# ECON 3130 Notes

Jinzhou Wu

December 18, 2025

## Contents

# 1 Probabilities and Events

**Date:** Aug 27, 2025

## 1.1 Definitions

**Definition** (Experiment)**.** A process that results in uncertain outcomes.

**Definition** (Outcome)**.** A possible result of the experiment.

**Definition** (Sample Space)**.** The set of all possible basic outcomes.

**Definition** (Event)**.** A set of basic outcomes, i.e., a subset of the sample space.

**Definition** (Probability)**.** A numerical measure of the chance that an event will occur.

## 1.2 Probability Distributions

**Definition** (Probability Distribution)**.** A table assigning probabilities to all basic outcomes of an experiment, where the probabilities sum to 1.

**Example: Preferred Pet**

| Pet | Pr(Pet) |
|---|---|
| Dog | 0.44 |
| Cat | 0.30 |
| Other | 0.07 |
| None | 0.19 |
| **Total** | 1.00 |

**Probability Tables** When analyzing multiple attributes (e.g., gender and pet preference), we summarize data in a **probability table**.

**Rules:**

1. Two attributes: one on rows, one on columns.
2. Events are mutually exclusive and exhaustive.
3. Joint probabilities go in cells; marginal probabilities go in row and column totals.

**Example Table: Preferred Pet vs. Gender**

| | Dog | Cat | Other | None | Pr(Gender) |
|---|---|---|---|---|---|
| Female | 0.18 | 0.12 | 0.04 | 0.05 | 0.38 |
| Male | 0.32 | 0.15 | 0.05 | 0.09 | 0.62 |
| Refuse/Other | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Pr(Pet) | 0.50 | 0.27 | 0.09 | 0.14 | 1.00 |

## 1.3 Set Theory

Events can be represented as sets within a sample space.

- $A \cap B$ = Event that both $A$ and $B$ occur.
- $A \cup B$ = Event that $A$ or $B$ (or both) occur.
- $A^c$ = Complement of $A$, i.e., $A$ does not occur.

**Rules for operations on sets:**

- **Commutativity:**
$$A \cup B = B \cup A, \quad A \cap B = B \cap A$$

- **Associativity:**
$$A \cup (B \cup C) = (A \cup B) \cup C, \quad A \cap (B \cap C) = (A \cap B) \cap C$$

- **Distributive laws:**
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- **DeMorgan's Laws:**
$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

## 1.4 Probability Rules

**Rule 1: Intersection**
$$\Pr(A \cap B) = \Pr(A \mid B) \Pr(B)$$

**Rule 2: Union**
$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

**Rule 3: Complement**
$$\Pr(A^c) = 1 - \Pr(A)$$

## 1.5 Conditional Probability

**Definition** (Conditional Probability)**.** The probability of event $A$ given event $B$:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \quad \Pr(B) > 0$$

**Multiplication Rule**
$$\Pr(A \cap B) = \Pr(A \mid B) \Pr(B)$$

# 2 Discrete Random Variables

**Date:** Sep 3, 2025

## 2.1 Independence

**Definition** (Independence)**.** Two events $A$ and $B$ are **independent** if and only if

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

**Theorem** (Independence and Joint Events)**.** *If $A$ and $B$ are independent, then*

$$Prob(A|B) = Prob(A)$$

*The conditional probability = the marginal probability (of the unconditioned event).*

## 2.2 Probability Trees

A way of deriving probabilities from conditional information. They are most useful when one has information that can be arranged sequentially.

Probability Tree Rules:

- The tree is drawn on its side.
- It starts from a circle called a chance node.
- The branches of the tree at each node correspond to possible outcomes. We label the branches with conditional probabilities.
- The probabilities at the end of the tree are joint probabilities obtained by multiplication.

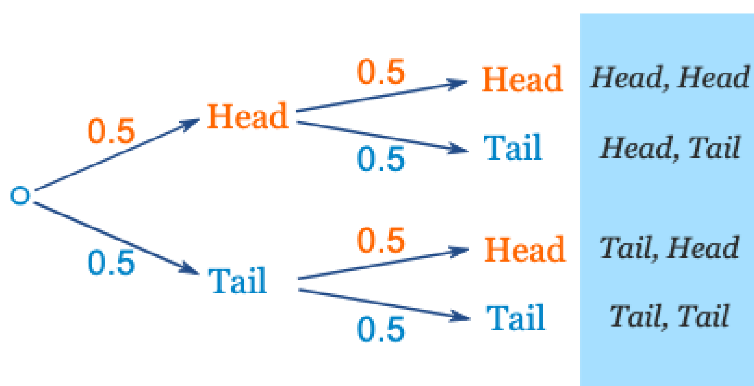



Figure 1: Example of a probability tree.

## 2.3 The Monty Hall Problem

A classic probability puzzle based on a game show:

- Three doors: behind one is a **car**, behind two are **goats**.

- You choose one door (say Door 1).
- The host, Monty Hall, who knows where the car is, opens another door showing a goat.
- You are given the choice to **stay** or **switch**.

Should you switch?

**Solution** Let $C_i$ = event that the car is behind door $i$. Let $H$ = event that Monty opens **Door 3**. Initially:

$$\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \tfrac{1}{3}.$$

Assume you pick Door 1 and Monty opens Door 3 revealing a goat. By Bayes' theorem:

$$\Pr(C_i \mid H) = \frac{\Pr(H \mid C_i)\Pr(C_i)}{\Pr(H)}.$$

Monty's behavior:

$$\Pr(H \mid C_1) = \tfrac{1}{2}, \qquad \Pr(H \mid C_2) = 1, \qquad \Pr(H \mid C_3) = 0.$$

Thus, the total probability of $H$:

$$\Pr(H) = \Pr(H \mid C_1)\Pr(C_1) + \Pr(H \mid C_2)\Pr(C_2) + \Pr(H \mid C_3)\Pr(C_3),$$

$$\Pr(H) = \tfrac{1}{2} \cdot \tfrac{1}{3} + 1 \cdot \tfrac{1}{3} + 0 = \tfrac{1}{2}.$$

Finally, the posterior probabilities:

$$\Pr(C_1 \mid H) = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \tfrac{1}{3}, \qquad \Pr(C_2 \mid H) = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \tfrac{2}{3}.$$

The probability that your original choice is correct is $\frac{1}{3}$, while switching gives $\frac{2}{3}$. **Always switch.**

## 2.4 Random Variables

**Definition** (Random Variable)**.** An experiment with numerical outcomes (or outcomes that can be mapped to numbers).

**Notation** We denote a random variable with a capital letter, like $X$ or $Y$.

**Probability Mass Function (PMF)**

$$f(x) = Pr(X = x)$$

Every discrete random variable has a PMF that describes the probability of each possible value.

## 2.5 Measures of Central Tendency

**Expected Value**

$$E[X] = \sum_{x \in S} x \cdot f(x)$$

**Mode**

$$Mode[X] = \text{value (or values) that maximize } f(x)$$

**Median**

$$Median[X] = \text{value } m \text{ such that } Pr(X \leq m) = Pr(X \geq m) = 0.5$$

## 2.6 Measures of Dispersion

**Variance**

$$Var[X] = E[(X - E[X])^2]$$
$$Var[X] = \sum_{x \in S} (x - E[X])^2 \cdot f(x)$$

A Variance Identity

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - E[2XE[X]] + E[(E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= \boxed{E[X^2] - (E[X])^2} \end{aligned}$$

Variance of a Linear Function of X

$$\begin{aligned} Var[aX + b] &= E[(aX + b - E[aX + b])^2] \\ &= E[(aX + b - aE[X] - b)^2] \\ &= E[(aX - aE[X])^2] \\ &= E[a^2(X - E[X])^2] \\ &= a^2E[(X - E[X])^2] \\ &= \boxed{a^2Var[X]} \end{aligned}$$

**Standard Deviation**

$$\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$$

## 2.7 Conditional Expectations

**Conditional Expectations.** If $X$ and $Y$ are random variables, the *conditional expectation* of $X$ given that $Y$ takes on a certain value $y$ is:

$$E[X \mid Y = y] = \sum_x x \cdot \Pr(X = x \mid Y = y)$$

- This is the same formula as the one for expected value, except that we have replaced the marginal probability with a conditional probability.
- The conditional expectation of $X$ given $Y = y$ is the probability-weighted average of $X$ given $Y = y$.
- $E[X \mid Y = y]$ is a function that operates on values of $Y$.

## 2.8 Law of Iterated Expectations

**Theorem** (Law of Iterated Expectations)**.**

$$E[E[X \mid Y]] = E[X]$$

*The expected value of the conditional expectation of $X$ given $Y$ is equal to the expected value of $X$.*

- This is also called the **law of total expectation**.
- It is useful when you know the distribution of $X$ conditional on $Y$, but not the marginal distribution of $X$.
- You can compute $E[X]$ by first computing $E[X \mid Y]$ for each value of $Y$, then taking the expectation over $Y$.

$$E[X] = \sum_y E[X \mid Y = y] \cdot \Pr(Y = y) = \sum_y \sum_x x \cdot \Pr(X = x \mid Y = y) \cdot \Pr(Y = y)$$

# 3 More Discrete Random Variables

**Date:** Sep 8, 2025

## 3.1 Moments

**Definition** (Moment)**.** Numbers that represent qualities of a particular distribution. Expected values of a random variables raised to a different powers. $r$th moment centered around $b$ of a random variable $X$ is defined as

$$\mu_r = E[(X-b)^r]$$

The mean is the first moment centered around zero. The variance is the second moment centered around E[X].

## 3.2 Moment-generating function (MGF)

**Definition** (Moment-generating function)**.** The moment-generating function (MGF) of a random variable $X$ is defined as

$$M_X(t) = E[e^{tX}] = \sum_x e^{tx} f(x) \quad \text{(pmf)}$$

for all $t$ in an open interval containing 0 such that the expectation exists.

The mgf is not always defined for a random variable. If two random variables have the same mgf, then they also have the same pmf.

**Finding moments using MGF** The $r$th moment of a random variable $X$ can be found by taking the $r$th derivative of the mgf and evaluating it at $t = 0$:

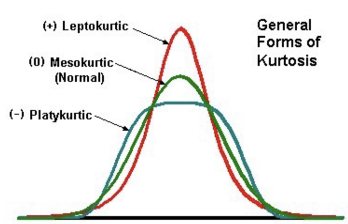
$$\mu_r = E[X^r] = M_X^{(r)}(0) = \sum_x x^r f(x)$$

## 3.3 Standardized moments

$$\hat{\mu}_k = \frac{E[(X-\mu)^k]}{E[(X-\mu)^2]^{k/2}}$$

kth centered moment divided by the standard deviation raised to the k. Standardized moments are unit-invariant.

Skew is standardized third moment. Negative skew = left tail. Positive skew = right tail. Kurtosis is standardized fourth moment.



Kurtosis measures how fat the tails are. Positive means more extreme values. Negative means fewer extreme values. Note that all three of these random variables have the same variance.

### 3.4 Bernoulli distribution

**Definition** (Beroulli distribution)**.** A Bernoulli random variable $X$ takes the value 1 with probability $p$ and the value 0 with probability $1-p$.

$$P(X=1)=p, \quad P(X=0)=1-p$$

The mean of a Bernoulli random variable is $E[X]=p$. The variance of a Bernoulli random variable is $Var(X)=p(1-p)$.

### 3.5 Binomial distribution

**Definition** (Binomial distribution)**.** A Binomial random variable $X$ with parameters $n$ and $p$ counts the number of successes in $n$ independent Bernoulli trials, each with success probability $p$.

The mean of a Binomial random variable is $E[X]=np$. The variance of a Binomial random variable is $Var(X)=np(1-p)$ (by the linearity of expectation).

- There are a fixed number of independent trials.
- Each trial has two basic outcomes (e.g., "success"/"failure").
- The probability of a success, p, is constant across trials.

**Normal approximation to the Binomial** When the number of trials $n$ is large, the Binomial distribution can be approximated by a Normal distribution with the same mean and variance:

$$X \sim \text{Binomial}(n,p) \approx N(np, np(1-p))$$

This approximation is most accurate when $np \geq 10$ and $n(1-p) \geq 10$.

**Binomial probability mass function (pmf)** The probability mass function (pmf) for a Binomial random variable $X$ is:

$$\Pr(X=k \mid n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \binom{n}{k}p^k(1-p)^{n-k}$$

where:

- $n$ = number of trials (maximum possible number of successes)
- $k$ = number of successes
- $n-k$ = number of failures
- $p$ = probability of a success on each trial
- $1-p$ = probability of a failure on each trial

$X$ can also be defined as the sum of $n$ independent Bernoulli random variables, each with parameter $p$:

$$X=\sum_{i=1}^{n} X_i, \quad X_i \sim \text{Bernoulli}(p)$$

# 4 Continuous Random Variables

**Date:** Sep 10, 2025

**Definition** (Continuous Random Variable)**.** A random variable $X$ is said to be continuous if there exists a function $f(x)$ such that for any set $A$,

$$P(X \in A) = \int_A f(x)\,dx$$

where $f(x)$ is called the probability density function (pdf) of $X$. Key properties of a legal pdf: $f(x) \geq 0 \ \forall x \in S$ and $\int_S f(x)dx = 1$.

## 4.1 Cumulative Distribution Function (CDF)

$$F(a) = \Pr(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

## 4.2 Uniform Random Variable

If $X \sim \text{Uniform}(a, b)$, then every value between $a$ and $b$ is equally likely. The probability density function (pdf) is:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function (CDF) is:

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b \end{cases}$$

The probability that $X$ falls between $a$ and $b$ is:

$$\Pr(a < X < b) = \int_a^b \frac{1}{b-a} dx = 1$$

The expected value and variance of a uniform random variable are:

$$E[X] = \frac{a+b}{2}, \quad \text{Var}[X] = \frac{(b-a)^2}{12}$$

## 4.3 Expected Value

The expected value (mean) of a continuous random variable $X$ with pdf $f(x)$ is defined as:

$$E[X] = \int_{-\infty}^{+\infty} x f(x)\,dx$$

## 4.4 Variance

The variance of a continuous random variable $X$ with pdf $f(x)$ is defined as:

$$\text{Var}[X] = \int_{-\infty}^{+\infty} (x - E[X])^2 f(x)\,dx$$

## 4.5 Moment-Generating Function (MGF)

The moment-generating function (MGF) of a continuous random variable $X$ with pdf $f(x)$ is defined as:

$$M(t) = \int_{-\infty}^{+\infty} e^{tx} f(x)\, dx, \qquad -h < t < h$$

where $h$ is a constant such that $M(t)$ is finite for $t$ in $(-h, h)$.

For discrete random variables, the MGF is:

$$M(t) = \sum_{x \in S} e^{tx} f(x)$$

The MGF characterizes the distribution if it is finite in some neighborhood around $t$.

The $r$-th moment of $X$ can be obtained by differentiating $M(t)$ $r$ times and evaluating at $t = 0$:

$$M^{(r)}(0) = E[X^r]$$

## 4.6 Quantiles

The $p$'th quantile $\pi_p$ is defined as the value such that

$$p = \int_{-\infty}^{\pi_p} f(x)\, dx$$

# 5 Normal Random Variables

**Date:** Sep 15, 2025

## 5.1 Definition and Properties

**Definition** (Normal Random Variable)**.** A normal random variable is a continuous random variable whose probability density function (pdf) is given by:

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

where $\mu$ is the mean and $\sigma^2$ is the variance ($\sigma$ is the standard deviation).

- The normal distribution is symmetric about its mean $\mu$.
- The total area under the pdf is 1.
- The standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$.
- Probabilities are computed as areas under the curve:

$$\Pr(a < X < b) = \int_a^b f(x)dx$$

- $\mathbb{E}(X) = \mu$
- $\mathrm{Var}(X) = \sigma^2$ (and thus $\mathrm{SD}(X) = \sigma$)
- 68%, 95%, 99.7% of outcomes are within 1, 2, 3 standard deviations of the mean, respectively.

A normal random variable $X$ is denoted by:

$$X \sim N(\mu, \sigma^2)$$

which we read as "$X$ is normally distributed with mean $\mu$ and variance $\sigma^2$."

If $\mu = 0$ and $\sigma = 1$, then the random variable is a **standard normal**, denoted $Z$.

## 5.2 Moment Generating Function (mgf) of the Normal

In general, the moment generating function (mgf) of a random variable $X$ is:

$$M(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx} f(x)dx$$

For the normal distribution:

$$M(t) = \int_{-\infty}^{+\infty} \frac{e^{tx}}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

This simplifies to:

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

The mgf can be used to compute all moments of the normal distribution by differentiating $M(t)$ with respect to $t$ and evaluating at $t = 0$.

### 5.3 Linear Combinations of Independent Normals

1. **Sums.** The sum of independent normally distributed random variables is also normally distributed.
   *Example:* If $X$ and $Y$ are independent and $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, then $W = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2. **Constants.** If $Z \sim N(0, 1)$, then for constants $a$ and $b$, $a + bZ \sim N(a, b^2)$. In general, multiplying a normal random variable by a constant and/or adding a constant yields another normal random variable.

3. **Standardizing.** If $X \sim N(\mu, \sigma^2)$, then the standardized variable $Z = \frac{X-\mu}{\sigma}$ follows a standard normal distribution: $Z \sim N(0, 1)$.

# 6 More Random Variables

**Date:** Sep 17, 2025

## 6.1 Negative Binomial Random Variables

**Definition** (Negative Binomial Random Variable)**.** A random variable $X$ is said to have a **negative binomial distribution** with parameters $r$ and $p$, denoted $X \sim \text{NegBin}(r, p)$, if it represents the number of independent Bernoulli trials needed to achieve **exactly $r$ successes**, where each trial has two possible outcomes (success or failure) and the probability of success on each trial is constant and equal to $p$.

The probability mass function is given by:

$$f(x; r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

for $x = r, r+1, r+2, \ldots$

- $f(x; r, p)$: The probability that the $r$-th success occurs on the $x$-th trial.
- $\binom{x-1}{r-1}$: The number of ways to arrange $r-1$ successes in the first $x-1$ trials.
- $p^r$: Probability of $r$ successes.
- $(1-p)^{x-r}$: Probability of $x-r$ failures.

The mean and variance of a negative binomial random variable $X \sim \text{NegBin}(r, p)$ are:

$$E[X] = \frac{r}{p}$$

$$\text{Var}[X] = \frac{r(1-p)}{p^2}$$

**Derivation:**

Let $X$ be the number of trials needed to get $r$ successes. $X$ can be written as the sum of $r$ independent geometric random variables $Y_i$ (each representing the number of trials needed to get the $i$-th success after the $(i-1)$-th success):

$$X = Y_1 + Y_2 + \cdots + Y_r$$

Each $Y_i$ is $\text{Geom}(p)$, so $E[Y_i] = 1/p$ and $\text{Var}[Y_i] = (1-p)/p^2$. By linearity of expectation and independence:

$$E[X] = r \cdot E[Y_1] = \frac{r}{p}$$

$$\text{Var}[X] = r \cdot \text{Var}[Y_1] = \frac{r(1-p)}{p^2}$$

## 6.2 Geometric Random Variables

A **geometric random variable** is a special case of the negative binomial distribution where $r = 1$. That is, it counts the number of trials needed to get the first success.

The probability mass function is:

$$f(x;p) = \binom{x-1}{1-1} p^1 (1-p)^{x-1} = p(1-p)^{x-1}$$

for $x = 1, 2, 3, \ldots$

The mean and variance are:

$$E[X] = \frac{1}{p}$$

$$\mathrm{Var}[X] = \frac{1-p}{p^2}$$

**Derivation:**

For $X \sim \mathrm{Geom}(p)$,

$$E[X] = \sum_{x=1}^{\infty} xp(1-p)^{x-1}$$

This is a standard result and can be shown using the formula for the expectation of a geometric series:

$$E[X] = \frac{1}{p}$$

Similarly, for the variance:

$$\mathrm{Var}[X] = \frac{1-p}{p^2}$$

## 6.3 Poisson Random Variables

A **Poisson random variable** is defined by a single parameter $\lambda$ that represents the rate of occurrences.

**Definition** (Poisson Random Variable)**.** A random variable $X$ is said to have a **Poisson distribution** with parameter $\lambda > 0$, denoted $X \sim \mathrm{Poisson}(\lambda)$, if it represents the number of occurrences of an event in a fixed interval of time or space, where occurrences happen independently and at a constant average rate $\lambda$.

The probability mass function is:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

for $k = 0, 1, 2, \ldots$

**Derivation from Binomial:**

Suppose we divide a unit interval into $n$ subintervals, each of length $1/n$. In each subinterval, the probability of an occurrence is $\lambda/n$, so the total expected number of occurrences in the unit interval is $\lambda$.

Let $Y \sim \mathrm{Bin}(n, \lambda/n)$ be the number of occurrences in $n$ subintervals.

The probability mass function is:

$$P(Y = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

As $n \to \infty$, the binomial distribution converges to the Poisson distribution:

$$P(X = x) = \lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x e^{-\lambda}}{x!}$$

This shows that the Poisson distribution can be viewed as the limiting case of the binomial distribution when the number of trials is large and the probability of success is small, but the expected number of successes remains fixed.

**Conditions for a Poisson process:**

- Numbers of occurrences in nonoverlapping subintervals are independent.
- Probability of exactly one occurrence in a sufficiently short subinterval of length $h$ is approximately $\lambda h$.
- Probability of two or more occurrences in a sufficiently short subinterval is essentially zero.

The mean and variance of a Poisson random variable $X \sim \text{Poisson}(\lambda)$ are:

$$E[X] = \lambda$$

$$\text{Var}[X] = \lambda$$

**Derivation:**

The mean is

$$E[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!}$$

This can be shown to equal $\lambda$ (using properties of the exponential and the fact that $\sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{\lambda}$).

The variance is

$$\text{Var}[X] = E[X^2] - (E[X])^2$$

It can be shown that $E[X^2] = \lambda^2 + \lambda$, so

$$\text{Var}[X] = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

# 7 Even More Random Variables

**Date:** Sep 22, 2025

## 7.1 Exponential Random Variables

An **exponential random variable** $W$ with rate parameter $\lambda > 0$ is a continuous random variable with cumulative distribution function (CDF)

$$F(w) = \begin{cases} 0 & \text{if } w < 0 \\ 1 - e^{-\lambda w} & \text{if } w \geq 0 \end{cases}$$

and probability density function (PDF)

$$f(w) = \begin{cases} \lambda e^{-\lambda w} & \text{if } w \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The expected value and variance of $W$ are given by

$$\mathbb{E}[W] = \frac{1}{\lambda}, \qquad \mathrm{Var}[W] = \frac{1}{\lambda^2}$$

The exponential distribution is commonly used to model the time between independent events that occur at a constant average rate. It is the time until the first event in a Poisson process.

## 7.2 Gamma Random Variables

A **gamma random variable** $X$ with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$ has pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}, \quad 0 \leq x < \infty$$

The gamma distribution generalizes the exponential distribution. If $\alpha = 1$, the gamma distribution reduces to the exponential distribution. The gamma distribution models the waiting time until the $\alpha$-th event in a Poisson process with rate $\lambda = 1/\theta$.

The mean and variance are:

$$\mathbb{E}[X] = \frac{t}{\lambda} = \alpha\theta, \qquad \mathrm{Var}[X] = \frac{t}{\lambda^2} = \alpha\theta^2$$

where $\Gamma(t)$ is the **gamma function**:

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy, \quad t > 0$$

**Derivation of the gamma function:**

We use integration by parts:

$$\int u\, dv = uv - \int v\, du$$

Let $u = y^{t-1}$, $dv = e^{-y}dy$, so $du = (t-1)y^{t-2}dy$, $v = -e^{-y}$.

Applying integration by parts:

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy = \left[-y^{t-1}e^{-y}\right]_0^\infty + \int_0^\infty (t-1) y^{t-2} e^{-y} dy$$

The boundary term vanishes for $t > 1$, so:

$$\Gamma(t) = (t-1)\int_0^\infty y^{t-2}e^{-y}dy = (t-1)\Gamma(t-1)$$

This gives the recurrence relation:

$$\Gamma(t) = (t-1)\Gamma(t-1)$$

For $t = 1$:

$$\Gamma(1) = \int_0^\infty e^{-y}dy = 1$$

**Gamma Random Variables When $\alpha$ is an Integer** If $\alpha$ is a positive integer, the gamma distribution describes the waiting time until the $\alpha$-th event in a Poisson process with rate $\lambda = 1/\theta$.

The cumulative distribution function (cdf) is:

$$F(x) = \Pr(X \leq x) = 1 - \Pr(X > x)$$

This is equivalent to the probability that at least $\alpha$ events have occurred by time $x$:

$$F(x) = 1 - \Pr(\text{fewer than } \alpha \text{ occurrences in } [0, x])$$

For a Poisson process, the probability of $k$ events in $[0, x]$ is:

$$\Pr(k \text{ events in } [0, x]) = \frac{(\lambda x)^k e^{-\lambda x}}{k!}$$

So,

$$F(x) = 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda x)^k e^{-\lambda x}}{k!}$$

Or, equivalently, using $\theta = 1/\lambda$:

$$F(x) = 1 - \sum_{k=0}^{\alpha-1} \frac{(x/\theta)^k e^{-x/\theta}}{k!}$$

This expresses the cdf of the gamma distribution when $\alpha$ is an integer in terms of the Poisson distribution.

# 8 Bivariate Distributions

**Date:** Oct 1, 2025

## 8.1 Linearity of Expectation:

For random variables $A$ and $B$, $\mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B]$.

$$\begin{aligned}
\mathbb{E}[A+B] &= \sum_a \sum_b (a+b) f(a,b) \\
&= \sum_a \sum_b a f(a,b) + \sum_a \sum_b b f(a,b) \\
&= \left( \sum_a \sum_b a f(a,b) \right) + \left( \sum_a \sum_b b f(a,b) \right) \\
&= \sum_a a \left( \sum_b f(a,b) \right) + \sum_b b \left( \sum_a f(a,b) \right) \\
&= \sum_a a f_A(a) + \sum_b b f_B(b) \\
&= \mathbb{E}[A] + \mathbb{E}[B]
\end{aligned}$$

where $f(a,b)$ is the joint probability mass function, and $f_A(a)$ and $f_B(b)$ are the marginal distributions of $A$ and $B$, respectively.

## 8.2 Covariance

**Definition.** For two random variables $X$ and $Y$, the covariance is defined as:

$$\text{Cov}[X,Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

For discrete random variables, this can be written as:

$$\text{Cov}[X,Y] = \sum_x \sum_y (x - \mathbb{E}[X])(y - \mathbb{E}[Y]) \Pr(X = x, Y = y)$$

- $\text{Cov}(X,Y) > 0$: When $X$ increases, $Y$ tends to increase (positive association).
- $\text{Cov}(X,Y) < 0$: When $X$ increases, $Y$ tends to decrease (negative association).
- $\text{Cov}(X,Y) = 0$: No (linear) association between $X$ and $Y$.

Covariance measures the linear association between two random variables.

## 8.3 Variance of a Sum

For any two random variables $X$ and $Y$,

$$\text{Var}(X+Y) = \text{Var}(X) + 2\,\text{Cov}(X,Y) + \text{Var}(Y)$$

**Derivation**

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}\left[(X+Y-\mathbb{E}[X+Y])^2\right] \\
&= \mathbb{E}\left([(X-\mathbb{E}[X])+(Y-\mathbb{E}[Y])]^2\right) \\
&= \mathbb{E}\left[(X-\mathbb{E}[X])^2+2(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])+(Y-\mathbb{E}[Y])^2\right] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2]+2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]+\mathbb{E}[(Y-\mathbb{E}[Y])^2] \\
&= \mathrm{Var}(X)+2\,\mathrm{Cov}(X,Y)+\mathrm{Var}(Y)
\end{aligned}$$

If $X$ and $Y$ are independent, then $\mathrm{Cov}(X,Y)=0$, so

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$$

**Intuition** If two random variables negatively covary, they will tend to "cancel each other out" and reduce the variance of the sum; if they positively covary, they will tend to magnify the variance of the sum.

**Additional Useful Facts about Covariance**

- For constants $a, b, c, d$,
$$\mathrm{Cov}(aX+b,\, cY+d)=ac\,\mathrm{Cov}(X,Y)$$
- Covariance is linear in each argument:
$$\mathrm{Cov}(X,\, Y+Z)=\mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z)$$
- If $X$ and $Y$ are independent, then $\mathrm{Cov}(X,Y)=0$.
- The converse is not true: $\mathrm{Cov}(X,Y)=0$ does **not** imply $X$ and $Y$ are independent.

**Independence Implies Zero Covariance** If $X$ and $Y$ are independent, then their joint pmf factors $f(x,y)=f_X(x)f_Y(y)$. Hence,

$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \sum_x\sum_y(x-\mu_X)(y-\mu_Y)f(x,y) \\
&= \sum_x\sum_y(x-\mu_X)(y-\mu_Y)f_X(x)f_Y(y) \\
&= \sum_x(x-\mu_X)f_X(x)\sum_y(y-\mu_Y)f_Y(y) \\
&= \sum_x(x-\mu_X)f_X(x)\cdot 0 \\
&= 0
\end{aligned}$$

where $\mu_X=\mathbb{E}[X]$ and $\mu_Y=\mathbb{E}[Y]$. Thus, independence implies zero covariance.

# 9 Bivariate Distributions Continued

**Date:** Oct 6, 2025

## 9.1 Correlation Coefficient

- Covariance varies with the measurement scale of the variables.
- The correlation coefficient $\rho$ is scale-invariant.
- $\rho$ ranges from $-1$ (perfect negative linear association) to 1 (perfect positive linear association).
- $\rho = 0$ indicates no linear relationship.

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}$$

## 9.2 Conditional Distributions

- For two discrete random variables $X$ and $Y$, the conditional probability mass function of $X$ given $Y$ is:
$$g(x|y) = \frac{f(x, y)}{f_Y(y)}, \quad \text{provided that } f_Y(y) > 0$$
- $f(x, y)$ is the joint probability mass function of $X$ and $Y$.
- $f_Y(y)$ is the marginal probability mass function of $Y$.
- $g(x|y)$ is not defined for values of $y$ where $f_Y(y) = 0$.

If $X$ and $Y$ are independent, then $g(x|y) = f_X(x)$ for all $y$ such that $f_Y(y) > 0$.

## 9.3 Continuous Bivariate Distributions

- A joint probability density function (pdf) $f(x, y)$ for continuous random variables $X$ and $Y$ must satisfy:
  1. $f(x, y) \geq 0$, and $f(x, y) = 0$ when $(x, y) \notin S$, where $S$ is the sample space.
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$
  3. For any region $A$, $\Pr[(X, Y) \in A] = \iint_A f(x, y)\, dx\, dy$
- The marginal pdfs are obtained by integrating out the other variable:
$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$
- Independence: $X$ and $Y$ are independent if and only if $f(x, y) = f_X(x) f_Y(y)$ for all $(x, y)$.

# 10 Functions of Random Variables

**Date:** Oct 8, 2025

## 10.1 Expectations

$$E[X] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) \, dy \, dx$$

## 10.2 The Bivariate Normal Distribution

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{q(x, y)}{2}\right]$$

where

$$q(x, y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$$

- $\mu_X, \mu_Y$ are the means of $X$ and $Y$
- $\sigma_X, \sigma_Y$ are the standard deviations
- $\rho$ is the correlation coefficient between $X$ and $Y$

## 10.3 Conditional Expectation of the Bivariate Normal

If $X$ and $Y$ are bivariate normal:

- The marginal distributions of $X$ and $Y$ are also normal.
- $X + Y$ and $X$ are jointly normal (as are $X + Y$ and $Y$).
- The conditional distribution $Y|X$ is also normal with:

$$\begin{aligned} E[Y|x] &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) \\ &= \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X + \rho\frac{\sigma_Y}{\sigma_X}x \end{aligned}$$

  Since $\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$, we can substitute to get:

$$E[Y|x] = \mu_Y - \frac{\sigma_{XY}}{\sigma_X^2}\mu_X + \frac{\sigma_{XY}}{\sigma_X^2}x$$

  where $\sigma_{XY}$ is the covariance between $X$ and $Y$.
  And $Var(Y|x) = \sigma_Y^2(1 - \rho^2)$.

- $X$ and $Y$ being Normal does not imply the joint distribution is Bivariate Normal.

## 10.4 Functions of a Random Variable

### 10.4.1 Distribution Function Technique

If $Y = u(X)$, then

$$G(y) = \Pr(Y \leq y) = \Pr(u(X) \leq y)$$

and the density is

$$g(y) = G'(y)$$

Suppose $X \sim N(0,1)$ and $Y = \exp(X)$ (log-normal distribution). Let $\Phi(x)$ be the cdf and $\phi(x)$ the pdf of the standard normal. Then $\Phi'(x) = \phi(x)$.

$$G(y) = \Pr(\exp(X) \leq y) = \Pr(X \leq \ln y) = \Phi(\ln y)$$

$$g(y) = G'(y) = \Phi'(\ln y) \cdot \frac{1}{y} = \phi(\ln y) \cdot \frac{1}{y}$$

for $y > 0$; otherwise, $g(y) = 0$.

### 10.4.2 Change-of-Variable Technique

If $Y = u(X)$, define $v()$ such that $X = v(Y)$.

Suppose $c_1 < x < c_2$ and $d_1 = u(c_1) < y < d_2 = u(c_2)$ ($v()$ is increasing),

$$G(y) = \int_{c_1}^{v(y)} f(x)dx, \qquad d_1 < y < d_2$$

$$G'(y) = g(y) = f(v(y)) \cdot v'(y), \qquad d_1 < y < d_2$$

**Log-normal example:**
$u(x) = \exp(x)$, $v(y) = \ln(y)$, $v'(y) = 1/y$

$$g(y) = \phi(\ln y) \cdot \frac{1}{y} \qquad \text{when } y > 0$$

Suppose $c_1 < x < c_2$ and $d_1 = u(c_1) > y > d_2 = u(c_2)$ ($v()$ is decreasing),

$$G(y) = \int_{v(y)}^{c_2} f(x)dx, \qquad d_2 < y < d_1$$

$$G'(y) = g(y) = f(v(y)) \cdot (-v'(y)), \qquad d_2 < y < d_1$$

Note: The negative sign appears because $v'(y) < 0$ when $v()$ is decreasing.
In general, we have:

$$g(y) = f(v(y)) \cdot |v'(y)|, \qquad y \in S_Y$$

## 11 More Functions of Random Variables

**Date:** Oct 15, 2025

### 11.1 Cauchy Distribution

**Definition** (Cauchy Distribution)**.** Let $W \sim U(-\pi/2, \pi/2)$, i.e., $W$ is uniformly distributed between $-\pi/2$ and $\pi/2$. Define $X = \tan W$. Then $X$ has the standard Cauchy distribution.

The probability density function (pdf) of the standard Cauchy distribution is:

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}$$

**Properties:**

- The Cauchy distribution is symmetric about $x = 0$.
- It has much heavier tails than the normal distribution.
- Neither the mean nor the variance exists.
- The median and mode are both 0.

**Derivation of the PDF:**
Let $W \sim U(-\pi/2, \pi/2)$, so its pdf is

$$f_W(w) = \frac{1}{\pi}, \quad w \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

Define $X = \tan W$. To find the pdf of $X$, use the change of variables formula:

$$g_X(x) = f_W(v(x)) \left|v'(x)\right|$$

where $v(x) = \arctan x$ and $v'(x) = \frac{1}{1+x^2}$.
Substituting, we get:

$$g_X(x) = f_W(\arctan x) \cdot \frac{1}{1+x^2} = \frac{1}{\pi} \cdot \frac{1}{1+x^2} = \frac{1}{\pi(1+x^2)}$$

**Notes:**

1. The graph of the pdf is called the witch of Agnesi.
2. $X$ is the ratio of two independent standard normals.
3. $X$ is a $t$ distribution with one degree of freedom.

### 11.2 Change-of-Variable Technique (2 Variables)

Suppose $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$, and we can invert these to get $X_1 = v_1(Y_1, Y_2)$ and $X_2 = v_2(Y_1, Y_2)$. Then, the joint pdf of $(Y_1, Y_2)$ is given by:

$$g(y_1, y_2) = |J|\, f\big(v_1(y_1, y_2), v_2(y_1, y_2)\big)$$

where $f$ is the joint pdf of $(X_1, X_2)$, and $J$ is the Jacobian determinant:

$$J = \begin{vmatrix} \frac{\partial v_1}{\partial y_1} & \frac{\partial v_1}{\partial y_2} \\ \frac{\partial v_2}{\partial y_1} & \frac{\partial v_2}{\partial y_2} \end{vmatrix}$$

# 12 The Central Limit Theorem

**Date:** Oct 20, 2025

## 12.1 Sample Mean

Suppose we have $n$ random variables $X_i$ that are independent and identically distributed (iid) with mean $\mu$ and standard deviation $\sigma$.

Define the sample mean as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The expected value and variance of the sample mean are:

$$E[\overline{X}] = \mu \qquad Var[\overline{X}] = \frac{\sigma^2}{n}$$

As $n \to \infty$, $Var[\overline{X}] \to 0$.

**Law of Large Numbers (LLN):** As $n$ increases, the sample mean $\overline{X}$ converges to the population mean $\mu$.

**Central Limit Theorem (CLT):** As $n$ increases, the distribution of $\overline{X}$ approaches a normal distribution, regardless of the underlying distribution of $X_i$.

## 12.2 Markov's Inequality

For a **nonnegative** random variable $X$ and any $a > 0$,

$$\Pr(X \geq a) \leq \frac{E[X]}{a}$$

**Proof:**

Let $f(x)$ be the probability mass function of $X$.

$$\begin{aligned}
E[X] &= \sum_{x} x f(x) \\
&= \sum_{x \geq a} x f(x) + \sum_{x < a} x f(x) \\
&\geq \sum_{x \geq a} x f(x) \\
&\geq \sum_{x \geq a} a f(x) \\
&= a \sum_{x \geq a} f(x) \\
&= a \Pr(X \geq a)
\end{aligned}$$

Therefore,

$$\Pr(X \geq a) \leq \frac{E[X]}{a}$$

### 12.3 Chebyshev's Inequality

For any random variable $X$ and any $r > 0$,

$$\Pr(|X - E[X]| \geq r) \leq \frac{Var[X]}{r^2}$$

**Proof:**
Define $A = \{x : |x - E[X]| \geq r\}$.

$$\begin{aligned}
Var[X] &= \sum_x (x - E[X])^2 f(x) \\
&= \sum_{x \in A} (x - E[X])^2 f(x) + \sum_{x \notin A} (x - E[X])^2 f(x) \\
&\geq \sum_{x \in A} (x - E[X])^2 f(x) \\
&\geq \sum_{x \in A} r^2 f(x) \\
&= r^2 \sum_{x \in A} f(x) = r^2 \Pr(A) = r^2 \Pr(|X - E[X]| \geq r)
\end{aligned}$$

Therefore,

$$\frac{Var[X]}{r^2} \geq \Pr(|X - E[X]| \geq r)$$

# 13 Proving the Central Limit Theorem

**Date:** Oct 22, 2025

- The Central Limit Theorem (CLT) states that for a random sample $X_1, X_2, \dots, X_n$ from a distribution with finite mean $\mu$ and variance $\sigma^2$, the standardized sample mean

$$W = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to the standard normal:

$$W \xrightarrow{d} N(0, 1)$$

as $n \to \infty$.

- This means that as the sample size $n$ increases, the distribution of the sample mean $\overline{X}$ approaches a normal distribution with mean $\mu$ and variance $\frac{\sigma^2}{n}$:

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

- The CLT is fundamental in statistics because it justifies the use of normal approximations for inference, even when the underlying distribution is not normal.
- Key conditions: finite mean and variance, and sufficiently large $n$ (typically $n \geq 30$).

## 13.1 The mgf Method

- If the sequence of mgfs $M_n(t)$ of random variables $X_n$ converges to an mgf $M(t)$ for $t$ in an open interval around 0, then the distributions of $X_n$ converge to the distribution with mgf $M(t)$.
- In the context of the Central Limit Theorem, we show that the mgf of the standardized sum converges to the mgf of the standard normal distribution.
- This establishes that the limiting distribution is normal.

## 13.2 Proof

### Big Picture

- Let $W = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$.
- We want to show that $W \xrightarrow{d} N(0, 1)$.
- The mgf of the standard normal is $\exp(t^2/2)$.
- Steps:
  1. Derive the mgf of $W$.
  2. Take the limit of that mgf as $n \to \infty$.

3. Show that's the mgf of the standard normal.

Let $W = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$.
The moment generating function (mgf) of $W$ is:

$$M_W(t) = E[\exp(tW)] = E\left[\exp\left(\frac{t}{\sqrt{n}\sigma}\left(\sum_{i=1}^{n} X_i - n\mu\right)\right)\right]$$

$$= E\left[\prod_{i=1}^{n} \exp\left(\frac{t}{\sqrt{n}\sigma}(X_i - \mu)\right)\right]$$

Since $X_i$ are independent,

$$= \prod_{i=1}^{n} E\left[\exp\left(\frac{t}{\sqrt{n}\sigma}(X_i - \mu)\right)\right]$$

Let $Y_i = \frac{X_i - \mu}{\sigma}$ (standardized), so

$$= \left(E\left[\exp\left(\frac{t}{\sqrt{n}}Y_i\right)\right]\right)^n$$

Let $m(t)$ be the mgf of $Y_i$:

$$m(t) = E[\exp(tY_i)]$$

So,

$$M_W(t) = \left(m\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Now, expand $m(t)$ using Taylor's theorem around $t = 0$:

$$m(t) = 1 + m'(0)t + \frac{m''(t_1)t^2}{2}$$

where $t_1$ is between 0 and $t$.
Since $E[Y_i] = 0$ and $E[Y_i^2] = 1$, we have $m(0) = 1$, $m'(0) = 0$, $m''(0) = 1$.
So,

$$m(t) = 1 + \frac{m''(t_1)t^2}{2}$$

$$= 1 + \frac{t^2}{2} + \frac{(m''(t_1) - 1)t^2}{2}$$

Plug in $t/\sqrt{n}$:

$$m\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \frac{(m''(t_1) - 1)t^2}{2n}$$

So,

$$M_W(t) = \left[1 + \frac{t^2}{2n} + \frac{(m''(t_1) - 1)t^2}{2n}\right]^n$$

As $n \to \infty$, $t_1 \to 0$ and $m''(t_1) \to m''(0) = 1$:

$$\lim_{n\to\infty} M_W(t) = \lim_{n\to\infty} \left[1 + \frac{t^2}{2n}\right]^n$$

Recall $\lim_{n\to\infty}(1 + x/n)^n = \exp(x)$:

$$\lim_{n\to\infty} M_W(t) = \exp\left(\frac{t^2}{2}\right)$$

This is the mgf of the standard normal distribution. Therefore, by the mgf method, $W \xrightarrow{d} N(0, 1)$ as $n \to \infty$.

# 14 Estimation

**Date:** Nov 3, 2025

## 14.1 Today's Agenda

- How and why do we make estimates?
  - Populations and samples
  - Estimating a population mean and the law of averages
  - Estimating and using variances, covariances, and correlations
- How can we quantify the quality of our estimates?

## 14.2 Overview

- **Big Picture**: We use random variables to represent uncertain outcomes (e.g., stock returns, household income).
- **Problem**: Typically, we don't know the probability distribution of interest (e.g., unknown expected value or variance).
- **Solution**: We estimate these properties with data by computing statistics that are good approximations to the unknowns.

## 14.3 Samples and Populations

- We collect a **sample** of representative outcomes.
- We view the sample as a representative subset of some **population** (the whole class of individuals we want to characterize).
- Two examples:

| | **Population** | **Sample** |
|---|---|---|
| Pet survey | Cornell students | 125 students who responded |
| Daily stock returns | Past and future realizations | Past realizations |

- When working with samples, we need:
  1. A clear definition of the population.
  2. A reasonable probability model of how the sample was generated.
- **Comparison**:

| **Population (Unknowns)** | **Sample (Observed)** |
|---|---|
| Random Variable $Y$ | Data $\{y_1, y_2, \ldots, y_N\}$ |
| Unknown distribution | Sample histogram |
| Mean $\mu = E[Y]$ | Sample average $\bar{y}$ |
| Variance $\sigma^2 = E[(Y-\mu)^2]$ | Sample variance $s^2$ |

- **Goal**: Statistical Inference — Saying something about the population using the sample data.

- **Problems with Sampling**:
  1. **Uncertainty**: Because you only have part of the population.
  2. **Bias**: Potential biases in the sample (e.g., media/poll bias).

### 14.4 Estimating Means, Variances, Covariances, Correlations

- Population targets:

$$\mu = E[Y], \qquad Var(Y) = E[(Y - E[Y])^2], \qquad Cov(X, Y) = E[(X - E[X])(Y - E[Y])].$$

- **Sample average** (estimator of $E[Y]$):

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(y_1 + y_2 + ... + y_n)$$

- **Sample variance** (estimator of $Var(Y)$):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

- **Sample covariance** (estimator of $Cov(X, Y)$):

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

- **Sample correlation**:

$$r = \frac{s_{xy}}{s_x s_y}$$

- **Note**: The $n - 1$ denominator is used for the same reason in both variance and covariance estimation.

### 14.5 Example: Portfolio Analysis

- Consider holding a mix of two assets: "Experts' portfolio" and "Eli Lilly stock".
- Does holding a mix reduce risk?
- Given:
  - Experts: $Var = 522.16$
  - Eli Lilly: $Var = 327$
  - $Cov(\text{Experts}, \text{Eli}) = 103.54$
  - Weights: $w_E = 0.65$, $w_L = 0.35$
- Variance of a two-asset mix (generic weights $w$ and $1 - w$):

$$\begin{aligned} Var(w \cdot \text{Eli} + (1-w) \cdot \text{Experts}) = {} & w^2 Var(\text{Eli}) + (1-w)^2 Var(\text{Experts}) \\ & + 2w(1-w)Cov(\text{Eli}, \text{Experts}) \end{aligned}$$

- Variance of the mix:

$$Var(Mix) = (0.65)^2(522.16) + (0.35)^2(327) + 2(0.65)(0.35)(103.54) = 307.65$$

- Expected return of the mix:

$$E[Mix] = 0.65(10.31) + 0.35(6.37) = 8.931$$

- Correlation between Experts and Eli Lilly:

$$\text{Corr}(\text{Experts}, \text{Eli}) \approx 0.25$$

- **Bottom Line**: A mixed portfolio often has lower variance than either component, even when they are positively correlated.
- **Financial Lessons**:
  1. Optimal holdings depend on risk-return preferences.
  2. Some portfolios are sub-optimal for any investor.

### 14.6 Quality of Estimates

- How good are our estimates?
- Let $\overline{X}$ be a random variable centered on the population mean.
- How do we get the standard deviation of $\overline{X}$?
  - **Method 1**: Resample many times (Expensive).
  - **Method 2**: Rely on Central Limit Theorem (CLT) since sample means are approximately normally distributed when $n$ is large (Cheap).
- **Standard Error (SE)**: The standard deviation of an estimator (like $\overline{X}$), estimated by plugging in sample estimates for unknown parameters (like $\sigma$).

$$SE(\overline{X}) = \frac{s}{\sqrt{n}}$$

- **Example (Birthweight)**:
  - Sample A ($n = 100, s = 24$): $SE(\overline{X}) = s/\sqrt{n} = 24/10 = 2.4$.
  - Sample B ($n = 400, s = 24$): $SE(\overline{X}) = s/\sqrt{n} = 24/20 = 1.2$.

# 15 Estimation and Consistency

**Date:** Nov 5, 2025

## 15.1 Quality of Estimates

- Let $\overline{X}$ be a random variable centered on the population mean.
- How do we get the standard deviation of $\overline{X}$?
- **Method 1**: Resample 100 times and look at the standard deviation of the observed distribution of $\overline{X}$. (Expensive)
- **Method 2**: Rely on the Central Limit Theorem (CLT). Since sample means are approximately normally distributed with variance $\sigma^2/n$ when $n$ is large. (Cheap)
- **Note**:
  - The SD of an estimator usually depends on unknown population parameters (like $\sigma$).
  - When we plug in estimates for these parameters, we call it the **Standard Error (SE)**.

$$SD(X_i) \neq s \neq SE(\overline{X})$$

## 15.2 Example: Virginia Governor's Election 2021

- **Poll**: October 20-26, 2021, Washington Post/GMU polled 1,107 registered voters.
- **Result**: Youngkin support estimate $\hat{p} = 0.45$.
- **Model**: $X_i = 1$ if individual $i$ supports Youngkin, 0 otherwise.
- Then $\hat{p} = \overline{X}$, and (approximately, by the CLT) $\overline{X} \sim N\Big(p, \frac{p(1-p)}{n}\Big)$.
- **Standard Error**:

$$SD(\hat{p}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.45(1-0.45)}{1107}} = 0.015$$

$$SD(\hat{p}) \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.45(1-0.45)}{1107}} = 0.015$$

- **True SD example**: If $p = 0.50$, then $SD(\hat{p}) = \sqrt{\frac{0.50(1-0.50)}{1107}} = 0.015$.

$$\text{True}\, SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.50(1-0.50)}{1107}} = 0.015$$

- **Margin of Error**: $\approx \pm 3$ percentage points (approx. 2 SEs for 95% confidence).
- **Outcome**: Youngkin actually won with more support. The poll was off by 5.6 points.

- **Probability Calculation**: If the true $p = 0.506$, then using $SD(\hat{p}) \approx 0.015$,

$$\overline{X} \sim N(0.506, 0.015^2)$$

and the probability of observing a result at least as extreme as 0.45 (two-sided around 0.506) is:

$$\Pr(\overline{X} \leq 0.45 \cup \overline{X} \geq 0.562) = 2 \times (1 - \Pr(\overline{X} \leq 0.562)) = 2 \times \left(1 - \Pr\left(Z \leq \frac{0.562 - 0.506}{0.015}\right)\right)$$

$$= 2 \times (1 - \Pr(Z \leq \frac{0.562 - 0.506}{0.015})) = 2 \times (1 - \Pr(Z \leq 3.73)) = 0.0002$$

## 15.3 Convergence in Probability

- Let $S_1, S_2, \ldots, S_n, \ldots$ be a sequence of random variables.
- $S_n \xrightarrow{p} \mu$ if and only if $\lim_{n\to\infty} Pr(|S_n - \mu| \geq \epsilon) = 0$ for all $\epsilon > 0$.
- Typically denoted as:
  - plim $S_n = \mu$
  - $S_n$ converges in probability to $\mu$.
  - $S_n$ is a **consistent estimator** of $\mu$.

## 15.4 Law of Large Numbers (LLN)

- If $Y_1, \ldots, Y_n$ are i.i.d., $E[Y_i] = \mu$, and $Var[Y_i] < \infty$, then:

$$\overline{Y} \xrightarrow{p} \mu$$

- That is, $\overline{Y}$ is a consistent estimator of $\mu$.

## 15.5 Chebyshev's Inequality

- For any random variable $X$ and any $r > 0$:

$$Pr(|X - E[X]| \geq r) \leq \frac{Var[X]}{r^2}$$

- **Proving LLN**:

$$Var[\overline{Y}] = \frac{\sigma^2}{n}$$

$$Pr(|\overline{Y} - \mu| \geq r) \leq \frac{\sigma^2/n}{r^2} \to 0 \text{ as } n \to \infty$$

## 15.6 Examples: Bias vs. Consistency

- Let $X_1, \ldots, X_n$ be i.i.d. with $E[X_i] = \mu$.
- $\overline{X}$ is unbiased and consistent for $\mu$.
- $X_1$ is unbiased for $\mu$, but not consistent (it does not converge to $\mu$ as $n \to \infty$).
- $\frac{1}{n+1}\sum_{i=1}^{n} X_i = \frac{n}{n+1}\overline{X}$ is biased for finite $n$, but consistent for $\mu$.

### 15.7 Consistency Guidelines

- **Generalized Chebyshev's Inequality**: For any random variable $W$, $Pr(|W| \geq r) \leq E[W^2]/r^2$.
- To prove an estimator $\tilde{Y}$ is consistent for $\mu$, define $W = \tilde{Y} - \mu$.
- **Mean Squared Error (MSE)**: $E[(\tilde{Y} - \mu)^2] = Bias(\tilde{Y})^2 + Var(\tilde{Y})$.
- **Sufficient Condition**: If $MSE \to 0$ as $n \to \infty$, the estimator is consistent.
- Equivalently, for an estimator sequence $\theta_n$,

$$\lim_{n\to\infty} E[\theta_n] = \mu \quad \text{and} \quad \lim_{n\to\infty} \mathrm{Var}[\theta_n] = 0 \quad \Rightarrow \quad \theta_n \xrightarrow{p} \mu.$$

# 16 Maximum Likelihood Estimation and Confidence Intervals

**Date:** Nov 10, 2025

## 16.1 Maximum Likelihood Estimation (MLE)

- **Setup**:
  - We have a random sample of i.i.d. observations.
  - We know the functional form of the pdf/pmf for each observation, say $f(x_i; \theta)$.
  - We do NOT know the parameters $\theta$.
- **Likelihood Function**: The joint probability (pdf or pmf) of the whole sample, viewed as a function of the parameters $\theta$.
$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$
- **MLE**: The estimator that maximizes $L(\theta)$ for the observed data.
- **Log Likelihood**: It is usually easier to maximize the log of the likelihood function.
$$\mathcal{L}(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

## 16.2 Examples

### 16.2.1 Example 1: Normal Distribution $N(\mu, \sigma^2)$

- Likelihood:
$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$
- Equivalent factorization:
$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$
- Log likelihood:
$$\mathcal{L}(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$
- **MLE Estimators**:
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$
- Note: $\hat{\sigma}^2_{MLE}$ is biased (denominator is $n$ instead of $n-1$), but consistent.

### 16.2.2 Example 2: Bernoulli($p$)

- PDF: $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$.
- Likelihood: $L(p) = \prod p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$.
- Log Likelihood:
$$\mathcal{L}(p) = (\sum x_i)\ln(p) + (n - \sum x_i)\ln(1-p)$$
- Maximizing with respect to $p$:
$$\frac{\partial \mathcal{L}}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$$
$$\Rightarrow \hat{p}_{MLE} = \frac{\sum X_i}{n} = \overline{X}$$

## 16.3 Properties of MLE

1. **Consistent**: Converges to true parameter as $n \to \infty$.
2. **Asymptotically Normal**.
3. **Asymptotically Efficient**: Has the Minimum Variance among all consistent estimators (Cramér-Rao Lower Bound).

- **Catch**: You must assume the correct distribution.

## 16.4 Confidence Intervals (Large Sample)

- From CLT: $\overline{X} \sim N(\mu, \sigma^2/n)$, so $Var(\overline{X}) = \sigma^2/n \approx s^2/n$.
- **Standard Error**: $SE(\overline{X}) = s/\sqrt{n}$.
- We choose $a, b$ such that:
$$\Pr(\overline{X} - a \leq \mu \leq \overline{X} + b) = 0.95,$$
i.e. $\Pr(\mu < \overline{X} - a) = 0.025$ and $\Pr(\mu > \overline{X} + b) = 0.025$.
- 95% Confidence Interval for $\mu$:
$$\left[\overline{X} - 1.96\frac{s}{\sqrt{n}}, \quad \overline{X} + 1.96\frac{s}{\sqrt{n}}\right]$$
- **Interpretation**: In repeated sampling, 95% of the intervals constructed this way will cover the true population mean $\mu$.

## 16.5 Example: Consumer Expenditure Survey (2013)

- For `bedroomq`: $n = 6679$, $\overline{X} = 2.778$, $s = 1.074$.
- Standard error:
$$SE(\overline{X}) = s/\sqrt{n} = 1.07/\sqrt{6679} = 0.013$$

## 16.6 Small Sample Confidence Intervals

For large $n$, the CLT implies

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

for any distribution of $X_i$ with finite variance.

If $X_i \sim N(\mu, \sigma^2)$ and $\sigma$ is unknown, then for any $n > 1$,

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

So a $1 - \alpha$ confidence interval is:

$$\overline{X} \pm t_{\alpha/2,\, n-1} \cdot SE(\overline{X}).$$

- Example ($n = 10$, $\overline{X} = 2.9$, $s = 1.1005$): $SE(\overline{X}) = 1.1005/\sqrt{10} = 0.348$ and a 99% CI uses $t_{0.005,9}$, so half-width $\approx 0.348 \cdot t_{0.005,9} \approx 1.131$.

# 17 Introduction to Hypothesis Testing

**Date:** Nov 12, 2025

## 17.1 Concepts

- **Goal**: Determine if an observed phenomenon is "real" or just due to chance.
- **Null Hypothesis ($H_0$)**: The observed difference reflects pure chance variation. (Default assumption).
- **Alternative Hypothesis ($H_1$ or $H_a$)**: The observed difference is real.
- **Generic test statistic**:

$$z = \frac{\text{Observed} - \text{Expected(Under Null)}}{SE(\overline{x})}.$$

## 17.2 Example: Quarter of Birth and Schooling (Angrist and Krueger 1991)

- **Question**: Do kids born in the 1st quarter get less schooling?
- **Data**: 1980 US Census.
- **Sample**: $n = 81,671$ (born in 1st quarter).
  - Mean schooling $\overline{X} = 12.69$ years.
  - Standard deviation $s = 3.310$ years.
- **Population**: National average $\mu_0 = 12.77$.

## 17.3 One Sample Test (Large Sample)

- **Hypotheses**:

$$H_0 : \mu = 12.77$$
$$H_1 : \mu \neq 12.77$$

- **Test Statistic**:

$$z = \frac{\text{Observed} - \text{Expected(Under Null)}}{SE(\overline{x})} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

- **Distribution**: Under $H_0$, $Z \sim N(0, 1)$ (approximately suitable for large samples).
- **Reminder**: The null distribution depends on the data-generating process, e.g.

$$X_i \sim \text{Normal}(\mu_0, \sigma^2), \quad X \sim \text{Binomial}(p = \mu_0, n), \quad X \sim \text{Poisson}(\lambda = \mu_0),$$

  but for sample means we often use a Normal approximation via the CLT:

$$\frac{\overline{X} - \mu_0}{SE(\overline{X})} \approx \text{Normal}(0, 1).$$

- **Calculation**:

$$z = \frac{12.69 - 12.77}{3.310/\sqrt{81671}} \approx \frac{-0.08}{0.0116} \approx -6.9$$

- **Decision**: Since $|z| > 1.96$, we reject $H_0$ at the 5% level.

### 17.4 Decision Rules and P-values

- **Decision Rule**: Reject if p-value $< \alpha$ (usually $\alpha = 0.05$).
- **P-value**: Probability of observing a test statistic as extreme as the one seen, assuming $H_0$ is true.
- **Statistical Significance**: If we reject $H_0$, the result is "statistically significant".

### 17.5 Two Sample Test (Independent Samples)

- **Setup**: Two populations with means $\mu_1, \mu_2$ and variances $\sigma_1^2, \sigma_2^2$.
- **Hypotheses**: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.
- **Assumption**: Equal variances ($\sigma_1^2 = \sigma_2^2$).
- Under the equal-variance assumption,

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

- **Test Statistic**:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  where $s_p$ is the pooled standard deviation:

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

- **Q3 vs Q4 Example**:
  - Q3: $n_1 = 86,856, \overline{X}_1 = 12.81, s_1 = 3.25$.
  - Q4: $n_2 = 80,844, \overline{X}_2 = 12.84, s_2 = 3.24$.
  - $z \approx -1.89$.
  - p-value $= 2 \times \Pr(Z > 1.89) = 0.059$.
  - Result: Fail to reject $H_0$ at 5% level (though close).

# 18 Small Sample Hypothesis Testing

**Date:** Nov 17, 2025

## 18.1 Large Sample Tests Review

- **One Sample**: $H_0 : \mu = \mu_0$. Test statistic $z = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.
- **Two Independent Samples**: $H_0 : \mu_1 = \mu_2$. Test statistic $z = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$.
- If the population SD is known and equal across groups, an alternative form is:

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

- Under an equal-variance model,

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

- **Approximation**: Uses Central Limit Theorem (CLT) for $n > 30$. Under $H_0$, $Z \sim N(0, 1)$.

## 18.2 Small Sample Tests (The t-test)

When sample sizes are small ($n < 30$), CLT does not apply.

- **Assumptions**: Data are approximately normal, independent, equal variances.
- **Distribution**: The test statistic follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.
- **One Sample t-test**:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \quad \text{(under } H_0 \text{ and Normal data).}$$

- **Test Statistic (Two Independent Samples)**:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- As $n \to \infty$, the t-distribution converges to the standard normal distribution using $z$.

## 18.3 Example: Two Independent Samples (Small $n$)

- Example with $n_1 = n_2 = 14$, $s_p = 10$, and sample means $\overline{x}_1 = 64.4$, $\overline{x}_2 = 67.9$:

$$t_{26} = \frac{64.4 - 67.9}{10\sqrt{\frac{1}{14} + \frac{1}{14}}} = -0.926.$$

$$t_{26} = -0.926.$$

### 18.4 Paired Samples

Used when we have pairs of comparable individuals or repeated measures on the same individual (longitudinal).

- **Definition**: Let $D_i = X_{1i} - X_{2i}$ be the difference for pair $i$.
- **Hypothesis**: $H_0 : \Delta = 0$ (mean difference is zero).
- **Test Statistic**:
$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$
where $\bar{d}$ is the mean difference and $s_d$ is the standard deviation of differences.
$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2}.$$
$$s_d = \sqrt{\frac{1}{n-1}\sum(d_i - \overline{d})^2}.$$
- **Distribution**: Under $H_0$, $t \sim t_{n-1}$.
- **Advantages**: Controls for individual variation, often resulting in lower standard errors compared to independent samples.

### 18.5 One-sided vs. Two-sided Tests

- **Two-sided**: $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.
  - Rejection region split between both tails.
- **One-sided**: $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ (or vice versa).
  - Used when we are willing to assume the effect can only go in one direction.
  - P-value is half of the two-sided p-value (easier to reject $H_0$).
  - **Note**: The conservative approach is to always use a two-sided test.

# 19 More Hypothesis Testing

**Date:** Nov 19, 2025

## 19.1 Type I and Type II Errors

| | $H_0$ **is True** | $H_0$ **is False** |
|---|---|---|
| **Accept** $H_0$ | No Error | Type II Error ($\beta$) |
| **Reject** $H_0$ | Type I Error ($\alpha$) | No Error |

- **Type I Error ($\alpha$)**: Rejecting $H_0$ when it is true.
  - We control this by setting the significance level $\alpha$ (usually 0.05).
- **Type II Error ($\beta$)**: Failing to reject $H_0$ when it is false.
  - Depends on $\alpha$, sample size ($n$), and the true effect size.
  - Power of the test $= 1 - \beta$.

## 19.2 Test of Equality of Variances

Before doing a two-sample t-test (equal variances), we can test if the variances are actually equal.

- **Hypotheses**: $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$.
- **Test Statistic**:
$$F = \frac{s_1^2}{s_2^2}$$
(Convention: Put the larger sample variance in the numerator so $F \geq 1$).
$$H_0 : \sigma_1^2 = \sigma_2^2, \qquad F = \frac{s_1^2}{s_2^2}.$$
- **Distribution**: Under $H_0$ (and normal populations), $F \sim F_{n_1-1,n_2-1}$.
- **Decision**: Reject if p-value $< \alpha$.

## 19.3 Two Sample Test with Unequal Variances (Welch's t-test)

If we reject equality of variances, or simply don't want to assume it:

- In general,
$$Var[\overline{X}_1 - \overline{X}_2] = Var[\overline{X}_1] + (-1)^2 Var[\overline{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \neq \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}.$$
- Correspondingly,
$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$
- **Hypotheses**: $H_0 : \mu_1 = \mu_2$.

- **Test Statistic**:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Distribution**: Approximate t-distribution with degrees of freedom (Satterthwaite):

$$d' = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}}.$$

## 19.4 Tests with Proportions (Large Samples)

Since binary outcomes are not normal, we rely on the CLT (large samples).

- **One Sample Proportion**:
  - $H_0 : p = p_0$.
  - Test Statistic: $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.
- **Two Independent Proportions**:
  - $H_0 : p_1 = p_2$.
  - Pooled proportion: $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$.
  - Test Statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

## 19.5 Big Picture: Which Mean Test?

- There are many tests of population means:
  - One sample vs. two samples
  - Small sample vs. large sample
  - Two samples: independent vs. paired; equal vs. unequal variance

## 19.6 Formula Summary (Common Cases)

- **One sample mean (large $n$)**:

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \approx N(0,1).$$

- **Two independent means (large $n$, unequal variances)**:

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0,1).$$

- **Two independent means (large $n$, equal variances)**:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \qquad z = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx N(0, 1).$$

- **Two independent binary outcomes (large samples)**:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}, \qquad s = \sqrt{\hat{p}(1 - \hat{p})}, \qquad z = \frac{\hat{p}_1 - \hat{p}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx N(0, 1).$$

- **Paired samples**:

$$t = \frac{\overline{d}}{s_d/\sqrt{n}}, \qquad s_d = \sqrt{\frac{1}{n - 1}\sum(d_i - \overline{d})^2}.$$

# 20 Univariate Nonparametric Tests

**Date:** Nov 24, 2025

## 20.1 Review of Distributions

We have covered several distributions derived from the Normal distribution:

- **Chi-square Distribution** ($\chi^2_k$):
  - If $Z_1, \dots, Z_k \sim N(0,1)$ are independent, then $X = \sum_{i=1}^k Z_i^2 \sim \chi^2_k$.
  - Notation: $V \sim \chi^2(n)$.
  - Mean: $k$, Variance: $2k$.
  - PDF:
$$f_k(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x > 0.$$
- **F Distribution** ($F_{m,n}$):
  - If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then $X = \frac{U/m}{V/n} \sim F_{m,n}$.
  - Used for testing equality of variances ($H_0 : \sigma_1^2 = \sigma_2^2$).
- **Student's t Distribution** ($t_n$):
  - If $Z \sim N(0,1)$ and $U \sim \chi^2_n$ are independent, then $X = \frac{Z}{\sqrt{U/n}} \sim t_n$.
  - Used for small sample tests of means.

## 20.2 K-Multinomial Random Variables

Let $Y \sim km(n, p_1, p_2, \dots, p_k)$.

- $Y$ is a vector of $k$ counts $(Y_1, \dots, Y_k)$.
- $\sum p_i = 1$.
- $n$ total trials. $Y_k$ is the number of times we observe the $k$-th outcome.

## 20.3 The Univariate Chi-square Test

We want to test if a sample comes from a specific multinomial distribution (Goodness of Fit).

- **Hypothesis**: The data is drawn from a multinomial distribution with probabilities $p_1, \dots, p_k$.
- **Test Statistic (Pearson's Chi-square)**:
$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} \sim \chi^2_{k-1}$$
- Here, $Y_i$ is the *Observed* count ($O_i$) and $np_i$ is the *Expected* count ($E_i$).
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

## 20.4 Proof for k=2

For $k = 2$, the Chi-square statistic is:

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

Since $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$, this simplifies to:

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1} + \frac{((n - Y_1) - n(1 - p_1))^2}{n(1 - p_1)} = \frac{(Y_1 - np_1)^2}{np_1(1 - p_1)}.$$

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1 - p_1)}$$

This is exactly the square of the z-statistic for a (single) proportion:

$$z_k = \frac{Y_k - np_k}{\sqrt{np_k(1 - p_k)}}.$$

In particular,

$$z_k^2 = \frac{(Y_k - np_k)^2}{np_k(1 - p_k)}.$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \implies z^2 = \frac{(Y_1/n - p_1)^2}{p_1(1 - p_1)/n} = \frac{(Y_1 - np_1)^2}{np_1(1 - p_1)}$$

Thus, for $k = 2$, the Chi-square test is equivalent to the two-sided z-test for proportions, and

$$Q_1 = z^2 \approx \chi^2(1).$$

## 21 Bivariate Nonparametric Tests

**Date:** Dec 1, 2025

### 21.1 Bivariate Chi-square Test (Test of Independence / Homogeneity)

Used to test if two (or more) groups have the same distribution of categorical outcomes.

- **Setup**: Data is arranged in a contingency table with $J$ groups (columns) and $K$ values (rows).
- **Hypothesis**: $H_0$: The distributions of the values are the same across groups.
- **Expected Counts ($E_{ij}$)**: Under $H_0$, the best estimate for the probability of being in category $i$ is the pooled proportion:
$$\hat{p}_i = \frac{\sum_{j=1}^{J} Y_{ij}}{N}$$
Then, $E_{ij} = \hat{p}_i \times (\text{Total count for Group } j)$.
$$E_{ij} = \hat{p}_i \sum_{k=1}^{K} Y_{kj}$$
$$\hat{p}_i = \frac{\sum_{j=1}^{J} Y_{ij}}{N}, \quad E_{ij} = \hat{p}_i \sum_{k=1}^{K} Y_{kj}.$$
$$E_{ij} = \frac{(\text{Row Total}_i) \times (\text{Column Total}_j)}{N}$$
- **Test Statistic**:
$$Q = \sum_{j=1}^{J} \sum_{i=1}^{K} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}}$$
- **Distribution**: Under $H_0$, $Q \sim \chi^2_{(J-1)(K-1)}$.
$$Q = \sum_{j=1}^{J} \sum_{i=1}^{K} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{((J-1)\times(K-1))}.$$
- **Condition**: Works well if expected counts in each cell are at least 5.

### 21.2 Fisher's Exact Test

- Used when sample sizes are small (e.g., cell counts $< 5$) where the Chi-square approximation fails.
- Does not rely on the CLT or large samples.
- Computationally intensive for large tables.

### 21.3 Median Tests

Useful when data is skewed or has outliers (t-tests might be invalid).

### 21.3.1 One Sample Median Test

- $H_0$ : Median$(X) = m_0$.
- Under $H_0$, we expect 50% of observations to be below $m_0$.
- Let $C$ = count of observations below $m_0$.
- Under $H_0$, $C \sim$ Binomial$(n, 0.5)$. We calculate the p-value using the Binomial distribution.

### 21.3.2 Two Sample Median Test

Tests if two populations have the same median.

1. Combine the two samples and compute the **pooled median**.
2. For each sample, count the number of observations above and below the pooled median.
3. Create a $2 \times 2$ contingency table with these counts.
4. Perform a Chi-square test (or Fisher's exact test) on this table.

# 22 Randomized Experiments

**Date:** Dec 3, 2025

## 22.1 Today's Agenda

- Define and work with randomized experiments and associated jargon:
  - Average Treatment Effect (ATE)
  - Treatment-on-the-Treated (TT)
  - Selection Bias (SB)
  - Heterogeneous Treatment Effects
  - Intent-to-Treat (ITT)
- Two case studies: Online advertising; Breza et al. (2021)

## 22.2 Advertising Effectiveness (Motivation)

- A common observational setup (e.g., comScore/HBR-style studies) compares outcomes across consumers who:
  1. saw no ads (control),
  2. saw only online display ads,
  3. saw only targeted search ads,
  4. saw both display and search ads,

  and then measures outcomes like website visits and (online/offline) sales.
- Key issue: these group comparisons may reflect selection, not causal effects.

## 22.3 Notation and Definitions

We are interested in the causal effect of a treatment (e.g., advertising).

- $Y_{0i}$: Potential outcome for individual $i$ if **not** treated.
- $Y_{1i}$: Potential outcome for individual $i$ if treated.
- $D_i \in \{0, 1\}$: Indicator variable. $D_i = 1$ if treated, 0 otherwise.
- **Observed Outcome** $Y_i$:

$$Y_i = (1 - D_i)Y_{0i} + D_i Y_{1i} = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

## 22.4 Treatment Effects

- **Average Treatment Effect (ATE)**: The expected difference in outcomes if the entire population were treated vs. not treated.

$$\text{ATE} = E[Y_{1i} - Y_{0i}]$$

- **Treatment on the Treated (TT)**: The effect for those who actually received the treatment.

$$\text{TT} = E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$$

- **Heterogeneous Treatment Effects**: Treatment effects $(Y_{1i} - Y_{0i})$ may differ across individuals (e.g., some groups respond more than others).

## 22.5 Stylized Example

- Suppose the potential outcomes are:

| $i$ | $Y_{0i}$ | $Y_{1i}$ |
|---|---|---|
| 1 | 50 | 50 |
| 2 | 40 | 40 |
| 3 | 100 | 110 |
| 4 | 150 | 160 |

- Estimated ATE:

$$\widehat{ATE} = \frac{1}{4}\sum_{i=1}^{4}(Y_{1i} - Y_{0i}) = \frac{0+0+10+10}{4} = 5.$$

- If we observe $D_3 = D_4 = 1$ and $D_1 = D_2 = 0$, then the "HBR effect" (confounding comparison)

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \frac{110+160}{2} - \frac{50+40}{2} = 90.$$

## 22.6 Selection Bias

When we compare the average outcomes of treated vs. control groups using observational data, we get:

$$\text{Observed Difference} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

This can be decomposed as:

$$= \underbrace{(E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1])}_{\text{Treatment on the Treated}} + \underbrace{(E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0])}_{\text{Selection Bias}}$$

- **Selection Bias**: The difference in baseline outcomes (without treatment) between those who were treated and those who were not.

- Example: People who click on ads might be more likely to buy the product anyway (even without the ad).

## 22.7 Randomization

If treatment $D_i$ is randomly assigned, it is independent of potential outcomes $Y_{0i}, Y_{1i}$.

- Independence implies Selection Bias is zero: $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$.
- Therefore, Observed Difference = ATE.

## 22.8 Intent-to-Treat (ITT)

What if not everyone assigned to the treatment group actually completes the treatment (non-compliance)?

- **ITT Effect**: The difference between the outcome of the group *assigned* to treatment and the group *assigned* to control.
- This preserves the benefits of randomization even if compliance is imperfect.

## 22.9 Case Studies

### 22.9.1 Online Advertising

Comparing sales between those who saw ads and those who didn't often yields a "confounding comparison" due to selection bias. Randomized field experiments (like allocating cookies/users to treatment/control) are needed to measure the true lift.

### 22.9.2 Breza et al. (2021): Social Media Messages

- **Research Question**: Can social media messages from health professionals influence behavior (holiday travel) and health outcomes (COVID-19 infections)?
- **Experiment**: A large-scale randomized control trial (RCT) using Facebook ads in the US before Thanksgiving and Christmas 2020.
- **Intervention**: Short video messages from doctors and nurses.
  - Quote: "This Thanksgiving, the best way to show your love is to stay home. If you do visit, wear a mask at all times... don't risk spreading COVID. Stay safe, stay home."

**Randomization Design** The study covered 820 counties across 13 states and used a two-stage randomization:

1. **Stage 1 (County Level)**: Counties were randomized into "high intensity" (75% treated zip codes) or "low intensity" (25% treated zip codes) groups.
2. **Stage 2 (Zip-code Level)**: Within counties, zip codes were randomized to Treatment (users see ads) or Control (no ads).

This design allows for measuring both direct effects and spillovers (though the primary analysis focused on the direct intent-to-treat effect).

**Results**

- **First Outcome (Travel)**: Checked mobile phone location data.
  - Result: Average distance traveled decreased by approx. 1 percentage point (0.993 ppt) in high-intensity counties compared to low-intensity counties in the 3 days before holidays.
- **Second Outcome (COVID-19)**: Infections measured 2 weeks post-holiday.
  - Result: ZIP codes with high-intensity treatment saw a **3.5% reduction** in COVID-19 infections compared to controls.

**Conclusion** Social media interventions by trusted professionals can be an effective low-cost public health tool. The study demonstrates the power of large-scale RCTs in digital platforms.