Introduction to Statistical Method

Inference on Proportions

Estimating Proportions

$$X = \begin{cases} 1 & \text{has trait} \\ 0 & \text{does not have trait} \end{cases}$$

$$p = \frac{\#\text{members with trait}}{\text{population size}} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

If we take a random sample $X_1, \dots X_n$ of X, the sample mean $\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased estimator for p.

So the random variable X follows a point binomial distribution with $E[X] = p$ and $VarX = p(1-p)$.

Then we can see from central limit theorem, $\hat{p}$ is approximately normally distributed with mean p and $Var = p(1-p)/n$.

Thus $\dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$ is approximately standard-normally distributed.

The $100(1-\alpha)\%$ confidence interval for p is $\hat{p} \pm z_{\alpha/2}\sqrt{p(1-p)/n}$

But the p unknown, so we replace it with $\hat{p}$.

But we should replace the $z_{\alpha/2}$ with $t_{\alpha/2}$ according to `ve401 note 7`. (we replace σ by S to get confidence interval for mean)

(we neglect the operation here)

Choosing the Sample Size

Given a $100(1-\alpha)\%$ confidence interval $p = \hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

Then $d = z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ and we derive $n = \dfrac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{d^2}$ as the sample size to ensure the confidence and $|p - \hat{p}| < d$. (requirement, have an estimate $\hat{p}$ and p beforehand).

Since $x(1-x) < 0.25$, then we deduce that $n = \dfrac{z_{\alpha/2}^2}{4d^2}$

Hypothesis Testing

Let $X_1..X_n$ be a random sample of size n from a Bernoulli distribution with parameter p and we let $\hat{p} = \overline{X}$ for the sample mean.

Test $H_0 : p = p_0$ based on statistic $Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ is called a large-sample test for proportion.

We reject H_0 at significance level α

- in favor of $H_1 : p \neq p_0$ if $|Z| > z_{\alpha/2}$
- in favor of $H_1 : p > p_0$ if $Z > z_\alpha$
- in favor of $H_1 : p < p_0$ if $Z < -z_\alpha$

Comparing Two Proportions

If we have different random sample of size n_1 and n_2 for $X^{(1)}$ and $X^{(2)}$, both are approximately normally distributed.

So mean p_1, p_2 and variance $p_1(1-p_1)/n_1, p_2(1-p_2)/n_2$.

Then for large samples, the estimator $\hat{p_1} - \hat{p_2}$ is approximately normal with mean $p_1 - p_2$ and variance $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$.

So the $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$: $\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p_1}(1-\hat{p_1})}{n_1} + \dfrac{\hat{p_2}(1-\hat{p_2})}{n_2}}$

Test for Comparing Two Proportions

Let $X_1^{(i)}..X_{n_i}^{(i)}, i = 1, 2$ be random samples of size n_i from two Bernoulli distributions with parameters p_i and $\hat{p_i} = \overline{X}_i$ for corresponding sample means.

Then $H_0 : p_1 - p_2 = (p_1 - p_2)_0$ based on statistic $Z = \dfrac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$ is called large-sample test for differences in proportions.

We reject H_0 at significance level α

- in favor of $H_1 : p_1 - p_2 \neq (p_1 - p_2)_0$ if $|Z| > z_{\alpha/2}$
- in favor of $H_1 : p_1 - p_2 > (p_1 - p_2)_0$ if $Z > z_\alpha$
- in favor of $H_2 : p_1 - p_2 < (p_1 - p_2)_0$ if $Z < -z_\alpha$

Equality of Proportions

So the test statistic $Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$ follows standard normal distribution.

Both $\hat{p_1}$ and $\hat{p_2}$ are estimators of p, then we estimate $\hat{p} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}$

Pooled Test

Let $X_1^{(i)} .. X_{n_i}^{(i)}, i = 1, 2$ be random samples of size n_i from two Bernoulli distributions with parameters p_i and $\hat{p}_i = \overline{X}_i$ for corresponding sample means.

The test $H_0 : p_1 = p_2$ based on statistic $Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ is called pooled large-sample test for equality of proportions.

We reject H_0 at significance level α

- in favor of $H_1 : p_1 - p_2 \neq (p_1 - p_2)_0$ if $|Z| > z_{\alpha/2}$
- in favor of $H_1 : p_1 - p_2 > (p_1 - p_2)_0$ if $Z > z_\alpha$
- in favor of $H_2 : p_1 - p_2 < (p_1 - p_2)_0$ if $Z < -z_\alpha$