity of Expectation**.

- General Rule: $E[aX + bY] = aE[X] + bE[Y]$

Key Insight: This rule holds **regardless** of whether X and Y are independent. Expectations are always additive/subtractive in this straightforward way.

Simple Cases:

- Sum: $E[X + Y] = E[X] + E[Y]$
- Difference: $E[X - Y] = E[X] - E[Y]$

2. Variance

The rule for variance includes a covariance term.

General Rule: $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\,\text{Cov}(X, Y)$

Special Case: Independent Variables

If X and Y are **independent**, then their covariance is zero ($\text{Cov}(X, Y) = 0$). The formula simplifies significantly:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

- **Simple Cases (for Independent Variables):**
  - Sum: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
  - Difference: $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$

3. Covariance

Covariance measures the joint variability of two random variables.

- **Rule for Sums:** Covariance is bilinear, meaning it acts like the distributive property in algebra.

$$\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$$

- **Covariance with Itself is Variance:**

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X)$$

This bilinearity is precisely how the general variance rule $\mathrm{Var}(X + Y) = \mathrm{Cov}(X + Y, X + Y)$ is derived. (Why?)

Summary Table

Property	Linear Combination	General Case	Independent Case ($Cov(X, Y) = 0$)
Expectation	$E[aX \pm bY]$	$a\mu_X \pm b\mu_Y$	(Same as General)
	$E[X + Y]$	$\mu_X + \mu_Y$	(Same as General)
	$E[X - Y]$	$\mu_X - \mu_Y$	(Same as General)
Variance	$\text{Var}(aX + bY)$	$a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\text{Cov}(X, Y)$	$a^2\sigma_X^2 + b^2\sigma_Y^2$
	$\text{Var}(X + Y)$	$\sigma_X^2 + \sigma_Y^2 + 2\,\text{Cov}(X, Y)$	$\sigma_X^2 + \sigma_Y^2$
	$\text{Var}(X - Y)$	$\sigma_X^2 + \sigma_Y^2 - 2\,\text{Cov}(X, Y)$	$\sigma_X^2 + \sigma_Y^2$

8. Introduction to Inference and Sampling

Population vs. Sample

Population: The entire group of individuals, objects, or data points that we are interested in studying.

- *Example:* All households in the United States.
- *Example:* All firms listed on the New York Stock Exchange.

Sample: A subset of the population from which we actually collect data.

- *Example:* A survey of 2,000 U.S. households.
- *Example:* The stock prices of 50 firms from the NYSE.

Why sample? It's often impossible or too expensive to collect data on the entire population. We use samples to make inferences about the population.

Parameters vs. Statistics

Parameter: A numerical characteristic of a **population**. These are typically unknown and what we want to estimate. They are considered fixed values.

- **Examples:** Population mean (μ), population variance (σ^2), population correlation (ρ).

Statistic: A numerical characteristic of a **sample**. We calculate statistics from our data. **A statistic is a random variable**, as its value depends on the particular sample drawn.

- **Examples:** Sample mean ($\bar{x}$), sample variance (s^2), sample correlation (r).

The core idea of inference is to use a **sample statistic** to learn about a **population parameter**.

Simple Random Sampling

Simple Random Sampling is the most basic sampling method.

Definition: A sample of size n where every possible sample of that size has an equal chance of being selected, and every individual in the population has an equal chance of being included.

Importance: SRS is the ideal. Statistical methods (like the ones we're learning) are built on the assumption of random sampling. If a sample is not drawn randomly, our inferences may be biased and incorrect.

9. The Concept of a Sampling Distribution

The Distribution of a Statistic

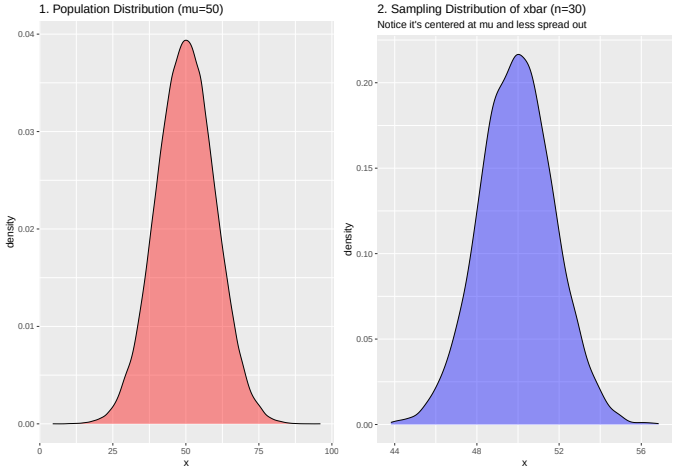
This is a crucial, but sometimes tricky, concept.

Imagine this thought experiment:

1. There is a population with an unknown mean μ.
2. We take a random sample of size n (e.g., n=100) and calculate its sample mean, $\bar{x}_1$.
3. We take a *different* random sample of size n and get a different sample mean, $\bar{x}_2$.
4. We repeat this process 10,000 times, getting $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_{10000}$.

The **Sampling Distribution** of the sample mean is the probability distribution of all these possible $\bar{x}$ values. It's the distribution of a statistic, not of the original data.

Sampling Distribution Visualization



Sampling Distribution Example

Imagine a tiny population that contains only four numbers. These are all the values that exist in our entire population.

```
## [1] "The population is: 2, 4, 6, 10"
```

```
## [1] "The true population mean (mu) is: 5.5"
```

The true mean of our population is **5.5**. In a real research problem, this is the value we want to estimate, but we don't know it.

Step 2: List All Possible Samples

Now, let's list every single possible sample of size $n = 2$ that we can draw from this population *without replacement*.

The number of combinations is "4 choose 2", which is 6. We can use R to list them all.

```
# Use the combn() function to get all unique combinations of size 2
all_possible_samples <- combn(population, 2)
print("All 6 possible samples of size n=2:")
```

```
## [1] "All 6 possible samples of size n=2:"
```

```
print(all_possible_samples)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    2    2    2    4    4    6
## [2,]    4    6   10    6   10   10
```

Step 3: Calculate the Sample Mean for Each Sample

For each of the 6 possible samples, we will now calculate its sample mean ($\bar{x}$).

```
## [1] "The mean of each of the 6 possible samples:"
```

```
## [1] 3 4 6 5 7 8
```

Step 4: The Sampling Distribution of the Sample Mean

The list of all possible sample means we just calculated (3, 4, 6, 5, 7, 8) **is the sampling distribution**. It's the distribution of all possible values the sample mean can take for a sample of size $n = 2$ from our population.

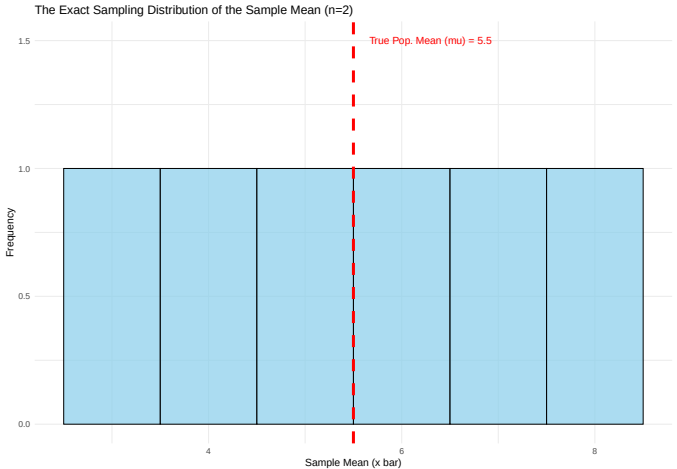
Let's organize it into a frequency table and visualize it.

```
## [1] "The Sampling Distribution (as a frequency table):"
```

```
##   Sample_Mean Frequency
## 1           3         1
## 2           4         1
## 3           5         1
## 4           6         1
## 5           7         1
## 6           8         1
```

Step 5: Visualization the Sampling Distribution of the Sample Mean

Now, let's plot this distribution with a histogram.



10. The Central Limit Theorem (CLT)

The Most Important Theorem in Statistics

The **Central Limit Theorem (CLT)** states:

If you take a sufficiently large random sample ($n \geq 30$ is a common rule of thumb) from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ will be approximately normally distributed, *regardless of the original population's distribution*.

Furthermore, the mean of this sampling distribution will be the population mean μ, and its standard deviation (called the **standard error**) will be $\sigma/\sqrt{n}$.

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Why Is The CLT Important?

The implications of the CLT are profound:

1. **We can use the normal distribution for inference on the mean even if the underlying data is not normal.** Many economic variables (like income) are highly skewed, but the CLT lets us work with their sample means.
2. **It provides a precise formula for the variance of the sample mean (σ^2/n).** This shows that as our sample size n increases, the sample mean $\bar{x}$ becomes a more precise estimator of the population mean μ (its sampling distribution gets narrower).

The CLT is the foundation that allows us to build confidence intervals and conduct hypothesis tests for the mean. And also for estimators that are functions of the mean.

11. Introduction to Estimation

Point Estimators

An **estimator** is a rule (a formula) for calculating an estimate of a population parameter based on sample data. The value it produces is called an **estimate**.

A **Point Estimator** is a formula that produces a single value as the estimate.

Common Point Estimators:

- The sample mean $\bar{x}$ is a point estimator for the population mean μ.
- The sample proportion $\hat{p}$ is a point estimator for the population proportion p.
- The sample variance s^2 is a point estimator for the population variance σ^2.

Desirable Properties of Estimators

How do we know if an estimator is "good"? We look for three properties (conceptually):

Unbiasedness: An estimator is unbiased if its expected value is equal to the true population parameter. $E[\theta] = \theta$.

- *Analogy:* On average, the shots hit the center of the target. There's no systematic over- or under-estimation.

Efficiency: Among all unbiased estimators, the most efficient one is the one with the smallest variance.

- *Analogy:* The shots are tightly clustered around the center. It's a precise estimator.

Consistency: An estimator is consistent if, as the sample size n approaches infinity, the value of the estimator converges to the true parameter value.

- *Analogy:* The more information you get, the closer you get to the truth.

12. Introduction to Hypothesis Testing

The Logic of a Statistical Test

Hypothesis Testing is a formal procedure for checking whether our sample data provides convincing evidence against a preconceived claim about the population.

The Logic: Proof by Contradiction

1. We start by assuming something is true about the population (the **Null Hypothesis**).
2. We then look at our sample data.
3. We ask: "If the null hypothesis were true, how likely is it that we would get sample data like this?"
4. If our sample data is very unlikely under the null hypothesis, we reject our initial assumption in favor of an alternative.

Null and Alternative Hypotheses

Every hypothesis test has two competing hypotheses:

Null Hypothesis (H_0): The claim being tested. It's the "status quo" or "no effect" hypothesis. It always contains an equality sign ($=$, $\leq$, or $\geq$).

- *Example:* The new drug has no effect on blood pressure ($\mu_{change} = 0$).
- *Example:* The mean income in a region is $50,000$ ($\mu = 50000$).

Alternative Hypothesis (H_A or H_1): The claim we are trying to find evidence *for*. It's what we conclude if we reject the null hypothesis. It never contains an equality sign ($\neq$, $<$, or $>$).

- *Example:* The new drug does have an effect ($\mu_{change} \neq 0$).
- *Example:* The mean income is not $50,000$ ($\mu \neq 50000$).

Test Statistics and P-Values (Conceptual)

How do we decide whether our data is "unlikely"?

Test Statistic: A value calculated from sample data that measures how far our sample statistic (e.g., $\bar{x}$) is from the parameter value claimed by the null hypothesis (μ_0). It's often standardized, like a Z-score.

$$\text{Test Statistic} = \frac{\text{Sample Statistic} - \text{Null Hypothesis Value}}{\text{Standard Error}}$$

P-Value: The probability of observing a test statistic as extreme (or more extreme) than the one calculated, *assuming the null hypothesis is true*.

- **Small P-Value (e.g., < 0.05):** The observed data is very unlikely if H_0 were true. We **reject** H_0. The evidence supports H_A.
- **Large P-Value (e.g., > 0.05):** The observed data is plausible if H_0 were true. We **fail to reject** H_0. There is not enough evidence to support H_A.

A Crucial Ingredient: The Standard Error

How do we measure if a sample result is "surprising"? We use the **Standard Error (SE)**.

Definition: The Standard Error of a statistic (like the sample mean) is the standard deviation of its sampling distribution. In simpler terms, it measures the typical or average distance between the sample statistic and the true population parameter.

Its Role: The SE tells us how much we expect a sample mean ($\bar{x}$) to naturally vary from the true population mean (μ).

- A small SE means our sample means will be tightly clustered around the true mean.
- A large SE means they will be more spread out.

One-Sided vs. Two-Sided Hypotheses

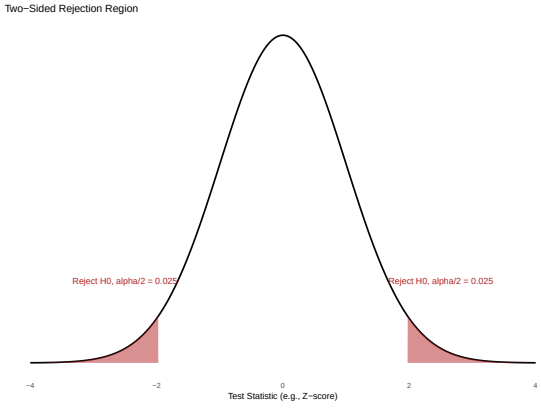
Two-Sided Hypotheses tests are the most common type of test.

The Question: Is the population parameter **different from** a specific value? (We don't care if it's higher or lower, just that it's not the same).

- **Null Hypothesis:** $H_0 : \mu = \mu_0$
- **Alternative Hypothesis:** $H_A : \mu \neq \mu_0$

Rejection Region Two-Sided Test

We are looking for an extreme result in **either direction**. The significance level (α, e.g., 0.05) is split between the two tails of the distribution.



One-Sided Hypothesis Testing (Right-Tailed)

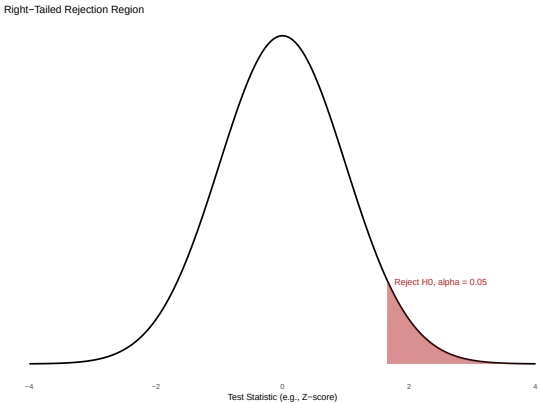
The Question: Is the population parameter **greater than** a specific value?

This is used when we have a strong reason to believe the effect can only go in one direction, or we are only interested in an effect in one direction.

- **Null Hypothesis:** $H_0 : \mu \leq \mu_0$
- **Alternative Hypothesis:** $H_A : \mu > \mu_0$

Rejection Region One-Sided Test (Right-Tailed)

The entire significance level (α) is placed in the **upper (right) tail**.



Example: Testing if a new fertilizer *increases* crop yield. We don't care if it decreases it.

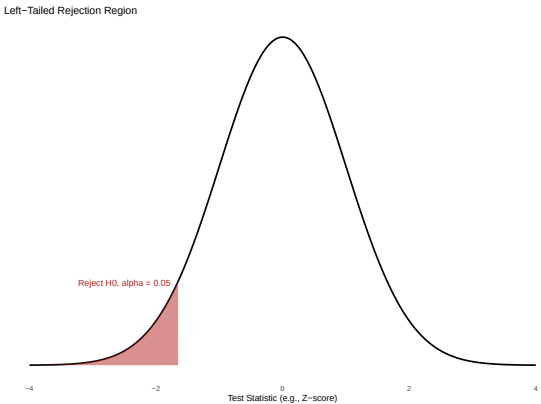
One-Sided Hypothesis Testing (Left-Tailed)

The Question: Is the population parameter **less than** a specific value?

- **Null Hypothesis:** $H_0 : \mu \geq \mu_0$
- **Alternative Hypothesis:** $H_A : \mu < \mu_0$

Rejection Region One-Sided Test (Left-Tailed)

The entire significance level (α) is placed in the **lower (left) tail**.



Example: Testing if a new manufacturing process *reduces* the number of product defects.

Which Test to Use?

Golden Rule: Unless you have a very strong, justifiable, pre-specified reason for expecting an effect in only one direction, **you should use a two-sided test.**

Feature	Two-Sided Test	One-Sided Test
Key Question	Is there a **difference**?	Is it **greater than** or **less than**?
Alternative (H_A)	$\mu \neq \mu_0$	$\mu > \mu_0$ **or** $\mu < \mu_0$
Rejection Region	Split into two tails	All in one tail (left or right)
When to Use	The default, conservative choice.	Only when there is a strong prior reason or you only care about one direction.
Power	Less powerful.	More powerful (if the effect is in the hypothesized direction).

Confidence Intervals

A confidence interval (CI) is a range of values, derived from sample data, that is likely to contain the value of an unknown **population parameter** (e.g., the true population mean μ or proportion p).

Ingredients:

- **Point Estimate:** A single value calculated from the sample that estimates the population parameter (e.g., sample mean $\bar{x}$). It's our "best guess," but it's almost certainly wrong.
- **Interval Estimate:** The confidence interval provides a range around the point estimate, acknowledging the uncertainty inherent in sampling.
- **Confidence Level:** The probability that the *method* used to construct the interval will capture the true population parameter. Common levels are 90%, 95%, and 99%.

Interpretation

The confidence level refers to the **long-run success rate of the method**, not the probability of a single interval being correct.

- **Correct Interpretation:** "We are **95% confident** that the method used to construct this interval from our sample captures the true population mean."
  - *Analogy:* Imagine throwing rings at a post (the true parameter). The 95% confidence level means that if you were to take many samples and throw many "ring" intervals, 95% of them would land on the post. You don't know if the *one* ring you just threw is a success or a miss.
- **Incorrect Interpretation:** It is **wrong** to say, "There is a 95% probability that the true population mean lies within *this specific* interval $[A, B]$." Once an interval is calculated, the true mean is either in it or it isn't; there is no probability involved for that specific interval.

Construction of a Confidence Interval

Most confidence intervals share a common structure.

General Formula:

$$CI = \text{Point Estimate} \pm \text{Margin of Error}$$

The Margin of Error (ME) quantifies the uncertainty of our estimate and is built from two pieces:

1. **Critical Value:**

- A number from a probability distribution (typically a **Z** or **t** distribution).
- It determines how many standard errors to go out from the point estimate to achieve the desired confidence level.
- For a 95% CI, the Z-critical value is $Z_{\alpha/2} = 1.96$. For a t-distribution, it also depends on the sample size (degrees of freedom).

Margin of Error

2. **Standard Error of the Estimate (SE):**

- An estimate of the standard deviation of the sampling distribution of the point estimate.
- It measures the typical amount of variability we expect in our point estimate from sample to sample.
- Example for a mean: $SE = s/\sqrt{n}$ (where s is the sample standard deviation and n is the sample size). (Why? See Variance Rules)

Common Examples

- **CI for a Population Mean (μ)** (when population σ is unknown, which is most common):

$$\bar{x} \pm t_{\alpha/2,df} \left(\frac{s}{\sqrt{n}} \right)$$

Where $\bar{x}$ is the sample mean, s is the sample standard deviation, and t is the critical value from the t-distribution with $n-1$ degrees of freedom.

- **CI for a Population Proportion (p):**

$$\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where $\hat{p}$ is the sample proportion and Z is the critical value from the standard normal distribution.