Part 1 – Question sheet

Problem 1 (10pts)

The reproducibility of a method for the determination of a pollutant in water was investigated by taking twelve samples from a single batch of water and determining the concentration of pollutant in each. The following results were obtained: 5.98, 8.80, 6.89, 8.49, 8.48, 7.47, 7.97, 6.94, 7.32, 6.64, 6.98, 7.94. It is expected that from this sample a 95% confidence interval for the concentration of pollutant will be obtained.

(i) There is a concern that the data may contain an outlier. Thus the following procedure has been performed on the data:

```
          Dixon test for outliers
data:  x
Q = 0.3625, p-value = 0.6326
alternative hypothesis: lowest value 5.98 is an outlier
```

(2pts) Describe what is the purpose of this procedure.

(1pts) Write the conclusion that follows from it.

(ii) After the test for an outlier, another preliminary procedure has been performed on the data

```
         One-sample Kolmogorov-Smirnov test
data:  x
D = 0.1414, p-value = 0.9432
alternative hypothesis: two-sided
```

(2pts) Explain what is the name and purpose of this procedure.

(1pts) What is the conclusion and why?

(iii) After these initial verifications, the confidence interval can be obtained from the following computations

```
mean(x)        sd(x)         qt(0.975,11)
7.491667       0.8572454     2.200985
```

(4pts) Based on the obtained values write down the confidence interval for the pollutant. What confidence level is used here? Name the distribution that is used for this purpose.

Problem 2 (10pts)

The following table gives the concentration of norepinephrine (μmol per gram creatinine) in the urine of healthy volunteers in their early twenties. Male: 0.48 0.36 0.28 0.55 0.45 0.46 0.47 0.25; Female 0.35 0.37 0.27 0.29 0.28 0.31 0.33. The problem is to determine if there is evidence that concentration of norepinephrine differs between genders. The following analyses have been performed on the data:

```
        F test to compare two variances
data:  M and F
F =7.90, num df =7, denom df =6, p-value =0.022
alter. hypothesis:
ratio of variances not equal to 1
95 percent confidence interval:
  1.386947 40.433419
ratio of variances
          7.899317
```

```
        Welch Two Sample t-test
data:  M and F
t = 2.4732, df = 8.953, p-value = 0.03551
alter. hypothesis:
difference in means not equal to 0
95 percent confidence interval:
 0.008309402 0.188119170
mean of x mean of y
0.4125000 0.3142857
```

- (4pts) Explain what is the purpose of each of the two procedure.
- (4pts) Write the conclusions that follow from the given analyses.
- (2pts) How, if at all, does the first procedure affect the second one?

Part 1 – Answer sheet

Problem 1 (10pts)

(2pts) Describe what is the purpose of this procedure

This is the Dixon test that is supposed to detect significantly too large or too little values (outliers) in data that are supposed to follow the normal distribution.

(1pts) Write a conclusion that follows from it.

Since the p-value is reported at 63% level there are no significant outliers in the data.

(2pts) Explain what is the name and purpose of this procedure.

This the Kolmogorov-Smirnov goodness-of-fit test that is examining if data follow prespecified distribution. In this case, data are tested against normal distribution.

(1pts) What is the conclusion and why?

The reported p-value of 94% clearly indicates that there is no evidence of deviation of the data from the normal distribution.

(4pts) Confidence interval and confidence level: $[6.95, 8.04]$ at 95% confidence level; Distribution: Student t-distribution with 11 degrees of freedom

Problem 2 (10pts)

(4pts) The purpose of the first procedure

This procedure which is called the ratio test for equality of variances in two sample problem is examining if the variability as measured by variances is the same within each of two samples.

The purpose of the second procedure

The second procedure which is called the Welch test for equality of the means in two samples problem. It is a version of the t-test for testing equality of two means for the case when variances within each of two samples are different.

(4pts) The conclusion from the first procedure

The p-value is below 5% so there is a significant evidence that the variance among males is different from the one within females. In fact, the confidence intervale for the ration of variances obtained from the procedure is about $[1.4, 40.4]$.

The conclusion form the second procedure

The p-value is also below 5% so there is a significant evidence that the average concentration of norepinephirine among males is different from the one within females. The confidence intervale for the difference of means obtained from the procedure is about $[0.0083, 0.1881]$.

(2pts) Consequences of the first procedure for the second one:

The purpose of the first procedure was to test if the variances are the same within male and female populations. The assumption of equal variances leads to more accurate and exact test for a difference in two sample problem. Since the procedure indicated a significant difference in variances a less accurate test (the Welch test) had to be used to test equality of the means.

Part 2 – Question sheet

Problem 1 (12pts)

Six analysts each made seven determinations of the paracetamol content of the same batch of tablets. The results are shown below

Analyst	Paracetamol content						
A	84.32	84.51	84.63	84.61	84.64	84.51	84.62
B	84.24	84.25	84.41	84.13	84.00	84.30	84.02
C	84.29	84.40	84.68	84.28	84.40	84.36	84.63
D	84.14	84.22	84.02	84.48	84.27	84.33	84.22
E	84.50	83.88	84.49	83.91	84.11	84.06	83.99
F	84.70	84.17	84.11	84.36	84.61	83.81	84.15

The following table has been produced as a result of analysis of these data:

```
Analysis of Variance Table

Response: xx
          Df  Sum Sq Mean Sq F value   Pr(>F)
An         5 0.86108 0.17222  4.2362 0.003941 **
Residuals 36 1.46351 0.04065
---
```

Based on this analysis answer the following questions.

(1pts) Describe what is the purpose of this procedure.

(2pts) Find the total sum of squares SS_T and determine the number of degrees-of-freedom associated with it.

(2pts) Determine the within-sample and between-sample estimators of variance.

(2pts) How are these estimators used to decide if there is a significant difference between analysts?

(2pts) What is the name of the distribution used in the above procedure? Specify its parameters. In what other important statistical procedure is this distribution also used?

(1pts) What is the conclusion following from the above analysis?

(2pts) The following row means (analysts' means) have been computed 84.54857 84.19286 84.43429 84.24000 84.13429 84.27286. Based on these values identify which analysts differ from others in their determinations using the least significant difference method (the 97.5% quantile of t-distriubution with 36 degrees of freedom is equal to `qt(0.975,36)=2.028094`).

Problem 2 (8pts)

It has been observed that measurements of concentration of a certain chemical have a standard deviation of 0.045. It is believed that the true concentration is either 2.0 or 2.06. Propose a statistical procedure based on a sample of measurements which would allow to decide between these two values and such that the chances of making any kind of error in the final claim are at most 5%. How big a sample size is needed for your procedure?

Part 2 – Question sheet

Problem 1

(1pts) Describe what is the purpose of this procedure

The procedure is called one-way analysis of variance (ANOVA) and is intended to detect any difference among analysts in the mean values of paracetamol determinations.

(2pts) The total sum of squares SS_T is $\boxed{0.86108+1.46351=2.32459}$ and its number of degrees of freedom is $\boxed{36+5=41}$.

(2pts) The within-sample estimator of variance is $\boxed{0.041}$.

The between-sample estimator of variance is $\boxed{0.172}$.

(2pts) How are these estimators used to decide if there is a significant difference between analysts?

The between-sample estimator of variance is divided by the within-sample estimator of it. If this ratio under the equality of means assumption (H_0 hypothesis) is distributed according to F-distribution with 5 and 36 degrees-of-freedom. Thus if this ratio exceeds the upper 95% quantile there is a significant evidence that the means for analysts are not equal.

(2pts) The name of distribution: $\boxed{F\text{-distribuion}}$; Parameters: $\boxed{\text{5 and 36 degrees of freedom}}$.

Used also in: $\boxed{\text{Testing for the equality of variances in two samples problems}}$.

(1pts) What is the conclusion following from the above analysis?

Since the p-value is only about 0.3% it is highly evident that the determination means differ among analysts.

(2pts) The least significant difference $\boxed{\sqrt{0.04065 * 2/7} * 2.028094 = 0.2185669}$

Analysts that significantly differ from A: $\boxed{\text{B,D,E,F}}$ from B: $\boxed{\text{A,C}}$ from C: $\boxed{\text{B,E}}$ from D: $\boxed{\text{A}}$ from E: $\boxed{\text{A,C}}$ from F: $\boxed{\text{A}}$.

Problem 2

(2pts) Name of the procedure: $\boxed{\text{One sample test for the mean on 5\% significance level}}$.

(4pts) Detailed description:

First, $T = \sqrt{n}(\bar{X} - 2.0)/0.045$ has to be evaluated and the claim that the concentration is 2.0 has to be rejected when $T > 1.96$. This procedure guarantees that rejecting the claim that the concentration is 2.0 when it is actually true is has chances equal to 5%. The chances of not rejecting this claim when the concentration is 2.06 (Type II Error) are given by $P(T < 1.645|\mu = 2.06) = \Phi(1.645 - 0.06 * sqrt(n)/0.045)$ which sets the equation

$$1.645 - 0.06 * \sqrt(n)/0.045 \leq -1.645.$$

(2pts) Sample size: $\boxed{n \geq (2 * 1.645 * 0.045)^2/(0.06)^2 \approx 6.09, \text{ i.e. } n = 7}$.