Expectation and Variance: Further Notes

1 Real-valued Random Variables

A *real-valued random variable* over a probability space, $\mathcal{S}$, is a function from $\mathcal{S}$ to the real numbers. We restrict attention to discrete spaces $\mathcal{S}$.

The *probability density function (pdf)* of a random variable, R, is the function $f_R : \mathbb{R} \to \mathbb{R}$ defined as:

$$f_R(r) = \sum_{\{s \mid R(s)=r\}} \Pr\{s\}.$$

The *cumulative density function* of R is the function $F_R : \mathbb{R} \to \mathbb{R}$ defined as:

$$F_R(r) = \sum_{\{s \mid R(s) \leq r\}} \Pr\{s\} = \sum_{\{r' \in \mathrm{range}(R) \mid r' \leq r\}} f_R(r').$$

When several random variables are mentioned in the same context, we tacitly assume they are defined on the same probability space.

2 Expectation

Definition. *The* expectation, $\mathrm{E}[R]$, *of a random variable* R, *is*

$$\mathrm{E}[R] = \sum_{s \in \mathcal{S}} R(s) \Pr\{s\},$$

providing this sum has a well-defined limit.

Absolute convergence of the sum on the right is necessary and sufficient for the expectation to have a well-defined *finite* value. Another way to say this is that $\mathrm{E}[R]$ is finite iff $\mathrm{E}[|R|]$ is finite.

A variable may also have a well-defined infinite or negative infinite value. For example, a random variable taking the value -2^n with probability 2^{-n} for all $n > 0$, has expectation $-\infty$. On the other hand, an random variable taking the values 2^n and -2^n, each with probability $2^{-n}/2$ for $n > 0$ does not have a well-defined expectation.

The next lemma gives an alternative formulation of expectation that many texts take as its definition. This formulation shows that the expectation of a random variable depends only on the pdf of the variable rather than on the behaviour of the variable on individual sample points.

Lemma 2.1.

$$\mathrm{E}[R] = \sum_{r \in \mathrm{range}(R)} r \Pr\{R = r\}.$$

The equality above holds in the strong sense that the expectation is well-defined iff the series on the right has a well-defined limit.

Proof. Let $[R = r]$ denote the event that $R = r$. We leave it to the reader to verify that if any of the sums in the following derivation are well-defined, then they all are:

$$\begin{aligned}
\sum_{r \in \text{range}(R)} r \cdot \Pr\{R = r\} &= \sum_r r \cdot \left[\sum_{s \in [R=r]} \Pr\{s\} \right] \\
&= \sum_r \sum_{s \in [R=r]} [r \cdot \Pr\{s\}] \\
&= \sum_r \sum_{s \in [R=r]} [R(s) \cdot \Pr\{s\}] \\
&= \sum_{s \in \mathcal{S}} R(s) \cdot \Pr\{s\}
\end{aligned}$$

The last equality follows from the fact that the events $[R = r]$ for $r \in \text{range}(R)$ are a partition of the sample space $\mathcal{S}$. ■

It follows easily from the definition that

$$\mathrm{E}[aR + b] = a\mathrm{E}[R] + b$$

for any $a, b \in \mathbb{R}$.

Theorem 2.2. *[Linearity of Expectation] Let R_0, R_1, ... , be random variables such that*

$$\sum_{i=0}^{\infty} \mathrm{E}[|R_i|]$$

converges. Then

$$\mathrm{E}\left[\sum_{i=0}^{\infty} R_i\right] = \sum_{i=0}^{\infty} \mathrm{E}[R_i].$$

Proof. We leave it to the reader to verify that, under the given convergence hypothesis, all the sums in the following derivation are absolutely convergent, which justifies rearranging them as follows::

$$
\begin{aligned}
\sum_{i=0}^{\infty} \mathrm{E}[R_i] &= \sum_{i=0}^{\infty} \sum_{s \in \mathcal{S}} [R_i(s) \cdot \Pr\{s\}] \\
&= \sum_{s \in \mathcal{S}} \sum_{i=0}^{\infty} [R_i(s) \cdot \Pr\{s\}] \\
&= \sum_{s \in \mathcal{S}} \left[\sum_{i=0}^{\infty} R_i(s) \right] \cdot \Pr\{s\} \\
&= \sum_{s \in \mathcal{S}} \left[\left[\sum_{i=0}^{\infty} R_i \right] (s) \right] \cdot \Pr\{s\} \\
&= \mathrm{E}\left[\sum_{i=0}^{\infty} R_i \right].
\end{aligned}
$$

■

Corollary 2.3. *[Finite Linearity of Expectation] Let $R_0, R_1, \ldots, R_n$ be random variables with finite expectations. Then*

$$
\mathrm{E}\left[\sum_{i=0}^{n} R_i \right] = \sum_{i=0}^{n} \mathrm{E}[R_i].
$$

Proof. Since $\mathrm{E}[R_i]$ is finite, so is $\mathrm{E}[|R_i|]$, and therefore so is their sum for $0 \le i \le n$. Hence the convergence hypothesis of Theorem 2.2 is trivially satisfied. ■

Exercise: Show that linearity of expectation fails for the sum of two variables, one with expectation $+\infty$ and the other with $-\infty$.

2.1 A Paradox

One of the simplest casino bets is on "red" or "black" at the roulette table. In each play at roulette, a small ball is set spinning around a roulette wheel until it lands in a red, black, or green colored slot. The payoff for a bet on red or black matches the bet; for example, if you bet \$10 on red and the ball lands in a red slot, you get back your original \$10 bet plus another matching \$10.

In the US, a roulette wheel has two green slots among 18 black and 18 red slots, so the probability of red is $p = 18/38 \approx 0.473$. In Europe, where roulette wheels have only one green slot, the odds for red are a little better —that is, $p = 18/37 \approx 0.486$—but still less than even. To make the game fair, we might agree to ignore green, so that $p = 1/2$.

There is a notorious gambling strategy which seems to guarantee a profit at roulette: bet \$10 on red, and keep doubling the bet until a red comes up. This strategy implies that a player will leave the game as a net winner of \$10 as soon as the red first appears. Of course the player may need an awfully large bankroll to avoid going bankrupt before red shows up—but we know that the mean time until a red occurs is $1/p$, so it seems possible that a moderate bankroll might actually work out. (In this setting, a "win" on red corresponds to a "failure" in a mean-time-to-failure situation, cf. Lecture 22 Notes.) In any case, we won't worry about bankruptcy and will assume we can keep doubling our bets indefinitely until a red comes up.

Suppose we have the good fortune to gamble against a fair roulette wheel. In this case, our expected win on any spin is zero, since at the ith spin we are equally likely to win or lose $10 \cdot 2^i$ dollars. So our expected win after any finite number of spins remains zero, and therefore our expected win using this gambling strategy is zero. This is just what we should have anticipated in a fair game.

But wait a minute. As long as there is a fixed, positive probability of red appearing on each spin of the wheel—even if the wheel is unfair—it's *certain* that red will eventually come up. So with probability one, we leave the casino having won \$10, and our expected dollar win is obviously \$10, not zero!

Something's wrong here. What?

3 Variance

For random variable R with mean μ,

$$\text{Var}[R] = \text{E}\left[(R-\mu)^2\right].$$

It follows from linearity of expectation, that

$$\text{Var}[R] = \text{E}\left[R^2\right] - \text{E}^2[R],$$

as the reader can verify.

Lemma 3.1. *For any $a, b \in \mathbb{R}$,*

$$\text{Var}[aR+b] = a^2\text{Var}[R].$$

Proof. Let $\mu = \text{E}[R]$, so $\text{E}[aR+b] = a\mu + b$. Then

$$\begin{aligned}
\text{Var}[aR+b] &= \text{E}\left[((aR+b)-(a\mu+b))^2\right] \\
&= \text{E}\left[(aR-a\mu)^2\right] \\
&= a^2\text{E}\left[(R-\mu)^2\right] \\
&= a^2\text{Var}[R].
\end{aligned}$$

■

Lemma 3.2. *If X, Y are independent,*

$$\begin{aligned}
\text{E}[XY] &= \text{E}[X]\,\text{E}[Y], \\
\text{Var}[X+Y] &= \text{Var}[X] + \text{Var}[Y].
\end{aligned}$$

Proof. See F97 Lecture 24 Notes. ■

The *standard deviation*, sometimes just called the *deviation*, of a random variable is the square root of its variance. Sometimes it's more convenient to work with the deviation than with the variance. The deviation is often denoted by the symbol σ and the variance by σ^2.

4 Markov and Chebyshev Bounds

4.1 Review

The theorems of Markov and Chebyshev give upper bounds on the probability that a random variable differs by a given amount from its mean. The Markov bound holds solely under the hypothesis that the variable is nonnegative and the expectation exists, *i.e.*, is finite. The Chebyshev bound holds solely under the hypothesis that the variance exists. (Remember that the variance can exist only if the expectation does.)

Theorem 4.1. *[Markov's Bound] If R is a non-negative random variable with finite expectation, then for all $x > 0$,*

$$\Pr\{R \geq x\} \leq \frac{\mathrm{E}[R]}{x}$$

Setting $x = c \cdot \mathrm{E}[R]$ allows Markov's Theorem to be expressed in an alternate form:

Corollary 4.2. *If R is a non-negative random variable with finite expectation, then for all $c > 0$*

$$\Pr\{R \geq c \cdot \mathrm{E}[R]\} \leq \frac{1}{c}$$

Theorem 4.3. *[Chebyshev's Bound] Let R be a random variable with finite variance, and let x be a positive real number. Then*

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} \leq \frac{\mathrm{Var}[R]}{x^2}.$$

4.2 The Weak Law of Large Numbers

The intuition behind the definition of expectation is that the average of a large number of random samples of a variable will be close to the expectation of the variable. This will happen even if the random variable never actually takes a value close to its expectation. Using Chebyshev's Theorem and the facts about variance and expectation, we are finally in a position to be more precise about this intuitive idea.

For example, suppose we want to estimate the fraction of the U.S. voting population who favor Al Gore over all other year 2000 presidential hopefuls. Let p be this unknown fraction. Let's suppose we have some random process—say throwing darts at voter registration lists—which will yield each voter with equal probability. Now we can define a Bernoulli variable, G, by the rule that $G = 1$ if a random voter most prefers Gore, and $G = 0$ otherwise. In this case, $G = G^2$, so

$$\mathrm{E}\left[G^2\right] = \mathrm{E}[G] = \Pr\{G = 1\} = p,$$

and

$$\mathrm{Var}[G] = \mathrm{E}\left[G^2\right] - \mathrm{E}^2[G] = p - p^2 = p(1-p).$$

To estimate p, we take a large number, n, of sample voters and count the fraction who favor Gore. We can describe this estimation as taking independent Bernoulli variables $G_1, G_2, \ldots, G_n$, each with the same expectation as G, computing their sum

$$S_n = \sum_{i=1}^{n} G_i,$$

and then using the average S_n/n as our estimate of p.

This estimate S_n/n is a random variable with two critical properties:

$$\begin{aligned} \mathrm{E}\left[\frac{S_n}{n}\right] &= \mathrm{E}[G], \\ \mathrm{Var}\left[\frac{S_n}{n}\right] &= \frac{\mathrm{Var}[G]}{n}. \end{aligned}$$

To prove this, note that by linearity of expectation

$$\mathrm{E}\left[\frac{S_n}{n}\right] = \frac{\mathrm{E}\left[\sum_{i=1}^{n} G_i\right]}{n} = \frac{\sum_{i=1}^{n} \mathrm{E}[G_i]}{n} = \frac{n\mathrm{E}[G]}{n} = \mathrm{E}[G].$$

Also, by Lemma 3.2, since the G_i's are independent, the variances will add, so

$$\begin{aligned} \mathrm{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \mathrm{Var}[S_n] && \text{by Lemma 3.1,} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}[G_i] \\ &= \left(\frac{1}{n}\right)^2 n\mathrm{Var}[G] \\ &= \frac{\mathrm{Var}[G]}{n}. \end{aligned}$$

Now Chebyshev's Bound tells us that

Theorem 4.4. *Let $S_n = \sum_{i=1}^{n} G_i$ where $G_1, \ldots, G_n$ are mutually independent variables with the same mean, μ, and deviation, σ. Then*

$$\Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq x\right\} \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

This theorem finally provides a precise statement about how the average of independent samples of a random variable approaches the mean. It generalizes to many cases when S_n is the sum of independent variables whose mean and deviation are not necessarily all the same, though we shall not develop such generalizations here.

A simple consequence of Theorem 4.4 is The Weak Law of Large Numbers. Let's first rename x to be ϵ—the traditional symbol for a small positive quantity. Then, with ϵ fixed, we can always choose n large enough to ensure that, with probability as close to one as desired, the average of n samples is within ϵ of the actual average. We can state this as a limit theorem:

Theorem 4.5. *[Weak Law of Large Numbers] Let $S_n = \sum_{i=1}^{n} G_i$ where $G_1, \ldots, G_n$ are mutually independent variables with the same mean and variance. For any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \mathrm{E}\left[\frac{S_n}{n}\right]\right| \geq \epsilon\right\} = 0,$$

or equivalently,

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \mathrm{E}\left[\frac{S_n}{n}\right]\right| < \epsilon\right\} = 1,$$

This Weak Law of Large Numbers serves as a helpful reminder of how to formulate estimation of the mean by average sampling. It is the most elementary of many such theorems. However, the Weak Law as it stands has no actual applications—first because it does not say anything about the *rate* at which the limits are approached (rate information is essential in applications), and second, because it provides no information about the way the value of the average may be expected to *oscillate* in the course of an experiment. There is a Strong Law of Large Numbers which deals with the oscillations. Such oscillations may not be important in our example of polling about Gore's popularity, but they are critical in gambling situations, where large oscillations can bankrupt a player, even though the player's average winnings are assured if he survives long enough. The problem here with the long run view, as the famous economist Keynes is alleged to have remarked, is that "In the long run, we are all dead."

4.3 Size of a Poll

Chebyshev's Theorem 4.4 allows us to calculate how many voters to poll if we want to get a reliable estimate of voters' preference for Gore.

Suppose, in particular, we want to know within 0.02 what fraction of the voters favor Gore. So we let $x = 1/50$ and conclude from Theorem 4.4 that we can, by choosing n large enough, reduce the probability that our estimate is off by more than ± 0.02 to as close to zero as we please.

For example, suppose further that we want to be within 0.02 of p with probability 0.95—ninety-five per cent "confidence level" is a standard used in many statistical applications. Then we choose n so that $\mathrm{Var}\left[G\right]/nx^2 \leq 1 - 0.95$. That is, we want

$$n \geq 20\mathrm{Var}\left[G\right] 50^2 = 50,000p(1-p).$$

Solving for the sample size n in terms of the unknown p that we are trying to estimate in the first place may not seem to be making progress. But it's easy to see that the maximum value of $p(1-p)$ in the interval $0 \leq p \leq 1$ occurs at $p = 1/2$, so we conclude that if we sample

$$n \geq 50,000(1 - 1/2)1/2 = 12,500$$

voters, we can say that 95% of the time, our estimate $S_{12,500}/12,500$ will be within 0.02 of the fraction of voters who favor Gore.

Note that this bound on poll size holds regardless of how large the total voting population may be—whether we are trying to determine the preferences of a few tens of thousands of Cambridge voters, or of the tens of millions of all American voters, the same poll size is adequate.

Now suppose a pollster dutifully checks with 12,500 randomly chosen voters and finds that 6,300 prefer Gore. It's tempting, but sloppy, to say that this means "With probability 0.95, the fraction of voters who prefer Gore is 0.504 ± 0.02." What's objectionable about this statement is that it talks about the probability of a real world fact, namely the actual value of the fraction p. But p is what it is, and it simply makes no sense to talk about the probability that it is something else. For example, suppose p is actually 0.53; then it's nonsense to ask about the probability that it is within 0.02 of 0.504—it simply isn't.

A more careful summary of what we have accomplished would be that we have described a probabilistic procedure for estimating the actual value of the fraction p, and the probability that *our estimation procedure* yields a value within 0.02 is 0.95. This is a bit of a mouthful, so special phrasing closer to the sloppy language is commonly used. The pollster would describe his conclusion by saying that "At the 95% *confidence level*, the fraction of voters who prefer Gore is 0.504 ± 0.02."

By the way, polling 12,500 voters is wildly excessive. We derived this bound on poll size solely by applying Chebyshev's Theorem to value of the variance of S_n/n. But in fact we know the exact distribution of S_n/n, namely, it has a binomial distribution with parameters n, p. By a more detailed calculation of probabilities of deviation from the mean specifically for the binomial distribution (cf. Lecture 21 Notes), we can show that the poll size could be more than an order of magnitude smaller than 12,500.