Wasserman: All of Statistics

Kenny Workman

August 6, 2025

1 Probability

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

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.

1.2 Random Variables

– (kenny) TODO fix counters

Definition 1.7 (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 1.8 (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 1.9. *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 1.10. *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 $y_1, y_2, \ldots$ be a sequence where $y_1 < y_2 \ldots$ 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 1.11. 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 1.12. 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 1.13. 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 1.14 (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 1.15 (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 1.16 (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 1.17 (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 1.18 (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 1.19 (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 1.20 (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 1.21 (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 1.22 (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, $\Gamma(3) = \int_0^{\infty} y^2e^{-y}dy$. Using integration by parts, this evaluates to $[-y^2e^{-y} - 2ye^{-y} - 2e^{-y}]_0^{\infty}$. 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 1.23 (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 1.24 (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 1.25 (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 1.26. *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:

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

Definition 1.27 (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. •

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

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

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 $N \sim \text{Poisson}(\lambda)$.
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, $f_Y(y) = \frac{1}{\sqrt{2\pi}y} e^{-\frac{(\ln y)^2}{2}}$ □

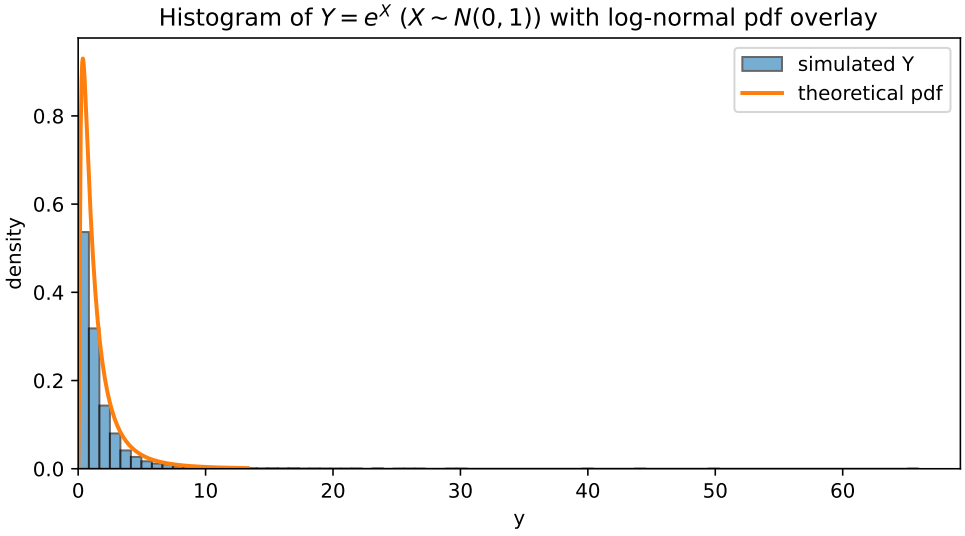

Figure 1: Histogram of $Y = e^X$ overlaid with its log-normal density.

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.

□

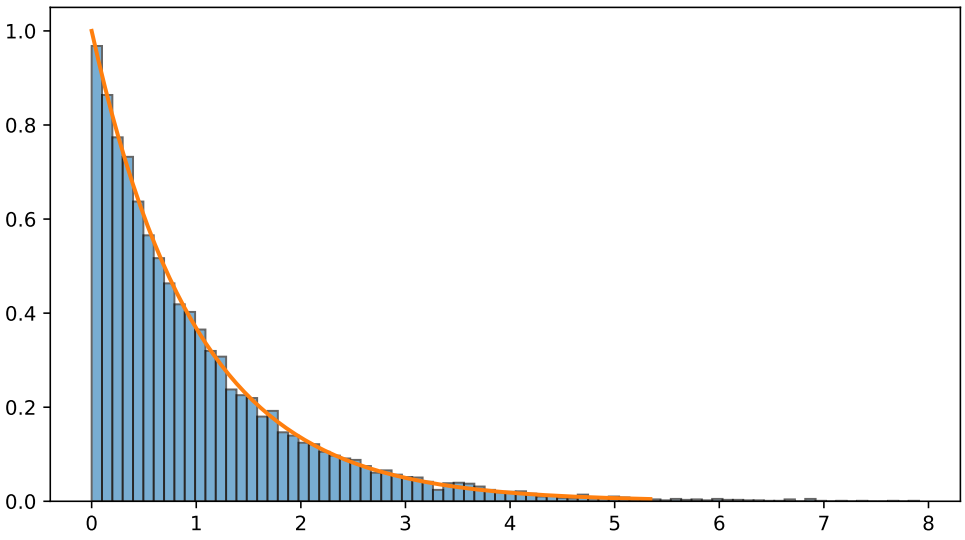

Figure 2: Histogram of generated exponentials overlayed against theoretical PDF

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.

1.3 Expectation

Definition 1.28. 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: $\mathbf{E}[X] = \mu = \mu_X$

Theorem 1.29. $\mathbf{E}\left[\sum_i X_i\right] = \sum_i \mathbf{E}\left[X\right]_i$

Theorem 1.30. *If $X_1 \dots X_i$ are independent,* $\mathbf{E}\left[\prod X_i\right] = \prod_i \mathbf{E}\left[X\right]_i$

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.

Definition 1.31. The variance of X is defined as

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

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

Theorem 1.32. *Assuming the variance of X is well-defined, it has the following properties:*

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

Proof.

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

□

Definition 1.33. 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

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

Theorem 1.34. *If $X_1 \dots X_n$ are i.i.d. and* $\mathbf{E}\left[X_i\right] = \mu$ *and* $\mathbf{Var}\left[X_i\right] = \sigma^2$*, 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.

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

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

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:

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

□

Definition 1.35. 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:

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

The correlation is then:

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

Theorem 1.36. *$Cov(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$ and $\rho_{X,Y}$ satisfies $-1 \leq \rho_{X,Y} \leq 1$. 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.*

Definition 1.37 (Conditional Expectation)**.**

Definition 1.38 (The Law of Iterated Expectation)**.** $\mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[X]$

Definition 1.39.

Example 1.40. 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:

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

Definition 1.41 (The Moment Generating Function)**.** $\psi_X t = \mathbf{E}\left[e^{tX}\right] = \int e^{tX} dF_X dx$ is the MGF or Laplace Transformation of X. 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 $t = 0$, $\psi'_X(0) = \frac{d}{dt}\,\mathbf{E}\left[e^{tX}\right]|_{t=0} = \mathbf{E}\left[\frac{d}{dt}e^{tX}\right]|_{t=0} = \mathbf{E}\left[Xe^{tX}\right]|_{t=0} = \mathbf{E}[X]$.