Dr Joshua Habiger
BIOS 872

Guanlin Zhang
Spring '17

Math Stat II Project
An Example of Exact and Approximate p-value

Guanlin Zhang

April 2, 2017

Our question:

Suppose that $X_1 \sim \text{binomial}(n_1, p_1)$ and $X_2 \sim \text{binomial}(n_2, p_2)$. Construct a valid p-value for testing $H_0 : p_1 = p_2$ based on the distribution of $X_1|X_1 + X_2 = t$. Construct an approximately valid p-value based on a normal approximation for $\hat{p}_1 - \hat{p}_2$ where $\hat{p}_i = X_i/n_i$. Use both p-values to analyze a real data set and compare the results.

Our solution:

Part I:

Let's first construct a valid p-value based on the conditional distribution of $X_1|X_1 + X_2 = t$:

We have $X_1 \sim \text{Binom}(n_1, p_1)$ and $X_2 \sim \text{Binom}(n_2, p_2)$, thus

$$P(X_1 = x_1) = \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1}$$
$$P(X_2 = x_2) = \binom{n_2}{x_2} p_2^{x_2}(1-p_2)^{n_2-x_2}$$

So the joint pmf of the sample (X_1, X_2) under the null $H_0 : p_1 = p_2 = p$ is:

$$\begin{aligned} f(x_1, x_2|p_1 = p_2 = p) &= \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} \binom{n_2}{x_2} p^{x_2}(1-p)^{n_2-x_2} \\ &= \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2}(1-p)^{n_1+n_2-(x_1+x_2)} \end{aligned}$$

It is easy to see by factorization theorem that $T = X_1 + X_2$ is a sufficient statistic under H_0, so we can compute the conditional distribution of $X_1|T = t$ as the following:

$$\begin{aligned} P_{H_0}(X_1 = x_1|T = t) &= \frac{P(X_1 = x_1, X_2 = t - x_1)}{P(T = t)} = \frac{\binom{n_1}{x_1}\binom{n_2}{t-x_1} p^t(1-p)^{n_1+n_2-t}}{\binom{n_1+n_2}{t} p^t(1-p)^{n_1+n_2-t}} \\ &= \frac{\binom{n_1}{x_1}\binom{n_2}{t-x_1}}{\binom{n_1+n_2}{t}} \end{aligned}$$

which is the pmf of a hypergeometric distribution with parameter $(n_1 + n_2, n_1, t)$, and we denote it as $\text{HG}(n_1 + n_2, n_1, t)$. So $P_{H_0}(X_1|T = t) \sim \text{HG}(n_1 + n_2, n_1, t)$ which does **NOT** depend on $p_1 = p_2 = p$ (because T is sufficient).

Let's consider the alternative hypothesis as $H_1 : p_1 > p_2$, then given $T = X_1 + X_2 = t$, it is reasonable to reject H_0 if $X_1 \geq x_1$ for some appropriate x_1. (If we consider $H_1 : p_1 < p_2$, we reject when $X_2 > x_2$ for some x_2, and the approach would be similar).

Also just for a sidenote, the likelihood ratio test would give us a very complicated tes statistic here so we will not use it. We skip the algebra detail for deriving the likehood ratio test statistic but just give the result to show it is very impractical to use here. The LRT in this case is:

$$\Lambda = \Big(\frac{\frac{X_1+X_2}{n_1+n_2}}{\frac{X_1}{n_1}}\Big)^{X_1}\Big(\frac{\frac{X_1+X_2}{n_1+n_2}}{\frac{X_2}{n_2}}\Big)^{X_2}\Big(\frac{1-\frac{X_1+X_2}{n_1+n_2}}{1-\frac{X_1}{n_1}}\Big)^{n_1-X_1}\Big(\frac{1-\frac{X_1+X_2}{n_1+n_2}}{1-\frac{X_2}{n_2}}\Big)^{n_2-X_2}$$

So move on to $P(X_1|X_1 + X_2 = t)$. Notice that:

$$\begin{aligned}&\text{If } t \leq n_1 :\\ &\qquad \text{then} \quad X_1 \in \{0, 1, 2, \ldots, t\}\\ &\text{If } t > n_1 :\\ &\qquad \text{then} \quad X_1 \in \{0, 1, 2, \ldots, n_1\}\end{aligned}$$

In conclusion, $X_1 \leq \min\{t, n_1\}$. So we have:

$$\begin{aligned}P(X_1 \geq x_1|T = t) &= \sum_{j=x_1}^{\min\{t,n_1\}} P(X_1 = j|T = t) = \sum_{j=x_1}^{\min\{t,n_1\}} \frac{\binom{n_1}{j}\binom{n_2}{t-j}}{\binom{n_1+n_2}{t}}\\ &= \frac{1}{\binom{n_1+n_2}{t}} \sum_{j=x_1}^{\min\{t,n_1\}} \binom{n_1}{j}\binom{n_2}{t-j}\end{aligned}$$

So given $X_1 = x_1, X_2 = x_2$, we can define a p-value as:

$$\begin{aligned}p_1(x_1, x_2) = P(X_1 \geq x_1|X_1 + X_2 = x_1 + x_2) &= \sum_{j=x_1}^{\min\{x_1+x_2,n_1\}} P(X_1 = j|X_1 + X_2 = x_1 + x_2)\\ = \frac{1}{\binom{n_1+n_2}{x_1+x_2}} &\sum_{j=x_1}^{\min\{x_1+x_2,n_1\}} \binom{n_1}{j}\binom{n_2}{x_1+x_2-j}\end{aligned}$$

Let's verify that $p_1(X_1, X_2)$ defined above is valid. We want to be careful with directly claiming it is valid from the result of the book, because the book only dealt with continuous distribution and used the fact that $F(\mathbf{X})$ follows uniform distribution, which we can not do here.

Our proof of p-value validation:

$$P_{H_0}(p_1(X_1, X_2) \leq \alpha) = \sum_t \underbrace{P(p_1(X_1, X_2) \leq \alpha|X_1 + X_2 = t)}_{X_1+X_2 \text{ sufficient}} \cdot P_{H_0}(X_1 + X_2 = t)$$

It will suffice to show for each given value $X_1 + X_2 = t$:

$$P(p_1(X_1, X_2) \leq \alpha|X_1 + X_2 = t) \leq \alpha$$

because then the p-value would be estimated as:

$$\begin{aligned} P_{H_0}(p_1(X_1, X_2) \le \alpha) &\le \sum_t \alpha \cdot P_{H_0}(X_1 + X_2 = t) = \alpha \sum_t P_{H_0}(X_1 + X_2 = t) \\ &= \alpha \end{aligned}$$

We have the following estimation:

$$\begin{aligned} P(p_1(X_1, X_2) \le \alpha | X_1 + X_2 = t) &= P\Big(\sum_{j=X_1}^{\min\{X_1+X_2, n_1\}} P(X_1 = j | X_1 + X_2) \le \alpha \Big| X_1 + X_2 = t \Big) \\ &= P\Big(\Big\{ x_1 : \sum_{j=x_1}^{\min\{t, n_1\}} P(X_1 = j | X_1 + X_2 = t) \le \alpha \Big\} \Big| X_1 + X_2 = t \Big) \\ &= P\Big(\Big\{ x_1 : x_1 \ge 1 - \alpha \text{ quantile of HG}(n_1 + n + 2, n_1, t) \Big\} \Big) \\ &\le \alpha \end{aligned}$$

Since t is any possiblevalue of $X_1 + X_2$ here, we have finished our proof of p-value validation.

Part II:

Now we construct an approximately valid p-value based on the normal approximation of $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$:

Since $X_1 \sim \text{Binom}(n_1, p_1), X_2 \sim \text{Binom}(n_2, p_2)$, we can approximate X_1 and X_2 with normal distribution: $X_1 \sim N\Big(n_1 p_1, n_1 p_1 (1 - p_1)\Big), X_2 \sim N\Big(n_2 p_2, n_2 p_2 (1 - p_2)\Big)$.

So we have:

$$\begin{aligned} \hat{p}_1 - \hat{p}_2 &= \frac{X_1}{n_1} - \frac{X_2}{n_2} \\ &\sim \frac{N(n_1 p_1, n_1 p_1 (1 - p_1))}{n_1} - \frac{N(n_2 p_2, n_2 p_2 (1 - p_2))}{n_2} \\ &\sim N\Big(p_1, \frac{p_1(1 - p_1)}{n_1}\Big) - N\Big(p_2, \frac{p_2(1 - p_2)}{n_2}\Big) \\ &\sim N\Big(p_1 - p_2, \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}\Big) \end{aligned}$$

The last one is due to the independence of X_1 and X_2. So under the null $H_0 : p_1 = p_2 = p$, we have:

$$\hat{p}_1 - \hat{p}_2 \sim N\Big(0, (\frac{1}{n_1} + \frac{1}{n_2}) p(1 - p)\Big)$$

Consider $H_1 : p_1 > p_2$, then we let $T = \hat{p}_1 - \hat{p}_2$ be the test statistic, and reject when $T \ge k$ for appropriate k.

Notice that we have:

$$P_{H_0}\Big(T \geq \frac{x_1}{n_1} - \frac{x_2}{n_2}\Big) = P_{H_0}\Big(\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}} \geq \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}}\Big)$$
$$\overset{\text{approximately}}{\simeq} P\Big(Z > \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}}\Big) \qquad (Z \sim N(0,1))$$

So we can define our p-value as:

$$p_2(x_1, x_2) = P\Big(Z > \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}}\Big)$$

We want to check the p value defined above is approximately valid:

$$\begin{aligned} P_{H_0}\Big(p_2(X_1, X_2) \leq \alpha\Big) &= P_{H_0}\Big(P(Z > \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}}) \leq \alpha\Big) \\ &= P_{H_0}\Big(\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}} \geq \Phi^{-1}(1-\alpha)\Big) \qquad (\Phi \text{ is the cdf of } N(0,1)) \\ &\overset{\text{approximately}}{\simeq} P\Big(Z \geq \Phi^{-1}(1-\alpha)\Big) \\ &\leq \alpha \end{aligned}$$

So our p value is **approximately** valid. However this does not prove that the p value is valid. In fact it is not, as we will see in the simulation below.

Part III:

Now we simulate some data and compare the type I error between the two test functions defined by each p-value.

Our test functions are:

$$\phi_1(X_1, X_2) = I(p_1(X_1, X_2) \leq \alpha)$$
$$\phi_2(X_1, X_2) = I(p_2(X_1, X_2) \leq \alpha)$$

We take $\alpha = 0.05$, and simulate (X_1, X_2) for 1000 times under the null hypothesis and compute the type I error rate.

We have the following code:

```
#generate X_1 and X_2.
#X_1 follows binom(30, 0.5), X_2 follows binom(50, 0.5)
#we simulate 1000 times
X_1 <- rbinom(1000, 30, 0.5)
X_2 <- rbinom(1000, 50, 0.5)
```

```r
#compute the p value under binomial distribution
p_binom <- rep(0, 1000)

for (i in 1:1000){
  for(j in X_1[i]: min(X_1[i]+ X_2[i], 30)){
    p_binom[i] <- p_binom[i] +
      choose(30, j)*choose(50, X_1[i]+X_2[i]-j)/choose(80, X_1[i]+X_2[i])
  }
}

#compute the p value under the normal approximation
p_norm <- rep(0, 1000)

for (i in 1:1000){
  p_norm[i] <- pnorm((X_1[i]/30- X_2[i]/50)/sqrt(0.25*(1/30+1/50)),
                     mean=0, sd=1, lower.tail=FALSE)
}

#compute the test function:
phi_binom <- ifelse(p_binom<=0.05, 1, 0)
phi_norm <- ifelse(p_norm<=0.05, 1, 0)
#compute the type I error:
Err_binom <- sum(phi_binom)/1000
Err_norm <- sum(phi_norm)/1000
cat("The Type I Error with binomial distribution is:", Err_binom )

## The Type I Error with binomial distribution is: 0.029

cat("The Type I Error with normal approximation is:", Err_norm)

## The Type I Error with normal approximation is: 0.053
```

We find that ϕ_1 alwyas give a lower type I error rate than ϕ_2, and as expected ϕ_1 always have a type I error rate $\leq \alpha = 0.05$. However the type I error rate for ϕ_2 is closer to 0.05, some times even larger than $\alpha = 0.05$(depends on the random result we get each time when we run the simulation), which also verifies that $p_2(X_1, X_2)$ is not valid, but only approximately valid.