## 6.7 Probability Distributions and Variance

Distributions of random variables

We have given meaning to the phrase expected value. For example, if we flip a coin 100 times, the expected number of heads is 50. But to what extent do we expect to see 50 heads. Would it be surprising to see 55, 60 or 65 heads instead? To answer this kind of question, we have to analyze how much we expect a random variable to deviate from its expected value. We will first see how to analyze graphically how the values of a random variable are distributed around its expected value. The *distribution function D* of a random variable X is the function on the values of X defined by

$$D(x) = P(X = x).$$

You probably recognize the distribution function from the role it played in the definition of expected value. The distribution function of X assigns to each value of X the probability that X achieves that value. When the values of X are integers, it is convenient to visualize the distribution function with a diagram called a *histogram*. In Figure 6.8 we show histograms for the distribution of the "number of heads" random variable for ten flips of a coin and the "number of right answers" random variable for someone taking a ten question test with probability .8 of getting a correct answer. What is a histogram? Those we have drawn are graphs which show for for each integer value x of X a rectangle of width 1, centered at x, whose height (and thus area) is proportional to the probability $P(X = x)$. Histograms can be drawn with non-unit width rectangles. When people draw a rectangle with a base ranging from $x = a$ to $x = b$, the area of the rectangle is the probability that X is between a and b.
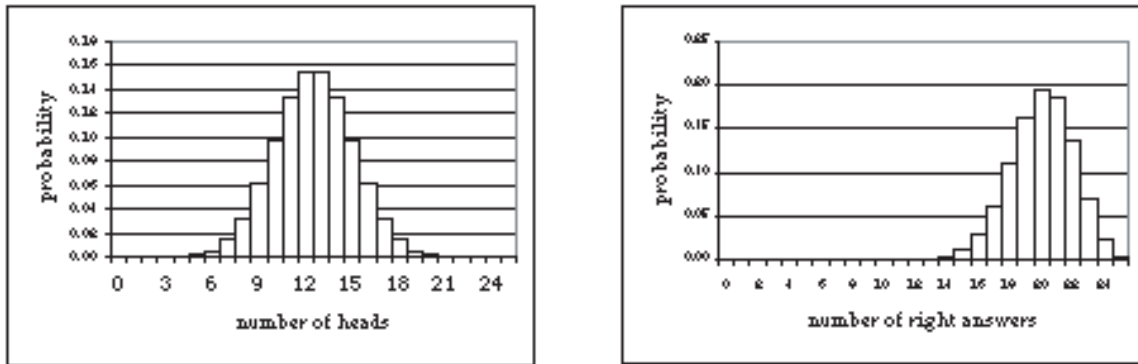
Figure 6.8: Two histograms.
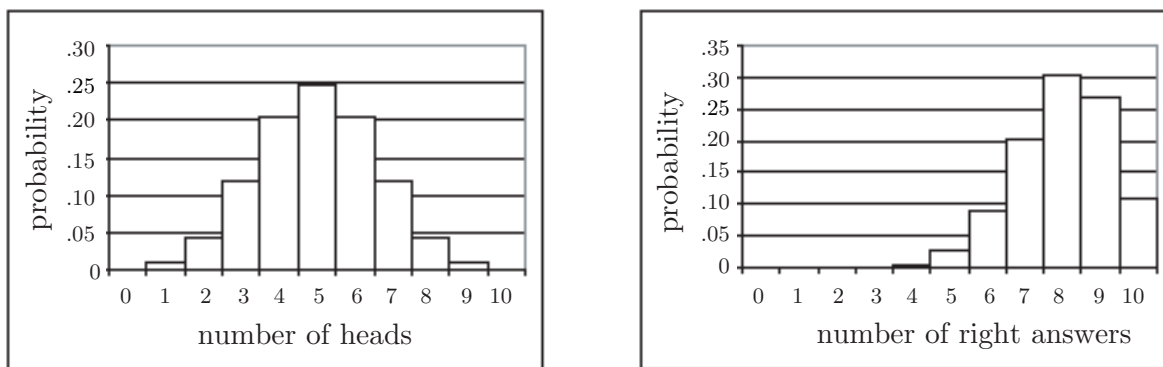
From the histograms you can see the difference in the two distributions. You can also see that we can expect the number of heads to be somewhat near the expected number, though as few heads as 2 or as many as 8 are not out of the question. We see that the number of right answers tends to be clustered between 6 and ten, so in this case we can expect to be reasonably close to the expected value. With more coin flips or more questions, however, will the results spread out? Relatively speaking, should we expect to be closer to or farther from the expected value? In Figure 6.9 we show the results of 25 coin flips or 25 questions. The expected number of heads is 12.5. The histogram makes it clear that we can expect the vast majority of our results

to have between 9 and 16 heads. Essentially all the results lie between 4 and 20 Thus the results
are not spread as broadly (relatively speaking) as they were with just ten flips. Once again the
test score histogram seems even more tightly packed around its expected value. Essentially all
the scores lie between 14 and 25. While we can still tell the difference between the shapes of the
histograms, they have become somewhat similar in appearance.

Figure 6.9: Histograms of 25 trials

In Figure 6.10 we have shown the thirty most relevant values for 100 flips of a coin and a 100
question test. Now the two histograms have almost the same shape, though the test histogram is
still more tightly packed around its expected value. The number of heads has virtually no chance
of deviating by more than 15 from its expected value, and the test score has almost no chance
of deviating by more than 11 from the expected value. Thus the spread has only doubled, even
though the number of trials has quadrupled. In both cases the curve formed by the tops of the
rectangles seems quite similar to the bell shaped curve called the normal curve that arises in so
many areas of science. In the test-taking curve, though, you can see a bit of difference between
the lower left-hand side and the lower right hand side.

Since we needed about 30 values to see the most relevant probabilities for these curves, while
we needed 15 values to see most of the relevant probabilities for independent trials with 25 items,

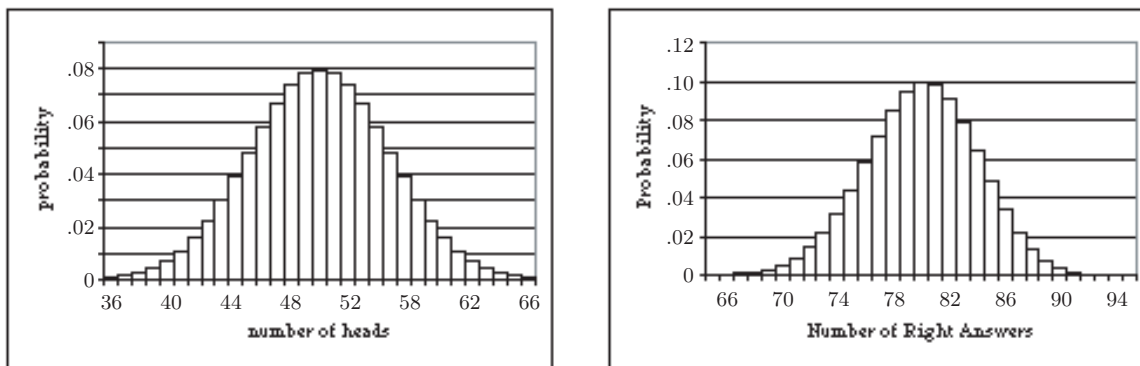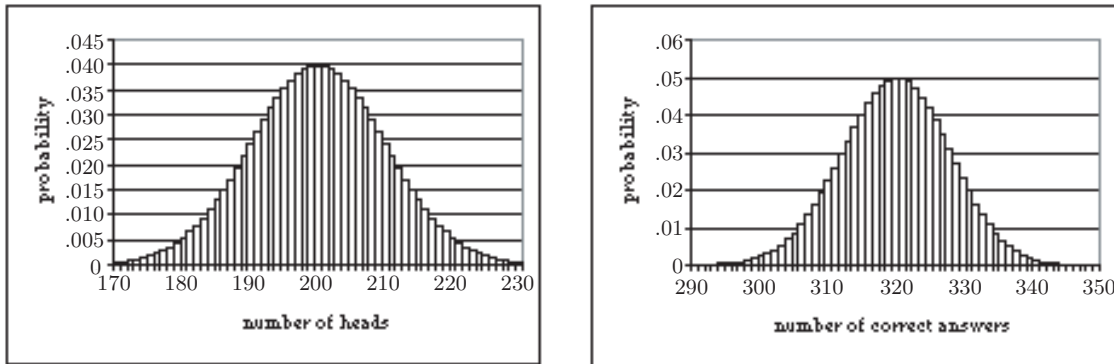Figure 6.10: One hundred independent trials

Figure 6.11: Four hundred independent trials

we might predict that we would need only about 60 values to see essentially all the results in four hundred trials. As Figure 6.11 shows, this is indeed the case. The test taking distribution is still more tightly packed than the coin flipping distribution, but we have to examine it closely to find any asymmetry. These experiments are convincing, and they suggest that the spread of a distribution (for independent trials) grows as the square root of the number of trials, because each time we quadruple the number of elements, we double the spread. They also suggest there is some common kind of bell-shaped limiting distribution function for at least the distribution of successes in independent trials with two outcomes. However without a theoretical foundation we don't know how far the truth of our observations extends. Thus we seek an algebraic expression of our observations. This algebraic measure should somehow measure the difference between a random variable and its expected value.

Variance

Exercise 6.7-1 Suppose the X is the number of heads in four flips of a coin. Let Y be the random variable $X - 2$, the difference between X and its expected value. Compute $E(Y)$. Doe it effectively measure how much we expect to see X deviate from its expected value? Compute $E(Y^2)$. Try repeating the process with X being the number of heads in ten flips of a coin and Y being $X - 5$.

Before answering these questions, we state a trivial, but useful lemma (which appeared as Problem 9 in Section 4 of this chapter and corollary showing that the expected value of an expectation is that expectation.

Lemma 6.25 *If X is a random variable that always takes on the value c, then $E(X) = c$.*

Proof: $E(X) = P(X = c) \cdot c = 1 \cdot c = c$. ∎

We can think of a constant c as a random variable that always takes on the value c. When we do, we will just write $E(c)$ for the expected value of this random variable, in which case our lemma says that $E(c) = c$. This lemma has an important corollary.

Corollary 6.26 $E(E(X)) = E(X)$.

Proof: When we think of $E(X)$ as a random variable, it has a constant value, μ. By Lemma 6.25 $E(E(x)) = E(\mu) = \mu = E(x)$. ∎

Returning to Exercise 6.7-1, we can use linearity of expectation and Corollary 6.26 to show that

$$E(X - E(X)) = E(X) - E(E(X)) = E(X) - E(X) = 0. \tag{6.42}$$

Thus this is not a particularly useful measure of how close a random variable is to its expectation. If a random variable is sometimes above its expectation and sometimes below, you would like these two differences to somehow add together, rather than cancel each other out. This suggests we try to convert the values of $X - E(X)$ to positive numbers in some way and then take the expectation of these positive numbers as our measure of spread. There are two natural ways to make numbers positive, taking their absolute value and squaring them. It turns our that to prove things about the spread of expected values, squaring is more useful. Could we have guessed that? Perhaps, since we see that the spread seems to grow with the square root, and the square root isn't related to the absolute value in the way it is related to the squaring function. On the other hand, as you saw in the example, computing expected values of these squares from what we know now is time consuming. A bit of theory will make it easier.

We define the **variance** $V(X)$ of a random variable X as the expected value of $(X - E(X))^2$. We can also express this as a sum over the individual elements of the sample space S and get that

$$V(X) = E(X - E(X))^2 = \sum_{s:s\in S} P(s)(X(s) - E(X))^2. \tag{6.43}$$

Now let's apply this definition and compute the variance in the number X of heads in four flips of a coin. We have

$$V(X) = (0-2)^2 \cdot \frac{1}{16} + (1-2)^2 \cdot \frac{1}{4} + (2-2)^2 \cdot \frac{3}{8} + (3-2)^2 \cdot \frac{1}{4} + (4-2)^2 \cdot \frac{1}{16} = 1.$$

Computing the variance for ten flips of a coin involves some very inconvenient arithmetic. It would be nice to have a computational technique that would save us from having to figure out large sums if we want to compute the variance for ten or even 100 or 400 flips of a coin to check our intuition about how the spread of a distribution grows. We saw before that the expected value of a sum of random variables is the sum of the expected values of the random variables. This was very useful in making computations.

Exercise 6.7-2 What is the variance for the number of heads in one flip of a coin? What is the sum of the variances for four independent trials of one flip of a coin?

Exercise 6.7-3 We have a nickel and quarter in a cup. We withdraw one coin. What is the expected amount of money we withdraw? What is the variance? We withdraw two coins, one after the other without replacement. What is the expected amount of money we withdraw? What is the variance? What is the expected amount of money and variance for the first draw? For the second draw?

Exercise 6.7-4 Compute the variance for the number of right answers when we answer one question with probability .8 of getting the right answer (note that the number of right answers is either 0 or 1, but the expected value need not be). Compute the variance for the number of right answers when we answer 5 questions with probability .8 of getting the right answer. Do you see a relationship?

In Exercise 6.7-2 we can compute the variance

$$V(X) = (0 - \frac{1}{2})^2 \cdot \frac{1}{2} + (1 - \frac{1}{2})^2 \cdot \frac{1}{2} = \frac{1}{4}.$$

Thus we see that the variance for one flip is $1/4$ and sum of the variances for four flips is 1. In Exercise 6.7-4 we see that for one question the variance is

$$V(X) = .2(0 - .8)^2 + .8(1 - .8)^2 = .16$$

For five questions the variance is

$$4^2 \cdot (.2)^5 + 3^2 \cdot 5 \cdot (.2)^4 \cdot (.8) + 2^2 \cdot 10 \cdot (.2)^3 \cdot (.8)^2 + 1^2 \cdot 10 \cdot (.2)^2 \cdot (.8)^3 +$$
$$0^2 \cdot 5 \cdot (.2)^1 \cdot (.8)^4 + 1^2 \cdot (.8)^5 = .8$$

The result is five times the variance for one question.

For Exercise 6.7-3 the expected amount of money for one draw is $.15. The variance is

$$(.05 - .15)^2 \cdot .5 + (.25 - .15)^2 \cdot .5 = .01.$$

For removing both coins, one after the other, the expected amount of money is $.30 and the variance is 0. Finally the expected value and variance on the first draw are $.15 and .01 and the expected value and variance on the second draw are $.15 and .01.

It would be nice if we had a simple method for computing variance by using a rule like "the expected value of a sum is the sum of the expected values." However Exercise 6.7-3 shows that the variance of a sum is not always the sum of the variances. On the other hand, Exercise 6.7-2 and Exercise 6.7-4 suggest such a result might be true for a sum of variances in independent trials processes. In fact slightly more is true. We say random variables X and Y are *independent* when the event that X has value x is independent of the event that Y has value y, regardless of the choice of x and y. For example, in n flips of a coin, the number of heads on flip i (which is 0 or 1) is independent of the number of heads on flip j. To show that the variance of a sum of independent random variables is the sum of their variances, we first need to show that the expected value of the product of two *independent* random variables is the product of their expected values.

Lemma 6.27 *If X and Y are independent random variables on a sample space S with values $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_m$ respectively, then*

$$E(XY) = E(X)E(Y).$$

Proof: We prove the lemma by the following series of equalities. In going from (6.44) to (6.45), we use the fact that X and Y are independent; the rest of the equations follow from

definitions and algebra.

$$
\begin{aligned}
E(X)E(Y) &= \sum_{i=1}^{k} x_i P(X = x_i) \sum_{j=1}^{m} y_j P(Y = y_j) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j P(X = x_i) P(y = y_j) \\
&= \sum_{z:\ z\text{is a value of } XY} z \sum_{(i,j): x_i y_j = z} P(X = x_i) P(Y = y_j) \qquad (6.44) \\
&= \sum_{z:\ z\text{is a value of } XY} z \sum_{(i,j): x_i y_j = z} P((X = x_i) \wedge (Y = y_j)) \qquad (6.45) \\
&= \sum_{z:\ z\text{is a value of } XY} z P(XY = z) \\
&= E(XY)
\end{aligned}
$$

∎

Theorem 6.28 *If X and Y are independent random variables then*

$$
V(X + Y) = V(X) + V(Y).
$$

Proof: Using the definitions, algebra and linearity of expectation we have

$$
\begin{aligned}
V(X + Y) &= E((X + Y) - E(X + Y))^2 \\
&= E(X - E(X) + Y - E(Y))^2 \\
&= E((X - E(X))^2 + 2(X - E(X))(Y - E(Y)) + (Y - E(Y))^2) \\
&= E(X - E(X))^2 + 2E((X - E(X))(Y - E(Y)) + E(Y - E(Y))^2
\end{aligned}
$$

Now the first and last terms and just the definitions of $V(X)$ and $V(Y)$ respectively. Note also that if X and Y are independent and b and c are constants, then $X - b$ and $Y - c$ are independent (See Problem 8 at the end of this section.) Thus we can apply Lemma 6.27 to the middle term to obtain

$$
= V(X) + 2E(X - E(X))E(Y - E(Y)) + V(Y).
$$

Now we apply Equation 6.42 to the middle term to show that it is 0. This proves the theorem. ∎

With this theorem, computing the variance for ten flips of a coin is easy; as usual we have the random variable X_i that is 1 or 0 depending on whether or not the coin comes up heads. We saw that the variance of X_i is 1/4, so the variance for $X_1 + X_2 + \cdots + X_{10}$ is $10/4 = 2.5$.

Exercise 6.7-5 Find the variance for 100 flips of a coin and 400 flips of a coin.

Exercise 6.7-6 The variance in the previous problem grew by a factor of four when the number of trials grew by a factor of 4, while the spread we observed in our histograms grew by a factor of 2. Can you suggest a natural measure of spread that fixes this problem?

For Exercise 6.7-5 recall that the variance for one flip was 1/4. Therefore the variance for 100 flips is 25 and the variance for 400 flips is 100. Since this measure grows linearly with the size, we can take its square root to give a measure of spread that grows with the square root of the quiz size, as our observed "spread" did in the histograms. Taking the square root actually makes intuitive sense, because it "corrects" for the fact that we were measuring expected squared spread rather than expected spread.

The square root of the variance of a random variable is called the *standard deviation* of the random variable and is denoted by σ, or $\sigma(X)$ when there is a chance for confusion as to what random variable we are discussing. Thus the standard deviation for 100 flips is 5 and for 400 flips is 10. Notice that in both the 100 flip case and the 400 flip case, the "spread" we observed in the histogram was ± 3 standard deviations from the expected value. What about for 25 flips? For 25 flips the standard deviation will be 5/2, so ± 3 standard deviations from the expected value is a range of 15 points, again what we observed. For the test scores the variance is .16 for one question, so the standard deviation for 25 questions will be 2, giving us a range of 12 points. For 100 questions the standard deviation will be 4, and for 400 questions the standard deviation will be 8. Notice again how three standard deviations relate to the spread we see in the histograms.

Our observed relationship between the spread and the standard deviation is no accident. A consequence of a theorem of probability known as the central limit theorem is that the percentage of results within one standard deviation of the mean in a relatively large number of independent trials with two outcomes is about 68%; the percentage within two standard deviations of the mean is about 95.5%, and the percentage within three standard deviations of the mean is about 99.7%.

What the central limit theorem says is that the sum of independent random variables with the same distribution function is approximated well by saying that the probability that the sum is between a and b is an appropriately chosen multiple of $\int_a^b e^{-cx^2}\, dx$ (where c is an appropriate constant) when the number of random variables we are adding is sufficiently large.[8] The distribution given by that multiple of the integral is called the *normal distribution.* Since many of the things we observe in nature can be thought of as the outcome of multistage processes, and the quantities we measure are often the result of adding some quantity at each stage, the central limit theorem "explains" why we should expect to see normal distributions for so many of the things we do measure. While weights can be thought of as the sum of the weight change due to eating and exercise each week, say, this is not a natural interpretation for blood pressures. Thus while we shouldn't be particularly surprised that weights are normally distributed, we don't have the same basis for predicting that blood pressures would be normally distributed, even though they are!

Exercise 6.7-7 If we want to be 95% sure that the number of heads in n flips of a coin is within $\pm 1\%$ of the expected value, how big does n have to be?

Exercise 6.7-8 What is the variance and standard deviation for the number of right answers for someone taking a 100 question short answer test where each answer is graded either correct or incorrect if the person knows 80% of the subject material for the test the test and answers correctly each question she knows? Should we be surprised if such a student scores 90 or above on the test?

---

[8]Still more precisely, if we let μ be the expected value of the random variable X_i and σ be its standard deviation (all X_i have the same expected value and standard distribution since they have the same distribution) and scale the sum of our random variables by $Z = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma \sqrt{n}}$, then the probability that $a \le Z \le b$ is $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$.

Recall that for one flip of a coin the variance is $1/4$, so that for n flips it is $n/4$. Thus for n flips the standard deviation is $\sqrt{n}/2$. We expect that 95% of our outcomes will be within 2 standard deviations of the mean (people always round 95.5 to 95) so we are asking when two standard deviations are 1% of $n/2$. Thus we want an n such that $2\sqrt{n}/2 = .01(.5n)$, or such that $\sqrt{n} = 5 \cdot 10^{-3}n$, or $n = 25 \cdot 10^{-6}n^2$. This gives us $n = 10^6/25 = 40,000$.

For Exercise 6.7-8, the expected number of correct answers on any given question is .8. The variance for each answer is $.8(1 - .8)^2 + .2(0 - .8)^2 = .8 \cdot .04 + .2 \cdot .64 = .032 + .128 = .16$. Notice this is $.8 \cdot (1 - .8)$. The total score is the sum of the random variables giving the number of points on each question, and assuming the questions are independent of each other, the variance of their sum is the sum of their variances, or 16. Thus the standard deviation is 4. Since 90% is 2.5 standard deviations above the expected value, the probability of getting that a score that far from the expected value is somewhere between .05 and .003 by the Central Limit Theorem. (In fact it is just a bit more than .01). Assuming that someone is just as likely to be 2.5 standard deviations below the expected score as above, which is not exactly right but close, we see that it is quite unlikely that someone who knows 80% of the material would score 90% or above on the test. Thus we should be surprised by such a score, and take the score as evidence that the student likely knows more than 80% of the material.

Coin flipping and test taking are two special cases of Bernoulli trials. With the same kind of computations we used for the test score random variable, we can prove the following.

Theorem 6.29 *In Bernoulli trials with probability p of success, the variance for one trial is $p(1 - p)$ and for n trials is $np(1 - p)$, so the standard deviation for n trials is $\sqrt{np(1 - p)}$.*

Proof: You are asked to give the proof in Problem 7.■

Important Concepts, Formulas, and Theorems

1. *Histogram.* *Histograms* are graphs which show for for each integer value x of a random variable X a rectangle of width 1, centered at x, whose height (and thus area) is proportional to the probability $P(X = x)$. Histograms can be drawn with non-unit width rectangles. When people draw a rectangle with a base ranging from $x = a$ to $x = b$, the area of the rectangle is the probability that X is between a and b.

2. *Expected Value of a Constant.* If X is a random variable that always takes on the value c, then $E(X) = c$. In particular, $E(E(X)) = E(X)$.

3. *Variance.* We define the *variance* $V(X)$ of a random variable X as the expected value of $(X - E(X))^2$. We can also express this as a sum over the individual elements of the sample space S and get that $V(X) = E(X - E(X))^2 = \sum_{s:s \in S} P(s)(X(s) - E(X))^2$.

4. *Independent Random Variables.* We say random variables X and Y are *independent* when the event that X has value x is independent of the event that Y has value y, regardless of the choice of x and y.

5. *Expected Product of Independent Random Variables.* If X and Y are independent random variables on a sample space S, then $E(XY) = E(X)E(Y)$.

6. *Variance of Sum of Independent Random Variables.* If X and Y are independent random variables then $V(X + Y) = V(X) + V(Y)$.

7. *Standard deviation.* The square root of the variance of a random variable is called the *standard deviation* of the random variable and is denoted by σ, or $\sigma(X)$ when there is a chance for confusion as to what random variable we are discussing.

8. *Variance and Standard Deviation for Bernoulli Trials.* In Bernoulli trials with probability p of success, the variance for one trial is $p(1-p)$ and for n trials is $np(1-p)$, so the standard deviation for n trials is $\sqrt{np(1-p)}$.

9. *Central Limit Theorem.* The central limit theorem says that the sum of independent random variables with the same distribution function is approximated well by saying that the probability that the random variable is between a and b is an appropriately chosen multiple of $\int_a^b e^{-cx^2}\, dx$, for some constant c, when the number of random variables we are adding is sufficiently large. This implies that the probability that a sum of independent random variables is within one, two, or three standard deviations of its expected value is approximately .68, .955, and .997.

Problems

1. Suppose someone who knows 60% of the material covered in a chapter of a textbook is taking a five question objective (each answer is either right or wrong, not multiple choice or true-false) quiz. Let X be the random variable that for each possible quiz, gives the number of questions the student answers correctly. What is the expected value of the random variable $X - 3$? What is the expected value of $(X - 3)^2$? What is the variance of X?

2. In Problem 1 let X_i be the number of correct answers the student gets on question i, so that X_i is either zero or one. What is the expected value of X_i? What is the variance of X_i? How does the sum of the variances of X_1 through X_5 relate to the variance of X for Problem 1?

3. We have a dime and a fifty cent piece in a cup. We withdraw one coin. What is the expected amount of money we withdraw? What is the variance? Now we draw a second coin, without replacing the first. What is the expected amount of money we withdraw? What is the variance? Suppose instead we consider withdrawing two coins from the cup together. What is the expected amount of money we withdraw, and what is the variance? What does this example show about whether the variance of a sum of random variables is the sum of their variances.

4. If the quiz in Problem 1 has 100 questions, what is the expected number of right answers, the variance of the expected number of right answers, and the standard deviation of the number of right answers?

5. Estimate the probability that a person who knows 60% of the material gets a grade strictly between 50 and 70 in the test of Exercise 6.7-4

6. What is the variance in the number of right answers for someone who knows 80% of the material on which a 25 question quiz is based? What if the quiz has 100 questions? 400 questions? How can we "correct" these variances for the fact that the "spread" in the histogram for the number of right answers random variable only doubled when we multiplied the number of questions in a test by 4?

7. Prove Theorem 6.29.

8. Show that if X and Y are independent and b and c are constant, then $X - b$ and $Y - c$ are independent.

9. We have a nickel, dime and quarter in a cup. We withdraw two coins, first one and then the second, without replacement. What is the expected amount of money and variance for the first draw? For the second draw? For the sum of both draws?

10. Show that the variance for n independent trials with two outcomes and probability p of success is given by $np(1-p)$. What is the standard deviation? What are the corresponding values for the number of failures random variable?

11. What are the variance and standard deviation for the sum of the tops of n dice that we roll?

12. How many questions need to be on a short answer test for us to be 95% sure that someone who knows 80% of the course material gets a grade between 75% and 85%?

13. Is a score of 70% on a 100 question true-false test consistent with the hypothesis that the test taker was just guessing? What about a 10 question true-false test? (This is not a plug and chug problem; you have to come up with your own definition of "consistent with.")

14. Given a random variable X, how does the variance of cX relate to that of X?

15. Draw a graph of the equation $y = x(1 - x)$ for x between 0 and 1. What is the maximum value of y? Why does this show that the variance (see Problem 10 in this section) of the "number of successes" random variable for n independent trials is less than or equal to $n/4$?

16. This problem develops an important law of probability known as *Chebyshev's law*. Suppose we are given a real number $r > 0$ and we want to estimate the probability that the difference $|X(x) - E(X)|$ of a random variable from its expected value is more than r.

    (a) Let $S = \{x_1, x_2, \ldots, x_n\}$ be the sample space, and let $E = \{x_1, x_2, \ldots, x_k\}$ be the set of all x such that $|X(x) - E(X)| > r$. By using the formula that defines $V(X)$, show that
    $$V(X) > \sum_{i=1}^{k} P(x_i)r^2 = P(E)r^2$$

    (b) Show that the probability that $|X(x) - E(X)| \geq r$ is no more than $V(X)/r^2$. This is called *Chebyshev's law*.

17. Use Problem 15 of this section to show that in n independent trials with probability p of success,
    $$P\left(\left|\frac{\#\ \text{of successes} - np}{n}\right| \geq r\right) \leq \frac{1}{4nr^2}$$

18. This problem derives an intuitive law of probability known as the *law of large numbers* from Chebyshev's law. Informally, the law of large numbers says if you repeat an experiment many times, the fraction of the time that an event occurs is very likely to be close to the probability of the event. In particular, we shall prove that for any positive number s, no matter how small, by making the number n independent trials in a sequence of independent trials large enough, we can make the probability that the number X of successes is between $np - ns$ and $np + ns$ as close to 1 as we choose. For example, we can make the probability that the number of successes is within 1% (or 0.1 per cent) of the expected number as close to 1 as we wish.

    (a) Show that the probability that $|X(x) - np| \geq sn$ is no more than $p(1 - p)/s^2 n$.

    (b) Explain why this means that we can make the probability that $X(x)$ is between $np - sn$ and $np + sn$ as close to 1 as we want by making n large.

19. On a true-false test, the score is often computed by subtracting the number of wrong answers from the number of right ones and converting that number to a percentage of the number of questions. What is the expected score on a true-false test graded this way of someone who knows 80% of the material in a course? How does this scheme change the standard deviation in comparison with an objective test? What must you do to the number of questions to be able to be a certain percent sure that someone who knows 80% gets a grade within 5 points of the expected percentage score?

20. Another way to bound the deviance from the expectation is known as Markov's inequality. This inequality says that if X is a random variable taking only non-negative values, then, for any $k \geq 1$,

$$P(X > kE(X)) \leq \frac{1}{k}.$$

Prove this inequality.