Chapter 14

Rules for Means and Variances; Prediction

14.1 Rules for Means and Variances

The result in this section is very technical and algebraic. And dry. But it is useful for understanding many of prediction results we obtain in this course, beginning later in this chapter.

We have independent random variables $W_1$ and $W_2$. Note that, typically, they are not identically distributed. The result below is an extremely special case of a much more general result. It will suffice, however, for our needs; thus, I see no reason to subject you to the pain of viewing the general result.

We need some notation:

- Let $\mu_1 [\mu_2]$ denote the mean of $W_1 [W_2]$.
- Let $\text{Var}(W_1) [\text{Var}(W_2)]$ denote the variance of $W_1 [W_2]$.
- Let $b$ denote any number. Define

$$W = W_1 - bW_2.$$ 

Result 14.1 (The mean and variance of $W$.) For the notation given above,

- The mean of $W$ is

$$\mu_W = \mu_1 - b\mu_2 \quad \text{(14.1)}$$

- The variance of $W$ is

$$\text{Var}(W) = \text{Var}(W_1) + b^2 \text{Var}(W_2). \quad \text{(14.2)}$$

In our two applications of this result in this chapter, the number $b$ will be taken to equal $\mu_1/\mu_2$; thus,

$$\mu_W = \mu_1 - (\mu_1/\mu_2)\mu_2 = \mu_1 - \mu_1 = 0,$$

for our applications.
14.2 Predicting for Bernoulli Trials

Predictions are tough, especially about the future—Yogi Berra.

We plan to observe $m$ Bernoulli trials and want to predict the total number of successes that will be obtained. Let $Y$ denote the random variable and $y$ the observed value of the total number of successes in the future $m$ trials. Similar to estimation, we will learn about point and interval predictions of the value of $Y$.

14.2.1 When $p$ is Known

Because prediction is a new idea in this course, I want to present a gentle introduction to it. Suppose that you have a favorite pair of dice, one colored blue and the other white. Let’s focus on the blue die. And let’s say your favorite number is 6—perhaps you play Risk a great deal; in Risk, 6 is the best outcome by far when one casts a die. Ghengis Khan playing Risk would roll a lot of 6’s.

You plan to cast the die 600 times and you want to predict the number of 6’s that you will obtain. You believe that the die is balanced; i.e., that the six possible outcomes are equally likely to occur.

OK. Quick. Don’t think of any of the wonderful things you have learned in this course. Give me your point (single number) prediction of how many 6’s you will obtain. I conjecture that your answer is 100. (I asked this question several times over the years to a live lecture and always—save once—received the answer 100 from the student who volunteered to answer. One year a guy said 72 and got a big laugh. I failed him because it is my job to make the jokes, such as they are. No, I didn’t really fail him, but I was more than a bit annoyed that he got a larger laugh than I did with my much cleverer anecdotes.)

My academic grandfather (my advisor’s advisor, who happened to be male) is Herb Robbins, a very brilliant and witty man. Herb was once asked what mathematical statisticians do, and he replied, “They find out what non-statisticians do and prove it’s optimal.”

Thus, I am going to argue that your answer of 100 is the best answer to the die question I posed above. In order to show that something is best mathematically, we find a way to measure good and whichever candidate answer has the largest amount of good is best. This is the approach for the glass-half-full people. More often, one finds a way to measure bad and whichever candidate answer has the smallest amount of bad is best.

We want to predict, in advance, the value that $Y$ will yield. We denote the point prediction by the single number $\hat{y}$. We adopt the criterion that we want the probability of being correct to be as large as possible. (Thus, we define being correct as good and seek to maximize the probability that we will get a good result.)

**Result 14.2 (The best point prediction of $Y$.)** Calculate the mean of $Y$, which is $mp$.

- **If $mp$ is an integer**, then it is uniquely the most probable value of $Y$ and our point prediction is $\hat{y} = mp$. 

338
• If $mp$ is not an integer, then the most probable value of $Y$ is one of the integers immediately on either side of $mp$. Check them both; whichever is more probable is the point prediction. If they are equally probably, I choose the even integer.

Below are some examples of this result.

• For my die example, $m = 600$ and $p = 1/6$, giving $mp = 600(1/6) = 100$. This is an integer; thus, 100 is the point prediction of $Y$. With the help of the website calculator (details not given), I find that $P(Y = 100) = 0.0437$. For comparison, $P(Y = 99) = 0.0436$ and $P(Y = 101) = 0.0432$. Thus, if 99 is your life-long favorite number, it is difficult for me to criticize using it as your point prediction. In the long-run, you will have one fewer correct point prediction for every 10,000 times you cast the blue die 600 times. That’s a lot of die casting!

• Suppose that $m = 200$ and $p = 0.50$. Then, $mp = 200(0.5) = 100$ is an integer; thus, 100 is the point prediction of $Y$. With the help of the website calculator, I find that $P(Y = 100) = 0.0563$.

• Suppose that $m = 300$ and $p = 0.30$. Then, $mp = 300(0.3) = 90$ is an integer; thus, 90 is the point prediction of $Y$. With the help of the website calculator, I find that $P(Y = 90) = 0.0502$.

• Suppose that $m = 20$ and $p = 0.42$. Then, $mp = 20(0.42) = 8.4$ is not an integer. The most likely value of $Y$ is either 8 or 9. With the help of the website calculator, I find that $P(Y = 8) = 0.1768$ and $P(Y = 9) = 0.1707$. Thus, $\hat{y} = 8$.

• Suppose that $m = 75$ and $p = 0.50$. Then, $mp = 75(0.50) = 37.5$ is not an integer. The most likely value of $Y$ is either 37 or 38. With the help of the website calculator, I find that $P(Y = 37) = 0.0912$ and $P(Y = 38) = 0.0912$. Because these probabilities are identical, I choose the even integer; thus, $\hat{y} = 38$.

• Suppose that $m = 100$ and $p = 0.615$. Then, $mp = 100(0.615) = 61.5$ is not an integer. The most likely value of $Y$ is either 61 or 62. With the help of the website calculator, I find that $P(Y = 61) = 0.0811$ and $P(Y = 62) = 0.0815$. Thus, $\hat{y} = 62$.

In each of the above examples we saw that the probability that a point prediction is correct is very small. As a result, scientists usually prefer a prediction interval. It is possible to create a one-sided prediction interval, but we will consider only two-sided prediction intervals.

We have two choices: using a Normal curve approximation or finding an exact interval. Even if you choose the exact interval, it is useful to begin with the Normal curve approximation.

**Result 14.3 (Predicting when $p$ is known.)** Let $Y$ denote the total number of successes in $m$ future Bernoulli trials. The Normal curve approximate prediction interval for $Y$ when $p$ is known is:

$$mp \pm z^* \sqrt{mpq}$$

(14.3)
where the value of $z^*$ is determined by the desired probability of getting a correct prediction interval. The value of $z^*$ is given in Table [12.1] on page 296. It is the same number that is used for the two-sided confidence interval for $p$.

I won’t indicate a proof of this result, other than to say it is the same derivation we used in Chapter 12 to find the approximate confidence interval for $p$, except that in the current situation we don’t need Slutsky’s theorem because the value of $p$ is known.

For our die example with $m = 600$, I want to have a prediction interval for which the probability it will be correct equals approximately 98%. From Table [12.1] I find that $z^* = 2.326$. The prediction interval is:

$$600(1/6) \pm 2.326 \sqrt{600(1/6)(5/6)} = 100 \pm 21.23 = [78.77, 121.23].$$

Let me make a couple of remarks concerning this answer. First, it makes no sense to predict that I will obtain a fractional number of 6’s; thus, I round-off my endpoints to obtain the closed interval $[79, 121]$. Second, I remember that this answer is not really an interval of numbers; more accurately, I predict that $Y$ will be one of the numbers in the set 79, 80, 81, . . . , 121. This more accurate statement is way too tedious for me; thus, I will abuse language and say that $[79, 121]$ is my approximate 98% prediction interval for $Y$.

You might be thinking: Hey, Bob, why did you use the Normal curve approximation; why not use exact binomial probabilities? Good question. If I am really serious about this prediction problem, I take my approximate answer, $[79, 121]$, as my starting point. I go to the binomial calculator website and find that for $m = 600$ and $p = 1/6$:

$$P(Y \leq 121) = 0.9894 \text{ and } P(Y \leq 78) = 0.0078.\,$$

Thus,

$$P(79 \leq Y \leq 121) = 0.9894 - 0.0078 = 0.9816,$$

which is very close to my target of 98% probability. Thus, I am really happy with the prediction interval $[79, 121]$.

### 14.2.2 When $p$ is Unknown

We now consider the situation in which $p$ is unknown. We will begin with point prediction. The first difficulty is that we cannot achieve our criterion’s goal: we cannot find the most probable value of $Y$. The most probable value, as we saw above, is at or near $mp$, but we don’t know what $p$ is. Thus, we won’t concern ourselves with point prediction.

Because $p$ is unknown, we cannot use Formula [14.3] as a prediction interval. In order to proceed, we need to add another ingredient to our procedure. We assume that we have past data from the process that will generate the $m$ future trials. We denote the past data as consisting of $n$ trials which yielded $x$ successes, giving $\hat{p} = x/n$ as our point estimate of the unknown $p$ and, as always, $\hat{q} = 1 - \hat{p}$. We will now use the Result [14.1] to derive a prediction interval for $Y$. 

340
It will be convenient to begin by defining a symbol \( r \), the ratio of the future number of trials to the past number of trials:

\[
\begin{align*}
r &= \frac{m}{n}. \\
\end{align*}
\] (14.4)

Note that if \( r \) is close to zero, then we are using a relatively large amount of past data to predict a relatively small number of future trials. Conversely, if \( r \) is large, then we are using a relatively small amount of past data to predict a relatively large number of future trials.

The algebra between this point and Result 14.4 is pretty intense. Unless you find that working through messy algebra improves your understanding, feel free to jump (skip, hop, dash) ahead to Result 14.4.

In Result [14.4] let \( W_1 = Y \), \( W_2 = X \) and \( b = r \). Thus,

\[
W = Y - rX.
\]

The mean of \( W \) is:

\[
\mu_W = \mu_Y - r\mu_X = mpq - (m/n)npq = mpq - mpq = 0
\]

and the variance of \( W \) is:

\[
\text{Var}(W) = \text{Var}(Y) + r^2 \text{Var}(X) = mpq + r(m/n)npq = mpq(1 + r).
\]

Let me make a couple of quick comments about this formula for the variance of \( W \). As you will see below, the larger the variance, the wider the prediction interval. We see that the variance is proportional to \( m \), the number of trials being predicted. This makes sense; more trials means more uncertainty. The variance is also proportional to \( (1 + r) \). To see why this makes sense, refer to my comments above immediately after Equation [14.4].

We can standardize \( W \):

\[
Z = \frac{W - \mu_W}{\sqrt{\text{Var}(W)}} = \frac{W - 0}{\sqrt{mpq\sqrt{1 + r}}} = \frac{Y - rX}{\sqrt{mpq\sqrt{1 + r}}}
\]

It can be shown that if \( m \) and \( n \) are both large and \( p \) is not too close to either 0 or 1, then probabilities for \( Z \) can be well approximated by the N(0,1) curve. Slutsky’s work can also be applied here; the result is:

\[
Z' = \frac{Y - rX}{\sqrt{rX[1 - (X/n)]}\sqrt{1 + r}}
\]

It can be shown that if \( m \) and \( n \) are both large and \( p \) is not too close to either 0 or 1, then probabilities for \( Z' \) can be well approximated by the N(0,1) curve. As a result, using the same algebra we had in Chapter 12 (the names have changed) we can expand \( Z' \) and get the following prediction interval for \( Y \).

**Result 14.4 (Predicting when \( p \) is unknown.)** The formula below is the approximate prediction interval for \( Y \), the total number of successes in \( m \) future Bernoulli trials, when \( p \) is unknown. In this formula, \( r \) is given in Equation [14.4]; \( x \) is the observed number of successes in the past data.
of \(n\) trials; \(\hat{q} = (n - x)/n\) is the proportion of failures in the past data; and \(z^*\) is determined by the desired probability of the interval being correct. The relationship between \(z^*\) and the desired probability is given in Table 12.1 on page 296.

\[
rx \pm z^*\sqrt{rx\hat{q}\sqrt{1 + r}} = rx \pm z^*\sqrt{rx(1 + r)\hat{q}}.
\]  

(14.5)

I will illustrate the use of this formula with real data from basketball.

On page 274 I introduced you to the data Katie Voigt collected and shared with me. I will use some of Katie’s data to illustrate the current method.

I will put myself in time on day 3 of Katie’s study before she collected the day’s data. I want to use the combined data from Katie’s first two days of shooting to predict the number of successes that she will achieve on her \(m = 100\) trials on day 3.

Let me explicitly review the necessary assumptions. Katie’s 300 shots on days 1–3 are Bernoulli trials with an unknown value of \(p\). Note, in particular, that I assume that her future trials are governed by the same process that generated her past data. Now, let’s get to Katie’s data.

On day 1, Katie obtained 56 successes and on day 2 she obtained 59 successes. Combining these days, the past data consist of \(n = 200\) trials with \(x = 56 + 59 = 115\) successes. I will use these data to obtain the approximate 95% prediction interval for \(Y\), the number of successes she will obtain on day 3. I will use Formula 14.5 to obtain my answer. Thus, I need to identify the values of the various symbols in the formula.

\[
r = m/n = 100/200 = 0.5; \ x = 115; \ z^* = 1.96; \ \text{and} \ \hat{q} = (200 - 115)/200 = 0.425.
\]

Next, I substitute these values into Formula 14.5 and obtain:

\[
0.5(115) \pm 1.96\sqrt{0.5(115)(0.425)}\sqrt{1 + 0.5} = 57.5 \pm 1.96\sqrt{24.4375}\sqrt{1.5} = 57.5 \pm 11.86 = [45.64, 69.36],
\]

which I will round to [46, 69].

This is a very wide interval. (Why do I say this? Well, in my opinion, a basketball player will believe that making 46 out of 100 is significantly different than making 69 out of 100.) A great thing about prediction (not shared by estimation or testing) is that we find out whether our answer is correct. It is no longer the case that only Nature knows! In particular, when Katie attempted the \(m = 100\) future shots, she obtained \(y = 66\) successes. The prediction interval is correct because it includes \(y = 66\).

You should note that there is no exact solution to this problem. Even a simulation experiment is a challenge. I will discuss how a simulation experiment is performed for a limited situation motivated by Katie’s data.

In order to do a simulation study we must specify three numbers: \(m, n\) and \(p\). (Even though the researcher does not know \(p\), Nature, the great simulator, must know it.) And, as we shall see, it is a two-stage simulation.

To be specific, I will simulate something similar to Katie’s problem. I will take \(m = 100, n = 200\) and \(p = 0.60\). (Of course, I don’t know Katie’s \(p\), but 0.60 seems to be not ridiculous.) A single run of the simulation experiment is as follows.
1. Simulate the value of $X \sim \text{Bin}(200, 0.60)$, our simulated past data for Katie.

2. Use our value of $X$ from step 1 to estimate $p$ and then compute the 95% prediction interval for $Y$, remembering to use the interval for $p$ unknown.

3. Simulate the value of $Y \sim \text{Bin}(100, 0.60)$ and see whether or not the interval from step 2 is correct.

I performed this simulation with 10,000 runs and obtained 9,493 (94.93%) correct prediction intervals! This is very close to the nominal (advertised) rate of 95% correct.

Before we leave this section, I will use Katie’s data to obtain one more prediction interval. I will put myself in time on day 9 of Katie’s study before she collected the day’s data. I want to use the combined data from Katie’s first eight days of shooting ($n = 800$) to predict the number of successes that she will achieve on her $m = 200$ trials on days 9 and 10 combined ($m = 200$).

Let me explicitly review the necessary assumptions. Katie’s 1,000 shots on days 1–10 are Bernoulli trials with an unknown value of $p$. Note, in particular, that I assume that her future trials are governed by the same process that generated her past data. Now, let’s get to Katie’s data.

On days 1–8, Katie obtained a total 505 successes. Thus, the past data consist of $n = 800$ trials with $x = 505$ successes. I will use these data to obtain the approximate 95% prediction interval for $Y$, the number of successes she will obtain on days 9 and 10 combined. I will use Formula 14.5 to obtain my answer. Thus, I need to identify the values of the various symbols in the formula.

$$r = m/n = 200/800 = 0.25; x = 505; z^* = 1.96; \text{ and } \hat{q} = (800 - 505)/800 = 0.369.$$ 

Next, I substitute these values into Formula 14.5 and obtain:

$$0.25(505) \pm 1.96\sqrt{0.25(505)(0.369)\left(1 + 0.25\right)} = 126.25 \pm 1.96\sqrt{46.5862}\sqrt{1.25} =$$

$$126.25 \pm 14.96 = [111.29, 141.21],$$

which I will round to [111, 141].

On days 9 and 10 combined, Katie achieved $y = 130$ successes. The prediction interval is correct because it includes $y = 130$.

### 14.3 Predicting for a Poisson Process

Compared to my above work on the binomial distribution, this section will be quite brief. I will consider only one of several possible problems, namely the following. I plan to observe a Poisson Process with unknown rate $\lambda$ for $t_2$ units of time. I have past data on the same process that gave $x$ successes in an observational period of $t_1$ units of time. After I derive the prediction interval, Formula 14.7, I will illustrate the method with the Baltimore homicide data introduced in Chapter 13 in Table 13.2 on page 331. In particular, I will put myself in time at the end of 2004, which means that my past data are for $t_1 = 2$ years and I will want to predict the total number of homicides for the year 2005, giving me a future observation of $t_2 = 1$ year.

343
Returning to the general problem, the observed number of successes in the past data, \( x \), is the observed value of a random variable \( X \) which has distribution \( \text{Poisson}(\lambda t_1) \). I want to predict the value of \( Y \) which has distribution \( \text{Poisson}(\lambda t_2) \).

For the binomial problem with \( p \) unknown, recall that the ratio \( r \), defined in Equation \[14.4\], played a key role in our prediction interval. We need a similar ratio for the current Poisson problem:

\[ r' = \frac{t_2}{t_1}. \] (14.6)

Note that, similar to the binomial problem, this is the ratio of the length of future observation of the process to the length of past observation of the process. I put a prime symbol on the notation for the current ratio to avoid confusion with the binomial problem’s ratio.

The algebra between this point and Result \[14.5\] is pretty intense. Unless you find that working through messy algebra improves your understanding, feel free to jump ahead to Result \[14.5\].

In Result \[14.1\] let \( W_1 = Y \), \( W_2 = X \) and \( b = r' \). Thus,

\[ W = Y - r'X. \]

The mean of \( W \) is:

\[ \mu_W = \mu_Y - r'\mu_X = \lambda t_2 - (t_2/t_1)\lambda t_1 = \lambda t_2 - \lambda t_2 = 0. \]

and the variance of \( W \) is:

\[ \text{Var}(W) = \text{Var}(Y) + (r')^2 \text{Var}(X) = \lambda t_2 + r'(t_2/t_1)\lambda t_1 = \lambda t_2(1 + r'). \]

We can standardize \( W \):

\[ Z = \frac{W - \mu_W}{\sqrt{\text{Var}(W)}} = \frac{W - 0}{\sqrt{\lambda t_2(1 + r')}} = \frac{Y - r'X}{\sqrt{\lambda t_2(1 + r')}}. \]

It can be shown that if \( t_1 \) and \( t_2 \) are both large and \( \lambda \) is not too close to 0, then probabilities for \( Z \) can be well approximated by the \( \text{N}(0,1) \) curve. Slutsky’s work can also be applied here. The idea is that in the denominator we replace the unknown \( \lambda \) by \( X/t_1 \). This seems sensible because the mean of \( X \) is \( \lambda t_1 \). The result is:

\[ Z' = \frac{Y - r'X}{\sqrt{r'X(1 + r')}}. \]

It can be shown that if \( t_1 \) and \( t_2 \) are both large and \( \lambda \) is not too close to 0, then probabilities for \( Z' \) can be well approximated by the \( \text{N}(0,1) \) curve. As a result, using the same algebra we had in Chapter 12 (the names have changed) we can expand \( Z' \) and get the following prediction interval for \( Y \).

**Result 14.5 (Predicting for a Poisson Process.)** A Poisson Process will be observed for \( t_2 \) units of time; let \( Y \) denote the number of successes that will occur. The process has been observed previously for \( t_1 \) units of time, during which \( x \) successes were counted. In the formula below, \( r' \) is given by Equation \[14.6\], and \( z^* \) is determined by the desired probability of the interval being correct.
The relationship between \( z^* \) and the desired probability is given in Table 12.1 on page 296. The formula below is the approximate prediction interval for \( Y \) when the rate of the Poisson Process is unknown.

\[
r'x \pm z^* \sqrt{r'x(1 + r')}
\] (14.7)

I will illustrate the use of this formula with the Baltimore homicide data, presented in Table 13.2 on page 331. The past data consist of the \( t_1 = 2 \) years, 2003 and 2004. The total number of homicides, \( x \), in those two years is \( 270 + 276 = 546 \). The future period of interest consists of \( t_2 = 1 \) year, 2005. Thus, \( r' = 1/2 = 0.5 \). I want to have an interval for which the probability it will be correct is approximately 95%; thus, my choice for \( z^* \) is 1.96. Substituting these values into Formula 14.7, I get:

\[
0.5(546) \pm 1.96 \sqrt{0.5(546)(1 + 0.5)} = 273 \pm 39.66 = [233.34, 312.66],
\]

which I round to [233, 313]. In words, at the beginning of 2005, using the 2003 and 2004 data, my 95% prediction interval for the number of homicides in Baltimore in 2005 is [233, 313]. This is a very wide interval! I imagine that if—after homicide totals of 270 and 276 in 2003 and 2004—the number declined to 233 or increased to 313 in 2005, many people would conclude that something must have changed. Our interval shows that such a large change is within the bounds of Poisson variation.

The actual number of homicides in 2005 turned out to be \( y = 269 \); thus, the prediction interval is correct.

As with Bernoulli trials with \( p \) unknown, there is no exact formula for a prediction interval. The only way to evaluate the quality of the Normal curve approximation contained in our prediction interval is to perform a simulation experiment. A simulation study is a challenge. For simplicity, I will limit my presentation of how a simulation experiment is performed to the context of the above Baltimore homicide example.

In order to do a simulation experiment we must specify one number, the rate of the Poisson Process. For the Baltimore process, I will specify that \( \lambda = 270 \) in my computer simulation. Each run of the computer simulation will consist of the following three steps:

1. Simulate the value of \( X \sim \text{Poisson}(2\lambda = 540) \).
2. Use our value of \( X \) from step 1 to compute the 95% prediction interval for \( Y \).
3. Simulate the value of \( Y \sim \text{Poisson}(\lambda = 270) \) and see whether or not the interval from step 2 is correct.

I performed this simulation with 10,000 runs and obtained 210 intervals that were too large and 253 intervals that were too small. Thus,

\[
10,000 - (210 + 253) = 9,537 — \text{or } 95.37\% - \text{of the intervals were correct.}
\]

This is a very good agreement between simulation results and the nominal probability of being correct, 95%.
14.4 Summary

This chapter introduces the notion of prediction, in the context of Bernoulli trials or a Poisson Process.

The first situation is that we plan to observe $m$ future Bernoulli trials and we want to predict the total number of successes, $Y$, that will be obtained. There are two situations of interest: $p$ known and $p$ unknown.

For the case in which $p$ is known, we first consider the point prediction $\hat{y}$. We adopt the criterion that we want to maximize the probability that the point prediction will be correct; i.e., we want to choose $\hat{y}$ to maximize $P(Y = \hat{y})$. The result is:

- If the mean of $Y$, $mp$, is an integer, then $\hat{y} = mp$ is the unique maximizer of the probability of obtaining a correct point prediction.
- If the mean of $Y$, $mp$, is not an integer, then calculate the probability of $Y$ equaling each of the two integers nearest to $mp$. Whichever of these two integers has the larger probability of occurring is the point prediction. If they have the same probability of occurring, then I arbitrarily decide that the even integer will be the point prediction.

We looked at several examples and found that in every case the probability that the point prediction will be correct is quite small. Thus, the somewhat tedious steps outlined above for finding the point prediction are arguably somewhat (no pun intended) pointless. As a result, we turn our attention to finding a prediction interval for the value of $Y$.

The first decision is to select the desired probability that the prediction interval will be correct. The popular choices—80%, 90%, 95%, 98% and 99%—are familiar from our work on confidence interval estimation. The approximate prediction interval for $Y$ is given in Formula 14.3, reproduced below:

$$mp \pm z^* \sqrt{mpq}.$$  

In this formula, the value of $z^*$ depends on the desired probability and is the same number we used for approximate confidence intervals.

There are three comments to remember about this approximate prediction interval. First, because we are predicting the number of successes—which must be an integer—the endpoints of the interval should be rounded-off to their nearest integers. For example, if the formula yields $[152.27, 191.33]$ we should round this to $[152, 191]$. Second, although for convenience I will always refer to the answer as an interval, it really isn’t. For example, if we predict that $Y$ will be in the interval $[152, 191]$, we actually are predicting that $Y$ will take on one of the values:

$$152, 153, 154, \ldots, 191.$$  

Third and finally, if $m \leq 1,000$, once we have a prediction interval, we should use the binomial calculator website to obtain the exact probability that it will be correct. (Recall that for $m > 1,000$, the calculator website does not yield exact binomial probabilities.)

Now we turn to the scientifically more interesting problem of finding a prediction interval for $Y$ when $p$ is unknown. This new problem cannot be solved unless we add a new ingredient to our
set-up. We assume that we have observed $n$ past Bernoulli trials from the same process that will generate the $m$ future Bernoulli trials; in particular, the $p$ for the past is the same as the $p$ for the future and the future is statistically independent of the past. The approximate prediction interval for $Y$ is given in Formula [14.5] reproduced below:

$$rx ± z^* \sqrt{rx\hat{q} \sqrt{1 + r}} = rx ± z^* \sqrt{rx(1 + r)\hat{q}}.$$  

In this formula, $r = m/n$, $x$ is the total number of successes in the $n$ past trials; $z^*$ is the same as it was for $p$ known; and $\hat{q} = (n - x)/n$ is the proportion of failures in the $n$ past trials.

Lastly, we learned how to obtain a prediction interval in the context of a Poisson Process. Assume that there is a Poisson Process with unknown rate $\lambda$. The process has been previously observed for $t_1$ units of time, yielding a total of $x$ successes. The process will be observed for $t_2$ units of time in the future and will yield $Y$ successes. The approximate prediction interval for $Y$ is given in Formula [14.7] reproduced below

$$r'x ± z^* \sqrt{r'x(1 + r')}.$$  

In this formula, $r' = t_2/t_1$; and $z^*$ is the same as it was for both of our earlier prediction intervals.

Finally, I want to comment on Formula [14.5] In particular, I want to explain why I present it in two different ways:

$$rx ± z^* \sqrt{rx\hat{q} \sqrt{1 + r}} \text{ and } rx ± z^* \sqrt{rx(1 + r)\hat{q}}.$$  

I will refer to these as the outside and inside versions, respectively, because the former has the term $(1 + r)$ within its own radical sign outside the first radical sign; obviously, inside has the term $(1 + r)$ inside its only radical sign. What possible reason do I have for doing this?

The term $rx$ is equal to

$$(m/n)x = m(x/n) = m\hat{p}.$$  

Thus, the outside version is equivalent to

$$m\hat{p} ± z^* \sqrt{m\hat{p}\hat{q} \sqrt{1 + r}}.$$  

This expression makes clear the relationship between the situations when $p$ is unknown, Formula [14.5] and $p$ is known, Formula [14.3] Namely, in the $p$ is known formula we replace the unknown $p$ and $q$ by their estimates from the previous data. To make this work, we need to include the correction term $\sqrt{1 + r}$. Somewhat whimsically, I like to think of this correction term as representing the cost of ignorance.

The inside version gives insight of the connection between the binomial and Poisson formulas for prediction. Recalling that I use $r$ [r'] to denote the ratio of future to past for the binomial [Poisson], the prediction formulas for binomial and Poisson are identical except that the binomial formula contains the term $\hat{q}$ under its radical sign.

The derivations of Formulas [14.5] and [14.7] involve some intense algebra and the use of the fairly advanced Result [14.1]. I include these things in these Course Notes for completeness; feel free to ignore them if you don’t find excessive algebra to be helpful.
14.5 Practice Problems

1. The entry

http://en.wikipedia.org/wiki/Mendelian_inheritance

in Wikipedia provides a colorful illustration of the results of a dihybrid cross involving brown or white cats with short or long tails. According to Mendelian inheritance, the probability that a cat created by such a cross will be brown with a short tail is $\frac{9}{16}$. For the sake of this question, let’s all agree that Mendelian inheritance is correct.

I plan to observe $m = 320$ future cats created by such a cross and I want to obtain a 98% prediction interval for the total number of these cats that will have brown fur and a short tail.

2. Refer to the previous problem. According to Mendelian inheritance, the probability that a cat created by such a cross will have a long tail—regardless of fur color—is 0.25.

I plan to observe $m = 400$ future cats created by such a cross and I want to obtain a 99% prediction interval for the total number of these cats that will have a long tail.

3. During Michael Jordan’s first season with the Chicago Bulls, 1984–85, he made 630 out of 746 free throw attempts. In his last season with the Chicago Bulls, 1997–98, he made 565 out of 721 free throw attempts. Use the data from his first season to predict the number of free throws he would make during his last season. Use the 99% prediction level.

4. The purpose of this example is to show you the folly of using a small amount of data to predict a large amount of the future. On day 1, Katie obtained 56 successes in 100 trials. Use these data to obtain the 98% prediction interval for the $m = 900$ future trials on days 2–10 combined.

After day 10, it was determined that Katie had made 579 of her shots on days 2–10 combined. Comment on your prediction interval.

5. Data for hockey player Wayne Gretzky were presented on page 333 in a Practice Problem. Use his data from the 1981–82 season—92 goals in 80 games—to predict his number of goals—71 in 80 games—in the 1982–83 season. Assume that we have a Poisson Process with unknown rate of $\lambda$ goals per game. Calculate the 95% prediction interval.

6. Refer to the previous problem. In the two seasons 1981–83 combined, Gretzky scored a total of $92 + 71 = 163$ goals. Given that Gretzky would play 74 games during the 1983–84 season, use the data from 1981–83 to obtain the 95% prediction interval for the number of goals he would score in 1983–84.

Given that Gretzky scored 87 goals in 1983–84, comment on your prediction interval.
14.6 Solutions to Practice Problems

1. This is a standard problem with \( p = 9/16 \) known and \( m = 320 \). The desired probability, 98%, gives \( z^* = 2.326 \). Using Formula [14.3] we get

\[
320 \left( \frac{9}{16} \right) \pm 2.326 \sqrt{320 \left( \frac{9}{16} \right) \left( \frac{7}{16} \right)} = 180 \pm 20.64 = [159.36, 200.64],
\]

which I round to [159, 201].

I go to the binomial calculator website and obtain

\[
P(Y \leq 201) = 0.9926 \quad \text{and} \quad P(Y \leq 158) = 0.0079.
\]

Thus,

\[
P(159 \leq Y \leq 201) = 0.9926 - 0.0079 = 0.9847.
\]

This is a bit larger than the target of 98%. Let’s see if we can do better. The idea is to make the prediction interval narrower, but not much narrower. Too much narrower and the probability of being correct would fall below the target of 98%.

First, I try the interval [160, 201]; i.e., I increase the lower bound by one and leave the upper bound alone. I find that its probability of including \( Y \) is 0.9819 (details not given; check if you need practice with the binomial calculator).

Next, I try [159, 200] and find that its probability of including \( Y \) is 0.9820.

Finally, I try [160, 200]; i.e., I take the original approximate interval, increase the lower bound by one and decrease the upper bound by one. I find that its probability of including \( Y \) is 0.9792. This is a bit smaller than the target probability, but it’s actually the closest of the three probabilities.

Any one of these modifications of the approximate interval is fine; I (very slightly) prefer the interval [160, 200] because of its symmetry around the point prediction 180.

Note: Usually at this time I tell you the value of \( Y \) and we can see whether the prediction interval is correct. Sorry, but acquiring 320 cats is too much for this dedicated author.

2. This is a standard problem with \( p = 0.25 \) known and \( m = 400 \). The desired probability, 99%, gives \( z^* = 2.576 \). Using Formula [14.3] we get

\[
400 \left( \frac{0.25}{2} \right) \pm 2.576 \sqrt{400 \left( \frac{0.25}{2} \right) \left( \frac{0.75}{2} \right)} = 100 \pm 22.31 = [77.69, 122.31],
\]

which I round to [78, 122].

I go to the binomial calculator website and obtain

\[
P(Y \leq 122) = 0.9946 \quad \text{and} \quad P(Y \leq 77) = 0.0039.
\]

Thus,

\[
P(78 \leq Y \leq 122) = 0.9946 - 0.0039 = 0.9907.
\]

This is very close to the target of 99%. Thus, I won’t bother to see what happens if I change either endpoint of the interval.
3. This is prediction with $p$ unknown. The past data give $x = 630$, $n = 746$ and $\hat{q} = 116/746 = 0.155$. The future number of trials is $m = 721$, giving $r = 721/746 = 0.9665$. (Note: A weakness with this method is that at the beginning of the 1997–98 season, nobody knew the eventual value of $m$. One way around this was to restate the problem as “Of his first $m = 500$ free throws during the season, predict his number of successes.” It is possible that the number $m$ is somehow related to how well he was shooting, but I will ignore this possibility.)

I will substitute these values into the prediction formula, Formula 14.5, noting that 99% gives $z^* = 2.576$:

$$rx \pm z^* \sqrt{rx \hat{q} \sqrt{1 + r}} = 608.90 \pm 2.576 \sqrt{608.90(0.155) \sqrt{1 + 0.9665}} = 608.90 \pm 35.09 = [573.81, 643.99],$$

which I round to $[574, 644]$. This prediction interval is too large, because $y = 565$. Perhaps using data from 13 years earlier was a bad choice for the past data.

In 1997–98, Jordan had his lowest free throw success percentage of his career; even worse than his two (misguided) later years in Washington.

4. This is prediction with $p$ unknown. The past data give $x = 56$, $n = 100$ and $\hat{q} = 44/100 = 0.44$. The future number of trials is $m = 900$, giving $r = 900/100 = 9$.

I will substitute these values into the prediction formula, Formula 14.5, noting that 98% gives $z^* = 2.326$:

$$rx \pm z^* \sqrt{rx \hat{q} \sqrt{1 + r}} = 504 \pm 2.326 \sqrt{504(0.44) \sqrt{1 + 9}} = 504 \pm 109.54 = [394.46, 613.54],$$

which I round to $[394, 614]$. This extremely wide prediction interval is correct, because $y = 579$.

5. We have $t_1 = t_2 = 80$ games; thus $r' = 80/80 = 1$. The 95% level gives us $z^* = 1.96$. The observed value of $X$ is 92. We substitute this information into Formula 14.7 and obtain:

$$r'x \pm z^* \sqrt{r'x(1 + r')} = 92 \pm 1.96 \sqrt{92(1 + 1)} = 92 \pm 26.6 = [65, 119],$$

after rounding. This very wide interval is correct because it includes $y = 71$. I opine that many hockey fans would interpret the change from 92 to 71 goals as significant, but it falls within the range of Poisson variation.

6. We have $t_1 = 80 + 80 = 160$ games and $t_2 = 74$ games; thus $r' = 74/160 = 0.4625$. The 95% level gives us $z^* = 1.96$. The observed value of $X$ is 163, giving us $r'x = 75.39$. We substitute this information into Formula 14.7 and obtain:

$$r'x \pm z^* \sqrt{r'x(1 + r')} = 75.39 \pm 1.96 \sqrt{75.39(1 + 0.4625)} = 75.39 \pm 20.6 = [55, 96],$$

after rounding. This very wide interval is correct because it includes $y = 87$.  

350
14.7 Homework Problems

1. Refer to Practice Problems 1 and 2. Let $A$ be the event that the cat has white fur and let $B$ be the event that the cat has a long tail.

   (a) Define a success to be that event ($A$ or $B$) occurs. Remember, or means and/or. According to Mendelian inheritance, what is the probability of a success?

   (b) For $m = 160$ cats created by such a cross, calculate the 95% prediction interval for the number of successes.

   (c) Use the binomial calculator website to determine the exact probability that your interval in (b) will be correct.

2. On days 1–19, Katie made 1,227 out of 1,900 attempted shots. Use these 1,900 observations to calculate the 99% prediction interval for the number of shots, out of $m = 100$, that she would make on day 20.

   Given that Katie made $y = 71$ shots on day 20, comment on your prediction interval.

3. Refer to the data on traffic accidents in Table 13.3 on page 336. Use the combined data from years 2005–2008, 405 crashes with bikes, to obtain the 95% prediction interval for the number of bike crashes in 2009.

   Given that the number of bike crashes in 2009 was $y = 115$, comment on your prediction interval.