

Math and Proof Techniques in Linear Models: Expectation Theorems

Robert Gulotty

February 18, 2025

There are several tricks used repeatedly in mathematical demonstrations and proofs in the world of statistics. This document walks through two expectation theorems, labeled T5 and T8, from a very good text by Goldberger. While that text offers proofs, I will show the steps at each stage. By copying the steps in this guide you will see how to write your own notes that fill in the gaps.

1 How and why to read this document

This document goes through two theorems in Goldberger and the associated proofs. You have may not have worked with ‘proofs’ much before, and there are several common problems people face when dealing with them:

- When is a proof necessary?
- How do I know when I am done or have proven enough?
- What is the structure of a proof?

Proofs are how scholars of mathematics and related fields communicate deductive reasoning. They explain how to go from premises to a conclusion by reference to accepted procedures. Getting used to their structure, how assumptions are stated, and how conclusions are drawn, takes time and effort.

One part of the difficulty is getting used to notation: math uses an unfamiliar set of symbols, and sometimes proofs use familiar English words in unfamiliar ways. For example, the sentence, ”If the butler was at the party, he is not guilty”, in math, is identical to the claim, ”either the butler was not at the party or he is not guilty (or both)”. For most non-logicians, finding that the butler was not at the party is materially irrelevant to the truth of the claim. A mathematical logician, by contrast, would only allow the claim to be false if they could be assured that the butler is guilty and that he nonetheless was at the party. In this class, we will appeal to procedures and steps that require how to add and subtract, sometimes multiply, and very rarely take a derivative or integral. Other times we will be applying definitions. This requires attention and care that takes time to learn.

The most difficult part, however, is that unlike most of the math you have likely seen, proof techniques are not mechanical. They require you to try things and be creative.

2 Example 1: T5

On page 45 of Goldberger we see the following classic result on the linearity of the expectation operator:

T5. LINEAR FUNCTION. Suppose that $Z = a + bX + cY$, where a, b, c are constants. Then

$$E(Z) = a + bE(X) + cE(Y),$$

$$V(Z) = b^2V(X) + c^2V(Y) + 2bcC(X, Y).$$

Proof. For the expectation,

$$\begin{aligned} E(Z) &= \sum_i \sum_j (a + bx_i + cy_j)f(x_i, y_j) \\ &= a \sum_i \sum_j f(x_i, y_j) + b \sum_i x_i \left[\sum_j f(x_i, y_j) \right] + c \sum_j y_j \left[\sum_i f(x_i, y_j) \right] \\ &= a \quad 1 \quad + b \sum_i x_i f_1(x_i) \quad + c \sum_j y_j f_2(y_j). \end{aligned}$$

For the variance, $V(Z) = E(Z^{*2})$, where

$$Z^* = Z - E(Z) = b[X - E(X)] + c[Y - E(Y)] = bX^* + cY^*.$$

Expanding the square gives Z^{*2} as a linear function of X^{*2} , Y^{*2} , and X^*Y^* . Use the rule for expectation of a linear function, extended to handle three variables. ■

Our goal is to unpack this proof. Go get your pencil and paper and follow along.

2.1 Notation

There are two math statements in T5, one statement is about the expectation, the other the variance. Both proofs are located in the text between italicized “Proof” and the ■ at the end. These two proofs contain many of the properties of probability and the tricks you will see all the time in linear models.

The following notation is introduced earlier in the text, it can be helpful to write these down first:

- Z, X, Y are random variables, z_k, x_i and y_j are values that these random variables produce.
- $f(z_k)$ is a probability density function evaluated at z_k , shorthand for $f(Z = z_k)$.
- $f(x_i, y_j)$ is a joint probability density function evaluated at x_i and y_j , shorthand for $f(X = x_i, Y = y_j)$.
- $f_1(x_i), f_2(y_j)$ are marginal probability density functions, sometimes written $f_X(x_i), f_Y(y_j)$.
- E stands for the expectation function, formally,

$$E(X) = \sum_i x_i f(x_i)$$

- V the variance function,

$$V(X) = \sum_i (x_i - E(X))^2 f(x_i)$$

- $C(X, Y)$ is the covariance function,

$$C(X, Y) = \sum_i \sum_j (x_i - E(X))(y_j - E(Y)) f(x_i, y_j)$$

3 Part 1: Linearity of the Expectation

We begin our discussion with the expectation result: Suppose that $Z = a + bX + cY$ and that a , b , and c are constants, then

$$E(Z) = a + bE(X) + cE(Y).$$

That is, the expectation function is *linear*. In general a function $F(x)$ is *linear* if $F(x+y) = F(x) + F(y)$ and $F(ax) = aF(x)$.

3.1 Proof Strategy

The mathematical statement has an “if A then B” structure:

If $\underbrace{Z \text{ is a linear function of random variables},}_{\text{A}}$

then $\underbrace{\text{the expectation of } Z \text{ is equal to a sum of expectations.}}_{\text{B}}$

The proof strategy here is called “direct proof”, starting from the *A* part (assumptions about Z) and showing that the *B* part ($E(Z) = a + bE(X) + cE(Y)$) follows.

The way that you do this in practice is to work both forwards and backwards. Forwards in the sense that we start applying definitions to the “A” part, backwards in the sense we apply definitions to the “B” part. We then hope we can meet in the middle.

To remind oneself, always start by writing the *B* part down with “we want to show that:”. In this case, write down

“We want to show that $E(Z) = a + bE(X) + cE(Y)$ ”.

Step 1: Applying definitions forwards and backwards

3.1.1 Forwards

Step 1 is to apply definitions from “A” to the left hand side of the equality. This is the most common first step in proofs. There are two things we can write out from the supposition at the beginning of the statement. 1) The definition of the expectation of Z and 2) the definition of Z .

The first line of the proof is to write out the expectation (see equation (1) below). Expectations are weighted averages, summing over all values of z , where the weights are the distribution (pdf or pmf) of Z .

The second line of the proof is to apply the definition of Z , displayed as equation (2) below. This means replacing Z with its definition in the statement, which is in terms of a , b , c , x_i and y_j . The sum in the expectation of Z sums over all possible values of Z . Because Z is a function of two other random variables, x and y , we need each possible combination of x and y , e.g. $(x_1, y_1), (x_1, y_2) \dots (x_2, y_1) \dots (x_i, y_j), \dots$

The dots indicate there need not be a finite number of these combinations. This long summation is captured in the double summation, moving from line 1 to 2.

$$E(Z) = \sum_k z_k f(z_k) \quad (1)$$

$$= \sum_i \sum_j (a + bx_i + cy_j) f(x_i, y_j) \quad (2)$$

We now have the first step of Goldberger's proof. Note we have also replaced the distribution of Z with the joint distribution of X and Y . More on this below.

3.1.2 Aside on summation ordering

Double summations (and integrals) are used quite a bit, remember you can reverse the order like in the following example:

$$\begin{aligned} \sum_i [\sum_j x_i * y_j] &= \sum_i [x_i * y_1 + x_i * y_2 + x_i * y_3 + \dots] \\ &= [x_1 * y_1 + x_1 * y_2 + \dots] + [x_2 * y_1 + x_2 * y_2 + \dots] + \dots \\ &= [x_1 * y_1 + x_2 * y_1 + \dots] + [x_1 * y_2 + x_2 * y_2 + \dots] + \dots \\ &= \sum_j [x_1 * y_j + x_2 * y_j + \dots] \\ &= \sum_j [\sum_i x_i * y_j] \end{aligned}$$

This reversal is a useful tool which we will apply in a minute.

3.1.3 Backwards

We also want to apply the definition of expectation to what we want to get in “B”, this will give us a direction.

$$\begin{aligned} &= a + bE(X) + cE(Y) \\ &= a + b \sum_i x_i f_X(x_i) + c \sum_j y_j f_Y(y_j) \end{aligned}$$

Comparing this result to what we had above, we now recognize that the “B” part will require isolating distributions in terms of just X and Y, moving b and c out of the summation. In equation (2) they are all jammed together and things are written in terms of joint distributions, but here we have just the distribution of X and Y alone (recall these are called the marginal distributions).

3.1.4 Aside on distributive property

It turns out when we want to get constants out of sums we use the distributive property of multiplication. For example, if a is a constant:

$$\sum_{i \in \{1,2,3\}} ax_i = ax_1 + ax_2 + ax_3 = a(x_1 + x_2 + x_3) = a \sum_{i \in \{1,2,3\}} x_i$$

Note that we usually do not list the items we sum over, as there may be an infinite number of them. Also recall that sums can be distributed:

$$\sum_i (x_i + y_i) = \sum x_i + \sum y_i$$

Step 2: Moving constants outside summations

Returning to our proof, we continue below in 6 steps. We first *distribute* the multiplication (2-3), then the summations (3-4), then bring out the constants (4-5). From (4-5) we also reverse the order of summation signs. This was done to isolate terms in the summations. We also use the distributive property described above with the constant c . We do so again on lines (5-6) with x_i and y_j , because we recognize that x_i are constant when it comes to summing over the j indexed objects, and y_j is a constant when summing over the i indexed objects. This gives us the second step of Goldberger's proof.

$$= \sum_i \sum_j (a + bx_i + cy_j) f(x_i, y_j) \quad (2)$$

$$= \sum_i \sum_j (af(x_i, y_j) + bx_i f(x_i, y_j) + cy_j f(x_i, y_j)) \quad (3)$$

$$= \sum_i \sum_j af(x_i, y_j) + \sum_i \sum_j bx_i f(x_i, y_j) + \sum_i \sum_j cy_j f(x_i, y_j) \quad (4)$$

$$= a \sum_i \sum_j f(x_i, y_j) + b \sum_i \sum_j x_i f(x_i, y_j) + c \sum_j \sum_i y_j f(x_i, y_j) \quad (5)$$

$$= a \sum_i \sum_j f(x_i, y_j) + b \sum_i x_i \left[\sum_j f(x_i, y_j) \right] + c \sum_j y_j \left[\sum_i f(x_i, y_j) \right] \quad (6)$$

Step 3: Using definitions of joint distributions

At this point, we have some things that look close to the definitions of $E(X)$ and $E(Y)$ used in the “B” part, but with joint distributions rather than marginal distributions. To move forward, we will need to use axioms of probability and properties of joint distributions.

3.1.5 Aside on joint distributions

Consider the following example of a joint distribution between two Bernoulli random variables X and Y.

	Y = 0	Y = 1	
X = 0	$f(X=0, Y=0)$	$f(X=0, Y=1)$	$f_1(X=0)$
X = 1	$f(X=1, Y=0)$	$f(X=1, Y=1)$	$f_1(X=1)$
	$f_2(Y=0)$	$f_2(Y=1)$	

Filling in example values, perhaps $f(x_i, y_j)$ is as follows

	Y = 0	Y = 1	
X = 0	0.01	0.05	0.06
X = 1	0.25	0.69	0.94
	0.26	0.74	

We can read this as “the probability of getting a 0 from both X and Y is .01”.

Note the sum of the joint distribution across the rows or columns produces a marginal distribution: $\sum_i f(x_i, y_j) = f_2(y_j)$ and $\sum_j f(x_i, y_j) = f_1(x_i)$. Finally, it is an axiom of probability that the sum of all probabilities must always be 1: $\sum_i \sum_j f(x_i, y_j) = 1$, as we can see by summing all of the entries on the interior of the table in the case of the joint distribution.

We apply these results to go from line 6 to line 7.

$$= a \sum_i \sum_j f(x_i, y_j) + b \sum_i x_i \sum_j f(x_i, y_j) + c \sum_j y_j \sum_i f(x_i, y_j) \quad (6)$$

$$= a * 1 + b \sum_i x_i f_1(x_i) + c \sum_j y_j f_2(y_j) \quad (7)$$

Step 4: Using definitions of expectation

The last step is to place back in the definition of the expectation operator, reversing step 1 above and applying our “backward” result from 2.1.3.

$$= a * 1 + b \sum_i x_i f_1(x_i) + c \sum_j y_j f_2(y_j) \quad (8)$$

$$= a + bE(X) + cE(Y) \quad (9)$$

Now we have “B” and we are then done with part one of the proof. Make sure you follow each step before continuing onward.

4 Part 2: Linearity of the variance

4.1 Proof Strategy

The second mathematical statement again has an “if A then B” structure. If Z is that function of random variables and constants, then we can write its variance as a sum of variances and covariances. Again, the proof strategy is a “direct proof”, starting from definitions of the A part and showing that B follows. However, here we use our second trick, and that is when working with variances it is useful to subtract the mean of our variables, as it simplifies algebra. We replace, do the algebra, then substitute it back in.

Step 1: Forward and backward definitions of V and C

Recall typical definition of variance is

$$V(X) = E[(X - E(X))^2]$$

The definition of the covariance, which is the “B” part, is:

$$C(X, Y) = E[(X - E(X))(Y - E(Y))]$$

The inside part have the form $X - E(X)$, which is a pain to keep writing. Here we will replace it with a new term, here $X^* = X - E(X)$. Note,

$X^* = X - E(X)$ can be rewritten $X^* + E(X) = X$. We demonstrate that trick here, replacing Z with Z^* :

$$\begin{aligned} V(Z) &= E[(Z - E(Z))^2] \\ &= E[Z^{*2}] \end{aligned}$$

Similarly,

$$\begin{aligned} C(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[X^*Y^*] \end{aligned}$$

This trick of using a symbol to represent more complicated objects is used throughout math and statistics.

Step 2: Using definition of Z

We can now use the definition of Z^* to rewrite in terms of X^* and Y^* . First we replace Z with its definition (10-11). We then use the results from part one on the linearity of the expectation operator to distribute the expectation across the summation (11-12). We then remove the constants. (12-13), as the expectation is linear (see above). Finally, we regroup terms (13-16), and replace $X - E(X)$ and $Y - E(Y)$ with X^* and Y^*

$$V(Z) = E[(Z - E(Z))^2] \tag{10}$$

$$= E[(a + bX + cY - E(a + bX + cY))^2] \tag{11}$$

$$= E[(a + bX + cY - E(a) - E(bX) - E(cY))^2] \tag{12}$$

$$= E[(a + bX + cY - a - bE(X) - cE(Y))^2] \tag{13}$$

$$= E[((a - a) + (bX - bE(X)) + (cY - cE(Y)))^2] \tag{14}$$

$$= E[b(X - E(X)) + c(Y - E(Y))]^2 \tag{15}$$

$$= E[(bX^* + cY^*)^2] \tag{16}$$

Note the a drops out. The variance of X and the variance of $X+5$ are the same.

Step 3: Expand the square

Studying variances requires taking squares (16-18). The algebra here shows up quite often. In (18-20) we redistribute the expectation operator, use the result from step one to pull out the constants, noting that the constants are now squared. Lastly we plug back in the $X - E(X)$ for X^* and we see the definition of $V()$ and $C()$ again.

$$= E[(bX^* + cY^*)^2] \quad (16)$$

$$= E[bX^* * bX^* + bX^*cY^* + cY^*bX^* + cY^* * cY^*] \quad (17)$$

$$= E[b^2X^{*2} + 2bcX^*Y^* + c^2Y^{*2}] \quad (18)$$

$$= E[b^2X^{*2}] + E[2bcX^*Y^*] + E[c^2Y^{*2}] \quad (19)$$

$$= b^2E[X^{*2}] + 2bcE[X^*Y^*] + c^2E[Y^{*2}] \quad (20)$$

$$= b^2E[(X - E(X))^2] + 2bcE[(X - E(X))(Y - E(Y))] + c^2E[(Y - E(Y))^2] \quad (21)$$

$$= b^2V(X) + 2bcC(X, Y) + c^2V(Y) \quad (22)$$

■

5 Discussion

Here we have unpacked a six line proof to a mere 11 pages and 22 steps. Showing all these steps would make textbooks very long, but it is generally assumed you worked through them at least once. In the long run you will become comfortable with some of these manipulations, so you will not need to reshew each step. You will apply the linearity of the expectation repeatedly throughout the class, including in the next example proof.

6 Example 2: T8

On page 48 of Goldberger we have the “Law of Iterated Expectations” or LIE, among the most important results in probability:

T8. LAW OF ITERATED EXPECTATIONS. The (marginal) expectation of $Z = h(X, Y)$ is the expectation of its conditional expectations:

$$E(Z) = E_X[E(Z|X)].$$

(Note: The symbol E_X , read as “the expectation over X ,” is the expectation taken in the marginal distribution of X . The subscript may be omitted if there is no risk of confusion.)

Proof.

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)f(x, y) dy dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)[g_2(y|x)f_1(x)] dy dx \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} h(x, y)g_2(y|x) dy \right] f_1(x) dx \\ &= \int_{-\infty}^{\infty} E(Z|x)f_1(x) dx. \blacksquare \end{aligned}$$

Note, here we have generalized how Z can depend on X and Y . Above, it was assumed that Z was a linear function of X and Y . Here it is an arbitrary function $h()$.

6.1 Aside on conditional distributions

Here we will assume that you are somewhat familiar with conditional distributions and conditional expectations. In this context, $E(X|Y)$ and $V(X|Y)$ are random variables.

Given a joint distribution $f(x, y)$ and marginals $f_1(x)$, $f_2(y)$, we can define two conditional distributions $g_1(x|y) = \frac{f(x,y)}{f_2(y)}$ and $g_2(y|x) = \frac{f(x,y)}{f_1(x)}$. A conditional expectation is written:

$$E(y|x) = \int_{-\infty}^{\infty} yg_2(y|x)dy$$

Returning to our two Bernoulli random variables X and Y,

	Y = 0	Y = 1	
X = 0	0.01	0.05	0.06
X = 1	0.25	0.69	0.94
	0.26	0.74	

In this case we can calculate the conditional expectation, for instance, $E(Y|x=0)$. Note that $g_2(y|x) = \frac{f(x,y)}{f_1(x)}$, so in this case $g_2(y|x=0) = \frac{f(x=0,y)}{f_1(x=0)}$ so,

$$g_2(y|x=0) = \begin{cases} \frac{.01}{.06} = \frac{1}{6} & \text{for } Y = 0 \\ \frac{.05}{.06} = \frac{5}{6} & \text{for } Y = 1 \end{cases}$$

$$\begin{aligned} E(Y|x=0) &= \sum_{y=0,1} yg_2(y|x=0) \\ &= 0 * g_2(y|x=0) + 1 * g_2(y|x=0) \\ &= 0 * \frac{1}{6} + 1 * \frac{5}{6} \\ &= \frac{5}{6} \end{aligned}$$

Similarly $E(Y|x=1)$. Here $g_2(y|x=1) = \frac{f(x=1,y)}{f_1(x=1)}$ so,

$$g_2(y|x=1) = \begin{cases} \frac{.25}{.94} & \text{for } Y = 0 \\ \frac{.69}{.94} & \text{for } Y = 1 \end{cases}$$

$$\begin{aligned}
E(Y|x=1) &= \sum_{y=0,1} yg_2(y|x=1) \\
&= 0 * g_2(y|x=1) + 1 * g_2(y|x=1) \\
&= 0 * \frac{.25}{.94} + 1 * \frac{.69}{.94} \\
&= \frac{69}{94}
\end{aligned}$$

6.2 Proof Strategy

Here the proof strategy is again to apply definitions following a similar strategy as above, isolating different parts of the summation. Here we have integrals, for clarity, I will label them with x and y and use square brackets to indicate order of operations. We know the goal is to show B, which has the form of $E[E(Z|x)]$, the idea is to work forward and backwards with definitions, and then hope we can see a connection between them.

6.3 Step 1: Using definition of conditional expectation

In (23) first write out the definition of $E[Z]$, following the steps from (1-2) above. We know we need to eliminate the joint distribution, as what we are after is an expectation with respect to the marginal distribution and an expectation in terms of a conditional distribution. In (23-24) then use the definition of conditional distribution, that is that given a joint distribution $f(x, y)$, a marginal distribution $f_1(x)$, we can define a conditional distribution (recall we label it g_2 to remember that we are talking about the original second argument y):

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)}$$

$$g_2(y|x)f_1(x) = f(x, y)$$

$$E(Z) = \int_{x=-\infty}^{\infty} \left[\int_{y=-\infty}^{\infty} h(x, y)f(x, y)dy \right] dx \quad (23)$$

$$= \int_{x=-\infty}^{\infty} \left[\int_{y=-\infty}^{\infty} h(x, y)[g_2(y|x)f_1(x)]dy \right] dx \quad (24)$$

6.4 Step 2: Moving constants outside integrations

We now notice that $f_1(x)$ is a constant with regards to y . We can then move it outside the integral sign (24-25).

$$= \int_{x=-\infty}^{\infty} \left[\int_{y=-\infty}^{\infty} h(x, y)[g_2(y|x)f_1(x)]dy \right] dx \quad (24)$$

$$= \int_{x=-\infty}^{\infty} f_1(x) \left[\int_{y=-\infty}^{\infty} h(x, y)[g_2(y|x)]dy \right] dx \quad (25)$$

6.5 Step 3: Applying Definitions

We immediately see that the thing in the square brackets is the definition of the conditional expectation function from page 47 of the Golberger text ($E(Z|x) = \int_{-\infty}^{\infty} h(x, y)g_2(y|x)dy$). The outer part is also an expectation (weights \times values), and we are done.

$$= \int_{x=-\infty}^{\infty} f_1(x) \left[\int_{y=-\infty}^{\infty} h(x, y)[g_2(y|x)]dy \right] dx \quad (25)$$

$$= \int_{x=-\infty}^{\infty} f_1(x) [E(Z|x)] dx \quad (26)$$

$$= E [E(Z|x)] \quad (27)$$

7 Discussion

In your homework you will prove similar claims using the techniques we have just discussed. The Law of Iterated Expectations is a very important result in its own right. It shows up in proofs when our independent variables, X , are stochastic. Here is an example of one of the most common LIE tricks.

Prop.: if $E(\epsilon|X) = 0$ then $Cov(\epsilon, f(X)) = 0$ for *any arbitrary function* f .

Proof: We will show that $E(\epsilon f(X)) = 0$, which is sufficient. Assume $E(\epsilon|X) = 0$. The Law of Iterated Expectations tells us the first equality. The second comes from the fact that $f(X)$ is a constant when conditioning on the information in X . The next lines follow from our first assumption.

$$E(\epsilon f(X)) = E[E(\epsilon f(X)|X)] \quad (28)$$

$$= E[f(X)E(\epsilon|X)] \quad (29)$$

$$= E[f(X) * 0] \quad (30)$$

$$= 0 \quad (31)$$

Finally LIE has practical use as well. Suppose there is new democracy that fails with probability p at each regular election. We are interested in how many elections it will go before failing. Call $F = A$ the event it fails in the first election, and $F = A^c$ the event it does not fail in the first election. Call C the time to failure.

$$E[C] = E[E[C|F]] = \sum_{F \in \{A, A^c\}} E[C|F]p(F) = E[C|A]p(A) + E[C|A^c]p(A^c)$$

By assumption, $E[C|A] = 1$, as this failure right away. Meanwhile, $E[C|A^c] = 1 + E[C]$, as once we make it past the first year, we aren't learning anything about future failure rates.

$$E[C] = 1 * p(A) + (1 + E[C])p(A^c) \quad (32)$$

$$E[C] = p + (1 + E[C])(1 - p) \quad (33)$$

$$E[C] = 1 + (1 - p)E[C] \quad (34)$$

$$E[C](1 - (1 - p)) = 1 \quad (35)$$

$$E[C] = \frac{1}{p} \quad (36)$$

If on each election there is a .2 chance of failure, the democracy will last 5 elections.