Mathematical Statistics I
Chapter 4: Expected Values

Jesse Wheeler

Contents

1 Discrete random variables

Introduction

- This material comes primarily from Rice (2007, Chapter 4).

- We will cover the ideas of expected value, variance, as well has higher-order moments.

- This includes topics such as conditional expectation, which is one of the fundamental ideas behind many branches of statistics and machine learning.

- For instance, most regression / prediction algorithms are built with the idea of minimizing some conditional expectation.

Expectation: Discrete random variables

Definition: Expectation of discrete random variables
Let X be a discrete random variable with pmf $p(x)$, which takes values in the space $\mathcal{X}$. The *expected value* of X is
$$E(X) = \sum_{x \in \mathcal{X}} x \, p(x),$$
provided that $\sum_{x \in \mathcal{X}} |x| \, p(x) < \infty$; otherwise, the expectation is not defined.

- This is not the most mathematically precise definition of expectation, but a more complete treatment of the topic is outside the scope of this course (See Resnick, 2019).

- The concept of the expected value parallels the notion of a *weighted average*.

- That is, we weight each possibility $x \in \mathcal{X}$ by their corresponding probability: $\sum_x x\,p(x)$.

- $E(X)$ is also referred to as the *mean* of X, and is typically denoted μ or μ_X.

- If the function p is thought of as a weight, then $E(X)$ is the center; that is, if we place the mass $p(x_i)$ at the points x_i, then the balancing point is $E(X)$.

- Like with the pmf and cdf, we often use subscripts to denote which probability law we are using for the expectation, it if is not clear: $E_X(X)$.

Roulette
A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet \$1 that an odd number comes up, you win or lose \$1 according to whether that event occurs. If X denotes your net gain, $X = 1$ with probability 18/38 and $X = -1$ with probability 20/28. The expected value of X is

$$E(X) = 1 \times \frac{18}{38} + (-1) \times \frac{20}{38} = -\frac{1}{19}.$$

- As you might imagine, the expected value coincides in the limit with the actual average loss per game, if you play many games (Chapter 5).

- Most casino games have a negative expected value by design; you may win some money, but if a large number of games are played, the house will come out on top.

Geometric Random Variable
Suppose that items are produced in a plant are independently defective with probability p. If items are inspected one by one until a defective item is found, then how many items must be inspected on average?
Let X denote the number of items inspected, up-to and including the first defective item. X is geometrically distributed, which as pmf

$$p(k) = P(X = k) = p\,(1 - p)^{k-1}.$$

Therefore

$$E(X) = \sum_{k=1}^{\infty} kp\,(1 - p)^{k-1}$$
$$= p\sum_{k=1}^{\infty} k\,(1 - p)^{k-1}.$$

To work out this summation, we will use a trick that is sometimes useful for infinite series. First, lets define $q = 1 - p$, and note that $0 < q < 1$. Then, the sum becomes

$$E(X) = p\sum_{k=1}^{\infty} k\,q^{k-1}.$$

You might notice that the summand is a power-rule derivative:

$$\frac{d}{dq}q^k = k\,q^{k-1}.$$

This fact is going to be useful, because the left-hand side of this derivative equation is a geometric sum, which we know how to calculate:

$$\sum_{k=1}^{\infty} q^k = \sum_{k=1}^{\infty} q \, q^{k-1} = q \sum_{j=0}^{\infty} q^j = \frac{q}{1-q}.$$

Thus, what we would like to do is write

$$\frac{d}{dq}\Big(\frac{q}{1-q}\Big) = \frac{d}{dq}\Big(\sum_{k=1}^{\infty} q^k\Big) \stackrel{?}{=} \sum_{k=1}^{\infty} \frac{d}{dq} q^k = \sum_{k=1}^{\infty} k q^{k-1}.$$

Now we can easily calculate the left-hand side to be $\frac{1}{(1-q)^2}$, and therefore we want to make the conclusion

$$\sum_{k=1}^{\infty} k \, q^{k-1} \stackrel{?}{=} \frac{d}{dq}\Big(\frac{q}{1-q}\Big) = \frac{1}{(1-q)^2}.$$

The question is: **Can we move the derivative inside of the infinite sum?** For this particular case, the answer is *yes*. For more details, see Slide 21 from Chapter 3. In this class, all of the sums (and integrals) we will consider will be "well-behaved" and will satisfy the necessary conditions.
With this sorted out, we can now use our trick to finish the calculation:

$$E(X) = p \sum_{k=1}^{\infty} k \, (q)^{k-1}$$
$$= p \frac{1}{(1-q)^2}$$
$$= \frac{p}{p^2} = \frac{1}{p}.$$

Poisson Distribution
The Poisson(λ) distribution has pmf $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$, for all $k \geq 0$. Thus, if $X \sim \text{Pois}(\lambda)$, then what is $E[X]$?

$$E[X] = \sum_{k=0}^{\infty} \frac{k \, \lambda^k}{k!} e^{-\lambda}$$
$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k \, \lambda^{k-1} \cdot \lambda}{k!}$$
$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$
$$= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$$
$$= \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

2   Continuous random variables

Expectation: Continuous random variables

Definition: Expectation of continuous random variables

Let X be a continuous random variable with pdf $f(x)$, which takes values in the space $\mathcal{X}$. The *expected value* of X is

$$E(X) = \int_{x \in \mathcal{X}} x f(x) \, dx.$$

provided that $\int_{x \in \mathcal{X}} |x| \, f(x) \, dx < \infty$, otherwise the expectation is undefined.

- As before, this is not the most mathematically precise definition of expectation, but a more complete treatment of the topic is outside the scope of this course (See Resnick, 2019).

- We can still think of $E(X)$ as the center of mass of the density.

Exponential(λ) expectation
Let X have an Exponential(λ) density, with $\lambda > 0$. Thus, the pdf of X is given by

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leq x < \infty$$

Find $E[X]$. *Solution.*

$$E[X] = \int_x x \, f_X(x) \, dx$$
$$= \int_0^\infty \frac{1}{\lambda} x e^{-x/\lambda} \, dx$$

- To solve this integral, we can use integration by parts:

$$\int u \, dv = uv - \int v \, du.$$

- We let $u = x$, and $dv = \frac{1}{\lambda} e^{-x/\lambda}$.

- Then, $du = dx$, $v = \int \frac{1}{\lambda} e^{-x/\lambda} \, dx = -e^{-x/\lambda}$.

- Plugging this in, we get:

$$E[X] = -x e^{-x/\lambda} \big|_0^\infty - \int_0^\infty -e^{-x/\lambda} \, dx$$
$$= \int_0^\infty -e^{-x/\lambda} \, dx$$
$$= \lambda e^{-x/\lambda} \big|_0^\infty = \lambda.$$

Gamma Density
If X follows a gamma density with parameters α and λ, then the pdf of X is

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

Find $E(X)$.

Solution: By definition, the expected value of X is

$$E(X) = \int_0^\infty (x) \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, dx.$$

Combining the factors of x in the integrand, we obtain

$$E(X) = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x}\, dx.$$

Now we will apply the "integration by density function" trick: we will re-write the integrand so that it corresponds to the density function of some random variable, and use the fact that the density function must integrate to one. Specifically, note that if we let $\alpha^* = \alpha + 1$, then $\alpha = \alpha^* - 1$, and we can express the integral as:

$$\begin{aligned}
E(X) &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x}\, dx \\
&= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha^*-1} e^{-\lambda x}\, dx \\
&= \Big(\frac{\lambda^\alpha}{\Gamma(\alpha)}\Big)\Big(\frac{\Gamma(\alpha^*)}{\lambda^{\alpha^*}}\Big) \int_0^\infty \frac{\lambda^{\alpha^*}}{\Gamma(\alpha^*)} x^{\alpha*-1} e^{-\lambda x}\, dx \\
&= \Big(\frac{\lambda^\alpha}{\Gamma(\alpha)}\Big)\Big(\frac{\Gamma(\alpha^*)}{\lambda^{\alpha^*}}\Big)
\end{aligned}$$

Where the last step is a result of the fact that the integrand (and support of the integral) matches the density of a $\mathrm{Gamma}(\alpha^*, \lambda)$ distribution. Now using the fact that $\alpha* = \alpha + 1$, and that $\Gamma(x+1) = x\Gamma(x)$, we obtain

$$\begin{aligned}
E(X) &= \frac{\lambda^\alpha\, \Gamma(\alpha+1)}{\Gamma(\alpha)\, \lambda^{\alpha+1}} \\
&= \frac{\alpha}{\lambda}
\end{aligned}$$

3   Expectation of functions of random variables

Functions of random variables

- We are often interested in functions of random variables: $Y = g(X)$.

- Ideas that we have already covered enable us to calculate $E(Y)$.

- For instance, you could use the change-of-variables theorem to get the density of Y, then use the definition to calculate $E[Y]$.

- Fortunately, we don't have to do this. We can instead calculate $E[Y]$ by integrating (or summing) with respect to X:
$$E[g(X)] = \int_{x \in \mathcal{X}} g(x) f(x)\, dx.$$

- We will justify this for the discrete case.

Theorem 4.1: Expectation of transformed random variables

Suppose that X is a random variable and that $Y = g(X)$ for some function g. Then,

- If X is discrete with pmf $p(x)$:
$$E(Y) = \sum_x g(x)\, p(x),$$

provided that $\sum_x |g(x)| p(x) < \infty$.

5

- If X is continuous with pdf $f(x)$:

$$E(Y) = \int_{-\infty}^{\infty} g(x)f(x)\,dx,$$

provided that $\int |g(x)|f(x)\,dx < \infty$.

Functions of random variables: proof

Proof: By definition of expectation,
$$E(Y) = \sum_i y_i p_Y(y_i).$$

Now let A_i denote the set of x's that are mapped to y_i by g. That is, A_i is the pre-image of y_i, meaning that $x \in A_i$ if $g(x) = y_i$. Then,
$$p_Y(y_i) = \sum_{x \in A_i} p(x),$$

and we can express the expectation as

$$\begin{aligned}
E(Y) &= \sum_i y_i p_Y(y_i) \\
&= \sum_i y_i \sum_{x \in A_i} p(x) \\
&= \sum_i \sum_{x \in A_i} y_i\, p(x) \\
&= \sum_i \sum_{x \in A_i} g(x)\, p(x) \\
&= \sum_x g(x)\, p(x)
\end{aligned}$$

Here, the second to last step is because for all $x \in A_i$, $g(x) = y_i$ by definition. The final step is a result of the fact that the A_i are disjoint, and every x belongs to some A_i, and thus the sum over i and $x \in A_i$ is the sum of all x.

- The proof for the continuous case is similar, but does require a measure-theoretic approach to integration.

- One important thing to note is that $g\big(E(X)\big)$ is not usually equal to $E\big(g(x)\big)$.

- For example, let Z be a standard normal. We know that $E[Z] = 0$, because it's symmetric. However, $P\big(|Z| > 0\big) = 1$, thus we can readily deduce that $E\big[|Z|\big] \geq 0 = \big|E[Z]\big|$.

- This idea can be extended to show that if for all non-negative random variables X that have finite expectation, if $g(x) \leq x$ for some function g, then $E[g(X)] \leq E[X]$.

Expected value of indicator functions

- Another important example of expectations is *indicator* random variables.

- For example, suppose that X is a random variable. Then $Y = 1[X \in A]$ for some $A \subset \mathcal{X}$ is a random variable.

Indicator Random Variable

Let X follow a standard normal distribution, and $A = [-1, 1]$. Then $Y = 1[X \in A]$ is defined as the random variables such that $Y(\omega) = 1$ if $X(\omega) \in A$, and $Y(\omega) = 0$ otherwise.

- Expectations of indicator variables are *probabilities.* Let $Y = 1[X \in A]$.

$$
\begin{aligned}
E(Y) &= E\big(1[X \in A]\big) \\
&= \int_{x \in \mathcal{X}} 1[X \in A]\, f(x)\, dx \\
&= \int_{x \in A} f(x)\, dx = P(X \in A).
\end{aligned}
$$

- This fact is useful for deriving some important inequalities.

- First, we will show that the expectations of interest actually exist.

- Let X be a continuous random variable with expectation $E(X)$. From our definition, this implies that $\int |x|\, f(x)\, dx < \infty$.

- Now suppose that for some random variable $Y = g(X)$ such that $|Y| \leq |X|$. Then we can deduce that $\int |y|\, f(x)\, dx < \infty$, and therefore $E[Y]$ exists.

- Now suppose that φ is a non-decreasing, non-negative function, and that for some $a \in \mathbb{R}$, $\varphi(a) > 0$. Then, for all $x \geq a$, $\varphi(x)/\varphi(a) \geq 1$.

- Define $Y = 1[X \geq a]$. Note that for all possible outcomes $\omega \in \Omega$,

$$
Y = 1[X \geq a] \leq \varphi(X)/\varphi(a)1[X \geq a] \leq \varphi(X)/\varphi(a).
$$

- Taking expectations of everything (which we argued preserves inequalities),

$$
E\big(1[X \geq a]\big) = P(X \geq a) \leq \frac{E\big[\varphi(X)\big]}{\varphi(a)} = E\left[\varphi(X)/\varphi(a)\right].
$$

- This inequality is known as *Markov's (general) inequality*, and is very useful for bounding the probability of particular events.

- Specifically, if $\varphi(x) = |x|^p$, with $p > 0$, then because $|X|$ is always positive, φ is non-negative, non-decreasing, and therefore

$$
P(|X| \geq a) \leq \frac{E\big[|X|^p\big]}{a^p},
$$

- If we restrict ourselves to the case where X is non-negative, we get the most standard version of the inequality:

$$
P(X \geq a) \leq E(X)/a.
$$

Markov's Inequality in Action
Suppose that an individual is taken randomly from a population that has an average salary of \$50,000. If we assume that salary from the population is approximately independently and identically distributed, we can provide an upper-bound for the probability that the individual is wealthy.
Let X_i be the salary of individual i, randomly drawn from said population. Even though all we know is the average salary, Markov's inequality tells use that:

$$
P(X \geq 200,000) \leq \frac{50,000}{200,000} = \frac{1}{4}.
$$

- Returning to expectations of functions of random variables, we can extend to the multi-variate case

Theorem 4.2: functions of multiple variables
Suppose that $X_1, \ldots, X_n$ are jointly distributed RVs and $Y = g(X_1, \ldots, X_n)$. Then

- IF X_i are discrete with pmf $p(x_1, \ldots, x_n)$, then

$$E(Y) = \sum_{x_1, \ldots, x_n} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n).$$

- If X_i are continuous with pdf $f(x_1, \ldots, x_n)$, then

$$E(Y) = \int_{\mathcal{X}_1, \ldots, \mathcal{X}_n} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \, dx_1 \, \ldots, dx_n.$$

In both cases, we need the sum (or integral) of $|g|$ to converge.

- The proof for the discrete case of Theorem 4.2 follows directly that of Theorem 4.1

- An immediate consequence of Theorem 4.2 is the following

Corollary 4.2.1
If X and Y are independent random variables, and g and h are fixed functions, then

$$E\big[g(X)h(Y)\big] = \Big(E\big[g(X)\big] E\big[h(Y)\big]\Big),$$

provided that the expectations on the right-hand side exist.

Example: Breaking sticks
A stick of unit-length is broken randomly (uniformly) in two places. What is the average length of the middle piece?
We will interpret this problem to mean that the locations of the two break-points are independent uniform random variables, U_1 and U_2, and we need to computing $E|U_1 - U_2|$.
Solution: Theorem 4.2 implies that we do not need to find the density function of $U_1 - U_2$. Instead, we just need to integrate $|u_1 - u_2|$ against the joint density: $f(u_1, u_2) = 1$, with $0 \leq u_1, u_2 \leq 1$. Thus

$$E|U_1 - U_2| = \int_0^1 \int_0^1 |u_1 - u_2| \, du_1 \, du_2$$

Splitting this integral into two regions, one where $u_1 \geq u_2$ and one where $u_2 > u_1$, we get

$$E|U_1 - U_2| = \int_0^1 \int_0^{u_1} (u_1 - u_2) \, du_2 \, du_1 + \int_0^1 \int_{u_1}^1 (u_2 - u_1) \, du_2 \, du_1$$
$$= \frac{1}{3}$$

Logically, this result makes sense as it suggests that if we have two break points (three pieces) and the break points are uniform and random, the middle piece, on average, will be 1/3 of the length of the original stick.

Linear Combinations of Random Variables

- A useful property of expectation is that it is a *linear operator*.

Theorem 4.3: Linear combinations

If $X_1, \ldots, X_n$ are jointly distributed random variables with expectations $E(X_i)$, respectively, and $Y = a + \sum_{i=1}^{n} b_i X_i$, then,

$$E(Y) = a + \sum_{i=1}^{n} b_i E(X_i).$$

Proof. We will show this for the continuous case with $n = 2$. The proof for the discrete case is similar, and this argument is readily extended to the case that $n > 2$. First, we will argue that the expectation is well-defined. By definition,

$$E|Y| = \int |a + b_1 x_1 + b_2 x_2| f(x_1, x_2) \, dx_1 \, dx_2$$

and by the triangle inequality $|a + b_1 x_1 + b_2 x_2| \leq |a| + |b_1||x_1| + |b_2||x_2|$, and the fact that $E[X_i]$ exists (which implies that $E|X_i| < \infty$), we see that $E|Y| < \infty$. Now we can calculate the expected value. By Theorem 4.2,

$$E(Y) = \int (a + b_1 x_1 + b_2 x_2) f(x_1, x_2) \, dx_1 \, dx_2$$

$$= a \int f(x_1, x_2) \, dx_1 \, dx_2 + b_1 \int x_1 f(x_1, x_2) \, dx_1 \, dx_2$$

$$+ b_2 \int x_2 f(x_1, x_2) \, dx_1 \, dx_2$$

Note that the first integral is equal to 1, because it's the integral of a joint pdf over the support. We'll focus now on the second integral, which is evaluated in a similar way as the third integral due to symmetry.

$$b_1 \int x_1 f(x_1, x_2) \, dx_1 \, dx_2 = b_1 \int x_1 \left(\int f(x_1, x_2) \, dx_2 \right) dx_1$$

$$= b_1 \int x_1 f(x_1) \, dx_1$$

$$= b_1 E[X_1].$$

Thus, applying the same idea to the third integral, we get

$$E(Y) = a + b_1 E(X_1) + b_2 E(X_2).$$

- The previous theorem is extremely useful for calculating expected values.

- An obvious example is *sums* of random variables, such as the arithmetic average.

- It's also useful because some distributions can be expressed as the sum of other distributions.

- For instance, we saw in a previous example that the sum of two exponential random variables has a Gamma distribution. Thus, if we know the mean of an exponential, we can readily calculate the mean of a Gamma distribution.

Expectation of a binomial distribution

Let Y follow a Binomial(p, q) distribution. Find the expected value of Y.

Solution. The pmf of a binomial distribution is given by

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

Therefore, to find the expected value directly, we need to calculate the sum

$$E(Y) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

It's not immediately clear how one can calculate this sum directly. Instead, we can use the fact that a binomial random variable is defined by the sum of independent Bernoulli distributed random variables. That is, let $X_1, X_2, \ldots, X_n$ be Bernoulli random variables with parameter p. Then

$$Y \stackrel{d}{=} \sum_{i=1}^{n} X_i$$

where the symbol $\stackrel{d}{=}$ is used to indicate that Y and $\sum_i X_i$ have the same distribution[1]. Then it is very easy to calculate the expected value of Y, as it is the sum of expected values of X_i:

$$E[X_i] = p(1) + (1-p)(0) = p.$$

Thus,

$$E[Y] = \sum_i E[X_i] = \sum_i p = np.$$

Example: Baseball Card Collection

Suppose that you collect baseball cards, that there are n distinct cards, and that on each trial you are equally likely to get a card of any of the types. How many trials would you expect to go through until you had a complete set of cards?

- Let X_1 denote the number of trials up to and including the trial on which the first type of card is collected. Since our first draw is guaranteed to give us a new type of card, we have $X_1 = 1$.

- Now let X_2 be the number of trials form that point up to and including the trail on which the next card of a new type is obtained.

- We can continue this definition, letting X_i be the number of trials needed to obtain the ith type of card, after $i - 1$ types of cards have already been obtained. Thus, the total number of trials needed until all types of cards have been obtained is $X = \sum_{i=1}^{n} X_i$.

- What is the distribution of X_r, for $1 \leq r \leq n$?

- When counting the number of trials to find the rth type, we have already found $r - 1$ unique types, leaving $n - r + 1$ types that we have not yet collected. We can see then that X_r is a geometric distributed random variable with $p = (n - r + 1)/n$ representing the probability of success. The

---

[1] they are not necessarily equal to each-other, as they are not necessarily defined using the same probability space. Instead, we only require that they have the same distribution.

expected value of a geometric random variable is $1/p$, and therefore we have

$$E[X] = \sum_{i=1}^{n} E[X_i]$$

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

$$= n \sum_{i=1}^{n} \frac{1}{n-i+1} = n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{1}\right)$$

$$= n \sum_{k=1}^{n} \frac{1}{k}$$

- As far as I am aware, there is not a way to express this final sum succinctly, but we can easily calculate the sum on a computer for moderate values of n. For very large values of n, there are some additional approximations that can be useful.

- For instance, if we suppose that there are roughly 1200 MLB players (40 players per team, 30 teams), and assume that all players are equally as likely to appear on a card (not a good assumption), then the expected number of cards we would need to purchase until we had a card for every player is $1200 \sum_{k=1}^{1200} \frac{1}{k}$. In R, we could calculate this using

$$\texttt{1200 * sum(1/(1:1200))} = 9201.25$$

- For most $n < 1 \times 10^8$, modern computers can calculate this sum almost instantly.

- For very large n, we could use the famous approximation

$$\sum_{k=1}^{n} \frac{1}{k} = \log n + \gamma + \epsilon_n,$$

where $\gamma \approx 0.577$ is "Euler's Constant" (which is often defined as the limit of difference between harmonic series and $\log n$), and $\epsilon_n \to 0$.

Example: Group Testing
Suppose that a large number, n of blood samples are screened for a rare disease. If each sample is taken individually, n tests will be required. An alternative approach is group individuals into m groups of size k, pool the blood samples for each group together and perform a test on the pooled sample. If the pooled test is negative, we know all individuals in the group do not have the rare disease; however, if the test is positive, we can then do tests on each individual in the smaller group. What is the expected number of tests that will be conducted using this approach?

- To solve a problem like this, it's important to give names to quantities of interest.

- We have n individuals we need to test, and m groups of size k, such that $n = mk$.

- Let X_i denote the number of tests conducted on the ith group. Thus, the total number of tests is $X = \sum_i X_i$.

- If a group tests negative, then $X_i = 1$. If a group tests positive, then we test all members of the group, so $X_i = k$.

- Let's let p denote the probability that an individual tests negative (assuming independence, we could let p be 1 minus the proportion of individuals that have the rare disease).

- The probability that a group tests negative is therefore p^k; in this case, the total number of tests is 1. The probability that a group tests positive is $1 - p^k$, and in this case, the total number of tests is $k + 1$ (one for the group, k for each individual test).

- Thus, the expected number of tests is

$$E[X_i] = p^k + (k+1)(1 - p^k) = k + 1 - kp^k.$$

- This expectation is the same for all groups, thus

$$E[X] = \sum_i E[X_i] = mE[X_i] = mk + m - mkp^k = n\left(1 + \frac{1}{k} - p^k\right).$$

- We can see now that the expectation of this group testing scenario is n times a proportion $\left(1 + \frac{1}{k} - p^K\right)$. Specifically, if we fix p at 1 minus the rate of disease occurrence in a large population, then the number of tests is a function of group size k.

- Consider using Desmos to play with the value p (start with something like 0.99) as a function of k.

Example: Counting DNA "words"
Within DNA patterns, we might be interested in finding the number of times a particular combination of letters (or "word") occurs in a DNA sequence. This can be useful for determining if a region of DNA has unusually large occurrences of specific sequences. Assume each sequence is randomly composed of letters A, C, G, T, and that for each location in the sequence, each letter has probability $1/4$. For example, consider occurrence of the "word" $TATA$.

$$ACTATATAGATATA$$

In the above sequence, we would count $TATA$ 3 times (counting overlaps). In a sequence of length N, what is the expected number of times a word of length q occurs?
Solution. To solve this problem, we will use indicator functions.

- Let I_n denote the event that the start of a word starts at position n of the sequence, for $n \in \{1, 2, \ldots, N - q + 1\}$.

- Thus, $I_n = 1$ if the word occurs in position n, and $I_n = 0$ otherwise.

- The total number of words in a sequence of length N is therefore

$$W = \sum_{n=1}^{N-q+1} I_n,$$

- because I_n only has two possibilities, it is Bernoulli distributed.

- Our task is now to find the value of p, the probability that a word occurs at position n.

- Our assumption that we made is that all letters are independent, and equally as likely to occur at any given position.

- Therefore, the probability that the first letter occurs at position n is $1/4$, and the probability that the second letter occurs at the position $n + 1$ is $1/4$, and so on.

- By the multiplication principle, $P(I_n = 1) = (1/4)^q$, and the expected value is $E[I_n] = (1/4)^q$.

- Thus, the expected value of W is

$$E[W] = \sum_{n=1}^{N-q+1} E[I_n] = (N - q + 1)(1/4)^q.$$

Expected value as a predictor

- One useful property of the expectation is that it serves as a good predictor for the value of a random variable.

- Suppose X is a random variable with well-defined expectation, and that we want to make a prediction for the value of X.

- Denote our predicted value of X as b.

- One common way to measure accuracy using the Mean-Squared Error (MSE), which is defined as:

$$\text{MSE}(b) = E\big[(X - b)^2\big].$$

- Here, the closer b is to X, the smaller $(X - b)^2$ is. We take the expectation because X is random.

- By this measure, the best predictor would minimize this error.

Theorem: Expectation and MSE
If X is a random variable, then the value b that minimizes $E\big[(X - b)^2\big]$ is $b = E[X]$:

$$\underset{b}{\arg\min}\, E\big[(X - b)^2\big] = E[X].$$

Proof.

$$
\begin{aligned}
E\big[(X - b)^2\big] &= E\big[(X - E(X) + E(X) - b)^2\big] \\
&= E\big[(X - E[X]) + (E[X] - b)^2\big] \\
&= E\big[(X - E[X])^2\big] + \big(E[X] - b\big)^2 + 2E\big[(X - E[X])(E[X] - b)\big].
\end{aligned}
$$

- Above, we have just expanded the square, and used the fact that both $E[X]$ and b are constants.

- Using this same idea, because $(E[X] - b)$ is just a constant, we can pull it out of the expectation in the last term, leaving $c \times E\big[X - E[X]\big]$ for some constant c, but $E\big[X - E[X]\big] = 0$.

- Thus, we have the equality

$$E\big[(X - b)^2\big] = E\big[(X - E[X])^2\big] + \big(E[X] - b\big)^2.$$

- As a reminder, we are trying to minimize this quantity with respect to b.

- The first term does not involve b, so it has no role in the minimization.

- The second term is a quadratic function of b, meaning it has a unique global minimum.

- You could take the derivative and set equal to zero at this point, or just notice that if $b = E[X]$, the quadratic function is equal to 0, which is the smallest it could be.

- Thus, $b = E[X]$ minimizes the MSE.

$\square$

Some comments on expected values

- An important thing to notice about the theorem for linear combinations is that we do not require independence.

- The last example demonstrates this principle. Though I_n is Bernoulli distributed, $\sum_n I_n$ is *NOT* binomial distributed, because the I_n are not independent.

- As an example, if our word is $TATA$, then $I_1 = 1$ implies that $I_2 = 0$, since a $TATA$ at position 1 implies that the second letter starts with A, and thus $TATA$ cannot occur at position 2.

- Despite this, we can still calculate the expected value of a sum by taking the sum of expected values.

- The expected value can be used as an indication of the central value of the density or frequency function.

- Because of this, the expected value is sometimes referred to as a *location parameter*.

- The expected value is not the only type of location parameter. For instance, the *median* is also a type of location parameter.

- We have seen a lot of parallel between the expected value of a discrete random variable and that of a continuous random variable. This is not a coincidence.

- Specifically, we generally just "swap" and integration with summation, and pdf with pmfs.

- With a more rigorous definition of expectation, we could define expectation as a *Lebesgue-Stieltjes* integral, with respect to some measure P.

- That is, $E(X) = \int_\Omega X dP$, where P is a probability measure. If the probability measure is a counting measure, then the integral *is* a sum.

- Note that this definition does not require the existence of a pdf; in fact, there distributions where the expectation is well-defined, but the pdf is not. These types of distributions do not come up often in standard examples.

4   Variance and Standard Deviation

Variance

- The expected value is useful for summarizing the average or expected behavior of a random variable.

- We are also often interested in the "spread" of a random variable.

- That is, if the expected value is the center (or location) of a distribution, we want an indication of how dispersed a distribution is around this center.

- The two most common ways to express this idea is the *variance* and *standard deviation* of a random variable.

Definition: Variance
If X is a random variable with expected value $E(X)$, then the *variance* of X is

$$\mathrm{Var}(X) = E\left[\left(X - E(X)\right)^2\right],$$

provided the expectation exists.

- Letting $\mu = E[X]$, we can use the identity $g(x) = (x - \mu)^2$, and our expression for $E\big[g(X)\big]$ to get a way of calculating the variance.

- If X is a discrete random variable, then by Theorem 4.1,

$$\text{Var}(X) = \sum_i (x_i - \mu)^2 \, p(x_i),$$

- If X is a continuous random variable, then

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx$$

Definition: Standard deviation
If X is a random variable, then the standard deviation of X is the square-root of the variance, provided it exists.

- The variance is often denoted by σ^2, and the standard deviation σ.

- Because $\big(X - E(X)\big)^2 \geq 0$, $\text{Var}(X) \geq 0$.

- Formally, the variance is the mean of the squared distance between X and $E[X]$. If most values of X are close to the mean, this value is small; and vice-versa if most values of X are far away from $E[X]$.

- By this definition, the units for the variance are squared units.

- That is, if X is measured in meters, then the variance is measured in square-meters, and the standard deviation is measured in meters.

Theorem 4.4: linear transformation of a single variable
Let X be a random variable, and assume that $\text{Var}(X)$ exists. Then if $Y = a + bX$, then $\text{Var}(X) = b^2\text{Var}(X)$.

Proof.

$$E\Big[\big(Y - E(Y)\big)^2\Big] = E\Big[\big(a + bX - (a + bE[X])\big)^2\Big]$$
$$= E\Big[b^2\big(X - E[X]\big)^2\Big]$$
$$= b^2 E\Big[\big(X - E[X]\big)^2\Big]$$
$$= b^2\text{Var}(X)$$

$\square$

- This result makes a lot of sense: adding a constant only "shifts" a distribution, it does not affect the spread.

- The multiplier does change the spread, and because we're squaring the difference, the multiplier is also squared.

- From this result, we can also see that the standard deviation also changes in a natural way.

- Specifically, if σ_Y, σ_X denote the standard deviations of X and Y, respectively, then

$$\sigma_Y = |b|\sigma_X.$$

- We take the absolute value, because variance and standard deviation are always positive, though the multiplier b might be negative.

Example: Bernoulli distribution

Let X be a Bernoulli(p) distributed random variable. What is the variance of X?

Solution. We'll calculate the variance using the definition of expectation. For a discrete random variable, that means summing the possible values by the corresponding probabilities:

$$
\begin{aligned}
\mathrm{Var}(X) &= \sum_x (x-p)^2 \, p(x) \\
&= (0-p)^2(1-p) + (1-p)^2(p) \\
&= p^2(1-p) + p(1-p)^2 \\
&= p^2 - p^3 + p - 2p^2 + p^3 \\
&= p(1-p).
\end{aligned}
$$

- Note that $p(1-p)$ is a quadratic function of p, that is maximized at $p = 1/2$.

- When $p = 0$ or $p = 1$, the variance is 0, because the value will have value $X = 0$ or $X = 1$ with probability 1.

Example: Normal distribution

Let $X \sim N(\mu, \sigma^2)$. What is $\mathrm{Var}(X)$?

Solution. Since we are already familiar with the normal distribution, we suspect that the variance is σ^2, both because it's the standard notation for variance, and because we call σ the standard deviation parameter for the distribution. However, let's quickly demonstrate that this is indeed the case. If we denote $E(X) = \mu$, then

$$
\begin{aligned}
\mathrm{Var}(X) &= E\left[\left(X - E[X]\right)^2\right] \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, dx.
\end{aligned}
$$

Making the change of variables $z = (x-\mu)/\sigma$, we have

$$
\begin{aligned}
\mathrm{Var}(X) &= \frac{\sigma^3}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} \, dz \\
&= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 e^{-z^2/2} \, dz.
\end{aligned}
$$

We now make another transformation, $u = z^2/2$, which implies that $du = z\,du$ and $z = \sqrt{2u}$, and thus

$$
\begin{aligned}
\mathrm{Var}(X) &= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} \frac{2u}{\sqrt{2u}} e^{-u} \, du \\
&= \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} u^{1/2} e^{-u} \, du.
\end{aligned}
$$

Now we can use the Gamma-distribution to help us evaluate this integral. Namely we need to find values of α and λ such that the integrand matches the pdf of a Gamma(α, λ) distribution:

$$
f(u) = \frac{\lambda^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\lambda u}, \quad u \geq 0.
$$

Thus if we let $\alpha = 3/2$ and $\lambda = 1$, then we find that the integral on the right hand side is

$$
\begin{aligned}
\text{Var}(X) &= \frac{2\sigma^2}{\sqrt{\pi}} \times \frac{\Gamma(\alpha)}{\lambda^{\alpha}} \\
&= \frac{2\sigma^2}{\sqrt{\pi}} \times \Gamma(3/2) \\
&= \frac{2\sigma^2}{\sqrt{\pi}} \frac{\sqrt{\pi}}{2} = \sigma^2.
\end{aligned}
$$

- Using the definition of variance, we will derive a very famous inequality.

Theorem 4.5: Chebyshev's Inequality
Let X be a random variable with $E[X] = \mu$, and $\text{Var}(X) = \sigma^2$. Then for any $t > 0$,

$$
P\big(|X - \mu| > t\big) \leq \frac{\sigma^2}{t^2}.
$$

Proof. The proof of the inequality is rather trivial using Markov's inequality. Let $Y = (X - \mu)^2$. Because Y is non-negative, we can use the most standard version of Markov's inequality:

$$
\begin{aligned}
P(Y > t^2) &\leq \frac{E[Y]}{t^2} \\
P\big((X - \mu)^2 > t^2\big) &\leq \frac{E\big[(X - \mu)^2\big]}{t^2} \\
P\big(|X - \mu| > t\big) &\leq \frac{\sigma^2}{t^2}.
\end{aligned}
$$

$\square$

- This theorem bounds the probability that the difference between X and $E[X]$ is larger than t.

- If σ^2 is small, then the probability that X deviates far away from the mean is also small.

- By letting $t = k\sigma$, we get a bound on the probability that a variable will be k-standard deviations away from the mean:
$$
P\big(|X - \mu| > k\sigma\big) \leq \frac{1}{k^2},
$$

- For instance, the probability that any arbitrary random variable X will be more than 4σ away from $E[X]$ is less than $1/16$.

- While applicable to all random variables with well-defined variances, it is not the most optimal bound we can achieve.

- For instance, if $X \sim N(\mu, \sigma^2)$, then $P(|X - \mu| > 1.96 \times \sigma) = 0.05 < 1/4$

Corollary: zero variance
Let X be a random variable with $\text{Var}(X) = 0$. Then $P(X = \mu) = 1$.

Proof. Suppose that $P(X = \mu) \neq 1$. Since P is a probability measure, we can deduce $P(X = \mu) < 1$ from this assumption. Thus, there must exist some $\epsilon > 0$ such that $P(|X - \mu| > \epsilon) > 0$. However, this leads us to a contradiction: using Chebyshev's inequality, we know that for all $\epsilon > 0$, $P(|X - \mu| > \epsilon) = 0$, and therefore our assumption must be false, implying that $P(X = \mu) = 1$. $\square$

Theorem 4.6: Variance Calculation

Let X be a random variable such that $\mathrm{Var}(X)$ exists. Then

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

where $\mu = E(X)$.

Proof.

$$\begin{aligned}
\mathrm{Var}(X) &= E\big[(X - \mu)^2\big] \\
&= E\big[X^2 - 2\mu X + \mu^2\big] \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2.
\end{aligned}$$

$\square$

- Theorem 4.6 is sometimes useful to help us calculate the variance of a random variable.

- Other times, the variance is known, and the theorem helps us calculate $E(X^2)$.

Example: Uniform distribution

Let $X \sim U(0, 1)$. Use Theorem 4.6 to find $\mathrm{Var}(X)$.

Solution. the pdf of X is $f(x) = 1[0 \le x \le 1]$. Then

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 x\, dx = 1/2.$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_0^1 x^2\, dx = 1/3.$$

Thus,

$$\mathrm{Var}(X) = E[X^2] - \big(E[X]\big)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

4.1 Bias-Variance Tradeoff

Measurement Error

- Often, values of interest cannot be known precisely, but instead must be determined by experimental procedures.

- For instance: measurements of weight, length, voltage, or intervals of time can be complex, and generally involve potential sources of error.

- The National Institute of Standards and Technology (NIST) in the US are charged with developing and maintaining measurement standards.

- Statisticians have historically been employed by these organizations to help with this endeavor.

- Typically, there are two main types of measurement error: *random* vs *systematic*.

- For instance, a sequence of repeated independent measurements made from the same instrument or experimental procedure may not give the same value each time. These uncontrollable differences are modeled as *random* error.

- However, there may be a *systematic* error that affects all measurements, such as poorly calibrated instruments, or errors that are associated with the method of measurement.

- Suppose that the true value of a quantity being measured is x_0. We have a random measurement X, which is modeled as
$$X = x_0 + \beta + \epsilon.$$

- Here, β is the systematic error, and ϵ is the random component of the error.

Definition: Bias

Let x_0 be the true value of a measurement, modeled as a random variable X such that
$$X = x_0 + \beta + \epsilon,$$
where $E(\epsilon) = 0$, $\text{Var}(\epsilon) = \sigma^2$. Then, we have
$$E[X] = x_0 + \beta.$$

The value $\beta = E(X - x_0)$ is called the *bias* of the random variable, and we say that X is an unbiased estimate of x_0 if $\beta = 0$.

- The two factors that impact the quality of our estimator is the bias β and the variance σ^2.

- If both $\beta = 0$ and $\sigma^2 = 0$, then we get a perfect measurement.

- Ideally, we want an estimator that minimizes the bias and the variance, though as we will see (Math 4451) there is a principle known as the *bias-variance* trade-off, which suggests that efforts to minimize bias often result in larger variance (and vice-versa).

- Many approaches in statistics we will cover next semester aim at finding estimators that are unbiased ($\beta = 0$), while having minimum variance as possible (that is, the minimum-variance unbiased estimator (MVUE)).

Theorem 4.7: Mean Squared Error

Let X be a random variable representing a random estimate for value x_0. The mean-squared error of the estimator X is defined as $\text{MSE}(X) = E\big[(X - x_0)^2\big]$. If β is the bias of the estimator and σ^2 the variance, then
$$\text{MSE}(X) = \beta^2 + \sigma^2.$$

Proof.

$$
\begin{aligned}
E\big[(X - x_0)^2\big] &= \text{Var}(X - x_0) + \big[E(X - x_0)\big]^2 \\
&= \text{Var}(X) + \beta^2 \\
&= \sigma^2 + \beta^2.
\end{aligned}
$$

$\square$

5   Covariance and Correlation

Covariance

- The variance of a random variable is a measure of its variability.

- The *covariance* of two random variables is a measure of their joint-variability.

- It's also used to measure how closely associated two random variables are.

Definition: Covariance

If X and Y are jointly distributed random variables with expectations μ_X and μ_Y, the covariance of X, Y is:

$$\text{Cov}(X, Y) = E\big[(X - \mu_x)(Y - \mu_y)\big]$$

- The covariance is the average value of the product of the deviation of X from it's mean, and Y from it's mean.

- If X and Y are positively associated, we expect that if a value of X is larger than it's mean, then the value of Y is also larger than it's mean.

- In this case, the covariance is positive.

- Example: Suppose X is a random variable representing height of an adult male, and Y is the weight. In this case, we expect heights larger than average will also have weights larger than average, so the covariance is positive.

Calculating Covariance

Let X and Y be random variables. Then

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y].$$

Proof.   • Using the notation that $E[X] = \mu_X$ and $E[Y] = \mu_Y$, we have:

$$\begin{aligned}
\text{Cov}(X, Y) &= E\big[(X - \mu_X)(Y - \mu_Y)\big] \\
&= E\big[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y\big] \\
&= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y \\
&= E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

$\square$

- One important example is when X and Y are independent:

- In this case, we have shown that $E[XY] = E[X]E[Y]$.

- Therefore, $\text{Cov}(X, Y) = E[X]E[Y] - E[X]E[Y] = 0$.

- **however**, the inverse is not true: Just because $\text{Cov}(X, Y) = 0$ does *not* imply X and Y are independent.

Example: Calculating Covariance

Let (X, Y) be jointly defined random variables is joint pdf $f(x, y) = 2x + 2y - 4xy$, for all $0 \leq x, y \leq 1$. Calculate the covariance $\mathrm{Cov}(X, Y)$.

Solution:

- You might notice that $f(x, y)$ was one of the copula functions considered in the last chapter, specifically the bivariate construction using the function $H(x, y)$, marginal uniform densities, and parameter $\alpha = -1$. This isn't necessary, but useful to note that the marginal distributions are uniform.

- We will use the identity $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.

- First, we have

$$E[X] = \int_0^1 x f_X(x) \, dx$$
$$= \int_0^1 x \left(\int_0^1 f(x, y) \, dy \right) \, dx$$
$$= \int_0^1 x \left(\int_0^1 (2x + 2y - 4xy) \, dy \right) \, dx$$
$$= \int_0^1 x(1) \, dx = 1/2$$

- By symmetry, we also have $E[Y] = E[X] = 1/2$.

- (Note in the calculation above, we find that X and Y are marginally Uniform$(0, 1)$ distributed).

- Now to calculate $E[XY]$:

$$E[XY] = \iint xy f(x, y) \, dx \, dy$$
$$= \int_0^1 \int_0^1 xy \, (2x + 2y - 4xy) dx dy$$
$$= \int_0^1 \left(\frac{2}{3}x^3 y + x^2 y^2 - \frac{4}{3}x^3 y^2 \right)\Big|_0^1 dy$$
$$= \int_0^1 \left(\frac{2}{3}y + y^2 - \frac{4}{3}y^2 \right) dy$$
$$= \left(\frac{1}{3}y^2 + \frac{1}{3}y^3 - \frac{4}{9}y^3 \right)\Big|_0^1$$
$$= \frac{1}{3} + \frac{1}{3} - \frac{4}{9} = \frac{2}{9}.$$

- Therefore, the covariance $\mathrm{Cov}(X, Y)$ is given by:

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{2}{9} - \frac{1}{4} = -\frac{1}{36}$$

Covariance Properties

- Covariance has several useful properties that can help with calculations.

- One of them is that the covariance is *bilinear* operator.

- You can also show that covariance is an inner-product for a particular inner-product space.

Theorem: Bilinear Covariance

Let X_i, $i = 1, 2, \ldots, n$ and Y_j, $j = 1, 2, \ldots, m$ be a collection of random variables, and a, c, b_i, d_j be real numbers for all i and j. Then:

$$\text{Cov}\Big(a + \sum_{i=1}^n X_i, c + \sum_{j=1}^m Y_i\Big) = \sum_{i=1}^n \sum_{j=1}^m b_i d_j \text{Cov}(X_i, Y_j).$$

In particular,

$$\text{Cov}(aX + bW, cY + dZ) = ac\,\text{Cov}(X, Y) + ad\,\text{Cov}(X, Z)$$
$$+ bc\,\text{Cov}(W, Y) + bd\,\text{Cov}(W, Z)$$

Additional properties of the covariance include:

- $\text{Cov}(X, X) = \text{Var}(X)$. Therefore,

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

- More generally,

$$\text{Var}\Big(a + \sum_{i=1}^n b_i X_i\Big) = \sum_{i=1}^n \sum_{j=1}^n b_i b_j \text{Cov}(X_i, X_j).$$

- If the X_i are independent, this implies that $\text{Var}\big(\sum_i X_i\big) = \sum_i \text{Var}(X_i)$.

Example: Variance of Binomial RV

Let X follow a Binomial(n, p) distribution. Calculate $\text{Var}(X)$.

- If we are trying to do this from definition, we need to calculate:

$$\text{Var}(X) = \sum_{k=0}^\infty (k - np)^2 \binom{n}{k} p^k (1 - p)^{n-k}.$$

- Directly calculating this sum is no easy task!

- We instead can use the fact that a binomial random variable can be expressed as the sum of independent Bernoulli(p) random variables.

- Let $X_1, X_2, \ldots, X_n$ be independent Bernoilli(p) random variables. Then $\sum_i X_i \overset{d}{=} X$, and

$$\begin{aligned}
\text{Var}(X) &= \text{Var}\Big(\sum_i X_i\Big) \\
&= \sum_{i=1}^n \text{Var}(X_i) \quad \text{(because of independence)} \\
&= \sum_{i=1}^n p(1 - p) = np(1 - p).
\end{aligned}$$

Example: Random Walk

A similar example is a *Random Walk*. Suppose we start a random process at $x_0 = 0$, and at each time point t_i, we take a random "step", following a X_i distribution, where $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. That is, our position after one step is $S(1) = x_0 + X_1$, and after two steps, $S(2) = x_0 + X_1 + X_2$, and so on. What's the mean and variance of the position after N steps?

- We let $S(N)$ denote the position after the Nth step.

- $S(N) = x_0 + \sum_{i=1}^{N} X_i$.

- By linearity of expectation,

$$E\big[S(N)\big] = x_0 + \sum_{i=1}^{N} E[X_i] = x_0 + n\mu$$

- Because the X_i are independent,

$$\text{Var}\big(S(N)\big) = \sum_{i=1}^{N} \text{Var}(X_i) = n\sigma^2.$$

- Thus, we expect our position to be $x_0 + n\mu$, with uncertainty measured by the standard deviation $\sqrt{n}\sigma$.

- If $\mu > 0$, then we will expect that our position will be larger than where we started (x_0), particularly when n is large.

- Random walks have found applications in many areas of science.

- A small extension to a random walk is when the time-steps are continuous, and the steps being normally distributed.

- This type of process is known as Brownian motion, named after a Biologist in the 1800's who used a similar idea to describe the spontaneous motion of pollen grains suspended in water. Einstein later explained this as a result of collisions of the grains with randomly moving water molecules.

- Random walks are still a popular model for modeling the evolution of stock-markets over time: short term behavior is generally unpredictable, but long term trends do follow a pattern.

- When we are interested in multiple random variables, covariance is often expressed as a matrix.

- Let $X_1, X_2, \ldots, X_n$ be random variables, and we denote $\boldsymbol{X}$ to be the random (column) vector, $\boldsymbol{X} = (X_1, \ldots, X_n)^T$.

- Then, the *variance-covariance* matrix is defined as:

$$\Sigma = \text{Var}(\boldsymbol{X}) = \text{Cov}(\boldsymbol{X}, \boldsymbol{X}) = E\big[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T\big].$$

- In particular, the (i,j)th entry $\Sigma_{i,j} = \text{Cov}(X_i, X_j)$.

- Σ is a symmetric, positive definite matrix.

Correlation

Definition: Correlation

If X and Y are jointly distributed random variables, and the variances and covariances exist, and the variances are non-zero, then the correlation of X and Y is:

$$\text{Cor}(X, Y) = \rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

- By how correlation is defined, it is a unit-less measure.

- Also, $-1 \le \rho \le 1$ (HW problem)?

6   Conditional Expectation

Conditional Expecation

- The idea of conditional distributions can be extended to conditional expectations.

  Definition: Conditional Expectation

  Let X and Y be jointly defined random variables. The *conditional expectation* of Y given $X = x$ is

  $$E[Y|X = x] = \begin{cases} \sum_y y\, p_{Y|X}(y|x) & \text{if } Y|X = x \text{ is discrete} \\ \int y f_{Y|X}(y|x)\, dy & \text{if } Y|X = x \text{ is continuous} \end{cases}$$

- In particular, for some function h, we have

  $$E[h(Y)|X = x] = \int h(y)\, f_{Y|X}(y|x)\, dy,$$

  and similar for the discrete case.

Theorem: Law of total expectation

(also called the tower property or the tower law)

$$E(Y) = E\big[E(Y|X)\big].$$

- We will show the desired identity holds for the discrete case, and note that the continuous case is justified in a similar fashion.

- Note that the outer-expectation is taken with respect to the random variable X, and the conditional expectation $E[Y|X = x]$ is a function of x.

- That is, for discrete or continuous random variables, the pmf / pdf used for the expectation is that for the variable X, and we can think of the expectation of Y as the expected value of the function $E[Y|X = x]$ for the random variable X.

$$E(Y) = \sum_y y\, p_Y(y) \quad \text{(definition)}$$

$$= \sum_y y \sum_x p_{Y|X}(y|x)\, p_X(x) \quad \text{(total prob.)}$$

$$= \sum_y \sum_x y\, p_{Y|X}(y|x)\, p_X(x)$$

$$= \sum_x p_X(x) \sum_y y\, p_{Y|X}(y|x)$$

$$= \sum_x p_X(x) E[Y|X = x]$$

$$= E\big[E(Y|X)\big]$$

Example: System Failure

Suppose that in a system, a component and backup unit both have mean lifetimes equal to μ. If the component fails, the system automatically substitutes the backup unit, but there is a probability p that something will go wrong and the backup won't be used correctly. Let T be the total lifetime of the system. Find the expected lifetime of the system.

Solution. Let X be an indicator random variable such that $X = 1$ if the substitution of the backup works correctly, and $X = 0$ if it does not. Then,

$$E[T|X = 1] = 2\mu, \quad E[T|X = 0] = \mu.$$

Using the law of total expectation, we have:

$$E[T] = E\big[E(T|X)\big] = E[T|X = 1]P(X = 1) + E[T|X = 0]P(X = 0) = \mu(2 - p).$$

Example: Random Sums

Let N be a random variable denoting the number of events, and $X_1, \ldots, X_N$ be the "size" of the events, which we assume to be independent and have the same mean: $E[X_i] = \mu$. For example, maybe N is the number of customers entering a store, and X_i is how long customer i spends in the store. Find the expected value of the random sum,

$$T = \sum_{i=1}^N X_i.$$

Solution. For any fixed $N = n$, we have

$$E[T|N = n] = E \sum_{i=1}^n X_i = \sum_{i=1}^n E[X_i] = n\mu.$$

Using the law of total probability,

$$E[T] = E\big[E(T|N)\big] = E[N\mu] = \mu E[N].$$

- This idea of a random sum is a useful model for a variety of situations.

- Following our example, it can be used to model the total amount of time that passes if there are a random number of events, each with random length of time.

- Other examples are useful in business (e.g., insurance): random number of events (insurance claims), each of a random size.

Theorem: Law of total variance
$$\text{Var}(Y) = \text{Var}\big[E(Y|X)\big] + E\big[\text{Var}(Y|X)\big].$$

Proof.   • First we explain what the notation above means. First, the variance of Y given $X = x$. Per our previous calculations on variance, we have

$$\text{Var}(Y|X = x) = E[Y^2|X = x] - \big[E(Y|X = x)\big]^2,$$

which is defined for all values of x. Therefore, just like we did for the conditional expectation, we can define $\text{Var}(Y|X)$ as a random variable by letting X be random. Thus, the following is a random variable,

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

Using the linearity of expectation, we have

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

• Similarly, we can write

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

• Finally, we can use the law of total expectation to write:

$$\begin{aligned}
\text{Var}(Y) &= E[Y^2] - \big(E[Y]\big)^2 \\
&= E\big[E[Y^2|X]\big] - \big[E\big(E[Y|X]\big)\big]^2
\end{aligned}$$

• Combining everything together, we get

$$\begin{aligned}
\text{Var}(Y) &= E\big[E[Y^2|X]\big] - \big[E\big(E[Y|X]\big)\big]^2 \\
&= E\big[E[Y^2|X]\big] + 0 - \big[E\big(E[Y|X]\big)\big]^2 \\
&= E\big[E[Y^2|X]\big] - E\big[\big(E[Y|X]\big)^2\big] + E\big[\big(E[Y|X]\big)^2\big] - \big[E\big(E[Y|X]\big)\big]^2 \\
&= E\big[\text{Var}(X|Y)\big] + \text{Var}\big(E[Y|X]\big)
\end{aligned}$$

$\square$

Example: Random Sums
Continuing the random sum example from before, let's assume that the X_i have the same variance, $\text{Var}(X_i) = \sigma^2$, and assume that $\text{Var}(N) < \infty$. If $T = \sum_{i=1}^{N} X_i$ represents the sum of N elements, then find $\text{Var}(T)$.
Solution. By the law of total variance,

$$\text{Var}(T) = E\big[\text{Var}(T|N)\big] + \text{Var}\big(E[T|N]\big).$$

We previously argued that $E[T|N] = NE[X_i] = N\mu$, thus

$$\text{Var}\big(E[T|N]\big) = \text{Var}(N\mu) = \mu^2 \text{Var}(N).$$

Also, because $\text{Var}(T|N = n) = \text{Var}\big(\sum_i X_i\big) = n\text{Var}(X)$, we have the random variable

$$\text{Var}(T|N) = N\text{Var}(X) = N\sigma^2,$$

and

$$E\big[\text{Var}(T|N)\big] = E[\sigma^2 N] = \sigma^2 E[N].$$

Putting it together, we have

$$\text{Var}(T) = E\big[\text{Var}(T|N)\big] + \text{Var}\big(E[T|N]\big) = \sigma^2 E[N] + \mu^2 \text{Var}(N).$$

6.1 Prediction

Prediciton

- A major topic in statistics is prediction: Can I use information about one variable to make inference on another?

- This is a primary outcome of many disciplines, including machine learning:

  - How will certain events impact large financial markets?
  - What will the impact be of a new medical treatment on health outcomes?
  - For AI: given an input question, what's the output that matches our training data?

- These are all types of conditional expectations.

- The first case we will consider is where there is a variable Y of interest (which is random), and we take a measurement X, which is also random.

  - For example, suppose we are interested in the volume of a tree, Y. This often is difficult to measure exactly, but we can measure the tree diameter X quickly. We want to predict Y given X.

- First, consider making a prediction c for the variable Y. As previously discussed, we may want to minimize

$$\text{MSE}(c) = E\big[(Y - c)^2\big] = \text{Var}(Y) + (\mu - c)^2$$

where $\mu = E[Y]$.

- The first part of the MSE does not depend on c, and we can't control it.

- The second part is minimized when $c = \mu = E[Y]$.

- Now instead of some constant c, consider using another variable X to make a prediction.

- Specifically, we want to predict Y using some function of X: $h(X)$.

- We might want to pick the function h such that the MSE $E[(Y - h(X))]^2$ is minimized.

- Using the law of total expectation, we get:

$$\text{MSE}(h) = E\big[(Y - h(X))^2\big] = E\Big[E\big((Y - h(X))^2 | X\big)\Big]$$

- The outer expectation is taken with respect to X.

- For every $X = x$, the inner expectation is minimized by setting $h(x) = E[Y|X = x]$.

- Thus, the minimizing function h is equal to:

$$h(X) = E[Y|X].$$

- Thus, for some prediction model $Y = h(X; \theta) + \epsilon$, the best predictor function h (in terms of MSE) is chosen such that $h(X; \theta) = E[Y|X]$. In other words, we are just fitting a conditional expectation.

- The practical limitation of the optimal prediction scheme above is that it requires knowing the joint distribution of Y and X, which is typically not known.

- For this reason, we generally make some assumptions about the relationship between the variables, or otherwise restrict the family of functions from which h comes from.

- A common approach is to pick the optimal *linear* predictor of Y.

- That is, rather than finding the best function h among all functions, we try to find the best function of the form $h(x) = \alpha + \beta x$.

- In this case, h depends on only two parameters, $\theta = (\alpha, \beta)$.

- Now we can calculate the best linear predictor analytically:

$$
\begin{aligned}
E\big[(Y - h(X; \theta))^2\big] &= E\big[(Y - \alpha - \beta X)^2\big] \\
&= \mathrm{Var}\big(Y - \alpha - \beta X\big) + \big[E(Y - \alpha - \beta X)\big]^2 \\
&= \mathrm{Var}\big(Y - \beta X\big) + \big[E(Y - \alpha - \beta X)\big]^2
\end{aligned}
$$

- Notably, α does not impact the first term, so we can select α to minimize the second term.

- Using the linearity of expectation, the second term (prior to squaring it) is equal to

$$
E(Y - \alpha - \beta X) = \mu_Y - \alpha - \beta \mu_X,
$$

- Thus, if $\alpha = \mu_Y - \beta \mu_X$, then the squared term is zero (which is a global minimum), making it the most optimal choice for α.

- For the first term, we can use the properties of variance to calculate

$$
\mathrm{Var}(Y - \beta X) = \sigma_Y^2 + \beta^2 \sigma_X^2 - 2\beta \sigma_{XY}.
$$

- This is a quadratic function of β, and we can find the minimum by taking the derivative with respect to β and setting it equal to zero, giving

$$
\beta = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{XY}}{\sigma_X^2} \frac{\sigma_X \sigma_Y}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \frac{\sigma_X \sigma_Y}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X}.
$$

- Putting these results together, we get the best estimate of Y to be:

$$
\hat{Y} = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X).
$$

- The MSE of this predictor is

$$
\begin{aligned}
\mathrm{MSE}(\alpha, \beta) &= E\big[(Y - \alpha - \beta X)^2\big] \\
&= \mathrm{Var}\big(Y - \alpha - \beta X\big) + \big[E(Y - \alpha - \beta X)\big]^2 \\
&= \mathrm{Var}\big(Y - \beta X\big) \\
&= \sigma_Y^2 + \big(\frac{\sigma_{XY}}{\sigma_X^2}\big)^2 \sigma_X^2 - 2\big(\frac{\sigma_{XY}}{\sigma_X^2}\big)\sigma_{XY} \\
&= \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} \\
&= \sigma_Y^2 - \rho^2 \sigma_Y^2 = \sigma^2(1 - \rho^2)
\end{aligned}
$$

- One thing to note is that the best linear predictor for Y given X only depends on the joint distribution of (X, Y) through their means, variances, and covariance.

- Thus, in practice, we don't need the entire joint distribution for a linear predictor.

- Also noteworthy is that the optimal linear predictor of $E[Y|X]$ matches the conditional mean if Y and X are jointly distributed following a bivariate normal distribution (See Example 4.1.1 B, Rice, 2007).

- This idea is later useful to demonstrate that minimizing the MSE for prediction problems is equivalent to performing maximum likelihood estimation under the assumption that the errors are normally distributed (Chapter 8 topic).

- The estimator we derived is also *unbiased*, meaning it's the best linear *unbiased* estimator (BLUE).

7 Moment Generating Functions

Moment Generating Functions

Definition: The Moment-Generating Function
The *moment-generating function* (mgf) of a random variable X is $M(t) = E[e^{tX}]$. If X is discrete, this means
$$M(t) = \sum_x e^{tx} \, p(x).$$

If X is continuous, then
$$M(t) = \int_{-\infty}^{\infty} e^{tX} \, f(x) \, dx.$$

- Despite it's appearance, the mgf is a very useful tool that can dramatically simplify certain calculations.

- The expectation (and consequently the mgf), *doesn't necessarily exist* for particular values of t.

- In the continuous case, the existence of the expectation depends on how rapidly the tails of the density decrease.

Theorem: MGF Uniqueness
If the moment-generating function exists for t in an open interval containing 0, it uniquely determines the probability distribution.

- We won't prove the theorem above because it does require some technical details regarding Laplace transforms. The implications are that if two random variables have the same mgf in an open interval containing zero, they have the same distribution.

- For some problems, we can find the mgf and then use that to find the unique probability distribution that it corresponds with.

- The name *moment generating function* comes from the fact that it can be used to find moments of a distribution.

  Definition: Moments

  Let X be a random variable. Then $E[X^r]$ is called the rth *moment*, if it exists.

- We have already encountered the first and second moments. Trivially, we have $E[X] = \mu$ is the first moment, and $\text{Var}(X) = E[X^2] - (E[X])^2$ is the difference between the second and first moments.

- The rth *central moment* (rather than ordinary moment) are defined as
$$E\big[(X - E[X])^r\big].$$

- The variance *is* the second central moment.

- The third central moment is called *skewness*, and is used to measure the asymmetry of a density about its mean; if a density is symmetric about the mean, then the skewness is zero. (HW problem?)

Theorem: Derivatives of the mgf
If the moment-generating function exists in an open interval containing zero, then the rth derivative of $M(t)$ evaluated at 0 is the rth moment:

$$M^{(r)}(0) = E(X^r).$$

Proof.   - The proof will be done for the continuous case. The discrete case follows in a similar fashion.

- We first assume that $M(t)$ exists in some neighborhood containing 0.

- Thus, by assumption, $E[e^{tX}]$ exists (for that same neighborhood), and by definition of expectation, $\int_X |e^{tx}| f_X(x) \, dx < \infty$.

- This gives us a bounding function to apply our version of the dominated convergence theorem.

- Now consider the first derivative:

$$\frac{d}{dt} M(t) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f(x) \, dx$$
$$= \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f(x) \, dx$$
$$= \int_{-\infty}^{\infty} x \, e^{tx} f(x) \, dx.$$

- Here, we swapped the derivative and integral using the dominated convergence theorem.

- Now, we evaluate the derivative $M'(t)$ at $t = 0$, which is in the existing neighborhood.

$$M'(0) = \int_{-\infty}^{\infty} x f(x) \, dx = E[X].$$

- Thus, we have shown that $M^{(r)}(0) = E[X^r]$ for $r = 1$.

- We could proceed by induction, however the same basic argument works for larger r. In particular, it's easy to see (here is where you'd apply the induction argument) that

$$M^{(r)}(t) = \int_{-\infty}^{\infty} x^r e^{tx} f(x) \, dx,$$

and therefore
$$M^{(r)}(0) = \int_{-\infty}^{\infty} x^r f(x) \, dx = E[X^r].$$

$\square$

- This last theorem is extremely useful for finding moments of random variables.

- Without the theorem, we have to calculate infinite sums or indefinite integrals. Now, we can just find the MGF (often given already), and do some differentiation (easy).

Example: Poisson Distribution

Suppose X has a Poisson(λ) distribution. Find $E[X]$ and $\text{Var}(X)$.

Solution. To calculate the mean and variance, it would require evaluating the following sums:

$$E[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda}, \qquad \text{Var}(X) = \sum_{k-0}^{\infty} (k - E[X])^2 \frac{\lambda^k}{k!} e^{-\lambda}.$$

- Although we have calculated the first sum before, the second sum remains very challenging to compute.

- Instead of calculating the sum directly, we will use the MGF.

$$\begin{aligned} M(t) &= \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda} e^{\lambda \exp(t)} \\ &= e^{\lambda(\exp(t) - 1)}. \end{aligned}$$

- Note that this is true for all $t \in \mathbb{R}$, so our neighborhood containing zero that $M(t)$ is defined on is $\mathbb{R}$.

- Differentiating with respect to t, we get:

$$\begin{aligned} M'(t) &= \lambda e^t e^{\lambda(\exp(t) - 1)} \\ M''(t) &= \lambda e^t e^{\lambda(\exp(t) - 1)} + \lambda^2 e^{2t} e^{\lambda(\exp(t) - 1)}. \end{aligned}$$

- Therefore, we find by evaluating at $t = 0$,

$$\begin{aligned} M'(0) &= E[X] = \lambda \\ M''(0) &= E[X^2] = \lambda^2 + \lambda \end{aligned}$$

- From this, we can calculate the variance:

$$\text{Var}(X) = E[X^2] - \left[E(X)\right]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Example: Gamma Distribution

Let $X \sim \text{Gamma}(\alpha, \lambda)$, and find $E[X]$ and $\text{Var}(X)$.

Solution. First, recall the pdf of the Gamma distribution:

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

- Applying the MGF definition, we have

$$\begin{aligned} M(t) &= \int_0^{\infty} e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x} \, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha - 1} e^{x(t - \lambda)} \, dx. \end{aligned}$$

- We can solve this integral using "integration by pdf".

31

- In particular, we note that the integrand is just the density of a Gamma$(\alpha, \lambda - t)$ distribution without the normalizing constant.

- Since this integrates for $\lambda - t > 0$, as long as $t < \lambda$ (the neighborhood containing 0), we have

$$M(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{\Gamma(\alpha)}{(\lambda - t)^\alpha} \right) = \left(\frac{\lambda}{\lambda - t} \right)^\alpha.$$

- What's left now is to differentiate with respect to t, and evaluate at $t = 0$, giving:

$$M'(0) = E(X) = \frac{\alpha}{\lambda}$$
$$M''(0) = E(X^2) = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

- Therefore, we can find the variance as

$$\mathrm{Var}(X) = E[X^2] - \big[E(X) \big]^2 = \frac{\alpha(\alpha + 1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

Example: Standard Normal Distribution
Find the mgf of a standard normal distribution.
Solution.

- First recall the pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

- The mgf is:

$$M(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tx} e^{-x^2/2} \, dx$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2 + tx} \, dx.$$

- When dealing with normal pdf / integrals, a common strategy is completing the square. That is, we will write:

$$\frac{x^2}{2} - tx = \frac{1}{2}(x^2 - 2tx + t^2) - \frac{t^2}{2} = \frac{1}{2}(x - t)^2 - t^2/2.$$

- Thus, the integral can be written as:

$$M(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2 + t^2/2} \, dx$$
$$= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} \, dx$$

- The integral can once again be computed using "integration by density", in this case, it's the integral for the pdf of a $N(t, 1)$ distribution. Therefore:

$$M(t) = e^{t^2/2}.$$

Theorem: MGF of linear transformations
If X is a random variable with mgf $M_X(t)$, and $Y = a + bX$, the Y has the mgf $M_Y(t) = e^{at} M_X(bt)$.

Proof.

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

$\square$

Example: MGF of General Normal Distribution
If Y follows a general normal distribution with mean μ variance σ^2, then the distribution of Y is the same as the distribution of $\mu + \sigma X$, where X is a standard normal distribution ($Y \stackrel{d}{=} \mu + \sigma X$).
By the previous theorem on linear transformations, and uniqueness of the mgf, we have the mgf of Y:

$$
M_Y(t) = e^{\mu t}M_X(\sigma t) = e^{\mu t}e^{\sigma^2 t^2/2}.
$$

Theorem: MGF of independent variables
If X and Y are independent random variables with mgf's M_X and M_Y, respeictively, and $Z = X + Y$, then $M_Z(t) = M_X(t)M_Y(t)$ is the mgf of Z, where the values t are the common interval where both mgf's exist.

Proof.

$$
\begin{aligned}
M_Z(t) &= E(e^{tZ}) \\
&= E(e^{tX+tY}) \\
&= E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t).
\end{aligned}
$$

$\square$

- We extend the idea of the mgf to more than one variable.

- For instance, if (X, Y) are jointly distributed (not-independent), we define the joint mgf as:

$$
M_{XY}(s,t) = E(e^{sX+tY}).
$$

- Similar to the uni-variate case, the joint mgf (if it exists) uniquely determines the joint distribution. Also, the joint mgf can be used to find $E(XY)$ and higher-order moments.

- It can be shown that X and Y are independent if and only if their joint mgf factors into the product of the mgf of the marginal distributions.

- For more than two random variables, e.g., $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$, the joint mgf is is

$$
M_{\boldsymbol{X}}(\boldsymbol{t}) = E[e^{\boldsymbol{t}^T\boldsymbol{X}}].
$$

- While the mgf is very useful, the primary limitation is that the mgf may not exist.

- For this reason, we can often consider a similar function known as the *characteristic function*.

Definition: the characteristic function

If X is a random variable, the *characteristic function* of X is defined to be

$$\phi(t) = E(e^{itX}),$$

where $i = \sqrt{-1}$.

- We won't really use this function in this class, because it requires some experience with complex analysis.

- However, one thing of note is that $|e^{itX}| \leq 1$ for all t, and as such the expectation always exists (unlike the mgf).

- This function has many similar properties to the mgf. For instance, it uniquely determines a probability distribution, can be used to find moments, etc.

Final comments

- We're going to skip section 4.6 of Rice (2007) (might return to this later).

- However, it's fairly interesting material that discusses approximation methods.

- For instance, suppose we have a random variable X, and we only know the mean μ_X and variance σ_X^2.

- Now suppose we have $Y = g(X)$, and we want to make inference on Y.

- Even with limited information, we can use a Taylor series approximation to get

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X),$$

and taking expectations, derive $\mu_y \approx g(\mu_X)$ and $\sigma_Y^2 \approx \sigma_X^2 [g'(\mu_X)]^2$.

- This is sometimes called the propagation of error, and works well if g is well approximated by a linear function near μ_X.

Acknowledgments

References

Resnick S (2019). *A probability path*. Springer. 2, 3

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 1, 16, 18