Chapter 11 R Probability Examples

Bret Larget

March 26, 2014

Abstract

This document shows some probability examples and R code that goes beyond the scope of the Lock^5 textbook.

1 2 × 2 Tables

To illustrate the ideas, we begin with an artificial example where each of a sample of 20 individuals is characterized by sex and whether or not they have one or more pierced ears. Here is the data in a table.

	Ears		
Sex	**Pierced**	**not Pierced**	**Total**
Female	10	2	12
Male	1	7	8
Total	11	9	20

Let p_F and p_M be the population proportions of individuals with at least one pierced ear for females and males, respectively, and assume that the 20 indivuals are a random sample from a population of interest. Here is a hypothesis we can test to assess the evidence that a larger proportion of females have at least one pierced ear than do males.

$$H_0 : p_F = p_M$$
$$H_A : p_F > p_M$$

Earlier, we computed a p-value for this test by taking an array of 11 ones and 9 zeros, sampling 12 of these *without replacement*, taking the difference in sample proportions of ones for the samples of size 12 and 8, and seeing what proportion of these were at least as large as the observed difference

$$\frac{10}{12} - \frac{1}{8} \doteq 0.7083$$

From this process, we can define X to be the number of individuals in the sample of size 12 with pierced ears if we were to sample 12 individuals at random from 20 of which 11 have a pierced ear. In the sample, $X = 10$. For the data to at least as extreme, X would need to be 10 or larger. The only outcome with a more extreme difference in proportions would be if $X = 11$ and of the data looked like this.

	Ears		
Sex	**Pierced**	**not Pierced**	**Total**
Female	11	1	12
Male	0	8	8
Total	11	9	20

There are only 11 individuals with pierced ears, so that is the maximum that can be in the first sample. The minimum is 3 which would happen if all 8 individuals from the sample of 8 had pierced ears.

The p-value, then, is $\mathsf{P}(X = 10 \cup X = 11)$ which is $\mathsf{P}(X = 10) + \mathsf{P}(X = 11)$ as these two events are disjoint. One way to think about computing these probabilities is to think about all ways to choose 12 balls from 20 and to count how many of these ways have 10 (or 11) of the color associated with pierced ears. Using the binomial coefficient

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

for the number of ways to choose k items from n, we find the following.

$$\mathsf{P}(X = 10) = \frac{\binom{11}{10}\binom{9}{2}}{\binom{20}{12}} = \frac{11 \times 36}{125970} \doteq 0.0031$$

The p-value is $\mathsf{P}(X = 10) + \mathsf{P}(X = 11)$, so we also need to calculate

$$\mathsf{P}(X = 11) = \frac{\binom{11}{11}\binom{9}{1}}{\binom{20}{12}} = \frac{1 \times 9}{125970} \doteq 7.1446 \times 10^{-5}$$

and their sum

$$\text{P-value} = \sum_{x=10}^{11} \frac{\binom{11}{x}\binom{9}{12-x}}{\binom{20}{12}} \doteq 0.0032$$

Summary in context.— There is very strong evidence that the population proportion of people with at least one pierced ear is higher for females than males ($p = 0.0032$, 1-sided Fisher's Exact Text).

Calculations in R.— Here are severl ways to do this calculation in R. First, using `choose()`.

```
choose(11, 10) * choose(9, 2)/choose(20, 12) + choose(11, 11) * choose(9, 1)/choose(20,
    12)

## [1] 0.003215
```

A second way recognizes that X is a hypergeometric random variable with parameters $m = 11$ ones, $n = 9$ zeros, and a sample size of $k = 12$, using `dhyper()`.

```
sum(dhyper(10:11, m = 11, n = 9, k = 12))

## [1] 0.003215
```

A third way uses the fact that $\mathsf{P}(X \geq 10) = 1 - \mathsf{P}(X \leq 9)$.

```r
1 - phyper(9, m = 11, n = 9, k = 12)

## [1] 0.003215
```

Finally, there is a builtin function named `fisher.test()` that takes the data as a matrix and the direction of the alternative hypothesis refers to the upper left element.

```r
mat = matrix(c(10, 1, 2, 7), nrow = 2, ncol = 2)
mat

##      [,1] [,2]
## [1,]   10    2
## [2,]    1    7

fisher.test(mat, alternative = "greater")

##
##  Fisher's Exact Test for Count Data
##
## data:  mat
## p-value = 0.003215
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  2.634   Inf
## sample estimates:
## odds ratio
##       26.6
```

1.1 Another Problem

Exercise 11.34 on page 652 categorizes members of the Rock and Roll Hall of Fame on whether or not the member is a performer or not, and on whether or not the member (individual or group) has at least one female member or not. Here is the data.

	Female members	No female members	**Total**
Performer	32	149	181
Not a performer	9	83	92
Total	41	232	273

The observed proportion of performer members with at least one female member is $32/181 \doteq 0.1768$ while the observed proportion of nonperformer members with at least one female member is $9/92 \doteq 0.0978$. This data is the entire population, so inference from a sample to a population does not make sense. But we can ask how unusual it would be to see a difference as large as

$$\frac{32}{181} - \frac{9}{92} \doteq 0.079$$

when taking random samples without replacement of sizes 181 and 92 from the total. In this example, it would be much more tedious to compute the probabilities of each outcome at least as extreme as the real data individually, so we do not.

```
1 - phyper(31, 41, 232, 181)
```

```
## [1] 0.058
```

```
sum(dhyper(32:41, 41, 232, 181))
```

```
## [1] 0.058
```

2 Normal Distribution Problems

The density of a normal distribution with mean μ and standard deviation σ is

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that a normal random variable X is between a and b where $-\infty \leq a < b \leq +\infty$ is

$$\int_a^b f(x \mid \mu, \sigma) \, \mathrm{d}x$$

Using the change of variable

$$z = \frac{x - \mu}{\sigma}, \qquad \mathrm{d}z = \mathrm{d}x / \sigma$$

it follows that

$$\int_a^b f(x \mid \mu, \sigma) \, \mathrm{d}x = \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \phi(z) \, \mathrm{d}z$$

where

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(z^2)}$$

is a normal density with $\mu = 0$ and $\sigma = 1$ called the *standard normal density*. Hence, every probability calculation for an arbitrary normal distribution can be rewritten as an equivalent calculation for a standard normal distribution.

$$\mathsf{P}\left(a < X < b\right) = \mathsf{P}\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right)$$

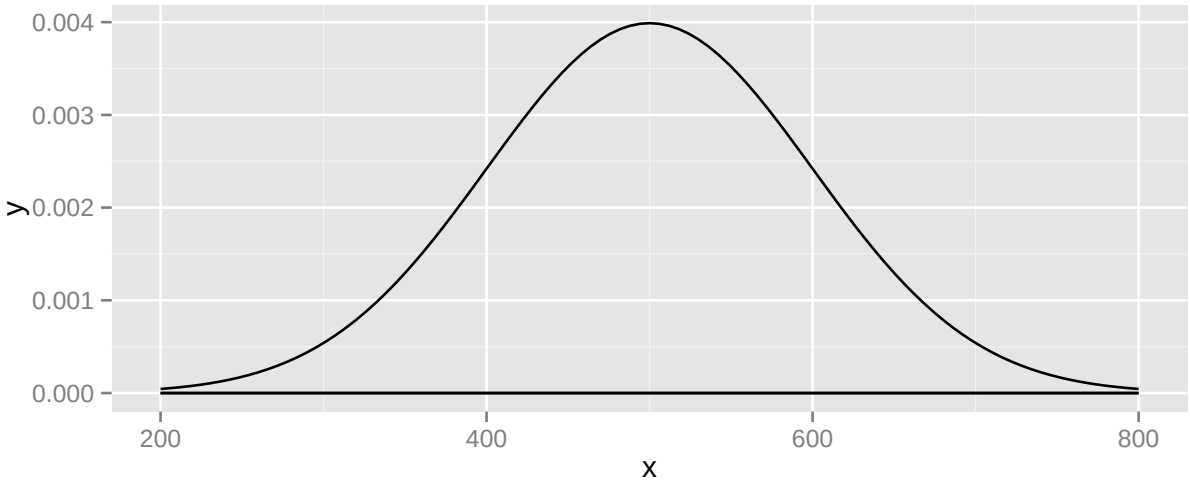
2.1 Finding Probabilities

Probabilities from normal distributions are computed by finding areas under a normal curve. The R function `pnorm()` will tell the area to the left. Here are several examples from a normal curve with $\mu = 500$ and $\sigma = 100$, drawn below.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(data.frame(x = c(200, 800)), aes(x = x)) + stat_function(fun = dnorm,
    args = list(mean = 500, sd = 100)) + geom_segment(aes(x = 200, xend = 800,
    y = 0, yend = 0))
```



P $(X < 400)$

```
pnorm(400, 500, 100)
```

```
## [1] 0.1587
```

P $(X > 650)$

```
1 - pnorm(650, 500, 100)
```

```
## [1] 0.06681
```

P $(|X - 500| > 150)$

```
pnorm(350, 500, 100) + (1 - pnorm(650, 500, 100))
```

```
## [1] 0.1336
```

2.2 Finding Quantiles

The 0.05 and 0.95 quantiles.

```
qnorm(c(0.05, 0.95), 500, 100)
```

```
## [1] 335.5 664.5
```

2.3 Finding Coverage Probabilities

Suppose that a sample mean from an independent sample from a large normal population has a standard error of 20. Then, a 95% confidence interval using the $2 \times$ SE rule covers the true mean if $|\bar{X} - \mu| < 40$, or $-40 < \bar{X} - \mu < 40$. As $\bar{X} - \mu \sim N(0, 20)$, we calculate the coverage probability as

```
pnorm(40, 0, 20) - pnorm(-40, 0, 20)

## [1] 0.9545
```

Note this is slightly larger than 0.95, but it also depended on knowing the SE perfectly.

2.4 Power Calculation

Suppose that the standard error of the sample mean is 20 and the distribution is normal; in symbols, $\bar{X} \sim N(\mu, 20)$. For the hypotheses $H_0 : \mu = 50$ versus $H_A : \mu < 50$, do the following.

1. Assume H_0 is true. For what value c is it true that if $\bar{X} = c$, the p-value calculated from a normal distribution will be 0.01?

   ```
   qnorm(0.01, 50, 20)

   ## [1] 3.473
   ```

2. If H_0 is true, what is the probability that the p-value is less than 0.01?

   ```
   pnorm(qnorm(0.01, 50, 20), 50, 20)

   ## [1] 0.01
   ```

   Or, think it through!

3. If $\bar{X} = 28.7$, what is the p-value?

   ```
   pnorm(28.7, 50, 20)

   ## [1] 0.1434
   ```

4. If $\mu = 44$, what is the probability that the p-value will be less than 0.01?

   ```
   pnorm(qnorm(0.01, 50, 20), 44, 20)

   ## [1] 0.02136
   ```

3 Binomial Calculations

The binomial distribution with parameters n and p has probability mass function

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad \text{for } x = 0, 1, 2, \ldots, n$$

Here are some sample calculations for a distribution with $n = 50$ and $p = 0.3$. Note that `dbinom()` computes binomial probabilities at individual outcomes, `pbinom()` computes the cumulative distribution function, the sum of probabilities less than or equal to a value, and `qbinom()` finds quantiles.

1. $\mathsf{P}(X = 14)$.

```
dbinom(14, 50, 0.3)
```

```
## [1] 0.1189
```

2. $\mathsf{P}(X = x)$ for $x = 5, \ldots, 10$.

```
dbinom(5:10, 50, 0.3)
```

```
## [1] 0.0005509 0.0017709 0.0047705 0.0109891 0.0219783 0.0386190
```

3. $\mathsf{P}(X \leq 18)$ two ways.

```
pbinom(18, 50, 0.3)
```

```
## [1] 0.8594
```

```
sum(dbinom(0:18, 50, 0.3))
```

```
## [1] 0.8594
```

4. $\mathsf{P}(14 \leq X \leq 18)$.

```
pbinom(18, 50, 0.3) - pbinom(13, 50, 0.3)
```

```
## [1] 0.5316
```

```
sum(dbinom(14:18, 50, 0.3))
```

```
## [1] 0.5316
```

5. $\mathsf{P}(X \geq 20)$.

```
1 - pbinom(19, 50, 0.3)
```

```
## [1] 0.0848
```

```
sum(dbinom(20:50, 50, 0.3))
```

```
## [1] 0.0848
```

6. The 0.1 and 0.9 quantiles.

```r
qbinom(c(0.1, 0.9), 50, 0.3)
```

```
## [1] 11 19
```

7. A number c so that $\mathsf{P}(X \leq c) \geq 0.4$ and $\mathsf{P}(X \geq c) \geq 0.6$. (This is the 0.4 quantile.)

```r
qbinom(0.4, 50, 0.3)
```

```
## [1] 14
```

```r
pbinom(qbinom(0.4, 50, 0.3), 50, 0.3)
```

```
## [1] 0.4468
```

```r
1 - pbinom(qbinom(0.4, 50, 0.3) - 1, 50, 0.3)
```

```
## [1] 0.6721
```