Last modified December 3, 2019

STA257: Probability and Statistics I

University of Toronto — Fall 2019

Jeff Shen

Contents

1 Probability and Counting **2**
1.1 Introduction . . . 2
1.2 Set Theory . . . 2
1.3 Probability Measures . . . 2
1.4 Counting . . . 3
1.5 Conditional Probability and Independence . . . 4
1.6 Law of Total Probability . . . 5
1.7 Equations . . . 6

2 Random Variables **7**
2.1 Discrete Random Variables . . . 7
2.2 Continuous Random Variables . . . 8
2.3 Transformations of Random Variables . . . 10
2.4 Distributions . . . 10

3 Expected Values **12**
3.1 Mean and Variance . . . 12
3.2 Covariance and Correlation . . . 12
3.3 Conditional Expectation . . . 13
3.4 Moment Generating Functions . . . 15
3.5 Equations . . . 16

4 Joint Distributions **17**
4.1 Joint and Marginal Distributions . . . 17
4.2 Independence in Joint Distributions . . . 18
4.3 Conditional Distributions . . . 18
4.4 Functions of Joint Distributions . . . 18
4.5 Order Statistics . . . 18
4.6 Equations . . . 18

5 Limit Theorems **18**
5.1 Law of Large Numbers . . . 18
5.2 Central Limit Theorem . . . 19

1 Probability and Counting

1.1 Introduction

Experiments: situations where outcome is random. e.g. flipping a coin is an experiment.
Sample space: set of all possible outcomes, denoted S or Ω. The number of elements in the sample space (cardinality) is denoted $|\Omega|$.
Event: a subset of a sample space.
Outcome: a particular element of a sample space. e.g. $s_1 \in \Omega$.

1.2 Set Theory

1.2.1 Definitions

- **union**: $A \cup B$. Elements in either A or B.
- **intersection**: $A \cap B$. Elements in both A and B.
- **complement** of A: A^c. Elements not in A.
- **empty set**: $\varnothing$. Set with no elements in it.
- A and B are **disjoint**: $A \cap B = \varnothing$. There are no elements in the intersection of A and B.

1.2.2 Laws of Set Theory

1. **commutativity**: $A \cup B = B \cup A$, $A \cap B = B \cap A$
2. **associativity**: $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap B)$
3. **distributivity**: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

1.3 Probability Measures

Probability measure: a function which maps subsets of Ω, which can be defined on any space, to real numbers $\mathbb{R}$.

1.3.1 Axioms of Probability Measures

- $P(\Omega) = 1$
- $\forall A \in \Omega, P(A) \geq 0$
- if $A_1, A_2, \ldots A_n, \ldots$ are mutually disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

1.3.2 Properties of Probability Measures

- $\forall A \in \Omega, P(A^c) = 1 - P(A)$

  Proof:

$$\begin{aligned} 1 &= P(\Omega) && \text{by Axiom 1} \\ &= P(A \cup A^c) && \text{by definition of complement} \\ &= P(A) + P(A^c) && \text{by Axiom 3 (since } A, A^c \text{ are disjoint)} \end{aligned}$$

  Rearrange this to see that $P(A^c) = 1 - P(A)$.

- $P(\varnothing) = 0$

  Proof:

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \varnothing) && \text{since } \Omega \cup \varnothing = \Omega \\ &= P(\Omega) + P(\varnothing) && \text{by Axiom 3 (since } \Omega, \varnothing \text{ are disjoint)} \end{aligned}$$

  So $P(\varnothing) = 0$.

- For $A, B \subseteq \Omega, A \subseteq B \implies P(A) \leq P(B)$

  Proof:

$$\begin{aligned} P(B) &= P(A \cup (B \cap A^c)) \\ &= P(A) + P(B \cap A^c) && \text{by Axiom 3 (since } A, A^c \text{ are disjoint)} \end{aligned}$$

  But note that $P(B \cap A^c) \geq 0$ by Axiom 2.

  Then $P(B) = P(A) + P(B \cap A^c) \geq P(A)$.

- For $A, B \subseteq \Omega, P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  Proof:

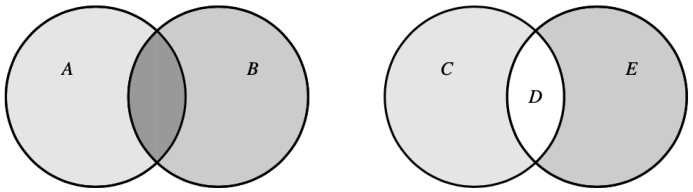
  Case 1: A, B are disjoint. Then $A \cap B = \varnothing \implies P(A \cap B) = 0$.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) && \text{by Axiom 3 (since } A, B \text{ are disjoint)} \\ &= P(A) + P(B) + P(A \cap B) && \text{since we can add 0 wherever we want} \end{aligned}$$

  Case 2: A, B not disjoint. Then $A \cap B \neq \varnothing$.

  Let $C = A \cap B^c$, $D = A \cap B$, $E = A^c \cap B$.

  Then C, D, E are disjoint, and $A = C \cup D$, $B = D \cup E$, and $A \cup B = C \cup D \cup E$.



$$\begin{aligned} P(A) + P(B) - P(A \cap B) &= P(C \cup D) + P(D \cup E) - P(D) && \text{by how we defined } C, D, E \\ &= P(C) + P(D) + P(D) + P(E) - P(D) && \text{by Axiom 3 and disjointness of } C, D, E \\ &= P(C) + P(D) + P(E) \\ &= P(C \cup D \cup E) && \text{by Axiom 3 and disjointness of } C, D, E \\ &= P(A \cup B) \end{aligned}$$

1.4 Counting

Multiplication principle: if there are m ways to do one thing, and n ways to do another thing, then there are mn ways to do both things.
Permutation: ordered arrangement of objects.

- Sampling **with replacement** means that duplicate item selection is allowed. (can pick the same object twice). For a set of size n and a **sample size** (number of items selected) r, there are n^r possible selections.
- Sampling **without replacement** means that each item is selected once at most. For a set of size n and a sample size r, there are $n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}$ possible selections. In particular, there are $n(n-1)\dots(1) = n!$ ways to order n elements.

Combination: arrangement of objects *without regard to order*. Think about this as the ways to select objects without replacement, divided by the ways that those objects can be ordered. For a set of size n and a sample size r, we express the combination as follows:

$$\binom{n}{r} = \frac{n(n-1)\dots(n-r+1)}{r!} = \frac{n!}{(n-r)!\,r!}$$

Binomial expansion:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}$$

In particular, for $a = b = 1$,

$$(1+1)^n = 2^n = \sum_{k=0}^{n} \binom{n}{k} (1)^k (1)^{n-k} = \sum_{k=0}^{n} \binom{n}{k}$$

The number of ways to group n objects into r classes, with n_i objects in the ith classes, where $1 < i < r$, is given by the following formula:

$$\binom{n}{n_1 n_2 \dots n_r} = \frac{n!}{n_1!\, n_2! \dots n_r!}$$

Proof: There are $\binom{n}{n_1}$ ways to select the first class, $\binom{n-n_1}{n_2}$ ways to select the second class, and so on. Repeat this for all r classes, and then apply the multiplication rule. Then we have that

$$\begin{aligned}
\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r} &= \frac{n!}{n_1!\,(n-n_1)!}\,\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}\cdots\frac{(n-n_1-\cdots-n_{r-1})!}{n_r!\,0!} \\
&= \frac{n!}{n_1!\,\cancel{(n-n_1)!}}\,\frac{\cancel{(n-n_1)!}}{n_2!\,\cancel{(n-n_1-n_2)!}}\cdots\frac{\cancel{(n-n_1-\cdots-n_{r-1})!}}{n_r!\,0!} \\
&= \frac{n!}{n_1!\, n_2! \dots n_r!}
\end{aligned}$$

1.5 Conditional Probability and Independence

Conditional probability: probability that some event will occur given that another event has occured:

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

Two events are **independent** if knowing that one has occured gives no information about the likelihood of the other occuring. Events A, B are independent if and only if any of the following hold:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

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

From the definition of conditional probability, we can derive **Bayes' Rule**, which describes the probability of an event given prior knowledge of conditions related to the event:

$$\begin{aligned} & P(A|B)\,P(B) = P(A \cap B) = P(B \cap A) = P(B|A)\,P(A) \\ \implies & P(A|B)\,P(B) = P(B|A)\,P(A) \\ \implies & P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \end{aligned}$$

1.6 Law of Total Probability

Let $B_1, B_2, \ldots, B_n$ be sets satisfying the following conditions:

1. $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$ (the B_i's are a **partition** of Ω)
2. $\forall i, j \in [1, n] \subseteq \mathbb{N}, i \neq j \implies B_i \cap B_j = \varnothing$ (pairwise disjoint)
3. $\forall i \in [1, n] \subseteq \mathbb{N}, P(B_i) > 0$ (strictly positive probability),

we can conclude that for any event A, we have that

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

Proof: Since all B_i's are pairwise disjoint, all $A \cap B_i$'s must also be pairwise disjoint. Then

$$\begin{aligned} P(A) &= P(A \cap \Omega) \\ &= P\left(A \cap \left(\bigcup_{i=1}^{n} B_i\right)\right) && \text{by condition 1} \\ &= P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) \\ &= \sum_{i=1}^{n} P(A \cap B_i) && \text{by Axiom 3 of probability measures} \\ &= \sum_{i=1}^{n} P(A|B_i)\,P(B_i) && \text{by definition of conditional probability} \end{aligned}$$

Under the same conditions, we can also conclude an alternate form of Bayes' Rule:

$$P(B_j|A) = \frac{P(A|B_j)\,P(B_j)}{\sum_{i=1}^{n} P(A|B_i)\,P(B_i)}$$

1.7 Equations

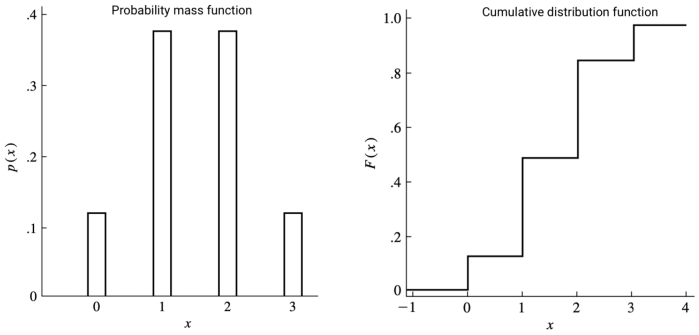
Axioms	$P(\Omega) = 1$
	$P(A) \geq 0$
	for mutually disjoint sets, $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
Properties	$P(A^c) = 1 - P(A)$
	$P(\varnothing) = 0, P(\Omega) = 1$
	$A \subseteq B \implies P(A) \leq P(B)$
	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Permutations	with replacement: n^r
	without replacement: $\frac{n!}{(n-r)!}$
Combinations	$\binom{n}{r} = \frac{n!}{(n-r)!\,r!}$
Binomial Expansion	$(a+b)^n = \sum_{i=1}^{n} \binom{n}{k} a^k\, b^{n-k}$
Grouping n objects into r classes	$\binom{n}{n_1 n_2 \ldots n_r} = \frac{n!}{n_1!\, n_2! \ldots n_r!}$
Conditional probability	$P(A\|B) = \frac{P(A \cap B)}{P(B)}$
Bayes' Rule	$P(A\|B) = \frac{P(B\|A)\, P(A)}{P(B)}$
	$P(B_j\|A) = \frac{P(A\|B_j)\, P(B_j)}{\sum_{i=1}^{n} P(A\|B_i)\, P(B_i)}$
Law of Total Probability	$P(A) = \sum_{i=1}^{n} P(A\|B_i)\, P(B_i)$

2 Random Variables

2.1 Discrete Random Variables

Random variable: function from Ω to $\mathbb{R}$.
Discrete random variable: random variable which can only take a finite number of values, or a countably infinite number of values.
Probability mass function (PMF): describes probability properties of a random variable. For some discrete random variable X taking on values $x_1, x_2, \ldots$, the probability mass function is some function p such that $p(x_i) = P(X = x_i)$, and satisfies that $\sum_{i=1}^{\infty} p(x_i) = 1$.
Cumulative distribution function (CDF): defined as some function F such that $F(x) = P(X \leq x)$, where $-\infty < x < \infty$. It satisfies the following conditions:

- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- is right-continuous (can have jumps, but must be continuous when approaching values from the right side). Continuity implies right-continuity.
- is non-decreasing.



For random variables X, Y taking on values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ respectively, X and Y are **independent** if $\forall i, j, P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$.

2.1.1 Bernoulli

$X \sim \text{Bern}(p)$. Was there a success?
X takes on the values 1 and 0 with probabilities p and $q = 1 - p$ respectively. For some event A, Bernoulli random variables are often used as an **indicator random variable**. That is, $I_A = 1$ if A occurs, and $I_A = 0$ otherwise.

2.1.2 Binomial

$X \sim \text{Bin}(n, p)$. Something was done n times. How many successes were there?
Binomial random variables can be thought of as the sum of n **i.i.d (independent and identically distributed)** Bernoulli random variables. That is, $X = X_1 + X_2 + \cdots + X_n$, where $X_i \sim \text{Bern}(p)$.

2.1.3 Geometric

$X \sim \text{Geom}(p)$. Keep going until there is a success.
Constructed from a (countably) infinite sequence of Bernoulli trials. Count the number of trials it took until finally getting a success. Some people don't include the kth trial (the last trial, which is a success), but we will. So k is the trial on which we have a success.

2.1.4 Negative Binomial

$X \sim \text{NB}(n, p)$. Keep going until there are n successes.
A natural extension to the geometric distribution—can be thought of as the sum of n i.i.d geometric random variables. That is, $X = X_1 + X_2 + \cdots + X_n$, where $X_i \sim \text{Geom}(p)$.

2.1.5 Hypergeometric

$X \sim \text{HG}(n, r, m)$. Total population of n in which r are marked/tagged. We randomly pick m from the population without replacement. How many are tagged?

2.1.6 Poisson

$X \sim \text{Pois}(\lambda)$. Large number of trials, each will a small probability of success.
λ is called the **rate parameter**. Think of dividing some interval into a large number of subintervals, such that the probability of an event occuring in each subinterval is small, but there are a large number of subintervals.
Assume that what happens in one subinterval is independent of what happens in the other subintervals, that the probability of some event occuring is the same for each subinterval, and that events do not occur simultaneously. Then this can be derived by setting up a binomial distribution in such a way that $n \to \infty$, $p \to 0$, and $np = \lambda$.

DERIVE POISSON FROM BINOMIAL

2.2 Continuous Random Variables

Continuous random variables can take on a continuum of values rather than just a finite (or countably infinite) number.
Probability density function (pdf). Analogous to the pmf for discrete random variables. Satisfies the following conditions:

- $f(x) \geq 0$
- f is piecewise continuous
- $\int_{-\infty}^{\infty} f(x)dx = 1$

For some a, b such that $a < b$, the probability that X is in the interval (a, b) is given by the area under $f(x)$ from a to b: $P(a < X < b) = \int_a^b f(x)dx$. Note that the probability of taking on a particular value is 0: $P(X = c) = \int_c^c f(x)dx = 0$. This implies that $P(a \leq X \leq b) = P(a < X < b)$.
For an (infinitessimally) small value of dx, the probability that X is in the interval $(x, x + dx)$ is proportional to $f(x)$: $P(x \leq X \leq x + dx) = f(x)dx$.
The **cumulative distribution function (cdf)** can be expressed in terms of the pdf:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

If f is continuous at x, then by the Fundamental Theorem of Calculus,

$$f(x) = F'(x)$$
$$\implies P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a).$$

If F is strictly increasing on some interval I, with $F = 0$ to the left of I and $F = 1$ to the right of I, then the **inverse** function F^{-1} is well defined:

$$y = F(x)$$
$$\implies x = F^{-1}(y)$$

The p**th quantile** of F is the value x_p such that $x_p = F^{-1}(p)$. Note that the **percentile** is the same as the quantile, but expressed as a percentage from 0 to 100, whereas the quantile is expressed as a value between 0 and 1. Special quantiles are the **median** ($p = \frac{1}{2}$) and the **lower and upper quartiles** ($p = \frac{1}{4}, \frac{3}{4}$) respectively.

2.2.1 Uniform

$U \sim \text{Unif}(a, b)$. All intervals of the same length are equally probable. Probability is proportional to the length of the subinterval: longer subintervals are more probably.
Is a "universal" distribution because it can be used to generate all other distributions. For some strictly increasing and continuous cdf F,

$$\begin{aligned} U \sim \text{Unif}(0,1) &\implies X = F^{-1}(U) \sim F \\ X \sim F &\implies U = F(X) \sim \text{Unif}(0,1) \end{aligned}$$

That is, plugging any random variable into its own cdf returns a uniform random variable, and plugging a uniform random variable into the inverse of the cdf yields a random variable distributed according to that cdf.

2.2.2 Exponential

$X \sim \text{Exp}(\lambda)$. Characterized by (and is the only distribution which has) the **memoryless property**:

$$P(X \geq s + t \,|\, X \geq s) = P(X \geq t)$$

Proof: Let $X \sim \text{Expo}(\lambda)$. Consider the **survival function** $P(X \geq s)$.

$$P(X \geq s) = 1 - P(X \leq s) = 1 - (1 - e^{-\lambda s}) = e^{-\lambda s}$$

Then by definition of conditional probability,

$$P(X \geq s + t \,|\, X \geq s) = \frac{P(X \geq s + t,\, X \geq s)}{P(X \geq s)}$$

But clearly, the term in the denominator is,

$$P(X \geq s + t,\, X \geq s) = P(X \geq s + t) = e^{-\lambda(s+t)} = e^{-\lambda s} e^{-\lambda t}$$

Substituting this into the previous equation,

$$P(X \geq s + t \,|\, X \geq s) = \frac{\cancel{e^{-\lambda s}} e^{-\lambda t}}{\cancel{e^{-\lambda s}}} = e^{-\lambda t} = P(X \geq t)$$

From this, we conclude that the probability of surviving t more units of time is independent of how many units of time it has already survived.

2.2.3 Gamma

$X \sim \text{Gamma}(\alpha, \lambda)$. α is called the shape parameter (which varies the shape of the density), and λ is called the scale parameter. Can be thought of as the sum of α i.i.d $\text{Exp}(\lambda)$ variables. That is, $X = X_1 + X_2 + \cdots + X_\alpha$, where $X_i \sim \text{Exp}(\lambda)$. Thus, the case of $\alpha = 1$ coincides with the exponential density.
Based on the **gamma function**, which is an extension of the factorial function (to complex numbers). It is defined as follows:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx, \text{ where } z > 0.$$

The gamma function has the following identity:

$$\Gamma(x + 1) = x\,\Gamma(x).$$

This identity implies that for any positive integer n,

$$\Gamma(n) = (n-1)!.$$

watch the stat110 video for gamma https://www.youtube.com/w

2.2.4 Beta

watch the stat110 beta videos https://www.youtube.com/watch?

2.2.5 Standard Normal

2.2.6 General Normal

2.3 Transformations of Random Variables

2.4 Distributions

pmf/pdf, cdf, ex, var, range of support for k

Bernoulli	$\mathrm{Bern}(p)$	$p(k) = \begin{cases} q = 1 - p & k = 0 \\ p & k = 1 \end{cases}$
Binomial	$\mathrm{Bin}(n, p)$	$p(k) = \binom{n}{k} p^k \, q^{n-k}$, for $k = 1, 2, \ldots$
Geometric	$\mathrm{Geom}(p)$	$p(k) = q^{k-1} \, p$, for $k = 1, 2, \ldots$
Negative Binomial	$\mathrm{NB}(r, p)$	$p(k) = \binom{k-1}{r-1} p^r \, q^{k-r}$, for $k = 1, 2, \ldots$; $r \leq k$
Hypergeometric	$\mathrm{HG}(n, r, m)$	$p(k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$, for $k \leq m, r \leq n$
Poisson	$\mathrm{Pois}(\lambda)$	$p(k) = \frac{e^{-\lambda} \, \lambda^k}{k!}$, for $k = 0, 1, \ldots$; $\lambda > 0$

Uniform	$\mathrm{U}(a, b)$	$f(x) = \frac{1}{b-a}$, for $a \leq x \leq b$
Exponential	$\mathrm{Exp}(\lambda)$	$f(x) = \lambda \, e^{-\lambda \, x}$, for $x \geq 0$
Gamma	$\mathrm{Gamma}(\alpha, \lambda)$	$g(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \, t^{\alpha-1} \, e^{-\lambda \, t}$, for $t \geq 0$; $x > 0$a
Beta	$\mathrm{Beta}(a, b)$	$f(u) = \frac{\Gamma(a+b)}{\Gamma(a) \, \Gamma(b)} \, u^{a-1} \, (1-u)^{b-1}$, for $0 \leq u \leq 1$
Standard Normal	$\mathrm{N}(0, 1)$	$\varphi(z) = \frac{1}{\sqrt{2\pi}} \, e^{-z^2/2}$
Normal	$\mathrm{N}(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \, e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2} = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$

3 Expected Values

3.1 Mean and Variance

For two independent random variables X, Y, the expectation of the product is the product of the expectation:

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$$

3.2 Covariance and Correlation

3.2.1 Covariance

Covariance allows us to talk about the variance for sums of random variables. For two jointly distributed random variables X, Y, the **covariance of X and Y** is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}((X - \mathrm{E}X)(Y - \mathrm{E}Y)).$$

An alternate way to write the covariance is given by

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}((X - \mathrm{E}X)(Y - \mathrm{E}Y)) \\
&= \mathrm{E}(XY - Y\mathrm{E}(X) - X\mathrm{E}(Y) + \mathrm{E}(X)\mathrm{E}(Y)) \\
&= \mathrm{E}(XY) - \mathrm{E}(Y\mathrm{E}(X)) - \mathrm{E}(X\mathrm{E}(Y)) + \mathrm{E}(\mathrm{E}(X)\mathrm{E}(Y)) && \text{by linearity of expectation} \\
&= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) - \mathrm{E}(Y)\mathrm{E}(X) + \mathrm{E}(X)\mathrm{E}(Y) && \text{by taking out constants} \\
&= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).
\end{aligned}$$

If X and Y are independent, then the covariance is zero. Note that the converse is not true.
Consider the following cases:

- Note that the covariance of some random variable with itself reduces down to its variance:

$$\begin{aligned}
\mathrm{Cov}(X, X) &= \mathrm{E}((X - \mathrm{E}(X))(X - \mathrm{E}(X))) \\
&= \mathrm{E}((X - \mathrm{E}(X))^2) \\
&= \mathrm{Var}(X).
\end{aligned}$$

- For some constant c, the covariance of X and c is zero:

$$\begin{aligned}
\mathrm{Cov}(X, c) &= \mathrm{E}((X - \mathrm{E}(X))(c - \mathrm{E}(c))) \\
&= \mathrm{E}((X - \mathrm{E}(X))(c - c)) \\
&= \mathrm{E}(0) = 0.
\end{aligned}$$

- The covariance of a scaled random variable is a scaled covariance:

$$\begin{aligned}
\mathrm{Cov}(cX, Y) &= \mathrm{E}((cX - \mathrm{E}(cX))(Y - \mathrm{E}(Y))) \\
&= \mathrm{E}((cX - c\mathrm{E}(X))(Y - \mathrm{E}(Y))) \\
&= c\,\mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \\
&= c\,\mathrm{Cov}(X, y).
\end{aligned}$$

- For random variables X, Y, and Z, the covariance is "distributive":

$$\begin{aligned}
\mathrm{Cov}(X, Y + Z) &= \mathrm{E}((X - \mathrm{E}(X))((Y - \mathrm{E}(Y)) + (Z - \mathrm{E}(Z)))) \\
&= \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y)) + (X - \mathrm{E}(X))(Z - \mathrm{E}(Z))) \\
&= \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) + \mathrm{E}((X - \mathrm{E}(X))(Z - \mathrm{E}(Z))) \\
&= \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z).
\end{aligned}$$

More generally, the covariance of sums is

$$\text{Cov}\left(\sum_i a_i X_i, \sum_j b_j Y_j\right) = \sum_i \sum_j a_i b_j \text{Cov}(X_i, Y_j)$$

The covariance can be used to deal with the variance of sums:

$$\begin{aligned}\text{Var}(X+Y) &= \text{Cov}(X+Y, X+Y) \\ &= \text{Cov}(X,X) + \text{Cov}(X,Y) + \text{Cov}(Y,X) + \text{Cov}(Y,Y) \\ &= \text{Var}(X) + 2\text{Cov}(X,Y) + \text{Var}(Y).\end{aligned}$$

Note that if X and Y are independent, then the covariance term becomes zero, and we conclude that the variance of the sum is the sum of the variance. More generally,

$$\begin{aligned}\text{Var}(X_1 + \ldots + X_n) &= \sum_i \text{Var}(X_i) + 2\sum_i \sum_{j \le i} \text{Cov}(X_i, X_j) \\ &= \sum_i \text{Var}(X_i) + \sum_{i \ne j} \text{Cov}(X_i, X_j)\end{aligned}$$

3.2.2 Correlation

Given jointly distributed X and Y for which the variances and covariances exist, and the variances are nonzero, the **correlation of X and Y** is defined as

$$\begin{aligned}\rho &= \frac{\text{Cov}(X,Y)}{\text{SD}(X)\text{SD}(Y)} \\ &= \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \text{Cov}\left(\frac{X - \text{E}(X)}{\text{SD}(X)}, \frac{Y - \text{E}(Y)}{\text{SD}(Y)}\right) \qquad \text{(standardize, then take the covariance)}\end{aligned}$$

The correlation is always between -1 and 1.
Proof: Assume that X and Y have been standardized. Then $\text{Var}(X) = \text{Var}(Y) = 1$.

$$\begin{aligned}0 \le Var(X+Y) &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y) = 1 + 1 + 2\rho = 2(1+\rho) \\ &\implies \rho \ge -1 \\ 0 \le Var(X-Y) &= \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y) = 1 + 1 - 2\rho = 2(1-\rho) \\ &\implies \rho \le 1\end{aligned}$$

3.2.3 LOTUS

3.2.4 Inequalities

3.3 Conditional Expectation

This is the same as expectation, but conditioning on some additional information that we are given. For random variables X, Y, the **conditional expectation** of Y given X is, in the discrete case,

$$\text{E}(Y|X=x) = \sum_y y\, p_{Y|X}(y|x)$$

and in the continuous case,

$$\text{E}(Y|X=x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y\, \frac{f_{X,Y}(x,y)}{f_x(x)}dy$$

If the conditional expectation of Y given $X = x$ is defined for all x in the domain of X, then the expectation is a well-defined function of X. Then we write $(\mathrm{E}(Y|X))$, which is a random variable that is a function of X. This computes, for some fixed X, the expectation of Y.
Law of Total Expectation, also called **Law of Iterated Expectation** or **Adam's Law**:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y|X))$$

Proof: Let $g(X) = \mathrm{E}(Y|X)$. Want to show that $\mathrm{E}(Y) = \mathrm{E}(g(X))$.

$$\begin{aligned}
\mathrm{E}(g(X)) &= \sum_x g(X)\, P(X = x) && \text{by LOTUS} \\
&= \sum_x E(Y|X = x)\, P(X = x) && \text{by definition of } g \\
&= \sum_x \left(\sum_y P(Y = y|X = x) \right) P(X = x) && \text{by definition of expectation} \\
&= \sum_y \sum_x y\, P(Y = y, X = x) && \text{by definition of the joint PMF and swapping order of summation} \\
&= \sum_y y \sum_x P(Y = y, X = x) && \text{since } y \text{ does not depend on } x \\
&= \sum_y y P(Y = y) && \text{by definition of the marginal PMF} \\
&= \mathrm{E}(Y) && \text{by definition of expectation}
\end{aligned}$$

Conditional expectations satisfy the following properties:

- Taking out what is known: for any function h of X, $\mathrm{E}(h(x)\, Y|X) = h(x)\, E(Y|X)$
- If X, Y are independent, then $\mathrm{E}(Y|X) = \mathrm{E}(Y)$
- The **residual,** $Y - \mathrm{E}(Y|X)$, is uncorrelated with any function of X: $\mathrm{E}((Y - \mathrm{E}(Y|X))h(x)) = 0$
  Proof:

$$\begin{aligned}
\mathrm{E}((Y - \mathrm{E}(Y|X))h(x)) &= \mathrm{E}(Yh(x) - \mathrm{E}(Y|X)h(x)) \\
&= \mathrm{E}(Yh(x)) - \mathrm{E}(\mathrm{E}(Y|X)h(x)) && \text{by linearity} \\
&= \mathrm{E}(Yh(x)) - \mathrm{E}(\mathrm{E}(Yh(x)|X)) && \text{by the reverse of the first property} \\
&= \mathrm{E}(Yh(x)) - \mathrm{E}(Yh(x)) && \text{by Iterated Expectation} \\
&= 0
\end{aligned}$$

The **conditional variance** of Y given X is the same as the normal variance but conditioned on X wherever appropriate:

$$\mathrm{Var}(Y|X) = \mathrm{Var}(Y^2|X) - (\mathrm{Var}(Y|X))^2 = \mathrm{E}((Y - \mathrm{E}(Y|X))^2|X)$$

Law of Total Variance, also called **Law of Iterated Variance** or **Eve's Law**:

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

Proof: Just as we defined $\mathrm{E}(Y|X)$ before to be a random variable as a function of X, we can do the same with $\mathrm{Var}(Y|X)$ if $\mathrm{Var}(Y|X = x) = E(Y^2|X = x) - (\mathrm{E}(Y)|X = x)^2$ is defined for all x in the domain of X. Then

$$\begin{aligned}
&\mathrm{Var}(Y|X) = \mathrm{E}(Y^2|X) - (\mathrm{E}(Y|X))^2 && \text{by definition of variance} \\
\implies &\mathrm{E}(\mathrm{Var}(Y|X)) = \mathrm{E}(\mathrm{E}(Y^2|X)) - \mathrm{E}(\mathrm{E}(Y|X))^2) && \text{by linearity of expectation}
\end{aligned}$$

and

$$\text{Var}(\text{E}(Y|X)) = \text{E}((\text{E}(Y|X))^2) - (\text{E}(\text{E}(Y|X)))^2 \qquad \text{by definition of variance}$$

and

$$\begin{aligned}\text{Var}(Y) &= \text{E}(Y^2) - (\text{E}(Y))^2 && \text{by definition of variance}\\ &= \text{E}(\text{E}(Y^2|X)) - (\text{E}(\text{E}(Y|X)))^2 && \text{by iterated expectation}\end{aligned}$$

Then

$$\text{Var}(Y) = \text{E}(\text{E}(Y^2|X)) - (\text{E}(\text{E}(Y|X)))^2$$

If we add and subtract $\text{E}((\text{E}(Y|X))^2)$, this expression remains the same:

$$\text{Var}(Y) = \text{E}(\text{E}(Y^2|X)) - \text{E}((\text{E}(Y|X))^2) + \text{E}((\text{E}(Y|X))^2) - (\text{E}(\text{E}(Y|X)))^2$$

Then by grouping the first two and the last two terms together:

$$\text{Var}(Y) = \text{E}(\text{Var}(Y|X)) + \text{Var}(\text{E}(Y|X))$$

3.4 Moment Generating Functions

A **moment generating function (MGF)** is an alternate way of specifying the distribution of a function (apart from the pdf/cdf). For some random variable X, the MGF is defined as

$$M_X(t) := E(e^{tX}).$$

wherever the expectation exists. The MFG *uniquely determines* the probability distribution if it exists in some open interval around zero.
The **n-th moment** is defined as $E(X^n)$, and the **n-th central moment** is defined as $E\left((X - E(X))^n\right)$. By these definitions, it is clear where the "generating" part of the name comes from; the Maclaurin expansion of e^{tX} is

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \dots,$$

and thus, by linearity of expectation,

$$\begin{aligned}E(e^{tX}) &= 1 + tE(X) + \frac{t^2E(X^2)}{2!} + \frac{t^2E(X^3)}{3!} + \dots\\ &= 1 + tm_1 + \frac{t^2}{2!}m_2 + \frac{t^3}{3!}m_3 + \dots\end{aligned}$$

where m_i is the ith moment.
Thus, the ith moment is the coefficient of $\frac{t^i}{i!}$, and it can be found by taking the ith derivative of the series expansion (lower powered terms will disappear, and higher powered terms will equal zero when evaluated at zero).
For random variables X with MGF M_X and Y with MGF M_Y, if X and Y are independent, then the MGF of $X + Y$ is

$$\begin{aligned}M_{X+Y}(t) &= E(E^{t(X+Y)})\\ &= E(e^{tX+tY})\\ &= E(e^{tX}e^{tY})\\ &= E(e^{tX})E(e^{tY}) && \text{by independence}\\ &= M_X(t)M_Y(t).\end{aligned}$$

For random variable X with MGF M_X and random variable $Y = a + bX$, Y has MGF

$$\begin{aligned} M_Y(t) &= E(e^{tY}) \\ &= E(e^{t(a+bX)}) \\ &= E(e^{at+btX}) \\ &= E(e^{at}e^{btX}) \\ &= e^{at}E(btX) \\ &= e^{at}M_X(bt). \end{aligned}$$

3.5 Equations

means, variances, and mgfs for each distribution.

4 Joint Distributions

4.1 Joint and Marginal Distributions

Joint distributions deal with probabilities of multiple random variables defined on the same sample space. For random variables X, Y, the **joint CDF** of X and Y is

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

The probability that (X, Y) is in some rectangle $[x_1, x_2] \times [y_1, y_2]$ is given by

$$P(x_1 \leq X \leq x_2,\, y_1 \leq Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1).$$

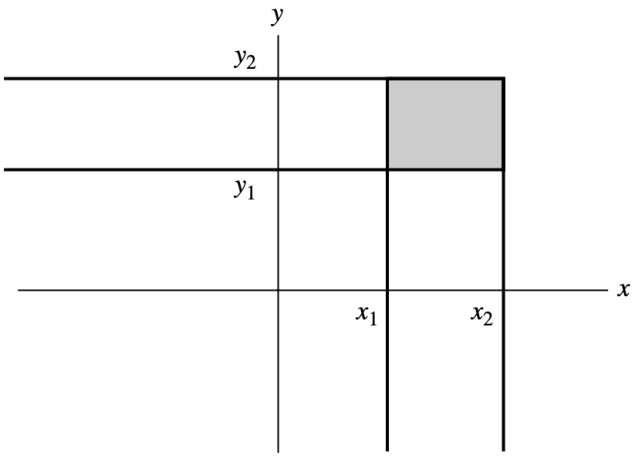


Figure 1: The probability of the shaded rectangle can be found by taking the probability of the (x_2, y_2) rectangle, subtracting the probabilities of the (x_1, y_2) and (x_2, y_1) rectangles, and then adding back in the probability of the (x_1, y_1) rectangle to account for the double subtraction.

4.1.1 Discrete

Suppose X and Y are discrete random variables defined on the same sample space, with X taking on values $x_1, x_2, \ldots$, and Y taking on values $y_1, y_2, \ldots$. Then the **joint pmf** $p(x, y)$ is

$$p(x_i, y_i) = P(X = x_i,\, Y = y_i).$$

The pmf of one random variable is called its **marginal probability function**. For X and Y jointly distributed, the marginal pmf of X is given by

$$p_X(x) = \sum_i p(x, y_i).$$

In general, to get a marginal pmf, fix a value for the random variable whose marginal pmf we want, and then sum the joint pmf over all possible values for all other (jointly distributed) random variables. That is, for $X_1, \ldots, X_n$ jointly distributed,

$$p_{X_i}(x_i) = \sum_{j : x_j \neq x_i} p(x_1, \ldots, x_n).$$

For example, the two-dimensional marginal pmf for X_1 and X_2 would be given by

$$p_{X_1, X_2}(x_1, x_2) = \sum_{x_3 \ldots x_n} p(x_1, \ldots, x_n).$$

4.1.2 Continuous

Suppose X and Y are continuous random variables. For a set A defined as $A = (X, Y) \mid X \leq x,\, Y \leq y$, the **joint cdf** can be expressed as

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, dv\, du.$$

Then by the Fundamental Theorem of Calculus, the **joint density** is given by

$$f(x, y) = \partial_x \partial_y F(x, y)$$

wherever the derivative is well-defined. The **joint density** $f(x, y)$ satisfies the following properties:

- $f(x, y) \geq 0$ (non-negative)
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$

For some two-dimensional set A,

$$P((X, Y) \in A) = \iint_{A} f(x, y)\, dx\, dy.$$

The **marginal cdf** of X, denoted F_X, is

$$F_X(x) = P(X \leq x) = \lim_{y \to \infty} F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, y)\, dy\, du,$$

and the **marginal density** of X, denoted f_x, is

$$f_X(x) = F_X'(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

4.2 Independence in Joint Distributions

4.3 Conditional Distributions

4.3.1 Discrete

4.3.2 Continuous

4.4 Functions of Joint Distributions

4.5 Order Statistics

4.6 Equations

5 Limit Theorems

5.1 Law of Large Numbers

Given i.i.d. random variables $X_1, X_2, \ldots$, each with mean μ and variance σ^2, let the sample mean be defined as $\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then for $\epsilon > 0$, the **weak law of large numbers** says that

$$P(|\overline{X_n} - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

In other words, as the sample size increases, it is increasingly likely that the sample mean is close to the true mean.

Proof: Note that

$$\begin{aligned}
\text{Var}(\overline{X_n}) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) && \text{by definition of } \overline{X_n} \\
&= \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n} X_i) && \text{by pulling out the } \frac{1}{n} \\
&= \frac{1}{n^2} n\sigma^2 && \text{by independence} = \frac{\sigma^2}{n}
\end{aligned}$$

Then

$$\begin{aligned}
P(|\overline{X_n} - \mu| > \epsilon) &\leq \frac{\text{Var}(\overline{X_n})}{\epsilon^2} && \text{by Chebyshev's inequality} \\
&= \frac{\sigma^2}{n\epsilon^2} \to 0 && \text{as } n \to \infty
\end{aligned}$$

5.2 Central Limit Theorem

Given i.i.d random variables $X_1, X_2, \ldots$, each with mean μ, variance σ^2, and MGF $M(t)$ defined on some interval around zero. Let $S_n = \sum_{i=1}^{n} X_i$. Then the **central limit theorem** says that

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), -\infty < x < \infty$$

In other words, for large n, the distribution of $\frac{S_n}{\sigma\sqrt{n}}$ approaches the standard normal distribution.
Proof:i Without loss of generality, assume that $\mu = 0$ and $\sigma^2 = 1$ (we can always normalize to achieve this). So, we want to show that the MGF of $\frac{S_n}{\sqrt{n}}$ approaches the MGF of the standard normal. By definition, the MGF is

$$\begin{aligned}
\text{E}\left(e^{t\frac{S_n}{\sqrt{n}}}\right) &= \text{E}\left(t\frac{X_1}{\sqrt{n}}\right)\ldots\text{E}\left(t\frac{X_n}{\sqrt{n}}\right) && \text{by independence of the } X_i\text{'s} \\
&= \left(\text{E}\left(t\frac{X_1}{\sqrt{n}}\right)\right)^n && \text{since all the MGFs are equivalent by i.i.d.} \\
&= \left(\text{M}\left(\frac{t}{\sqrt{n}}\right)\right)^n && \text{by definition of the MGF}
\end{aligned}$$

This is an indeterminate form. We can proceed by taking the logarithm:

$$\log\left(M_{\frac{S_n}{\sqrt{n}}}(t)\right) = \log\left(\left(\text{M}\left(\frac{t}{\sqrt{n}}\right)\right)^n\right) = n\log\left(\text{M}\left(\frac{t}{\sqrt{n}}\right)\right)$$

Now taking the limit, we get

$$\lim_{n\to\infty} n\log\left(\text{M}\left(\frac{t}{\sqrt{n}}\right)\right) = \lim_{n\to\infty} \frac{\log\left(\text{M}\left(\frac{t}{\sqrt{n}}\right)\right)}{1/n}$$

Remember that $n \in \mathbb{Z}$, and we cannot do calculus on $\mathbb{Z}$. We perform a change of variables, where we let $Y = \frac{1}{\sqrt{n}}$. Then $Y \in \mathbb{R}$. Rewriting the above limit in terms of Y, we get

$$\lim_{y\to 0} \frac{\log\left(M\left(yt\right)\right)}{y^2} = \lim_{y\to 0} \frac{M'\left(yt\right)t}{2yM\left(yt\right)} \qquad \text{by L'Hopital's Rule and the Chain Rule}$$

Note that

- $M(0) = 1$
- $M'(0) = \mu = 0$
- $M''(0) = \text{E}(X^2) = \text{Var}(X) + (\text{E}(X))^2 = \text{Var}(X) = 1$ since $\text{E}(X) = 0$

Then the limit becomes

$$\begin{aligned}\frac{t}{2}\lim_{y\to 0}\frac{M'(yt)}{y} &= \frac{t^2}{2}\lim_{y\to 0}\frac{M''(yt)}{1} && \text{by L'Hopital's Rule and the Chain Rule}\\ &= \frac{t^2}{2}\end{aligned}$$

Now taking the exponent to reverse our logarithm operation, we find that

$$M_{\frac{S_n}{\sqrt{n}}}(t) = e^{\frac{t^2}{2}}$$

which is just the MGF of the standard normal, as we wanted.