





Baseball Statistics


1 A sea of numbers


 Question 1: Make a histogram of the number of home runs (HR). Describe the shape of the distribution.

 Question 2: Just for fun, let's see who hit the most home runs in 2002. The variable HR records the number hit for each player, whose ID is stored in `playerID`:

```
> playerID[HR == max(HR)]  
  
[1] rodrial01  
1217 Levels: abbotpa01 abernbr01 abreubo01 acevejo01 aceveju01 ... zitoba01
```

The answer is the cryptically coded `rodrial01` which you may recognize as Alex Rodriguez of the Texas Rangers. (The real names of the player ID's is available from the original dataset.)

 Question 3: Make a histogram of the number of doubles (DOUBLE). Describe the shape of the distribution. Is it the same “shape” as the home run distribution? Who hit the most doubles this year?

 Question 4: The number of singles is not included. You can find it by subtracting the number of doubles, triples and home runs from the number of hits. Do this, then describe the shape of its distribution. Is it like that of the home runs or doubles? Who hit the most singles this year?

2 Measures of Batting success

A key statistic for baseball is the batting average. According to the “Baseball Almanac” (<http://www.baseball-almanac.com/>) this is defined as the number of hits divided by the number of at bats or in the variable names, $AVG = H/AB$.

```
> AVG = H/AB  
  
> hist(AVG, main = "histogram of batting averages")
```

Figure 1 shows a distribution that is skewed right, but not so much.

This batting average statistic is almost as old as the game (the late 1800's). As at bats do not include walks, hit by pitches and other events, a slightly better statistic to compare is the “On Base Percentage”. The formula for this is more complicated but not much so:

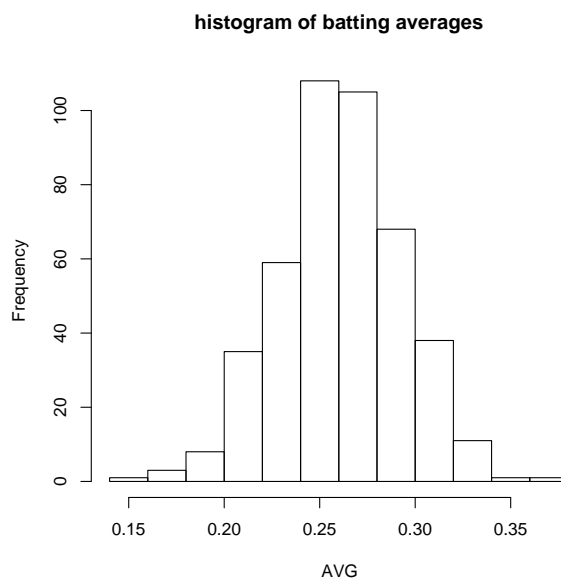




Figure 1: Histogram of batting averages for 2002 Major League Baseball season


```
> OBP = (H + BB + HBP)/(AB + BB + HBP + SF)
```


That is we add base on balls and hit by pitches to the hits column and divide by all our at bats plus these and sacrifice flies. A histogram (Figure ??) of our data shows one outlier (Barry Bonds) and a fairly symmetric distribution otherwise.

 Question 5: Define `OBP` and produce a histogram of the variable. Describe its shape. Is it similar to the shape of the batting average variable?

 Question 6: The shape of two distributions can be compared with boxplots. Sometimes it helps to use the `scale()` function first so that only z-scores are compared. This focuses on the shape, as the center and spread has been standardized. Produce this boxplot and compare shapes as in the previous exercise.

```
> boxplot(AB, OBP, names = c("AB", "OBP"))
```

 Question 7: Make a histogram of `OPB-AVG`. What is this measuring? Describe the shape of this new variable. Can this variable be negative?

 Question 8: An alternative to the batting average (`AVG`) and the on base percentage (`OBP`) is the Slugging Percentage. This is defined by:

$$SLG = (1 \cdot \text{SINGLE} + 2 \cdot \text{DOUBLE} + 3 \cdot \text{TRIPLE} + 4 \cdot \text{HR}) / AB$$

Compute `SLG`, make a histogram and discuss its shape relative to the shape of the batting average and the on base percentage distributions.

3 Bernoulli trials

Baseball can be kind of complicated when all the possibilities are taken account of. To simplify, we'll assume a batter either has a success (gets on base, etc.) or a failure. That is, there is no gray area.

As fans, we expect players with a high on base percentage to have a better “chance” of success (getting on base). This makes sense, they have done this previously in the past. What do we mean by “chance” though? For this we need to have an understanding of probability.

The simplest way to think about the “chance” that a batter gets on base, is to imagine somebody tossing a coin to determine this: heads they get on base (a success), and tails they don't (a failure). In our case the coin is a little weighted and should be heads with probability p and tails with probability $1 - p$.

The model using a coin toss implies a few things that may or may not be appropriate for the real-life scenario. In particular, each time we toss the coin we assume:

- the same probability of heads (called p),
- the upcoming outcome does not depend on any or all of the previous outcomes (independence).



Question 9: Is it a reasonable assumption to assume that each time a player is batting they have the same chance of getting on base? Why or why not?



Question 10: Does it make sense that each time a player is batting their previous attempts do not influence the outcome of the current attempt? Why or why not?



Question 11: Someone is “due” to get on base if they haven't had a success for quite some time. Does this fit with the model?

Even if you have doubts about the assumptions, we can still construct a *probability model* that incorporates the assumptions and later decide if the model is a “good” one for analyzing or predicting our data. The model described above – tossing a coin with the associated assumptions of independence and identical trials – is an example of a *Bernoulli trials* model.

To illustrate we can do a simple simulation. Suppose $p = 0.3$ and our batter bats 10 times here are the results (1 is a head, 0 a tail)


```
> p = 0.3
> x = sample(c(0, 1), 10, replace = TRUE, prob = c(1 - p, p))
> x

[1] 0 1 0 0 1 0 0 0 0 1
```

Read 1 as a hit and 0 as an out. We used the `sample()` command to select 10 times from the values of 0 or 1. We need to specify that we replace the selected value so each time we pick

from the same set of outcomes and we need to specify the probabilities of selecting 0 or 1 in this case.

If you try this, chances are your sample will be completely different, yet similar in some ways. It is hard to compare samples of Bernoulli trials directly, usually we summarize them, in particular knowing the number of 1's (or successes) and the number of trials tells us a lot about the sequence of trials (not the order though).

 **Question 12:** Type in the above commands and find a sample of 10 times batting. Is your sequence identical to the one above? How many times did your batter get on base? Was it also 0 times?

4 Binomial distributions

The number of successes in next 10 times at bat is of course a random number until the next 10 at bats actually happen. To describe a random number, or variable, at best we can specify the range and probabilities for the possible values. We may also use numeric summaries such as a mean or standard deviation, or in this case, to summarize this distribution.

For concreteness, we define a *Binomial random variable* with n trials and success probability p to be the number of successes in n Bernoulli trials with success probability p . In the example above, we have a Binomial random variable with $p = 0.3$ and $n = 10$.

Due to the reasonably simple nature of things, we can specify explicitly the probability that our Binomial random variable takes on some value. We do this with its distribution. In particular, if X is a binomial random variable with success probability p and n trials then for any k between 0 and n we have the probabilities

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Where

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

is the *binomial coefficient*.

In particular, if we have $n = 10$, $p = 0.3$ and $X = 3$. The probability of this is

$$P(X = 3) = \frac{10!}{3!7!} (0.3)^3 (0.7)^7 = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} (0.3)^3 (0.7)^7$$

As you suspect, we can simplify our work with the computer:

```
> dbinom(3, 10, 0.3)
```

```
[1] 0.2668279
```

The R software has a built in function, `dbinom`, which returns the distribution of the binomial.

We plot all the values at once in Figure 2 with

```
> n = 10
> p = 0.3
> plot(0:n, dbinom(0:n, n, p), type = "h")
```

(The argument `type="h"` forces a plot with lines not points.)

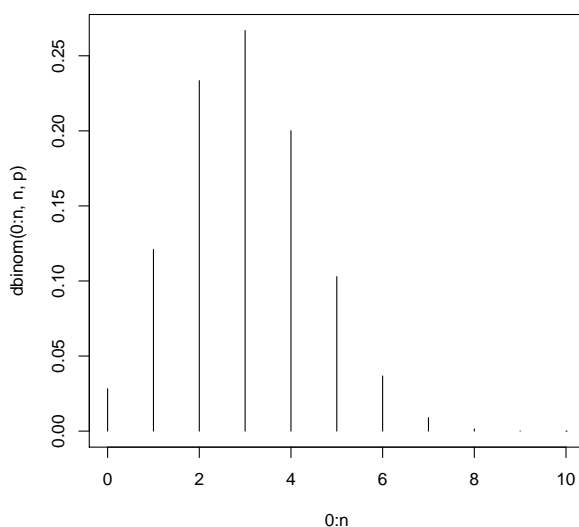


Figure 2: Distribution of binomial random variable with $n = 10$, $p = 0.3$

We see that the distribution peaks at a value of 3 as we might expect, but that other values are certainly possible, even the remote possibility of getting 10 successes.

Suppose a player has 200 at bats and is batting .320. If we suspect that the player is a “300” hitter (has a value of $p = 0.3$) is their current average “unusual”? How can we answer this?

We can compute the probability of having 64 successes in 200 trials with $p = 0.3$. (Why 64?) It is

```
> dbinom(64, 200, 0.3)
```

```
[1] 0.05002625
```

That’s small. Yet, we expect it to be small. The maximum value for any given number of successes is simply

```
> max(dbinom(0:200, 200, 0.3))
```

```
[1] 0.06146172
```

So its close to the max possible. A better way to compare how likely something is for the binomial is to ask about a range of possibilities. In this case we might want to know the probability of 64 *or more* successes in 200 trials.

```
> sum(dbinom(64:200, 200, 0.3))
```

```
[1] 0.2921207
```

If X is binomial, questions such as $P(X > 64)$ can usually be answered in terms of probabilities specified by $P(X \leq k)$. For instance, $P(X > 64) = 1 - P(X \leq 63)$. The `pbinom()` function computes $P(X \leq k)$, or when given the argument `lower.tail=FALSE` will find $P(X > k)$.

For our example, we have $P(X \geq 64)$ so is answered by $P(X > 63)$:

```
> pbinom(63, 200, 0.3, lower.tail = FALSE)
```

```
[1] 0.2921207
```

This value indicates that it is not unusual for a “300” hitter to average “320” over their first 200 at bats.

4.1 z-scores

One way to measure unusual is with the number of standard deviations away from the mean (the z-score). Recall, the z-score for a data point using the population parameters is the number of standard deviations away from the mean:

$$\text{z-score} = \frac{x_i - \mu}{\sigma}.$$

For the binomial distribution it is known that

$$\begin{aligned} \mu &= np && \text{for the binomial distribution} \\ \sigma &= \sqrt{np(1-p)}. \end{aligned}$$


What is an “unusual” z-score? For bell shaped distributions—such as the binomial for most cases—let’s say that z-scores more than 2 or less than -2 are unusual. (These happen about 5% of the time.)


So in our example, we are


```
> n = 200
> p = 0.3
> (64 - n * p)/sqrt(n * p * (1 - p))
```

```
[1] 0.6172134
```

standard deviations from the mean. This is not an unusual z-score.

 **Question 13:** Is it “unusual” for a “250” hitter ($p = 0.25$) to average “300” over their first 100 at bats? How about over 600 at bats? Why is there a difference?

 Question 14: What is the probability that a $p = 0.3$ hitter has 0 successes in 4 at bats?


 Question 15: If a player is a “300” hitter and has 600 at bats during the season, what is the range of possibilities for their batting average so that it is not too “unusual”. By this, so that they are in the middle 80% of the distribution? (Answer this using the quantile finding function `qbinom`)

5 Fun with the binomial

How often will things happen in the sport of baseball? For example, a no-hitter, a perfect game, a .400 season? We can get a good idea using the binomial model, and the following fact:

If you repeat independently a chance event which has probability p of success the expected number of times until the first success is $1/p$

So, if you toss a coin, you expect to have the first heads by the second toss. If you roll a die, you expect to have the first six after six times.

 Question 16: A no hitter. In a game of baseball, a team needs to record 27 outs. If each time a batter is up, they have a chance $p = .260$ of getting a hit, what is the probability of no hit in 27 independent chances? How many games do you expect to play before the first no hitter?


If there are 14 American League teams and 16 National League teams and each plays 162 games, then there is a total of


```
> (14 + 16) * 162
```

```
[1] 4860
```

chance at a no hitter. (Neglect the extra innings games, the chance of two no hitters, etc. which could throw this off.)

Would you expect a no hitter each season? Explain

 Question 17: A perfect game is a no hitter in which no opposing player even makes it to first base even by error, walk or other means. Assume the chance a player gets on base by any means is estimated by the mean of the on base percentage, $p = .330$. Compute the probability of a perfect game, how many games before a perfect game would be expected, and how many perfect games there would have been in the last 100 years of baseball with approximately 4000 games per season. (There have been 15 in the last 100 years.)

 Question 18: What is the probability of a “400 season”? That is, a players batting average being .400 or more.

Suppose the players batting averages are distributed by the normal distribution with mean $\mu = .2608$ and standard deviation $\sigma = .032$. What is the probability the normal

distribution with these parameters assigns to the value 0.400 or more? Use `pnorm()` with the argument `lower.tail=FALSE`.

How many players would you expect to play a season before having a “400 season”?

There are about 400 players a year with enough at bats to qualify to hit 400. How many seasons would pass before we expect to see the first “400 season”?

There have been 14 “400 season” in the last 100 years. Is this consistent with your calculation? If not, does this suggest the normal distribution is the wrong distribution to describe the player’s batting averages?

5.1 The normal approximation

Before the ubiquity of the computer or calculators a calculation like finding the *exact* probability of batting “320” or better over 200 at bats was impossible. However, good approximations were quite easy to do. The reason being is that the binomial distribution is approximately a “normal” distribution when n is large enough. (“Large enough” is typically taken to be that np and $n(1 - p)$ are both bigger than 5.) Thus to compute binomial distributions for all n and p , it was enough to know **just** the normal distributions. This was tabulated and appears in most introductory statistical books now.

We can view this approximation quite easily with the computer. For concreteness, the approximation is this:

If X is a binomial random variable with n and p as parameters, then the probability $P(X \leq b)$ is given approximately by the area to the left of b for the “normal distribution” with mean np and standard deviation $\sqrt{np(1 - p)}$.

Graphically, we can illustrate this with two overlaid plots in Figure 3.

The figure shows the probability that a binomial random variable, X , is less than or equal to 5 using the areas of 6 rectangles, each with base length 1, and height given by a probability. The normal curve, with parameters chosen by $\mu = np$ and $\sigma = \sqrt{np(1 - p)}$, is overlaid showing that the shaded area is approximately given by the area to the left of 5 for the normal curve.

To illustrate in terms of functions we can use the `pnorm()` function to return the probabilities for the normal distribution:

```
> n = 8
> p = 1/2
> k = 5
> pbinom(k, n, p)

[1] 0.8554688

> pnorm(k, mean = n * p, sd = sqrt(n * p * (1 - p)))

[1] 0.76025
```

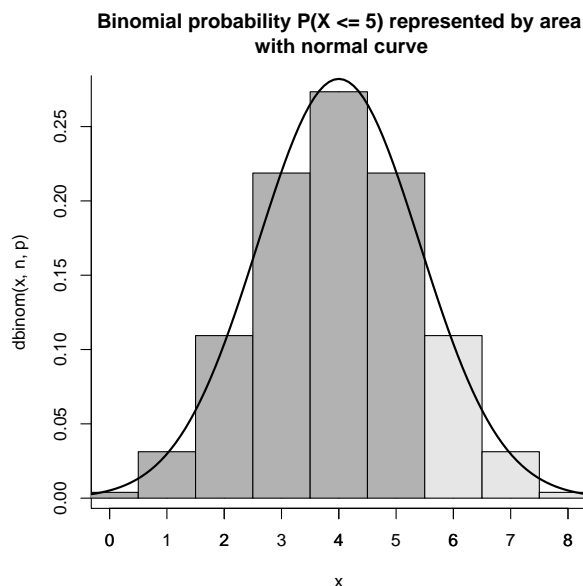



Figure 3: Probability $P(X \leq 5)$ for a binomial random variable with $n = 8$ and $p = 1/2$. The probability is the sum of the areas of the darker-shaded rectangles and this is well approximated by the area under the drawn curve to the left of 5 (or even better $5 \frac{1}{2}$).

```
> pnorm(k + 1/2, mean = n * p, sd = sqrt(n * p * (1 - p)))
[1] 0.8555778
```


The approximation is better with the extra $1/2$ (a continuity correction), but when n is large, this difference becomes negligible.


For the example from above of hitting “320” or better


```
> n = 200
> p = 0.3
> pbinom(0.32 * n, n, p, lower.tail = FALSE)
[1] 0.2420945

> pnorm(0.32 * n, n * p, sqrt(n * p * (1 - p)), lower.tail = FALSE)
[1] 0.268547
```

Notice the value is not exact, but is a good approximation.

 **Question 19:** Usually you see the normal approximation in terms of the z-score. In particular, the z-score is approximately normal with mean 0 and variance 1. Suppose $n = 600$, $p = .25$ and find the probability of 125 or fewer successes. First find the z-score and store it in `z`, then use the command `pnorm(z)`.

 Question 20: Mike Piazza was a “320” hitter in the 1990’s. In the year 2002 he had 478 at bats and 134 hits for a “280” average (`batting[playerID=="piazzmi01",]`). What is the probability a 320 hitter has this type of average or worse? Find both the exact answer using the binomial distribution, then compare with the normal approximation.

 Question 21: The overall batting average in 2002 (for these players) was $p = 0.2674$ from `sum(H)/sum(AB)`. Let’s assume that we think all batters have this success probability and their different batting averages are simply due to randomness. Compute the z-scores for all the players at once, then analyze the distribution of the z-scores using a histogram. (You may want $z = (H - AB \cdot p) / \sqrt{AB \cdot p \cdot (1-p)}$.)

If you draw your histogram using the argument `probability=TRUE`, then you can add the normal curve with the command

```
> curve(dnorm(x), add=TRUE)
```

Do so, then comment if the normal distribution seems to describe this data set.