Statistical Methods for Decision Making
Residency-II

For BABI Program[1]

[1] Great Lakes Institute of Management, Gurgaon

February 18, 2018

GREAT LAKES INSTITUTE OF MANAGEMENT

- Revisited Sampling Distribution and CLT
- Point estimate and Interval Estimate
- Confidence Interval
- How to set up Null and alternative hypothesis
- Confidence interval approach of hypothesis testing
- Test statistic approach of hypothesis testing
- p-value approach of hypothesis testing

Note: Hypothesis Testing uses Sampling Distribution

GREAT LAKES INSTITUTE OF MANAGEMENT

- Applications to Cash Transfer Study (including test of variance)
- Type-I error, Type-II error
- Power of the Test
- Test of association
- Poisson Distribution (if time permits)

A **Type I error** occurs if we reject the null hypothesis H_0 (in favor of the alternative hypothesis H_A) when the null hypothesis H_0 is true. We denote $\alpha = P(\mathit{Type\ I\ Error})$.

A **Type II error** occurs if we fail to reject the null hypothesis H_0 when the alternate hypothesis H_A is true. We denote $\beta = P(\mathit{Type\ II\ Error})$.

The **power of a hypothesis test** is the probability of making the correct decision if the alternative hypothesis is true. That is, the power of a hypothesis test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_A is the hypothesis that is true.

1. Minimize the probability of committing a Type I error. That, is minimize $\alpha = P(\textit{Type I Error})$. Typically, a significance level of $\alpha \leqslant 0.05$ is desired.

2. Maximize the power (at a value of the parameter under the alternative hypothesis that is scientifically meaningful). Typically, we desire power to be 0.80 or greater. Alternatively, we could minimize $\beta = P(\textit{Type II Error})$, aiming for a Type II error rate of 0.20 or less.

Say, you hypothesize that Below Poverty Level Families consume on average 2000 calories under public distribution system. So, your null hypothesis is

$$H_0 : \ \mu = 2000 \ \textit{calories}$$

The Alternative Hypothesis is:

$$H_A : \ \mu = 2400 \ \textit{calories}$$

Cheat Sheet for Hypothesis Testing

1	Type of test	Test of mean μ or test of proportion p
2	Type of sample	one sample or two sample
3	Specify Null as	Status Quo
4	Specify Alternate as	Whatever you want to prove
5	Tails	If $H_a < \mu$ then left tail, $H_a > \mu$ is right tail , not equal is both
6	Confidence level	Fix the alpha (probability of type—I error)
7	Is sigma known? and/or Is $n > 30$?	If answer "Yes" to both , then Z otherwise t test
8	"critical values" or areas	Based on 6 and 7, find "critical values" or areas that define Left, Right or Both tails
9	Acceptance Region	Shade the area where H_a is "true"
10	Test Statistic	Calculate test statistics using fomula
11	Inference	If test statistic in the above step is in shaded portion then H_a holds, not otherwise

Reference:
`http://www.math.wayne.edu/~bert/courses/1020/hypothesis.testing.pdf`

Test For	*Null Hypothesis (Ho)*	*Test Statistic*	*Distribution*	*Use When*
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$	Standard normal (Z)	n is at least 30
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$	t_{n-1}	n is less than 30
Population proportion (p)	$p = p_0$	$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Standard normal (Z)	$n \times p_0$ and $n(1-p_0)$ are at least 5
Difference of two population means ($\mu_x - \mu_y$)	$\mu_x - \mu_y = 0$	$\frac{(\bar{x}-\bar{y})-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	Standard normal (Z)	n_1 and n_2 are both at least 30
Mean of difference (before – after)	$\mu_d = 0$	$\frac{\bar{d}-0}{s/\sqrt{n}}$	Standard normal (Z)	30 or more pairs of data
Mean of difference (before – after)	$\mu_d = 0$	$\frac{\bar{d}-0}{s/\sqrt{n}}$	t_{n-1}	Less than 30 pairs of data
Difference of two population proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1-\hat{p}_2)-0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_1}\right)}}$	Standard normal (Z)	$n \times \hat{p}$ and $n(1-\hat{p})$ are at least 5 for each group

Hypothesis testing

The R-Command for t-Test:

$t.test(x, y = NULL, alternative = c(\text{“}two.sided\text{”}, \text{“}less\text{”}, \text{“}greater\text{”}), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)$

Source: `https://www.rdocumentation.org/packages/stats/versions/3.4.3/topics/t.test`

$prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, correct = TRUE)$

Source: `https://www.rdocumentation.org/packages/stats/versions/3.4.3/topics/prop.test`

Definition: The chi-squared distribution with r degrees of freedom is the distribution of a random variable that is the sum of the squares of r independent standard normal random variables. WeÊ$\frac{1}{4}$ll call this distribution $\chi^2(r)$.

Definition: The mean of the chi-squared distribution is r (the same as degrees of freedom) and the variance is $2 \star r$.

Note: the value of the chi square statistic is sensitive to the sample size. Refer (`http://uregina.ca/~gingrich/ch11a.pdf` and `http://uregina.ca/~gingrich/ch10.pdf`

The chi-square statistic, denoted with the Greek χ^2, is found by comparing the observed counts from a sample with expected counts derived from a null hypothesis. The formula for computing the statistic is

$$\chi^2 = \sum \frac{(Obeserved - Expected)^2}{Expected}$$

where the sum is over all cells of the table.

Chi-square Goodness-of-Fit Test

To test a hypothesis about the proportions of a categorical variable, based on a table of observed counts in k cells:

H_0: Specifies proportions, p_i, for each cell

H_a: At least one p_i is not as specified

- Compute the expected count for each cell using $n \cdot p_i$, where n is the sample size and p_i is given in the null hypothesis.
- Compute the value of the chi-square statistic,

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

- Find the p-value for χ^2 using the upper tail of a chi-square distribution with $k-1$ degrees of freedom.

The chi-square distribution is appropriate if the sample size is large enough that each of the expected counts is at least 5.

Chi-square goodness—of—fit test

Chi-square Test for Association

To test for an association between two categorical variables, based on a two-way table that has r rows as categories for variable A and c columns as categories for variable B:

Set up hypotheses:

$$H_0 : \text{Variable A is not associated with variable B}$$

$$H_a : \text{Variable A is associated with variable B}$$

Compute the expected count for each cell in the table using

$$\text{Expected count} = \frac{\text{Row total} \cdot \text{Column total}}{\text{Sample size}}$$

Compute the value for a chi-square statistic using

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Find a p-value using the upper tail of a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

The chi-square distribution is appropriate if the expected count is at least five in each cell.

Happy Learning!!!