Wasserman: All of Statistics

Kenny Workman

January 30, 2026

1 Basics

Definition 1.1. The **sample space** Ω is the set of outcomes from an experiment. Each point is denoted ω and subsets, eg. $A \subset \Omega$ are called **events**.

Definition 1.2 (Axioms of Probability)**.** A function $\mathbb{P} : \Omega \to \mathbb{R}$ that assigns a real number to each event $A \subset \Omega$ is called a **probability function** or **probability measure** if it satsifies these three axioms:

1. **Non-negativity**. $\mathbb{P}(A) \geq 0$ for every event A
2. **Normalization**. $\mathbb{P}(\Omega) = 1$.
3. **Additivity**. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \emptyset$.

It is incredible, and not obvious, that much of probability is built up from these only these three axioms

Example 1.3. It's actually tricky to show $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ with these three facts:

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}(AB^c \cap AB \cap A^cB) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) + P(AB) - P(AB) \\
&= \mathbb{P}(AB^c \cup AB) + \mathbb{P}(A^cB \cup AB) - P(AB) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - P(AB)
\end{aligned}$$

Another simple idea is that events that are identical at the limit should have identical probabilities.

Theorem 1.4 (Continuity of Events)**.** *If $A_n \to A$ then $\mathbb{P}(A_n) \to \mathbb{P}(A)$.*

Proof. Let A_n be monotone increasing: $A_1 \subset A_2 \subset \dots$. Let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$.

Construct disjoint sets B_i from each A_i where $B_1 = A_1$ and $B_n = \{\omega \in \Omega : \omega \in A_n, \omega \notin \bigcup_{i=1}^{i-1} A_i\}$. It will be shown that (1) each pair of B_i are disjoint, (2) $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$ and (3) $A = \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ (Exercise 1.1).

From Axiom 3: $\mathbb{P}(A_n) = \mathbb{P}(\bigcup_{i=1}^{n} A_i) = \mathbb{P}(\bigcup_{i=1}^{n} B_i) = \sum_{i=1}^{n} \mathbb{P}(B_i)$.

Then $\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \sum^{n} \mathbb{P}(B_n) = \sum^{\infty} \mathbb{P}(B_n) = \mathbb{P}(\bigcup^{\infty} B_n) = \mathbb{P}(A)$

□

Definition 1.5 (Conditional Probability)**.** If $\mathbb{P}(B) > 0$, then the probability of A given B is

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

Theorem 1.6 (Total Probability)**.** *If $A_1 \cdots A_k$ partition Ω, $\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$*

Note. It can be difficult to assign a probability to every subset of Ω. In practice, we only assign values to select subsets described by a **sigma algebra** denoted A. This is a subset algebra with three properties:

- Non empty. $\emptyset \in A$. "Measure of the impossible."
- Closed over unions. $A_1, A_2 \cdots \in A \implies \cup_i A_i \in A$.
- Closed over complements. $A \in A \implies A^c \in A$.

Every set in A is considered **measurable** (by its membership in the sigma algebra). A along with Ω comprises a **measurable space**, denoted by the pair (A, Ω). If the measure on A is a probability function, importantly $\mathbb{P}(\Omega) = 1$, this space is also a **probability space**, denoted by the triple $(\Omega, A, \mathbb{P})$.

When Ω is the real line, the measure is often the Lebesgue measure, assigning intuitive values of "set length", eg. $[a, b] \mapsto b - a$.

Why is this important? While overly pedantic at first glance, this is the structure that explains why continuous density functions (next section) have nonzero probabilities when integrated over intervals but assign 0 probability to single points. The continuous measure, eg. Lebesgue measure, defined on the underlying probability space assigns positive values to sets and 0 to single points.

Exercise 1.1. Fill in the details for Theorem 1.2 and extend to the case where A_n is monotone decreasing.

Proof. For any pair B_{n+1} and B_n, because $B_n \subset A_n$ and $B_{n+1} \cap A_n = \emptyset$, it follows that $B_{n+1} \cap B_n = \emptyset$.
Let $\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i$. Then $\bigcup_{i=1}^{n+1} B_i = (A_{n+1} \setminus \bigcup_{i=1}^{n} A_i) \bigcup (\bigcup_{i=1}^{n} A_i) = \bigcup_{i=1}^{n+1} A_i$.
For the monotone decreasing case, let A_n be a sequence where $A_1 \supset A_2 \supset A_3 \ldots$.
Observe $A_1^c \subset A_2^c \ldots$ and $\lim_{n\to\infty} A_n = \Omega \setminus \bigcup^{\infty} A_i^c$. Construct disjoint B_n^c from A^c in the same way.
Then $\lim_{n\to\infty} \mathbb{P}(A_n) = 1 - \sum^{\infty} \mathbb{P}(B_i^c) = 1 - \mathbb{P}(A^c) = \mathbb{P}(A)$ □

Exercise 1.3. Let Ω be a sample space and $A_1, A_2, \ldots$ be events. Define $B_n = \cup_{i=n}^{\infty} A_i$ and $C_n = \cap_{i=n}^{\infty} A_i$.

(a) Show $B_1 \supset B_2 \supset B_3 \ldots$ and $C_1 \subset C_2 \subset C_3 \ldots$

(b) Show $\omega \in \cap_{n=1}^{\infty} B_n$ iff ω is in an infinite number of the events

(c) Show $\omega \in \cup n = 1^{\infty} C_n$ iff ω belongs to all of the events, except possibly a finite number of those events.

Proof.

(a) Certainly $\cup_{i=1}^{\infty} A_i \supset \cup_{i=2}^{\infty} A_i \ldots$ and $\cap_{i=1}^{\infty} A_i \subset \cap_{i=2}^{\infty} \ldots$.

(b) Forward. Assume $\omega \in \cap_{n=1}^{\infty} B_n$. If ω does not belong to an infinite number of events A_i, there exists some index j past which $\omega \notin B_j$. Then certainly $\omega \notin \cap_{n=1}^{\infty} B_n$. Reverse. ω belonging to infinite events means there cannot exist such a j described previously so $\omega \in B_n$ for all n. Indeed $\omega \in \cap_{n=1}^{\infty} B_n$

(c) Forward. Assume $\omega \in \cup_{n=1}^{\infty} C_n$. Then $\omega \in C_j = \cap_{i=j}^{\infty} A_i$ for some j. This is another way of saying ω is in every single event except for perhaps a finite number in $A_{i<j}$. Reverse. Let j be the index of the largest event that ω is not in. Then $\omega \in C_{n>j}$ and certainly $\omega \in \cup^{\infty} C_n$.

□

Note. The key idea above is this notion of "infinitely often" (i.o.) and "all but finitely often" (eventually) which are two distinct structures of infinite occurence in sequences. Consider an ω that exists in every other event (eg. just the odd indices) for infinite events and revisit its inclusion in $\cap^{\infty} B_i$ and $\cup^{\infty} C_i$.

Note. $\lim \cap \cup A_n$ is also referred to as the limit infimum of A_n. Similarly, $\lim \cup \cap A_n$ is referred to as the limit supremum of A_n.

Exercise 1.7. Let $\mathbb{P}(\bigcup^{n} A_i) \leq \sum^{n} \mathbb{P}(A_i)$. Then $\mathbb{P}(A_{n+1} \cup (\bigcup^{n} A_i)) \leq \mathbb{P}(A_{n+1}) + (\sum^{n} \mathbb{P}(A_i)) - \mathbb{P}(A_{n+1} \cap (\bigcup^{n} A_i)) \leq \sum^{n+1} \mathbb{P}(A_i)$

Note. Expand a bit on the Boole inequality.

Exercise 1.9. For fixed B s.t. $\mathbb{P}(B) > 0$, show $\mathbb{P}(. \mid B)$ satisfies the three axioms of probability.

Proof.

- **Non-negativity**. If $\mathbb{P}(B) > 0$ and $\mathbb{P}(AB) > 0$ for any $A \subset \Omega$, certainly $\frac{\mathbb{P}(AB)}{\mathbb{P}(B)} > 0$.
- **Normalization**. $\frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$
- **Additivity**. Let $AB \cap CB = \emptyset$, then $\mathbb{P}(AB \cap CB) = \mathbb{P}(AB) + \mathbb{P}(CB)$. Indeed $\frac{\mathbb{P}(AB \cap CB)}{B} = \frac{\mathbb{P}(AB)}{B} + \frac{\mathbb{P}(CB)}{B}$

□

Exercise 1.11. Suppose A and B are independent events. Show that A^c and B^c are also independent.

Proof. We are given $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$. Then $\mathbb{P}(A^c)\mathbb{P}(B^c) = (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) = 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(B) = 1 - \mathbb{P}(A \cup B) = \mathbb{P}(A^c B^c)$. The second to last equality uses independence of $P(AB)$. The last equality uses the property of set complements $P(A \cup B) = P(A^c \cap B^c)$. □

Exercise 1.13. Suppose a fair coin is tossed repeatedly until heads and tails is each encoutered exactly once. Describe Ω and compute the probablity exactly three tosses are needed.

Proof.

- The sample space is the set of binary strings with exactly one 0 and 1. For strings of length greater than 2, these are repeated strings of 0 or 1 capped with a 1 or 0 respectively.
- By independence, each n-string has an identical probability $\frac{1}{2}^n$. There are two such 3-strings: 001 and 110. Using additivity, $\mathbb{P}(3 \text{ tosses}) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$

□

Exercise 1.15. The probability a child has blue eyes is $\frac{1}{4}$. Assume independence between children. Consider a family with 3 children.

- If it is known that at least one of the children have blue eyes, what is the probablity that at least two of the children have blue eyes?
- If it is know that the youngest child has blue eyes, what is the probability that at least two of the children have blue eyes?

Proof.

- Straightforward conditional probability. Let A be the event where at least one child has blue eyes and B be the event where at least two children have blue eyes. Consider first, $\mathbb{P}(A) = 1 - \mathbb{P}(\text{no child has blue eyes}) = 1 - \frac{27}{64} = \frac{37}{64}$. Compute $\mathbb{P}(A \cap B)$ by enumerating events 101, 111, 110 and using additivity: $2 \cdot \frac{1}{4}^2 \cdot \frac{3}{4} + \frac{1}{4}^3 = \frac{10}{64}$. Then $\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{10}{64} \cdot \frac{64}{37} = \frac{10}{37}$
- Similar procedure. Let A be the event where the youngest child has blue eyes and B be as before. Using independence, $\mathbb{P}(A) = \frac{1}{4}$. (To see this rigorously, enumerate the sample space and see $\mathbb{P}(\Omega \mid \text{first child blue}) = 1$). Now $\mathbb{P}(B \cap A)$ describe events 110, 101, 111 only. $\frac{7}{64} \cdot \frac{4}{1} = \frac{7}{16}$.

□

Exercise 1.17. Show $\mathbb{P}(ABC) = \mathbb{P}(A \mid BC)\mathbb{P}(B \mid C)\mathbb{P}(C)$

Proof. By straightforward application of the definition of conditional probability: $\frac{\mathbb{P}(ABC)}{\mathbb{P}(BC)} \frac{\mathbb{P}(BC)}{\mathbb{P}(C)} \mathbb{P}(C) = \mathbb{P}(ABC)$ □

Exercise 1.19. Suppose 50% of computer users are Windows. 30% are Mac. 20% are Linux. Suppose 65% of Mac users, 82% of Windows users and 50% of Linux users get a virus. We select a person at random and learn they have the virus. What is the probability they are a Windows user?

Proof. Let each $\omega \in \Omega$ be a distinct user. Then $W, M, L \subset \Omega$ are the users with Windows, Mac + Linux machines. $V, N \subset \Omega$ are the users with and without viruses.

We want $\mathbb{P}(W \mid V) = \dfrac{\mathbb{P}(V \mid W)\mathbb{P}(W)}{\mathbb{P}(V)}$. Compute $\mathbb{P}(V) = \sum_{X=\{W,M,L\}} \mathbb{P}(V \mid X)\mathbb{P}(X) = 0.705$. Then $\mathbb{P}(W \mid V) = \dfrac{0.82 \cdot 0.50}{0.705} = 0.581$. □

Exercise 1.20. A box contains 5 coins, each with a different probability of heads: $0, 0.25, 0.5, 0.75, 1$. Let C_i be the event with coin i and H_i be the event that heads is recovered on toss i.Suppose you select a coin at random and flip it.

- What is the posterior probability $\mathbb{P}(C_i \mid H_1)$ for each coin?
- What is $\mathbb{P}(H_2 \mid H_1)$?
- Let B_i be the event that the first heads is recovered on flip i. What is $\mathbb{P}(C_i \mid B_i)$ for each coin?

Proof.

- $\mathbb{P}(H_1) = \frac{1}{2}$. For each coin, $\mathbb{P}(C_i \mid H) = \dfrac{\mathbb{P}(H \mid C_i)\mathbb{P}(C_i)}{\mathbb{P}(H)}$. $\mathbb{P}(H)$ can be worked out using total probability: $\sum_i \mathbb{P}(H|C_i)\mathbb{P}(C_i) = \frac{1}{2}$. Then eg. the posterior $\mathbb{P}(C_4 \mid H) = \frac{3}{4} \cdot \frac{1}{5} \cdot \frac{2}{1} = \frac{3}{10}$.
- Note that both tosses are conditionally independent: $\mathbb{P}(H_2H_1 \mid C_i) = \mathbb{P}(H_2 \mid C_i)\mathbb{P}(H_1 \mid C_i)$. $\mathbb{P}(H_2 \mid H_1) = \dfrac{\mathbb{P}(H_2H_1)}{\mathbb{P}(H_1)} = \dfrac{\sum_i \mathbb{P}(H_2H_1 \mid C_i)\mathbb{P}(C_i)}{\sum_i \mathbb{P}(H_1 \mid C_i)\mathbb{P}(C_i)}$. Because $\mathbb{P}(C_i)$ is uniform, we can simply to $\dfrac{\sum_i \mathbb{P}(H_2H_1 \mid C_i)}{\sum_i \mathbb{P}(H_1 \mid C_i)}$. The result is $\dfrac{\sum_i p_i^2}{\sum_i p_i}$.
- Similar idea to (a).

□

Note. Important to see that independent events are not conditionally independent in general. Try to construct an example.

2 Random Variables

– (kenny) TODO fix counters

Definition 2.1 (Random Variable)**.** A random variable X is a function mapping the sample space to real numbers: $X : \Omega \to \mathbb{R}$.

It is important to think of the relationship between the random variable and its underlying sample space when computing probabilities: eg. $\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x))$ and $\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$.

Definition 2.2 (Cumulative Distribution Function)**.** The CDF is the function $F_X : \mathbb{R} \to [0, 1]$ where $F_X(x) = \mathbb{P}(X \leq x)$. Equivalently $F_X(x) = \mathbb{P}(X^{-1}((-\infty, x])$.

The CDF contains "all the information" in a random variable. This is articulated by the following theorem:

Theorem 2.3. *For random variables X and Y with CDFs F and G, if $F(x) = G(X) \forall x \in [0, 1]$, then $X = Y$ ($\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for each $A \subset \mathbb{R}$).*

And the behavior of the CDF, including "all of its information" is uniquely determined by just three properties:

Theorem 2.4. *A function $F : \mathbb{R} \to [0, 1]$ is a CDF iff it satisfies three properties:*

- *Non-decreasing.* $x_2 > x_1 \implies F(x_2) \geq F(x_1)$
- *Normalization.* $\lim_{y\to 0} F(y) = 0$ *and* $\lim_{y\to 1} F(y) = 1$
- *Right-continuous. For any* $x \in \mathbb{R}$, $F(x) = F^+(x)$ *where* $F^+(x) = \lim_{y\to x, y>x} F(y)$

Proof. Starting with (iii) from the text, let $A = (-\infty, x]$ and $y_1, y_2, \dots$ be a sequence where $y_1 < y_2 \dots$ and $\lim_i y_i = x$. By the definition of the CDF, $F(y_i) = \mathbb{P}(A_i)$ and $F(x) = \mathbb{P}(A)$, where $\lim_i F(y_i)$ is equivalent to $\lim_{y\to x, y>x} F(y)$. Observe $\cap_i A_i = A$ so $\mathbb{P}(A) = \mathbb{P}(\cap_i A_i) = \lim_i \mathbb{P}(A_i) = \lim_i F(y_i) = F(x)$ as desired.

To see (ii), $\lim_{y\to-\infty} F(y) = 0$, define a sequence $y_1, y_2 \cdots$ where $y_1 > y_2 \cdots$ as before and $y_1 = y$. Let $A_i = (\infty, y_i]$. Then $\cap_i A_i = \emptyset$ and $\mathbb{P}(\cap_i A_i) = \mathbb{P}(\emptyset) = 0$. Indeed $\lim_{y\to-\infty} F(y) = \lim_i \mathbb{P}(A_i) = \mathbb{P}(\cap_i A_i) = 0$. A similar argument shows the limit to the other direction.

For (iii), if $x_2 > x_1$ then $P((-\infty, x_2]) \geq P((-\infty, x_1])$ and $F(x_2) \geq F(x_1)$. □

The interesting direction is the reverse: a function satisfying these properties uniquely determines a probability function. It is difficult to show in general. A concrete example is the Cantor function (Devil's staircase) which satisfies non-decreasing, normality and right-continuous properties but from which is difficult to derive a measure that satisfies eg. countable additivity.

Note. A deeper measure theory course will approach this problem by defining the probablity function on an algebra of subsets rather than on each subset directly. Refer to tools like Caratheodory's extension theorem.

It is from these random variables that we build "distributions", essentially functions $\mathbb{R} \to [0, 1]$ that obey the three probability axioms.

Definition 2.5. If X "takes" countably many values (eg. has a countable range) it is **discrete**. $f_X(x) = \mathbb{P}(X = x)$ is its **probability mass function** or PMF.

Definition 2.6. X is **continuous** if it has some f_X that obeys three properties:

- $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$
- $\forall x \in \mathbb{R} : f_X(x) \geq 0$
- $\mathbb{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x$

f_X is called the **probability density function** or PDF. Additionally, $F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x$ and $f_X(x) = F_X'(x)$ for all points x where F_X is differentiable.

The formal relation between the density function and the sample space is a bit tricky, especially when X is continuous. In practice, we often just produce a function and deal with it directly while assuming the underlying sample space with a well defined measure is lurking around.

Note. We learned the probability function is defined on a well-defined sample space by measuring events / sets.

Definition 2.7. The quartile function (or inverse CDF) is $F^{-1}(q) = \inf\{x : q < F(x)\}$

We call $F^{-1}(\frac{1/4}{)}$ the first quartile, $F^{-1}(\frac{1/2}{)}$ the second quartile (or median), etc.

We will proceed with some important mass functions.

Definition 2.8 (The Point Mass Distribution)**.** If $X \sim \sigma_a$ (reads "X has a point mass distribution at a"), $f_X(a) = 1$ while $f_X(x) = 0$ for all $x \neq a$.

$$F_X(x) = \begin{cases} 0, & x < a, \\ 1, & x \geq a \end{cases}$$

Definition 2.9 (The Uniform Distribution)**.** Suppose X has a mass function:

$$f(x) = \begin{cases} \frac{1}{k}, & x \in \{1 \dots k\} \\ 0, & \text{o.w.} \end{cases}$$

X then has a uniform distribution on $\{1 \dots k\}$.

Definition 2.10 (The Bernoulli Distribution)**.** If $X \sim Bernoulli(p)$, the PMF of X is $f(x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$ and $p \in [0, 1]$.

Here is the first instance of a parameterized random variable.

Definition 2.11 (The Binomial Distribution)**.** Binomial variables model the number of successful flips for n identical trials with probability p for each. We say $X \sim Binomial(n, p)$ with PMF:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{1 \dots n\} \\ 0 & \text{o.w.} \end{cases}$$

The following represent different ideas of unbounded "counting": trials until success and trials in some interval of time.

Definition 2.12 (The Geometric Distribution)**.** Here we have the idea of flipping a coin until our first success. $X \sim Geometric(p)$ with PMF: $f(x) = (1-p)^{x-1}p$

The probability value of each term is a geometric series. Indeed $p\sum^{\infty}(1-p)^x = \frac{p}{1-(1-p)} = 1$.

Definition 2.13 (The Poisson Distribution)**.** If $X \sim Poisson(\lambda)$ with PMF $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$

λ can be thought of as some interval of time. X then measures the number of events in this interval: decaying particles or mRNA translation.

Similarly to the geometric distribution, each term in the poisson is a Taylor polynomial, derived from the power series expansion of the exponential function. Indeed $e^{-\lambda}\sum^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1$.

Note. For distributions that count trials in some interval - some time or number of trials - the sum of variables equals a single variable that accumulates the interval.

If $X_1 \sim Binomial(n_1, p)$ and $X_2 \sim Binomial(n_2, p)$, then $X_1 + X_2 \sim Binomial(n_1 + n_2, p)$.

If $X_1 \sim Poisson(\lambda_1)$ and $X_2 \sim Poisson(\lambda_2)$, then $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$.

Note. Recall Ω really lurking around. Eg. let $X \sim Bernoulli$ and $\mathbb{P}(X = 1)$ is $\mathbb{P}(\omega \in [0, p]) = p$.

For the continuous distributions, useful to think of integration.

Definition 2.14 (The Continuous Uniform Distribution)**.** If X has a uniform distribution on the interval $[a, b]$ with PDF:

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{o.w.} \end{cases}$$

and CDF:

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x \geq b \end{cases}$$

Definition 2.15 (The Normal (Gaussian) Distribution)**.** $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

Note. If $X \sim N(0, 1)$ we say that X has a **standard Normal distribution**. We often denote X as Z with ϕ and Φ as the PDF and CDF.

There is no closed form function for Φ, so we use precomputed values from tables or rely on statistical programs. Calculations with Normal distributions then proceed by reexpressing X as some function of Z and using these values.

The following facts are essential when manipulating these variables:

- If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- If $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma)$
- If $X_i \sim N(\mu_i, \sigma_i)$ are independent, then $X = \sum_i X_i \sim N(\sum_i \mu, \sum_i \sigma)$

Definition 2.16 (The Exponential Distribution)**.** If $X \sim Exp(\beta)$, then $f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$.

Indeed $\int$

Note (The Gamma Function)**.** We often want a continuous extension of the factorial to real arguments, where $\Gamma(x) = (x-1)!$ for $x \in \mathbb{Z}^+$. This is the **gamma function** and defined $\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy$.

Evaluating the integral for $\Gamma(1), \Gamma(2) \ldots$ is a useful exercise to convince oneself of agreement with the factorial.

For example, $\Gamma(3) = \int_0^\infty y^2 e^{-y} dy$. Using integration by parts, this evaluates to $[-y^2 e^{-y} - 2ye^{-y} - 2e^{-y}]_0^\infty$. Using L'Hopital's, the first two terms drop out and we are left with $\Gamma(3) = 2 = (3-1)!$ as desired.

Equipped with the gamma function, we can now develop the gamma distribution.

Definition 2.17 (Gamma Distribution)**.** Let $\alpha, \beta > 0$. A continuous random variable X is said to have a *Gamma distribution* with shape parameter α and scale parameter β, denoted

$$X \sim \Gamma(\alpha, \beta),$$

if its probability density function is

$$f_X(x) = \frac{1}{\beta^\alpha \, \Gamma(\alpha)} \, x^{\alpha-1} \, e^{-\,x/\beta}, \quad x \geq 0.$$

If $X_i \sim \Gamma(\alpha_i, \beta)$ are independent, $\sum_i X_i = \Gamma(\sum \alpha_i, \beta)$.

The exponential distribution is then just a special case of a gamma distribution with $\alpha = 1$.

Note. The Gamma-normalization comes from evaluating

$$\int_0^\infty x^{\alpha-1} \, e^{-x/\beta} \, dx.$$

We make the substitution

$$x = \beta \, t, \quad dx = \beta \, dt,$$

so that

$$\int_0^\infty x^{\alpha-1} e^{-x/\beta} \, dx = \int_0^\infty (\beta t)^{\alpha-1} \, e^{-t} \, (\beta \, dt) = \beta^{\alpha-1} \, \beta \int_0^\infty t^{\alpha-1} e^{-t} \, dt = \beta^\alpha \, \Gamma(\alpha).$$

Hence in the density

$$f(x) = \frac{1}{\beta^\alpha \, \Gamma(\alpha)} \, x^{\alpha-1} \, e^{-x/\beta}$$

the factor $\beta^\alpha \, \Gamma(\alpha)$ is exactly the normalizing constant that makes $\int_0^\infty f(x) \, dx = 1$.

Definition 2.18 (X^2 Distribution)**.** X has a X^2 distribution with p degrees of freedom if the PDF is

$$f(x) = \frac{1}{\Gamma(\frac{p}{2}) 2^{\frac{p}{2}}} x^{\frac{p}{2}-1} e^{\frac{x}{2}}$$

Let $p > 0$. A random variable X is said to have a χ^2 distribution with p degrees of freedom, denoted $X \sim \chi^2_p$, if its probability density function is

$$f_X(x) = \frac{1}{2^{\,p/2} \, \Gamma\left(\frac{p}{2}\right)} \, x^{\,\frac{p}{2}-1} \, e^{-x/2}, \quad x > 0.$$

This distribution is the sum of squared, independent normals. If $Z_i \sim N(0,1)$ then $\sum_i Z_i^2 \sim \mathrm{X}^2_p$.

Definition 2.19 (Independence of Random Variables)**.** If $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$, we say X and Y are independent, written $X \perp\!\!\!\perp Y$.

Exercise 2.5. Let X and Y be discrete random variables. Show X and Y are independent iff $f_{X,Y}(x,y) = f_X(x)f_Y(y)$

Proof. If $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for every subset A, B, let $A = \{x\}$ and $B = \{y\}$ for every possible pair of elements. Then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. To see the reverse, $\mathbb{P}(X \in A, Y \in B) = \sum_{x\in A}\sum_{y\in B} f_{X,Y}(x,y) = \sum_{x\in A} f_X(x)\sum_{y\in B} f_Y(y) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ □

Theorem 2.20. *Suppose the range of X and Y is a (potentially infinite) rectangle. If we can express $f_{X,Y} = g(x)h(y)$, then X and Y are independent.*

Proof. Start by computing the marginals. $f_X = \int g(x)h(y)dy = g(x)(\int h(y)dy)$ and $f_Y = \int g(x)h(y)dx = h(y)(\int g(x)dx)$.

Then $f_X f_Y = g(x)(\int h(y)dy)h(y)(\int g(x)dx) = g(x)h(y)(\int\int h(y)g(x)dxdy)$. Because $g(x)h(y) = f_{X,Y}$, the integration term evaluates to 1. Then $f_X f_Y = g(x)h(y)(\int\int h(y)g(x)dxdy) = g(x)h(y)(1) = f_{X,Y}$ which is exactly the condition for independence of X and Y. □

In the above problem, notice the significance of requiring the range to be a rectangle. Any other region would produce integration limits in one variable that are functions of the other variable and you can no longer pull out the integration terms from the maringals:

$$f_X(x) \;=\; g(x)\underbrace{\Big[\int_{y=x^2}^{1} h(y)\,dy\Big]}_{\text{a function of } x}.$$

Definition 2.21 (Transformation of Continuous R.V.)**.** When Y and X are continuous.

- Find $A_y = \{x : r(x) \le y\}$ for each $y \in R$
- Then $F_Y(y) = \mathbb{P}(r(X) \le y) = \int_{A_y} f_X dx$
- $f_Y = F_Y'$

Exercise 2.1. Show $\mathbb{P}(X = x) = F(x^+) - F(x^-)$

Proof. The key here is to see $\lim_{z<x, z\to x} F(z) = \mathbb{P}(X \in \cup_i(\infty, z_i]) = \mathbb{P}(X < x)$ for some sequence $z_1, z_2, \cdots$ where $\lim_i z_i = x_i$. While $\lim_{y>x, y\to x} F(y) = \mathbb{P}(X \in \cap_i(\infty, y_i]) = \mathbb{P}(X \le x)$.

Pay attention to the behavior of converging sets and the boundary. In the right-continous case, the sequence is approaching the boundary x from above and each sequence is closed on x. Therefore in the limit, they include x.

In the left-continuous case, the sequence is approaching the boundary x from below and each sequence excludes x. Therefore in the limit, they exclude x.

To conclude $F(x^+) - F(x^-) = \mathbb{P}(X \le x) - \mathbb{P}(X < x) = \mathbb{P}(X = x)$. Of course, if X is continuous, $F(x^+) = F(x^-)$ and $\mathbb{P}(X = x) = 0$, showing once again that every real value has no probability mass. □

Exercise 2.4. Let X have density

$$f_X(x) = \begin{cases} \frac{1}{4}, & 0 < x < 1, \\ \frac{3}{8}, & 3 < x < 5, \\ 0, & o.w. \end{cases}$$

- Find the CDF of f_X
- Let $Y = \frac{1}{X}$. Find f_Y.

Proof.

- $$F_X(x) = \begin{cases} \frac{1}{4}x, & 0 < x < 1, \\ \frac{1}{4}, & 1 < x < 3, \\ \frac{3}{8}(x-3) + \frac{1}{4}, & 3 < x < 5 \\ 1, & x \geq 5 \end{cases}$$

- $$F_Y(y) = \begin{cases} 0, & y < \frac{1}{5}, \\ \frac{3}{8}(5 - \frac{1}{y}), & \frac{1}{5} < y < \frac{1}{3}, \\ \frac{6}{8}, & \frac{1}{3} < y < 1, \\ \frac{1}{4}(1 - \frac{1}{y}) + \frac{6}{8}, & 1 < y, \end{cases}$$

  Then to compute $f_Y(y) = F_Y'(y)$, we differentiate:

$$\begin{cases} 0, & y < \frac{1}{5}, \\ \frac{3}{8y^2}, & \frac{1}{5} < y < \frac{1}{3}, \\ 0, & \frac{1}{3} < y < 1, \\ \frac{1}{4y^2}, & 1 < y, \end{cases}$$

□

Exercise 2.7. Let X and Y be independent and suppose each is $Uniform(0,1)$. Let $Z = min\{X, Y\}$. Find the density $f_Z(z)$.

Proof. $\mathbb{P}(Z > z) = \mathbb{P}(X > z, Y > z) = \mathbb{P}(X > z)\mathbb{P}(Y > z)$. Then, $\mathbb{P}(Z > z) = (1-z)^2$ and $F_Z = 1 - \mathbb{P}(Z > z) = 1 - (1-z)^2$. $f_Z = F_Z' = -2z + 2$ □

Exercise 2.9. Let $X \sim Exp(\beta)$. Find $F(x)$ and $F^{-1}(q)$.

Proof. $f(x) = \frac{1}{\beta}e^{1/\beta}$. So $F(x) = \int_0^x f(x)dx = 1 - e^{-x/\beta}$.

F is a bijection over the interval $[0, \infty)$ so we can find a genuine inverse F^{-1} as $-\beta \ln(1-q)$.

□

Plugging in a few numbers to get a feel for $F^{-1}(q)$, we see that $F^{-1}(0.99) = \beta 4.6$ and $F^{-1}(0.9999) = \beta 9.2$, confirming that linear changes in sample space value have exponential effect in probability and that eg. increasing β decreases likelihood of events by stretching the density.

Exercise 2.11. Flip a coin once with probability heads of p. Let X and Y be the number of heads and tails.

- Show X and Y are independent
- Let $N \sim Poisson(\lambda)$ be the number of coin flips. Show now that X and Y are independent

Proof. **(a) One toss.** Because $Y = 1 - X$,

$$\mathbb{P}\{Y = 1 \mid X = 1\} = 0 \neq \mathbb{P}\{Y = 1\} = 1 - p,$$

so X and Y are dependent.

(b) Random number $N \sim$ Poisson(λ).
Step 1 (conditional pmf). Given $N = k$,

$$\mathbb{P}\{X = x, Y = y \mid N = k\} = \mathbf{1}_{\{x+y=k\}} \binom{k}{x} p^x (1-p)^y.$$

Step 2 (unconditional pmf). Summing over k, only the term $k = x + y$ remains:

$$\mathbb{P}\{X = x,\, Y = y\} = e^{-\lambda}\frac{\lambda^{x+y}}{(x+y)!}\binom{x+y}{x}p^x(1-p)^y = e^{-\lambda}\frac{(\lambda p)^x}{x!}\,e^{-\lambda}\frac{\big(\lambda(1-p)\big)^y}{y!}.$$

Step 3 (marginals). Hence

$$X \sim \text{Poisson}(\lambda p), \qquad Y \sim \text{Poisson}\big(\lambda(1-p)\big),$$

and $\mathbb{P}\{X = x,\, Y = y\} = \mathbb{P}\{X = x\}\mathbb{P}\{Y = y\}$, so X and Y are independent. □

Exercise 2.13. Let $X \sim Normal(0, 1)$ and $Y = e^X$.

- Find f_Y and plot it.
- Generate 10,000 random draws from X. Create a histogram of these draws and compare to the density plot.

Proof. Because $r = e^x$ is a strictly monotonically increasing function, we can apply $f_Y = f_X(s(x))s'(x)$ where $s = r^{-1}$. Then $f_Y(y) = f_X(ln(y))\frac{1}{y}$. Using the standard normal density, $f_Y(y) = \frac{1}{\sqrt{2\pi}y}e^{-\frac{(\ln y)^2}{2}}$ □

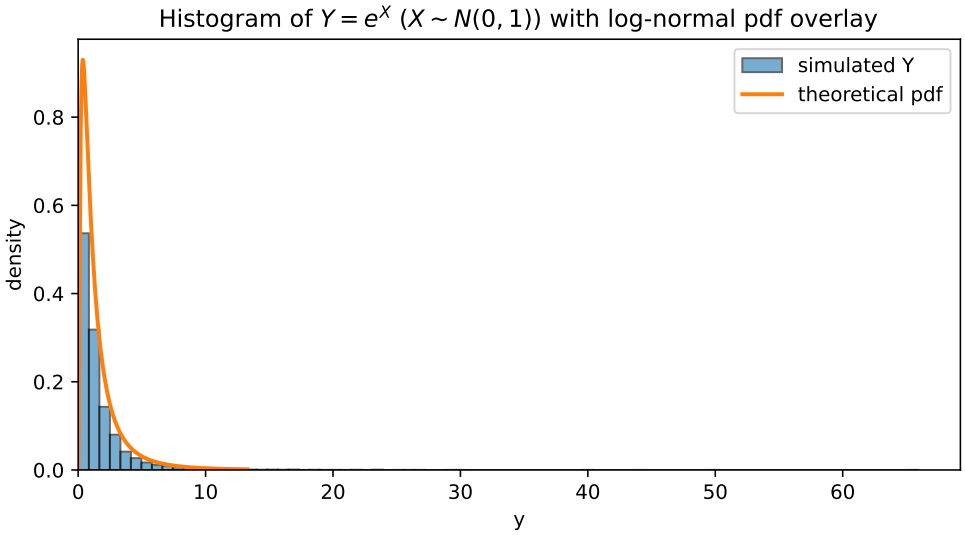


Figure 1: Histogram of $Y = e^X$ overlaid with its log-normal density.

Note. It is worth understanding why $f_Y = f_X(s(x))s'(x)$ can be used when r is a strict monotonically increasing or decreasing function. This condition forces s to be differentiable and single-valued for the single-variable change-of-variable integration.

Exercise 2.15.

- Let X have a continuous, strictly increasing CDF F. Let $Y = F^{-1}(X)$. Find the density of Y.
- Now let $U \sim Uniform(0, 1)$. Let $X = F^{-1}(U)$, where F is no longer the CDF of X but is still continuous and strictly increasing. Show $F_X = X$.
- Write a program to generate $Exponential(\beta)$ random variables from $Uniform(0, 1)$

Proof.

- $\mathbb{P}(Y \leq y) = \mathbb{P}(F(X) \leq y) = \mathbb{P}(X \leq F^{-1}(y)) = F(F^{-1}(y))$ So $F_y = 1$.
- $\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F_U(F(x)) = F(x)$
- Using the fact that $X = F^{-1}(U)$ has CDF F, we compute the exponential CDF and find its inverse: $F_X^{-1}(q) = -\beta \ln(1 - \beta^2 q)$. A histogram of generated values, overlayed against the exponential PDF, can be found below.

□

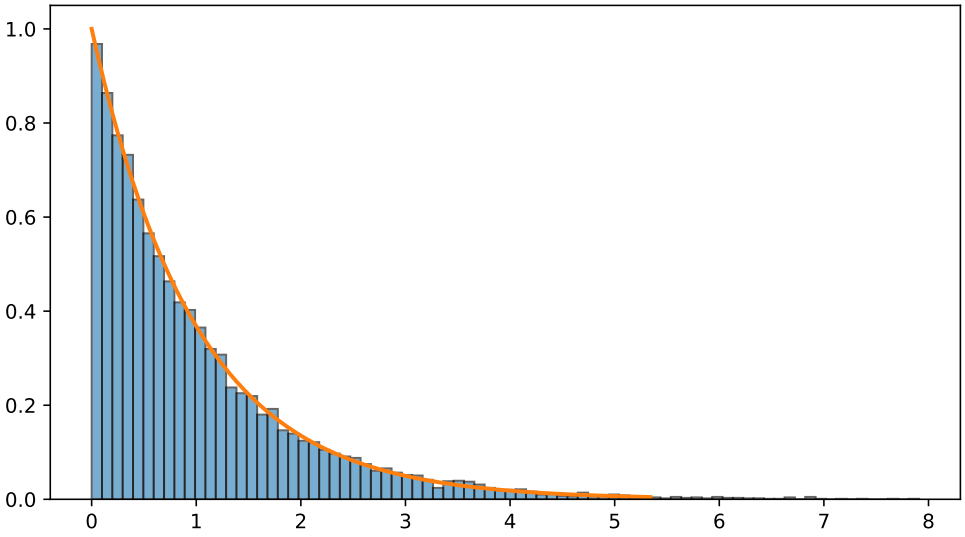


Figure 2: Histogram of generated exponentials overlayed against theoretical PDF

Exercise 2.16. Let $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$ be independent random variables. Find the density of X given $X + Y = n$. Use the fact that $X + Y \sim Poisson(\lambda + \mu)$ and $\mathbb{P}(X = x, X + Y = n) = \mathbb{P}(X = x, Y = n - x)$.

Proof. We are interested in the quantity $\mathbb{P}(X = x, X + Y = n | X + Y = n)$. Observe $\mathbb{P}(X + Y = n) = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}$. And $\mathbb{P}(X = x, X + Y = n) = \mathbb{P}(X = x, Y = n - x) = \mathbb{P}(X = x)\mathbb{P}(Y = n - x) = e^{-\lambda} \frac{\lambda^x}{x!} e^{-\mu} \frac{\mu^{n-x}}{(n-x)!}$.

Our conditional distribution is then the expression:

$$\frac{e^{-\lambda} e^{-\mu} \frac{\lambda^x}{x!} \frac{\mu^{n-x}}{(n-x)!}}{e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}}$$

Simplifying we begin to see the shape of the binomial:

$$\frac{\lambda^x}{x!} \frac{\mu^{n-x}}{(n-x)!} \frac{n!}{(\lambda+\mu)^n}$$

$$\frac{n!}{(n-x)!x!} \frac{\lambda^x \mu^{n-x}}{(\lambda+\mu)^n}$$

$$\frac{n!}{(n-x)!x!} \frac{\lambda^x \mu^{n-x}}{(\lambda+\mu)^{n-x}(\lambda+\mu)^x}$$

This is $\binom{n}{x}(\frac{\lambda}{\lambda+\mu})^x(\frac{\mu}{\lambda+\mu})^(n-x)$ or $Binomial(n, \frac{\lambda}{\lambda+\mu})$. □

Exercise 2.20.

3 Expectation

3.1 Expectation of a Random Variable

Definition 3.1. The expected value (or mean or first moment) of X is defined as

$$\mathbf{E}[X] = \int x dF_X(x) = \begin{cases} \sum_x x f(x), & \text{if X is discrete} \\ \int_x x f(x) dx, & \text{if X is continuous} \end{cases}$$

Assuming the sum or integral is well-defined, we use the following notation to denote the expected value of X: $\mathbf{E}[X] = \mu = \mu_X$

3.2 Properties of Expectation

Theorem 3.2. $\mathbf{E}\left[\sum_i X_i\right] = \sum_i \mathbf{E}[X]_i$

Theorem 3.3. *If $X_1 \dots X_i$ are independent,* $\mathbf{E}\left[\prod X_i\right] = \prod_i \mathbf{E}[X]_i$

Note. Work out the above briefly and note why we need independence for the product and not the sum. Eg. $\int (x+y) f_{X,Y}$ vs. $\int xy f_{X,Y}$. Integration is additive but factorization of the joint PDF/PMF is necessary for the product terms.

3.3 Variance and Covariance

Definition 3.4. The variance of X is defined as

$$\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]$$

and is denoted as σ_X^2 or σ^2 or $\mathbf{Var}[X]$

Theorem 3.5. *Assuming the variance of X is well-defined, it has the following properties:*

- $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2$
- $\mathbf{Var}[aX + b] = a^2 \mathbf{Var}[X]$
- *If $X_1 \dots X_n$ are independent,* $\mathbf{Var}\left[\sum_i a_i X_i\right] = \sum_i a_i^2 \mathbf{Var}[X_i]$

Proof.

- $\mathbf{E}\left[(X-\mu)^2\right] = \mathbf{E}\left[X^2 - 2X\mu + \mu^2\right] = \mathbf{E}\left[X^2\right] - \mu^2$
- $\mathbf{E}[(aX+b) - \mathbf{E}[aX+b]]^2 = \mathbf{E}\left[(aX + b - a\mu + b)^2\right] = \mathbf{E}\left[(a(X-\mu))^2\right] = a^2 \mathbf{E}\left[(X-\mu)^2\right] = a^2 \mathbf{Var}[X]$
- $\mathbf{Var}\left[\sum_i a_i X_i\right] = \mathbf{E}\left[(\sum_i a_i X_i - \mathbf{E}\left[\sum_i a_i X_i\right])^2\right]$. Using additivity of expectation, $\mathbf{E}\left[\sum_i a_i X_i\right] = \sum_i a_i \mathbf{E}[X]_i$. Then our expression becomes $\mathbf{E}\left[(\sum a_i X_i - a_i \mathbf{E}[X]_i)^2\right]$. Expanding this expression, we arrive at $\mathbf{E}\left[\sum_i (a_i X_i - a_i \mathbf{E}[X]_i)^2 + \sum_{i,j} a_i a_j (X_i - \mathbf{E}[X]_i)(X_j - \mathbf{E}[X]_j)\right]$. The first set of terms become $\sum_i a_i^2 \mathbf{Var}[X]_i$ and the second set of terms drop out when expanded as every pair of variables are independent. $(\mathbf{E}[X_i X_j] - \mathbf{E}[X]_i \mathbf{E}[X]_j = 0)$.

□

Definition 3.6. Let $X_1 \dots X_n$ be random variables. The **sample mean** is then

$$\bar{X_n} = \frac{1}{n} \sum_i X_i$$

And the **sample variance** is

$$S_n^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X_n})^2$$

Theorem 3.7. *If $X_1 \dots X_n$ are i.i.d. and $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2$, then $\mathbf{E}\left[\bar{X_n}\right] = \mu$, $\mathbf{Var}\left[\bar{X_n}\right] = \frac{\sigma^2}{n}$, and $\mathbf{E}\left[S_n^2\right] = \sigma^2$.*

Proof.

$$\mathbf{E}\left[\bar{X_n}\right] = \frac{1}{n}\sum_i \mathbf{E}[X_i] = \mu$$

$$\mathbf{Var}\left[\bar{X_n}\right] = \frac{1}{n^2}\sum_i \mathbf{Var}[X_i] = \frac{\sigma^2}{n}$$

Notice $\sum_i (X_i - \bar{X_n})^2 = \sum_i (X_i^2 - 2X_i\bar{X_n} + \bar{X_n}) = \sum_i (X_i^2) - 2\sum_i X_i\bar{X_n} + \sum_i \bar{X_n}^2$. The inner term becomes $2\bar{X_n}n\bar{X_n} = 2n\bar{X_n}^2$. So:

$$\mathbf{E}\left[S_n^2\right] = \mathbf{E}\left[\frac{1}{n-1}\sum_i (X_i - \bar{X_n})^2\right] = \frac{1}{n-1}\sum_i \mathbf{E}\left[X_i^2\right] - \mathbf{E}\left[\bar{X_n}^2\right] = \frac{1}{n-1} n[(\sigma^2 - \mu^2) - (\frac{\sigma^2}{n} - \mu^2)] = \sigma^2$$

□

Note. So what's up with the $\frac{1}{n-1}$?

Natural way to introduce "degrees of freedom". Consider the vector of residuals $r_i = X_i - \bar{X_n}$. We actually "use up" one of these residuals in the following way.

Notice the sum of our residuals evaluates to 0.

$$\sum_{i=1}^{n}(X_i - \bar{X_n}) = \sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0$$

This is just algebra and comes from the fact that our mean is not the true mean rather estimated from data. So after picking $n-1$ such r_i, the last r_n must equal $-\sum_{i=0}^{n-1} r_i$ for this identity to hold.

We then say that the sum of residuals, $\sum_i r_i$, used within the S_n^2 statistic has only $n-1$ "degrees of freedom". It is common shorthand to also say the variance estimate itself (S_n^2) also has $n-1$ degrees of freedom.

Definition 3.8. Let X and Y be r.v.s with means μ_X, μ_Y and standard deviations σ_X, σ_Y. The **covariance** of X and Y is then:

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

The correlation is then:

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

Theorem 3.9. *$Cov(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$ and $\rho_{X,Y}$ satisfies $-1 \le \rho_{X,Y} \le 1$. If $Y = aX + b$, where a, b are constants, then $\rho_{X,Y} = \begin{cases} -1, & a < 0 \\ 1, & a > 0 \end{cases}$. If X, Y are independent, then $Cov(X,Y) = 0$, although the converse need not be true.*

Theorem 3.10 (Variance of a sum)**.** $\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2Cov(X,Y)$. *More generally* $\mathbf{Var}[\sum_i a_i X_i] = \sum_i a_i^2 \mathbf{Var}[X_i] + \sum\sum_{i<j} 2 * a_i * a_j Cov(X_i, X_j)$

3.4 Expectation and Variance of Important Random Variables

3.5 Conditional Expectation

Definition 3.11 (Conditional Expectation)**.**

Definition 3.12 (The Law of Iterated Expectation)**.** $\mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[X]$

Definition 3.13.

Example 3.14. Suppose we pick a county from the US at random and choose n people from it. Let X be the number of these people with a disease. Let Q be the proportion of people in the county with the disease. Then X given $Q = q$ is $Binomial(n, q)$. $\mathbf{E}[X|Q = q] = nQ$ and $\mathbf{Var}[X|Q = q] = nQ(1 - Q)$.

Suppose now $Q \sim Uniform(0, 1)$. This is a **hierarchical model**. $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X]] = \mathbf{E}[nQ] = \frac{n}{2}$. $\mathbf{Var}[X] = \mathbf{Var}[\mathbf{E}[X|Q]] + \mathbf{E}[\mathbf{Var}[X|Q]]$.

$$\mathbf{Var}[\mathbf{E}[X|Q]] = \mathbf{Var}[nQ] = n^2\,\mathbf{E}[Q] = n^2\frac{1}{12}$$

$$\mathbf{E}[\mathbf{Var}[X|Q]] = \mathbf{E}[nQ(1-Q)] = n\,\mathbf{E}\left[Q - Q^2\right] = \int q - q^2 dq = \frac{n}{6}$$

Then:

$$\mathbf{Var}[X] = \frac{n}{6} + \frac{n^2}{12}$$

3.6 Moment Generating Functions

Definition 3.15 (The Moment Generating Function)**.** The MGF, or Lapace Transformation, of X, is $\psi_X t = \mathbf{E}\left[e^{tX}\right] = \int e^{tX} dF_X dx$ where t varies over $\mathbb{R}$.

We will use the MGF to compute the moments of X. Assuming ψ is well defined on the open interval around $t = 0$, $\psi'_X(0) = \frac{d}{dt}\,\mathbf{E}\left[e^{tX}\right]|_{t=0} = \mathbf{E}\left[\frac{d}{dt}e^{tX}\right]|_{t=0} = \mathbf{E}\left[Xe^{tX}\right]|_{t=0} = \mathbf{E}[X]$.

Swapping differentiation with expectation when ψ is well defined in this interval is a fact we will assume for now but (hopefully) will return to later.

Theorem 3.16 (Properties of the MGF)**.**
If $Y = aX + b$, then $\psi_Y(t) = e^b \psi_X(at)$

Theorem 3.17. *Let X and Y be random variables. If $\psi_X(t) = \psi_Y(t)$ for all t, then X and Y are equal in distribution.*

I view the above fact as another way of saying if X and Y have identical moments, they must have the same distribution.

3.7 Exercises

Exercise 2.1. Assume we have some fortune c and we play a game where each turn we half or double our money with even probability. Compute the expected value of the resulting fortune after n turns.

Proof. Let X_n be our random variable and see $\mathbb{P}(X_n = c * 2^{n-2x}) = \binom{n}{x}2^{-n}$. The density is binomial with support $c * 2^{n-2x}$ ranging over $x = 0$ to $x = n$. The 2^{-n} expression comes from simplifying the standard $\frac{1}{2}^x \frac{1}{2}^{n-x}$. Similar for 2^{n-2x}

Then $\mathbf{E}[X_n] = \sum_{x=0}^{n} c * 2^{n-2x}\binom{n}{x}2^{-n}$. Simplifying the obvious things, we have $c\sum_{x=0}^{n} 2^{-2x}\binom{n}{x}$. □

Exercise 2.2. Show $\mathbf{Var}[X] = 0$ iff exists some constant c where $\mathbb{P}(X = c) = 1$.

Proof. □

Exercise 2.3. Let $X_1 \dots X_n$ be i.i.d. Uniform$(0,1)$ and $Y_n = \max\{X_1 \dots X_n\}$. Compute $\mathbf{E}[Y_n]$.

Proof. Observe $F_{Y_n} = \mathbb{P}(\max\{X_1 \dots X_n\} \leq y) = y^n$. (It is helpful to see also that $\mathbb{P}(\min\{X_1 \dots X_n\} \leq y) = 1-(1-y)^n$). Then $f_{Y_n} = dF_{Y_n} = \frac{d(y^n)}{dy} = ny^{n-1}$. $\mathbf{E}[Y_n] = \int y d_{Y_n} dy = \int ny^n dy = [\frac{n}{n+1} y^{n+1}]_{y=0}^{1} = \frac{n}{n+1}$. □

Exercise 2.4.

Proof. □

Note. Another approach is invoking the "rule of the lazy statistician", eg. $EY = \int r(x) f_{X_1 \dots X_n} dx_1 \dots dx_n$. In two dimensions, this calculation is trivial as the density can be evaluated as a piecewise integral over two halves of the unit square (those halves separated by a line through the diagonal). $2 \int \int_{x_1 > x_2} x_1 dx_2 dx_1 = 2 * \frac{1}{3} = \frac{2}{3}$ as expected for $\mathbf{E}[Y_2]$

Exercise 2.5. Flip a fair coin until you encounter a heads. Compute the expected value of the number of tosses.

Proof. Let X be the random variable holding the number of tosses. See that $\mathbb{P}(X = x) = \frac{1}{2}^x$. Then $\mathbf{E}[X] = \sum_{x=1}^{\infty} \frac{x}{2^x}$. To compute this series, recognize a general form of a geometric series can be expressed as $G(r) = \sum_{x=1}^{\infty} r^x = \frac{1}{1-r}$.

We need to rearrange things a bit: First, we take the derivative to pull out an x: $\frac{d(G(r))}{dr} = \sum_{x=1}^{\infty} x r^{x-1} = (1-r)^{-2}$. Then we multiply by r: $\sum_{x=1}^{\infty} x r^x = r * (1-r)^{-2}$. When $r = \frac{1}{2}$, this expression is equivalent to the series of our expectation. $\mathbf{E}[X] = 2$ □

Exercise 2.7. Let X be a continuous random variable where $\mathbb{P}(X < 0) = 0$ and the expectation exists. Show $\mathbf{E}[X] = \int_{x=0}^{\infty} \mathbb{P}(X \geq x) dx$.

Proof. Observe $\int_{x=0}^{\infty} \mathbb{P}(X \geq x) = \int_{x=0}^{\infty} (1-F(x)) dx$. This evaluates to $[(1-F(x))x]_{x=0}^{\infty} - \int_{x=0}^{\infty} (-f_X(x)) x dx$ using integration by parts. Observing that $\lim_{x \to \infty} x\big(1-F_X(x)\big) = 0$, this expression simplifies to $\int_{x=0}^{\infty} x f_X(x) dx = \mathbf{E}[X]$ as desired. □

Note. This is the tail-sum expression of expectation that will come in handy later.

Exercise 2.12. Compute $\mathbf{E}[X]$ and $\mathbf{Var}[X]$ when X is Poisson, Exponential.

Proof. Let $X \sim Poisson(\lambda)$. Recall $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ using the Maclaurin series for e^x. We are interested in $\sum_{x=0}^{\infty} x f_X(x) = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda}$ for $\mathbf{E}[X]$. Notice the Maclaurin series for xe^x is $\frac{n x^n}{n!}$. Then $e^{-\lambda} \sum_{x=0}^{\infty} \frac{x \lambda^x}{x!} = e^{-\lambda} \lambda e^{\lambda} = \lambda$ as desired.

To compute $\mathbf{Var}[X]$, we are instructed to first find $\mathbf{E}[X(X-1)]$. Notice $\mathbf{E}\left[X^2\right] = \mathbf{E}[X(X-1)] + \mathbf{E}[X]$.

$$\mathbf{E}[X(X-1)] = \sum_{x=0}^{\infty} \frac{x(x-1)\lambda^x}{x!} e^{-\lambda}$$

Simplify and notice this looks like a "shifted" form of the Maclaurin series for e^{λ}. The first two terms are 0, so we can start our counter at $x = 2$:

$$\sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} e^{-\lambda}$$

Factoring out a λ^2 we arrive at the familiar series:

$$e^{-\lambda} \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{(x-2)}}{(x-2)!} = e^{-\lambda} \lambda^2 e^{\lambda} = \lambda^2$$

Indeed, $\mathbf{E}\left[X^2\right] = \mathbf{E}[X(X-1)] + \mathbf{E}[X] = \lambda^2 + \lambda$. So $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$ a desired.

Now let $X \sim Exp(\beta)$. Recall $f_X(x) = \frac{1}{\beta}e^{-\frac{x}{\beta}}$. Then $\mathbf{E}[X] = \int_{x>0} \frac{x}{\beta}e^{-\frac{x}{\beta}}dx$ Using integration by parts, $\frac{1}{\beta}\int xe^{\frac{-x}{\beta}} = \frac{1}{\beta}[-\beta xe^{\frac{-x}{\beta}} - \beta^2 e^{\frac{x}{\beta}}\Big|_{x=0}^{\infty}] = \frac{1}{\beta} * \beta^2 = \beta$ as desired.

Now we approach $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2$. First we compute $\mathbf{E}\left[X^2\right] = \int_{x>0} \frac{x^2}{\beta}e^{-\frac{x}{\beta}}dx = \frac{1}{\beta}\int_{x>0} x^2 e^{-\frac{x}{\beta}}$. Again using integration by parts we arrive at:

$$\frac{1}{\beta}(-\beta x^2 e^{\frac{-x}{\beta}} - \int(-\beta)2xe^{\frac{-x}{\beta}})\Big|_{x=0}^{\infty} =$$

$$\frac{1}{\beta}(-\beta x^2 e^{\frac{-x}{\beta}} + 2\beta(-\beta xe^{\frac{-x}{\beta}} - \beta^2 e^{\frac{-x}{\beta}}))\Big|_{x=0}^{\infty} =$$

$$\frac{1}{\beta}(-\beta x^2 e^{\frac{-x}{\beta}} - 2\beta^2 xe^{\frac{-x}{\beta}} - 2\beta^3 e^{\frac{-x}{\beta}}))\Big|_{x=0}^{\infty} =$$

$$\frac{1}{\beta}(2\beta^3) = 2\beta^2$$

Then $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = 2\beta^2 - \beta^2 = \beta^2$ as desired. □

Exercise 2.13. Suppose we generate a random variable in the following way. We flip a fair coin. If heads, $X \sim Uniform(0,1)$, and if tails, $X \sim Uniform(3,4)$. Find the mean and standard deviation of X.

Proof. Let $Y \sim Bernoulli(0.5)$ represent the coin flip. $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]] = 0.5 * \mathbf{E}[X|Y=0] + 0.5 * \mathbf{E}[X|Y=1] = 0.5 * 0.5 + 0.5 * 3.5$

$\mathbf{Var}[X] = \mathbf{E}[\mathbf{Var}[X|Y]] + \mathbf{Var}[\mathbf{E}[X|Y]]$. $\mathbf{E}[\mathbf{Var}[X|Y]] = 0.5 * \frac{1}{12} + 0.5 * \frac{1}{12} = \frac{1}{12}$. $\mathbf{Var}[\mathbf{E}[X|Y]] = \mathbf{E}\left[(\mathbf{E}[X|Y] - \mathbf{E}[X])^2\right] = 0.5*(0.5-2)^2 + 0.5*(3.5-2)^2 = 1.5^2$. $\mathbf{Var}[X] = \mathbf{E}[\mathbf{Var}[X|Y]] + \mathbf{Var}[\mathbf{E}[X|Y]] = \frac{1}{12} + 2.25 = 2.33$ $\sigma_X = 1.53$ □

Exercise 2.21. Let X, Y be random variables. Suppose $\mathbf{E}[Y|X] = X$. Show $Cov(X,Y) = Var(X)$.

Proof. $Cov(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$. Recognize $\mathbf{E}[XY] = \mathbf{E}[\mathbf{E}[XY|X]]$. Evaluating the inner expectation $\mathbf{E}[XY|X] = X\,\mathbf{E}[Y|X] = X^2$. Then, $\mathbf{E}[XY] = \mathbf{E}\left[X^2\right]$. Similarly, $\mathbf{E}[X]\mathbf{E}[Y] = \mathbf{E}[X]\mathbf{E}[\mathbf{E}[Y|X]] = \mathbf{E}[X]\mathbf{E}[X]$. So the expression collapses to $\mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \mathbf{Var}[X]$. □

Exercise 2.23. Find the MGFs for Poisson, Normal and Gamma distributions.

Exercise 2.24. Let $X \sim Poisson(\lambda)$. Then $\psi_X(t) = \mathbf{E}\left[e^{Xt}\right] = \sum_{x=0}^{\infty} e^{Xt}e^{-\lambda}\frac{\lambda^x}{x!}$. Recognize $\sum_{x=0}^{\infty}\frac{e^{xt}\lambda^x}{x!} = \sum_{x=0}^{\infty}\frac{(e^t\lambda)^x}{x!}$ is the Maclaurin series for $e^{\lambda e^t}$. Then $e^{-\lambda}\sum_{x=0}^{\infty} e^{Xt}\frac{\lambda^x}{x!} = e^{-\lambda}e^{e^t\lambda} = e^{\lambda(e^t-1)}$.

Let $X \sim Normal(\mu, \sigma^2)$.

Exercise 2.6. Prove the rule of the lazy statistician in the discrete case.

Proof. Let $Y = r(X)$. $\mathbf{E}[Y] = \sum_{y\in Y} y\mathbb{P}(Y=y)$. For any given y, $y*\mathbb{P}(Y=y) = \sum_{x\in r^{-1}(y)} r(x)\mathbb{P}(X=x) = r(x)\sum_{x\in r^{-1}(y)}\mathbb{P}(X=x)$. Then $\sum_{y\in Y} y\mathbb{P}(Y=y) = \sum_{y\in Y}\sum_{x\in r^{-1}(y)} r(x)\mathbb{P}(X=x) = \sum_{x\in X} r(x)\mathbb{P}(X=x)$. □

Exercise 2.15. Let

$$f_{X,Y} = \begin{cases} \frac{1}{3}(x+y), & 0 \le x \le 1, 0 \le y \le 2 \\ 0, & o.w. \end{cases}$$

and find $\mathbf{Var}[2X+3Y+8]$.

Proof. The main idea is $\mathbf{Var}[2X+3Y+8] = 4\,\mathbf{Var}[X] + 9\,\mathbf{Var}[Y] + 2*2*3Cov(X,Y)$. The rest is tedious algebra.

The following calculations are important: $\mathbf{E}[X] = \frac{5}{9}$ $\mathbf{E}\left[X^2\right] = \frac{7}{18}$ $\mathbf{E}[Y] = \frac{11}{9}$ $\mathbf{E}\left[Y^2\right] = \frac{16}{9}$ $\mathbf{Var}[X] = \frac{13}{162}$ $\mathbf{Var}[Y] = \frac{23}{81}$ $\mathbf{E}[XY] = \frac{2}{3}$ $Cov(X,Y) = -\frac{1}{81}$

Plugging things back in, we arrive at $\frac{245}{81}$. □

Exercise 2.16. Let $r(x)$ and $s(y)$ be functions of x and y. Show that $\mathbf{E}[r(X)s(Y)|X] = r(X)\,\mathbf{E}[s(Y)|X]$. Then show $\mathbf{E}[r(X)|X] = r(X)$.

Proof. $\mathbf{E}[r(X)s(Y)|X] = \int r(x')s(y)f_{X,Y|X}(x', y|x)dx'dy$. Pay careful attention to the use of x' and x. We integrate over all x' in X where as x is provided by the conditioning X (and may be fixed in future calculations).

$\int r(x')s(y)f_{X,Y|X}(x', y|x)dx'dy = \int r(x')s(y)f_{Y|X}(y|x')f_{X|X}(x'|x)dx'dy$ by the chain rule of conditional densities. Now observe when x is fixed, eg. in the expression $\mathbf{E}[r(X)s(Y)|X = x]$, $f_{X|X}(x'|x)$ becomes $1_{X=x}$ and X degenerates to $X = x$. Indeed, $\mathbf{E}[r(X)s(Y)|X = x] = r(x)\int s(y)f_{Y|X}(y|x')dx'dy = r(x)\,\mathbf{E}[s(Y)|X]$. When X is not fixed, $\mathbf{E}[r(X)s(Y)|X] = r(X)\,\mathbf{E}[s(Y)|X]$ □

Exercise 2.19.

Proof. $\mathbf{E}\left[\bar{X}_n\right] = \frac{1}{n} * n * \mathbf{E}[X_i] = \mathbf{E}[X_i]$ $\mathbf{Var}\left[\bar{X}_n\right] = \frac{1}{n^2} * n * \mathbf{Var}[X_i] = \frac{1}{n}\mathbf{Var}[X_i]$ □

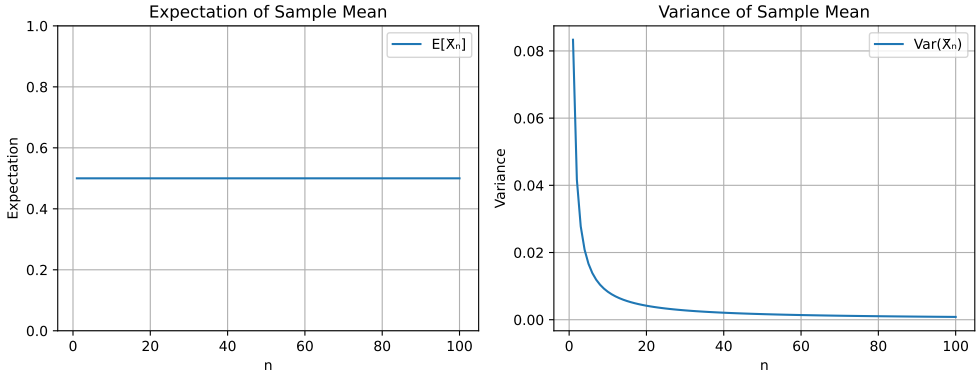


Figure 3: Expectation and variance of statistic as function of n

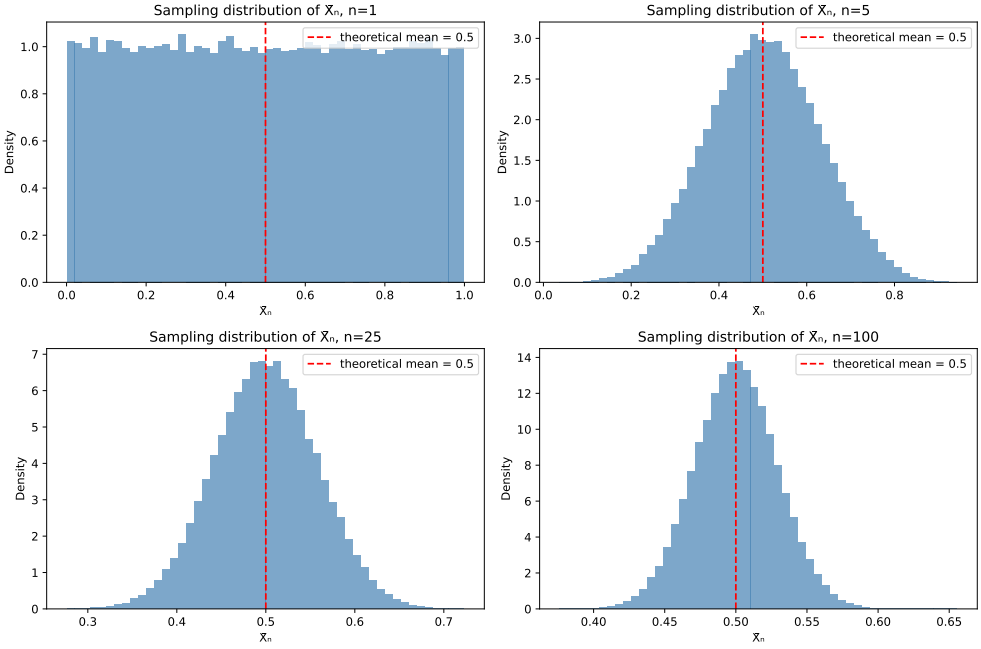


Figure 4: Sampling distribution for increasing n

Exercise 2.22. Let $0 < a < b < 1$ and $X \sim Uniform(0,1)$. Let

$$Y = \begin{cases} 1, & 0 < X < b, \\ 0, & \text{otherwise}, \end{cases} \qquad Z = \begin{cases} 1, & a < X < 1, \\ 0, & \text{otherwise}. \end{cases}$$

- Show Y and Z are not independent.
- Evaluate $\mathbf{E}[Y|Z]$

Proof.

- $\mathbb{P}(Y = 1, Z = 1) = b - a \neq \mathbb{P}(Y = 1)\mathbb{P}(Z = 1) = a * (1 - a)$
- $\mathbf{E}[Y|Z] = \begin{cases} \frac{b-a}{2-a}, & Z = 1, \\ 1, & Z = 0. \end{cases}$

□

4 Inequalities

4.1 Probability Inequalities

Definition 4.1 (Markov's Inequality)**.** For some nonnegative r.v. X and $t > 0$:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbf{E}[X]}{t}$$

Definition 4.2 (Chebyshev's Inequality)**.** Let $\mu = \mathbf{E}[X]$ and $\sigma^2 = \mathbf{Var}[X]$, then

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

and

$$\mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2}$$

where $Z = \frac{X-\mu}{\sigma}$. In particular, $\mathbb{P}(Z \geq 2) \leq \frac{1}{4}$ and $\mathbb{P}(Z \geq 3) \leq \frac{1}{9}$

Proof. $\frac{\mathbf{E}\left[(X-\mu)^2\right]}{t^2} \geq \mathbb{P}((X - \mu)^2 \geq t^2) = \mathbb{P}(|X - \mu| \geq t)$. The second case comes from $t = k\sigma$: $\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2}$. Then $\mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2}$ □

Definition 4.3 (Hoeffding's Inequality)**.** Let $Y_1 \dots Y_n$ be independent observations s.t. $\mathbf{E}[Y_i] = 0$ and $a_i \leq Y_i \leq b_i$. Let $\epsilon > 0$. For any $t > 0$, $\mathbb{P}((\sum_{i=0}^n Y_i) \geq \epsilon) \leq e^{-t\epsilon} \prod_{i=0}^n e^{t^2 \frac{(a_i - b_i)}{8}}$

Definition 4.4 (Bernoulli Presentation of Hoeffding's Inequality)**.** Let $X_1 \dots X_n \sim Bernoulli(p)$ with $\epsilon > 0$. Then:

$$\mathbb{P}(|\bar{X_n} - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Proof. Using Markov's bound, $\mathbb{P}(\sum_{i=0}^n Y_i \geq t\epsilon) = \mathbb{P}(e^{\sum_{i=0}^n Y_i} \geq e^{t\epsilon}) \leq \frac{\mathbf{E}\left[e^{\sum_{i=0}^n Y_i}\right]}{e^{t\epsilon}} = e^{-te} \prod_{i=0}^n e^{tY_i}$.

Now consider $\mathbf{E}\left[e^{tY_i}\right]$. We know $a_i \leq Y_i \leq b_i$, so we can express Y_i as a convex combination $Y_i = (1 - \alpha)a_i + \alpha b_i$ where $\alpha = \frac{Y_i - a_i}{b_i - a_i}$. Because e^x is a convex function, $e^{tY_i} \leq (1 - \alpha)e^{ta_i} + \alpha e^{tb_i}$. Because $\mathbf{E}[Y_i] = 0$, $\mathbf{E}[\alpha] = \frac{-a_i}{b_i - a_i}$. Indeed, $\mathbf{E}\left[e^{tY_i}\right] \leq \mathbf{E}\left[(1 - \alpha)e^{ta_i} + \alpha e^{tb_i}\right] = \frac{b_i e^{a_i} - a_i e^{b_i}}{b_i - a_i} = e^{g(u)}$ for some u, g.

We will make use of the exact form of Taylor's theorem: if g is a smooth function then there is a number $\xi \in (0, u)$ such that $g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\xi)$.

Some algebra is needed to $\frac{b_i e^{a_i} - a_i e^{b_i}}{b_i - a_i} = e^{g(u)}$ □

Note. Our bounds, like Hoeffding's, are statements about tail probabilities $\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon) \leq \alpha$. But confidence intervals require a something more like 'μ must lie within the ϵ neighborhood of $\hat{\mu}$'.

In general, recognize:

$$\begin{aligned} &\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon) \leq \alpha \\ &= \mathbb{P}(|\hat{\mu} - \mu| < \epsilon) > \alpha \\ &= \mathbb{P}(\mu \in (\hat{\mu} - \epsilon, \hat{\mu} + \epsilon)) > \alpha \end{aligned}$$

Immediately recognize the placement of terms to recover this statement about containment in an interval to avoid getting lost in the algebra. Reason about these statements geometrically.

Note. Hoeffding's provides an easy way to obtain confidence intervals. Let $\epsilon_n = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$, then

$$\mathbb{P}(|\bar{X_n} - p| \geq \epsilon_n) \leq \alpha$$

Which is exactly

$$\mathbb{P}(p \in (\bar{X_n} - \epsilon_n, \bar{X_n} + \epsilon_n)) \geq \alpha$$

Definition 4.5 (Mill's Inequality)**.** Let $Z \sim Normal(0, 1)$. For $t > 0$, $\mathbb{P}(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{\frac{-t^2}{2}}}{t}$

This result comes from another manipulation of the Markov's bound. We work this out in detail in the exercises.

Exercise 2.1. Let $X \sim Exponential(\beta)$. Find $\mathbb{P}(|X - \mu_X| < k\sigma_X)$ for $k > 1$. Compare this to the bound you get from Chebyshev's inequality.

Proof. Notice $\mathbb{P}(|X - \mu_X| > k\sigma_X) = \mathbb{P}(|X - \beta| > k\beta) = 1 - \mathbb{P}(|X - \beta| < k\beta)$. Write the central event $|X - \beta| < k\beta \iff \beta - k\beta < X < \beta + k\beta$. If $k > 1$, then $\beta - k\beta$ lies outside our support so our event simplifies to $\mathbb{P}(X < \beta + k\beta) = \mathbb{P}(X < \beta(1 + k)) = \int_{x=0}^{\beta(1+k)} f_X(x)dx = -e^{1+k} + 1$. $\mathbb{P}(|X - \mu_X| > k\sigma_X) = 1 - (-e^{1+k} + 1) = e^{1+k}$.

Chebyshev's bound is simply $\frac{1}{k^2}$. □

Note. Notice this bound goes like e^k, while the Chebyshev bound goes like $\frac{1}{k^2}$. For distributions with long exponential tails, this bound rapidly becomes quite poor.

Exercise 2.2. Let $X \sim Poisson(\lambda)$. Use Chebyshev's to show $\mathbb{P}(X \geq 2\lambda) \leq \frac{1}{\lambda}$

Proof. $\mathbb{P}(|X - \lambda| \geq \lambda) \leq \frac{\lambda}{\lambda^2} = \mathbb{P}(X \geq 2\lambda) \leq \frac{1}{\lambda}$. (Notice $\{|X - \lambda| \geq \lambda\} \supset \{X \geq 2\lambda\}$) □

Exercise 2.3. Let $X_1 \dots X_n \sim Bernoulli(p)$ and $\bar{X_n} = n^{-1} \sum_{i=0}^{n} X_i$. Bound $\mathbb{P}(|\bar{X_n} - p| \geq \epsilon)$ using Chebyshev's and Hoeffding's inequality. Show that Hoeffding's inequality produces a tighter bound when n is large.

Proof. By Chebyshev's (recall for i.i.d. X_i, $\mathbf{E}\left[n^{-1} \sum_{n=0}^{i} X_i\right] = \mathbf{E}[X_i]$), $\mathbb{P}(|X_n - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{n4\epsilon^2}$, where the second bound comes from the fact that $\frac{1}{4}$ is the largest value of $\mathbf{Var}[X]$ over p. We use the Bernoulli presentation of Hoeffding inequality and arrive at $2e^{2*n\epsilon^2}$

Let $\epsilon = 0.2$ and examine the bounds for $n = 10, 100, 1000$. The Chebyshev bound goes like $0.625, 0.0625, 0.00625$ while the Hoeffding goes like $0.8, 0.00067, 3.6e - 35$. Notice Chebyshev starts out stronger and quickly becomes order(s) of magnitude weaker. □

Exercise 2.4. Let $X_1 \dots X_n \sim Bernoulli(p)$. Fix $\alpha > 0$. Define $\epsilon_n = \sqrt{\frac{1}{2n} log(\frac{2}{\alpha})}$, $\hat{p} = n^{-1} \sum_{i=0}^{n} X_i$, and $C_n = (\hat{p} - \epsilon, \hat{p} + \epsilon)$.

- Use Hoeffding's to show $\mathbb{P}(C_n \text{ contains } p) \geq 1 - \alpha$
- Fix $\alpha = 0.05$ and $p = 0.4$. Conduct a simulation study with a computer to see how often the interval contains p (called coverage) for different values of n between 1 and 10000. Plot the coverage as a function of n.

- Plot the length of the interval versus n. Suppose we want the interval to be less than 0.05. How large should n be?

Proof.

- Using the Bernoulli presentation of Hoeffding's given, $\mathbb{P}(|\hat{X}_n - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$. Equivalently, $\mathbb{P}(|\hat{X}_n - p| \leq \epsilon) \geq 1 - 2e^{-2n\epsilon^2}$. Notice $\hat{p_n} = \bar{X_n}$ from this original definition. Plugging in ϵ_n, we have $\mathbb{P}(|\hat{p_n} - p| \leq \epsilon_n) \geq 1 - \alpha$ which is the same as saying $\mathbb{P}(C_n \text{ contains } p) \geq 1 - \alpha$ as desired.

- The interval shrinks roughly like $\sqrt{n}$. $\sqrt{\frac{log(\frac{2}{\alpha})}{2n}} \leq 0.025$ then $n \geq 2960$

□

Exercise 2.5. Prove Mill's inequality: $\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{t^2}{2}}}{t}$

Proof. Observe $\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t)$. We'll begin with one side. Let $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}$ be the PDF of Z. $\mathbb{P}(Z > t) = \int_{x=t}^{\infty} \phi(x)dx$. Because $\frac{x}{t} > 1$ when $x > t$, $\int_{x=t}^{\infty} \phi(x)dx \leq \int_{x=t}^{\infty} \frac{x}{t}\phi(x)dx$. Notice $\phi'(x) = -x\phi(x)$. Then $\int_{x=t}^{\infty} \frac{x}{t}\phi(x)dx = \frac{1}{t}[-\phi(x)]_{x=t}^{\infty} = \frac{\phi(t)}{t}$.

Then $\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t) \leq 2\frac{\phi(t)}{t} = \sqrt{\frac{2}{\pi}} \frac{1}{t} e^{\frac{-t^2}{2}}$ as desired. □

4.2 Inequalities for Expectation

5 Convergence

5.1 Types of Convergence

Proof. Now we prove (b). Fix $\epsilon > 0$ and let x be a continuity point of F. $F_n(x) = \mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X \geq x + \epsilon) \leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon) \leq F(x + \epsilon) + \mathbb{P}(|X - X_n| \geq \epsilon)$

Similarly, $F(x - \epsilon) = \mathbb{P}(X \leq x - \epsilon, X_n \leq x) + \mathbb{P}(X \leq x - \epsilon, X_n \geq x) \leq F_n(x) + \mathbb{P}(|X - X_n| \geq \epsilon)$.

We use these bounds to construct the inequality $F(x-\epsilon)-\mathbb{P}(|X-X_n| \geq \epsilon) \leq F_n(x) \leq F(x)+\mathbb{P}(|X-X_n| \geq \epsilon)$. Take the limit as n goes to ∞. Notice that $F_n(x)$ is a sequence with no guarantees of convergence, so we have to account for the largest and smallest elements. By our assumption $\mathbb{P}(|X - X_n| \geq \epsilon) \to 0$ as $n \to \infty$.

$F(x - \epsilon) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x + \epsilon)$

Now take the limit as $\epsilon \xrightarrow{0}$ (this statement holds for arbitrary $\epsilon > 0$). $F(x^-) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x^+)$. Using continuinty of F at x, $F(x^-) = F(x^+)$ and $F_n(x) = F(x)$ as desired. □

Note (Convergence in probability does not imply convergence in quadratic mean)**.** Let $U \sim Uniform(0, 1)$ and $X_n = \sqrt{n} I_{(0, \frac{1}{n})}(U)$. First see $X_n \xrightarrow{P} 0$. Indeed as n grows sufficiently large, for arbitrary ϵ, $\mathbb{P}(|X_n| \geq \epsilon) = \mathbb{P}(0 \geq U < n) = n \to 0$. However, $\mathbf{E}\left[X_n^2\right] = \int_{u=0}^{\frac{1}{n}} n du = 1$.

Note (Convergence in distribution does not imply convergence in probability)**.** Let $X \sim Normal(0, 1)$. For each n, let $X_n = -X$. $F_n(x) = F(x)$ for any x. But $\mathbb{P}(|X - X_n| \geq \epsilon) = \mathbb{P}(|2X| \geq \epsilon) = \mathbb{P}(|X| \geq \frac{\epsilon}{2}) \neq 0$. Indeed the symmetrical shape of the Gaussian preserves the CDF but any given values are negatives of each other and will never converge.

5.2 The Law of Large Numbers

A crowned jewel of probability. The sample mean approaches the expectation of the underlying distribution.

Definition 5.1. Let $X_1 \dots X_n$ be i.i.d. with $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2$. Let $\bar{X_n} = n^{-1} \sum_{i=0}^{n} X_i$. Then $\bar{X_n} \xrightarrow{P} \mu$.

The sample mean is a **random variable** so it will never numerically equal the expectation. It will cluster closer and closer to it.

5.3 The Central Limit Theorem

While the LLN tells us the sample mean clusters around the true mean, it does not give us tools to approximate statements about probability.

Definition 5.2 (CLT)**.** Let $X_1 \dots X_n$ be i.i.d. with $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2$. Let $\bar{X_n} = n^{-1}\sum_{i=0}^{n} X_i$. Then $Z_n = \frac{\sqrt{n}(\bar{X_n}-\mu)}{\sigma} \xrightarrow{D} N(0,1)$

In other words, $\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \mathbb{P}(Z \le z) = \phi(z)$

Note we use this to approximate **probability statements** not the distribution $\bar{X_n}$ itself.

Note. By scaling/shifting Z, the CLT really gives us a family of limiting distributions. We often refer to $\sqrt{n}(\bar{X_n} - \mu)$ as the 'canonical scaling'. This is because it leads to a non-degenerate limit, approximated by $N(0, \sigma^2$ where neither mean or variance depend on n.

Another way of thinking about this 'canonical' member is as result of the minimal number of steps needed to 'remove the dependence on n' from both mean and variance. If one scales by $\sqrt{n}$, the mean is now $\sqrt{n}\mu$. If one shifts by μ, the n remains in the variance.

These approximations are not perfect and indeed we can bound the error:

Definition 5.3 (The Berry-Esseen Inequality)**.** $sup_x|\mathbb{P}(Z_n \le x) - \phi(x)| \le \frac{37}{4}\frac{\mathbf{E}\left[|X_i-\mu|^3\right]}{\sqrt{n}\sigma^3}$

Where the *sup* bounds the difference across all possible x in the domain.

5.4 Delta Method

This is like the CLT for functions of the sample mean.

Exercise 2.1. Let $X_1 \dots X_n$ be i.i.d. with $EX = \mu$ and $\mathbf{Var}[X] = \sigma^2$. Let $S_n = \frac{1}{1-n}\sum_{i=1}^{n}(X_i - \bar{X_n})$

- Show $\mathbf{E}[S_n] = \sigma^2$

Proof.

- Recognize $\sum_{i=1}^{n}(X_i-\bar{X_n})^2 = \sum_{i=1}^{n}(X_i^2-2X_i\bar{X_n}+\bar{X_n}^2) = \sum_{i=1}^{n}(X_i^2)-2n\bar{X_n}\bar{X_n}+n\bar{X_n}^2 = \sum_{i=1}^{n}(X_i^2)-n\bar{X_n}^2 = nX_i^2-n\bar{X_n}^2$. Our original expectation simplifies like $\mathbf{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X_n})^2\right] = \frac{1}{n-1}\mathbf{E}\left[nX_i^2\right] - \mathbf{E}\left[n\bar{X_n}^2\right] = \frac{1}{n-1}(n(\sigma^2-\mu^2) - n(\frac{\sigma^2}{n}-\mu^2)) = \sigma^2$
-

□

Exercise 2.2. Let $X_1 \dots X_n$ be a sequence. Show $X_n \xrightarrow{QM} b$ if and only if $\mathbf{E}[X_n] \to b$ and $\mathbf{Var}[X_n] \to 0$

Proof. The key identity is $\mathbf{E}\left[(X_n-b)^2\right] = \mathbf{E}\left[X_n^2\right] - (\mathbf{E}[X_n]^2 + \mathbf{E}[X_n]^2) - 2\mathbf{E}[X_n]b + b^2 = \mathbf{Var}[X_n] + (\mathbf{E}[X_n]-b)^2$. The reverse implication follows immediately. To see the forward argument, notice $\mathbf{Var}[X_n] \le \mathbf{E}\left[(X_n-b)^2\right] \to 0$ and $(\mathbf{E}[X_n]-b)^2 \le \mathbf{E}\left[(X_n-b)^2\right] \to 0$. Then $\mathbf{Var}[X_n] \to 0$ and $\mathbf{E}[X_n] \to b$. □

Exercise 2.3. Let $X_1 \dots X_n$ be i.i.d. and let $\mu = \mathbf{E}[X_1]$. Suppose the variance is finite. Show $\bar{X_n} \xrightarrow{QM} \mu$

Proof. Expand $\mathbf{E}\left[(\bar{X_n}-\mu)^2\right] = \mathbf{Var}[X_n] + \mathbf{E}\left[\bar{X_n}^2\right] - 2\mu\mathbf{E}\left[\bar{X_n}\right] + \mu^2 = \frac{\sigma^2}{n} \to 0$ □

Exercise 2.4. Let $X_1, X_2 \dots$ be a sequence of r.v.s such that $\mathbb{P}(X_n = \frac{1}{n}) = 1 - \frac{1}{n^2}$ and $\mathbb{P}(X_n = n) = \frac{1}{n^2}$. Does X_n converge in probability? Does X_n converge in quadratic mean?

Proof. Intuitively X_n approaches 0 with increasing probability. Using Markov's, $\mathbb{P}(|X_n - 0| > \epsilon) \le \frac{\mathbf{E}[X_n]}{\epsilon}$ where $\mathbf{E}[X_n] = \frac{1}{n}(1-\frac{1}{n^2}) + n(\frac{1}{n^2}) \to 0$. We can conclude $X_n \xrightarrow{P} 0$. However, $\mathbf{E}\left[(X_n-0)^2\right] = \frac{1}{n^2}(1-\frac{1}{n^2}) + n^2(\frac{1}{n^2}) \to 1$. So X_n does not converge in quadratic mean. □

Exercise 2.5. Let $X_1 \dots X_n$ be i.i.d. $Bernoulli(p)$. Show $\frac{1}{n}\sum_{i=0}^{n} X_i^2$ converges in probability and quadratic mean to p.

Proof. $\mathbf{E}\left[(\frac{1}{n}\sum_{i=0}^{n} X_i)^2 - 2p\frac{1}{n}\sum_{i=0}^{n} X_i^2 + p^2\right] = \mathbf{E}\left[(\frac{1}{n}\sum_{i=0}^{n} X_i)^2\right] - 2p^2 + p^2$. Notice $\mathbf{E}\left[(\frac{1}{n}\sum_{i=0}^{n} X_i)^2\right]$ has n terms of expectation p (the diagonals) and $n^2 - n$ terms of expectation p^2. Then $\mathbf{E}\left[(\frac{1}{n}\sum_{i=0}^{n} X_i)^2\right] = \frac{1}{n^2}((n^2-n)p^2 + np)$. The entire expectation is then $\frac{1}{n^2}((n^2-n)p^2 + np) - p^2 = \frac{1}{n}(p^2 - p) \to 0$. Because convergence in qm implies convergence in probability, this sum also converges in probability. □

Exercise 2.6. Assume average height of men is 68 inches and standard deviation 2.6 inches. Draw 100 men at random. Find the approximate probability the average height will be at least 68 inches.

Proof. By the CLT, we know $\frac{\sqrt{n}(X_n - \mu)}{\sigma} \xrightarrow{D} Z$ so $\mathbb{P}(\bar{X}_n \geq 68) = \mathbb{P}(Z \geq 0)$. Indeed $\mathbb{P}(Z \geq 0) = 1 - \mathbb{P}(Z \leq 0) = \frac{1}{2}$ □

6 Models, Statistical Inference and Learning

6.1 Introduction

Statistical inference is the process of using data to infer the distribution that generated the data. A typical question is:

Given a sample $X_1 \dots X_n \sim F$, how do we infer F ?

6.2 Parametric and Nonparametric Models

6.3 Fundamental Concepts in Inference

Many inferential problems are one of three types: point estimation, confidence sets or hypothesis testing.

Definition 6.1. The bias of an estimator is defined by $bias(\hat{\theta}_n) = \mathbf{E}\left[\hat{\theta}_n\right] - \theta$

We say that $\hat{\theta}_n$ is **unbiased** if $bias(\hat{\theta}_n) = 0$.

Definition 6.2. An estimator $\hat{\theta}_n$ of a parameter θ is **consistent** if $\hat{\theta}_n \xrightarrow{P} \theta$.

Definition 6.3. The distribution of $\hat{\theta}_n$ is called the **sampling distribution**. The **standard error** of $\hat{\theta}_n$ is defined as $\sqrt{\mathbf{Var}\left[\hat{\theta}_n\right]}$.

Intuitively the bias is a good measure of the mean of our estimator, but does not rule out things like long tails. Asymptotically, consistency is a stronger condition and lets us know if the entire distribution of the estimator clusters around the parameter instead of just the mean.

Theorem 6.4. $\text{MSE}(\hat{\theta}) = \mathbf{Var}\left[\hat{\theta}\right] + \text{bias}^2(\hat{\theta})$

Proof. $\mathbf{E}\left[(\hat{\theta} - \theta)^2\right] = \mathbf{E}\left[(\hat{\theta} - \bar{\hat{\theta}} + \bar{\hat{\theta}} - \theta)^2\right] = \mathbf{E}\left[(\hat{\theta} - \bar{\hat{\theta}})^2\right] + \mathbf{E}\left[2(\hat{\theta} - \bar{\hat{\theta}})(\bar{\hat{\theta}} - \theta)\right] + \mathbf{E}\left[(\bar{\hat{\theta}} - \theta)^2\right]$. The middle terms drop out ($\mathbf{E}\left[\hat{\theta} - \bar{\hat{\theta}}\right]) = 0$) and we are left with $\mathbf{Var}\left[\hat{\theta}\right] + \text{bias}^2(\hat{\theta})$ □

Definition 6.5. A $1 - \alpha$ **confidence interval** for a parameter θ is some interval $C_n = (a, b)$ where $a = a(X_1 \dots X_n)$ and $b = b(X_1 \dots X_n)$ are functions of the data such that $\mathbb{P}(\theta \in C) \geq 1 - \alpha$.

Exercise 2.1. Let $X_1 \dots X_n \sim Poisson(\lambda)$ and $\hat{\lambda} = n^{-1}\sum_{i=1}^{n} X_i$. Find the bias, se and MSE of the estimator.

Proof. $\text{bias}^2(\hat{\lambda}) = \mathbf{E}\left[\hat{\lambda}\right] - \lambda = 0$. $\text{se}(\hat{\lambda}) = \sqrt{\frac{\lambda}{n}}$.

We will compute the MSE directly for practice rather than using the decomposition of bias and variance. $\text{MSE}(\hat{\lambda}) = \mathbf{E}\left[(\hat{\lambda} - \lambda)^2\right] = \mathbf{E}\left[\hat{\lambda}^2\right] - 2\,\mathbf{E}\left[\hat{\lambda}\right]\mathbf{E}\left[\lambda\right] + \mathbf{E}\left[\lambda^2\right] = \frac{\lambda}{n} + \lambda^2 - 2\lambda^2 + \lambda^2 = \frac{\lambda}{n}$. □

Exercise 2.2. Let $X_1 \dots X_n \sim Uniform(\theta)$ and $\hat{\theta} = max\{X_1 \dots X_n\}$. Find the bias, se and MSE of the estimator.

Proof. Let $M = max\{X_1 \dots X_n\}$. The CDF for M is $\mathbb{P}(M \leq x) = (\frac{x}{\theta})^n$. The density is then $n\frac{x^{n-1}}{\theta^n}$. $\mathbf{E}[M] = \int_{x=0}^{\theta} n\frac{x^n}{\theta^n}dx = [\frac{1}{\theta^n}\frac{n}{n+1}x^{n+1}]_{x=0}^{\theta} = \frac{n}{n+1}\theta$ $\text{bias}^2(\hat{\theta}) = -\frac{1}{n+1}\theta$ □

7 Estimating the CDF and Statistical Functionals

Definition 7.1. Consider i.i.d. $X_1 \dots X_n \sim F$. The empirical distribution function $\hat{F}_n$ is defined $\hat{F}_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x)$

Theorem 7.2. *For any fixed value x,*

$\mathbf{E}\left[\hat{F_n}(x)\right] = F(x)$

$\mathbf{Var}\left[\hat{F_n}(x)\right] = n^{-1}(1 - F(x))F(x)$

$\text{MSE}\,\hat{F_n}(x) = n^{-1}(1 - F(x))F(x)$

$\hat{F_n}(x) \xrightarrow{P} F(x)$

Proof. $\mathbf{E}\left[\hat{F_n}(x)\right] = n^{-1}n\,\mathbf{E}[F_n(x)] = F(x)$. We can see $\mathbf{E}[F_n(x)] = F(x)$ by computing expectation of the indicator directly, where $\mathbf{E}[F_n(x)] = \mathbb{P}(X_i \leq x) = F(x)$ or by treating $F_n(x)$ as a coin flip with probability of heads of $F(x)$.

$\mathbf{Var}\left[\hat{F_n}(x)\right] = n^{-2}n\,\mathbf{Var}[F_n(x)] = n^{-1}(1 - F(x))F(x)$ where $\mathbf{Var}\left[\hat{F_n}(x)\right] = (1 - F(x))F(x)$ similarly because $\hat{F_n}(x)$ can be thought of as Bernoulli.

$\text{MSE}\,\hat{F_n}(x) = \mathbf{E}\left[(\hat{F_n}(x) - F(x))^2\right] = \mathbf{Var}\left[\hat{F_n}(x)\right]$.

$\hat{F_n}(x) \xrightarrow{MSE} F(x) \implies \hat{F_n}(x) \xrightarrow{P} F(x)$ by basic properties of convergence □

Recognize for fixed x, the value of the empirical distribution function can be treated as a sample mean of i.i.d. Bernoulli ($p = \mathbb{P}(X_i \leq x)$). The above proof shows this idea in action.

Definition 7.3 (Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)**.** Let $X_1 \dots X_n \sim F$. Then for any $\epsilon > 0$,

$$\mathbb{P}(\sup_x |F(x) - \hat{F_n}(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Note. We recall that Hoeffding's inequality was introduced to give a tight bound on the sample mean of i.i.d. Bernoulli:

$$\mathbb{P}(|\bar{X_n} - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Recall, when $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$, the bound is exactly α. This provides the $1 - \alpha$ confidence interval we desire.

The inequality we introduce next uses this same idea.

Definition 7.4 (Nonparametric $1 - \alpha$ Confidence Band for F)**.**

The supremum guarantees this holds even for the largest value of F.

Exercise 2.1. Let $X_1 \dots X_n \sim Bernoulli(p)$ and $Y_1 \dots Y_m \sim Bernoulli(q)$. Find the estimator, estimated SE and approximate 95 percent CI for p and $p - q$.

Proof. We start with p. Our estimator is exactly the plug-in estimator for the mean: $\hat{p}_n = n^{-1}\sum_{i=1}^{n} X_i$. The $\text{SE}(\hat{p}_n) = \frac{\sqrt{\sigma}}{n}$. Then our estimated standard error is $\hat{\text{SE}}(\hat{p}_n) = \frac{\sqrt{\hat{\sigma}_n}}{n}$. We can use the plug-in estimator for the variance of $\hat{p}_n$ to estimate $\hat{\sigma}_n$. This estimator is $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. The estimated error is then $\hat{\text{SE}}(\hat{p}_n) = \frac{\sqrt{n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}{n}$. To find the CI, we recognize $\hat{p}_n \sim Normal(p, \text{SE}(\hat{p}_n))$ as n grows large by

C.L.T. Let $z_{95} = \Phi^{-1}(95)$, then $(\frac{\hat{p}_n - p}{\sqrt{\hat{\mathrm{SE}}(\hat{p}_n)}} - z_{95}, \frac{\hat{p}_n - p}{\sqrt{\hat{\mathrm{SE}}(\hat{p}_n)}} + z_{95})$ is our approximate 90 percent confidence interval.

Now let $\theta = p - q$. $\hat{\theta}_n = n^{-1}\sum_{i=1}^n X_i - m^{-1}\sum_{j=1}^m Y_j$. $\hat{\mathrm{SE}}(\hat{\theta}_n) = \frac{\sqrt{n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}}{n} + \frac{\sqrt{m^{-1}\sum_{j=1}^m (Y_i - \bar{Y}_m)^2}}{m}$. $\hat{\theta}_n$ is similarly distributed over $Normal(\theta, \hat{\mathrm{SE}}(\hat{\theta}_n))$ $(\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\mathrm{SE}}(\hat{\theta}_n)}} - z_{95}, \frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\mathrm{SE}}(\hat{\theta}_n)}} + z_{95})$ is our approximate 90 percent confidence interval. □

Exercise 2.2. Let $X_1 \dots X_n \sim F$ and $\hat{F}_n$ be the empirical distribution function. For fixed x, use the CLT to find the limiting distribution of $\hat{F}_n(x)$

Proof. For fixed x, let $Y_i = I(X_i \leq x)$ so $\hat{F}_n(x) = \bar{Y}_n$. Then, by CLT, $\bar{Y}_n \approx N(\mu, \frac{\sigma^2}{n})$ where $\mu = \mathbb{P}(X_i \leq x)$ and $\sigma^2 = \mathbb{P}(X_i \leq x)(1 - \mathbb{P}(X_i \leq x))$. This approximation of $\hat{F}_n(x)$ is $N(\mathbb{P}(X_i \leq x), \frac{\mathbb{P}(X_i \leq x)(1 - \mathbb{P}(X_i \leq x))}{n})$. We can center and scale to find a limiting distribution that is not degenerate (mean and variance are independent of n): $\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))$. □

Proof. Let x and y be distinct points. Find $\mathbf{Cov}\left[\hat{F}_n(x), \hat{F}_n(y)\right]$. □

Exercise 2.3. $\mathbf{Cov}\left[\hat{F}_n(x), \hat{F}_n(y)\right] = \mathbf{Cov}\left[n^{-1}\sum_{i=1}^n 1(X_i \leq x), n^{-1}\sum_{j=1}^n 1(X_j \leq y)\right]$ Because covariance is bilinear, this is $n^{-2}\sum_{i=1}^n \sum_{j=1}^n \mathbf{Cov}\left[1(X_i \leq x), 1(X_j \leq y)\right]$. When $i \neq j$, the indicators are independent so the covariance drops out. We are left with the diagonals: $n^{-2} n\, \mathbf{Cov}\left[1(X_1 \leq x), 1(X_1 \leq y)\right]$. $\mathbf{Cov}\left[1(X_1 \leq x), 1(X_1 \leq y)\right] = \mathbf{E}\left[1(X_1 \leq x)1(X_1 \leq y)\right] - F(x)F(y) = \mathbb{P}(X_1 \leq \min\{x, y\})) - F(x)F(y) = F(\min\{x, y\}) - F(x)F(y)$. Our final answer is then $n^{-1}(F(\min\{x, y\}) - F(x)F(y))$.

Exercise 2.4. Let $X_1 \dots X_n \sim F$ and $\hat{F}_n$ be the empirical distribution function. Let $a < b$ be fixed numbers and define $\theta = T(F) = F(b) - F(a)$. Let $\hat{\theta} = T(\hat{F}_n) = \hat{F}_n(b) - \hat{F}_n(a)$. Find the estimated standard error of $\hat{\theta}$ and an expression for an approximate $1 - \alpha$ confidence interval for θ.

Proof. First, recognize $\hat{\theta} = \hat{F}_n(b) - \hat{F}_n(a)$ can be rewritten as a sample mean of Bernoulli, $Y_i = 1(X_i \leq b) - 1(X_i \leq a)$ where $\mathbf{E}\left[Y_i\right] = \theta$. Then $\hat{\theta} = \bar{Y}_n$ and $\mathrm{SE}\left(\hat{\theta}\right) = \sqrt{\frac{\theta(1-\theta)}{n}}$. θ is unknown so our approximate error is $\hat{\mathrm{SE}}\left(\hat{\theta}\right) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$.

To find our confidence intervals, recognize $\bar{Y}_n \xrightarrow{d} N(\theta, \frac{\theta(1-\theta)}{n})$ by CLT. Then, $\frac{\sqrt{n}\bar{Y}_n - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0, 1)$. Thus $\mathbb{P}(|\frac{\sqrt{n}\bar{Y}_n - \theta}{\sqrt{\theta(1-\theta)}}| \leq z_{1-\frac{\alpha}{2}}) \approx 1 - \alpha$. Moving terms around: $\mathbb{P}(|\hat{\theta} - \theta| \leq \sqrt{\frac{\theta(1-\theta)}{n}} z_{1-\frac{\alpha}{2}}) \approx 1 - \alpha$. □

Exercise 2.5. Using the earthquake magnitude data provided, estimate $F(x)$ then compute and plot the 95 percent confidence interval for F. Then compute an approximate 95 percent CI for $F(4.9) - F(4.3)$.

Proof. Using the DKW inequality, the 95% confidence band is

$$L(x) = \hat{F}_n(x) - \epsilon_n, \quad U(x) = \hat{F}_n(x) + \epsilon_n$$

where $\epsilon_n = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$ with $\alpha = 0.05$. See Fig. 5.

For $\theta = F(4.9) - F(4.3)$, we use the plug-in estimate $\hat{\theta} = \hat{F}_n(4.9) - \hat{F}_n(4.3)$ with Wald interval

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

giving the 95% CI $[0.495, 0.557]$. See `code/7.7.py` for implementation.

□

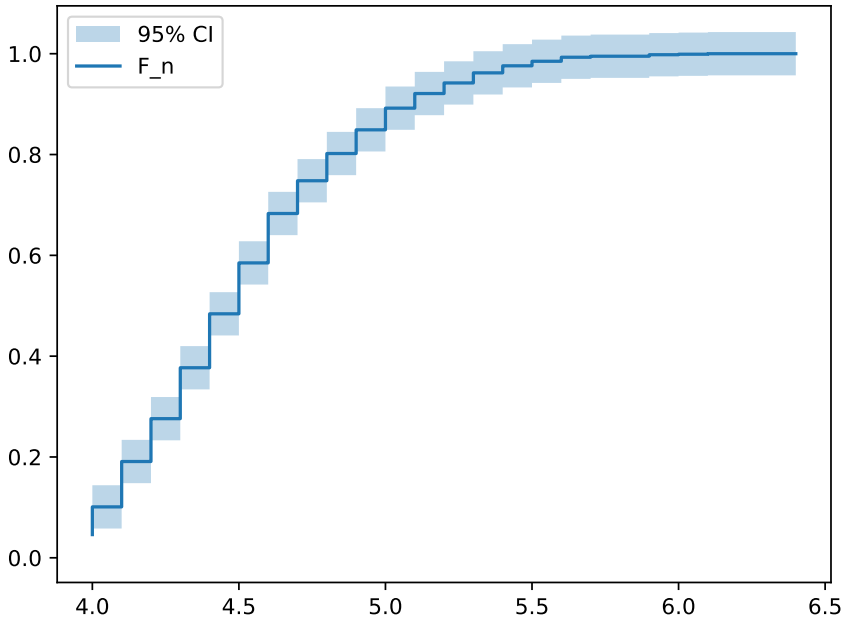


Figure 5: 95% confidence band for earthquake magnitude CDF

8 The Bootstrap

8.1 Simulation

If $Y_1 \dots Y_n \sim G$, by the law of large numbers:

$$\frac{1}{B}\sum_{i=1}^{B} Y_i \xrightarrow{P} \mathbf{E}\left[Y_i\right]$$

as $B \to \infty$. This is also true for arbitrary functions $g(Y_i)$:

$$\frac{1}{B}\sum_{i=1}^{B} g(Y_i) \xrightarrow{P} \mathbf{E}\left[g(Y_i)\right]$$

Recognize the $\mathbf{Var}\left[Y_i\right]$ can be approximated in this way:

$$\frac{1}{B}\sum_{i=1}^{B} Y_i^2 - (\frac{1}{B}\sum_{i=1}^{B} Y_i)^2 \xrightarrow{P} \mathbf{E}\left[Y_i^2\right] - \mathbf{E}\left[Y_i\right]^2 = \mathbf{Var}\left[X_i\right]$$

So if we somehow have a simulation where we can draw from G as many times as we like, we can get arbitrarily close to $\mathbf{E}\left[g(Y_i)\right]]$ using the LLN.

Note. Developing the bootstrap in this way, it is clear it is a method for estimating the **sampling behavior** of an estimator and a generic one at that. It is not restricted to the variance but arbitrary functions of our estimator, like the CDF as we will see in pivotal confidence intervals.

Recall, we can always construct the estimator so we can compute it from data in practice. If we can't do this, its not a useful estimator in the first place.

8.2 Bootstrap Variance Estimation

So when are we able to simulate the distribution that generates a random variable?

Well if we know that $\hat{F}_n$ generates our data, which is the case with plug-in estimates, we have a full description of the distribution in "this world". Simulation just amounts to picking out items from the set $\{X_1 \dots X_n\}$ we used to construct $\hat{F}_n$ in the first place.

Our bootstrap estimation is then quite a simple procedure:

Definition 8.1.

- Draw n observations $X_1^* \dots X_n^*$ with replacement from $X_1 \dots X_n$.
- Compute T_n^* from $X_1^* \dots X_n^*$.
- Repeat 1. and 2. to get $T_{n,1}^* \dots T_{n,B}^*$.

While maybe obvious, drawing n observations $X_1^* \dots X_n^*$ with replacement from $X_1 \dots X_n$ is not the same as outright using the set $X_1 \dots X_n$ even though they both use the same n and come from the same set. The former introduces randomness distributed exactly as $\hat{F}_n$. The difference is incredibly important when understanding eg. pivotal confidence intervals.

Note. A common use of the bootstrap is computing $\mathbf{Var}_{\hat{F}_n}[T_n]$. This is:

$$\frac{1}{B}\sum_{i=1}^{B} T_{n,i}^* - (\frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*)^2$$

Recall this is the variance of the **sampling distribution** of some statistic, possibly the variance itself, not of the original observed data

8.3 Bootstrap Confidence Intervals

We often use the bootstrap to build confidence intervals for our sampling distribution. We'll explore some techniques.

Normal Interval. The simplest technique assumes T_n can be approximated by a normal distribution: $T_n z_{1-\alpha} \hat{se}_{\text{boot}}$. Recall its actually quite difficult to estimate the standard error for many estimators.

Pivotal Intervals.

Definition 8.2 (Pivot Intervals)**.** Let $\theta = T(F)$ and $\hat{\theta}_n = T(\hat{F}_n)$. We define our **pivot** as $R = \hat{\theta}_n - \theta$.

Let $H(r) = \mathbb{P}(R \leq r)$ be the CDF of the pivot. $\mathbb{P}(a \leq \theta \leq b) = 1 - \alpha$ if $a = \hat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2})$ and $b = \hat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2})$ making this an exact confidence interval.

However because H is a function of F, we often compute a bootstrap estimation $\hat{H}(r) = B^{-1}\sum_{i=1}^{B} 1(R_{n,b}^* \leq r)$ and build an approximate confidence interval.

Where $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$. Denote R_β^* as the β sample quantile of $(R_{n,i}^* \dots R_{n,b}^*)$ and $\hat{\theta}_\beta^*$ as the β sample quantile of $(\theta_{n,i}^* \dots \theta_{n,b}^*)$. Notice $R_\beta^* = \hat{\theta}_\beta^* - \hat{\theta}$.

Our confidence interval is then $(2\hat{\theta} - \hat{\theta}_{1-\frac{\alpha}{2}}^*, 2\hat{\theta} - \hat{\theta}_{\frac{\alpha}{2}}^*)$.

Percentile Intervals . This is defined using direct bootstrap quantiles:

$$(\theta_{\frac{\alpha}{2}}^*, \theta_{1-\frac{\alpha}{2}}^*)$$

Which is very intuitive. Just grab the chunk of bootstraped θs that correspond to α of density and use the endpoints for the interval.

Note. The pivotal and percentile intervals look very similar to each other. What is the 'big idea' difference between them.

$$\theta + (\theta - q) = 2\theta - q$$

Exercise 2.3. Let $X_1, \ldots, X_n \sim t_3$ where $n = 25$. Let $\theta = T(F) = (q_{.75} - q_{.25})/1.34$ where q_p denotes the pth quantile. Do a simulation to compare the coverage and length of the following confidence intervals for θ: (i) Normal interval with standard error from the bootstrap, (ii) bootstrap percentile interval, and (iii) pivotal bootstrap interval.

Proof. See `code/8.3.py`. □

Exercise 2.5. Let $X_1 \ldots X_n$ be distinct observations without ties. Let $X_1^* \ldots X_n^*$ be a bootstrap sample and $\bar{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$. Find $\mathbf{E}\left[\bar{X}_n^* | X_1 \ldots X_n\right], \mathbf{Var}\left[\bar{X}_n^* | X_1 \ldots X_n\right], \mathbf{E}\left[\bar{X}_n^*\right], \mathbf{Var}\left[\bar{X}_n^*\right]$.

Proof. $\mathbf{E}\left[\bar{X}_n^* | X_1 \ldots X_n\right] = \mathbf{E}\left[n^{-1} \sum_{i=1}^n X_i^* | X_1 \ldots X_n\right] = n^{-1} \sum_{i=1}^n \mathbf{E}\left[X_i^* | X_1 \ldots X_n\right]$. Recognize $\mathbf{E}\left[X_i^* | X_1 \ldots X_n\right]$, the expectation of a draw from our observed data, is $\sum_{i=1}^n X_i n^{-1} = \bar{X}_n$. $n^{-1} \sum_{i=1}^n \mathbf{E}\left[X_i^* | X_1 \ldots X_n\right] = \bar{X}_n$.

$\mathbf{E}\left[\bar{X}_n^*\right] = \mathbf{E}\left[\mathbf{E}\left[\bar{X}_n^* | X_1 \ldots X_n\right]\right] = \mathbf{E}\left[\bar{X}_n\right] = \mu$.

$\mathbf{Var}\left[\bar{X}_n^* | X_1 \ldots X_n\right] = n^{-2} \sum_{i=1}^n \mathbf{Var}\left[X_i | X_1 \ldots X_n\right] = n^{-2} \sum_{i=1}^n (n^{-1} \sum_{i=1}^n X_i^2 - \bar{X}_n{}^2) = n^{-2}(\sum_{i=1}^n X_i^2 - n\bar{X}_n{}^2) = n^{-1}(n^{-1}(\sum_{i=1}^n X_i^2 - \bar{X}_n{}^2))$. We can use the same algebraic property to simplify sample variance: $n^{-1}(n^{-1}(\sum_{i=1}^n X_i^2 - \bar{X}_n{}^2)) = n^{-2}(\sum_{i=1}^n (X_i - \bar{X}_n)^2))$

□

Exercise 2.7. Let $X_1 \ldots X_n \sim Uniform(0, \theta)$ and $\hat{\theta} = \max\{X_1 \ldots X_n\}$. Generate a dataset of size 50 with $\theta = 1$

(a) Compute the distribution for $\hat{\theta}$ and compare to the bootstrap histogram.

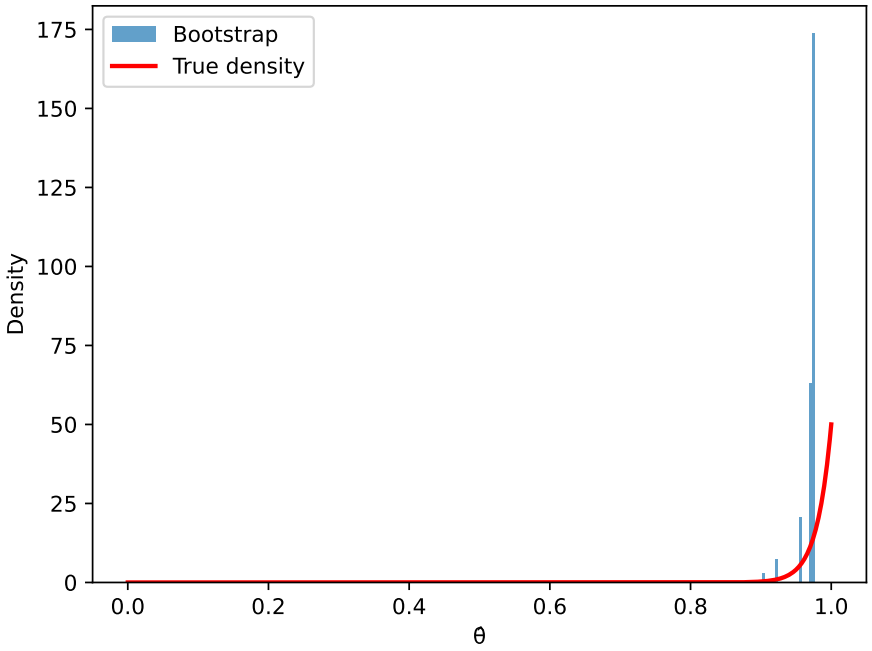
(b) This is a case where the bootstrap is poor. Show $\mathbb{P}(\hat{\theta} = \theta) = 0$ and $\mathbb{P}(\hat{\theta}^* = \hat{\theta}) \approx 0.632$.

Note. Wasserman writes $\mathbb{P}(\hat{\theta} = \hat{\theta})$ but almost certainly means $\mathbb{P}(\hat{\theta} = \theta)$.

Proof.

(a) We can find the density of $\hat{\theta}$ by starting with the CDF $F(x) = \prod_i \mathbb{P}(X_i \leq x) = (\frac{x}{\theta})^n$. Then $f(x) = \frac{n x^{n-1}}{\theta^n}$.

(b) $\hat{\theta}$ has continuous density and θ is a scalar so the first expression has value 0. The second expression is $1 - (\frac{n-1}{n})^n$, the complement of missing the maximum for each of the n draws in our bootstrap replication.



□

9 Parametric Inference

We are now concerned with models of the form

$$\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$$

9.1 Parameter of Interest

9.2 Method of Moments

Suppose our parameter θ has k components: $(\theta_1 \dots \theta_k)$. For $1 \leq j \leq k$, our jth moment is

$$\alpha_j(\theta) = \int x^j dF_\theta(x)$$

Our jth sample moment is:

$$\hat{\alpha}_j = n^{-1} \sum_i X_i^j$$

Definition 9.1 (Method of moments estimator)**.** The method of moments estimator $\hat{\theta}_n$ is the value of θ that satisfies the following system of k equations:

$$\alpha_1(\hat{\theta}_n) = \hat{\alpha}_1$$
$$\vdots$$
$$\alpha_k(\hat{\theta}_n) = \hat{\alpha}_k$$

9.3 Maximum Likelihood

Definition 9.2 (Maximum likelihood method)**.**

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

Where the log-likelihood is $\log \mathcal{L}_n(\theta)$

Definition 9.3. The maximum likelihood estimator (MLE) is exactly the $\hat{\theta}$ that maximizes $\mathcal{L}_n(\theta)$.

This is simply the joint density of data parameterized by θ. $\mathcal{L}_n : \theta \mapsto [0, \infty)$.

Example 9.4 (MLE of i.i.d. Normals)**.** $X_1 \dots X_n \sim Normal(\mu, \sigma^2)$

9.4 Properties of Maximum Likelihood Estimators

Under specific conditions, the MLE $\hat{\theta}$ possess properties that make it an appealing estimator:

1. The MLE is consistent: $\hat{\theta} \xrightarrow{P} \theta_*$ where θ_* is the true value of the estimated parameter θ.

9.5 Consistency of Maximum Likelihood Estimators

Definition 9.5 (Kullback-Leibler distance)**.** If f, g are densities:

$$D(f, g) = \int f(x) \log(\frac{f(x)}{g(x)}) dx$$

denotes the KL distance between f and g.

Work out that $D(f, f) = 0$ and $D(f, g) \geq 0$.

Denote $M_n(\theta) = n^{-1} \sum_i \log(\frac{f(X_i;\theta)}{f(X_i;\theta_*)})$. Notice by law of large numbers:

$$n^{-1} \sum_i \log(\frac{f(X_i;\theta)}{f(X_i;\theta_*)}) \xrightarrow{P} \mathbf{E}\left[(\log(\frac{f(X_i;\theta)}{f(X_i;\theta_*)})\right]$$

Where:

$$\mathbf{E}\left[(\log(\frac{f(X_i;\theta)}{f(X_i;\theta_*)})\right] = \int \log(\frac{f(X_i;\theta)}{f(X_i;\theta_*)}) f(X_i;\theta_*) = -\int f(X_i;\theta_*) \log(\frac{f(X_i;\theta_*)}{f(X_i;\theta)}) dx = -D(\theta_*, \theta)$$

Note. Our $D(\theta_*, \theta)$ notation is shorthand for $D(f(X_i;\theta_*), f(X_i;\theta))$.

Now it is not enough to show M_n converges to D for single values of θ. We need to show this for all θ and that M_n is well-behaved: θ_* is actually its global maximum and it doesn't do weird things as we get far away.

Theorem 9.6.

9.6 Equivariance of Maximum Likelihood Estimators

Theorem 9.7 (Equivariance)**.** *Let $\tau = g(\theta)$ be a function of θ and $\hat{\theta}_n$ be the MLE of θ. Then $\hat{\tau}_n$ is the MLE of τ.*

Proof. Let $h = g^{-1}$. $\mathcal{L}(\tau) = \prod_i f(X_i;\tau) = \prod_i f(X_i; h(\tau)) = \prod_i f(X_i;\theta) = \mathcal{L}(\theta)$. Then $\mathcal{L}(\theta) \leq \mathcal{L}(\hat{\theta}_n) \implies \mathcal{L}(\tau) \leq \mathcal{L}(\hat{\tau})$ □

Essentially we can reparameterize our likelihood by some function of θ and its 'shape' won't change.

9.7 Asymptotic Normality of Maximum Likelihood Estimators

The distribution of $\hat{\theta}_n$ is approximately normal. This allows it to compute it analytically instead of relying on methods like the bootstrap.

We need some definitions:

Definition 9.8 (Fisher Information)**.** The **score function** is:

$$s(X;\theta) = \frac{d\log(f(X;\theta))}{d\theta}$$

The **Fisher information** is:

$$I_n(\theta) = \sum_i \mathbf{Var}\left[s(X;\theta)\right]$$