M.S. Math Bootcamp: Probability & Statistics

Noah Kochanski

Department of Statistics
University of Michigan

August 20, 2024

Outline

Acknowledgements

Most of the material in these slides was adapted from *Mathematical Statistics & Data Analysis*, 3rd ed., by John A. Rice, as well as *Statistical Inference*, 2nd ed., by Casella and Berger.

Please do not distribute these slides.

Probability

Intro To Probability

- Probability is concerned with situations where outcomes are random.

Intro To Probability

- Probability is concerned with situations where outcomes are random.
- The set of all possible outcomes is called the *sample space*, which we may denote by Ω.

Intro To Probability

- Probability is concerned with situations where outcomes are random.
- The set of all possible outcomes is called the *sample space*, which we may denote by Ω.
  - Ex: Flip a coin. $\Omega = \{H, T\}$.
  - Ex: Roll a die. $\Omega = \{1, 2, 3, 4, 5, 6\}$

Intro To Probability

- Probability is concerned with situations where outcomes are random.
- The set of all possible outcomes is called the *sample space*, which we may denote by Ω.
  - Ex: Flip a coin. $\Omega = \{H, T\}$.
  - Ex: Roll a die. $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - A single element of Ω is denoted ω, e.g. $\omega \in \Omega$.

Intro To Probability

- Probability is concerned with situations where outcomes are random.
- The set of all possible outcomes is called the *sample space*, which we may denote by Ω.
    - Ex: Flip a coin. $\Omega = \{H, T\}$.
    - Ex: Roll a die. $\Omega = \{1, 2, 3, 4, 5, 6\}$
- An *event* E is a subset of Ω, i.e. $E \subseteq \Omega$.

Intro To Probability

- Probability is concerned with situations where outcomes are random.
- The set of all possible outcomes is called the *sample space*, which we may denote by Ω.
    - Ex: Flip a coin. $\Omega = \{H, T\}$.
    - Ex: Roll a die. $\Omega = \{1, 2, 3, 4, 5, 6\}$
- An *event* E is a subset of Ω, i.e. $E \subseteq \Omega$.
- Two events E_1, E_2 are said to be *disjoint* if their intersection is the empty set $E_1 \cap E_2 = \emptyset$.

Probability Measure (Informal)

A *probability measure* is a function

$$P : \{\text{subsets of } \Omega\} \to [0, 1]$$

that satisfies:

Probability Measure (Informal)

A *probability measure* is a function

$$P : \{\text{subsets of } \Omega\} \to [0, 1]$$

that satisfies:

- $P(\Omega) = 1$

Probability Measure (Informal)

A *probability measure* is a function

$$P : \{\text{subsets of } \Omega\} \to [0, 1]$$

that satisfies:

- $P(\Omega) = 1$
- $P(\emptyset) = 0$

Probability Measure (Informal)

A *probability measure* is a function

$$P : \{\text{subsets of } \Omega\} \to [0, 1]$$

that satisfies:

- $P(\Omega) = 1$
- $P(\emptyset) = 0$
- If A_1, A_2 disjoint, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ (this extends to countable sums).

Independent Events

- Given a sample space Ω and a probability measure P, we say two events E_1, E_2 are independent if

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

Independent Events

- Given a sample space Ω and a probability measure P, we say two events E_1, E_2 are independent if

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

- There can be many different probability measures associated with the same space Ω!

Independent Events

- Given a sample space Ω and a probability measure P, we say two events E_1, E_2 are independent if

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

- There can be many different probability measures associated with the same space Ω!
- However, in "reality" only one probability measure is used at any given time.

Example: Die Roll

- We have $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Consider the probability measure defined as follows:
    - $P(\{1\}) = P(\{2\}) = P(\{3\}) = \frac{1}{9}$
    - $P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{2}{9}$

Try It

- We have $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Consider the probability measure defined as follows:
  - $P(\{1\}) = P(\{2\}) = P(\{3\}) = \frac{1}{9}$
  - $P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{2}{9}$
- Using this information, we can compute the probability of any subset. Compute the probability of
  - $P(\{1, 5\})$
  - $P(\{2, 3, 6\})$

Try It

- We have $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Consider the probability measure defined as follows:
  - $P(\{1\}) = P(\{2\}) = P(\{3\}) = \frac{1}{9}$
  - $P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{2}{9}$
- Using this information, we can compute the probability of any subset. Compute the probability of
  - $P(\{1, 5\}) = P(\{1\} \cup \{5\}) = P(\{1\}) + P(\{5\}) = \frac{1}{9} + \frac{2}{9} = \frac{1}{3}$.
  - $P(\{2, 3, 6\}) = P(\{2\} \cup \{3\} \cup \{6\}) = \frac{1}{9} + \frac{1}{9} + \frac{2}{9} = \frac{4}{9}$

Conditional Probability

Let A, B be two events with $P(B) \neq 0$. Then the conditional probability of A given B is defined as

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

Conditional Probability: Example

- A coffee shop sells delicious, rare coffee (2 cups per person max). It's so good, that people who buy one cup very often buy another.

Conditional Probability: Example

- A coffee shop sells delicious, rare coffee (2 cups per person max). It's so good, that people who buy one cup very often buy another.
- Consider two experiments
    - Whether or not the person buys a cup a coffee.
    - Whether or not the person buys a second cup of coffee.

Conditional Probability: Example

- A coffee shop sells delicious, rare coffee (2 cups per person max). It's so good, that people who buy one cup very often buy another.
- Consider two experiments
    - Whether or not the person buys a cup a coffee.
    - Whether or not the person buys a second cup of coffee.
- The sample space for this sequence of two experiments is $\Omega = \{(Y, Y), (Y, N), (N, Y), (N, N)\}$

Conditional Probability: Example

- Consider two experiments
  - Whether or not the person buys a cup a coffee.
  - Whether or not the person buys a second cup of coffee.
- The sample space for each of the two experiments is $\Omega = \{(Y, Y), (Y, N), (N, Y), (N, N)\}$
- Suppose probabilities are known to be

First Cup/Second Cup	N	Y	Total
N	.5	0	.5
Y	.05	.45	.5
Total	.55	.45	1

Conditional Probability: Try It

First Cup/Second Cup	N	Y	Total
N	.5	0	.5
Y	.05	.45	.5
Total	.55	.45	1

- Let B be the event that a person buys a cup of coffee.

Conditional Probability: Try It

First Cup/Second Cup	N	Y	Total
N	.5	0	.5
Y	.05	.45	.5
Total	.55	.45	1

- Let B be the event that a person buys a cup of coffee.
- Let A be the event that a person buys a second cup of coffee.

Conditional Probability: Try It

First Cup/Second Cup	N	Y	Total
N	.5	0	.5
Y	.05	.45	.5
Total	.55	.45	1

- Let B be the event that a person buys a cup of coffee.
- Let A be the event that a person buys a second cup of coffee.
- Can you write the events A, B as subsets of Ω?

Conditional Probability: Try It

First Cup/Second Cup	N	Y	Total
N	.5	0	.5
Y	.05	.45	.5
Total	.55	.45	1

- Let B be the event that a person buys a cup of coffee.
- Let A be the event that a person buys a second cup of coffee.
- Can you write the events A, B as subsets of Ω?
- Compute $P(A \mid B)$ and $P(B \mid A)$.

Conditional Probability: Try It

First Cup/Second Cup	N	Y	Total
N	.5	0	.5
Y	.05	.45	.5
Total	.55	.45	1

- Let B be the event that a person buys a cup of coffee.
- Let A be the event that a person buys a second cup of coffee.
- Compute $P(A \mid B)$ and $P(B \mid A)$.
- $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{.45}{.50} = .9$

Conditional Probability: Try It

First Cup/Second Cup	N	Y	Total
N	.5	0	.5
Y	.05	.45	.5
Total	.55	.45	1

- Let B be the event that a person buys a cup of coffee.
- Let A be the event that a person buys a second cup of coffee.
- Compute $P(A \mid B)$ and $P(B \mid A)$.
- $P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{.45}{.45} = 1.0$

"And" Events

We can manipulate the conditional probability formula

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

to calculate

$$P(A \cap B) = P(A \mid B)P(B)$$

Law of Total Probability

- Let $B_1, \ldots, B_n$ be events such that
    - $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$

Law of Total Probability

- Let $B_1, \ldots, B_n$ be events such that
  - $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$
  - $B_i \cap B_j = \emptyset$ for all $i \neq j$.

Law of Total Probability

- Let $B_1, \ldots, B_n$ be events such that
  - $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$
  - $B_i \cap B_j = \emptyset$ for all $i \neq j$.
- Then

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

Law of Total Probability

- Let $B_1, \ldots, B_n$ be events such that
  - $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$
  - $B_i \cap B_j = \emptyset$ for all $i \neq j$.
- Then

$$
\begin{aligned}
P(A) &= \sum_{i=1}^{n} P(A \mid B_i) P(B_i) \\
&= \sum_{i=1}^{n} P(A \cap B_i)
\end{aligned}
$$

Law of Total Probability

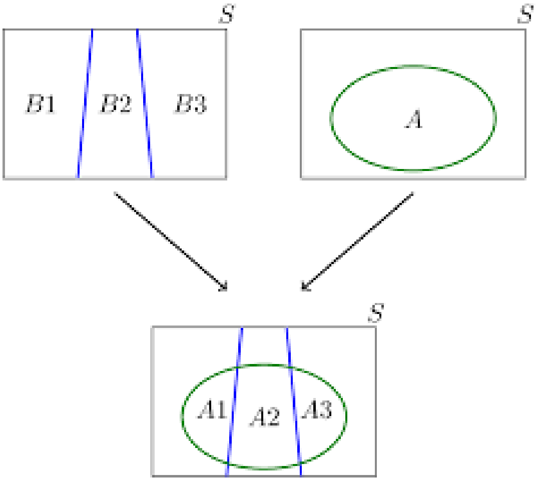


Image Source: `https://www.probabilitycourse.com/chapter1/1_4_2_total_probability.php`

Law of Total Probability: Example

- I have an urn with 7 red balls and 7 blue balls.

Law of Total Probability: Example

- I have an urn with 7 red balls and 7 blue balls.
- I pick two balls. Let R_2 be the event I pick red on my second draw. Let R_1 be the event I pick red on my first draw, B_1 be the event I pick blue on my first draw.

Law of Total Probability: Example

- I have an urn with 7 red balls and 7 blue balls.
- I pick two balls. Let R_2 be the event I pick red on my second draw. Let R_1 be the event I pick red on my first draw, B_1 be the event I pick blue on my first draw.
- The sample space for this experiment of picking two balls is

$$\Omega = \{(\text{blue}, \text{blue}), (\text{blue}, \text{red}), (\text{red}, \text{blue}), (\text{red}, \text{red})\}$$

Law of Total Probability: Example

- I have an urn with 7 red balls and 7 blue balls.
- I pick two balls. Let R_2 be the event I pick red on my second draw.
- Then

$$\begin{aligned} P(R_2) &= P(R_2 \mid B_1)P(B_1) + P(R_2 \mid R_1)P(R_1) \\ &= \frac{7}{13} \cdot \frac{1}{2} + \frac{6}{13} \cdot \frac{1}{2} \end{aligned}$$

Bayes' Rule

Bayes' Rule tells us that

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayes' Rule Derivation

Recall that

$$P(A \cap B) = P(A \mid B)P(B).$$

Bayes' Rule Derivation

Recall that

$$P(A \cap B) = P(A \mid B)P(B).$$

It follows that

$$P(B \cap A) = P(B \mid A)P(A).$$

Bayes' Rule Derivation

Recall that

$$P(A \cap B) = P(A \mid B)P(B).$$

It follows that

$$P(B \cap A) = P(B \mid A)P(A).$$

Note further that $P(A \cap B) = P(B \cap A)$.

Bayes' Rule Derivation

Recall that

$$P(A \cap B) = P(A \mid B)P(B).$$

It follows that

$$P(B \cap A) = P(B \mid A)P(A).$$

Note further that $P(A \cap B) = P(B \cap A)$. Then

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

Bayes' Rule Example

- Lie detector tests are fairly accurate on any given subject.

Bayes' Rule Example

- Lie detector tests are fairly accurate on any given subject.
- Let T denote the event that a subject tells the truth; let L denote the event that a subject lies.

Bayes' Rule Example

- Lie detector tests are fairly accurate on any given subject.
- Let T denote the event that a subject tells the truth; let L denote the event that a subject lies.
- Let $+$ be the event that the machine flags the subject, $-$ be the event that the machine doesn't.

Bayes' Rule Example

- Lie detector tests are fairly accurate on any given subject.
- Let T denote the event that a subject tells the truth; let L denote the event that a subject lies.
- Let $+$ be the event that the machine flags the subject, $-$ be the event that the machine doesn't.
- The underlying sample space being considered here is

$$\Omega = \{(\text{lie, flagged}), (\text{lie, not flagged}), (\text{truth, flagged}), (\text{truth, not flagged})\}$$

Bayes' Rule Example

- Lie detector tests are fairly accurate on any given subject.
- Let T denote the event that a subject tells the truth; let L denote the event that a subject lies.
- Let $+$ be the event that the machine flags the subject, $-$ be the event that the machine doesn't.
- Suppose that

$$P(+ \mid L) = .88$$

and

$$P(- \mid T) = .86$$

Bayes' Rule Example

- Lie detector tests are fairly accurate on any given subject.
- Let T denote the event that a subject tells the truth; let L denote the event that a subject lies.
- Let $+$ be the event that the machine flags the subject, $-$ be the event that the machine doesn't.
- Suppose that
$$P(+ \mid L) = .88$$
and
$$P(- \mid T) = .86$$
- Most people have no reason to lie, so generally $P(T) = .999$, $P(L) = .001$.

Try It

Given the information

$$\begin{aligned} P(+ \mid L) &= .88 \\ P(- \mid T) &= .86 \\ P(T) &= .999 \\ P(L) &= .001 \end{aligned}$$

Compute $P(T \mid +)$ using Bayes' Rule.

Try It

Given the information

$$
\begin{aligned}
P(+ \mid L) &= .88 \\
P(- \mid T) &= .86 \\
P(T) &= .999 \\
P(L) &= .001
\end{aligned}
$$

Compute $P(T \mid +)$ using Bayes' Rule.

$$P(T \mid +) = \frac{P(+ \mid T)P(T)}{P(+)}$$

Try It

Given the information

$$\begin{aligned} P(+ \mid L) &= .88 \\ P(- \mid T) &= .86 \\ P(T) &= .999 \\ P(L) &= .001 \end{aligned}$$

Compute $P(T \mid +)$ using Bayes' Rule.

$$P(T \mid +) = \frac{P(+ \mid T)P(T)}{P(+)} = \frac{P(+ \mid T)P(T)}{P(+ \mid L)P(L) + P(+ \mid T)P(T)}$$

Try It

Given the information

$$
\begin{aligned}
P(+ \mid L) &= .88 \implies P(- \mid L) = .12 \\
P(- \mid T) &= .86 \implies P(+ \mid T) = .14 \\
P(T) &= .999 \\
P(L) &= .001
\end{aligned}
$$

Compute $P(T \mid +)$ using Bayes' Rule.

$$
P(T \mid +) = \frac{P(+ \mid T)P(T)}{P(+)} = \frac{P(+ \mid T)P(T)}{P(+ \mid L)P(L) + P(+ \mid T)P(T)}
$$

Try It

Given the information

$$\begin{aligned} P(+ \mid L) &= .88 \implies P(- \mid L) = .12 \\ P(- \mid T) &= .86 \implies P(+ \mid T) = .14 \\ P(T) &= .999 \\ P(L) &= .001 \end{aligned}$$

Compute $P(T \mid +)$ using Bayes' Rule.

$$\frac{P(+ \mid T)P(T)}{P(+ \mid L)P(L) + P(+ \mid T)P(T)} = \frac{.14 \cdot .999}{.88 \cdot .001 + .14 \cdot .999} = .994$$

Random Variables

Random Variables

A random variable is a function

$$X : \Omega \to \mathbb{R}.$$

Random Variables

A random variable is a function

$$X : \Omega \to \mathbb{R}.$$

Note how this is different from a probability distribution, which is a function

$$P : \{\text{subsets of } \Omega\} \to [0, 1]$$

Induced Probability Distribution

Given a random variable X and a probability distribution P on sample space $\{\text{subsets of } \Omega\}$, P_X is an induced probability distribution on sample space $\mathbb{R}$

$$P_X : \{\text{subsets of } \mathbb{R}\} \to [0, 1]$$

Induced Probability Distribution

Given a random variable X and a probability measure P on sample space $\{\text{subsets of } \Omega\}$, P_X is an induced probability distribution on sample space $\mathbb{R}$

$$P_X : \{\text{subsets of } \mathbb{R}\} \to \mathbb{R}.$$

The probability measure P_X is defined as

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

Induced Probability Distribution

Given a random variable X and a probability measure P on sample space $\{\text{subsets of } \Omega\}$, P_X is the induced probability distribution

$$P_X : \{\text{subsets of } \mathbb{R}\} \to \mathbb{R}.$$

The probability measure P_X is defined as

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

This notation is really confusing, though, so we just write $P(X \in A)$. We call P_X the *distribution of* X.

Example

- Consider rolling a fair die. Clearly $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the probability measure assigns even probability to each outcome.

Example

- Consider rolling a fair die. Clearly $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the probability measure assigns even probability to each outcome.
- Let the random variable $X : \Omega \to \mathbb{R}$ be given by $X(\omega) = \text{remainder of } \omega \text{ divided by } 4$.

Example

- Consider rolling a fair die. Clearly $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the probability measure assigns even probability to each outcome.
- Let the random variable $X : \Omega \to \mathbb{R}$ be given by $X(\omega) = \text{remainder of } \omega \text{ divided by } 4$.
- What is $P(X \in [2, 3])$?

Example

- Consider rolling a fair die. Clearly $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the probability measure assigns even probability to each outcome.
- Let the random variable $X : \Omega \to \mathbb{R}$ be given by $X(\omega) = \text{remainder of } \omega \text{ divided by } 4$.
- What is $P(X \in [2, 3])$?
- Let's evaluate X. We have
  - X(1) $=$ 1, X(2) $=$ 2, X(3) $=$ 3, X(4)$=$0, X(5)$=$1, X(6)$=$2

Example

- Consider rolling a fair die. Clearly $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the probability measure assigns even probability to each outcome.
- Let the random variable $X : \Omega \to \mathbb{R}$ be given by $X(\omega) = \text{remainder of } \omega \text{ divided by } 4$.
- What is $P(X \in [2, 3])$?
- Let's evaluate X. We have
  - X(1) = 1, X(2) = 2, X(3) = 3, X(4)=0, X(5)=1, X(6)=2
- So $\omega = 2, 3, 6$ result in $X \in [2, 3]$.

Example

- Consider rolling a fair die. Clearly $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the probability measure assigns even probability to each outcome.
- Let the random variable $X : \Omega \to \mathbb{R}$ be given by $X(\omega) = \text{remainder of } \omega \text{ divided by } 4$.
- What is $P(X \in [2, 3])$?
- Let's evaluate X. We have
    - X(1) = 1, X(2) = 2, X(3) = 3, X(4)=0, X(5)=1, X(6)=2
- So $\omega = 2, 3, 6$ result in $X \in [2, 3]$.
- Therefore

$$P_X([2, 3]) = P(\{2, 3, 6\}) = 0.5$$

Try It

Consider a sample space $\Omega = \{1, 2, 3, 4\}$ with probability measure

$$
\begin{aligned}
P(\{1\} &= .5 \\
P(\{2\}) &= .25 \\
P(\{3\}) &= .125 \\
P(\{4\}) &= .125
\end{aligned}
$$

Try It

Consider a sample space $\Omega = \{1, 2, 3, 4\}$ with probability measure

$$\begin{aligned} P(\{1\} &= .5 \\ P(\{2\}) &= .25 \\ P(\{3\}) &= .125 \\ P(\{4\}) &= .125 \end{aligned}$$

Let $X(\omega) = \omega^2 - 5$.

Find $P(X \in [4, \infty))$.

Try It

Consider a sample space $\Omega = \{1, 2, 3, 4\}$ with probability measure

$$\begin{aligned} P(\{1\}) &= .5 \\ P(\{2\}) &= .25 \\ P(\{3\}) &= .125 \\ P(\{4\}) &= .125 \end{aligned}$$

Let $X(\omega) = \omega^2 - 5$.

Find $P(X \in [4, \infty))$.
Only $\omega = 3$ and $\omega = 4$ result in $X(\omega) \in [4, \infty)$. So we want $P(\{3, 4\}) = .25$

Cumulative Distribution Function

- The *cumulative distribution function* or *cdf* of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(c) = P(X \leq c) = P(X \in (-\infty, c))$$

Cumulative Distribution Function

- The *cumulative distribution function* or *cdf* of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(c) = P(X \leq c) = P(X \in (-\infty, c))$$

- The cdf completely characterizes the distribution of X.

Cumulative Distribution Function

- The *cumulative distribution function* or *cdf* of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(c) = P(X \leq c) = P(X \in (-\infty, c))$$

- The cdf completely characterizes the distribution of X.
- The cdf is monotonically increasing.

Cumulative Distribution Function

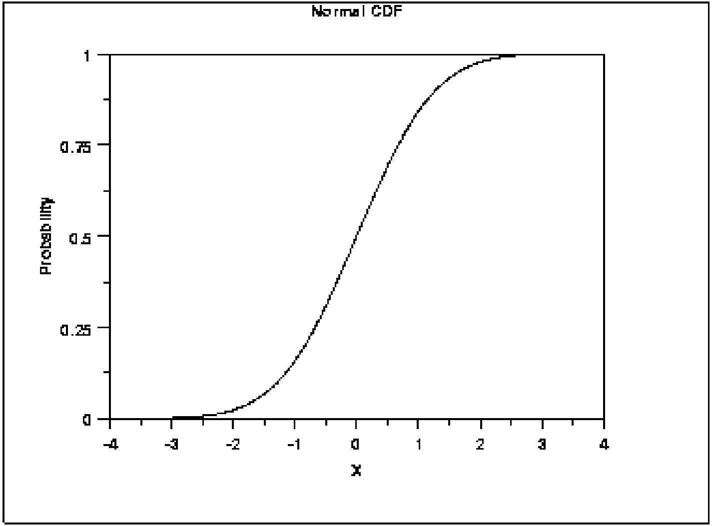


Image Source: https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

Discrete Random Variables

Abstract $\to$ Practical

- The underlying probability space Ω is a useful theoretical concept.

Abstract $\rightarrow$ Practical

- The underlying probability space Ω is a useful theoretical concept.
- However, in practice we don't care about Ω because we don't observe the results of the underlying experiment.

Abstract $\rightarrow$ Practical

- The underlying probability space Ω is a useful theoretical concept.
- However, in practice we don't care about Ω because we don't observe the results of the underlying experiment.
- We only observe random variable X, and so we only care about the distribution of X. From now on, P will denote distributions of a random variable.

Abstract $\rightarrow$ Practical

- $\Omega = \{\text{all possible underlying financial market conditions tomorrow}\}$

Abstract $\to$ Practical

- $\Omega = \{\text{all possible underlying financial market conditions tomorrow}\}$
- $X : \Omega \to \mathbb{R}$ is a function that returns the stock price of Nvidia at close tomorrow. In other words, $X(\omega)$ is a deterministic function of ω.

Discrete Random Variables

- Some random variable X only take on finitely (or countably many) values like $1, 2, 3, 4, \ldots$. These are called *discrete random variables*.
- Most discrete random variables *count* the number of occurrences of something.

Bernoulli Random Variable

- A *Bernoulli* random variable X with parameter p is such that

$$\begin{cases} P(X=1) = p \\ P(X=0) = 1-p \end{cases}$$

Binomial Random Variable

- Suppose that n independent trials are performed with each experiment resulting in a success or failure. The probability of an individual success is p.

Binomial Random Variable

- Suppose that n independent trials are performed with each experiment resulting in a success or failure. The probability of an individual success is p.
- The random variable $X = \#\text{ of successes}$ has a *binomial distribution*.

Binomial Random Variable

- Suppose that n independent trials are performed with each experiment resulting in a success or failure. The probability of an individual success is p.
- The random variable $X = \# \text{ of successes}$ has a *binomial distribution*.
- A binomial distribution has

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Poisson Random Variable

A Poisson random variable with parameter $\lambda > 0$ has distribution

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

for $k = 1, \ldots, \infty$.

Continuous Random Variables

Continuous Random Variables

- More generally, a random variable X can take on a continuous range of values in $\mathbb{R}$.
  - The height of a random person (in inches).
  - The value of company (in dollars)
  - The degree of improvement of a pharmaceutical drug.

Continuous Random Variables

- More generally, a random variable X can take on a continuous range of values in $\mathbb{R}$.
  - The height of a random person (in inches).
  - The value of company (in dollars)
  - The degree of improvement of a pharmaceutical drug.
- In situations like these, we want to know $P(X \in [a, b])$, $P(X > a)$, $P(X < b)$, etc.

Density

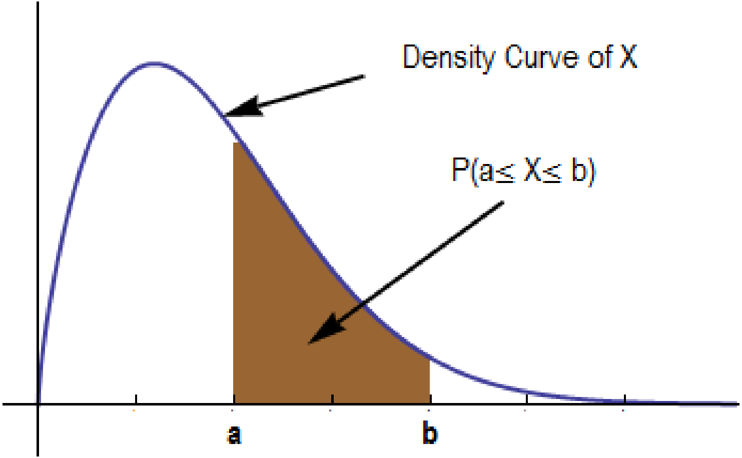
- The distribution of a continuous random variable X is often given to us in the form of a **density** function which we call $p(x)$.

Density

- The distribution of a continuous random variable X is often given to us in the form of a **density** function which we call $p(x)$.
- The density $p(x)$ gives probabilities via the relation

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

Why Area?



Food For Thought

Given density curve $p(x)$,

- How would I calculate $P(X > b)$?

Food For Thought

Given density curve $p(x)$,

- How would I calculate $P(X > b)$?
- How could I calculate $P(X < a)$?

Food For Thought

Given density curve $p(x)$,

- How would I calculate $P(X > b)$?
- How could I calculate $P(X < a)$?
- What is $\int_{-\infty}^{\infty} p(x)dx$?

Food For Thought

Given density curve $p(x)$,

- How would I calculate $P(X > b)$? $\int_b^\infty p(x)dx$
- How could I calculate $P(X < a)$?
- What is $\int_{-\infty}^{\infty} p(x)dx$?

Food For Thought

Given density curve $p(x)$,

- How would I calculate $P(X > b)$? $\int_b^\infty p(x)dx$
- How could I calculate $P(X < a)$? $\int_{-\infty}^a p(x)dx$
- What is $\int_{-\infty}^\infty p(x)dx$?

Food For Thought

Given density curve $p(x)$,

- How would I calculate $P(X > b)$? $\int_b^\infty p(x)dx$
- How could I calculate $P(X < a)$? $\int_{-\infty}^a p(x)dx$
- What is $\int_{-\infty}^\infty p(x)dx$? 1

Uniform Random Variable

A *uniform* random variable on the inteveral $[a, b]$ has density function

$$p(x) = \begin{cases} \frac{1}{b-a}, x \in [a, b] \\ 0, x \notin [a, b] \end{cases}$$

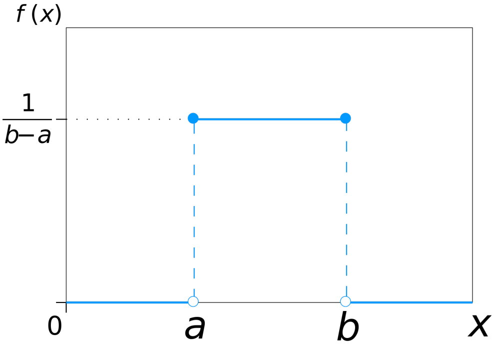


Image Source: `https://en.wikipedia.org/wiki/Continuous_uniform_distribution`

Try It

A *uniform* random variable on the interval $[a, b]$ has density function

$$p(x) = \begin{cases} \frac{1}{b-a}, x \in [a, b] \\ 0, x \notin [a, b] \end{cases}$$

Suppose that X is a uniform random variable on $[3, 10]$. Find

$$P(X > 5)$$

Try It

A *uniform* random variable on the interval $[a, b]$ has density function

$$p(x) = \begin{cases} \frac{1}{b-a}, x \in [a, b] \\ 0, x \notin [a, b] \end{cases}$$

Suppose that X is a uniform random variable on $[3, 10]$. Find

$$P(X > 5)$$

$$P(X > 5) = \int_5^\infty p(x)dx = \int_5^{10} p(x)$$

Try It

Suppose that X is a uniform random variable on $[3, 10]$. Find

$$P(X > 5)$$

$$\begin{aligned} P(X > 5) &= \int_5^\infty p(x)dx = \int_5^{10} p(x) \\ &= \int_5^{10} \frac{1}{7} dx \\ &= \frac{1}{7} x \Big|_5^{10} \end{aligned}$$

Try It

Suppose that X is a uniform random variable on $[3, 10]$. Find

$$P(X > 5)$$

$$\begin{aligned} P(X > 5) &= \int_5^\infty p(x)dx = \int_5^{10} p(x) \\ &= \int_5^{10} \frac{1}{7} dx \\ &= \frac{1}{7} x \Big|_5^{10} \\ &= \frac{10}{7} - \frac{5}{7} = \frac{5}{7} \end{aligned}$$

Note

How do the following quantities differ?

$$P(X > 5) \qquad \text{vs} \qquad P(X \geq 5)$$

Exponential Random Variable

An *exponential* random variable (with parameter $\lambda > 0$) has density function

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases}$$

Try It

An *exponential* random variable (with parameter $\lambda > 0$) has density function

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases}$$

Suppose $\lambda = 2$. Calculate $P(X \in [0, 2])$.

Try It

An *exponential* random variable (with parameter $\lambda > 0$) has density function

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases}$$

Suppose $\lambda = 2$. Calculate $P(X \in [0, 2])$.

$$\int_0^2 2e^{-2x} dx = -e^{-2x}|_0^2 = -e^{-4} + 1$$

Normal Random Variable

A *Gaussian* or *normal* random variable X with parameters μ, σ has density function

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

Normal Random Variable

A *Gaussian* or *normal* random variable X with parameters μ, σ has density function

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

We often denote such a random variable as $X \sim \mathcal{N}(\mu, \sigma^2)$.

The Distribution Zoo

`https://ben18785.shinyapps.io/distribution-zoo/`

Expectation

Discrete R.V's

The expectation of a discrete random variable is defined as

$$\mathbb{E}(X) = \sum_k kP(X = k).$$

Discrete R.V's

The expectation of a discrete random variable is defined as

$$\mathbb{E}(X) = \sum_k kP(X = k).$$

We also call $\mathbb{E}(X)$ the *mean* of X, which is often denoted μ.

Discrete R.V Example

The expectation of a discrete random variable is defined as

$$\mathbb{E}(X) = \sum_k kP(X = k).$$

Suppose X has distribution

$$\begin{aligned} P(X = 1) &= .5 \\ P(X = 2) &= .25 \\ P(X = 3) &= .2 \\ P(X = 4) &= .05. \end{aligned}$$

Discrete R.V Example

The expectation of a discrete random variable is defined as

$$\mathbb{E}(X) = \sum_k kP(X = k).$$

Suppose X has distribution

$$\begin{aligned} P(X = 1) &= .5 \\ P(X = 2) &= .25 \\ P(X = 3) &= .2 \\ P(X = 4) &= .05. \end{aligned}$$

Then $\mathbb{E}(X) = 1(.5) + 2(.25) + 3(.2) + 4(.05) = 1.8$

Try It

Suppose that a state lottery scratch-off game sells 99.9% losing cards, and .1% winning cards.

Try It

Suppose that a state lottery scratch-off game sells 99.9% losing cards, and .1% winning cards.

The payout for a winning ticket is 1000, and losing ticket is 0.

Try It

Suppose that a state lottery scratch-off game sells 99.9% losing cards, and .1% winning cards.

The payout for a winning ticket is 1000, and losing ticket is 0.

Suppose each ticket costs \$5. Let the random variable X be the *profit* earned from a random scratch-off. What is $\mathbb{E}(X)$?

Try It

Suppose that a state lottery scratch-off game sells 99.9% losing cards, and .1% winning cards.

The payout for a winning ticket is 1000, and losing ticket is 0.

Suppose each ticket costs \$5. Let the random variable X be the *profit* earned from a random scratch-off. What is $\mathbb{E}(X)$?

$\mathbb{E}(X) = .999(-5) + .001(1000) = -\$3.995$

Continuous R.V's

If X is a continuous random variable with density $p(x)$, then we define

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx$$

Try It

If X is a continuous random variable with density $p(x)$, then we define

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx$$

What is the mean of a uniform random variable on $[0, 10]$?

Try It

If X is a continuous random variable with density $p(x)$, then we define

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx$$

What is the mean of a uniform random variable on $[0, 10]$?

$\mathrm{E(X)} = \int_{-\infty}^{\infty} x\frac{1}{10}dx = \int_{0}^{10} x\frac{1}{10}dx = \frac{x^2}{20}\Big|_0^{10} = 5$

Expectation of Functions of R.V's

- Suppose X is a random variable, and $f : \mathbb{R} \to \mathbb{R}$ is a function.
- What is $\mathbb{E}(f(X))$?

Expectation of Functions of R.V's

- Suppose X is a random variable, and $f : \mathbb{R} \to \mathbb{R}$ is a function.
- What is $\mathbb{E}(f(X))$?
- For a discrete r.v., we have

$$\mathbb{E}(f(X)) = \sum_k f(k)P(X = k)$$

Expectation of Functions of R.V's

- Suppose X is a random variable, and $f : \mathbb{R} \to \mathbb{R}$ is a function.
- What is $\mathbb{E}(f(X))$?
- For a discrete r.v., we have

$$\mathbb{E}(f(X)) = \sum_k f(k)P(X = k)$$

- For a continuous r.v., we have

$$\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} f(x)p(x)dx$$

Example

- I have a model that generates a random variable X which represents a stock's price tomorrow.

Example

- I have a model that generates a random variable X which represents a stock's price tomorrow.
- Suppose tomorrow comes and the price turns out to be \$70.

Example

- I have a model that generates a random variable X which represents a stock's price tomorrow.
- Suppose tomorrow comes and the price turns out to be \$70.
- I may want to calculate $\mathbb{E}((X - 70)^2)$, i.e. the mean squared error of my estimate, to improve my model for the future.

Try It

- Let X be a discrete random variable on $0, 1, 2$ with probabilities $\frac{1}{2}, \frac{3}{8}, \frac{1}{8}$, respectively.

Try It

- Let X be a discrete random variable on $0, 1, 2$ with probabilities $\frac{1}{2}, \frac{3}{8}, \frac{1}{8}$, respectively.
- Calculate $\mathbb{E}(X^2)$.

Try It

- Let X be a discrete random variable on $0, 1, 2$ with probabilities $\frac{1}{2}, \frac{3}{8}, \frac{1}{8}$, respectively.
- Calculate $\mathbb{E}(X^2)$.

$$\mathbb{E}(X^2) = 0^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{1}{8} = \frac{7}{8}$$

Linearity of Expectation

- The expectation operation $\mathbb{E}$ is linear.

Linearity of Expectation

- The expectation operation $\mathbb{E}$ is linear.
- More explicitly, this means that

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

Linearity of Expectation

- The expectation operation $\mathbb{E}$ is linear.
- More explicitly, this means that

$$\begin{aligned} \mathbb{E}(X + Y) &= \mathbb{E}(X) + \mathbb{E}(Y) \\ \mathbb{E}(\alpha X) &= \alpha \mathbb{E}(X) \end{aligned}$$

Linearity of Expectation

- The expectation operation $\mathbb{E}$ is linear.
- More explicitly, this means that

$$\begin{aligned} \mathbb{E}(X + Y) &= \mathbb{E}(X) + \mathbb{E}(Y) \\ \mathbb{E}(\alpha X) &= \alpha \mathbb{E}(X) \end{aligned}$$

- Why is this? Because integration is linear.

Linearity of Expectation

- The expectation operation $\mathbb{E}$ is linear.
- More explicitly, this means that

$$\begin{aligned}\mathbb{E}(X+Y) &= \mathbb{E}(X)+\mathbb{E}(Y)\\ \mathbb{E}(\alpha X) &= \alpha\mathbb{E}(X)\end{aligned}$$

- Why is this? Because integration is linear.

$$\int_a^b \alpha f(x) + \beta g(x) dx = \alpha \int_a^b f(x) + \beta \int_a^b g(x)$$

Variance

The *variance* of a random variable X is defined as

$$\mathrm{Var}(X) = \mathbb{E}\big[X - \mathbb{E}(X)\big]^2$$

Variance

The *variance* of a random variable X is defined as

$$\mathrm{Var}(X) = \mathbb{E}\big[X - \mathbb{E}(X)\big]^2$$

In words, the variance of a r.v. is "*the average squared distance between X and its mean*".

Variance

The *variance* of a random variable X is defined as

$$\mathrm{Var}(X) = \mathbb{E}\big[X - \mathbb{E}(X)\big]^2$$

There's a lot of symbols in the formula above. It's easier to remember that $\mathbb{E}(X)$ is just a constant, μ, and rewrite as

$$\mathrm{Var}(X) = \mathbb{E}\big[X - \mu\big]^2$$

Joint Distributions

Joint Distribution

Often we have multiple random variables, e.g. X, Y, whose values we want to consider together.

Joint Distribution

Often we have multiple random variables, e.g. X, Y, whose values we want to consider together.

We're interested in quantities of the form

$$P(X = x, Y = y)$$

for discrete random variables and

$$P(X \in [a, b], Y \in [c, d])$$

for continuous random variables.

Joint Distribution

The notation with the comma

$$P(X = x, Y = y), P(X \in [a, b], Y \in [c, d])$$

is shorthand for denoting that both of these occur simultaneously.

Joint Density

Continuous random variables have a *joint density* $p(x, y)$ that satisfies

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) dxdy = 1$$

Joint Density

Continuous random variables have a *joint density* $p(x, y)$ that satisfies

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) dxdy = 1$$

and

$$P((X, Y) \in A) = \int \int_A p(x, y) dxdy$$

for any set $A \subseteq \mathbb{R}^2$.

Independence

Two discrete r.v.'s X, Y are *independent* if

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Independence

Two discrete r.v.'s X, Y are *independent* if

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Two continuous r.v.'s X, Y are independent if their joint density factors into a product of their marginal densities, i.e.

$$p(x, y) = p(x)p(y).$$

Expectation

For any function $g : \mathbb{R}^2 \to \mathbb{R}$ we define

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) p(x, y) dx dy$$

Covariance

The *covariance* between two random variables X, Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big]$$

Covariance

The *covariance* between two random variables X, Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big]$$

Fact: If random variables X, Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Parameter Estimation

Statistical Estimation (Informal)

Classical statistical frameworks consist of three main components.

1. A *parameter* θ, typically a population quantity of interest for the researcher.

Statistical Estimation (Informal)

Classical statistical frameworks consist of three main components.

1. A *parameter* θ, typically a population quantity of interest for the researcher.
2. A *model* $p(x \mid \theta)$, i.e. a distribution that depends on θ.

Statistical Estimation (Informal)

Classical statistical frameworks consist of three main components.

1. A *parameter* θ, typically a population quantity of interest for the researcher.
2. A *model* $p(x \mid \theta)$, i.e. a distribution that depends on θ.
3. Observed *data* $x_1, \ldots, x_n$ which are assumed to be draws of random variables $X_1, \ldots, X_n$ that follow the model.

Statistical Estimation (Informal)

Classical statistical frameworks consist of three main components.

1. A *parameter* θ, typically a population quantity of interest for the researcher.
2. A *model* $p(x \mid \theta)$, i.e. a distribution that depends on θ.
3. Observed *data* which are assumed to be draws of random variables $X_1, \ldots, X_n$ that follow the model.

An *estimator* $\hat{\theta}(X_1, \ldots, X_n)$ is a function of the data whose value we hope is close to θ.

Example: Polling

1. Let p denote the proportion of U.S. eligible voters who support term limits for politicians.

Example: Polling

1. Let p denote the proportion of U.S. eligible voters who support term limits for politicians.
2. If we pick an eligible voter at random from the population, their support on this issue can be modeled as $\mathrm{Bernoulli}(p)$.

Example: Polling

1. Let p denote the proportion of U.S. eligible voters who support term limits for politicians.
2. If we pick an eligible voter at random from the population, their support on this issue can be modeled as $\mathrm{Bernoulli}(p)$.
3. We sample 1,078 eligible voters, and find that 670 support term limits. A natural estimator is

$$\hat{p} = \frac{\#\text{ in favor}}{\#\text{ polled}} = \frac{670}{1078} \approx .622$$

Example: Quality Control

1. Let μ be the true mean number of calories in a can of Diet Coke produced by Coca-Cola.

Example: Quality Control

1. Let μ be the true mean number of calories in a can of Diet Coke produced by Coca-Cola.
2. With known standard deviation σ, assume a randomly selected can has a number of calories distributed as $\mathcal{N}(\mu, \sigma^2)$.

Example: Quality Control

1. Let μ be the true mean number of calories in a can of Diet Coke produced by Coca-Cola.
2. With known standard deviation σ, assume a randomly selected can has a number of calories distributed as $\mathcal{N}(\mu, \sigma^2)$.
3. The FDA samples 178 cans for testing, and computes the number of calories in each with a high-precision instrument. A natural estimator is

$$\hat{\mu} = \frac{1}{178} \sum_{i=1}^{178} X_i$$

   where X_i is the number of calories in can number i.

Observations About Estimators (Informal)

(Informal) Definition: An *estimator* $\hat{\theta}$ is a real-valued function of random variable X, i.e.

$$\hat{\theta} : \mathbb{R} \to \mathbb{R}$$

Observations About Estimators (Informal)

(Informal) Definition: An *estimator* $\hat{\theta}$ is a real-valued function of random variable X, i.e.

$$\hat{\theta} : \mathbb{R} \to \mathbb{R}$$

Here's an alternative perspective. As $X : \Omega \to \mathbb{R}$, an estimator $\hat{\theta}(X)$ is the *composition of functions* that is itself a random variable:

$$\hat{\theta}(X) : \Omega \to \mathbb{R}$$

We can thus talk about the distribution, expectation, variance etc. of the estimator $\hat{\theta}(X)$.

Lots of Data?

- If we have many random variables, suppose each $X_i : \Omega \to \mathbb{R}$ for $i = 1, \ldots, n$. Then the estimator

$$\hat{\theta} : \mathbb{R}^n \to \mathbb{R}$$

takes in all n numbers and outputs a single number.

Lots of Data?

- If we have many random variables, suppose each $X_i : \Omega \to \mathbb{R}$ for $i = 1, \ldots, n$. Then the estimator

$$\hat{\theta} : \mathbb{R}^n \to \mathbb{R}$$

  takes in all n numbers and outputs a single number.
- Define a new underlying sample space $\tilde{\Omega} = \Omega \times \Omega \times \cdots \times \Omega$. Then

$$\hat{\theta}(X_1, \ldots, X_n) : \tilde{\Omega} \to \mathbb{R}$$

  is a random variable defined on this new sample space.

Questions To Ask of An Estimator

1. Do we have $\mathbb{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$?

Questions To Ask of An Estimator

1. Do we have $\mathbb{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$?
2. How large is the variance of $\hat{\theta}(X_1, \ldots, X_n)$?

Questions To Ask of An Estimator

1. Do we have $\mathbb{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$?
2. How large is the variance of $\hat{\theta}(X_1, \ldots, X_n)$?
3. What is $P(|\hat{\theta} - \theta| > \epsilon)$?

Example: Quality Control

1. Let μ be the true mean number of calories in a can of Diet Coke produced by Coca-Cola. If the FDA determines $\mu < 5$, Coca-Cola gets to say it's zero calories.

Example: Quality Control

1. Let μ be the true mean number of calories in a can of Diet Coke produced by Coca-Cola. If the FDA determines $\mu < 5$, Coca-Cola gets to say it's zero calories.
2. With known standard deviation σ, assume a randomly selected can has a number of calories distributed as $\mathcal{N}(\mu, \sigma^2)$.

Example: Quality Control

1. Let μ be the true mean number of calories in a can of Diet Coke produced by Coca-Cola. If the FDA determines $\mu < 5$, Coca-Cola gets to say it's zero calories.
2. With known standard deviation σ, assume a randomly selected can has a number of calories distributed as $\mathcal{N}(\mu, \sigma^2)$.
3. The FDA samples 178 cans for testing, and computes the number of calories in each with a high-precision instrument. Call these numbers $X_1, \ldots, X_{178}$.

Example: Quality Control

1. Let μ be the true mean number of calories in a can of Diet Coke produced by Coca-Cola. If the FDA determines $\mu < 5$, Coca-Cola gets to say it's zero calories.
2. With known standard deviation σ, assume a randomly selected can has a number of calories distributed as $\mathcal{N}(\mu, \sigma^2)$.
3. The FDA samples 178 cans for testing, and computes the number of calories in each with a high-precision instrument. Call these numbers $X_1, \ldots, X_{178}$.
4. We will consider the estimator

$$\hat{\mu} = \frac{1}{178} \sum_{i=1}^{178} X_i.$$

Try It: Questions About $\hat{\mu}$

1. Assuming each $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and the random variables are independent, for $\hat{\mu} = \frac{1}{178}\sum_{i=1}^{178} X_i$ calculate $\mathbb{E}(\hat{\mu})$.
2. Calculate $\mathrm{Var}(\hat{\mu})$ (you will need to use the fact that for independent random variables Y, Z, we have $\mathrm{Var}(\alpha Y + \beta Z) = \alpha^2 \mathrm{Var}(Y) + \beta^2 \mathrm{Z}.$)

Try It

Assuming each $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and the random variables are independent, for $\hat{\mu} = \frac{1}{178}\sum_{i=1}^{178} X_i$ calculate $\mathbb{E}(\hat{\mu})$.
By linearity, we have

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(\frac{1}{178}\sum_{i=1}^{178} X_i)$$

Try It

Assuming each $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and the random variables are independent, for $\hat{\mu} = \frac{1}{178}\sum_{i=1}^{178} X_i$ calculate $\mathbb{E}(\hat{\mu})$.

By linearity, we have

$$\begin{aligned}
\mathbb{E}(\hat{\mu}) &= \mathbb{E}(\frac{1}{178}\sum_{i=1}^{178} X_i) \\
&= \frac{1}{178}\mathbb{E}(\sum_{i=1}^{178} X_i)
\end{aligned}$$

Try It

Assuming each $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and the random variables are independent, for $\hat{\mu} = \frac{1}{178}\sum_{i=1}^{178} X_i$ calculate $\mathbb{E}(\hat{\mu})$.

By linearity, we have

$$\begin{aligned}\mathbb{E}(\hat{\mu}) &= \mathbb{E}(\frac{1}{178}\sum_{i=1}^{178} X_i) \\ &= \frac{1}{178}\mathbb{E}(\sum_{i=1}^{178} X_i) \\ &= \frac{1}{178}\sum_{i=1}^{178}\mathbb{E}(X_i)\end{aligned}$$

Try It

Assuming each $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and the random variables are independent, for $\hat{\mu} = \frac{1}{178} \sum_{i=1}^{178} X_i$ calculate $\mathbb{E}(\hat{\mu})$.

By linearity, we have

$$\begin{aligned}
\mathbb{E}(\hat{\mu}) &= \mathbb{E}(\frac{1}{178} \sum_{i=1}^{178} X_i) \\
&= \frac{1}{178} \mathbb{E}(\sum_{i=1}^{178} X_i) \\
&= \frac{1}{178} \sum_{i=1}^{178} \mathbb{E}(X_i) \\
&= \frac{1}{178} \sum_{i=1}^{178} \mu = \mu
\end{aligned}$$

Try It

Calculate $\mathrm{Var}(\hat{\mu})$ (you will need to use the fact that for independent random variables Y, Z, we have $\mathrm{Var}(\alpha Y + \beta Z) = \alpha^2 \mathrm{Var}(Y) + \beta^2 \mathrm{Z}$, the variance of a $\mathcal{N}(\mu, \sigma^2)$ is σ^2.)

Using the fact given

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(\frac{1}{178} \sum_{i=1}^{178} X_i)$$

$$= \mathrm{Var}\left(\frac{X_1}{178} + \frac{X_2}{178} + \cdots + \frac{X_{178}}{178}\right)$$

Try It

Calculate $\mathrm{Var}(\hat{\mu})$ (you will need to use the fact that for independent random variables Y, Z, we have $\mathrm{Var}(\alpha Y + \beta Z) = \alpha^2 \mathrm{Var}(Y) + \beta^2 \mathrm{Z}$, the variance of a $\mathcal{N}(\mu, \sigma^2)$ is σ^2.)
Using the fact given

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= \mathrm{Var}(\frac{1}{178}\sum_{i=1}^{178} X_i) \\
&= \mathrm{Var}\left(\frac{X_1}{178} + \frac{X_2}{178} + \cdots + \frac{X_{178}}{178}\right) \\
&= \frac{1}{178^2}\mathrm{Var}(X_1) + \cdots + \frac{1}{178^2}\mathrm{Var}(X_{178})
\end{aligned}
$$

Try It

Calculate $\mathrm{Var}(\hat{\mu})$ (you will need to use the fact that for independent random variables Y, Z, we have $\mathrm{Var}(\alpha Y + \beta Z) = \alpha^2 \mathrm{Var}(Y) + \beta^2 \mathrm{Z}$, the variance of a $\mathcal{N}(\mu, \sigma^2)$ is σ^2.)

Using the fact given

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= \mathrm{Var}(\frac{1}{178}\sum_{i=1}^{178} X_i) \\
&= \mathrm{Var}\left(\frac{X_1}{178} + \frac{X_2}{178} + \cdots + \frac{X_{178}}{178}\right) \\
&= \frac{1}{178^2}\mathrm{Var}(X_1) + \cdots + \frac{1}{178^2}\mathrm{Var}(X_{178}) \\
&= \frac{1}{178^2}\sigma^2 + \cdots + \frac{1}{178^2}\sigma^2
\end{aligned}
$$

Try It

Calculate $\mathrm{Var}(\hat{\mu})$ (you will need to use the fact that for independent random variables Y, Z, we have $\mathrm{Var}(\alpha Y + \beta Z) = \alpha^2 \mathrm{Var}(Y) + \beta^2 \mathrm{Z}$, the variance of a $\mathcal{N}(\mu, \sigma^2)$ is σ^2.)

Using the fact given

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= \mathrm{Var}\left(\frac{1}{178}\sum_{i=1}^{178} X_i\right) \\
&= \mathrm{Var}\left(\frac{X_1}{178} + \frac{X_2}{178} + \cdots + \frac{X_{178}}{178}\right) \\
&= \frac{1}{178^2}\mathrm{Var}(X_1) + \cdots + \frac{1}{178^2}\mathrm{Var}(X_{178}) \\
&= \frac{1}{178^2}\sigma^2 + \cdots + \frac{1}{178^2}\sigma^2 \\
&= 178\frac{\sigma^2}{178^2} = \frac{\sigma^2}{178}
\end{aligned}$$

Likelihood Function

- Let X be a continuous random variable with density function p parameterized by parameter θ.

Likelihood Function

- Let X be a continuous random variable with density function p parameterized by parameter θ.
    - Denoted $p_\theta(x)$.

Likelihood Function

- Let X be a continuous random variable with density function p parameterized by parameter θ.
  - Denoted $p_\theta(x)$.
- The *likelihood function* is defined as

$$\mathcal{L}(\theta) = p_\theta(x)$$

Likelihood Function

- Let X be a continuous random variable with density function p parameterized by parameter θ.
  - Denoted $p_\theta(x)$.
- The *likelihood function* is defined as

$$\mathcal{L}(\theta) = p_\theta(x)$$

  - Viewed as a function of θ, rather than x.

Problem Formulation

- Data X is generated from some distribution, parameterized by θ.

Problem Formulation

- Data X is generated from some distribution, parameterized by θ.
  - E.g., $X \sim \mathcal{N}(\theta, 1)$.

Problem Formulation

- Data X is generated from some distribution, parameterized by θ.
  - E.g., $X \sim \mathcal{N}(\theta, 1)$.
- We observe X, but θ is unknown.

Problem Formulation

- Data X is generated from some distribution, parameterized by θ.
  - E.g., $X \sim \mathcal{N}(\theta, 1)$.
- We observe X, but θ is unknown.
- Maximum likelihood estimation (MLE) picks $\hat{\theta}$ to estimate θ as

$$\hat{\theta} = \arg\max \mathcal{L}(\theta)$$

Easy Example: $\mathcal{N}(\theta, 1)$

- Suppose $X \sim \mathcal{N}(\theta, 1)$.

Easy Example: $\mathcal{N}(\theta, 1)$

- Suppose $X \sim \mathcal{N}(\theta, 1)$.
- The density function for x is

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\dfrac{(x-\theta)^2}{2}}$$

Easy Example: $\mathcal{N}(\theta, 1)$

- Suppose $X \sim \mathcal{N}(\theta, 1)$.
- The density function for x is

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

- Suppose you observe $X = 10$. What value of θ maximizes the likelihood function?

Easy Example: $\mathcal{N}(\theta, 1)$

- Suppose $X \sim \mathcal{N}(\theta, 1)$.

Easy Example: $\mathcal{N}(\theta, 1)$

- Suppose $X \sim \mathcal{N}(\theta, 1)$.
- The density function evaluated at $X = 10$ is

$$p_\theta(10) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(10-\theta)^2}{2}}$$

Easy Example: $\mathcal{N}(\theta, 1)$

- Suppose $X \sim \mathcal{N}(\theta, 1)$.
- The density function evaluated at $X = 10$ is

$$p_\theta(10) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(10-\theta)^2}{2}}$$

Now we differentiate with respect to θ.

$$\begin{aligned}\frac{d}{d\theta}\mathcal{L}(\theta) = \frac{d}{d\theta}p_\theta(x) &= \frac{d}{d\theta}\frac{1}{\sqrt{2\pi}} e^{-\frac{(10-\theta)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(10-\theta)^2}{2}} \cdot (10-\theta) = 0\end{aligned}$$

Easy Example: $\mathcal{N}(\theta, 1)$

- Suppose $X \sim \mathcal{N}(\theta, 1)$.
- The density function for x is

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

We don't need to actually observe X. More generally if the random variable $X = x$ after the experiment is performed, then

$$\frac{d}{d\theta}\mathcal{L}(\theta) = \frac{\partial}{\partial\theta} p_\theta(x) = \frac{\partial}{\partial\theta} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} =$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \cdot (x-\theta) = 0$$

Typical Example

- Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. What is the MLE of θ?

Typical Example

- Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. What is the MLE of θ?
- The joint density function is

$$p_\theta(x_1, \ldots, x_n) = p_\theta(x_1)p_\theta(x_2)\cdots p_\theta(x_n)$$

Typical Example

- Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. What is the MLE of θ?
- The joint density function is

$$p_\theta(x_1, \ldots, x_n) = p_\theta(x_1) p_\theta(x_2) \cdots p_\theta(x_n)$$

$$= \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1 - \theta)^2}{2}} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_2 - \theta)^2}{2}} \right) \cdots \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_n - \theta)^2}{2}} \right)$$

Typical Example

- Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. What is the MLE of θ?
- The joint density function is

$$p_\theta(x_1, \ldots, x_n) = p_\theta(x_1)p_\theta(x_2)\cdots p_\theta(x_n)$$
$$= \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_1-\theta)^2}{2}}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_2-\theta)^2}{2}}\right)\cdots\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_n-\theta)^2}{2}}\right)$$
$$= \frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

Typical Example

- Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. What is the MLE of θ?
- We want

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \cdot \left(-\sum_{i=1}^{n} x_i - \theta\right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \cdot \left(n\theta - \sum_{i=1}^{n} x_i\right) = 0
\end{aligned}
$$

Typical Example

Clearly, to have

$$\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \cdot \left(n\theta - \sum_{i=1}^{n} x_i\right) = 0$$

we must have

$$\left(n\theta - \sum_{i=1}^{n} x_i\right) = 0$$

$$\implies \theta = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

The End