Review of Statistics: Primer

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.
Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.

The textbook comes with online resources and study guides. Other references will be given from time to time.

Content

- **Single Random Variable**
  - discrete and continuous random variables
  - expectations, mean, variance
- **Multiple Random Variables**
  - conditional probabilities
  - conditional means, variance, covariance
  - independence, correlation
- **The Normal Distribution**
  - properties of the normal distribution
  - standardizing and the z-score
  - computing probabilities
- **The Central Limit Theorem**
  - distribution of the sample mean
  - Law of Large Numbers (LLN)
  - Central Limit Theorem (CLT)

Random Variable

What is a random variable?

- **Example:** Flipping a coin.
- While the outcome is uncertain, a random variable has a **distribution**.
- For any subset of the sample space, the distribution describes the probability that the random variable takes a value in that subset.
- A fair coin has a 1 in 2 probability of landing head. The probability that the random variable X is in $\{H\}$ is $1/2$. The probability that X is in $\{T\}$ is $1/2$. The probability that X is in $\{H, T\}$ is 1.

Random Variable

- Suppose we want to know about the education levels of people in California.
- Different people have different levels of education. The education level of a randomly selected person in the population is uncertain.
- The education level is a **random variable** with a distribution that describes the probability that a randomly selected person has an education level in certain range, e.g. 80% have a high school diploma, 30% have a college degree.
- In general, we do not know the exact distribution of the random variable. Statistics is a set of methods used to extract information from fixed samples and make inferences about the underlying distribution of the random variable.

Sample Space

- Let X denote a random variable. The **sample space** Ω is the set of all possible values of X.
  - Coin flipping: The sample space is $\{H, T\}$.
  - Rolling a die: The sample space is $\{1, 2, 3, 4, 5, 6\}$.
  - Racing the 100m dash at a competition: The sample space may be the range $[9.5, 11]$, measured in seconds.
- **Discrete random variable:** The sample space of X is **countable**.
- **Continuous random variable:** The sample space of X is **uncountable**.
  - Flipping a coin and rolling a die are represented by discrete random variables.
  - Racing the 100m dash is represented by a continuous random variable.

Probability

- Let Ω denote the sample space of random variable X.
- Let E_i denote any subset of the sample space Ω — an event.
- We are interested in the probability $\Pr(X)$ that X takes values in $E_i \in \Omega$.
- The distribution of the random variable X is a map from $\cup_{i=1}^{n} E_i$ onto $[0, 1]$.
- **Axioms of probability:** The probability $\Pr(X)$ satisfies:
  - $\Pr(\Omega) = 1$
  - $\Pr(\emptyset) = 0$
  - $0 \leq \Pr(X) \leq 1$
  - If $E_1, E_2, \ldots, E_n$ are pairwise disjoint, then $\Pr(\cup_{i=1}^{n} E_i) = \sum_i \Pr(E_i)$.

Discrete Random Variables

- Let X be a **discrete random variable**. The distribution of X is described by the **probability mass function (pmf)**, $p_X(X)\colon E \subseteq \Omega \rightarrow [0, 1]$.
- For each element x of the sample space, $x \in \Omega$, the probability mass function, $p_X(\cdot)$, describes the probability that X takes value x: $p_X(x) = \Pr(X = x)$.
- The pmf can be used to recover the probability that X takes values in any subset E of the sample space Ω.

$$\Pr(E) = \sum_{x \in E} \Pr(X = x) = \sum_{x \in E} p_X(x)$$

Discrete Random Variables

- Compute the probability of getting an value in one roll of a fair die.
- Let X denote the outcome of the die. The distribution of X has the following pmf:

$$p_X(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

- Use the pmf to compute $\Pr(E)$ for $E = \{2, 4, 6\}$:

$$\begin{aligned} \Pr(E) = \Pr(x \in \{2, 4, 6\}) &= \sum_{x \in \{2,4,6\}} p_X(x) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \end{aligned}$$

Continuous Random Variables

- If X is a **continuous random variable**, the sample space Ω is uncountable, so we cannot use a probability mass function to describe its distribution.
- If we attempted to assign a probability to each element in the sample space, we would be faced with one of two consequences:
  - If we assigned a finite probability to an uncountable subset of the sample space, the sum of the probabilities would tend to infinity.
  - The only way for the sum of probabilities to tend to 1 would be to assign a zero probability to each element of an uncountable subset of the sample space.
- Either way, it would not be useful and would lead to absurd calculations.
  - What is the probability of drawing the number π from the subset of the real line $[3, 4]$? Answer: zero.
  - What is the probability of drawing any number from a subset of the real line? Answer: zero.
- We cannot use a pmf to describe the distribution of a continuous random variable.
- Instead, we reason in terms of non-overlapping dense subsets of the real line that are made arbitrarily small.

Continuous Random Variables.

- If X is a **continuous random variable**, we use the probability density function (pdf), $f_X(\cdot)$ to describe the distribution of X.
- The pdf is related to the probability measure $\Pr$ via the following equation:

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

- This identity can be used to calculate $\Pr(E)$ for any set $E \subseteq \Omega$.

Continuous Random Variables

- **Example:** Let X be a continuous random variable with pdf f_X given

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- This distribution is called the **uniform distribution** on $[0, 1]$.
- Calculate $\Pr([0, 0.5])$:

$$\begin{aligned} \Pr(X \in [0, 0.5]) &= \int_0^{0.5} f_X(x)\,dx \\ &= \int_0^{0.5} 1\,dx \\ &= x\Big|_0^{0.5} \end{aligned}$$

$$= 0.5 - 0 = 0.5$$

Review

- **Random Variable:**
  - Describes an event whose outcome is unknown.
  - To each outcome, associate a probability.
- **Discrete Random Variable:**
  - Its sample space is countable.
  - The probability distribution is described by a probability mass function (pmf).

$$\Pr(X \in \{x\}) = p_X(x)$$

- **Continuous Random Variable:**
  - Its sample space is uncountable.
  - The probability distribution is described by probability density function (pdf).

$$\Pr(X \in [a, b]) = \int_a^b f_X(x)\,dx$$

Expectation

- The expectation can be interpreted as a generalization of the arithmetic mean.
- The weighted average of the random variable X, where the weights are the probability associated with each possible value of X.
- The expectation of X is denoted $\mathrm{E}[X]$.

Discrete R.V	Continuous R.V
$\sum_{x \in \Omega} x \cdot p_X(x)$	$\int_\Omega x \cdot f_X(x)\, dx$

- Note that the difference between discrete and continuous is just summation vs. integral.
- In a population, the expectation may be denoted μ_X. In a sample, it may be denoted $\overline{X}$.

Expectation

- Consider a lottery that pays:
  - $\$100$ with probability $1/2$
  - $\$400$ with probability $1/4$
  - $\$0$ with probability $1/4$
- The payout of this lottery can be represented by a random variable X with pmf:

$$p_X(x) = \begin{cases} 1/2 & \text{if } x = \{100\} \\ 1/4 & \text{if } x \in \{0, 400\} \\ 0 & \text{otherwise} \end{cases}$$

- Calculate the expected value of this lottery $\mathrm{E}[X]$ – The mean value of the lottery prize.

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in \{0,100,400\}} x \cdot p_X(x) \\ &= 0 \cdot \frac{1}{4} + 100 \cdot \frac{1}{2} + 400 \cdot \frac{1}{4} = 50 + 100 = 150 \end{aligned}$$

- A gambler can expect to win $\$150$ by playing this lottery.

Expectation

- Let X be continuously distributed with sample space $[0, 1]$ and pdf:

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- Calculate the expectation:

$$\begin{aligned} \mathrm{E}[X] &= \int_\Omega x \cdot f_X(x)\, dx \\ &= \int_0^1 x \cdot 1\, dx \\ &= \frac{x^2}{2}\bigg|_0^1 = \frac{1^2}{2} - \frac{0^2}{2} = \frac{1}{2} \end{aligned}$$

- The expected value of X is 0.5. You can expect to win half a dollar from playing the lottery many, many times.

Expectation

- Consider the mean of any function $g(X)$, denoted $\mathrm{E}[g(X)]$.
- $g(X)$ is a random variable with distribution derived from X.
- The formula for calculating $\mathrm{E}[g(X)]$ is basically the same as for calculating $\mathrm{E}[X]$.

Discrete R.V	Continuous R.V
$\sum_{x \in \Omega} g(x) \cdot p_X(x)$	$\int_\Omega g(x) \cdot f_X(x)\, dx$

- The pmf/pdf is multiplied by $g(x)$ instead of just x.

Expectation

- **Linearity of Expectations:**
$$\mathrm{E}[ag(X) + bh(X)] = a\,\mathrm{E}[g(X)] + b\,\mathrm{E}[h(X)]$$
for any $a, b \in \mathbb{R}$.
- This property applies to two different random variables X and Y:
$$\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]$$

Expectation

- Suppose that X has pdf
$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
- Calculate the **second moment** of X, $\mathrm{E}[X^2]$:
$$\begin{aligned} \mathrm{E}[X^2] &= \int_\Omega x^2 \cdot f_X(x)\,dx \\ &= \int_0^1 x^2 \cdot 1\,dx \\ &= \frac{x^3}{3}\bigg|_0^1 = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3} \end{aligned}$$

Variance

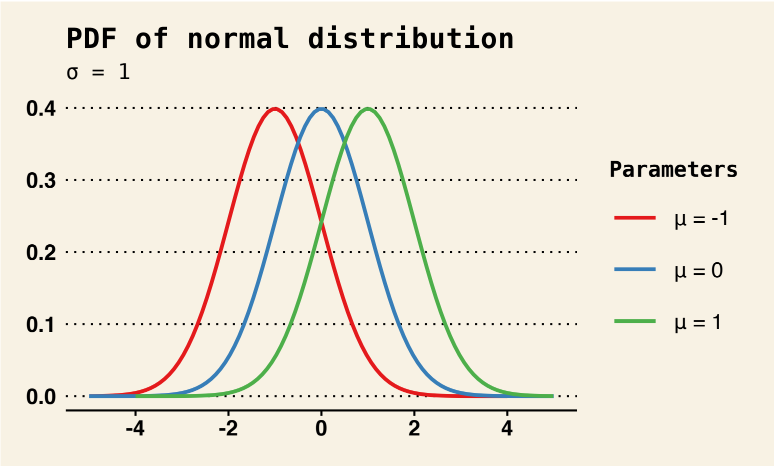
- The variance of a random variable X is a measure of the spread of the random variable X – a measure of how far on average X is from its mean.
$$\mathrm{var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$
- Using linearity of the expectation the expression above can be simplified:
$$\begin{aligned} \mathrm{var}(X) &= \mathrm{E}[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2] \\ &= \mathrm{E}[X^2] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \end{aligned}$$
- From the definition, the variance is always positive $\mathrm{var}(X) \geq 0$.
- The last expression is often convenient.

Variance

- The linearity of the expectation and the formula $\mathrm{var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$ imply several convenient properties of the variance.
- For any constants $a, b \in \mathbb{R}$ and any random variable X, we have:
$$\begin{aligned} \mathrm{var}(X + a) &= \mathrm{var}(X) \\ \mathrm{var}(aX) &= a^2\,\mathrm{var}(X) \end{aligned}$$
- Combining these gives:
$$\mathrm{var}(aX + b) = a^2\,\mathrm{var}(X)$$

Variance

Changes in the mean μ do not affect the variance σ:



Variance

- Suppose we have a lottery that pays $\$200$ with probability $1/2$ and nothing otherwise. The payout of this lottery is a random variable X with pmf:

$$p_X(x) = \begin{cases} \frac{1}{2} & \text{if } x \in \{0, 200\} \\ 0 & \text{otherwise} \end{cases}$$

- The expected payout of this lottery is $\mathrm{E}[X] = \$100$.
- How much we can expect our winnings to deviate from the expected value?
- This can be computed directly:

$$\mathrm{var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

or indirectly:

$$\mathrm{var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

Variance

- The payout of this lottery is a random variable X with pmf:

$$p_X(x) = \begin{cases} \frac{1}{2} & \text{if } x \in \{0, 200\} \\ 0 & \text{otherwise} \end{cases}$$

- Direct Calculation:

$$\begin{aligned} \mathrm{var}(X) &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= \mathrm{E}[(X - 100)^2] \\ &= \sum_{x \in \{0,200\}} (x - 100)^2 \cdot p_X(x) \\ &= (0 - 100)^2 \cdot \frac{1}{2} + (200 - 100)^2 \cdot \frac{1}{2} \\ &= 10,000 \end{aligned}$$

Variance

- The payout of this lottery is a random variable X with pmf:

$$p_X(x) = \begin{cases} \frac{1}{2} & \text{if } x \in \{0, 200\} \\ 0 & \text{otherwise} \end{cases}$$

- Indirect Calculation:

$$\begin{aligned} \mathrm{var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\ &= \sum_{x \in \{0,200\}} x^2 \cdot p_X(x) - (100)^2 \\ &= 0^2 \cdot \frac{1}{2} + 200^2 \cdot \frac{1}{2} - 100^2 \\ &= 10,000 \end{aligned}$$

Standard Deviation

- **Standard deviation of X**:
$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\mathrm{var}(X)}$$
- The standard deviation is the square root of the variance.

Bivariate Random Variables

- Consider the relationship between two random variables.
  - We care about the relationship between education and income
  - We care about the relationship between consumption of a medicine and a health outcome
- Note that:
  - not everyone has the same education/income
  - not everyone who takes a medicine will have the same health outcome

Bivariate Random Variables

- Let (X, Y) be a pair of joint random variables. Let Ω denote the sample space and $\Pr(\cdot)$ a probability measure defined on subsets of the sample space $E \subseteq \Omega$, that is $\Pr(\cdot) : \Omega \to [0, 1]$.
- Example: Let X denote income and Y denote age. The probability that a randomly selected person from the population has an income between $\$0$ and $\$100,000$ and is between 40 and 42 years old is denoted:
$$\Pr(\{0 \leq X \leq 100,000,\ 40 \leq Y \leq 42\})$$
- For two random variables, X, Y, we can represent the probability mass function (pmf) as a table.

Discrete Random Variables

- Let X be a random variable that describes whether a person gets 4 hours of sleep a night; 8 hours a sleep a night; or 12 hours of sleep a night. Let Y be a random variable that describes whether a person drinks 1 or 2 cups of coffee a day.
- The joint pmf of X and Y can be described with the table below

$p(x, y)$	1 cup	2 cups
4 hours	0	$1/6$
8 hours	$1/3$	$1/3$
12 hours	$1/6$	0

- The probability that a randomly selected person gets 8 hours of sleep and drinks 1 cup of coffee a day is $1/3$.
- Exercise: What is the probability that a randomly selected person gets 8 hours of sleep?

Continuous Random Variables

- As with single continuous random variables, the distribution of a continuous random variable is defined by a probability density function, $f_{XY}(x, y)$.
- As before, the joint pdf will be related to the joint probability measure $\Pr(\cdot)$

$$\Pr(\{a \leq X \leq b, c \leq Y \leq d\}) = \int_a^b \int_c^d f_{XY}(x, y)\, dy\, dx$$

Continuous Random Variables

- Consider two sprinters in the 100m dash. Let X denote the finish time of the favorite competitor and Y denote the finish time of the underdog competitor. Suppose their times follow the following joint pdf:

$$f_{XY}(x, y) = \begin{cases} 1 & \text{if } 9.5 \leq x \leq 10.5 \text{ or } 10 \leq y \leq 11 \\ 0 & \text{otherwise} \end{cases}$$

- Calculate the probability that the favorite competitor runs faster than 10 seconds and that the underdog competitor runs faster than 10.5 seconds, $\Pr(\{X \leq 10, Y \leq 10.5\})$.

$$\begin{aligned} \Pr(\{X \leq 10, Y \leq 10.5\}) &= \int_{9.5}^{10} \int_{10}^{10.5} f_{XY}(x, y)\, dy\, dx \\ &= \int_{9.5}^{10} \int_{10}^{10.5} 1\, dy\, dx \\ &= \int_{9.5}^{10} 0.5\, dx \\ &= 0.5 \times 0.5 \\ &= 0.25 \end{aligned}$$

Expectations

- Consider the average or expected value that some function, $g(X, Y)$, of the joint random variables. Calculate $\mathrm{E}[g(X, Y)]$.
- Examples of useful functions $g(x, y)$:
  - Expected value of X:

$$g(x, y) = x \implies \mathrm{E}[g(X, Y)] = \mathrm{E}[X]$$

  - Average difference between X and Y:

$$g(x, y) = x - y \implies \mathrm{E}[g(X, Y)] = \mathrm{E}[X - Y]$$

  - Range of interest:

$$g(x, y) = \mathbb{1}\{x \leq a, y \leq b\} \implies \mathrm{E}[g(X, Y)] = \Pr(\{X \leq a, Y \leq b\})$$

  - Covariance between X and Y:

$$g(x, y) = (x - \mu_X)(y - \mu_Y) \implies \mathrm{E}[g(X, Y)] = \mathrm{cov}(X, Y)$$

Expectations

- The formula for calculating expected value is the same as before:

Discrete R.V	Continuous R.V
$\sum_{a,b \in \Omega} g(a, b) p_{XY}(a, b)$	$\int_X \int_Y g(a, b) f_{XY}(a, b)\, db\, da$

- The function is evaluated at each point in the outcome space and weighted by the probability associated with that outcome.
- By linearity of the expectation, for any two functions $g(x, y)$ and $h(x, y)$ and any $a, b \in \mathbb{R}$:

$$\mathrm{E}[a \cdot g(X, Y) + b \cdot h(X, Y)] = a\, \mathrm{E}[g(X, Y)] + b\, \mathrm{E}[h(X, Y)]$$

- For instance,

$$\mathrm{E}[aX + bY] = a\, \mathrm{E}[X] + b\, \mathrm{E}[Y]$$

Expectations

- Consider the 100m dash example again.

$$f_{XY}(x,y) = \begin{cases} 1 & \text{if } 9.5 \le x \le 10.5, 10 \le y \le 11 \\ 0 & \text{otherwise} \end{cases}$$

- Calculate $\mathrm{E}[X-Y]$, the expected difference in finishing times between the favorite competitor and the underdog:

$$\begin{aligned} \mathrm{E}[X-Y] &= \int_{9.5}^{10.5}\int_{10}^{11}(x-y)f_{XY}(x,y)\,dy\,dx \\ &= \int_{9.5}^{10.5}\int_{10}^{11} x\,dy\,dx - \int_{9.5}^{10.5}\int_{10}^{11} y\,dy\,dx \\ &= \int_{9.5}^{10.5} x\left(y\Big|_{10}^{11}\right)dx - \int_{9.5}^{10.5}\left(\frac{y^2}{2}\Big|_{10}^{11}\right)dx \\ &= 1\cdot\frac{x^2}{2}\Big|_{9.5}^{10.5} - \frac{21}{2}\cdot x\Big|_{9.5}^{10.5} \\ &= -0.5 \end{aligned}$$

Covariance

- The covariance measures how much the variables X and Y co-vary, i.e. how they vibrate together.
- The **covariance** between X and Y is:

$$\mathrm{cov}(X,Y) = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$$

- Simplify the expression:

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathrm{E}[XY - X\,\mathrm{E}[Y] - Y\,\mathrm{E}[X] + \mathrm{E}[X]\,\mathrm{E}[Y]] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] - \mathrm{E}[Y]\,\mathrm{E}[X] + \mathrm{E}[X]\,\mathrm{E}[Y] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \end{aligned}$$

- The population covariance is often denoted σ_{XY}.
- The sample covariance is often denoted s_{XY}.

Covariance

- Covariance between X and Y:

$$f_{XY}(x,y) = \begin{cases} 1 & \text{if } 9.5 \le x \le 10.5 \text{ or } 10 \le y \le 11 \\ 0 & \text{otherwise} \end{cases}$$

- We know $\mathrm{E}[X]=10$ and $\mathrm{E}[Y]=10.5$, so calculate $\mathrm{E}[XY]$:

$$\begin{aligned} \mathrm{E}[XY] &= \int_{9.5}^{10.5}\int_{10}^{11} xy f_{XY}(x,y)\,dy\,dx \\ &= \int_{9.5}^{10.5} x\int_{10}^{11} y\,dy\,dx \\ &= \int_{9.5}^{10.5} x\cdot\left(\frac{y^2}{2}\Big|_{10}^{11}\right)dx \\ &= \frac{21}{2}\cdot\left(\frac{x^2}{2}\Big|_{9.5}^{10.5}\right) = 105 \end{aligned}$$

- Thus, $\mathrm{cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = 105 - 10\cdot 10.5 = 105 - 105 = 0$.
- There is no measurable association between competitors A and B.

Covariance

- Important properties of the covariance follow from $\mathrm{cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$:
  1. **Linearity:** $\mathrm{cov}(aX,bY) = ab\;\mathrm{cov}(X,Y)$

$$\mathrm{cov}(aX,bY) = \mathrm{E}[(aX)(bY)] - \mathrm{E}[aX]\,\mathrm{E}[bY] = ab\,(\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y])$$

  2. **Symmetry:** $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$ and $\mathrm{cov}(X,X) = \mathrm{var}(X)$

$$\mathrm{cov}(X,X) = \mathrm{E}[XX] - \mathrm{E}[X]\,\mathrm{E}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{var}(X)$$

  3. **Addition Rule:** $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X,Y)$

$$\begin{aligned} \mathrm{var}(X+Y) &= \mathrm{E}[(X+Y)^2] - \mathrm{E}[(X+Y)]^2 \\ &= \mathrm{E}[X^2 + 2XY + Y^2] - (\mathrm{E}[X] + \mathrm{E}[Y])^2 \\ &= \underbrace{\mathrm{E}[X^2]}_{\mathrm{var}(X)} + 2\underbrace{\mathrm{E}[XY]}_{\mathrm{cov}(X,Y)} + \underbrace{\mathrm{E}[Y^2]}_{\mathrm{var}(Y)} - \underbrace{(\mathrm{E}[X])^2}_{\mathrm{var}(X)} - 2\underbrace{\mathrm{E}[X]\,\mathrm{E}[Y]}_{\mathrm{cov}(X,Y)} - \underbrace{(\mathrm{E}[Y])^2}_{\mathrm{var}(Y)} \\ &= \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X,Y) \end{aligned}$$

Covariance

- Useful rule:

$$\begin{aligned}\mathrm{var}(aX + bY) &= \mathrm{var}(aX) + \mathrm{var}(bY) + 2\,\mathrm{cov}(aX, bY)\\ &= a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y) + 2ab\,\mathrm{cov}(X, Y)\end{aligned}$$

Correlation

- The population **correlation coefficient** is denoted ρ:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  where

$$\sigma_{XY} = \mathrm{cov}(X, Y), \quad \sigma_X = \sqrt{\mathrm{var}(X)}, \quad \sigma_Y = \sqrt{\mathrm{var}(Y)}$$

- The sample **correlation coefficient** is denoted r:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Conditioning and Independence

- Given two joint random variables, X and Y, we may be interested in characteristics of the distribution of Y conditional on X taking a certain value.
- Conditional expectation of Y given $X = x$ $\mathrm{E}[Y|X = x]$:
- Examples:
  - The average income of college graduates:

$$\mathrm{E}[\text{Income} \mid \text{Education = College Graduate}]$$

  - The average sales price of a home with floor size 1200 sq. ft:

$$\mathrm{E}[\text{Sales Price} \mid \text{Sqft} = 1200]$$

  - The average lifespan for smokers:

$$\mathrm{E}[\text{Lifespan} \mid \text{Smoker} = 1]$$

- Knowing the conditional expectation is particularly useful for predictions when we observe the X variable before we observe the Y variable.

Conditioning and Independence

- **Conditional expectation:**
  1. Calculate the marginal probability $\Pr(X = x)$.
  2. Fix the X variable at value $X = x$.
  3. Divide by the probability that $X = x$.

	Discrete R.V	Continuous R.V
$\mathrm{E}[Y\|X = x] =$	$\dfrac{\sum_y y \cdot p_{XY}(y, x)}{\sum_y p_{XY}(x, y)}$	$\dfrac{\int_Y y \cdot f_{XY}(x, y)\,dy}{\int_Y f_{XY}(x, y)\,dy}$

  where the marginal distribution of X at value x is:

$$\Pr(X = x) = \sum_y p_{XY}(x, y) \text{ for a discrete r.v.}$$

$$f_X(x) = \int_Y f_{XY}(x, y)\,dy \text{ for a continuous r.v.}$$

- To compute the marginal distribution, X is fixed at value x, while Y is varied.

Conditioning and Independence

- Let Y be hours of sleep and X cups of coffee drunk per day. The joint pmf $p(x, y)$ is:

$p(x, y)$	1 cup	2 cups
4 hours	0	$1/6$
8 hours	$1/3$	$1/3$
12 hours	$1/6$	0

- Compute the expected number of hours of sleep for someone who drinks 2 cups of coffee.

1. Calculate the probability that one random person drinks 2 cups of coffee:
$$\Pr(X = x) = p_{XY}(x, y) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$
2. Fix $X = 2$ and calculate $\sum_y y \cdot p_{XY}(2, y)$:
$$\sum_y y \cdot p_{XY}(2, y) = 4 \cdot \frac{1}{6} + 8 \cdot \frac{1}{3} + 12 \cdot 0 = \frac{10}{3}$$
3. Calculate $\mathrm{E}[Y|X = 2]$
$$\mathrm{E}[Y|X = 2] = \frac{10}{3} \cdot \frac{2}{1} = \frac{20}{3} \approx 6.67$$

Conditioning and Independence

- If X does not help predict Y, we say that X is **independent** of Y and denote $X \perp\!\!\!\perp Y$.
- Examples of independent random variables:
  - Knowing that one coin flip came up heads doesn't help predict the next coin flip: successive coin flips are independent.
  - Knowing the numbers that came up in a game of roulette cannot help to devise a better game strategy.

Conditioning and Independence

- If X and Y are independent, $X \perp\!\!\!\perp Y$, then:
$$\Pr(a \le X \le b, c \le Y \le d) = \Pr(a \le X \le b) \cdot \Pr(c \le Y \le d) \quad \forall (a, b, c, d)$$
$$\mathrm{E}[g(Y)|X = x] = \mathrm{E}[g(Y)] \quad \forall x \in \Omega, g(\cdot) : \Omega \to \mathbb{R}$$
- Examples of variables that seem independent but may not be:
  - weather and average house prices.
  - academic achievements and average house prices.

Review

- **Multiple Random Variables:**
  - Describe multiple events whose outcomes are unknown.
  - Have probabilities that the outcomes jointly take values in arbitrary subsets of the joint sample space.
- **Expectations:**
  - As before describe the "average" value of a function of the joint random variables.
  - The covariance function is a particular expectation we are interested in as it describes how two variables "move with" each other.
- **Conditioning and Independence:**
  - The conditional expectation is the average value of Y for individuals who have $X = x$.
  - If knowing X does not give us any information on the distribution of Y we say that X and Y are independent.

The Normal Distribution

Normal Distribution

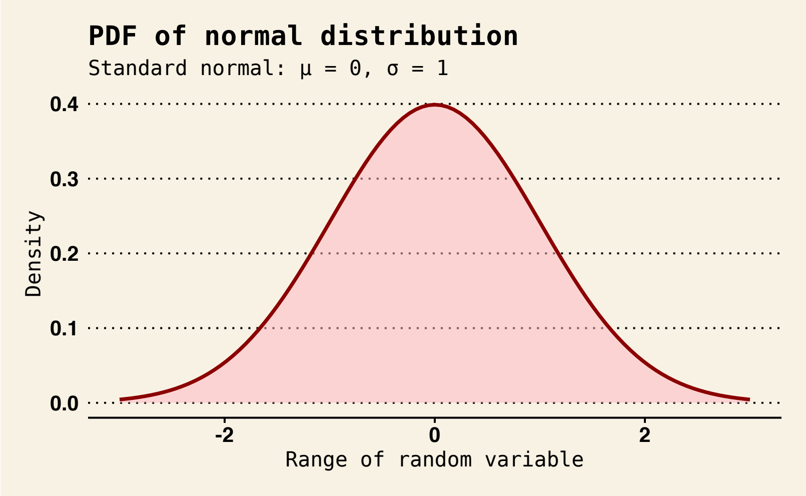
A random variable X follows a normal distribution with mean μ and variance σ^2 if it is continuously distributed with probability density function (pdf) given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

It is denoted $X \sim N(\mu, \sigma^2)$.

- **Standard normal distribution:** $Z \sim N(0,1)$. That is, $\mu = 0$ and $\sigma^2 = 1$.

The Normal Distribution



Properties of Normal Distribution

Property 1:

- If $X \sim N(\mu, \sigma^2)$ then $(X-\mu)/\sigma \sim N(0,1)$.
- Thus, we can express probabilities for any normal random variable in terms of $Z \sim N(0,1)$.
- Exercise: Show that if $X \sim N(2, 100)$ then $\Pr(X \geq 22) = \Pr(Z \geq 2)$.

$$\begin{aligned}\Pr(X \geq 22) &= \Pr\left(\frac{X-2}{10} \geq \frac{22-2}{10}\right) \\ &= \Pr(Z \geq 2)\end{aligned}$$

  Here we use the fact that $\mu = 2$ and $\sigma = \sqrt{100} = 10$.
- We calculate $\Pr(Z \geq 2)$ by looking up a table or using dedicated software.

Properties of Normal Distribution

Property 2:

- If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are jointly normal, then $W = aX + bY$ is also normally distributed for any $a, b \in \mathbb{R}$.
- Calculate $\mathrm{E}[W]$:
  - By linearity of expectation we have that $\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]$
  - So, $\mu_W = \mathrm{E}[W] = a\mu_X + b\mu_Y$.
- Calculate $\mathrm{var}(W)$:

$$\begin{aligned}&\mathrm{var}(aX + bY) = a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y) + 2ab\,\mathrm{cov}(X,Y) \\ \Longrightarrow\ &\mathrm{var}(W) = \mathrm{var}(aX + bY) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\end{aligned}$$

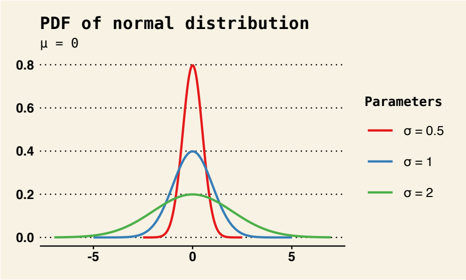
- Putting it together:

$$W \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\right)$$

Properties of Normal Distribution

Property 3:

- The distribution of $X \sim N(0, \sigma^2)$ is symmetric around zero:
$$\Pr(X \geq x) = \Pr(X \leq -x)$$
- Thus, we can compute:
$$\Pr(|X| \geq c) = \Pr(X \geq c) + \Pr(X \leq -c) = 2\Pr(X \geq c)$$



The Sample Mean

- **Example:** In an election the Green candidate is running against the Red candidate. We randomly selection $n = 100$ voters from the population and ask them who they plan on voting for. Answers are recorded as
$$X_i = \begin{cases} 1 & \text{if they plan to vote Green} \\ 0 & \text{if they plan to vote Red} \end{cases}$$
- The sample mean $\overline{X}_n$ and sample variance s_n^2 are:
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = 0.55$$
$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = 0.25$$

Questions:

- Will the Green candidate win the election?
- Would another poll confirm these results?
- How can we measure the uncertainty the sample mean?

The Sample Mean: As a Random Variable

- **Random Sampling:** $\overline{X}_n$ and s_n^2 are random variables.
- Each time a random sample is drawn, different members of the population are drawn.
- We want to use this random sample to learn about the population.
  - The sample $\{X_i\}_{i=1}^n$ is made up of n different random variables.
  - Each X_i is sampled from the same population distribution, so that

$$\mathrm{E}[X_i] = \mu_X$$
$$\mathrm{var}(X_i) = \sigma_X^2$$

The Sample Mean: As a Random Variable

- **Random sampling:** Learn about the population from one sample.

	Population	Sample
Measure of location	$\mathrm{E}[X]$	$\overline{X}_n$
Measure of dispersion	$\mathrm{var}(X)$	s_n^2

1. **Expectation:** of $\overline{X}_n$ is given $\mathrm{E}[\overline{X}_n] = \mathrm{E}[X]$.
$$\mathrm{E}[\overline{X}_n] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \mathrm{E}[X]$$
2. **Variance:** $\overline{X}_n$ is given $\mathrm{var}(\overline{X}_n) = \sigma_X^2/n$.
$$\mathrm{var}(\overline{X}_n) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X) = \sigma_X^2/n$$

- Note that $\mathrm{var}(\overline{X}_n) \to 0$ as $n \to \infty$. This is the basis of the **Law of Large Numbers** which states that $\overline{X}_n \to \mu_X$ as $n \to \infty$.

The Sample Mean: Central Limit Theorem

- **Distribution of $\overline{X}_n$:**
- Needed to make inferences about $\mathrm{E}[X]$ and compute probabilities like

$$\Pr\left(|\overline{X}_n - \mathrm{E}[X]| > 0.05\right)$$

- **Central Limit Theorem:** For n sufficiently large,

$$\overline{X}_n \sim N(\mu_x, \sigma_X^2/n)$$

- How large is "sufficiently large"?
- In practice, the central limit theorem provides a good approximation for the distribution of $\overline{X}_n$ when $n > 30$.

The Sample Mean: Central Limit Theorem

- For n sufficiently large,

$$\frac{\overline{X}_n - \mathrm{E}[\overline{X}_n]}{\sqrt{\mathrm{var}(\overline{X}_n)}} = \frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1)$$

- For n sufficiently large, we can approximate the population standard deviation σ_X with its sample estimate s_n:

$$\frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \approx \frac{\overline{X}_n - \mu_X}{s_n/\sqrt{n}} \sim N(0,1)$$

- The distribution of the sample mean $\overline{X}_n$ can be used for inference about μ_X. We can build confidence intervals and conduct hypothesis tests.

The Sample Mean: Central Limit Theorem

- Consider the polling example with $n = 100$
- What is the probability that the sample has mean $\overline{X}_n = 0.55$ and variance $s_n^2 = 0.25$, if the true proportion of Green voters in the population is $\mathrm{E}[X] = 0.5$?
- By the central limit theorem:

$$\begin{aligned}
\Pr(\overline{X}_n \geq 0.55) &= \Pr\left(\frac{\overline{X}_n - \mu_X}{s_n/\sqrt{n}} \geq \frac{0.55 - \mu_X}{s_n/\sqrt{n}}\right) \\
&= \Pr\left(\frac{\overline{X}_n - \mu_X}{s_n/\sqrt{n}} \geq \frac{0.55 - 0.5}{0.5/10}\right) \\
&= \Pr(Z \geq 1) \\
&\approx 0.159
\end{aligned}$$