Mathematical Statistics I

Chapter 4: Expected Values

Jesse Wheeler

Contents

1 Discrete random variables **1**

2 Continuous random variables **3**

3 Expectation of functions of random variables **4**

4 Variance and Standard Deviation **13**

5 Bias-Variance Tradeoff **17**

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

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

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

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

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

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

The question is: **Can we move the derivative inside of the infinite sum?** For this particular case, the answer is *yes*. In more advanced analysis classes, you learn methods for justifying this step rigorously using uniform convergence. Specifically, what we need is for uniform convergence of the partial sums and their derivatives. Fortunately for this class, all of the sums (and integrals) we will consider will be "well-behaved" and will satisfy these conditions.
With this sorted out, we can now use our trick to finish the calculation:

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

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 $E[X]$?

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

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.

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 Gamma(α^*, λ) 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 analog.

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 $\sum_x |g(x)|p(x) < \infty$.
- 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 $E\big[|Z|\big] \geq 0 = \big|E[Z]\big|$.

- An immediate consequence is 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

- An interesting example is *indicator* functions.

- 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.

- Example: 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*:

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

- This fact is useful for deriving some important inequalities.

- 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, if Y has a pdf, 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 both sides,

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

- 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

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

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 $E(X_i)$, repsectively, 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,

$$\begin{aligned} E(Y) &= \int (a + b_1 x_1 + b_2 x_2) f(x_1, x_2)\, dx_1\, dx_2 \\ &= a \int f(x_1, x_2)\, dx_1\, dx_2 + b_1 \int x_1 f(x_1, x_2)\, dx_1\, dx_2 \\ &\quad + b_2 \int x_2 f(x_1, x_2)\, dx_1\, dx_2 \end{aligned}$$

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.

$$\begin{aligned} b_1 \int x_1 f(x_1, x_2)\, dx_1\, dx_2 &= b_1 \int x_1 \left(\int f(x_1, x_2)\, dx_2 \right) dx_1 \\ &= b_1 \int x_1 f(x_1)\, dx_1 \\ &= b_1 E[X_1]. \end{aligned}$$

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 $X_1, X_2, \dots, X_n$ be Bernoulli random variables with parameter p. Then

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

where the symbol $\overset{d}{=}$ 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 $X = \sum_{i=1}^{n} X_i$.

[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.

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

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} E[X_i] \\ &= \sum_{i=1}^{n} \frac{n}{n-i+1} \\ &= n \sum_{i=1}^{n} \frac{1}{n-i+1} = n\Big(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{1}\Big) \\ &= n \sum_{k=1}^{n} \frac{1}{k} \end{aligned}$$

- 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

`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

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

- We can see now that the expectation of this group testing scenario is n times a proportion $\Big(1 + \frac{1}{k} - p^K\Big)$. 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 $n \in \{1, 2, \dots, N - q + 1\}$.
- 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

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

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

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

provided the expectation exists.

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

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

  where $\mu = E[X]$.
- 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.

$$\begin{aligned}
E\Big[\big(Y - E(Y)\big)^2\Big] &= E\Big[\big(a + bX - (a + bE[X]\big)^2\Big] \\
&= E\Big[b^2\big(X - E[X]\big)^2\Big] \\
&= b^2 E\Big[\big(X - E[X]\big)^2\Big] \\
&= b^2 \mathrm{Var}(X)
\end{aligned}$$

□

- 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

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

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 $du = zdu$ 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

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

- 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 $\mathrm{Var}(X) = \sigma^2$. 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:

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

□

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

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

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

- 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 $f(x) = 1[0 \leq x \leq 1]$. 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,
$$\mathrm{Var}(X) = E[X^2] - \big(E[X]\big)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

5 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.

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

□

Acknowledgments

- Compiled on September 2, 2025 using R version 4.5.1.
- Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin.
- We acknowledge students and instructors for previous versions of this course / slides.

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