Mathematical Statistics I

Chapter 4: Expected Values

Jesse Wheeler

Outline

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).

Expectation: Discrete random variables II

- 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: $E_X(X)$.

Expectation: Discrete random variables III

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).

Expectation: Discrete random variables IV

- 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?

Solution: $\underline{p(k) = P(X=k) = p(1-p)^{k-1}}$, $X \sim \text{Geom}(p)$

$$E[X] = \sum_{x \in \mathcal{X}} x\, p(x)$$

$$= \sum_{x=1}^{\infty} x\, p(1-p)^{x-1}$$

$$= p \sum_{x=1}^{\infty} x(1-p)^{x-1} \longleftarrow$$

$$= p\sum_{x=1}^{\infty} x q^{x-1}$$

$q = (1-p)$.

$$\frac{d}{dq} q^{x} = x q^{x-1}$$

we know $\sum_{x=1}^{\infty} q^{x} = \sum_{x=1}^{\infty} q q^{x-1} = q \sum_{j=0}^{\infty} q^{j}$

$$= \frac{q}{1-q}$$

$$p\sum_{x=1}^{\infty} x q^{x-1} = p\sum_{x=1}^{\infty} \frac{d}{dq} q^{x}$$

$$\overset{?}{=} p\frac{d}{dq}\left(\sum_{x=1}^{\infty} q^{x}\right) = p\frac{d}{dq}\left(\frac{q}{1-q}\right)$$

$$= \frac{p}{(1-q)^2}$$

$$\sum x\, p(x) = \frac{p}{(1-q)^2}$$

$$= \frac{p}{p^2} = \frac{1}{p} = E[X]$$

Poisson Distribution

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

Solution:

$$E[X] = \sum_{k=0}^{\infty} k\, p(k)$$

$$= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\frac{1}{(k-1)!}$$

$$= e^{-\lambda} \sum_{k=1}^{\infty} \frac{k \lambda^{k-1} \cdot \lambda}{k!}$$

$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

$$= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^{j}}{j!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

$E[X]$

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).

Expectation: Continuous random variables II

- 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

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

Find $E[X]$.

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

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

$\bullet\ u = x, \quad dv = \frac{1}{\lambda} e^{-x/\lambda}$

$du = dx \quad v = \int \frac{1}{\lambda} e^{-x/\lambda}\, dx = -e^{-x/\lambda} + C$

$$\int x \cdot \frac{1}{\lambda} e^{-x/\lambda}\, dx = -x e^{-x/\lambda} \Big|_0^{\infty} + \int_0^{\infty} e^{-x/\lambda}\, dx$$

$$= \int_0^{\infty} -e^{-x/\lambda}\, dx$$

$$= \lambda e^{-x/\lambda} \Big|_0^{\infty} = \lambda$$

Expectation: Continuous random variables IV

Solution.

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 $E(X)$.

"integration by distribution"

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

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

Expectation: Continuous random variables VI

Solution. $\alpha^* = \alpha+1$ $\quad x^\alpha = x^{(\alpha^*-1)}$

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

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha^*)}{\lambda^{\alpha^*}} \underbrace{\int_0^\infty \frac{\lambda^{\alpha^*}}{\Gamma(\alpha^*)} x^{\alpha^*-1} e^{-\lambda x}\, dx}_{1}$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}}$$

$$= \frac{\alpha}{\lambda}$$

$\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)$

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.

Functions of random variables II

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:

Functions of random variables: proof II

- 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.

Expected value of indicator functions II

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

- This fact is useful for deriving some important inequalities.
- First, we will show that the expectations of interest actually exist.

Expected value of indicator functions III

- 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 $x \geq a$, $\varphi(x)/\varphi(a) \geq 1$.

Expected value of indicator functions IV

- 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),

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

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

Expected value of indicator functions V

- Specifically, if $\varphi(x) = |x|^p$, with $p > 0$, then because $|X|$ is always positive, φ is non-negative, non-decreasing, and therefore
$$P(|X| \geq a) \leq \frac{E\left[|X|^p\right]}{a^p},$$
- If we restrict ourselves to the case where X is non-negative, we get the most standard version of the inequality:
$$P(X \geq a) \leq E(X)/a.$$

Expected value of indicator functions VI

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:

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

Expected value of indicator functions VII

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

Expected value of indicator functions VIII

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.

Expected value of indicator functions X

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:

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,

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

Linear Combinations of Random Variables II

Proof.

Linear Combinations of Random Variables III

- 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.

Linear Combinations of Random Variables IV

Expectation of a binomial distribution

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

Solution:

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?

Linear Combinations of Random Variables VI

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?

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?

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.

Expected value as a predictor II

Theorem: Expectation and MSE

If X is a random variable, then the value b that minimizes $E\left[(X-b)^2\right]$ is $b = E[X]$:

$$\underset{b}{\operatorname{argmin}}\, E\left[(X-b)^2\right] = E[X].$$

Proof:

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.

Some comments on expected values II

- 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.

Some comments on expected values III

- 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.

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.

Variance II

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.

Variance III

- 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,

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

- If X is a continuous random variable, then

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

Variance IV

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

Variance V

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

Variance VI

Theorem 4.4: linear transformation of a single variable

Let X be a random variable, and assume that $\mathrm{Var}(X)$ exists. Then if $Y = a + bX$, then $\mathrm{Var}(X) = b^2\mathrm{Var}(X)$.

Proof.

Variance VII

- 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.

Variance VIII

Example: Bernoulli distribution

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

Variance IX

Example: Normal distribution

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

Variance X

- 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 $t > 0$,

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

Variance XI

- 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 $1/16$.

Variance XII

- 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 $\mathrm{Var}(X) = 0$. Then $P(X = \mu) = 1$.

Variance XIII

Theorem 4.6: Variance Calculation

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

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

where $\mu = E(X)$.

Variance XIV

- 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 $\mathrm{Var}(X)$.

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.

Measurement Error II

- 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.

Measurement Error III

- 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.

Measurement Error IV

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 $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$. 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 $\beta = 0$.

Measurement Error V

- 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)).

Measurement Error VI

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 $\mathsf{MSE}(X) = E\big[(X - x_0)^2\big]$. If β is the bias of the estimator and σ^2 the variance, then

$$\mathsf{MSE}(X) = \beta^2 + \sigma^2.$$

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:

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

Covariance II

- 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.

Covariance III

Calculating Covariance

Let X and Y be random variables. Then

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

Proof.

Covariance IV

- One important example is when X and Y are independent:
- In this case, we have shown that $E[XY] = E[X]E[Y]$.
- Therefore, $\mathrm{Cov}(X, Y) = E[X]E[Y] - E[X]E[Y] = 0$.
- **however**, the inverse is not true: Just because $\mathrm{Cov}(X, Y) = 0$ does *not* imply X and Y are independent.

Covariance V

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:

Covariance VI

Solution cont...

Covariance Properties

- Covariance has several useful properties that can help with calculations.
- One of theme is that the covariance is bilinear operator.
- You can also show that covariance is an inner-product for a particular inner-product space.

Covariance Properties II

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:

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

In particular,

$$\begin{aligned} \mathrm{Cov}(aX + bW, cY + dZ) &= ac\,\mathrm{Cov}(X, Y) + ad\,\mathrm{Cov}(X, Z) \\ &\quad + bc\,\mathrm{Cov}(W, Y) + bd\,\mathrm{Cov}(W, Z) \end{aligned}$$

Covariance Properties III

Additional properties of the covariance include:

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

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

- More generally,

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

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

Covariance Properties IV

Example: Variance of Binomial RV

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

Solution.

Covariance Properties V

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 $\mathrm{Var}(X_i) = \sigma^2$. 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?

Covariance Properties VI

- 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:

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

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

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

- By how correlation is defined, it is a unit-less measure.
- Also, $-1 \leq rho \leq 1$ (HW problem)?

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

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

Conditional Expecation II

- 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.

Conditional Expecation III

Theorem: Law of total expectation

(also called the tower property or the tower law)

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

Proof.

Conditional Expecation IV

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.

Conditional Expecation V

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.

Conditional Expecation VI

Theorem: Law of total variance

$$\mathrm{Var}(Y) = \mathrm{Var}\big[E(Y|X)\big] + E\big[\mathrm{Var}(Y|X)\big].$$

Proof.

Conditional Expecation VII

Example: Random Sums

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

Solution.

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.

Prediciton II

- 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

$$\mathsf{MSE}(c) = E\left[(Y - c)^2\right] = \mathrm{Var}(Y) + (\mu - c)^2$$

where $\mu = E[Y]$.

Prediciton III

- 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.

Prediciton IV

- 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.

Prediciton V

- 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, $\theta = (\alpha, \beta)$.

Prediciton VI

- 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

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

Prediciton VII

- 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

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

Prediciton VIII

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

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

Prediciton IX

- The MSE of this predictor is

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

Prediciton X

- 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).

Prediciton XI

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

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.

Moment Generating Functions II

- 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.

Moment Generating Functions III

- 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 $\mathrm{Var}(X) = E[X^2] - (E[X])^2$ is the difference between the second and first moments.

Moment Generating Functions IV

- 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?)

Moment Generating Functions V

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.

Moment Generating Functions VI

- 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 $\mathrm{Var}(X)$.

Solution.

Moment Generating Functions VII

Example: Gamma Distribution

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

Solution.

Moment Generating Functions VIII

Example: Standard Normal Distribution

Find the mgf of a standard normal distribution.

Solution:

Moment Generating Functions IX

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.

Moment Generating Functions X

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:

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

Moment Generating Functions XI

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.

Moment Generating Functions XII

- 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.

Moment Generating Functions XIII

- 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.

Moment Generating Functions XIV

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.

Final comments II

- 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.

References and Acknowledgements

Resnick S (2019). *A probability path*. Springer.

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

- Compiled on October 28, 2025 using R version 4.5.1.

- We acknowledge students and instructors for previous versions of this course / slides.