Solutions Chapter 5 (Two samples problems) – Mathematical Statistics

Exercise 1

a. We want to compare two population proportions with two independent samples:
use the formula $\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$, where:
n_1 = 500, n_2 = 500, difference in proportions $\hat{p}_1 - \hat{p}_2 = \frac{140}{500} - \frac{100}{500} = 0.08$, standard error $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.0269$ and $c = 2.575$ is such that $\Phi(c) = 1 - \frac{1}{2}\alpha = 0.995$
99%-CI$(p_1 - p_2) = (0.08 - 2.575 \times 0.0269, 0.08 + 2.575 \times 0.0269) \approx (0.011, 0.149)$

b. If we would use this interval we can state that there is a significant proportional difference: at a confidence level of 99% the difference in mortality rate is between +1.1% and +14.9%, so the difference 0% $(= p_1 - p_2)$ is excluded by this interval.

c. For $\hat{p}_1 = \frac{140}{500} = 0.28$ we have: $\hat{p}_1 \pm c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 0.28 \pm 0.052$, so 99%-CI$(p_1) = (0.228, 0.332)$
And for $\hat{p}_2 = \frac{100}{500} = 0.2$: $\hat{p}_2 \pm c\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.20 \pm 0.046$, so 99%-CI$(p_2) = (0.154, 0.246)$
These intervals show some overlap, which implies that possibly $p_1 = p_2$.
Remark: it is best to base a statement on the difference $p_1 - p_2$ on an interval estimate of $p_1 - p_2$, and not on two separate intervals of p_1 and p_2,respectively.

Exercise 2

1. Let X_1 and X_2 be the numbers of rats, that died among the untreated and the treated rats, resp. X_1 and X_2 are independent and $B(500, p_1)$- resp. $B(500, p_2)$-distributed with mortality rates p_1 and p_2.
2. We test $H_0: p_1 = p_2$ versus $H_1: p_1 > p_2$, with α = between 1% and 10%.
3. Test statistic: $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ with $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$
4. Under H_0 Z has a $N(0,1)$-distribution.
5. Outcome of Z: $\hat{p} = \frac{140+100}{500+500} = 0.24$, so: $z = \frac{0.28 - 0.20}{\sqrt{0.24 \cdot 0.76\left(\frac{1}{500} + \frac{1}{500}\right)}} \approx 2.96$.
6. Reject H_0 if the p-value = $P(Z \geq 2.96) \leq \alpha$. The p-value = 1 - $\Phi(2.96)$ = 1 – 0.9985 = 0.15%
(*Or: reject H_0 if $Z \geq c$: Significance level α between 1% and 10%, so c = between 1.28 and 2.33*)
7. The p-value is less than every value of α between 1% and 10%, so reject H_0.
(Or: z =2.96 is not in the Rejection Region for all α between 1% and 10%, so reject H_0.)
8. We consider the statement that the mortality rate of rats decreases when using the medicine to be proven, at all levels of significance between 1% and 10%.

Exercise 3

This exercise deals with a comparison of two population proportions. Using the testing procedure in 8 steps:

1. We define X_1 and X_2 to be the numbers of dissertations within 6 years among 229 female and 795 male PhD`s, resp. X_1 and X_2 are independent and both binomially distributed with success probabilities p_1 and p_2.
2. Test $H_0: p_1 = p_2$ against $H_1: p_1 \neq p_2$ with α = 5%.
3. Test statistic: $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ with $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$
4. Under H_0 Z has a $N(0,1)$-distribution.

5. Observed value of Z: $\hat{p} = \frac{98 + 423}{229 + 795} = 0.509$, so: $z = \frac{\frac{98}{229} - \frac{423}{795}}{\sqrt{0.509 \cdot 0.491\left(\frac{1}{229} + \frac{1}{795}\right)}} \approx -2.77.$
6. Two-sided test: reject H_0 if $Z \leq -c$ or $Z \geq c$ Significance level 5%: $\Phi(c) = 0.975$ if $c = 1.96$
7. The observed value -2.77 lies in the rejection region, so reject H_0.
8. We showed, at a 5% level of significance, that the proportions of dissertations within 6 years for female and male PhD's are different.

Exercise 4

The condition for the confidence interval = the width = $2 \cdot c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq 0.02,$

where $c = 1.96$, $n_1 = n_2 = n$ and $p(1-p) \leq \frac{1}{4}$.

$1.96\sqrt{\frac{1/4}{n} + \frac{1/4}{n}} = \frac{1.96}{\sqrt{2n}} \leq 0.01$, so $\sqrt{2n} \geq 196$, implying: $n \geq 19208$

Exercise 5

a. Obviously we have two independent samples, drawn from two (separate) subpopulations of students, one subpopulation has to answer questions before and the other after the introduction. We will try to compare the mean scores μ_1 and μ_2 in those subpopulations by assessing the means in the samples.

b. This is a case of paired samples: there is only one population and one person is assessed twice: pairwise dependence of the observations. We are interested in the issues of a (systematic) difference μ of the two scores.

c. We have one random sample in this case.

d. The researcher evaluates two methods of measuring a fluid with a **fixed** concentration: the outcomes of all measurements are independent. If there is a systematic difference in the outcome of the two methods, that is, in μ_1 and μ_2, the samples will show a difference in sample means. We will have to conduct a test on the difference $\mu_1 - \mu_2$, based on two independent samples.

Exercise 6

a. Let $X_1, X_2, \ldots, X_9$ the crop quantities of variety A and $Y_1, Y_2, \ldots, Y_{11}$ the crop quantities of variety B.
Let us notate the sample means as $\overline{X}_1 = \frac{1}{9}\sum_{i=1}^{9} X_i$ and $\overline{X}_2 = \frac{1}{11}\sum_{j=1}^{11} Y_j$ and the sample variances as S_1^2 and S_2^2.
Observed values (simple calculator!): A: $\overline{x}_1 = 35.0$ and $s_1 = 2.598$ and B: $\overline{x}_2 = 39.0$ and $s_2 = 3.286$.
The standard deviations differ less than a factor 2, indicating, according to the rough rule of thumb, that the variances can be assumed equal.

b. To be more precise we will conduct the F-test to get the assumption of equal variance confirmed:
1. Probability model: the crop quantities $X_1, \ldots, X_9, Y_1, \ldots, Y_{11}$ are independent with $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_j \sim N(\mu_2, \sigma_2^2)$.
2. Test H_0: $\sigma_1^2 = \sigma_2^2$ (or $\sigma_1 = \sigma_2$) against H_1: $\sigma_1^2 \neq \sigma_2^2$ with $\alpha = 5\%$.
3. Test statistic $F = \frac{S_1^2}{S_2^2}$.
4. Distribution under H_0: $F \sim F_{11-1}^{9-1}$
5. Observed value: $F = \frac{s_1^2}{s_2^2} = \frac{2.598^2}{3.286^2} \approx 0.625$

6. We have a two-sided test: reject H_0 if $F \le c_1$ or $F \ge c_2$.
   $P(F_{10}^8 \ge c_2) = \frac{\alpha}{2} = 0.025$, so according to the F_{10}^8: $c_2 = 3.72$
   $P(F_{10}^8 \le c_1) = P\left(F_8^{10} \ge \frac{1}{c_1}\right) = \frac{\alpha}{2} = 0.025$, so $\frac{1}{c_1} = 4.30$, or $c_1 \approx 0.23$
7. Since $F = 0.625$ does not lie in the Rejection Region, we cannot reject H_0.
8. At a significance level of 5% we cannot prove that the variances of the crop quantities are different.

c. 1. Model assumptions ("statistical assumptions"):
We have two independent random samples of crop quantities here, one drawn from a $N(\mu_1, \sigma^2)$-distribution for variety A and the other from a $N(\mu_2, \sigma^2)$-distribution for variety B (equal σ's!)
Stated more formally: the crop quantities $X_1, \dots, X_9, Y_1, \dots, Y_{11}$ are independent, where $X_i \sim N(\mu_1, \sigma^2)$ and $Y_j \sim N(\mu_2, \sigma^2)$.

2. We will test H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \ne \mu_2$ with α = 5%

3. Test statistic $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{9}+\frac{1}{11}\right)}}$ with $S^2 = \frac{8S_1^2 + 10S_2^2}{9+11-2}$

4. T is under H_0 t-distributed with $df = n_1 + n_2 - 2 = 18$

5. Observed: $s^2 = \frac{8\times 2.598^2 + 10\times 3.286^2}{18} \approx 9.00$, so $t = \frac{35.0 - 39.0}{\sqrt{9.00\left(\frac{1}{9}+\frac{1}{11}\right)}} = -2.97$

6. This test is two-tailed: **reject H_0 if $T \le -c$ or $T \ge c$.**
   where $c = 2.101$, taken from the t_{18}-table.

7. t = -2.97 lies in the Rejection Region, so reject H_0.

8. The mean crop quantities of the two varieties are significantly different at a 5% level.

6./7. Using the p-value at the observed $t = -2.97$: $p - \text{value} = 2 \cdot P(T \ge |t|) = 2 \cdot P(T_{18} \ge 2.97)$
$P(T_{18} \ge 2.97)$ lies between 0.1% and 0.5%, so the p-value is between 0.2% and 1% < α: reject H_0.

d. We can use the formula of the interval bounds: $\overline{X}_1 - \overline{X}_2 \pm c\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$, in which
$n_1 = 9, n_2 = 11, \overline{x}_1 = 35.0\ \overline{x}_2 = 39.0$ (a.) , $s^2 = 9.00$ and $c = 2.101$ (b), so that:
95%-CI($\mu_1 - \mu_2$) = (-4.0 – 2.8, -4.0 + 2.8) = (-6.8, -1.2)

e. The difference 0 of $\mu_1 - \mu_2$ is not contained in the confidence interval in d.: it is completely negative. So "at a confidence level of 95%" one can state that the difference in expected crop quantities differ, confirming the conclusion in c.

f. Testing H_0: $\mu_1 - \mu_2 = \Delta_0$ against H_0: $\mu_1 - \mu_2 \ne \Delta_0$ implies that the test statistic is $T = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sqrt{S^2\left(\frac{1}{9}+\frac{1}{11}\right)}}$

The rejection region, determined in c., remains the same: $\boldsymbol{T \le -c}$ **or** $\boldsymbol{T \ge c}$**,** where $c = 2.101$.
Not rejecting H_0 implies $-c < T < c$: under H_0 we have $P(-c < T < c) = 1 - 0.05 = 0.95$
Hence, we have $-c < \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sqrt{S^2\left(\frac{1}{9}+\frac{1}{11}\right)}} < c$

Solving Δ_0 from the inequality results in: $(\overline{X}_1 - \overline{X}_2) - c\sqrt{S^2\left(\frac{1}{9}+\frac{1}{11}\right)} < \Delta_0 < (\overline{X}_1 - \overline{X}_2) - c\sqrt{S^2\left(\frac{1}{9}+\frac{1}{11}\right)}$,
the same formula and value of c. that we used for determining the 95%-$CI(\mu_1 - \mu_2)$ in part d.
<u>Note</u>: a similar relation can also be derived for the test and CI of p_1- p_2, but the test statistic to be used cannot use the equality of the proportions if H_0: $p_1 - p_2 = \Delta_0$: in that case one should choose for the test statistic $Z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$.

Exercise 7

a. We have 2 observations for each store: we need to apply a paired samples approach: we will consider the differences $Z_i = Y_i - X_i$ only and apply the 8 steps procedure:

1. The differences (sales increase: sales numbers **before-after** the campaign) $Z_1, Z_2, \dots, Z_7$are independent and all $N(\mu, \sigma^2)$-distributed, where the expected increase in sales μ is unknown and the variance σ^2 of the increases is unknown as well.
   (the notation of the mean will be $\overline{z} = 516.0$ and the accompanying standard deviation is $s_Z = 622.7$)
2. We will test $H_0: \mu = 0$ against $H_1: \mu > 0$ with α_0 = 0.05.
3. Test statistic: $T = \frac{\overline{Z}}{S_Z/\sqrt{n}} = \frac{\overline{Z}}{S_Z/\sqrt{7}}$
4. Under H_0: $T \sim t_6$
5. Outcome of T: $t = \frac{516.0}{622.7/\sqrt{7}} \approx 2.19$
6. We will reject H_0 if $T \geq c$. Since $\alpha_0 = 0.05$, we will find in the t_6-table: $c = 1.943$.
7. Outcome $t = 2.19$ lies in the rejection region $\Longrightarrow$ reject H_0.
8. At a 5% significance level we have proven that the sales numbers after the ad campaign increased.

b. Changes in the procedure are indicated below:

1.
2. We will test $H_0: \mu = 0$ against $\boldsymbol{H_1: \mu \neq 0}$ with α_0 = 0.05.
3.
4.
5.
6. We will reject H_0 if $\boldsymbol{T \leq -c}$ **or** $T \geq c$.
   Since $\alpha_0 = 0.05$, the critical value $\boldsymbol{c = 2.447}$ from the t_6-table is such that $P(T_6 \geq c) = \frac{\alpha_0}{2} = 0.025$.
7. Outcome $t = 2.19$ does **not lie in the rejection region $\Longrightarrow$ we fail to reject H_0.**
8. At a 5% significance level we **failed to prove** that the sales numbers before and after the ad campaign differ.

Exercise 8

a. Since the two groups of cockerels are treated differently we can assume that the samples are independent. If both samples are random and drawn from normal distributions with the same variance, we can apply the 2 samples t −procedure with equal variances.

b. We will apply the following formula (on the formula sheet!), interchanging $\bar{X}$ and $\bar{Y}$ (such that $\bar{Y}$-$\bar{X}$ is positive):

$\left(\bar{Y} - \bar{X} - c\sqrt{S^2\left(\frac{1}{20} + \frac{1}{20}\right)}, \bar{Y} - \bar{X} + c\sqrt{S^2\left(\frac{1}{20} + \frac{1}{20}\right)}\right)$, where

$c = 1.68$ from the *t*-table with *df* = 20 + 20 – 2 = 38 (using the t_{40} − table)

$s^2 = \frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1 + n_2 - 2} = \frac{19 \times 50.80^2 + 19 \times 42.73^2}{38} = 2203.25$ $(= 46.94^2,\ so\ s\ lies\ between\ s_1 and\ s_2)$

Result after substitution: 90%-CI$(\mu_1 - \mu_2) = (38.45 - 24.94, 38.45 + 24.94) = (13.51, 63.39)$

c. The assumptions are, in detail:

1. 2 independent random samples: $X_1, \dots, X_{20}\ and\ Y_1, \dots, Y_{20}$ are independent.
2. $X_1, \dots, X_{20}$ is a random sample drawn from the $N(\mu_1, \sigma_1^2)$ − distribution
3. $Y_1, \dots, Y_{20}$ is a random sample drawn from the $N(\mu_2, \sigma_2^2)$ − distribution
4. The variances are equal: $\sigma_1^2 = \sigma_2^2$

For short: $X_1, \dots, X_{20}, Y_1, \dots, Y_{20}$ are independent and $X_i \sim N(\mu_1, \boldsymbol{\sigma^2})$ and $Y_j \sim N(\mu_2, \boldsymbol{\sigma^2})$.
We will apply the testing procedure for the F-test to check the "equal variances"-assumption:

1. Assumptions: the increase of the weights $X_1, \dots, X_{20}, Y_1, \dots, Y_{20}$ are independent and $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_j \sim N(\mu_2, \sigma_2^2)$
2. Test $H_0: \sigma_1^2 = \sigma_2^2$ (*or* $\sigma_1 = \sigma_2$) against $H_1: \sigma_1^2 \neq \sigma_2^2$ with $\alpha = 10\%$
3. Test statistic $F = \frac{S_X^2}{S_Y^2}$
4. Distribution under H_0: $F \sim F_{20-1}^{20-1}$
5. Observed value: $F = \frac{s_X^2}{s_Y^2} = \frac{50.8^2}{42.73^2} \approx 1.41$
6. It is a two-sided test: reject H_0 if $F \leq c_1$ or $F \geq c_2$.
   $P(F_{19}^{19} \geq c_2) = \frac{\alpha}{2} = 0.05$, so (according to the F_{19}^{20}-table) $c_2 = 2.16$
   (or using interpolation of the table values in F_{19}^{20} and F_{19}^{15}: $c_2 = 2.17$)
   $P(F_{19}^{19} \leq c_1) = P\left(F_{19}^{19} \geq \frac{1}{c_1}\right) = \frac{\alpha}{2} = 0.05$, so $\frac{1}{c_1} = 2.16$, or $c_1 \approx 0.46$
7. Since $F = 1.41$ is not in the Rejection Region, we will not reject H_0.
8. The variances of the increase of the weights are not statistically significantly different at a 10% significance level.

Exercise 9

a. We want to show that $\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2}$ has a $F_{n_2-2}^{n_1-1}$-distribution. We know the following, using the given model:

1. $\frac{(n_1-1)S_X^2}{\sigma_1^2} \sim \chi_{n_1-1}^2$ and $\frac{(n_2-1)S_Y^2}{\sigma_2^2} \sim \chi_{n_2-1}^2$. The samples are independent and so are these variables.
2. Definition of an F-distribution: if $V \sim \chi_k^2$ and $W \sim \chi_l^2$ are independent, then $F = \frac{V/k}{W/l} \sim F_l^k$

Combining 1. and 2.: $F = \frac{U/f}{V/g} = \frac{\frac{(n_1-1)S_X^2}{\sigma_1^2} \Big/ (n_1-1)}{\frac{(n_2-1)S_Y^2}{\sigma_2^2} \Big/ (n_2-1)} = \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2}$ has a $F_{n_2-2}^{n_1-1}$-distribution.

b. In the F-table we can find values c_1 and c_2, such that $P\left(F_{n_2-2}^{n_1-1} \leq c_1\right) = P\left(F_{n_2-2}^{n_1-1} \geq c_1\right) = \frac{1}{2}\alpha$

So $P\left(c_1 < \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} < c_2\right) = 1 - \alpha$

$$P\left(c_1 \cdot \frac{S_Y^2}{S_X^2} < \frac{\sigma_2^2}{\sigma_1^2} < c_2 \cdot \frac{S_Y^2}{S_X^2}\right) = 1 - \alpha$$

$$P\left(\frac{1}{c_2} \cdot \frac{S_X^2}{S_Y^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{c_1} \cdot \frac{S_X^2}{S_Y^2}\right) = 1 - \alpha$$

Hence $(1-\alpha)100\% - BI\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left(\frac{1}{c_2} \cdot \frac{S_X^2}{S_Y^2}, \frac{1}{c_1} \cdot \frac{S_X^2}{S_Y^2}\right)$ is the stochastic confidence interval of $\frac{\sigma_1^2}{\sigma_2^2}$,
in which c_1 and c_2 are determined as follows: $P\left(F_{n_2-2}^{n_1-1} \geq c_2\right) = \frac{\alpha}{2}$ and

$$P\left(F_{n_2-2}^{n_1-1} \leq c_1\right) = P\left(\boldsymbol{F_{n_1-2}^{n_2-1}} \geq \frac{\mathbf{1}}{\boldsymbol{c_1}}\right) = \frac{\alpha}{2}.$$

For $n_1 = 6$, $n_2 = 10$ and level of confidence 95%:
From $P(F_9^5 \geq c_2) = \frac{\alpha}{2} = 0.025$ it follows that $c_2 = 4.48$ and
from $P\left(F_5^9 \geq \frac{1}{c_1}\right) = \frac{\alpha}{2} = 0.025$: $\frac{1}{c_1} = 6.68$ $(c_1 = \frac{1}{6.68} \approx 0.150$.