Examples in R

R. Trîmbiţaş
"Babeş-Bolyai" University

March 16, 2016

1 Inferences about one population

z-test

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

t-test

$$T = \frac{\bar{X} - m}{\frac{s'}{\sqrt{n}}} = \frac{\bar{X} - m}{\sqrt{\frac{\bar{m}_2}{n-1}}}$$

has a Student distribution with $n-1$ dfs.

z-test for proportions (large sample)

$$Z = \frac{k - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

χ^2 test for variance

$$X^2 = \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2 = \frac{(n-1)s'^2}{\sigma^2} \tag{2}$$

has a χ^2 distribution with $n-1$ dfs.

Problem 1 *The lifetime of 40 horses is given by the list:*

23.7	23.2	25.0	25.0	21.3	23.9	23.5	26.0
27.5	27.5	21.3	24.2	20.2	20.5	24.0	28.8
21.6	25.2	23.3	27.5	20.6	24.1	25.7	27.5
28.9	24.3	19.3	21.7	20.6	31.4	22.1	26.4
23.4	26.8	21.6	19.6	19.5	25.5	23.4	23.4

(a) *Does the selection allow us to say that the mean is 24 (or is* $\neq 24$*),* $\alpha = 5\%$*?*

(b) *Test the hypothesis $\sigma^2 = 10$ versus the two-sided alternative.*

(c) *Find the 90% confidence interval for mean and variance.*

Solution. Data in file `horses.dat`; solution in `horsestst.R`.

```
> horses<-scan("horses.dat");
Read 40 items
> library(TeachingDemos)
> z.test(horses,mu=24,sd=sd(horses))
One Sample z-test
data:  horses
z = -0.054674, n = 40.00000, Std. Dev. = 2.89200, Std. Dev. of the
sample mean = 0.45726, p-value = 0.9564
alternative hypothesis: true mean is not equal to 24
95 percent confidence interval:
23.07879 24.87121
sample estimates:
mean of horses
23.975
> sigma.test(horses,sigmasq=10)
One sample Chi-squared test for variance
data:  horses
X-squared = 32.617, df = 39, p-value = 0.4904
alternative hypothesis: true variance is not equal to 10
95 percent confidence interval:
5.61209 13.78923
sample estimates:
var of horses
8.363462
> z.test(horses,mu=24,sd=sd(horses),conf.level=0.90)
One Sample z-test
data:  horses
z = -0.054674, n = 40.00000, Std. Dev. = 2.89200, Std. Dev. of the
sample mean = 0.45726, p-value = 0.9564
alternative hypothesis: true mean is not equal to 24
90 percent confidence interval:
23.22287 24.72713
sample estimates:
mean of horses
23.975
> sigma.test(horses,sigmasq=10,conf.level=0.90)
One sample Chi-squared test for variance
data:  horses
X-squared = 32.617, df = 39, p-value = 0.4904
alternative hypothesis: true variance is not equal to 10
90 percent confidence interval:
```

```
5.976941 12.693911
sample estimates:
var of horses
8.363462
> t.test(horses,mu=24)
One Sample t-test
data:  horses
t = -0.054674, df = 39, p-value = 0.9567
alternative hypothesis: true mean is not equal to 24
95 percent confidence interval:
23.0501 24.8999
sample estimates:
mean of x
23.975
```

■

Problem 2 *A manufacturer of gunpowder has developed a new powder, which was tested in eight shells. The resulting muzzle velocities, in meters per second, were as follows:*

1001.7	*975.0*	*978.33*	*988.33*
998.33	*1001.7*	*979.0*	*968.33*

Find a 95% confidence interval for the true average velocity μ for shells of this type. Assume that muzzle velocities are approximately normally distributed. The manufacturer claims that the new gun powder produce an average velocity of not less than 1000 m/s. Do the sample data provide sufficient evidence to contradict the manufacturer's claim at 0.025 level of significance?

Solution. See file `bullets.R`. ■

Problem 3 *Researchers have shown that cigarette smoking has a deleterious effect on lung function. In their study of the effect of cigarette smoking on the carbon monoxide diffusing capacity (DL) of the lung, Ronald Knudson, W. Kaltenborn and B. Burrows found that current smokers had DL readings significantly lower than either ex-smokers or nonsmokers.The carbon monoxide diffusing capacity for a random sample of current smokers was as follows:*

103.750	*88.602*	*73.003*	*123.086*	*91.052*
92.295	*61.675*	*90.677*	*84.023*	*76.014*
100.615	*88.017*	*71.210*	*82.115*	*89.222*
102.754	*108.579*	*73.154*	*106.755*	*90.479*

Do these data indicate that the mean DL reading for current smokers is lower than 100, the average DL reading for nonsmokers?

Problem 4 *Ten human fossiles unearthed during an archaeological digging were classified as 6 females and 4 males. Do these data confirm the assertion that the sexes ratio is 1/1 ($\alpha = 1\%$).*

Solution. Use the binomial test:

```
> binom.test(6,10,0.5,"greater",conf.level=0.99)
Exact binomial test
data:  6 and 10
number of successes = 6, number of trials = 10, p-value = 0.377
alternative hypothesis: true probability of success is greater than 0.5
99 percent confidence interval:
0.2183385 1.0000000
sample estimates:
probability of success
0.6
```

■

Problem 5 *The Romanian National Authority for Consumer's Protection claim that an amount of 15% of beans from a controlled sample has bean weavils. To check the claim for $\alpha = 10\%$ one considered a sample of 200 bean seeds and 7 seeds had weavils. Is the claim of OPC in doubt?*

Proof. χ^2-test for proportions or z-test.

```
prop.test(7, 200, p=.15, alt="less", correct=T)
1-sample proportions test with continuity correction
data:  7 out of 200, null probability 0.15
X-squared = 19.853, df = 1, p-value = 4.182e-06
alternative hypothesis: true p is less than 0.15
95 percent confidence interval:
0.00000000 0.06647341
sample estimates:
p
0.035
> p0<-0.15
> Z<-(17/200-0.15)/sqrt(p0*(1-p0)/200);Z
[1] -2.574384
> p<-pnorm(Z);p
[1] 0.005020943
> qnorm(0.1)
[1] -1.281552
```

■

2 Inferences about two populations

z-test, known variances or large samples

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3}$$

t-test, $\sigma_1^2 = \sigma_2^2$, equal variances

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{4}$$

where

$$S_p = \sqrt{\frac{(n_1 - 1)s_1'^2 + (n_2 - 1)s_2'^2}{n_1 + n_2 - 2}}.$$

T statistics has a Student distribution with $n = n_1 + n_2 - 2$ degrees of freedom.

t-test, $\sigma_1^2 \neq \sigma_2^2$, different variances

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}}, \tag{5}$$

has approximately a Student distribution with n dfs, where n is computed from

$$\frac{1}{n} = \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1}, \tag{6}$$

and

$$c = \frac{\frac{s_1'^2}{n_1}}{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}. \tag{7}$$

(*Saitherwaite's approximation*)

F-test or Snedecor-Fisher test

$$F = \frac{s_1'^2}{s_2'^2} \tag{8}$$

has an(a) F (Snedecor-Fisher) distribution with $(n_1 - 1, n_2 - 1)$ dfs.

Problem 6 *At a large university, a mathematics placement exam is administered to all students. Samples of 36 male and 30 female students are randomly selected from this year's student body and the following scores recorded:*

Male	72	68	75	82	81	60	75	85	80	70	
	71	84	68	85	82	80	54	81	86	79	
	99	90	68	82	60	63	67	72	77	51	
	61	71	81	74	79	76					*:*
Female	81	76	94	89	83	78	85	91	83	83	
	84	80	84	88	77	74	63	69	80	82	
	89	69	74	97	73	79	55	76	78	81	

(a) Describe each set of data with a histogram (use the same class intervals on both histograms), the mean, and standard deviation.

(b) Construct 95% confidence interval for the mean score for all male students. Do the same for all female students.

(c) *Do the results found in part (b) show that the mean scores for males and females could be the same? Justify your answer. Be careful!*

(d) *Construct the 95% confidence interval for the difference between the mean scores for male and female students.*

(e) *Do the results found in part (d) show the mean scores for male and female students could be the same? Explain.*

(f) *Explain why the results in part b cannot be used to draw conclusions about the difference between the two means.*

Solution. See file `students.R`. ■

Problem 7 *Two methods were used to study the latent heat of ice fusions. Both methods A (an electrical method) and method B (a method of mixtures) were conducted with the specimen cooled to -0.72° C. The data represents the change in total heat from -0.72° C to water at 0° C, in calories per gram of mass.*

Method A	*Method B*
80.04	80.03
80.03	79.98
80.04	79.94
80.01	79.95
80.03	79.88
80.02	79.93
80.01	79.95
80.00	79.99
80.05	
80.05	
79.98	
80.00	
80.02	

(a) *Is there a significant difference in the mean values at the 0.05 level?*

(b) *Find a 95% confidence interval for the difference of mean and a 90% confidence interval for the ratio of variance.*

(c) *Are the variances equal ($\alpha = 90\%$)?*

Problem 8 *Two brands of paint are to be tested. Brand A is less expensive than brand B. Several chips are painted and exposed to weather conditions for a period of six months. Each chip is then judged on several qualities and a score is determined. The paint samples scored (higher scores are better) are shown in the accompanying table.*

Paint A	84	86	91	93	84	88	
Paint B	90	88	92	94	84	85	92

(a) *What brand we choose?*

(b) Is the variance of first brand less than the variance of the second brand?

Problem 9 *Ten soldiers were selected at random from each of two companies to participate in a rifle-shooting competition. Their scores are shown in the accompanying table. Can you conclude that company B has a higher mean score than company A? ($\alpha = 5\%$)*

Company A	72	29	62	60	68	59	61	73	38	48
Company B	75	43	63	63	61	72	73	82	47	43

What can you say about variances?

Problem 10 *To test the effects of sport hours on students, we employ a pair of tests at beginning and at the end of semester respectively. The score is given in the table. Is there any improvement?*

Before	*29*	*22*	*25*	*29*	*26*	*24*	*31*	*46*	*34*	*28*
After	*30*	*26*	*25*	*35*	*33*	*36*	*32*	*54*	*50*	*43*