

Two sample inference for proportions

Testing

Suppose we wish to test equality of two proportions p_1 and p_2 , based on two independent samples from $\text{Ber}(p_1)$ and $\text{Ber}(p_2)$, with respective sample size n_1 and n_2 . Let the samples be X_1, \dots, X_{n_1} where $X_i \sim \text{Ber}(p_1)$ and Y_1, \dots, Y_{n_2} where $Y_i \sim \text{Ber}(p_2)$. So we consider hypotheses

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2.$$

Take the two sample proportions $\bar{X}_{n_1} =: \hat{p}_{(1)}$ and $\bar{Y}_{n_2} =: \hat{p}_{(2)}$; a reasonable testing procedure (analogous to what was done for exactly normal observations, with two samples) would be to consider the difference $\hat{p}_{(1)} - \hat{p}_{(2)}$ and reject H_0 if this difference is "too large" in absolute value. To obtain a benchmark for what is "too large", we have to find the standard deviation of $\hat{p}_{(1)} - \hat{p}_{(2)}$:

$$\text{Var}(\hat{p}_{(1)} - \hat{p}_{(2)}) = \text{Var}(\hat{p}_{(1)}) + \text{Var}(\hat{p}_{(2)}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \quad (50)$$

Under the null hypothesis we have $p_1 = p_2 =: p$, and since the primary concern in testing is to ensure the bound on the error of first kind (under H_0), we might assume this equality:

$$\text{Var}(\hat{p}_{(1)} - \hat{p}_{(2)}) = p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

The next step is to form a normalized expression from $\hat{p}_{(1)} - \hat{p}_{(2)}$ and conjecture the asymptotic normality:

$$\hat{Z}_{n,0} := \frac{\hat{p}_{(1)} - \hat{p}_{(2)}}{\text{SD}(\hat{p}_{(1)} - \hat{p}_{(2)})} = \frac{\hat{p}_{(1)} - \hat{p}_{(2)}}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

This is not yet a Z-statistic since we do not know p (we have to estimate it).

1.20 Lemma *Suppose that $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$ in such a way that $n_1/n_2 \rightarrow c$ where $c > 0$. Then, under the null hypothesis $p_1 = p_2 = p$, we have $\hat{Z}_{n,0} \rightsquigarrow Z$.*

Proof. We have

$$\begin{aligned} \hat{Z}_{n,0} &= \frac{\hat{p}_{(1)} - p}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} + \frac{\hat{p}_{(2)} - p}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{\sqrt{n_1}(\hat{p}_{(1)} - p)}{\sqrt{p(1-p)}} \cdot \frac{1}{\sqrt{n_1 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} + \frac{\sqrt{n_2}(\hat{p}_{(2)} - p)}{\sqrt{p(1-p)}} \cdot \frac{1}{\sqrt{n_2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \end{aligned}$$

Furthermore

$$\begin{aligned}\sqrt{n_1 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} &= \sqrt{\left(1 + \frac{n_1}{n_2} \right)} \rightarrow \sqrt{(1+c)}, \\ \sqrt{n_2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} &= \sqrt{\left(\frac{n_2}{n_1} + 1 \right)} \rightarrow \sqrt{(1/c+1)}\end{aligned}$$

Applying the CLT to the two normalized expressions $\sqrt{n_i} (\hat{p}_{(i)} - p) / \sqrt{p(1-p)}$, it follows that

$$\hat{Z}_{n,0} \rightsquigarrow Z_1 \frac{1}{\sqrt{(1+c)}} + Z_2 \sqrt{\frac{c}{1+c}}$$

where Z_1, Z_2 are independent standard normal. The above linear combination of Z_1, Z_2 is zero mean normal with variance

$$\frac{1}{1+c} + \frac{c}{1+c} = 1$$

which proves that $\hat{Z}_0 \rightsquigarrow N(0, 1)$. ■

Remark. *It takes slightly more work to establish this result under the weaker assumption: $n_1+n_2 \rightarrow \infty$.*

To obtain a Z-statistic, we have to estimate the success probability p , common to both samples. The best way to do this is to **pool the samples** into one larger sample of size $n = n_1 + n_2$ and set

$$\hat{p}_n = \frac{1}{n} \left(\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i \right) = \frac{n_1}{n} \hat{p}_{(1)} + \frac{n_2}{n} \hat{p}_{(2)}.$$

It is obvious that as soon as $n \rightarrow \infty$ we have $\hat{p}_n \rightarrow p$ by the LLN. Thus if we form the Z-statistic

$$\hat{Z}_{n,\text{pool}} = \frac{\hat{p}_{(1)} - \hat{p}_{(2)}}{\text{SE}(\hat{p}_{(1)} - \hat{p}_{(2)})} = \frac{\hat{p}_{(1)} - \hat{p}_{(2)}}{\sqrt{\hat{p}_n(1 - \hat{p}_n) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (51)$$

then we know immediately from our earlier results on convergence in law and in probability that $\hat{Z}_{n,\text{pool}} - \hat{Z}_{n,0} \rightarrow_{\text{Pr}} 0$ and thus

$$\hat{Z}_{n,\text{pool}} \rightsquigarrow Z.$$

Thus we obtain an asymptotic α -test: the test

$$\phi(X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}) = \begin{cases} 1 & \text{if } \left| \hat{Z}_{n,\text{pool}} \right| > z_{\alpha/2}^* \\ 0 & \text{otherwise} \end{cases} \quad (52)$$

with $\hat{Z}_{n,\text{pool}}$ given by (51) is the **standard (two sided) two sample Z-test for the equality of proportions**.

Clearly, the one sided variant of this test for hypotheses $H_0 : p_1 = p_2$, $H_a : p_1 - p_2 > 0$ rejects if $\hat{Z}_{n,\text{pool}} > z_{\alpha}^*$.

Confidence interval

Suppose a confidence interval for $p_1 - p_2$ is desired. Obviously one would start from the difference $\hat{p}_{(1)} - \hat{p}_{(2)}$ and build the interval around this. We already found above in (50) that

$$\text{SD}(\hat{p}_{(1)} - \hat{p}_{(2)}) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

For the confidence interval we cannot assume that $p_1 = p_2$, hence there is no further simplification here and one cannot use a pooled estimate. To obtain the estimate $\text{SE}(\hat{p}_{(1)} - \hat{p}_{(2)})$, we have to estimate p_1 and p_2 separately and set

$$\text{SE}(\hat{p}_{(1)} - \hat{p}_{(2)}) = \sqrt{\frac{\hat{p}_{(1)}(1-\hat{p}_{(1)})}{n_1} + \frac{\hat{p}_{(2)}(1-\hat{p}_{(2)})}{n_2}}.$$

The appropriate Z-statistic as a basis for the confidence interval is then

$$\hat{Z}_n = \frac{\hat{p}_{(1)} - \hat{p}_{(2)}}{\text{SE}(\hat{p}_{(1)} - \hat{p}_{(2)})} \quad (53)$$

and the interval is now

$$\begin{aligned} & [\hat{p}_{(1)} - \hat{p}_{(2)} - m, \hat{p}_{(1)} - \hat{p}_{(2)} + m] \quad \text{where} \\ m &= z_{\alpha/2}^* \text{SE}(\hat{p}_{(1)} - \hat{p}_{(2)}) \end{aligned} \quad (54)$$

Clearly under the assumptions of Lemma 1.20 this is a confidence interval for $p_1 - p_2$ of asymptotic level $C = 1 - \alpha$.

Relation of standard test and confidence interval. The standard confidence interval (54) yields a valid asymptotic α -test for the two sided hypotheses $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$, -just check whether the value $0 = p_1 - p_2$ is covered or not. However this is not the standard test; the standard test (52) uses the pooled estimate \hat{p}_n in $\text{SE}(\hat{p}_{(1)} - \hat{p}_{(2)})$. Thus it is preferable, according to the philosophy that one should use every bit of information about the null hypothesis, in the interest of achieving the α bound on the error of first kind for smaller sample size. From a strictly formal, asymptotic point of view, or if sample size is large enough, the "non-pooled" Z-statistic (53) is sufficient.