Lab

Huamei Dong

04/12/2016

1. Z test or T test for one mean (one sample) or two means (two samples)
2. Chi square test for two categorical data
3. ANOVA F test for comparing two means or more than two means
4. T test for simple linear regression slope
5. ANOVA F test for simple linear regression slope
6. Sample size calculation

1. Z test or T test for comparing two means

```
>birth<-read.table(“births.txt”, as.is=T, header=T, sep=“\t”)
>birth_smoker<-subset(birth,smoke=="smoker”)
>birth_nonsmoker<-subset(birth,smoke=="nonsmoker”)
>hist(birth$weight)
>hist(birth_smoker$weight)
>hist(birth_nonsmoker$weight)
```

> **t.test(birth$weight~birth$smoke,var.equal=T)**

```
        Two Sample t-test

data:  birth$weight by birth$smoke
t = 1.5517, df = 148, p-value = 0.1229
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1095531  0.9105531
sample estimates:
mean in group nonsmoker    mean in group smoker
                 7.1795                  6.7790
```

> **t.test(birth$weight~birth$smoke,var.equal=F)**

```
        Welch Two Sample t-test

data:  birth$weight by birth$smoke
t = 1.4967, df = 89.277, p-value = 0.138
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1311663  0.9321663
sample estimates:
mean in group nonsmoker    mean in group smoker
                 7.1795                  6.7790
```

>**t.test(birth$weight~birth$smoke)**

```
        Welch Two Sample t-test

data:  birth$weight by birth$smoke
t = 1.4967, df = 89.277, p-value = 0.138
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1311663  0.9321663
sample estimates:
mean in group nonsmoker    mean in group smoker
                 7.1795                  6.7790
```

> **t.test(birth_smoker$weight,birth_nonsmoker$weight,var.equal=T)**

```
        Two Sample t-test

data:  birth_smoker$weight and birth_nonsmoker$weight
t = -1.5517, df = 148, p-value = 0.1229
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9105531  0.1095531
sample estimates:
mean of x mean of y
   6.7790    7.1795
```

> **t.test(birth_smoker$weight,birth_nonsmoker$weight,var.equal=F)**

```
        Welch Two Sample t-test

data:  birth_smoker$weight and birth_nonsmoker$weight
t = -1.4967, df = 89.277, p-value = 0.138
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9321663  0.1311663
sample estimates:
mean of x mean of y
   6.7790    7.1795
```

> **t.test(birth_smoker$weight,birth_nonsmoker$weight)**

```
        Welch Two Sample t-test

data:  birth_smoker$weight and birth_nonsmoker$weight
t = -1.4967, df = 89.277, p-value = 0.138
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9321663  0.1311663
sample estimates:
mean of x mean of y
   6.7790    7.1795
```

Homework (1) Conduct two sample t test by hand and compare your result with this. Hint: Use r to calculate the sample mean and standard deviation for the weights from smokers and sample mean and standard deviation from nonsmokers. Then use what you have learned in Chapter 5 to find T statistics.

2. ANOVA F test for comparing two or more means

> **oneway.test(birth$weight~birth$smoke,var.equal=T)**

One-way analysis of means

data: birth$weight and birth$smoke
F = 2.4077, num df = 1, denom df = 148, p-value = 0.1229

> **oneway.test(birth$weight~birth$smoke,var.equal=F)**

One-way analysis of means (not assuming equal variances)

data: birth$weight and birth$smoke
F = 2.2401, num df = 1.000, denom df = 89.277, p-value = 0.138

>**oneway.test(birth$weight~birth$smoke)**

One-way analysis of means (not assuming equal variances)

data: birth$weight and birth$smoke
F = 2.2401, num df = 1.000, denom df = 89.277, p-value = 0.138

3. Chi square test

```
>table1<-table(birth$sexBaby,birth$smoke)
>table1
        nonsmoker smoker
 female        49     19
 male          51      31
> chisq.test(table1)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table1
X-squared = 1.2139, df = 1, p-value = 0.2706
```

Homework (2): Conduct the chi-square test by hand and compare your result with this.

4. T test for Simple Linear Regression's slope

> **reg1<-lm(birth$weight~birth$smoke)**
> **summary(reg1)**

```
Call:
lm(formula = birth$weight ~ birth$smoke)
Residuals:
   Min     1Q  Median     3Q    Max
-5.5495 -0.5590  0.2605  0.9505  2.9505

Coefficients:
                     Estimate       Std. Error       t value       Pr(>|t|)
(Intercept)            7.1795         0.1490         48.178        <2e-16 ***
birth$smokesmoker     -0.4005         0.2581         -1.552         0.123
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.49 on 148 degrees of freedom
Multiple R-squared:  0.01601,       Adjusted R-squared:  0.009359
F-statistic: 2.408 on 1 and 148 DF,  p-value: 0.1229
```

Homework(3) : Conduct a T test to test whether the slope for smoker is zero by hand and compare your result with this. (Here response variable is numerical and explanatory is categorical)
Hint: For linear regression of **numerical** response variable against **categorical** explanatory data, the t test for the slope of simple linear regression is just like two sample t test with two samples having equal variances.

So you should use $s_{pooled}=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ to calculate standard error

$$SE=\sqrt{\frac{S_{pooled}^2}{n_1}+\frac{S_{pooled}^2}{n_2}}$$

instead of using $SE=\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$.

Then you can calculate T statistics $T=\frac{\bar{x}_1-\bar{x}_2}{SE}$ and find p-value.

If the simple linear regression is **numerical** response variable against **numerical** explanatory variable, then you can used $\hat{\beta}_1=\frac{s_y}{s_x}R$ and $SE(\hat{\beta}_1)=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\sqrt{\frac{sum\ of\ residuals'\ squares}{(n-1)(n-2)(s_x)^2}}$

to calculate statistics $T=\frac{\hat{\beta}_1-0}{SE(\hat{\beta}_1)}$ (See Example 1 in lecture from April 7).

5. ANOVA F test for simple linear regression's slope

> **fit1<-aov(birth$weight~birth$smoke)**
> **summary(fit1)**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
birth$smoke	1	5.3	5.347	2.408	0.123
Residuals	148	328.7	2.221		

6. Sample size calculation

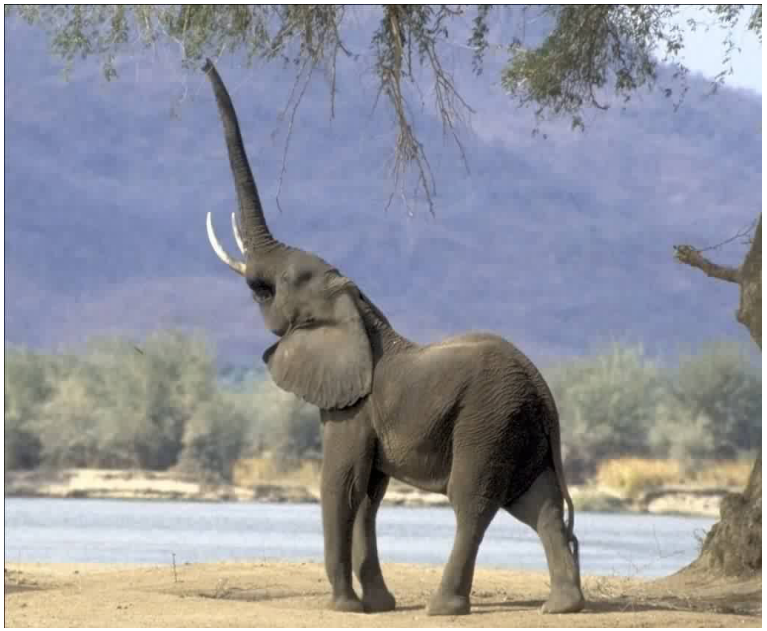
Sample size estimation can be estimated by confidence level, standard error and margin of error. For example, when you would like to sample a group of students in some university and measure their weight. You would like to the population mean weight in the university.

Suppose a 95% confidence interval for the true mean is

$$(\bar{x} - z^* SE,\ \ \bar{x} + z^* SE) = (\bar{x} - z^* \frac{s}{\sqrt{n}},\ \ \bar{x} + z^* \frac{s}{\sqrt{n}})$$

and you want your margin of error $z^* SE$ to be with in 5%. Then you can estimate your sample size using

$$1.96 \frac{s}{\sqrt{n}} \leq 0.05$$



Think fun:

(1)How can you weigh an elephant in the Zoo? You are provided with a huge wooden box (similar to a boat, but shape is rectangular prism), a big pond, a marker, a small scale, lots of pebbles)

(2) The relation between type I, type II error and crying wolf story.