Exercise 13

Applied Statistics 2021, IT University of Copenhagen

Preparation

- Read pages 283–284, 305–307, 321–325 from Verzani (2014).

Exercises

1. One sample t-test (T)

We perform a t-test for the null hypothesis $H_0 : \mu = 10$ by means of a dataset consisting of $n = 16$ elements with sample mean 11 and sample variance 4. We use significance level 0.05.

a. Should we reject the null hypothesis in favor of $H_1 : \mu \neq 10$?

Solution: We assume that the dataset is a realization of a random sample from an $N(\mu, \sigma^2)$ distribution. As a test statistics we will use

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

$$S_n = \sqrt{Var(X)} = 2$$
$$t = \frac{11 - 10}{2/\sqrt{16}} = 2$$

Under the null hypothesis the test statistic T has a $t(n-1)$ distribution. So our decision rule with significance level α is to reject H_0 in favor of H_1 whenever $T \leq -t_{n-1,\alpha/2}$ or $T \geq t_{n-1,\alpha/2}$.

Since $t_{n-1,\alpha/2} = 2.13$, we can not reject our H_0 hypothesis as $t_{n-1,\alpha/2} > t$

```
qt(1-0.025, 16-1)
```

```
## [1] 2.13145
```

We can also calculate the one-tailed p-value by calculating the $t(n-1)$ distribution $P(T \geq t) = 0.03$, which is larger than our confidance level $\alpha/2$.

```
pt(2, 16-1, lower.tail = FALSE)
```

```
## [1] 0.0319725
```

b. What if we test against $H_1 : \mu > 10$?

Solution: We calculate one-tailed t-test using the same formula:

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

Under the null hypothesis the test statistic T has a $t(n-1)$ distribution. So our decision rule with significance level α is to reject H_0 in favor of H_1 whenever $T \leq -t_{n-1,\alpha}$ or $T \geq t_{n-1,\alpha}$.

$t_{n-1,\alpha} = 1.75$ which tells us that we must reject our H_0 hypothesis as $t_{n-1,\alpha} < t$

```
qt(1-0.05, 16-1)
```

```
## [1] 1.75305
```

We can also calculate the one-tailed p-value by the $t(n-1)$ distribution distribution $P(T \geq t) = 0.03$, which is smaller than our confidence level α.

```
pt(2, 16-1, lower.tail = FALSE)
```

```
## [1] 0.0319725
```

2. Easter Eggs (T)

Assume that you got six similar Easter eggs with 20g of chocolate reported in each. After taking one more lecture in Applied Statistics, you want to further investigate whether it is plausible that the eggs really contain 20g chocolate or if the egg producer is cheating. You weight the eggs and obtain the following six observations for the chocolate weight:

Chocolate contents (g)
20.1, 19.1, 18.2, 20.2, 19.6, 19.1

You may assume that you measurement is a realization of a random sample from a normal distribution $N(\mu, \sigma^2)$, where μ represents the true average contents.

(a) Formulate the appropriate null hypothesis and alternative hypothesis.

Solution: Our null hypothesis is that the expected weight μ is 20, $H_0 : \mu = 20$.

We want to investigate if the producer is cheating us, so the alternative hypothesis must be if there is less than 20g of chocolate in the candy; $H_1 : \mu < 20$

(b) Which test is appropriate for testing the hypothesis? Explain why.

Solution: We can use a one-tailed t-test, as the measurements are a realization of a random sample from a normal distribution $N(\mu, \sigma^2)$. So, we use

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

(c) Compute the value of the test statistic and report your conclusion at significance level $\alpha = 0.05$.

Solution: First we compute sample average and standard deviation and use it to get the t test value:

$$\bar{X}_n = \frac{20.1 + 19.1 + 18.2 + 20.2 + 19.6 + 19.1}{6} = 19.3833$$

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} = \sqrt{\frac{1}{6}\sum_{i=1}^{6}(x_i - 19.3833)^2} = 0.7467708$$

$$t = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} = \frac{19.3833 - 20}{0.7467708/\sqrt{6}} = -2.022734$$

Under the null hypothesis the test statistic T has a $t(n-1)$ distribution. As we are using one-tailed test our decision rule with significance level α is to reject H_0 in favor of H_1 whenever $T \leq -t_{n-1,\alpha}$.

The left critical value is

$$t_{n-1,\alpha} = t_{5,0.05} = -2.015$$

, which tells us that we must reject our H_0 hypothesis as $t < -t_{n-1,\alpha}$, so the the manufacturer is seemingly cheating us.

```
qt(0.05,5)
```

```
## [1] -2.015048
```

(d) Compute the corresponding left tail p-value. It is likely to observe these measurements under the null hypothesis?

Solution: The left tail p-value $P(T \leq t) = 0.05$ shows that we are likely to observe measurements at least this extreme in 5% of the cases under the null hypothesis.

```
pt(-2.022734, 6-1)
```

```
## [1] 0.04951219
```

3. Two-sample t-test (T, Home Exercise)

The data in Table 28.3 (pp. 425, Dekking et al. (2010)) represents salaries (in pounds Sterling) in 72 randomly selected advertisements in The Guardian (April 6, 1992). When the range was given in the advertisement, the midpoint of the range is reproduced in the table. The data are salaries corresponding to two kinds of occupations ($n = m = 72$): (1) Creative, media, and marketing and (2) education. The sample mean and sample variance of the two datasets are, respectively:

$$\begin{aligned}
&(1)\ \bar{x}_{72} = 17410 \text{ and } s_x^2 = 41258741,\\
&(2)\ \bar{y}_{72} = 19818 \text{ and } s_y^2 = 50744521,
\end{aligned}$$

Suppose that the datasets are modeled as realizations of normal distributions with expectations μ_1 and μ_2, which represent the salaries for occupations (1) and (2).

a. Test the null hypothesis that the salary for both occupations is the same at level $\alpha = 0.05$ under the assumption of equal variances. Formulate the proper null and alternative hypotheses, compute the value of the test statistic, and report your conclusion.

Solution: First we formulate the null and alternative hypotheses:

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$.

Using formula from the textbook (pp. 417, Dekking et al. (2010)) we find the pooled-variance:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}(\frac{1}{n}+\frac{1}{m}) = \frac{(72-1)\cdot 41258741 + (72-1)\cdot 50744521}{72+72-2}\cdot(\frac{1}{72}+\frac{1}{72}) = 1277823$$

```
((71*41258741+71*50744521)/(72+72-2))*((1/72)+(1/72))
```

```
## [1] 1277823
```

And use it to get the test statistic for the null hypothesis:

$$T_p = \frac{\bar{X_n} - \bar{Y_m}}{S_p} = \frac{17410 - 19818}{\sqrt{1277823}} = -2.13$$

```
(17410 - 19818)/sqrt(1277823)
```

```
## [1] -2.130204
```

The critical values are $\pm t_{142,0.025} = \pm 1.98$.

```
qt(0.025,142)
```

```
## [1] -1.976811
```

Since $t_p < -t_{142,0.025}$ we must reject the null hypothesis.

b. Do the same without the assumption of equal variances.

Solution: For samples with unequal variances we use the nonpooled variance (pp. 420, Dekking et al. (2010)):

$$S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m} = \frac{41258741}{72} + \frac{50744521}{72} = 1277823$$

```
41258741/72+50744521/72
```

```
## [1] 1277823
```

$$T_d = \frac{\bar{X_n} - \bar{Y_m}}{S_d} = \frac{17410 - 19818}{\sqrt{1277823}} = -2.13$$

As we assume unequal variance, the test statistics has approximately the $t(v)$ distribution under the null hypothesis, where

$$v = \frac{(\frac{S_X^2}{n} + \frac{S_Y^2}{m})^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

```
v <- (41258741/72+50744521/72)^2/(41258741^2/(72^2*71)+50744521^2/(72^2*71))
qt(0.025,v)
```

```
## [1] -1.976992
```

Since $t_d < -t_{v,0.025}$ we must reject the null hypothesis.

c. As a comparison, one carries out an empirical bootstrap simulation for the nonpooled studentized mean difference. The bootstrap approximations for the critical values are $c_l^* = -2.004$ and $c_u^* = 2.133$. Report your conclusions about the salaries on the basis of the bootstrap results.

Solution: The observed value $t_d = -2.130$ is smaller than the left critical value $c_l^* = -2.004$, so we must reject the null hypothesis. With 95% confidence level we can say that the salaries for these two occupations are not the same.

4. Significance test for the mean (R)

The United States Department of Energy conducts weekly phone surveys on the price of gasoline sold in the United States. Suppose one week the sample average was \$4.03, the sample standard deviation was \$0.42, and the sample size was 800. Perform a one-sided significance test of $H_0 : \mu = 4.00$ against the alternative $H_\text{A} : \mu > 4.00$.

Solution: Since we are considering a large sample, our test statistic can be assumed to approximately follow a standard normal distribution.

```
obs <- (4.03 - 4)/(0.42/sqrt(800))
pt(obs, 799, lower.tail = F)
```

```
## [1] 0.02184248
```

The p-value is 0.02 which is very small and discredits the claim of $\mu = 4.00$ per gallon.

5. Two-sample tests of centre (R)

The data set `normtemp (UsingR)` contains body measurements for 130 healthy, randomly selected individuals. The variable `temperature` contains normal body temperature data and the variable `gender` contains gender information, with male coded as 1 and female as 2. Can you assume that the groups have similar standard deviation? Is the sample difference across two groups statistically significant? Is the conclusion the same if you made a different assumption of the standard deviation?

Solution: We start by checking whether we can assume that the groups have similar variance.

```
women <- subset(normtemp$temperature, normtemp$gender == 2)
men <- subset(normtemp$temperature, normtemp$gender == 1)
var(women)/var(men)
```

```
## [1] 1.132131
```

Looks like women temperature variance is about 13% higher, therefore we do not assume equal variance. Let's use t.test to check if the temperatures differ:

```
res <- t.test(men, women)
res
```

```
##
##  Welch Two Sample t-test
##
## data:  men and women
## t = -2.2854, df = 127.51, p-value = 0.02394
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.53964856 -0.03881298
## sample estimates:
## mean of x mean of y
##  98.10462  98.39385
```

P-value is less than 0.025 so we must reject H_0.

Now let's check if the conclusion is the same if we assume that the variances are equal:

```
res <- t.test(men, women, var.equal = T)
res
```

```
##
##  Two Sample t-test
##
## data:  men and women
## t = -2.2854, df = 128, p-value = 0.02393
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.53963938 -0.03882216
## sample estimates:
## mean of x mean of y
##  98.10462  98.39385
```

P-value is still less than 0.025 so we must reject H_0.

6. Bootstrapping in two-sample tests (R, Home Exercise)

For the `babies (UsingR)` data set, the variable `age` contains the recorded mom's age and `dage` contains the dad´s age for several different cases in the sample. Do a significance test of the null hypothesis of equal age against a one-sided alternative that dads are older in the sample population. Use a non-normal model with bootstrapping.

Solution: First, we formulate the hypotheses and check whether we can assume equal variances:

$H_0 : \mu_{dad} = \mu_{mom}$

$H_A : \mu_{dad} > \mu_{mom}$

```
x <- babies$dage
y <- babies$age
var(x)/var(y)
```

```
## [1] 1.741825
```

Variances seem to differ significantly, therefore we will assume unequal variance.

```
s_d <- sqrt(var(x)/length(x) + var(y)/length(y))
t_d <- (mean(x) - mean(y)) / s_d; t_d
```

```
## [1] 11.0671
```

```
alpha <- 0.05; t_d_s_samples = c();
for (i in 1:10000) {
  x_s <- sample(x, length(x), replace=TRUE)
  y_s <- sample(y, length(y), replace=TRUE)
  s_d_s <- sqrt(var(x_s)/length(x) + var(y_s)/length(y))
  t_d_s <- ( (mean(x_s)-mean(y_s)) - (mean(x)-mean(y)) ) / s_d_s
  t_d_s_samples <- c(t_d_s_samples, t_d_s)
}
quantile(t_d_s_samples, probs=c(1-alpha))
```

```
##      95%
## 1.605792
```

```
mean(t_d > t_d_s_samples)
```

```
## [1] 1
```

The decision rule with significance level α is : reject $H_0 : \mu_X = \mu_Y$ in favour of $H_1 : \mu_X > \mu_Y$ whenever $T_d \geq c_u^*$. With $t = 11.067$ we must reject H_0 at $\alpha = 0.05$, and we find the p-value $P(T_d \geq t_d) = 0.00$.

Dekking, F. M., C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester. 2010. *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer-Verlag.

Verzani, John. 2014. *Using R for Introductory Statistics*. CRC Press.