Chapter 21

Deviation from the Mean

21.1 Why the Mean?

In the previous chapter we took it for granted that expectation is important, and we developed a bunch of techniques for calculating expected (mean) values. But why should we care about the mean? After all, a random variable may never take a value anywhere near its expected value.

The most important reason to care about the mean value comes from its connection to estimation by sampling. For example, suppose we want to estimate the average age, income, family size, or other measure of a population. To do this, we determine a random process for selecting people —say throwing darts at census lists. This process makes the selected person's age, income, and so on into a random variable whose *mean* equals the *actual average* age or income of the population. So we can select a random sample of people and calculate the average of people in the sample to estimate the true average in the whole population. Many fundamental results of probability theory explain exactly how the reliability of such estimates improves as the sample size increases, and in this chapter we'll examine a few such results.

In particular, when we make an estimate by repeated sampling, we need to know how much confidence we should have that our estimate is OK. Technically, this reduces to finding the probability that an estimate *deviates* a lot from its expected value. This topic of *deviation from the mean* is the focus of this final chapter.

21.2 Markov's Theorem

Markov's theorem is an easy result that gives a generally rough estimate of the probability that a random variable takes a value *much larger* than its mean.

The idea behind Markov's Theorem can be explained with a simple example of *intelligence quotient*, IQ. This quantity was devised so that the average IQ measurement would be 100. Now from this fact alone we can conclude that at most $1/3$ the population can have an IQ of 300 or more, because if more than a third had an IQ of 300, then the average would have to be *more* than $(1/3)300 = 100$, contradicting the fact that the average is 100. So the probability that a randomly chosen person has an IQ of 300 or more is at most 1/3. Of course this is not a very strong conclusion; in fact no IQ of over 300 has ever been recorded. But by the same logic, we can also conclude that at most 2/3 of the population can have an IQ of 150 or more. IQ's of over 150 have certainly been recorded, though again, a much smaller fraction than 2/3 of the population actually has an IQ that high.

But although these conclusions about IQ are weak, they are actually the strongest general conclusions that can be reached about a random variable using *only* the fact that it is nonnegative and its mean is 100. For example, if we choose a random variable equal to 300 with probability 1/3, and 0 with probability 2/3, then its mean is 100, and the probability of a value of 300 or more really is 1/3. So we can't hope to get a better upper bound based solely on this limited amount of information.

Theorem 21.2.1 (Markov's Theorem)**.** *If R is a nonnegative random variable, then for all $x > 0$*

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

Proof. For any $x > 0$

$$\begin{aligned}
\mathrm{E}[R] &::= \sum_{y \in \mathrm{range}(R)} y \Pr\{R = y\} \\
&\geq \sum_{\substack{y \geq x,\\ y \in \mathrm{range}(R)}} y \Pr\{R = y\} && \text{(because } R \geq 0\text{)} \\
&\geq \sum_{\substack{y \geq x,\\ y \in \mathrm{range}(R)}} x \Pr\{R = y\} \\
&= x \sum_{\substack{y \geq x,\\ y \in \mathrm{range}(R)}} \Pr\{R = y\} \\
&= x \Pr\{R \geq x\}. && (21.1)
\end{aligned}$$

Dividing the first and last expression (21.1) by x gives the desired result. ■

Our focus is deviation from the mean, so it's useful to rephrase Markov's Theorem this way:

Corollary 21.2.2. *If R is a nonnegative random variable, then for all $c \geq 1$*

$$\Pr\{R \geq c \cdot \mathrm{E}[R]\} \leq \frac{1}{c}. \tag{21.2}$$

This Corollary follows immediately from Markov's Theorem(21.2.1) by letting x be $c \cdot \mathrm{E}[R]$.

21.2.1 Applying Markov's Theorem

Let's consider the Hat-Check problem again. Now we ask what the probability is that x or more men get the right hat, this is, what the value of $\Pr\{G \geq x\}$ is.

We can compute an upper bound with Markov's Theorem. Since we know $\mathrm{E}[G] = 1$, Markov's Theorem implies

$$\Pr\{G \geq x\} \leq \frac{\mathrm{E}[G]}{x} = \frac{1}{x}.$$

For example, there is no better than a 20% chance that 5 men get the right hat, regardless of the number of people at the dinner party.

The Chinese Appetizer problem is similar to the Hat-Check problem. In this case, n people are eating appetizers arranged on a circular, rotating Chinese banquet tray. Someone then spins the tray so that each person receives a random appetizer. What is the probability that everyone gets the same appetizer as before?

There are n equally likely orientations for the tray after it stops spinning. Everyone gets the right appetizer in just one of these n orientations. Therefore, the correct answer is $1/n$.

But what probability do we get from Markov's Theorem? Let the random variable, R, be the number of people that get the right appetizer. Then of course $\mathrm{E}[R] = 1$ (right?), so applying Markov's Theorem, we find:

$$\Pr\{R \geq n\} \leq \frac{\mathrm{E}[R]}{n} = \frac{1}{n}.$$

So for the Chinese appetizer problem, Markov's Theorem is tight!

On the other hand, Markov's Theorem gives the same $1/n$ bound for the probability everyone gets their hat in the Hat-Check problem in the case that all permutations are equally likely. But the probability of this event is $1/(n!)$. So for this case, Markov's Theorem gives a probability bound that is way off.

21.2.2 Markov's Theorem for Bounded Variables

Suppose we learn that the average IQ among MIT students is 150 (which is not true, by the way). What can we say about the probability that an MIT student has an IQ of more than 200? Markov's theorem immediately tells us that no more than $150/200$ or $3/4$ of the students can have such a high IQ. Here we simply applied Markov's Theorem to the random variable, R, equal to the IQ of a random MIT student to conclude:

$$\Pr\{R > 200\} \leq \frac{\mathrm{E}[R]}{200} = \frac{150}{200} = \frac{3}{4}.$$

But let's observe an additional fact (which may be true): no MIT student has an IQ less than 100. This means that if we let $T ::= R - 100$, then T is nonnegative and $\mathrm{E}[T] = 50$, so we can apply Markov's Theorem to T and conclude:

$$\Pr\{R > 200\} = \Pr\{T > 100\} \leq \frac{\mathrm{E}[T]}{100} = \frac{50}{100} = \frac{1}{2}.$$

So only half, not 3/4, of the students can be as amazing as they think they are. A bit of a relief!

More generally, we can get better bounds applying Markov's Theorem to $R - l$ instead of R for any lower bound $l > 0$ on R.

Similarly, if we have any upper bound, u, on a random variable, S, then $u - S$ will be a nonnegative random variable, and applying Markov's Theorem to $u - S$ will allow us to bound the probability that S is much *less* than its expectation.

21.2.3 Problems

Class Problems

Problem 21.1.
A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers the normal body temperature of a cow, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, **but no lower**, were actually found in the herd.

(a) Prove that at most 3/4 of the cows could have survived.

Hint: Let T be the temperature of a random cow. Make use of Markov's bound.

(b) Suppose there are 400 cows in the herd. Show that the bound of part (a) is best possible by giving an example set of temperatures for the cows so that the average herd temperature is 85, and with probability 3/4, a randomly chosen cow will have a high enough temperature to survive.

21.3 Chebyshev's Theorem

There's a really good trick for getting more mileage out of Markov's Theorem: instead of applying it to the variable, R, apply it to some function of R. One useful choice of functions to use turns out to be taking a power of $|R|$.

In particular, since $|R|^\alpha$ is nonnegative, Markov's inequality also applies to the event $[|R|^\alpha \geq x^\alpha]$. But this event is equivalent to the event $[|R| \geq x]$, so we have:

Lemma 21.3.1. *For any random variable R, $\alpha \in \mathbb{R}^+$, and $x > 0$,*

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

Rephrasing (21.3.1) in terms of the random variable, $|R - \mathrm{E}[R]|$, that measures R's deviation from its mean, we get

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} \leq \frac{\mathrm{E}[(R - \mathrm{E}[R])^\alpha]}{x^\alpha}. \tag{21.3}$$

The case when $\alpha = 2$ is turns out to be so important that numerator of the right hand side of (21.3) has been given a name:

Definition 21.3.2. The *variance*, $\text{Var}[R]$, of a random variable, R, is:

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

The restatement of (21.3) for $\alpha = 2$ is known as *Chebyshev's Theorem*.

Theorem 21.3.3 (Chebyshev)**.** *Let R be a random variable and $x \in \mathbb{R}^+$. Then*

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

The expression $\text{E}\left[(R - \text{E}[R])^2\right]$ for variance is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression, $R - \text{E}[R]$, is precisely the deviation of R above its mean. Squaring this, we obtain, $(R - \text{E}[R])^2$. This is a random variable that is near 0 when R is close to the mean and is a large positive number when R deviates far above or below the mean. So if R is always close to the mean, then the variance will be small. If R is often far from the mean, then the variance will be large.

21.3.1 Variance in Two Gambling Games

The relevance of variance is apparent when we compare the following two gambling games.

Game A: We win \$2 with probability $2/3$ and lose \$1 with probability $1/3$.

Game B: We win \$1002 with probability $2/3$ and lose \$2001 with probability $1/3$.

Which game is better financially? We have the same probability, 2/3, of winning each game, but that does not tell the whole story. What about the expected return for each game? Let random variables A and B be the payoffs for the two games. For example, A is 2 with probability 2/3 and -1 with probability 1/3. We can compute the expected payoff for each game as follows:

$$\text{E}[A] = 2 \cdot \frac{2}{3} + (-1) \cdot \frac{1}{3} = 1,$$
$$\text{E}[B] = 1002 \cdot \frac{2}{3} + (-2001) \cdot \frac{1}{3} = 1.$$

The expected payoff is the same for both games, but they are obviously very different! This difference is not apparent in their expected value, but is captured by variance. We can compute the $\text{Var}[A]$ by working "from the inside out" as follows:

$$\begin{aligned}
A - \text{E}[A] &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ -2 & \text{with probability } \frac{1}{3} \end{cases} \\
(A - \text{E}[A])^2 &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ 4 & \text{with probability } \frac{1}{3} \end{cases} \\
\text{E}\left[(A - \text{E}[A])^2\right] &= 1 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3} \\
\text{Var}[A] &= 2.
\end{aligned}$$

Similarly, we have for Var $[B]$:

$$\begin{aligned}
B - \mathrm{E}[B] &= \begin{cases} 1001 & \text{with probability } \frac{2}{3} \\ -2002 & \text{with probability } \frac{1}{3} \end{cases} \\
(B - \mathrm{E}[B])^2 &= \begin{cases} 1{,}002{,}001 & \text{with probability } \frac{2}{3} \\ 4{,}008{,}004 & \text{with probability } \frac{1}{3} \end{cases} \\
\mathrm{E}\left[(B - \mathrm{E}[B])^2\right] &= 1{,}002{,}001 \cdot \frac{2}{3} + 4{,}008{,}004 \cdot \frac{1}{3} \\
\mathrm{Var}[B] &= 2{,}004{,}002.
\end{aligned}$$

The variance of Game A is 2 and the variance of Game B is more than two million! Intuitively, this means that the payoff in Game A is usually close to the expected value of \$1, but the payoff in Game B can deviate very far from this expected value.

High variance is often associated with high risk. For example, in ten rounds of Game A, we expect to make \$10, but could conceivably lose \$10 instead. On the other hand, in ten rounds of game B, we also expect to make \$10, but could actually lose more than \$20,000!

21.3.2 Standard Deviation

Because of its definition in terms of the square of a random variable, the variance of a random variable may be very far from a typical deviation from the mean. For example, in Game B above, the deviation from the mean is 1001 in one outcome and -2002 in the other. But the variance is a whopping 2,004,002. From a dimensional analysis viewpoint, the "units" of variance are wrong: if the random variable is in dollars, then the expectation is also in dollars, but the variance is in square dollars. For this reason, people often describe random variables using standard deviation instead of variance.

Definition 21.3.4. The *standard deviation*, σ_R, of a random variable, R, is the square root of the variance:

$$\sigma_R ::= \sqrt{\mathrm{Var}[R]} = \sqrt{\mathrm{E}[(R - \mathrm{E}[R])^2]}.$$

So the standard deviation is the square root of the mean of the square of the deviation, or the *root mean square* for short. It has the same units —dollars in our example —as the original random variable and as the mean. Intuitively, it measures the average deviation from the mean, since we can think of the square root on the outside as canceling the square on the inside.

Example 21.3.5. The standard deviation of the payoff in Game B is:

$$\sigma_B = \sqrt{\mathrm{Var}[B]} = \sqrt{2{,}004{,}002} \approx 1416.$$

The random variable B actually deviates from the mean by either positive 1001 or negative 2002; therefore, the standard deviation of 1416 describes this situation reasonably well.

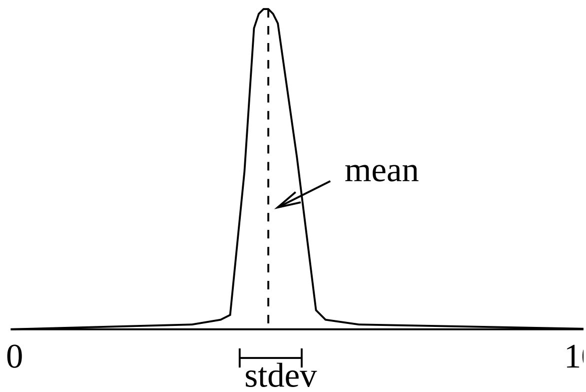

Figure 21.1: The standard deviation of a distribution indicates how wide the "main part" of it is.

Intuitively, the standard deviation measures the "width" of the "main part" of the distribution graph, as illustrated in Figure 21.1.

It's useful to rephrase Chebyshev's Theorem in terms of standard deviation.

Corollary 21.3.6. *Let R be a random variable, and let c be a positive real number.*

$$\Pr\{|R - \mathrm{E}[R]| \geq c\sigma_R\} \leq \frac{1}{c^2}.$$

Here we see explicitly how the "likely" values of R are clustered in an $O(\sigma_R)$-sized region around $\mathrm{E}[R]$, confirming that the standard deviation measures how spread out the distribution of R is around its mean.

Proof. Substituting $x = c\sigma_R$ in Chebyshev's Theorem gives:

$$\Pr\{|R - \mathrm{E}[R]| \geq c\sigma_R\} \leq \frac{\mathrm{Var}[R]}{(c\sigma_R)^2} = \frac{\sigma_R^2}{(c\sigma_R)^2} = \frac{1}{c^2}.$$

■

The IQ Example

Suppose that, in addition to the national average IQ being 100, we also know the standard deviation of IQ's is 10. How rare is an IQ of 300 or more?

Let the random variable, R, be the IQ of a random person. So we are supposing that $\mathrm{E}[R] = 100$, $\sigma_R = 10$, and R is nonnegative. We want to compute $\Pr\{R \geq 300\}$.

We have already seen that Markov's Theorem 21.2.1 gives a coarse bound, namely,

$$\Pr\{R \geq 300\} \leq \frac{1}{3}.$$

Now we apply Chebyshev's Theorem to the same problem:

$$\Pr\{R \geq 300\} = \Pr\{|R - 100| \geq 200\} \leq \frac{\text{Var}[R]}{200^2} = \frac{10^2}{200^2} = \frac{1}{400}.$$

So Chebyshev's Theorem implies that at most one person in four hundred has an IQ of 300 or more. We have gotten a much tighter bound using the additional information, namely the variance of R, than we could get knowing only the expectation.

21.4 Properties of Variance

The definition of variance of R as $\text{E}\left[(R - \text{E}[R])^2\right]$ may seem rather arbitrary. A direct measure of average deviation would be $\text{E}[\,|R - \text{E}[R]|\,]$. But the direct measure doesn't have the many useful properties that variance has, which is what this section is about.

21.4.1 A Formula for Variance

Applying linearity of expectation to the formula for variance yields a convenient alternative formula.

Lemma 21.4.1.

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

for any random variable, R.

Here we use the notation $\text{E}^2[R]$ as shorthand for $(\text{E}[R])^2$.

Proof. Let $\mu = \text{E}[R]$. Then

$$\begin{aligned}
\text{Var}[R] &= \text{E}\left[(R - \text{E}[R])^2\right] && \text{(Def 21.3.2 of variance)}\\
&= \text{E}\left[(R - \mu)^2\right] && \text{(def of } \mu)\\
&= \text{E}\left[R^2 - 2\mu R + \mu^2\right] && \\
&= \text{E}\left[R^2\right] - 2\mu\,\text{E}[R] + \mu^2 && \text{(linearity of expectation)}\\
&= \text{E}\left[R^2\right] - 2\mu^2 + \mu^2 && \text{(def of } \mu)\\
&= \text{E}\left[R^2\right] - \mu^2 && \\
&= \text{E}\left[R^2\right] - \text{E}^2[R]. && \text{(def of } \mu)
\end{aligned}$$

■

For example, if B is a Bernoulli variable where $p ::= \Pr\{B = 1\}$, then

Lemma 21.4.2.

$$\text{Var}[B] = p - p^2 = p(1 - p). \tag{21.4}$$

Proof. By Lemma 20.3.3, $\text{E}[B] = p$. But since B only takes values 0 and 1, $B^2 = B$. So Lemma 21.4.2 follows immediately from Lemma 21.4.1. ■

21.4.2 Variance of Time to Failure

According to section 20.3.3, the mean time to failure is $1/p$ for a process that fails during any given hour with probability p. What about the variance? That is, let C be the hour of the first failure, so $\Pr\{C = i\} = (1-p)^{i-1}p$. We'd like to find a formula for $\text{Var}[C]$.

By Lemma 21.4.1,

$$\text{Var}[C] = \text{E}\left[C^2\right] - (1/p)^2 \tag{21.5}$$

so all we need is a formula for $\text{E}\left[C^2\right]$:

$$\begin{aligned}\text{E}\left[C^2\right] &::= \sum_{i\geq 1} i^2(1-p)^{i-1}p \\ &= p\sum_{i\geq 1} i^2 x^{i-1} \qquad \text{(where } x = 1-p\text{).}\end{aligned} \tag{21.6}$$

But (17.2) gives the generating function $x(1+x)/(1-x)^3$ for the nonnegative integer squares, and this implies that the generating function for the sum in (21.6) is $(1+x)/(1-x)^3$. So,

$$\begin{aligned}\text{E}\left[C^2\right] &= p\,\frac{(1+x)}{(1-x)^3} \qquad \text{(where } x = 1-p\text{)} \\ &= p\,\frac{2+p}{p^3} \\ &= \frac{1-p}{p^2} + \frac{1}{p^2},\end{aligned} \tag{21.7}$$

Combining (21.5) and (21.7) gives a simple answer:

$$\text{Var}[C] = \frac{1-p}{p^2}\,. \tag{21.8}$$

It's great to be able apply generating function expertise to knock off equation (21.8) mechanically just from the definition of variance, but there's a more elementary, and memorable, alternative. In section 20.3.3 we used conditional expectation to find the mean time to failure, and a similar approach works for the variance. Namely, the expected value of C^2 is the probability, p, of failure in the first hour times 1^2, plus $(1-p)$ times the expected value of $(C+1)^2$. So

$$\begin{aligned}\text{E}\left[C^2\right] &= p\cdot 1^2 + (1-p)\,\text{E}\left[(C+1)^2\right] \\ &= p + (1-p)\left(\text{E}\left[C^2\right] + \frac{2}{p} + 1\right),\end{aligned}$$

which directly simplifies to (21.7).

21.4.3 Dealing with Constants

It helps to know how to calculate the variance of $aR + b$:

Theorem 21.4.3. *Let R be a random variable, and a a constant. Then*

$$\text{Var}[aR] = a^2 \text{Var}[R]. \tag{21.9}$$

Proof. Beginning with the definition of variance and repeatedly applying linearity of expectation, we have:

$$\begin{aligned}
\text{Var}[aR] &::= \text{E}\left[(aR - \text{E}[aR])^2\right] \\
&= \text{E}\left[(aR)^2 - 2aR\,\text{E}[aR] + \text{E}^2[aR]\right] \\
&= \text{E}\left[(aR)^2\right] - \text{E}[2aR\,\text{E}[aR]] + \text{E}^2[aR] \\
&= a^2\,\text{E}\left[R^2\right] - 2\,\text{E}[aR]\,\text{E}[aR] + \text{E}^2[aR] \\
&= a^2\,\text{E}\left[R^2\right] - a^2\text{E}^2[R] \\
&= a^2\left(\text{E}\left[R^2\right] - \text{E}^2[R]\right) \\
&= a^2\,\text{Var}[R] && \text{(by Lemma 21.4.1)}
\end{aligned}$$

■

It's even simpler to prove that adding a constant does not change the variance, as the reader can verify:

Theorem 21.4.4. *Let R be a random variable, and b a constant. Then*

$$\text{Var}[R + b] = \text{Var}[R]. \tag{21.10}$$

Recalling that the standard deviation is the square root of variance, this implies that the standard deviation of $aR + b$ is simply $|a|$ times the standard deviation of R:

Corollary 21.4.5.

$$\sigma_{aR+b} = |a|\,\sigma_R.$$

21.4.4 Variance of a Sum

In general, the variance of a sum is not equal to the sum of the variances, but variances do add for *independent* variables. In fact, *mutual* independence is not necessary: *pairwise* independence will do. This is useful to know because there are some important situations involving variables that are pairwise independent but not mutually independent.

Theorem 21.4.6. *If R_1 and R_2 are independent random variables, then*

$$\text{Var}[R_1 + R_2] = \text{Var}[R_1] + \text{Var}[R_2]. \tag{21.11}$$

Proof. We may assume that $\text{E}[R_i] = 0$ for $i = 1, 2$, since we could always replace R_i by $R_i - \text{E}[R_i]$ in equation (21.11). This substitution preserves the independence of the variables, and by Theorem 21.4.4, does not change the variances.

Now by Lemma 21.4.1, $\text{Var}[R_i] = \text{E}\left[R_i^2\right]$ and $\text{Var}[R_1 + R_2] = \text{E}\left[(R_1 + R_2)^2\right]$, so we need only prove

$$\text{E}\left[(R_1 + R_2)^2\right] = \text{E}\left[R_1^2\right] + \text{E}\left[R_2^2\right]. \tag{21.12}$$

But (21.12) follows from linearity of expectation and the fact that

$$\text{E}[R_1 R_2] = \text{E}[R_1]\,\text{E}[R_2] \tag{21.13}$$

since R_1 and R_2 are independent:

$$\begin{aligned}
\text{E}\left[(R_1 + R_2)^2\right] &= \text{E}\left[R_1^2 + 2R_1R_2 + R_2^2\right] \\
&= \text{E}\left[R_1^2\right] + 2\,\text{E}[R_1R_2] + \text{E}\left[R_2^2\right] \\
&= \text{E}\left[R_1^2\right] + 2\,\text{E}[R_1]\,\text{E}[R_2] + \text{E}\left[R_2^2\right] && \text{(by (21.13))} \\
&= \text{E}\left[R_1^2\right] + 2 \cdot 0 \cdot 0 + \text{E}\left[R_2^2\right] \\
&= \text{E}\left[R_1^2\right] + \text{E}\left[R_2^2\right]
\end{aligned}$$

■

An independence condition is necessary. If we ignored independence, then we would conclude that $\text{Var}[R + R] = \text{Var}[R] + \text{Var}[R]$. However, by Theorem 21.4.3, the left side is equal to $4\,\text{Var}[R]$, whereas the right side is $2\,\text{Var}[R]$. This implies that $\text{Var}[R] = 0$, which, by the Lemma above, essentially only holds if R is constant.

The proof of Theorem 21.4.6 carries over straightforwardly to the sum of any finite number of variables. So we have:

Theorem 21.4.7. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]. \tag{21.14}$$

Now we have a simple way of computing the variance of a variable, J, that has an (n, p)-binomial distribution. We know that $J = \sum_{k=1}^{n} I_k$ where the I_k are mutually independent indicator variables with $\Pr\{I_k = 1\} = p$. The variance of each I_k is $p(1 - p)$ by Lemma 21.4.2, so by linearity of variance, we have

Lemma (Variance of the Binomial Distribution)**.** *If J has the (n, p)-binomial distribution, then*

$$\text{Var}[J] = n\,\text{Var}[I_k] = np(1 - p). \tag{21.15}$$

21.4.5 Problems

Practice Problems

Problem 21.2.

A gambler plays 120 hands of draw poker, 60 hands of black jack, and 20 hands of stud poker per day. He wins a hand of draw poker with probability 1/6, a hand of black jack with probability 1/2, and a hand of stud poker with probability 1/5.

(a) What is the expected number of hands the gambler wins in a day?

(b) What would the Markov bound be on the probability that the gambler will win at least 108 hands on a given day?

(c) Assume the outcomes of the card games are pairwise independent. What is the variance in the number of hands won per day?

(d) What would the Chebyshev bound be on the probability that the gambler will win at least 108 hands on a given day? You may answer with a numerical expression that is not completely evaluated.

Class Problems

Problem 21.3.

The hat-check staff has had a long day serving at a party, and at the end of the party they simply return people's hats at random. Assume that n people checked hats at the party.

(a) What is the expected number of people who get their own hat back?

Let $X_i = 1$ be the indicator variable for the ith person getting their own hat back. Let $S_n = \sum_{i=1}^{n} X_i$, so S_n is the total number of people who get their own hat back.

(b) Write a simple formula for $\mathrm{E}[X_i X_j]$ for $i \neq j$. *Hint:* What is $\Pr\{X_j = 1 \mid X_i = 1\}$?

(c) Explain why you cannot use the variance of sums formula to calculate $\mathrm{Var}[S_n]$.

(d) Show that $\mathrm{E}\left[S_n^2\right] = 2$. *Hint:* $X_i^2 = X_i$.

(e) What is the variance of S_n?

(f) Use the Chebyshev bound to show that the probability that 11 or more people get their own hat back is at most 0.01.

Problem 21.4.

For any random variable, R, with mean, μ, and standard deviation, σ, the Chebyshev Bound says that for any real number $x > 0$,

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

Show that for any real number, μ, and real numbers $x \geq \sigma > 0$, there is an R for which the Chebyshev Bound is tight, that is,

$$\Pr\{|R| \geq x\} = \left(\frac{\sigma}{x}\right)^2. \tag{21.16}$$

Hint: First assume $\mu = 0$ and let R only take values $0, -x$, and x.

Problem 21.5.
Let R be a positive integer valued random variable such that

$$f_R(n) = \frac{1}{cn^3},$$

where

$$c ::= \sum_{n=1}^{\infty} \frac{1}{n^3}.$$

(a) Prove that $\mathrm{E}[R]$ is finite.

(b) Prove that $\mathrm{Var}[R]$ is infinite.

Homework Problems

Problem 21.6.
There is a "one-sided" version of Chebyshev's bound for deviation above the mean:

Lemma (One-sided Chebyshev bound)**.**

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

Hint: Let $S_a ::= (R - \mathrm{E}[R] + a)^2$, for $0 \leq a \in \mathbb{R}$. So $R - \mathrm{E}[R] \geq x$ implies $S_a \geq (x+a)^2$. Apply Markov's bound to $\Pr\left\{S_a \geq (x+a)^2\right\}$. Choose a to minimize this last bound.

Problem 21.7.
A man has a set of n keys, one of which fits the door to his apartment. He tries the keys until he finds the correct one. Give the expectation and variance for the number of trials until success if

(a) he tries the keys at random (possibly repeating a key tried earlier)

(b) he chooses keys randomly from among those he has not yet tried.

21.5 Estimation by Random Sampling

Polling again

Suppose we had wanted an advance estimate of the fraction of the Massachusetts voters who favored Scott Brown over everyone else in the recent Democratic primary election to fill Senator Edward Kennedy's seat.

Let p be this unknown fraction, and let's suppose we have some random process —say throwing darts at voter registration lists —which will select each voter with equal probability. We can define a Bernoulli variable, K, by the rule that $K = 1$ if the random voter most prefers Brown, and $K = 0$ otherwise.

Now to estimate p, we take a large number, n, of random choices of voters[1] and count the fraction who favor Brown. That is, we define variables $K_1, K_2, \dots$, where K_i is interpreted to be the indicator variable for the event that the ith chosen voter prefers Brown. Since our choices are made independently, the K_i's are independent. So formally, we model our estimation process by simply assuming we have mutually independent Bernoulli variables $K_1, K_2, \dots$, each with the same probability, p, of being equal to 1. Now let S_n be their sum, that is,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{21.17}$$

So S_n has the binomial distribution with parameter n, which we can choose, and unknown parameter p.

The variable S_n/n describes the fraction of voters we will sample who favor Scott Brown. Most people intuitively expect this sample fraction to give a useful approximation to the unknown fraction, p —and they would be right. So we will use the sample value, S_n/n, as our *statistical estimate* of p and use the Pairwise Independent Sampling Theorem 21.5.1 to work out how good an estinate this is.

21.5.1 Sampling

Suppose we want our estimate to be within 0.04 of the Brown favoring fraction, p, at least 95% of the time. This means we want

$$\Pr\left\{ \left| \frac{S_n}{n} - p \right| \leq 0.04 \right\} \geq 0.95\,. \tag{21.18}$$

So we better determine the number, n, of times we must poll voters so that inequality (21.18) will hold.

Now S_n is binomially distributed, so from (21.15) we have

$$\text{Var}\left[S_n\right] = n(p(1-p)) \leq n \cdot \frac{1}{4} = \frac{n}{4}$$

The bound of $1/4$ follows from the fact that $p(1-p)$ is maximized when $p = 1-p$, that is, when $p = 1/2$ (check this yourself!).

[1]We're choosing a random voter n times *with replacement*. That is, we don't remove a chosen voter from the set of voters eligible to be chosen later; so we might choose the same voter more than once in n tries! We would get a slightly better estimate if we required n *different* people to be chosen, but doing so complicates both the selection process and its analysis, with little gain in accuracy.

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[S_n\right] && \text{(by (21.9))}\\
&\leq \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (21.5.1))}\\
&= \frac{1}{4n} && (21.19)
\end{aligned}$$

Now from Chebyshev and (21.19) we have:

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \geq 0.04\right\} \leq \frac{\text{Var}\left[S_n/n\right]}{(0.04)^2} = \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \tag{21.20}$$

To make our our estimate with 95% confidence, we want the righthand side of (21.20) to be at most 1/20. So we choose n so that

$$\frac{156.25}{n} \leq \frac{1}{20},$$

that is,

$$n \geq 3,125.$$

A more exact calculation of the tail of this binomial distribution shows that the above sample size is about four times larger than necessary, but it is still a feasible size to sample. The fact that the sample size derived using Chebyshev's Theorem was unduly pessimistic should not be surprising. After all, in applying the Chebyshev Theorem, we only used the variance of S_n. It makes sense that more detailed information about the distribution leads to better bounds. But working through this example using only the variance has the virtue of illustrating an approach to estimation that is applicable to arbitrary random variables, not just binomial variables.

21.5.2 Matching Birthdays

There are important cases where the relevant distributions are not binomial because the mutual independence properties of the voter preference example do not hold. In these cases, estimation methods based on the Chebyshev bound may be the best approach. Birthday Matching is an example. We already saw in Section 18.5 that in a class of 85 students it is virtually certain that two or more students will have the same birthday. This suggests that quite a few pairs of students are likely to have the same birthday. How many?

So as before, suppose there are n students and d days in the year, and let D be the number of pairs of students with the same birthday. Now it will be easy to calculate the expected number of pairs of students with matching birthdays. Then we can take the same approach as we did in estimating voter preferences to get an estimate of the probability of getting a number of pairs close to the expected number.

Unlike the situation with voter preferences, having matching birthdays for different pairs of students are not mutually independent events, but the matchings are *pairwise independent*, as explained in Section 18.5. as we did for voter preference. Namely, let $B_1, B_2, \dots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event that the ith and jth people chosen have the same birthdays, that is, the event $[B_i = B_j]$. So our probability model, the B_i's are mutually independent variables, the $E_{i,j}$'s are pairwise independent. Also, the expectations of $E_{i,j}$ for $i \neq j$ equals the probability that $B_i = B_j$, namely, $1/d$.

Now, D, the number of matching pairs of birthdays among the n choices is simply the sum of the $E_{i,j}$'s:

$$D ::= \sum_{1 \le i < j \le n} E_{i,j}. \tag{21.21}$$

So by linearity of expectation

$$\mathrm{E}[D] = \mathrm{E}\left[\sum_{1 \le i < j \le n} E_{i,j}\right] = \sum_{1 \le i < j \le n} \mathrm{E}[E_{i,j}] = \binom{n}{2} \cdot \frac{1}{d}.$$

Similarly,

$$\begin{aligned}
\mathrm{Var}[D] &= \mathrm{Var}\left[\sum_{1 \le i < j \le n} E_{i,j}\right] \\
&= \sum_{1 \le i < j \le n} \mathrm{Var}[E_{i,j}] && \text{(by Theorem 21.4.7)} \\
&= \binom{n}{2} \cdot \frac{1}{d}\left(1 - \frac{1}{d}\right). && (by Lemma\ 21.4.2)
\end{aligned}$$

In particular, for a class of $n = 85$ students with $d = 365$ possible birthdays, we have $\mathrm{E}[D] \approx 9.7$ and $\mathrm{Var}[D] < 9.7(1 - 1/365) < 9.7$. So by Chebyshev's Theorem

$$\Pr\{|D - 9.7| \ge x\} < \frac{9.7}{x^2}.$$

Letting $x = 5$, we conclude that there is a better than 50% chance that in a class of 85 students, the number of pairs of students with the same birthday will be between 5 and 14.

21.5.3 Pairwise Independent Sampling

The reasoning we used above to analyze voter polling and matching birthdays is very similar. We summarize it in slightly more general form with a basic result we call the Pairwise Independent Sampling Theorem. In particular, we do not need to restrict ourselves to sums of zero-one valued variables, or to variables with the same distribution. For simplicity, we state the Theorem for pairwise independent variables with possibly different distributions but with the same mean and variance.

Theorem 21.5.1 (Pairwise Independent Sampling)**.** *Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{21.22}$$

Then

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

Proof. We observe first that the expectation of S_n/n is μ:

$$\begin{aligned}
\mathrm{E}\left[\frac{S_n}{n}\right] &= \mathrm{E}\left[\frac{\sum_{i=1}^{n} G_i}{n}\right] && \text{(def of } S_n\text{)}\\
&= \frac{\sum_{i=1}^{n} \mathrm{E}\left[G_i\right]}{n} && \text{(linearity of expectation)}\\
&= \frac{\sum_{i=1}^{n} \mu}{n}\\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

The second important property of S_n/n is that its variance is the variance of G_i divided by n:

$$\begin{aligned}
\mathrm{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \mathrm{Var}\left[S_n\right] && \text{(by (21.9))}\\
&= \frac{1}{n^2} \mathrm{Var}\left[\sum_{i=1}^{n} G_i\right] && \text{(def of } S_n\text{)}\\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\left[G_i\right] && \text{(pairwise independent additivity)}\\
&= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}. && (21.23)
\end{aligned}$$

This is enough to apply Chebyshev's Theorem and conclude:

$$\begin{aligned}
\Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq x\right\} &\leq \frac{\mathrm{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (21.23))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

■

The Pairwise Independent Sampling Theorem provides a precise general statement about how the average of independent samples of a random variable approaches the mean. In particular, it proves what is known as the Law of Large Numbers[2] : by choosing a large enough sample size, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

Corollary 21.5.2. *[Weak Law of Large Numbers] Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and the same finite deviation, and let*

$$S_n ::= \frac{\sum_{i=1}^{n} G_i}{n}.$$

Then for every $\epsilon > 0$,

$$\lim_{n \to \infty} \Pr\{|S_n - \mu| \leq \epsilon\} = 1.$$

21.6 Confidence versus Probability

So Chebyshev's Bound implies that sampling 3,125 voters will yield a fraction that, 95% of the time, is within 0.04 of the actual fraction of the voting population who prefer Brown.

Notice that the actual size of the voting population was never considered because *it did not matter*. People who have not studied probability theory often insist that the population size should matter. But our analysis shows that polling a little over 3000 people people is always sufficient, whether there are ten thousand, or million, or billion ... voters. You should think about an intuitive explanation that might persuade someone who thinks population size matters.

Now suppose a pollster actually takes a sample of 3,125 random voters to estimate the fraction of voters who prefer Brown, and the pollster finds that 1250 of them prefer Brown. It's tempting, **but sloppy**, to say that this means:

False Claim. *With probability 0.95, the fraction, p, of voters who prefer Brown is $1250/3125 \pm 0.04$. Since $1250/3125 - 0.04 > 1/3$, there is a 95% chance that more than a third of the voters prefer Brown to all other candidates.*

What's objectionable about this statement is that it talks about the probability or "chance" that a real world fact is true, namely that the actual fraction, p, of voters favoring Brown is more than $1/3$. 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 0.3; then it's nonsense to ask about the probability that it is within 0.04 of $1250/3125$ —it simply isn't.

This example of voter preference is typical: we want to estimate a fixed, unknown real-world quantity. But *being unknown does not make this quantity a random variable*, so it makes no sense to talk about the probability that it has some property.

[2]This is the *Weak* Law of Large Numbers. As you might suppose, there is also a Strong Law, but it's outside the scope of 6.042.

A more careful summary of what we have accomplished goes this way:

> We have described a probabilistic procedure for estimating the value of the actual fraction, p. The probability that *our estimation procedure* will yield a value within 0.04 of p 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 Brown is $1250/3125 \pm 0.04$.

So confidence levels refer to the results of estimation procedures for real-world quantities. The phrase "confidence level" should be heard as a reminder that some statistical procedure was used to obtain an estimate, and in judging the credibility of the estimate, it may be important to learn just what this procedure was.

21.6.1 Problems

Practice Problems

Problem 21.8.
You work for the president and you want to estimate the fraction p of voters in the entire nation that will prefer him in the upcoming elections. You do this by random sampling. Specifically, you select n voters independently and randomly, ask them who they are going to vote for, and use the fraction P of those that say they will vote for the President as an estimate for p.

(a) Our theorems about sampling and distributions allow us to calculate how confident we can be that the random variable, P, takes a value near the constant, p. This calculation uses some facts about voters and the way they are chosen. Which of the following facts are true?

1. Given a particular voter, the probability of that voter preferring the President is p.
2. Given a particular voter, the probability of that voter preferring the President is 1 or 0.
3. The probability that some voter is chosen more than once in the sequence goes to zero as n increases.
4. All voters are equally likely to be selected as the third in our sequence of n choices of voters (assuming $n \geq 3$).
5. The probability that the second voter chosen will favor the President, given that the first voter chosen prefers the President, is greater than p.
6. The probability that the second voter chosen will favor the President, given that the second voter chosen is from the same state as the first, may not equal p.

(b) Suppose that according to your calculations, the following is true about your polling:

$$\Pr\{|P - p| \leq 0.04\} \geq 0.95.$$

You do the asking, you count how many said they will vote for the President, you divide by n, and find the fraction is 0.53. You call the President, and ... what do you say?

1. Mr. President, $p = 0.53$!
2. Mr. President, with probability at least 95 percent, p is within 0.04 of 0.53.
3. Mr. President, either p is within 0.04 of 0.53 or something very strange (5-in-100) has happened.
4. Mr. President, we can be 95% confident that p is within 0.04 of 0.53.

Class Problems

Problem 21.9.
A recent Gallup poll found that 35% of the adult population of the United States believes that the theory of evolution is "well-supported by the evidence." Gallup polled 1928 Americans selected uniformly and independently at random. Of these, 675 asserted belief in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is $675/1928 \approx 0.350$. Gallup claims a margin of error of 3 percentage points, that is, he claims to be confident that his estimate is within 0.03 of the actual percentage.

(a) What is the largest variance an indicator variable can have?

(b) Use the Pairwise Independent Sampling Theorem to determine a confidence level with which Gallup can make his claim.

(c) Gallup actually claims greater than 99% confidence in his estimate. How might he have arrived at this conclusion? (Just explain what quantity he could calculate; you do not need to carry out a calculation.)

(d) Accepting the accuracy of all of Gallup's polling data and calculations, can you conclude that there is a high probability that the number of adult Americans who believe in evolution is 35 ± 3 percent?

Problem 21.10.
Suppose there are n students and d days in the year, and let D be the number of pairs of students with the same birthday. Let $B_1, B_2, \ldots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event $[B_i = B_j]$.

(a) What are $\text{E}[E_{i,j}]$ and $\text{Var}[E_{i,j}]$?

(b) What is $\mathrm{E}[D]$?

(c) What is $\mathrm{Var}[D]$?

(d) In a 6.01 class of 500 students, the youngest student was born in 1995 and the oldest in 1975. Let S be the number of students in the class who were born on exactly the same day. What is the probability that $4 \leq S \leq 32$? (For simplicity, assume that the distribution of birthdays is uniform over the 7305 days in the two decade interval from 1975 to 1995.)

Problem 21.11.
A defendent in traffic court is trying to beat a speeding ticket on the grounds that —since virtually everybody speeds on the turnpike —the police have unconstitutional discretion in giving tickets to anyone they choose. (By the way, we don't recommend this defense :-))

To support his argument, the defendent arranged to get a random sample of trips by 3,125 cars on the turnpike and found that 94% of them broke the speed limit at some point during their trip. He says that as a consequence of sampling theory (in particular, the Pairwise Independent Sampling Theorem), the court can be 95% confident that the actual percentage of all cars that were speeding is $94 \pm 4\%$.

The judge observes that the actual number of car trips on the turnpike was never considered in making this estimate. He is skeptical that, whether there were a thousand, a million, or 100,000,000 car trips on the turnpike, sampling only 3,125 is sufficient to be so confident.

Suppose you were were the defendent. How would you explain to the judge why the number of randomly selected cars that have to be checked for speeding *does not depend on the number of recorded trips*? Remember that judges are not trained to understand formulas, so you have to provide an intuitive, nonquantitative explanation.

Problem 21.12.
The proof of the Pairwise Independent Sampling Theorem 21.5.1 was given for a sequence $R_1, R_2, \ldots$ of pairwise independent random variables with the same mean and variance.

The theorem generalizes straighforwardly to sequences of pairwise independent random variables, possibly with *different* distributions, as long as all their variances are bounded by some constant.

Theorem (Generalized Pairwise Independent Sampling)**.** *Let $X_1, X_2, \ldots$ be a sequence of pairwise independent random variables such that* $\mathrm{Var}[X_i] \leq b$ *for some* $b \geq 0$

and all $i \geq 1$*. Let*

$$A_n ::= \frac{X_1 + X_2 + \cdots + X_n}{n},$$
$$\mu_n ::= \mathrm{E}[A_n].$$

Then for every $\epsilon > 0$*,*

$$\Pr\{|A_n - \mu_n| > \epsilon\} \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}. \tag{21.24}$$

(a) Prove the Generalized Pairwise Independent Sampling Theorem.

(b) Conclude that the following holds:

Corollary (Generalized Weak Law of Large Numbers)**.** *For every* $\epsilon > 0$*,*

$$\lim_{n\to\infty} \Pr\{|A_n - \mu_n| \leq \epsilon\} = 1.$$

Problem 21.13.

An *International Journal of Epidemiology* has a policy that they will only publish the results of a drug trial when there were enough patients in the drug trial to be sure that the conclusions about the drug's effectiveness hold at the 95% confidence level. The editors of the Journal reason that under this policy, their readership can be confident that at most 5% of the published studies will be mistaken.

Later, the editors are astonished and embarrassed to learn that *every one* of the 20 drug trial results they published during the year was wrong. This happened even though the editors and reviewers had carefully checked the submitted data, and every one of the trials was *properly performed and reported* in the published paper.

The editors thought the probability of this was negligible (namely, $(1/20)^{20} < 10^{-25}$). Explain what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Exam Problems

Problem 21.14.

Yesterday, the programmers at a local company wrote a large program. To estimate the fraction, b, of lines of code in this program that are buggy, the QA team will take a small sample of lines chosen randomly and independently (so it is possible, though unlikely, that the same line of code might be chosen more than once). For each line chosen, they can run tests that determine whether that line of code is buggy, after which they will use the fraction of buggy lines in their sample as their estimate of the fraction b.

The company statistician can use estimates of a binomial distribution to calculate a value, s, for a number of lines of code to sample which ensures that with 97% confidence, the fraction of buggy lines in the sample will be within 0.006 of the actual fraction, b, of buggy lines in the program.

Mathematically, the *program* is an actual outcome that already happened. The *sample* is a random variable defined by the process for randomly choosing s lines from the program. The justification for the statistician's confidence depends on some properties of the program and how the sample of s lines of code from the program are chosen. These properties are described in some of the statements below. Indicate which of these statements are true, and explain your answers.

1. The probability that the ninth line of code in the *program* is buggy is b.

2. The probability that the ninth line of code chosen for the *sample* is defective, is b.

3. All lines of code in the program are equally likely to be the third line chosen in the *sample*.

4. Given that the first line chosen for the *sample* is buggy, the probability that the second line chosen will also be buggy is greater than b.

5. Given that the last line in the *program* is buggy, the probability that the next-to-last line in the program will also be buggy is greater than b.

6. The expectation of the indicator variable for the last line in the *sample* being buggy is b.

7. Given that the first two lines of code selected in the *sample* are the same kind of statement —they might both be assignment statements, or both be conditional statements, or both loop statements,. . . —the probability that the first line is buggy may be greater than b.

8. There is zero probability that all the lines in the *sample* will be different.