Chapter 6 Inference for categorical data

Huamei Dong

03/15/2016

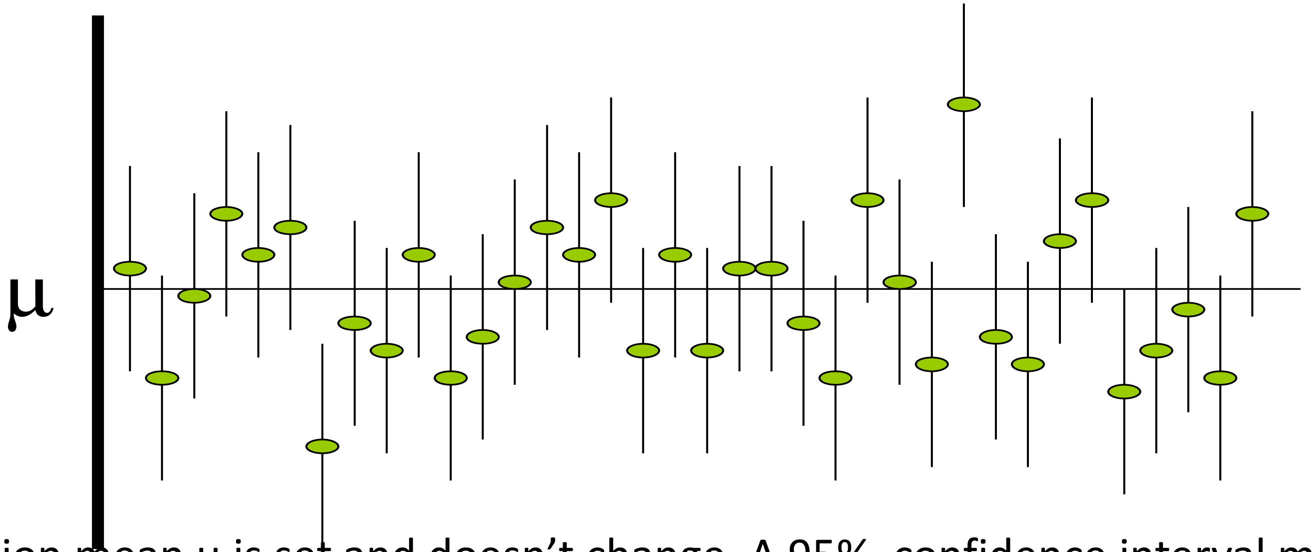
1. **Quick Summary**
2. **Sample Proportion** $\hat{p}$
3. **Sampling distribution of**
4. **Confidence interval for one proportion**
5. **Hypothesis test for one proportion**

1. Quick summary about confidence interval and hypothesis test

1.1 Confidence interval
1.2 Hypothesis test
1.3 The relation between Z test and T test
1.4 The relation between one sided test and two sided test
1.5 The relation between confidence interval and hypothesis test
1.6 The relation between type I error and type II error
1.7 Compare one population mean with a number
1.8 Compare two population means

1.1 Interpretation of confidence intervals

What does 95% confidence interval mean?



(1) The population mean μ is set and doesn’t change. A 95% confidence interval means that if we were to select 100 random samples from the population and use these 100 samples to calculate 100 different confidence intervals for μ, approximately 95 of the intervals would cover the true population mean. Therefore for a specific confidence interval from one random sample, we have 95% chance to capture the true population mean.

(2) We would like our confidence interval to capture μ. The wider our confidence interval is, the more possible we are going to capture μ and therefore the higher the confidence level will be.

1.2 Hypothesis test

The logic behind the hypothesis test:
(1) Assume H_0 is true. You calculate the probability of obtaining your sample data.

(2)If this probability is small, that is, p-value is small, (how small is called small? Compare to the significance level. Usually it is 0.05, 0.1 etc.), then

(a) we think this probability is negligible or equal to zero.

(b) Zero probability of obtaining your sample data means it is impossible to obtain your sample data. But now you did observe your sample data. Contradiction!

(c) So something is wrong. All you reasoning is correct. The only possible wrong doing is your assumption. So your assumption “H_0 is true” is wrong. So H_A is right.

(3) If p-value is not negligible comparing to the significance level, there is no contradiction. This doesn’t mean H_0 is right. So we can only prove H_0 wrong (i.e., H_A right), we can’t prove H_0 right.

(4) The analogy:

(a) If you go to school, I go to school. (If null is true, p-value should be big.)

(b) If I do not go to school, that implies you do not go. (If p-value is not big or **small**, that implies null is not true or **wrong**.)

(c) If I go, that doesn't imply you go.(If p-value is big, that doesn't imply null is true.)

(5) We use p-value to decide if we reject H_0 (accept H_A) or not. So our p-value should be related to our H_0 and H_A .

(6) The more our Z value or T value is favorable to H_A, the stronger evidence we have to reject H_0 and prove H_A.

1.3 The relation between Z test and T test

(a) When population distribution is nearly normal, population mean is μ and population standard deviation is σ. No matter sample size big or small, we have $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}=Z\sim N(0,1)$ that is $\overline{X}\sim N(\mu,\frac{\sigma}{\sqrt{n}})$ or Z test.

(b) When population distribution is nearly normal, population mean is μ, usually in reality population standard deviation σ is unknown, we have to use sample standard deviation *S* to estimate σ.

$$\frac{\overline{X}-\mu}{s/\sqrt{n}}=t_{n-1}$$

(c) When population distribution is nearly normal, population mean is μ, standard deviation σ is unknown. If sample size is large enough (usually larger than 30) , we think *S* will be close enough to σ. Then $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ *is almost Z*

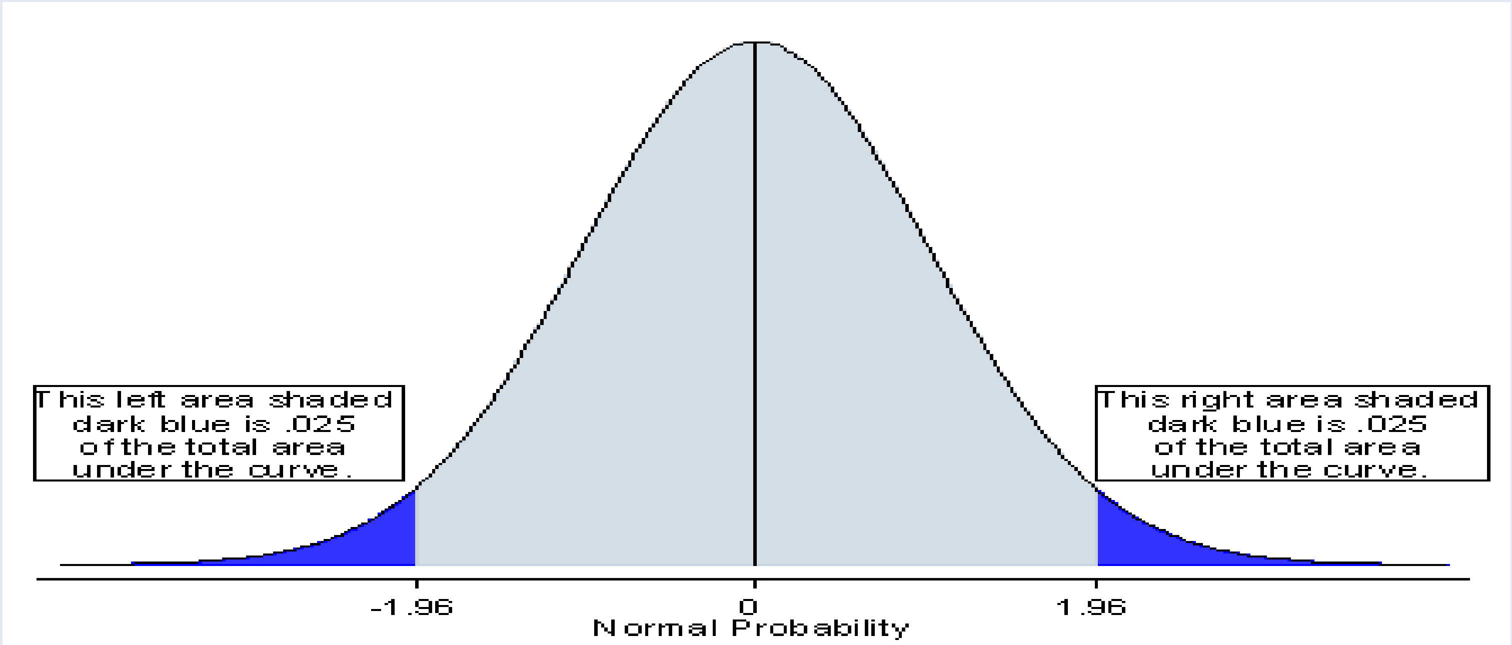
(d) When population distribution is not nearly normal, population mean is μ, standard deviation is σ. When sample size is large, as long as it is not too skewed, by central limited theorem,

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \quad is\ \ almost\ \ Z$$

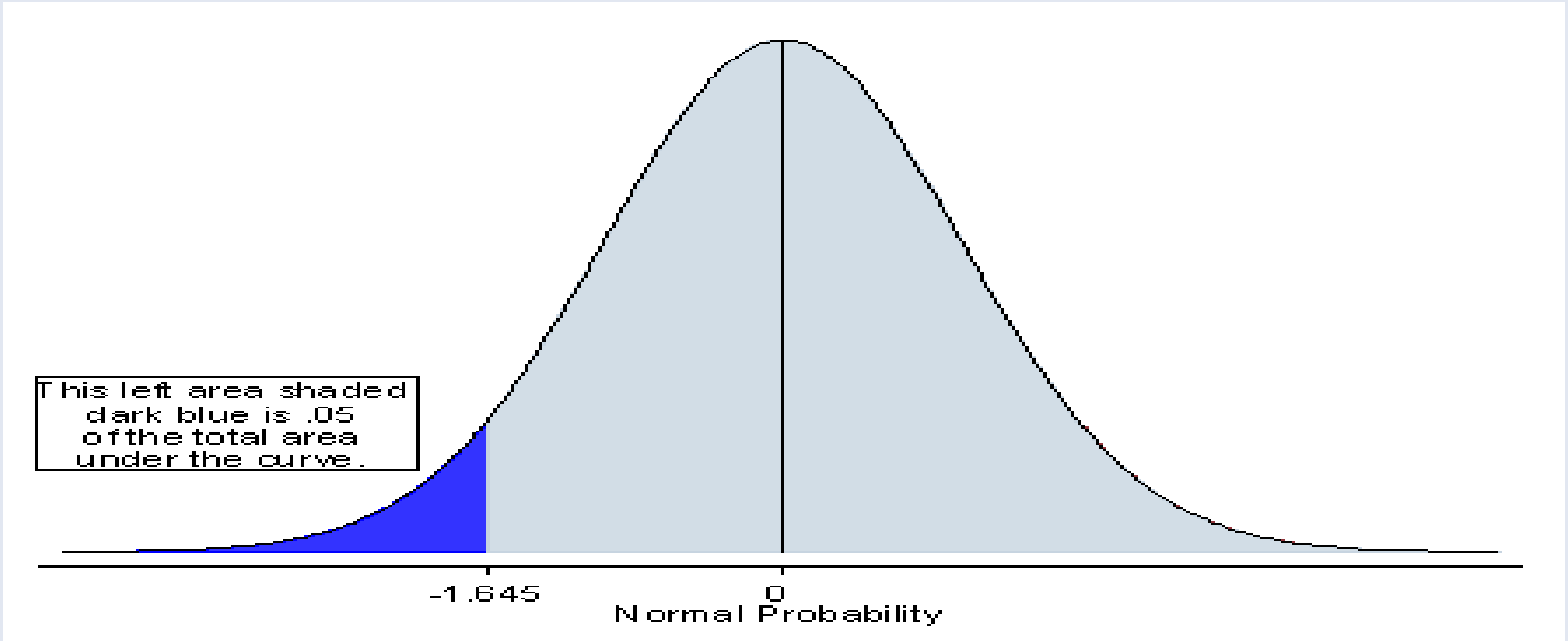
(e) When population distribution is not nearly normal, population mean is μ, standard deviation σ is unknown. When sample size is large, we think *S* will be close enough to σ. Also because sample size is large, as long as it is not too skewed, by central limited theorem,

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \quad is\ \ almost\ \ Z$$

1.4 The relation between one-sided test and two sided test


This left area shaded dark blue is .025 of the total area under the curve.

This right area shaded dark blue is .025 of the total area under the curve.


This left area shaded dark blue is .05 of the total area under the curve.

1.5 The relation between confidence interval and hypothesis test

(1) There is actually mathematical equivalence between confidence intervals and test of hypothesis. For instance, for a two-sided Z test, any value of Z that is between -1.96 and 1.96 would result in a p-value greater than 0.05 and the null hypothesis would not be rejected. On the other hand, H_0 would be rejected for any value of z that is either less than -1.96 or greater than 1.96.

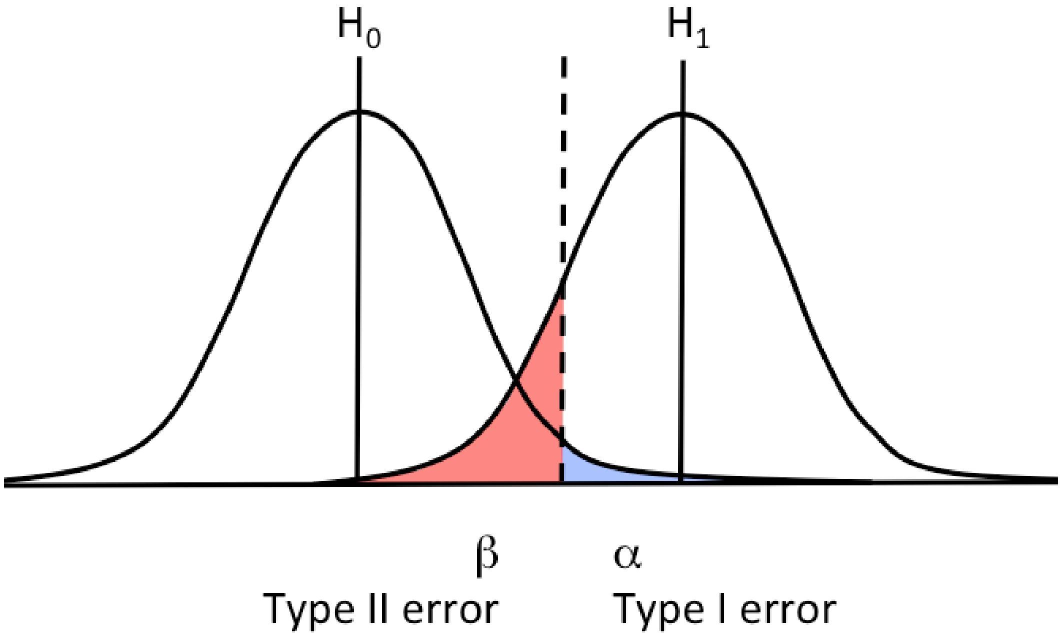
(2) The Z value is calculated using $\overline{X},\ \mu\ (actually\ \mu_0\ because\ we\ assume\ H_0\ is\ true), s\ and\ n.$ Then we compare Z value with -1.96 and 1.96.

(3) On the other hand, the true population mean μ calculated from the inequality

$-1.96 \leq Z = \frac{\overline{X} - \mu}{s/\sqrt{n}} \leq 1.96\ using\ \overline{X},\ s,\ n,\ -1.96\ and\ 1.96$ provides us 95% confidence interval.

1.6 Type I error and Type II error

When we do hypothesis test, we are making assumption for parameter using the sample data we have. Similar to confidence interval, we make errors in hypothesis tests. There are two kinds of error involved. One is Type 1 error and another one is Type 2 error.



The smaller type I error has to be obtained with the price of bigger type II error.

1.7 Compare one population mean with a number

When the null hypothesis is that the population mean equals to some number, for example, in a Z test,

$$H_0: \quad \mu=2,$$

then calculate Z as below

$$Z=\frac{\overline{X}-2}{s/\sqrt{n}}$$

and find p-value.

1.8 Compare two population means

When the null hypothesis is that one population mean equals to another population mean, for example, in a T test,

$$H_0 : \mu_1 = \mu_2,$$

then we calculate T using

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

and find p-value.

2. Sample proportion

Example 1 Find the sample proportion for the following data. Here 1 represents “success” and 0 represents “failure”.

0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,1,0,0,1,0,0,1,1,0,0,1,1,1,0,0

Answer:

$$\hat{p} = \frac{11}{30} = 0.367$$

3. Sampling distribution of $\hat{p}$

The sampling distribution of $\hat{p}$:

Assume the true population proportion p and sample size is n . If

(1) The sample observations are independent.

(2) $np \geq 10 \;\; and \;\; n(1-p) \geq 10$

Then the sampling distribution for $\hat{p}$ is nearly normal with mean p and standard error

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

4. Confidence Interval for a proportion

Constructing a confidence interval for a proportion.

(1) Verify the observations are independent and verify the success-failure condition using $\hat{p}$ and n .

(2) If the condition are met, the sampling distribution of $\hat{p}$ is nearly normal.

(3) Standard error can be approximated by

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

(4) Confidence interval is

$$\left(\hat{p} - z^{*}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \,,\ \hat{p} + z^{*}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Example2 :Use the data in Example 1, find a 95% confidence interval for P.

0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,1,0,0,1,0,0,1,1,0,0,1,1,1,0,0

Answer:

$$\hat{p} = \frac{11}{30} = 0.367$$

Verify condition is satisfied.

$$SE_{\hat{p}} = \sqrt{\frac{(0.367)(1-0.367)}{30}} = 0.088$$

Confidence interval is (0.367-1.96X0.088, 0.367+1.96X0.088).

That is (0.195, 0.539)

Example 3 In a study conducted to investigate the nonclinical factors associated with method of surgical treatment received for early-stage breast cancer, some patients underwent a modified radical mastectomy while others had a partial mastectomy accompanied by radiation therapy. (breast_cancer.xls)

(1) Construct a 95% confidence interval for the proportion of women under 55 who underwent a modified radical mastectomy.

(2) Construct a 95% confidence interval for the proportion of women under 55 who underwent a partial mastectomy accompanied by radiation therapy.

(3) Test whether the proportion of women under 55 who underwent a modified radical mastectomy is 0.2 at significance level of 0.05.

(4) Test whether the proportion of women under 55 who underwent a partial mastectomy accompanied by radiation therapy is 0.2 at significance level of 0.05.

```
Answer: > breast<-read.table(“breast_cancer.txt”, header=T, as.is=T, sep=“\t”)
        > table(breast)
```

```
           age
treatment  <55 >=55
  partial  292  366
  radical  397 1183
```

(1) The sample is random. The sample proportion is $\hat{p} = \frac{397}{1580} = 0.251$

The success-failure condition is satisfied because

$$n\hat{p} = 397 \quad and \quad n(1-\hat{p}) = 1183.$$

The standard error is

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.251(1-0.251)}{1580}} = 0.011$$

So a 95% confidence interval is (0.251-1.96x0.011, 0.251+1.96x0.011)=(0.229, 0.273)

(3) Now we need conduct the following hypothesis test (two-sided test):

H_0: p=0.2 vs. H_A: p≠0.2

(a) Check conditions: The sample is random.

The success-failure condition(satisfied): $np_0=1580(0.2)=316 \text{ and } n(1-p_0)=1580(0.8)=1264$

(b) Calculate standard error

$$SE_{\hat{p}}=\sqrt{\frac{p_0(1-p_0)}{n}}=\sqrt{\frac{(0.2)(0.8)}{1580}}=0.01$$

(c) Calculate Z value:

$$Z=\frac{point\ \ estimate-null\ \ value}{SE}=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}=\frac{0.251-0.2}{0.01}=5.1$$

(d) Using R to get the p-value is less than 0.001. So we reject $H_{0.}$ That is, the proportion of women under 55 who underwent a modified radical mastectomy is not 0.2.

Homework:

1. Finish part (2) and (4) in Example 3 and interpret your results

2. Please the data you got from our class to do the following hypothesis test: H_0: $p=0.4$, H_A: $p \neq 0.4$

and interpret your results.