

Test for the probability of a binomial distribution

Chao Cheng

August 30, 2022

Contents

1	Introduction	1
2	Normal approximation	2
2.1	Case 1	2
2.2	Case 2	2
3	Chi-square approximation	3
4	Exact test	3
4.1	Case 1: Clopper-Pearson test	3
4.1.1	Power analysis	4
4.1.2	Confidence interval	4
4.2	Case 2: Fisher's exact test	6
5	Approximated confidence interval for Case 1	7

1 Introduction

For an i.i.d sample from a bernoulli distribution

$$x_1, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p),$$

The likelihood of the data is

$$f(x_1, \dots, x_n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{n-\sum x_i}.$$

MLE for p is $\bar{x} = \frac{1}{n} \sum x_i$ and

$$\sum_{i=1}^n x_i \sim \text{Binom}(n, p).$$

So here are mainly two situations: One is to test the probability p against some given value p_0 . The other is to compare the probability between two independent random samples x_1, \dots, x_n and y_1, \dots, y_m .

Case1: One sample x_1, \dots, x_n from $\text{Bernoulli}(p)$, and test p against a given p_0 .

Case2: Two samples: x_1, \dots, x_2 from $\text{Bernoulli}(p_1)$ and y_1, \dots, y_m from $\text{Bernoulli}(p_2)$.
And test whether $p_1 = p_2$.

2 Normal approximation

2.1 Case 1

Note that

$$EX = p, \quad \text{Var}X = p(1 - p).$$

Then by CLT we have

$$\bar{x} \stackrel{\text{asympt}}{\sim} N\left(p, \frac{p(1-p)}{n}\right).$$

For $H_0 : p = p_0$, we propose a test statistic

$$Z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Then Z is asymptotically standard normal under H_0 .

Also we know that under H_1 :

$$\begin{aligned} Z &= \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{\bar{x} - p}{\sqrt{\frac{p(1-p)}{n}}} \cdot \sqrt{\frac{p(1-p)}{p_0(1-p_0)}} + \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &\sim N\left(\frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \frac{p(1-p)}{p_0(1-p_0)}\right). \end{aligned}$$

So the power of the test can be easily computed.

2.2 Case 2

So we have

$$\bar{x} \stackrel{\text{asympt}}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n}\right), \quad \text{and} \quad \bar{y} \stackrel{\text{asympt}}{\sim} N\left(p_2, \frac{p_2(1-p_2)}{m}\right).$$

A test statistic can be

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

where $\hat{p} = \frac{n\bar{x} + m\bar{y}}{n+m}$. This test statistic can be found at

<https://stats.stackexchange.com/questions/361015/proof-of-the-standard-error-of-the-distribution-between-two-normal-distributions/361048#361048>

<https://stats.stackexchange.com/questions/113602/test-if-two-binomial-distributions-are-statistically-different-from-each-other>

Here this $\hat{p}(1 - \hat{p})$ can be seen as an estimate for the variance $p(1 - p)$ when H_0 is true by directly plugging in \hat{p} . This is **NOT** a pooled variance for these two samples, which should always be no greater than $\hat{p}(1 - \hat{p})$.

The power of this test statistic is hard to compute under H_1 .

Note: One can also use the same idea in the “t-test.pdf” notes and propose the test statistic

$$T = \frac{\bar{x} - p_0}{\sqrt{S_x/n}},$$

where $S_x = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ for Case 1.

And for Case 2

$$T = \frac{\bar{x} - \bar{y} - \Delta}{\sqrt{(\frac{1}{n} + \frac{1}{m}) S_p}},$$

where $S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$ and $\Delta = p_1 - p_2$. But again it is hard to evaluate the testing power of these statistics.

3 Chi-square approximation

See the notes of “chisq_test.pdf” for details.

4 Exact test

4.1 Case 1: Clopper-Pearson test

The Clopper-Pearson method is an early method. It’s called exact method because it’s directly based on p.m.f of binomial distribution. Let $X = \sum_{i=1}^n x_i$. Then $X \sim \text{Binom}(n, p)$ and the p.m.f is

$$f(x; p) = P(X = x|p) = C_n^x p^x (1 - p)^{n-x}, \quad (1)$$

for $x = 0, 1, \dots, n$. One thing to point out is that though the $|p$ notation, (1) is frequentist’s point of view, not bayesian’s. Now let’s recall that p-value is the probability under H_0 that something **as or more extreme than** what we have observed happens. Then after observing $X = x_0$, for one-sided test:

- $H_0 : p \leq p_0$ against $H_1 : p > p_0$ for some given p_0 . The p-value at this observed x_0 is

$$p_{val}(x_0) = \sum_{x=x_0}^n f(x; p_0). \quad (2)$$

- $H_0 : p \geq p_0$ against $H_1 : p < p_0$ for some given p_0 . The p-value at this observed x_0 is

$$p_{val}(x_0) = \sum_{x=0}^{x_0} f(x; p_0). \quad (3)$$

For the two-sided test. This is a little complicated. Let index set

$$\mathcal{I} = \{x | P(X = x|p_0) \leq P(X = x_0|p_0), \quad 0 \leq x \leq n\}.$$

Then \mathcal{I} contains all possible realizations of X with its probability no greater than the probability of our observation. Then the p-value of $H_0 : p = p_0$ at this observed x_0 is given by

$$p_{val}(x_0) = \sum_{x \in \mathcal{I}} f(x; p_0). \quad (4)$$

4.1.1 Power analysis

The probability to reject H_0 of Clopper-Pearson test at given underlying p can be computed by

$$P(\text{Reject } H_0 | p) = \sum_{x=0}^n P(X = x | p) \cdot I_{\{p_{val}(x) \leq \alpha\}} = \sum_{x=0}^n f(x; p) \cdot I_{\{p_{val}(x) \leq \alpha\}}, \quad (5)$$

where α is the significant level of the test and $p_{val}(x)$ is computed for different types of H_0 based on (2), (3) and (4).

4.1.2 Confidence interval

First for the one-sided intervals:

- $(P_L, 1]$: From (2), $H_0 : p \leq p_0$ is rejected when probability of observing x_0 or more number of success at p_0 is small enough. Therefore the reject area

$$\text{Reject Area: } \left\{ x_0 : \sum_{x=x_0}^n f(x; p_0) \leq \alpha \right\}.$$

Hence the accept area

$$\text{Accept Area: } \left\{ x_0 : \sum_{x=x_0}^n f(x; p_0) > \alpha \right\}$$

Then we can construct the one-sided CI by increasing p_0 from 0 such that the first p_0 that satisfies this Accept area rule. Then that is the P_L . Therefore

$$\sum_{x=x_0}^n f(x; P_L) = \alpha. \quad (6)$$

- $[0, P_U)$: Similar idea, from (3) we can construct the accept area

$$\text{Accept Area: } \left\{ x_0; \sum_{x=0}^{x_0} f(x; p_0) > \alpha \right\}.$$

Therefore we decrease p_0 from 1 to find the first P_U that satisfies this Accept area rule. Therefore

$$\sum_{x=0}^{x_0} f(x; P_U) = \alpha. \quad (7)$$

Now for the two-sided intervals (P_L, P_U) : we apply the **equal-tail rule** and find P_L and P_U such that

$$\begin{aligned} \sum_{x=x_0}^n f(x; P_L) &= \alpha/2 \\ \sum_{x=0}^{x_0} f(x; P_U) &= \alpha/2. \end{aligned} \tag{8}$$

This interval can also be expressed as

$$S_{\leq} \cap S_{\geq},$$

or equivalently

$$(\inf S_{\geq}, \sup S_{\leq}),$$

where

$$\begin{aligned} S_{\leq} &\triangleq \left\{ \theta \mid P(\text{Binomial}(n, \theta) \leq x) > \frac{\alpha}{2} \right\} \\ S_{\geq} &\triangleq \left\{ \theta \mid P(\text{Binomial}(n, \theta) \geq x) > \frac{\alpha}{2} \right\}. \end{aligned}$$

One can utilize the relationship between the Binomial cumulative distribution function and **regularized incomplete beta function**, i.e. for $k = 0, \dots, n$

$$\begin{aligned} P(X \leq k) &= \sum_{i=0}^k C_n^i p^i (1-p)^{n-i} \\ &= \frac{\Gamma(n+1)}{\Gamma(n-k)\Gamma(k+1)} \int_0^{1-p} t^{n-k-1} (1-t)^k dt = pBeta(1-p; n-k, k+1) \\ &= \frac{\Gamma(n+1)}{\Gamma(n-k)\Gamma(k+1)} \int_p^1 t^k (1-t)^{n-k-1} dt = 1 - pBeta(p; k+1, n-k), \end{aligned}$$

where $pBeta(x; \alpha, \beta)$ represents the cumulative probability of $Beta(\alpha, \beta)$ distribution, cumulated from 0 to x . And it satisfies

$$pBeta(x; \alpha, \beta) = 1 - pBeta(1-x; \beta, \alpha), \quad \forall x \in [0, 1].$$

Similarly, for the quantile function $qBeta(p; \alpha, \beta)$, we can show that

$$1 - qBeta(p; \alpha, \beta) = qBeta(1-p; \beta, \alpha), \quad \forall p \in [0, 1].$$

So we can see that P_L and P_U are actually satisfying

$$\begin{aligned} 1 - \alpha/2 &= (\geq) P(X \leq x_0 - 1 | P_L) = pBeta(1 - P_L; n - x_0 + 1, x_0) \\ \implies 1 - P_L &= (\leq) qBeta(1 - \alpha/2; n - x_0 + 1, x_0) \\ \implies P_L &= (\geq) qBeta(\alpha/2; x_0, n - x_0 + 1) \end{aligned}$$

For P_L , it's taking inf, therefore

$$P_L = qBeta(\alpha/2; x_0, n - x_0 + 1). \tag{9}$$

Similarly, we have for P_U :

$$\begin{aligned} \alpha/2 &= P(X \leq x_0 | P_U) = pBeta(1 - P_U; n - x_0, x_0 + 1) \\ \implies 1 - P_U &= qBeta(\alpha/2; n - x_0, x_0 + 1) \\ \implies P_U &= qBeta(1 - \alpha/2; x_0 + 1, n - x_0) \end{aligned}$$

Therefore

$$P_U = qBeta(1 - \alpha/2; x_0 + 1, n - x_0). \quad (10)$$

Also, note that this cumulative probability is also related to F-distribution via

$$P(X \leq x_0) = F\left(x = \frac{1-p}{p} \frac{x_0 + 1}{n - x_0}; d_1 = 2(n - x_0), d_2 = 2(x_0 + 1)\right)$$

where $F(x; d_1, d_2)$ is the cumulative probability function of a F-distribution with degree of freedom d_1 and d_2 , cumulated from 0 to x . Then we have

$$\begin{aligned} P_L &= \left(1 + \frac{n - x_0 + 1}{x_0 \times qF\left(\frac{\alpha}{2}; 2x_0, 2(n - x_0 + 1)\right)}\right)^{-1} \\ P_U &= \left(1 + \frac{n - x_0}{(x_0 + 1) \times qF\left(1 - \frac{\alpha}{2}; 2(x_0 + 1), 2(n - x_0)\right)}\right)^{-1} \end{aligned} \quad (11)$$

where $qF(\alpha; d_1, d_2)$ is the quantile function of F-distribution.

4.2 Case 2: Fisher's exact test

Fisher's exact test is a method for testing proportion difference. A toy example of a 2×2 contingency table is shown in Table 1. When the margin of this table is fixed (n_x, n_y, n_1 and n_2), the probability for observing this table follows the hyper-geometric distribution

$$P(\#\{\text{Sample 1, Success}\} = a) = \frac{C_{n_x}^a C_{n_y}^{n_1 - a}}{C_n^{n_1}}.$$

	Success	Failure	Total
Sample 1	a	b	$n_x = a + b$
Sample 2	c	d	$n_y = c + d$
Total	$n_1 = a + c$	$n_0 = b + d$	$n = a + b + c + d$

Table 1: Data sample

We can compute the p-value based on the same idea from Section 4.1. Note that here the p-value is a **conditional** one since it is conditional on the fixed marginal values. To simplify the notation, denote X the number in cell (Sample 1, Success) and

$$f(x; n_x, n_y, n_1) = P(X = x | n_x, n_y, n_1) = \frac{C_{n_x}^x C_{n_y}^{n_1 - x}}{C_{n_x + n_y}^{n_1}}$$

Then for a observation with $X = x_0$ and fixed n_x, n_y, n_1 :

- $H_0 : p_x \geq p_y$ against $H_1 : p_x < p_y$. The p-value can be computed as

$$p_{val} = \sum_{i=0}^{x_0} f(i; n_x, n_y, n_1). \quad (12)$$

- $H_0 : p_x \leq p_y$ against $H_1 : p_x > p_y$. The p-value can be computed as

$$p_{val} = \sum_{i=x_0}^{n_1} f(i; n_x, n_y, n_1). \quad (13)$$

- $H_0 : p_x = p_y$ against $H_1 : p_x \neq p_y$. The p-value can be computed as

$$p_{val} = \sum_{i=a_L}^{a_U} f(i; n_x, n_y, n_1) I_{\{f(i; n_x, n_y, n_1) \leq f(x_0; n_x, n_y, n_1) \delta\}}, \quad (14)$$

where the summation limits $a_L = \max(0, n_1 - n_y)$ and $a_U = \min(n_x, n_1)$. Note that in an ideal world the red δ is just 1 in (14). But in actuality, since the computation involves large factorials, especially when sample size is large, the numerical results might be inaccurate. To ensure a convervative test, δ is set to 1.0000001 in R[?].

5 Approximated confidence interval for Case 1

Let z_α be the left α quantile of standard normal distribution. $\sum_{i=1}^n x_i$ is the number of success trials and $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$ is the MLE for p . Then a confidence interval for p can be constructed using various methods.

- Normal/Wald Approximation:

$$\hat{p} \pm z_{1-\alpha/2} \times \sqrt{\hat{p}(1-\hat{p})/n}.$$

- Agresti-Coull method: Define

$$\tilde{p} = \tilde{n}^{-1} \left(n\hat{p} + \frac{z_{1-\alpha/2}^2}{2} \right), \quad \tilde{n} = n + z_{1-\alpha/2}^2.$$

Then the CI is constructed as

$$\tilde{p} \pm z_{1-\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},$$

which is just the form of Normal Approximation with \tilde{p} and \tilde{n} plugged in.

- Wilson Score method: Find the roots p of

$$|p - \hat{p}| = z_{1-\alpha/2} \sqrt{p(1-p)/n}.$$

And the solutions form the CI

$$\left(1 + \frac{z_{1-\alpha/2}^2}{n} \right)^{-1} \left(\hat{p} + \frac{z_{1-\alpha/2}^2}{2n} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}} \right).$$

- Arcsin method:

$$\sin^2 \left(\arcsin \left(\sqrt{\hat{p}} \right) \pm \frac{z_{1-\alpha/2}}{2\sqrt{n}} \right).$$

The normal approximated one is the simplest and most introductory one, but its performance is only valid for large sample, not finite n . The Clopper-Pearson interval is an exact one, but it's always conservative, so the coverage probability is at least $1 - \alpha$. These other approximated all try to be more accurate than the normal approximated one and less conservative than Clopper-Pearson method. [{need reference here}](#) Though the [Arcsin method might be unstable when \$\hat{p}\$ is close to 0 or 1.](#)