Inference on Proportions

September 12, 2017

Variables

Our data now is of the form $X_1, X_2, \ldots, X_n$ which are iid from a Bernoulli distribution, that is: X_i is either a 0 (failure) or 1 (success). The probability of a success is $P(X_i = 1) = p$ where p is the population parameter that is unknown. The probability distribution function for X is given by

$$f_X(x) = P(X = x) = p^x(1-p)^{1-x} \quad x = 0, 1$$

.

Here the X's are coded as 0 and 1 based on responses:

- Democrat or Republican
- Control or Treatment
- Male or Female

Number of successes

Let Y be the sum of the X's, $Y = \sum_{i=1}^{n} X_i$. We can show that Y has a binomial distribution with parameters n and p, i.e.

$$f_Y(y) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \quad y = 0, 1, \ldots, n$$

Y is the number of successes in the n trials. $E(Y) = np$ and $Var(Y) = np(1-p)$.

Estimating p

The estimator $\hat{p} = \dfrac{Y}{n}$ is an unbiased estimator of p. $E(\hat{p}) = p$ and $Var(\hat{p}) = \dfrac{p(1-p)}{n}$. For large n we have

$$\hat{p} \to N\left(p, \frac{p(1-p)}{n}\right)$$

We can estimate the variance of $\hat{p}$ as $\hat{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n}$.

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for p is

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

Hypothesis Testing

To test $H_0 : p = p_0$ (or $H_0 : p \leq p_0$ or $H_0 : p \geq p_0$)
Test Statistic

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Under H_0 for large n, $T \to N(0, 1)$

Hypothesis Testing

Alternative Hypothesis		
$H_1 : p < p_0$	$H_1 : p > p_0$	$H_1 : p \neq p_0$
p-value		
$P(Z < t_{obs})$	$P(Z > t_{obs})$	$2 * P(Z > \lvert t_{obs} \rvert)$

Comparing two proportions

$$X_1, X_2, \ldots, X_n \quad iid \quad Bernoulli(p_x)$$
$$Y_1, Y_2, \ldots, Y_m \quad iid \quad Bernoulli(p_y)$$

Interest is in whether $p_x = p_y$.

Null Hypotheses

- Difference $H_0 : p_x - p_y = 0$
- Relative risk $H_0 : p_x/p_y = 1$
- Odds ratio $H_0 : \dfrac{p_x/(1-p_x)}{p_y/(1-p_y)} = 1.$

Difference

For large n and m and

$$\hat{p}_x = \frac{X}{n} \quad \hat{p}_y = \frac{Y}{m}$$

where X is the number of successes in the X's and Y is the number of successes in the Y's.

$$\hat{p}_x - \hat{p}_y \approx N\left(p_x - p_y, \frac{p_x(1-p_x)}{n} + \frac{p_y(1-p_y)}{m}\right)$$

Difference

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for $p_x - p_y$ is

$$\hat{p}_x - \hat{p}_y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n} + \frac{\hat{p}_y(1-\hat{p}_y)}{m}}$$

Difference

Hypothesis Test

To test $H_0 : p_x = p_y$, under the null hypothesis we can get a pooled estimate of $\hat{p}$ as

$$\hat{p} = \frac{n\hat{p}_x + m\hat{p}_y}{n + m}$$

We can then use test statistic

$$T = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

for $\hat{q} = 1 - \hat{p}$. Under H_0, $T \approx N(0, 1)$ when n, m are large.

Difference

Hypothesis Test

- For alternative $H_1 : p_x < p_y$ reject H_0 if T is too small
- For alternative $H_1 : p_x > p_y$ reject H_0 if T is too big
- For alternative $H_1 : p_x \neq p_y$ reject H_0 if T is too big or too small

Difference

Hypothesis Test n, m small

For two small samples, let $N = n + m$ be the total number of trials in both samples. Let $V = X + Y$ be the total number of successes in both samples.

	Success	Failure	Total
Sample 1	X	n-X	n
Sample 2	Y	m-Y	m
	V	N-V	N

The test statistic is $X =$ number of success from Sample 1.

When $H_0 : p_x = p_y$ is true, the successes are equally likely from the 2 samples. We have

$$P(X = i | X + Y = V) = \frac{\binom{n}{i}\binom{m}{V-i}}{\binom{N}{V}}$$

Upper p-value (for alternative $H_1 : p_x > p_y$):

$$P(X \geq x | X + Y = V)$$

for observed x number of successes in the 1st sample.

Example

The data in the table below are from an age discrimination case.

Age Group	Number Fired	Number Kept	
Young	1	24	25
Old	10	17	27
total	11	41	52

The question is whether the proportion of young people fired is less than the proportion of old people fired. $H_0 : p_y = p_o$ versus $H_1 : p_y < p_o$. The observed test statistic is $X = 1$. The p-value is thus

$$P(X \leq 1 | X + Y = 11) =$$

$$P(X = 0 | X + Y = 11) + P(X = 1 | X + Y = 11)$$

$$= \frac{\binom{25}{0}\binom{27}{11}}{\binom{52}{11}} + \frac{\binom{25}{1}\binom{27}{10}}{\binom{52}{11}} \approx 0.005$$