2DMT00 – Applied Statistics
Exercise Bundle

Richard Post
r.a.j.post@tue.nl

Kay Bogerd
k.m.bogerd@tue.nl

13-12-2019

Contents

Week 1: Confidence and prediction intervals

Questions

For the questions of this week we will use the IVF dataset (available on Canvas). Download this dataset and place it in the "myfolders" directory you created during the installation of your SAS University Edition. It should now be available under "My Folders" in SAS. Note: you might need to click the refresh button.

The IVF dataset contains three measurements of the infant motor profile (`IMP`) at different periods (`PER`). This is inconvenient when analyzing other variables, because every measurement is duplicated three times. Start by creating a new dataset without the `IMP`, `PER`, and `AGE` variables, that contains only a single observation of each child. You need to use this smaller dataset to answer the questions below.

Question 1.1

What is the total number of children considered in this study? How many of these children belong to a mother in the control group (`TRT = 0`), and how many received treatment-M (`TRT = 1`) or treatment-C (`TRT = 2`)?

Question 1.2

In this exercise you will analyze the age of mothers (`AGEM`) in the IVF dataset. For this question you may assume that the mother's age is normally distributed.

(a) Compute the mean and variance of `AGEM`. Use these to compute a 95% confidence interval for the average age of a mother, and a 95% prediction interval for a single new observation of `AGEM`.

(b) How many mothers were 40 years old or older when they became pregnant?

(c) Compute a 95% confidence interval for the variance of `AGEM`.

Question 1.3

In this exercise you will analyze the birth weight (`BW`) in the IVF dataset.

(a) Compute the first and third quartile of the birth weight together with a corresponding 95% confidence interval. What is the interquartile range of the birth weight?

(b) What is the percentage of children whose birth weight differs at most one interquartile range from the median?

(c) The birth weight data is not normally distributed. Draw a histogram of the Box-Cox transformed birth weight, use $\lambda \in \{-2, -1/2, 0, 1/2, 2\}$. What is the best value for λ such that the data becomes approximately normal?

(d) Using your answer from (c), compute a 95% prediction interval for a single new observation.

We also want to test whether boys (`SEX = 1`) or girls (`SEX = 0`) are typically heavier during birth based on data in the IVF dataset.

(e) What is the birth weight of the heaviest baby in this dataset, is this a boy or a girl?

(f) What percentage of children in this dataset are girls and boys? For each gender, compute the mean, variance, skewness, and kurtosis of the birth weight.

(g) Based on your answers in (f), do you think a prediction interval for one new observation will be wider if we know that this observation is from a boy or a girl?
Verify your answer by computing these 95% prediction intervals using the Box-Cox transformation you found in (c).

Question 1.4

In the table below are the measured values of 12 observations, you may assume that these are normally distributed. Create a new dataset containing the table below, and use this dataset to answer the following questions.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
Value	25.0	27.4	17.1	22.1	20.8	21.3	22.5	29.2	27.9	25.7	24.7	18.8

(a) Compute the mean, variance, skewness and kurtosis of the measured values.

(b) Create a 95% confidence interval for the mean, standard deviation and variance.

(c) Create a 99% prediction interval for a single new measurement.

Question 1.5

As the result of a study, a data-scientist finds a 95% confidence interval for $\log(\mu)$ equal to $(-0.137, 1.128)$. We want to use this confidence interval to draw conclusions about μ.

(a) Use the confidence interval of $\log(\mu)$ to derive a 95% confidence interval for μ.

(b) Test H_0: $\mu = 3$ vs. H_1: $\mu \neq 3$ at a significance level of $\alpha = 5\%$.

Question 1.6

In this exercise you will analyze the gestational age (`GA`) from the IVF dataset. The gestational age (`GA`) measures the duration in pregnancy in weeks. This is not normally distributed, but someone suggests that `LOG(44 - GA)` does have an approximate normal distribution.

(a) Use the transformation `LOG(44 - GA)` to compute a 95% prediction interval for a single new observation of the gestational age.

(b) Create a 95% confidence interval for the mean of the transformed variable `LOG(44 - GA)`. Can you use this confidence interval to create a 95% confidence interval for the mean of `GA`? Explain your reasoning.

(c) We want to test whether the median of `GA` is significantly different from 39. Create a 95% confidence interval for the median gestational age and compare this with the mean, is there a significant difference?

We also want to test whether a short gestational age can lead to more complications. Specifically, we will test whether preterm children (`GA <= 38`) have a greater chance to experience stress during delivery.

(d) Create an extra column to indicate whether a child was born preterm or not (i.e. add a new column that is 1 when `GA <= 38`, and 0 otherwise).

(e) Calculate the percentage of children that were in stress during delivery (`FIS`) for both the preterm and non-preterm children.

(f) Create an approximate 90%-confidence intervals for the percentage of children that experienced stress during delivery `FIS`, for both preterm children and non-preterm children.

Use these confidence intervals to test whether the percentage of children that experienced stress during delivery is significantly different from 32% at the $\alpha = 10\%$ level, for both preterm and non-preterm children.

(g) Because our sample size is not very large, someone suggests you to use an exact confidence interval in (f) instead. Does this affect your conclusion?

Question 1.7

Make use of the following macro to (randomly) sample 1000 datasets with each 10 measurements, from the empirical distribution of the IVF dataset. The simulated datasets are merged in a single dataset called `FINAL`, with the sample number indicated by the `SAMPLE` variable.

```
%macro samples(dataset=,ns=,n=);
proc surveyselect data=&dataset NOPRINT
    method=urs n=&n out=FINAL;
run;

data FINAL;
    set FINAL;
    sampleno=1;
run;


%do sn = 2 %to &ns;
    proc surveyselect data=&dataset NOPRINT
    method=urs n=&n out=SAMPLEI;
run;

data SAMPLEI;
    set SAMPLEI;
    sampleno= &sn;
run;

data FINAL;
    set Final SAMPLEI;
run;
%end;

proc datasets library=work NOPRINT;
    delete SAMPLEI;
run;
%mend;
```

To run this macro, you can use the following code

```
%samples(dataset=‘YOUR DATASET’, ns=1000, n=10);
```

Answer the following question using the dataset you obtain by running the `samples` macro.

(a) Create a dataset containing the sample mean and sample variance of `AGEM` per simulated study.

(b) The mean and variance of AGEM are now the distribution characteristics in our simulated study (do you see this?). Use the dataset you created in (a) to compute the standardized average and the sample variance rescaled by the real variance for each simulated study.

(c) Create a histogram for both variables you created in (b), do you recognize a distribution?

Question 1.8

In this exercise we will again analyze the birth weight (BW) in the IVF dataset.

(a) Perform a two-sided t-test to test H_0: $\mu = 3200$. Report the value of the test statistic, the p-value and your conclusion.

(b) Use your result of question (a) to derive the p-value of the one-sided t-test to test H_0: $\mu \leq 3200$. Verify your answer using `proc ttest`.

(c) In question 3 we saw that the birth weight was not normally distributed, are the results of (a) and (b) still reliable?

(d) Originally the IVF dataset contained 3 identical/duplicated observations per child. What would have been the impact of not excluding these duplicated observations on your analysis in (a) and (b)?

(e) While replacing the neonatal units in a hospital the procurement officer suspects that the average birth weight of the babies exceeds 3000g (in which case he should buy sturdier neonatal units). However, in practice he is not interested in testing H_0: $\mu \leq 3000$, but rather H_0: $|\mu - 3000| < \Delta$, where $\Delta = 200$ is a practically relevant margin. Derive the p-value to test H_0: $|\mu - 3000| < \Delta$. Hint: it is not necessary to compute new results.

Use the `samples` macro introduced in question 1.7, to simulate $ns = 1000$ datasets each containing $n = 200$ (re-)samples of the birth weight (BW).

(f) Run the one-sided t-test, to test H_0: $\mu \leq 3200$, on each simulated study and store the p-values in a new dataset. Hint: use the code below to output the results in a new dataset and not print all the individual results as output.

```
ods graphics off;
ods exclude all;
proc ttest data=FINAL h0=3200 sides=u alpha=0.05;
    var BW;
    by SAMPLENO;
    ods output ttests=POWER(keep=PROBT);
run;
ods exclude none;
ods graphics on;
```

(g) Estimate the power of the t-test you conducted in (b).

(h) How do your answers to (g) change when using only $n = 50$ or $n = 100$ (re-)samples in each simulated dataset.

Answers

Answer to Question 1.1

The IVF dataset contains $N_0 = 93$ observations in the control group (`TRT = 0`), $N_1 = 58$ observations in the treatment-M group (`TRT = 1`), and $N_2 = 102$ observations in the treatment-C group (`TRT = 2`). In total this are $N = N_0 + N_1 + N_2 = 253$ observations.

Answer to Question 1.2

(a) The mean of variance of `AGEM` are $\mu_{\texttt{AGEM}} = 32.6684$ and $\sigma^2_{\texttt{AGEM}} = 12.1374$. A 95% confidence interval for the mean is given by $(32.2370, 33.0998)$, and a 95% prediction interval for a new observation is given by $(25.7936, 39.5432)$.

(b) 3 mothers were 40 years old or older when the became pregnant.

(c) A 95% confidence interval for the variance is given by $(10.2686, 14.5703)$.

Answer to Question 1.3

(a) The first quartile is 2910 with 95% confidence interval $(2730, 3030)$ using `cipctlnormal` or $(2720, 3030)$ using the procedure discussed in lecture. The third quartile is 3780 with 95% confidence interval $(3670, 3850)$ using `cipctlnormal` and $(3650, 3850)$ using the procedure discussed in lecture. The interquartile range is $IQR = 3780 - 2910 = 870$.

(b) About $199/253 \approx 79\%$ differ at most one interquartile range from the median.

(c) A value of $\lambda = 2$ seems the most appropriate.

(d) A 95% prediction interval for the birth weight is $(1697.32, 4439.83)$.

(e) The heaviest baby had a birth weight of 4710g, this was a boy.

(f) About $119/253 \approx 47\%$ are girls, so 53% are boys. For the birth weight of girls we have: mean 3279.36, variance 348096.03, skewness -0.4139, and kurtosis 0.0090. For the birth weight of boys we have: mean 3305.50, variance 549339.15, skewness -0.5048, and kurtosis -0.1516.

(g) The variance of the birth weight is much larger for boys than for girls, so we will also get a wider prediction interval for boys.

For girls: the 95% prediction interval is $(1904.95, 4309.28)$, and for boys this is $(1478.59, 4555.98)$. The prediction interval for boys is about 1.3 times larger. Don't forget the use the transformation from (c) when creating the prediction intervals.

Answer to Question 1.4

(a) The sample mean equals 23.54, the sample variance equals 13.97, the sample skewness equals -0.15 and the sample excess kurtosis equals -0.85.

(b) A 95% confidence interval for the mean equals $(21.17, 25.92)$, for the variance $(7.01, 40.28)$ and for the standard deviation $(2.65, 6.35)$.

(c) A 99% prediction interval for a new observation equals $(11.46, 36.12)$.

Answer to Question 1.5

(a) A 95% confidence interval for μ is given by $(0.8720, 3.0895)$.

(b) Because the confidence interval contains 3 we cannot reject H_0 in this case.

Answer to Question 1.6

(a) A 95% prediction interval for a new observation of `GA` is given by $(33.1628, 42.1604)$.

(b) A 95% confidence interval for the mean of `LOG(44 - GA)` is given by $(1.4406, 1.5519)$.

No, if we only have the 95% confidence interval for the mean of `LOG(44 - GA)`, then we cannot compute a 95% confidence interval for the mean of `GA`.

(Since we assumed that `LOG(44 - GA)` has a normal distribution, one could use the method described here to create a confidence interval for `GA`.)

(c) A 95% confidence interval for the median of `GA` is given by $(39.4286, 39.8571)$. Because 39 is not contained in the confidence interval we conclude that the median of `GA` is a significantly different from 39.

(e) For preterm children the percentage is $15/65 \approx 23\%$, and for non-preterm children this is $70/187 \approx 37\%$.

(f) A approximate 90% confidence interval for the percentage of children in stress during delivery is: for preterm children $(0.1448, 0.3167)$, and for non-preterm children $(0.3161, 0.4325)$.

We see that the percentage of children that experienced stress during delivery is significantly different from 32% for the preterm children, but not for non-preterm children.

(g) An exact 90% confidence interval for the percentage of children in stress during delivery is: for preterm children $(0.1480, 0.3329)$, and for non-preterm children $(0.3153, 0.4364)$.

Using an exact confidence interval *does* change the conclusion in (f). Now we find that there is no significant difference between the children that experienced stress during delivery and 32% in both the preterm and non-preterm group.

Answer to Question 1.7

(a) -

(b) -

(c) You should recognize a χ^2_9 distribution for the rescaled sample variance and a t_9 distribution for the standardized average.

Answer to Question 1.8

(a) The test statistic $\frac{\bar{y}-3200}{\sqrt{\frac{s^2}{n}}} = 2.20$, the p-value equals $2 \cdot \mathbb{P}\left(X > \frac{\bar{y}-3200}{\sqrt{\frac{s^2}{n}}}\right) = 0.0285$. Thus, we conclude that the average birth weight is significantly ($\alpha = 0.05$) different from 3200.

(b) Now, the p-value equals $\mathbb{P}\left(X > \frac{\bar{y}-3200}{\sqrt{\frac{s^2}{n}}}\right) = \frac{0.0285}{2}$.

(c) You should ask yourself whether the central limit theorem can be applied. Is n large enough to assume that the sample average $\bar{y}$ follows a normal distribution? The macro presented in question 1.7 can be used to verify this. Load the macro first, and run the code below.

```
%samples(dataset='YOUR DATASET', ns=100, n=253);

proc means data=FINAL mean NOPRINT;
    var BW;
    by sampleno;
    output out=MEANSBW mean=BW_MEAN;
run;

proc univariate data=MEANSBW;
    hist BW_MEAN /normal;
run;
%mend;
```

Finally, you can conclude that it seems valid to apply the result of the CLT.

(d) Using a the t-test we judge whether an observed effect size, $\bar{y} - 3200$, is significantly large taking in to account the variability in the shape of the standard error, $\sqrt{\frac{s^2}{n}}$. Using a sample with 3 identical copies of every observed value will not change the effect size, while the standard error decreases with a factor $\frac{1}{\sqrt{3}}$. As a result, the p-value will be smaller and H_0 (if false) will be rejected faster.

(e) The null hypothesis is $H_0 : |\mu - 3000| < 200$, which is equivalent to $H_0 : 3000 - 200 < \mu < 3000 + 200$. The p-value equals $2 \cdot \min \left\{ \mathbb{P}\left(X < \frac{\bar{y}-2800}{\sqrt{\frac{s^2}{n}}} \right), \mathbb{P}\left(X > \frac{\bar{y}-3200}{\sqrt{\frac{s^2}{n}}} \right) \right\}$, for $X \sim \mathcal{N}(0,1)$. Since $\bar{y} > 3000$, the p-value equals $2 \cdot \mathbb{P}\left(X > \frac{\bar{y}-3200}{\sqrt{\frac{s^2}{n}}} \right) = 0.0285$, which equals the p-value computed in question (a).

(f) -

(g) Based on the 1000 simulations an estimate of the power is approximately 0.5.

(h) For $n = 100$ an estimate of the power is approximately 0.35 and for $n = 50$ approximately 0.25 .

Week 2:
Independent samples testing

Questions

For the questions of this week we will continue to use the IVF dataset (available on Canvas). If you haven't done so already, download this datasets and place it in the "myfolders" directory you created during the installation of your SAS University Edition. The dataset should be available under "My Folders" in SAS. Note: you might need to click the refresh button.

The IVF dataset contains three measurements of the infant motor profile (`IMP`) at different periods (`PER`). This is inconvenient when analyzing other variables, because every measurement is duplicated three times. Start by creating a new dataset without the `IMP`, `PER`, and `AGE` variables, that contains only a single observation of each child. You need to use this smaller dataset to answer the questions below.

Question 2.1

In this exercise we investigate whether stress during delivery (`FIS`) is more likely to occur with older/younger mothers. For this question you may assume that the mother's age (`AGEM`) is normally distributed.

(a) Test the homogeneity of variance in `AGEM` between children who experienced stress during delivery and those who didn't, using an F-test, Bartlett's test, and Levene's test. For each test, report the null hypothesis, the test statistic, and the p-value. Do all tests support the same conclusion?

(b) Use a t-test to test for a different average age of mothers between the groups in (a). Use your result in (a) to select either the Pooled or Satterthwaite approach. Report the null hypothesis, the test statistic, the p-value, and your conclusion.

Question 2.2

For this exercise we only consider those observations with a patient identity less or equal to 100 (`ID <= 100`) in the IVF dataset. Start by creating a new dataset that only contains these observations. In this question we will analyze the effect of treatment (`TRT`) on birth weight (`BW`).

(a) Use a Wilcoxon rank sum test to test for a difference in birth weight between the different treatment groups. Report the null hypotheses, the test statistic, and the p-value for each pair of treatments, what is your conclusion?

(b) Use your answer in (a) to compute the test statistic and the p-value for the Mann-Whitney U test. Does your conclusion in (a) change?

(c) What is the probability that a child from the control group has a higher birth weight than a child whose mother received treatment-M (`TRT=1`), and similarly between a child from the control group and a child whose mother received treatment-C (`TRT=2`)?

(d) Use a Kolmogorov-Smirnov test to test for a difference in birth weight between the different treatment groups. Report the null hypotheses, the test statistic, and the p-value for each pair of treatments, what is your conclusion?

(e) From week 1 we know that both the birth weight is not normally distributed, but becomes approximately normal after an appropriate Box-Cox transformation. Use this transformation to perform a t-test to test for a difference in birth weight between the different treatment groups. Report the null hypotheses, the test statistic, and the p-value for each pair of treatments, what is your conclusion?

(f) How do your answers in (a), (e), and (d) change when you consider the whole dataset instead of only the first 100 observations?

Question 2.3

In this exercise you will create a macro that outputs the results of a Mann-Whitney U test.

(a) Extend the macro below to output the results of a Mann-Whitney U test. Your macro should output the p-value, the value of the test statistic and the probability that a sample from class 0 is less than a sample from class 1.

```
%macro mann_whitney_u(dataset,class,var);
ods select none;
proc npar1way data=&dataset;
    var &var;
    class &class;
    exact wilcoxon / mc;
    ods output WilcoxonScores=OUT_SCORES;
    ods output WilcoxonTest=OUT_TEST;
run;
ods select all;

    --- YOUR CODE ---

%mend;
```

(b) Verify the correctness of your macro by using it to answer Question 2.2, parts (b), and (c), and compare your answer.

Question 2.4

In this exercise we will analyze the relation between birth weight (`BW`) and gestational age (`GA`) in the IVF dataset.

The gestational age (`GA`) measures the duration in pregnancy in weeks. We want to test whether a longer gestational age is likely to cause heavier babies, and vice versa.

(a) We call a baby heavy when his/her birth weight exceeds 4000g. Create a new column that indicates whether a baby is heavy.

(b) Use a t-test to test for a difference in mean gestational age between children born heavier than 4000g and children born 4000g or lighter. Report the null hypotheses, the test statistic, the p-value, and your conclusion. Hint: Use an F-test to select between the Pooled and Satterthwaite approach.

(c) We call a pregnancy late, when the gestational age is exceeds 41 weeks. Create a new column that indicates which pregnancies are late.

(d) Use a t-test to test for a difference in mean birth weight between pregnancies exceeding 41 weeks and the remaining pregnancies. Report the null hypotheses, the test statistic, the p-value, and your conclusion.

(e) From week 1 we know that both the birth weight and the gestational age are not normally distributed, but both become approximately normal using an appropriate transformation. Are your results in b and d reliable?

Question 2.5

In this exercise you will consider measurements of blood-pressure that are quantified as ‘High’ or ‘Low’. The effect of a new medicine has been measured in a sample of 200 patients, 100 of these patients were randomly selected and assigned to receive the new treatment, while the other patients got a placebo. The observations are summarized in the following contingency table.

	Low	High	
Control	77	23	100
Treatment	81	19	100
	158	42	200

Test for a treatment effect of this drug using the Chi-square test. Report the null-hypothesis, the test results, and your conclusion. Do you think it is appropriate to use this test?

Question 2.6

In this exercise we will test whether stress during delivery (`FIS`) is more/less likely depending on gender.

(a) Test whether girls (`SEX = 1`) or boys (`SEX = 0`) are more/less likely to experience stress during delivery. First compute $\widehat{p}_{\text{girl}}$ and $\widehat{p}_{\text{boy}}$, and use these to derive the value of the test statistic and the p-value. What is your conclusion?

(b) Create a SAS macro that answers (a) for an arbitrary dataset and column names. You can use the code below as a starting point.

```
%macro binary_hypothesis(dataset,var,class);
    --- YOUR CODE ---
%mend;
```

(c) Verify the correctness of your macro by running the code below and comparing the result to your answer in (a).

```
%binary_hypothesis(IVF_DATASET,FIS,SEX);
```

(d) Use a chi-squared test to test whether girls (`SEX = 1`) or boys (`SEX = 0`) are more/less likely to experience stress during delivery. How do your results compare to those of (a), what if you square the test statistic you obtained in (a)?

Question 2.7

In this exercise we will test whether stress during delivery (`FIS`) is depended on which treatment (`TRT`) the mother received.

(a) To test whether the treatment has an effect on the stress during delivery, perform a Fisher exact test and a chi-squared test for each pair of treatments. Report the null hypotheses, the test statistic, and the p-value for each pair of treatments, what is your conclusion?

(b) Instead of performing the chi-squared test pairwise as in (a) it is possible consider all three treatments simultaneously. Perform a single chi-square test to test whether the treatment has an effect on the stress during delivery. Report the null hypotheses, the test statistic, the p-value, and your conclusion. Has the conclusion changed from your answer in (a), can you explain why?

Question 2.8

In the table below is an example of content uniformity between two batches. Create a new dataset containing the table below, and use this dataset to answer the following questions.

Obs.	1	2	3	4	5	6	7	8	9	10
Batch 1	102	104	102	97	99	101	103	98	96	97
Batch 2	99	97	99	100	99	96	99	98	97	98

(a) Test the homogeneity of variance between the two batches, using an F-test, Bartlett's test, and Levene's test. For each test, report the null hypothesis, the test statistic, and the p-value. Do all tests support the same conclusion?

(b) Use a t-test to test for a difference in content uniformity between the two batches. Report the null hypotheses, the test statistic, and the p-value, what is your conclusion?

(c) Use a Wilcoxon rank sum test to test for a difference in content uniformity between the two batches. Report the null hypotheses, the test statistic, and the p-value, what is your conclusion?

(d) Use a Mann-Whitney U test to test for a difference in content uniformity between the two batches. Report the null hypotheses, the test statistic, and the p-value, what is your conclusion? What is the probability that a sample from batch 1 is smaller than a sample from batch 2?

(e) Use a Kolmogorov-Smirnov test to test for a difference in content uniformity between the two batches. Report the null hypotheses, the test statistic, and the p-value, what is your conclusion?

Answers

Answer to Question 2.1

(a) The null hypothesis is H_0: $\sigma^2_{\texttt{FIS=0}} = \sigma^2_{\texttt{FIS=1}}$. From an F-test, the statistic is $F = 1.03$ with p-value $p = 0.8490$. From an Bartlett's test, the statistic is $\chi^2 = 0.0288$ with p-value $p = 0.8653$. From an Bartlett's test, the statistic is $F = 0.02$ with p-value $p = 0.8813$. Therefore, all tests don't reject the null hypothesis.

(b) The null hypothesis is H_0: $\mu_{\texttt{FIS=0}} = \mu_{\texttt{FIS=1}}$. Since we didn't reject the null hypothesis in (a), we'll assume that the variances in both groups are equal and thus use the Pooled t-test. For this, the value of the statistic is $T = -0.83$ with p-value $p = 0.4065$. Therefore, we don't reject the null-hypothesis.

Answer to Question 2.2

(a) We do three tests, the null hypothesis are: H_0: $m_{\texttt{TRT=0}} = m_{\texttt{TRT=1}}$, H_0: $m_{\texttt{TRT=0}} = m_{\texttt{TRT=2}}$, and H_0: $m_{\texttt{TRT=1}} = m_{\texttt{TRT=2}}$, where m is used to denote the median. The statistic and normal approximated p-value (notice that the exact MC p-values are very close) for each pair are:

TRT	TRT	statistic (S)	p-value
0	1	256.0	0.1829
0	2	519.5	0.0164
1	2	261.5	0.4891

We conclude that babies from the control group have a different median birth weight than those that received the treatment-C.

(b) The p-values remain the same, the value of the statistics (U_1 presented, but U_2 is also a valid statistic) are:

TRT	TRT	statistic (U_1)	p-value
0	1	277	0.1829
0	2	538.5	0.0164
1	2	170.5	0.4891

(c) The probability that a child from the control group has a higher birth weight than a child whose mother received treatment-M is 0.6267.

The probability that a child from the control group has a higher birth weight than a child whose mother received treatment-C is 0.6886.

(d) We do three tests, the null hypothesis are: H_0: $F_{\texttt{TRT=0}} = F_{\texttt{TRT=1}}$, H_0: $F_{\texttt{TRT=0}} = F_{\texttt{TRT=2}}$, and H_0: $F_{\texttt{TRT=1}} = F_{\texttt{TRT=2}}$, where F is used to denote the (cumulative) distribution function. The statistic and (normal approximated) p-value for each pair are:

TRT	TRT	statistic (S)	p-value
0	1	0.3620	0.1700
0	2	0.4041	0.0227
1	2	0.2542	0.6566

We conclude that babies from the control group have a different birth weight distribution than those that received treatment-C.

(e) We again do three tests where the pairwise null hypothesis are if the means of the Box-Cox transformed variables are equal between the treatments. The (equal variance) statistic and p-value for each pair are:

TRT	TRT	statistic (T)	p-value
0	1	1.10	0.2771
0	2	2.15	0.0356
1	2	0.67	0.5053

Using this approach the conclusion does not change, we only see a significant difference between the control group and the treatment-C group.

(f) -

Answer to Question 2.3

(a) The following SAS macro performs the Mann-Whitney U test.

```
%macro mann_whitney_u(dataset,class,var);
ods select none;
proc npar1way data=&dataset;
    var &var;
    class &class;
    exact wilcoxon / mc;
    ods output WilcoxonScores=OUT_SCORES(
            rename=(SumOfScores=S));
    ods output WilcoxonTest=OUT_TEST(
            rename=(cValue1=P_VALUE)
            where=(Name1="P2_WIL"));
run;
ods select all;

data OUT_SCORES;
    set OUT_SCORES;
    CLASS_ID = _N_ - 1;
run;

proc transpose data=OUT_SCORES
        out=OUT_N(drop=_NAME_) prefix=N;
    id CLASS_ID;
    var N;
run;

proc transpose data=OUT_SCORES
        out=OUT_S(drop=_NAME_ _LABEL_) prefix=S;
    id CLASS_ID;
    var S;
run;

data RESULT;
    merge OUT_N OUT_S OUT_TEST(keep=P_VALUE);
    U0 = S0 - N0 * (N0+1)/2;
    U1 = S1 - N1 * (N1+1)/2;
    P0 = U0 / (N0*N1);
    P1 = U1 / (N0*N1);
run;

title "Mann Whitney U test";
proc print data=OUT_SCORES label noobs;
    var CLASS_ID CLASS;
    label CLASS_ID="class"
          CLASS="group identifier";
run;
title;
proc print data=RESULT label noobs;
    var P_VALUE U0 U1 P0 P1;
    label P_VALUE="p-value"
          U0="statistic (U0)"
          U1="statistic (U1)"
          P0="P(class0 >  class1)"
          P1="P(class0 <= class1)";
run;
%mend;
```

(b) See the answers of Question 2.8, (b) and (b).

Answer to Question 2.4

(a) -

(b) The F-value equals 5.30 giving rise to a p-value < 0.0001. Thus, the groups are clearly not homoskedastic. Therefore, we use a T-test assuming unequal variances, T-statistic equals -9.07 with a corresponding p-value < 0.0001. We reject the $H_0 : \mu_{GA|BW>4000} = \mu_{GA|BW\leq 4000}$ and conclude a difference in mean gestational age.

(c) -

(d) The F-value equals 1.77 giving rise to a p-value of 0.0407. Thus, the groups are clearly not homoskedastic. Therefore, we use a T-test assuming unequal variances, T-statistic equals -6.55 with a corresponding p-value < 0.0001. We reject the $H_0 : \mu_{BW|GA>41} = \mu_{BW|GA\leq 41}$ and conclude a difference in mean birth weight.

(e) Since $n = 253$ for both analyses (no missing values) the results are probably reliable based on the central limit theorem (CLT). To be absolutely sure it is wise to draw several fictional samples from the observed population (of course with replacement, do you see why?) and check for normality of the sample average distribution.

Answer to Question 2.5

The counts in each cell are > 5, so it seems reasonable to assume that the Chi-square approximation is appropriate. We test the $H_0 : p_{\texttt{High, control}} = p_{\texttt{High, treatment}}$, the test statistics equals 0.4822, the p-value based on the χ^2_1 approximation equals 0.4874, the exact

p-value equals 0.6029. We do not reject the H_0 based on this dataset, there is no indication that the treatment influences the probability of having a high blood pressure.

Answer to Question 2.6

(a) $\hat{p}_{\texttt{girl}} = 0.31092$, $\hat{p}_{\texttt{boy}} = 0.36090$ and $\hat{p} = 0.3373$, the value of the test statistic $\frac{\hat{p}_{\texttt{girl}} - \hat{p}_{\texttt{boy}}}{\sqrt{\hat{p}(1-\hat{p})(n_{\texttt{girl}}^{-1} + n_{\texttt{boy}}^{-1})}} = -0.8377$, such that the approximate (asymptotically valid) p-value equals 0.4022.

(b)

```
%macro binary_hypothesis(dataset,var,class);
proc means data=&dataset n sum noprint;
var &var;
class &class;
output out=OUT n=N sum=COUNT;
run;

data OUT0;
set OUT;
COUNT0 = COUNT;
N0 = N;
P0 = COUNT0 / N0;
where SEX = 0;
keep COUNT0 N0 P0;
run;

data OUT1;
set OUT;
COUNT1 = COUNT;
N1 = N;
P1 = COUNT1 / N1;
where SEX = 1;
keep COUNT1 N1 P1;
run;

data OUT;
merge OUT0 OUT1;
P = (COUNT0 + COUNT1) / (N0 + N1);
STAT = (P0 - P1) / sqrt(P * (1-P) * (1/N0 + 1/N1));
CHISQ = STAT**2;
P_VALUE = 2*min(cdf("normal", STAT, 0, 1),1-cdf("normal", STAT, 0, 1));
run;

proc print data=OUT;
var STAT CHISQ P_VALUE;
run;
%mend;
```

(c) -

(d) The value of the χ^2 statistic equals 0.7018 which equals $(-0.83774)^2$. The exact p-value equals 0.4256 and we cannot reject the $H_0 : p_{\texttt{boys}} = p_{\texttt{girls}}$.

Answer to Question 2.7

(a) We do three tests, the null hypothesis are: H$_0$: $p_{\texttt{TRT=0}} = p_{\texttt{TRT=1}}$, H$_0$: $p_{\texttt{TRT=0}} = p_{\texttt{TRT=2}}$, and H$_0$: $p_{\texttt{TRT=1}} = p_{\texttt{TRT=2}}$, where p is used to denote the proportion of children is stress during delivery. The statistic and p-value for each pair are:

		Fisher's exact test		Chi-squared test	
`TRT`	`TRT`	statistic (P)	(exact) p-value	statistic (χ^2)	p-value
0	1	53	0.1202	2.8533	0.0912
0	2	53	0.0347	4.9708	0.0258
1	2	41	0.8563	0.0458	0.8306

Using Fisher's exact and the Chi-squared test we see that children from the control group have a significantly different chance to experience stress during delivery than those of mothers that received treatment-C.

(b) We do a single test with null hypothesis H_0: $p_{\texttt{TRT=0}} = p_{\texttt{TRT=1}} = p_{\texttt{TRT=2}}$. This gives a statistic $\chi^2 = 5.7210$ and (an approximate) p-value of $p = 0.0572$. Now we cannot reject that there is no significant difference in `FIS` between the different treatments. However, the probability that the statistic would attain this value seems quite unlikely (close to 0.05), typically we would recommend the experiment designer to extend the research to be sure about our finding.

Answer to Question 2.8

(a) $H_0 : \sigma^2(f_1) = \sigma^2(f_2)$, F-statistic equals 5.36, $p = 0.0199$, Bartlett's statistic equals 5.4128, $p = 0.0200$, Levene's statistic equals 10.86, $p = 0.0040$, all tests reject homoskedasticity.

(b) $H_0 : \mu(f_1) = \mu(f_2)$, T-statistic: 1.73, p-value for T-test with unequal variances 0.1080. Thus, we do not reject the H_0 of equal means.

(c) Since we concluded that the batches are heteroskedastic, we test the null-hypothesis of no stochastic ordering. The statistic, $S_1 = 120.5$, and an exact (two sided) p-value equals 0.2483. Therefore, we cannot reject the null-hypothesis of no stochastic ordering.

(d) The H_0, p-value and conclusion are the same as in Question c. The Mann-Whitney-U equals 34.5 and the probability asked for equals 0.345. *Hint: Be aware of the fact that one could also ask for the probability that a sample of batch* 1 *is larger than a sample from batch* 2 *which equals* $1 - 0.345 = 0.655$.

(e) The $H_0 : F_1 = F_2$, the test statistic equals 0.5 and exact p-value equals 0.0964. Thus, we cannot reject the null-hypothesis of equal distributions.

Week 3:
Paired samples testing, correlation and copula's

Questions

For the questions of this week we will continue to use the IVF dataset (available on Canvas), and start with using the RCT dataset (also available on Canvas). If you haven't done so already, download these datasets and place them in the "myfolders" directory you created during the installation of your SAS University Edition. Both datasets should be available under "My Folders" in SAS. Note: you might need to click the refresh button.

Question 3.1

In this exercise we will simulate data (X, Y) and analyze this dataset subsequently. The marginal distributions of X and Y are Uniform and their dependence structure is modeled by a Gumbel copula with $\alpha = 5$. Use the SAS code below to generate the dataset `GumC`:

```
%Macro SIM_Gum(alpha=, nsim=, seed=);
proc iml;
call streaminit(&seed);
alpha=&alpha;

do i=1 to &nsim by 1;
U1=rand(’Uniform’);
U2=rand(’Uniform’);

start Func(x) global(U1,U2,alpha);
return(Exp(-((-Log(x))**alpha + (-Log(U1))**alpha)**(1/alpha))
*((-Log(x))**alpha + (-Log(U1))**alpha)**(-1 + 1/alpha)*
((-Log(U1))**(alpha-1))/U1-U2);
finish;

intervals = {0.00001 1};
U2C = froot("Func", intervals);

X=X//U1;
Y=Y//U2C;
YI=YI//U2;
end;

Total=X||Y||YI;

create GumC from Total [colname={’X’,’Y’,’YI’}];
append from Total;
close GumC;
quit;
%mend SIM_Gum;

%SIM_Gum(nsim=1000, alpha=5, seed=12345);
```

(a) Estimate Kendall's τ and Spearman's ρ correlation coefficients. Do you reject dependence of X and Y?

(b) What is the real value of τ for the distribution used to generate the data?

(c) Estimate α of Gumbel's copula using the expressions for Kendall's τ and Spearman's ρ.

(d) Estimate α of Frank's copula using the expressions for Kendall's τ and Spearman's ρ.

(e) Plot 1000 simulated points from both models fitted in (c) and (d) and compare these to a plot of the original data. Which copula do you prefer to explain the dependence structure between X and Y?

Question 3.2

In this exercise we will analyze the infant motor profile IMP from the IVF dataset. Start by adding some continuous noise to the IVF dataset:

```
DATA IVF;
SET IVF;
IMP = IMP + (ranuni(1)-0.5);
run;
```

(a) Create, as done in Lecture, a new dataset of the IMP over time per baby in the wide format using PROC TRANSPOSE.

(b) Report Pearson's ρ_p, Kendall's τ and Spearman's ρ_s for IMP4 and IMP18. Use each of these correlation coefficients to test for dependence and report the p-values. What does your conclusion mean in practice?

(c) Estimate the copula parameter for all copula's discussed during lecture using the expressions fors both Kendall's τ and Spearman's ρ_s, what do you conclude from the comparison of the two?

Question 3.3

In this exercise we will analyze the joint distribution of two random variables $X \sim U[0,1]$ and $Y \sim U[0,1]$. Hint: a random variable with distribution $U[0,1]$ has $F(x) = F^{-1}(x) = x$.

(a) Compute $\mathbb{P}(X \leq 0.7, Y \leq 0.7)$ if X and Y are independent.

(b) Compute $\mathbb{P}(X \leq 0.7, Y \leq 0.7)$ if the dependence between X and Y is described by a Gumbel copula for $\alpha \in \{1, 2, 5, 10\}$. Can you use your findings to explain what kind of structure is introduced by the Gumbel copula?

(c) Compute $\mathbb{P}(X \leq 0.7, Y \leq 0.7)$ if the dependence between X and Y is described with a Clayton copula for $\alpha \in \{2, 5, 10\}$.

(d) Compute $\mathbb{P}(X \leq 0.7, Y \leq 0.7)$ if the dependence between X and Y is described with Franks's copula for $\alpha \in \{-5, -1, 1, 5\}$.

(e) Compute $\mathbb{P}(X \leq 0.7, Y \leq 0.7)$ if the dependence between X and Y is described with a FGM copula for $\alpha \in \{-0.5, 0, 0.5\}$.

Question 3.4

In this exercise we will analyze the hemoglobin values of patients RESP from the RCT dataset.

(a) Create a new dataset consisting of the RESP over time per patient in the wide format.

(b) Report Pearson's ρ_p, Kendall's τ and Spearman's ρ_s for RESP1 and RESP2. Use each of these correlation coefficients to test for dependence and report the p-values. What does your conclusion mean in practice?

(c) Write a macro that estimates the parameters of all the copula structures discussed in the lecture (Gaussian, Gumbel, Clayton, Frank's and FGM) based on Kendall's τ and Spearman's ρ_s.

(d) Provide a 95% confidence interval for ρ_p based on the distribution of the Fishers's z. Why would you use this CI instead of the one based on the t-distribution? Hint: use `proc corr pearson fisher`.

(e) Use both $\hat{\rho_s}$ and $\hat{\tau}$ from (b) to estimate the FGM copula's α that best fits the data. Do you think the FGM copula is a reasonable choice to model the dependence between RESP1 and RESP2.

(f) Use $\hat{\tau}$ from question (b) to estimate the Clayton copula's α.

(g) Use $\hat{\rho_s}$ instead to estimate Clayton copula's α that fits the dependence between RESP1 and RESP2. Does this estimate agree with your answer in (f)?

(h) Estimate the α of Frank's copula from the data. Do you prefer Frank's or Clayton's copula to model the dependence between RESP1 and RESP2? Hint: simulate 700 points from both models and compare these to the scatter plot of the data.

Question 3.5

Let $Y = \alpha + \beta X^2$, where $X \sim \mathcal{N}(0,1)$ and $\alpha, \beta \in \mathbb{R}$. Derive an expression for Kendall's τ.

Question 3.6

In this exercise we will analyze the hemoglobin values of patients `RESP` from the RCT dataset of the fifth (`TIME = 5`) and sixth (`TIME = 6`) measurement. We only consider patients in the first center (`CENTER = 1`).

(a) Create a new dataset consisting of the `RESP` over time per patient in the wide format. Furthermore, create columns for the difference/log-difference/ratio between the fifth and sixth measurements. Investigate the shape of the distribution of these new variables by plotting a histogram.

(b) Perform a t-test, sign test and Wilcoxon signed rank test to test H$_0$: $F_1 = F_2$. Elaborate on this null-hypothesis for the different tests and different variables used. What do you conclude on mean/median RESP value at the fifth and sixth measurement?

(c) Which of the test results from the previous question seem reliable? In case the results are not reliable, report which assumptions were violated.

(d) Suppose we assume that the difference between the fifth and sixth measurements are normally distributed. Rank the t-test, the sign test and the Wilcoxon signed rank test (on these differences) in order of your preference and explain why?

(e) Suppose we assume that the distribution of the difference between the fifth and sixth measurements is not even symmetrically distributed. Can we use the t-test, the sign-test or the Wilcoxon-signed-rank test. If so, which null-hypothesis are we testing?

Question 3.7

In this exercise we will analyze the infant motor profile `IMP` at 4 and 18 months.

(a) Create a new dataset consisting of the `IMP` over time per child in the wide format. Furthermore, create columns for the difference between/log-difference between/ratio of the measurements at 4 and 18 months. Investigate the shape of the distribution of these new variables by plotting a histogram.

(b) Perform the t-test, sign test and Wilcoxon signed rank test to test H$_0$: $F_4 = F_{18}$. Elaborate on this null-hypothesis for the different tests and different variables used. What do you conclude on mean/median `IMP` value at 4 and 18 months?

(c) Which of the test results from the previous question seem reliable? In case the results are not reliable, report which assumptions were violated.

Question 3.8

Two coagulation tests exist for measuring thickness of blood; the Celite method and the Kaoline method. These methods measure the time (in seconds) it takes for blood to change from liquid to a gel, referred to as the clotting time. In total samples from 21 patients were collected and measured with both methods. Use the code presented below to import the data manually in SAS.

```
data COAG;
input Patient C K@@;
datalines;
1   120     132     8   145     133     15  117     123
2   114     116     9   120     123     16  125     108
3   129     135     10  129     116     17  136     131
4   128     115     11  126     127     18  151     119
5   155     134     12  136     140     19  130     129
6   105     56      13  135     140     20  136     124
7   114     114     14  125     114     21  113     112;
run;
```

(a) Compute the Pearson's, Kendall's and Spearman's correlation coefficient and test for dependence between the `C` and `K` measurements.

(b) Use the obtained results to estimate the associations of the Gumbel and FGM copula. Which of these models has your preference?

(c) Test if the `C` and `K` measurements are the same using the different test (on difference, log difference and ratio). Which test is most appropriate here?

(d) A blood thickness value below 120 means that the blood is too thin, create a new binary variable `TT` representing this event. Compute the phi-correlation of `TT` between the two methods.

(e) Test if the two `TT` measurements are the same on decision using the McNemar test. If so, would you indicate this relation as strong?

Question 3.9

In this exercise we will analyze the infant motor profile (`IMP`) at 10 and 18 months. Create a dataset where the `IMP` scores at 10 and 18 months are in different columns next to a column that indicates the patient number (`ID`). An `IMP` score lower than 85 points at age 10 and/or 18 months is considered a sign for a neurological problem.

(a) Use the sign test to determine if the `IMP` score at 18 months is larger than the `IMP` score at 10 months. Report the null hypothesis, the result on the binomial test statistic, the p-value and your conclusion.

(b) Investigate with the McNemar test whether there is difference in the probability of having a neurological problem between 10 and 18 months.

(c) Did you make sure to remove the missing data in the dataset before translating the `IMP` score below 85 to a binary variable? If not, please redo your analysis. To delete the missing values you can use the code below.

```
data WIDE_IVF;
    set WIDE_IVF;
    if cmiss (of IMP10 IMP18 ) then delete;
run;
```

(d) Report the Kappa statistic together with its 95% confidence interval and formulate precisely what this statistic means in this situation of neurological outcome.

Question 3.10

In this exercise you will investigate characteristics of the distribution of the κ statistic of an underlying two by two contingency table. The macro below can be used to simulate contingency tables of size n with p_{00}, p_{01} and p_{11}.

```
%macro kappatest(nsim=,ntab=,n00=,n01=,n11=);
    proc iml;
        G0 = (1-&n00-&n01-&n11);
        G  = {&n00 &n01 &n11};
        call randseed(1234);
        C = {"SIMID", "N00", "N01", "N11", "N10"};
        PROB = G||G0;
        X = RandMultinomial(&nsim, &ntab, PROB);
        ID = 1:&nsim;
        X = ID`||X;
        create CT from X[c=c];
        append from X;
        close CT;
    quit;

    data CT;
        set CT;
        EPO = 1 - N11/&ntab-N00/&ntab;
        EPE = 1 - (N00+N10)*(N00+N01)/(&ntab**2)
                - (N01+N11)*(N10+N11)/(&ntab**2);
        EKAPPA = 1 - EPO/EPE;
        PO = 1 - &n11 - &n00;
        PE = 1 - (&n00+(1-&n00-&n01-&n11))*(&n01+&n00)
               - (&n01+&n11)*((1-&n00-&n01-&n11)+&n11);
        KAPPA = 1 - PO/PE;
    run;
%mend;
```

(a) Use the macro to simulate 1000 studies where $n = 100$, $p_{00} = 0.05$, $p_{11} = 0.3$ and $p_{01} = 0.01$ and check the variability of your $\hat{\kappa}$ (e.g. by plotting a histogram using `proc univariate`). Can you understand that one should be careful when computing the κ statistic in studies with rare-events.

(b) Let $p_{00} = p_{11} = 0.3$, compute the κ for $p_{01} \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.35, 0.39\}$ to illustrate that the κ is sensitive to the row and column totals in a contingency table.

(c) Use $p_{00} = p_{11} = 0.3$ and $p_{01} = 0.1$, such that $\kappa \approx 0.23$ and perform the McNemar test on each of these datasets using the following statement.

```
data CT;
    set CT;
    MCN = (N01-N10)**2/(N01+N10);
    PMCN_NAPPROX = 1 - CDF('CHISQ',MCN,1);
    PMCN_EXACT   = 2*MIN(CDF('BINOM',N01,0.5,N01+N10),
                         1-CDF('BINOM',N01,0.5,N01+N10));
run;
```

What do you conclude from this experiment? Would you advice researchers to compute the Kappa statistic when performing a McNemar test and why?

Question 3.11

In this exercise we will consider measurements of blood-pressure that are quantified as 'High' or 'Low'. The effect of two new medicines have been measured in a sample of 50 patients. The paired (rather tripled) measurements $(y_{11}, y_{21}, y_{31}), (y_{21}, y_{22}, y_{32}), \ldots, (y_{501}, y_{502}, y_{503})$, consist of a control measurement (pre-treatment) and response to drug A-, and drug B-treatment. The observations are summarized in the following two contingency tables.

		Post treatment		
		High	Low	
Control	High	12	11	23
	Low	7	70	77
		19	81	100

Table 1: Drug A

		Post treatment		
		High	Low	
Control	High	10	13	23
	Low	4	73	77
		14	86	100

Table 2: Drug B

(a) Test for a treatment effect of drug A using an appropriate test. Report the null-hypothesis, the test results, and the conclusion of the hypothesis.

(b) Compute the Kappa statistic for agreement between the control measurement and the measurement after using drug B and formulate precisely what this statistic means in this situation of high/low blood-pressure.

(c) A Chi-square statistic can be computed on contingency Table 1. The p-value is less than 0.0001. Do you think this test is appropriate to verify an effect of drug A?

(d) Could we use the Chi-square test in a different (more appropriate) way? If so, report the test statistic, p-value and your conclusion.

Question 3.12

The McNemar test is used for paired measurements (y_1, y_2) to test $p_{1\cdot} = p_{\cdot 1}$. The null hypothesis implies that $\hat{p}_1 - \hat{p}_2 = \frac{N_{10} - N_{01}}{n}$ is close to zero, were the $N_{\cdot\cdot}$ are defined in the table below.

	$y_2 = 0$	$y_2 = 1$	
$y_1 = 0$	N_{00}	N_{01}	$N_{0\cdot}$
$y_1 = 1$	N_{10}	N_{11}	$N_{1\cdot}$
	$N_{\cdot 0}$	$N_{\cdot 1}$	n

Table 3: McNemar contingency table

(a) Derive the variance of $\frac{N_{10}-N_{01}}{n}$.

In Question 3.11 you demonstrated that data presented in the form of Table 3 can, after creating a new Table, be analyzed using a Chi-square test. However, this would ignore the paired structure and thus ignores information that is present.

(b) Create a new contingency table using the row and column totals of Table 3 such that the Chi-square test can be used to test $p_{1\cdot} = p_{\cdot 1}$. Present the statistic $\widehat{p}_{1\cdot} - \widehat{p}_{\cdot 1}$ expressed in the row and column totals of Table 3.

(c) Derive the variance of $\widehat{p}_{1\cdot} - \widehat{p}_{\cdot 1}$.

(d) Use the results of this exercise to explain to a researcher why the McNemar test should be used rather than the Chi-square test in case of paired measurements.

Answers

Answer to Question 3.1

(a) $\hat{\tau} = 0.81311$ and the p-value < 0.0001, $\hat{\rho_s} = 0.95131$ and the p-value < 0.0001.

(b) We simulated from a Gumbel copula with $\alpha = 5$, so $\tau = 1 - \frac{1}{5} = 0.8$.

(c) Based on Kendall's τ: $\hat{\alpha} = 5.350741$, based on Spearman's ρ_s: $\hat{\alpha} = 5.411538$.

(d) Based on Kendall's τ: $\hat{\alpha} = 19.607393$, based on Spearman's ρ_s: $\hat{\alpha} = 18.473158$.

(e) We simulate 1000 datapoints from both copulas using $\hat{\alpha}_{\text{Gumbel}} = 5.38$ and $\hat{\alpha}_{\text{Frank}} = 19$ running:

```
%SIM_Gum(nsim=1000, alpha=5.38, seed=6789);
%SIM_Frk(nsim=1000, alpha=19, seed=6789);
```

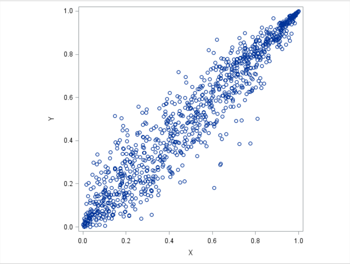


Figure 1: Original simulated data

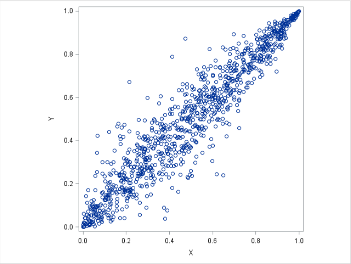


Figure 2: Estimated Gumbel copula

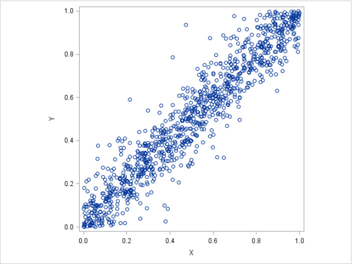


Figure 3: Estimated Frank copula

Of course we prefer the Gumbel copula. Make sure you focus on the simulated values using Frank's copula and see how this scatter plot deviates from the original data.

Answer to Question 3.4

(a) -

(b) $\hat{\rho_p} = 0.47800$, $\hat{\rho_s} = 0.50477$ and $\hat{\tau} = 0.35941$, the p-value to test independence based on the different correlation coefficients is < 0.0001 for all three methods.

(c) -

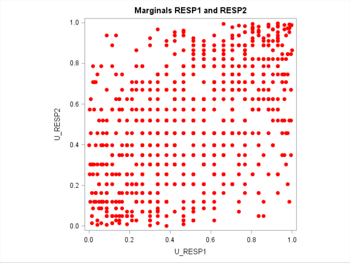
(d) A 95% confidence interval for the ρ_p based on the Fisher's z transformation equals $(0.419252, 0.532766)$.

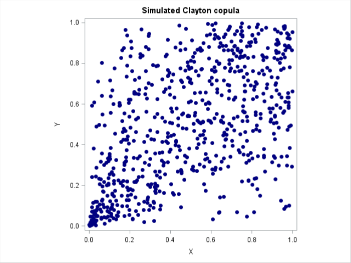
(e) The $\hat{\alpha}$ estimates for the FGM copula equal 1.617345 and $3 \cdot \hat{\rho_s} = 3 \cdot 0.50477 = 1.5143$ based on Kendall's τ and Spearman's ρ_s. Since the FGM copula parameter has a support of $[-1, 1]$ the copula with $\alpha = 1.5$ or $\alpha = 1.6$ does not exist. We can conclude that the FGM copula is not an appropriate way to model the dependency.

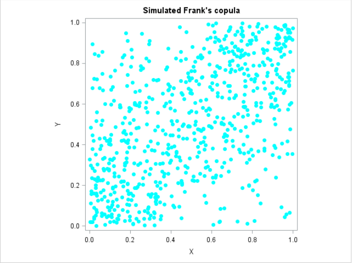
(f) $\hat{\alpha} = \frac{2\hat{\tau}}{1-\hat{\tau}} = \frac{2 \cdot 0.35941}{1-0.35941} = 1.12212$.

(g) $\hat{\alpha} = 1.094$ which is very close to the previous estimate. As a result of this observation we cannot rule out that the Clayton's copula appropriatly models the dependency.

(h) The $\hat{\alpha}$ estimates for Frank's copula equal 3.63 and 3.49 based on Kendall's τ and Spearman's ρ_s. We compare the plot of the `RESP2` percentiles versus the `RESP1` percentiles with simulations of Frank's and Clayton's copulas:



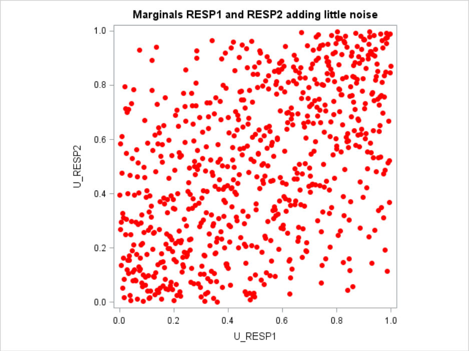




It is crucially to note that the RESP outcomes are discrete outcomes. Therefore this visual inspection is misleading: each point could represent one pair of observations or multiple. To overcome this issue we add little Uniform noise to the RESP outcomes using

```
DATA WIDE3;
SET WIDE3;
RESP1=RESP1 + 0.1*(ranuni(1)-0.5);
RESP2=RESP2 + 0.1*(ranuni(1)-0.5);
run;
```

Now the joint distribution of the percentiles do give information and we conclude that Frank's copula does better describe the dependency structure.



Answer to Question 3.6

(a) The histograms of the difference and the ratio suggest a distribution which is almost symmetric and perhaps even normal. The histogram of the log-difference suggests a symmetric distribution, but normality is not apparent.

(b) Let $z_i = \psi(y_{1j}, y_{2j})$.

i For the paired t-test, we assume z_j to be normal with mean μ and variance σ. Now, $H_0 : F_1 = F_2$ would correspond to $\mu = 0$ for the difference and log difference, and $\mu = 1$ for the ratio. The resulting t statistics are: -1.3998, -1.44898 and -0.52916 for the difference, log difference and ratio respectively. We fail to reject the null hypothesis in all cases.

ii For the sign test, we do not make distributional assumptions. The null hypothesis is given by: $H_0 : m(y_{1j}) = m(y_{2j})$ for all variables considered. The resulting M statistics are -15 for all variables considered. We fail to reject the null hypothesis in all cases.

iii For the Wilcoxon signed rank test, we assume z_j to have a symmetric distribution. The null hypothesis is given by: $H_0 : m(y_{1j}) = m(y_{2j})$. To compute the statistic, we need to consider the distance of z_j to the nominal value μ_0. For the difference and log difference, $\mu_0 = 0$. For the ratio, $\mu_0 = 1$. The resulting S statistics are -2741, -2748 and -1970 for the difference, log difference and ratio respectively. We fail to reject the null hypothesis in all cases.

(c) i For the paired t-test, we assume normality. Normality does not seem apparent for the ratio.

ii For the sign test, no assumptions are made.

iii For the Wilcoxon signed rank test, we need symmetric distributions. The histograms suggest slightly left-skewed distributions, but nothing extreme.

(d) This order is based on the power of the tests: 1). t-test, 2).Wilcoxon-signed rank test and 3). Sign test.

(e) The t-test is not directly appropriate in this case. However, if the sample size is large one could rely on the CLT. The assumption of symmetric differences is violated so the Wilcoxon-signed rank test cannot be used to test equality in median of both groups. Finally, the sign-test does not rely on any distributional assumptions, thus can be used here to test for equality in medians of the groups.

Answer to Question 3.9

(a) $H_0 : \mu_{IMP,10} \geq \mu_{IMP,18}$, the test statistic equals 112 (SAS reports $\frac{112-105}{2}$) and the corresponding (one-sided) p-value equals 0.3419 (equal to $P(X \geq 112)$ for $X \sim Bin(112 + 105, 0.5)$.

(b) The McNemar statistic equals 7.3636, resulting in an exact p-value of 0.0117. Therefore we reject the $H_0 : p_{10} = p_{18}$ and conclude that there is a difference in the probability of having a neurological problem between 10 and 18 months.

(c) -

(d) The $\hat{\kappa} = -0.0076$, with a corresponding 95% confidence interval equal to $(-0.0211, 0.0060)$. We conclude that there is 'poor' agreement between the observations at 10 and 18 months, as expected based on the answer on the previous question.

Answer to Question 3.11

(a) The McNemar test is very appropriate in this setting. The test statistic equals 0.8889 with a corresponding exact p-value of 0.4807. Therefore, the $H_0 : p_{high,pre} = p_{high,post}$ cannot be rejected and we conclude that there is no significant treatment effect of drug A.

(b) The kappa statistic equals $\hat{\kappa} = 0.4437$ from which we result that there is 'moderate' agreement between the blood-pressure before and after treatment B. Thus, there is no evidence for a large effect of treatment with drug B.

(c) The performed Chi-square test would test if the probability to have a high blood pressure after the treatment would be independent of the blood pressure before the treatment. This specific outcome can thus not be used to verify an effect of drug A.

(d) See Question 2.5.

Week 4: Randomness and ANOVA

Questions

For the questions of this week we will continue to use the IVF dataset and the RCT dataset (both available on Canvas).

Question 4.1

In the Netherlands, local drugstores do sell quite some dietary supplements. In this question we focus on two brands that claim to have a beneficial effect on ones IQ. Import the dataset with IQ levels for subjects that took `suppA` or `suppB` using the SAS code provided below.

```
data iq;
    input supplement$ iq@@;
    datalines;
    suppB 104.3 suppB 99.0  suppB 112.5 suppB 114.0
    suppB 132.4 suppB 109.4 suppB 98.8  suppB 98.9
    suppB 112.4 suppB 101.9 suppB 97.0  suppB 112.1
    suppB 100.7 suppB 100.5 suppB 114.8 suppB 100.6
    suppB 105.3 suppB 110.5 suppB 110.7 suppB 119.3
    suppA 97.0  suppA 106.7 suppA 108.1 suppA 97.1
    suppA 96.7  suppA 105.4 suppA 105.6 suppA 110.0
    suppA 106.3 suppA 99.7  suppA 108.4 suppA 106.3
    suppA 109.5 suppA 99.8  suppA 93.7  suppA 107.7
    suppA 102.7 suppA 106.3 suppA 97.7  suppA 107.3
    ;
```

(a) Write down a one-way ANOVA to test for a difference between the two groups in mathematical terms.

(b) Test with a one-way ANOVA if the mean IQ is different between the two groups.

(c) Compute a confidence interval for the overall mean, α_A and α_B manually.

(d) Perform a two sample t-test with homogeneous variance to test for a difference in IQ between the two groups. Compare the p-value and the test statistic with those of the ANOVA analysis, can you see and explain the relation?

(e) Perform a Kruskal-Wallis test to test for a difference in IQ between the two groups.

(f) Would you prefer presenting the results of the fixed effects ANOVA, the t-test with equal variance or the Kruskal-Wallis test for this dataset?

The product range has been expanded with a new dietary supplement `suppC`. Add the following data to the dataset.

```
    suppC 103.3 suppC 104.0 suppC 117.5 suppC 119.0
    suppC 135.4 suppC 113.4 suppC 103.8 suppC 103.9
    suppC 115.4 suppC 106.9 suppC 102.0 suppC 117.1
    suppC 105.7 suppC 105.5 suppC 119.8 suppC 105.6
    suppC 110.3 suppC 115.5 suppC 115.7 suppC 124.3
```

(g) Fit a one-way ANOVA to test for a difference between the three groups and derive the CI for the estimated differences between each pair of treatments using `lsmeans trt / diff cl;`.

(h) Compare these CI with those obtained by running `proc ttest` on all three pairs separately, can you explain the difference?

Question 4.2

Assume data generated from one of the following underlying model

$$Y_{ij} = \mu + \alpha_i + e_{ij}\,, \qquad i = 1, 2, \ldots, m\,, \qquad j = 1, 2, \ldots, n\,,$$

where Y_{ij} is the j-th measurement in the i-th group with overall mean μ and measurement error e_{ij}, with $e_{ij} \sim \mathcal{N}(0, \sigma_E^2)$. We consider the following two models:

(1) α_i is the expected deviation from the overall mean for group i, $\alpha_1 + \alpha_2 + ... + \alpha_m = 0$.

(2) α_i is an unknown random parameter such that *i.i.d.* $\alpha_i \sim \mathcal{N}(0, \sigma_G^2)$, which represents the variability between the groups.

Derive the distribution ($X \sim \ldots$), the expectation ($\mathbb{E}[X] = \ldots$) and the variance ($\mathbb{E}[(X - \mathbb{E}[X])^2] = \ldots$) of the statistics presented below for *both* models.

(a) $SS_W\,,$ (d) $MS_B\,,$ (g) $\overline{y}_{i.} = n^{-1}\sum_{j=1}^{n} y_{ij}\,,$

(b) $SS_B\,,$ (e) $\dfrac{MS_B}{MS_W}\,,$ (h) $\overline{\overline{y}} = n^{-1}m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}\,,$

(c) $MS_W\,,$ (f) $MS_G - MS_E\,,$ (i) $\overline{\overline{y}} - \overline{y}_{i.}\,.$

Question 4.3

In this question we will discuss the impact of violation of the ANOVA assumptions. Report the consequences for the estimates, for the test statistic and for the final conclusion of the violations presented in the scenarios below.

(a) Imagine the residuals are not normal distributed, but are following a distribution that is skewed to the right.

(b) Imagine the residuals are not normal distributed, but are following a distribution that is skewed to the left.

(c) Imagine the random effects are not normal distributed, but are following a distribution that is skewed to the right.

(d) Imagine the random effects are not normal distributed, but are following a distribution that is skewed to the left.

(e) Imagine the residual variance across groups are not equal.

Question 4.4

Use the blood thickness measurements dataset presented in Question 3.8.

(a) Assume a patient specific blood thickness mean and fit the one-way ANOVA model.

(b) Compute the intraclass correlation coefficient (ICC). What can you say about the variability between patients compared to the variability within one patient's record.

(c) Compute the EBLUP for patient 9.

(d) Save the EBLUPS for all patients in a new dataset and sort on patient ID. To test the randomness of this sequence of EBLUPS, which test would you propose and why?

(e) Perform the suggested test and report the test statistic and p-value. Which ANOVA assumption can be verified using this test?

(f) Similarly, verify the ANOVA assumption that the residuals are independently distributed.

Question 4.5

In this exercise you need the IVF dataset, use only the data of the first period (`PER = 4`). You can assume that the infant motor profile (`IMP`) can be modeled with an ANOVA model. We will consider the fixed effect treatment (`TRT`).

(a) Write down the introduced ANOVA model mathematically, so you should specify

$$Y_{ij} = \dots ,$$

and explain all the parameters used.

(b) Formulate the traditional null and alternative hypotheses for the effect of treatment. Report the value of the test statistic, the p-value and your conclusion.

(c) Answer question (b) using the Kruskal-Wallis test. Report the value of the test statistic, the p-value and your conclusion.

Question 4.6

In this exercise you need the RCT dataset, use only the data of the first day (`TIME = 1`). You can assume that the hemoglobin values (`RESP`) can be modeled with an ANOVA model. We will consider the random effect of (`ID`).

(a) Write down the introduced ANOVA model mathematically, so you should specify

$$Y_{ij} = \dots ,$$

and explain all the parameters used.

(b) Fit the ANOVA model, and report the EBLUP for patients 1 and 2.

(c) Save the conditional residuals for all measurements in a new dataset and sort on `CENTER`. Validate the ANOVA assumption of *i.i.d* residuals by using the conditional runs test on this sequence of residuals.

Question 4.7

The following dataset has been collected. Enter this dataset manually in SAS and use it to answer to questions below.

Obs.	1	2	3	4	5	6	7	8	9	10
	104.3	132.4	112.4	100.7	105.3	99.0	109.4	101.9	100.5	110.5

Obs.	11	12	13	14	15	16	17	18	19	20
	112.5	98.8	97.0	114.8	110.7	114.0	98.9	112.1	100.6	119.3

(a) Apply the conditional randomness test, report the p-value and your conclusion.

(b) Apply the unconditional randomness test, report the p-value and your conclusion.

(c) Compute the rank serial correlation η_1.

(d) Compute the normal autocorrelation r_1.

Question 4.8

Create a SAS macro that computes the serial correlation between two time series (one possible lagged).

```
%macro sercor(dataset,var1,var2,lag);
    --- YOUR CODE ---
%mend;
```

Use your macro to answer Question 4.9 **??** and (a), does your answer agree with the result you obtained previously?

Question 4.9

Time series data on the gas production[1] and the induced tremor catalog[2] in the north of the Netherlands is available online. In the dataset `GRONINGEN` (available on Canvas), $X(t)$ represents the cumulative production in Nm^3, scaled by a factor 10^{-15}, where t is the month after 1st of January, 1991, and $Y(t)$ the cumulative number of induced earthquakes (with magnitude greater than or equal to 1.5) up until that time.

(a) Estimate the autocorrelation of $X(t)$, does this estimate differ significantly from 0. Can you explain your results using your intuition?

(b) An empirical relation, proposed J. Hagoort[3], directly links the cumulative production since 1991 to the cumulative number of events since 1991:

$$\hat{Y}(t) = 0.027 \cdot X(t)^2 - 0.275 \cdot X(t)\,.$$

The value of $Y(t)$, while assuming this model, is provided as a variable in the `GRONINGEN` dataset. Create a column with the residuals of this model $Y(t) - \hat{Y}(t)$. Use an appropriate test to check whether this sequence of residuals is random. What is your final conclusion regarding the presented model? Do you think this model should be used for prediction in the future?

Question 4.10

Assume you have access to a time series dataset of two stocks $(S_A(t), S_B(t))$ over time in the period $[0, T]$. Which randomness test/statistic are appropriate to use in the different scenarios presented below.

(a) For S_A we would like to know whether the stock price is increasing over time.

(b) For S_B we would like to know whether the stock price was structurally higher in the first half of the period of interest ($[0, T/2]$).

[1] `https://www.nam.nl/feiten-en-cijfers/gaswinning.html`

[2] `https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus`

[3] `http://www.ondergroningen.nl/wp-content/uploads/2017/04/Empirical-Model-for-Induced-Earthquakes-in-the-Groningen-Gas-Field.pdf`

(c) For S_A we would like to check if the stock price at the end of December can be used to predict the stock price at the end of March in the year after. Hint: Could you think of two statistics, and which one do you prefer?

(d) Rumors are spreading regarding dependence between the two stock prices. Some brokers claim that the price of S_A can be used to predict the price of S_B one week later.

Answers

Answer to Question 4.1

(a) The ANOVA model is $y_{ij} = \mu + \alpha_i + e_{ij}$.

(b) The $\alpha_{suppA} = -4.1550$ and the p-value is 0.0747, which confirms that the effect is not significantly different from 0.

(c) The confidence intervals (see Slide 8 discussion hour W4) for μ, α_A, α_B are respectively $[103.38, 107.97]$, $[-4.372, 0.217]$, $[-0.217; 4.372]$.

(d) The t-test statistics is -1.83 with a corresponding p-value of 0.0747 which is the same value of the ANOVA analysis. In fact when only to groups are considered, the two tests are equivalent, since the two test statistics are the same. To show this consider the expression of the underlying F test of the ANOVA procedure: $F = \frac{MS_B}{MS_W} = \frac{SS_B}{\frac{SS}{n-2}}$. The numerator can be written as $SS_B = \sum_i n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$, with $i = 2$ (Hint: write $\bar{y}_{..} = \frac{n_1\bar{y}_{1.}+n_2\bar{y}_{2.}}{n}$ and multiply -and divide- each term by the total number of observations n. You will recognize the expression of the squared differences of the mean groups, averaged by the reciprocal sum of their observations). Similarly, the denominator can be re-written as $\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$, which equals the formula of s_p^2 of the t-test, which proves the equivalence of the two tests for this setting.

(e) The test statistics is 2.5916 with a corresponding p-value of 0.1074.

(f) The t-test with equal variances and the F test of ANOVA for two groups are equivalent. Normality seems not to be violated, as suggested by some explorative analysis, therefore the ANOVA outcome is more powerful than the Kruskal-Wallis test.

(g) The estimated differences and their corresponding confidence intervals are listed in the table below:

TRT	TRT	LS diff (S)	C.I.
A	B	-4.155	$[-9.03; 0.72]$
A	C	-8.605	$[-13.48; -3.73]$
B	C	-4.45	$[-9.32; 0.43]$

(h) The pairwise t-test differences and their corresponding confidence intervals are listed in the table below:

TRT	TRT	statistic (S)	C.I.
A	B	-1.83	$[-8.74; 0.43]$
A	C	-3.84	$[-13.15; -4.06]$
B	C	-1.61	$[-10.04; 1.14]$

The different amplitudes (wider for the LS estimates) have to do with the different computation of the standard error, that is derived from the combined data.

Answer to Question 4.4

(a) -

(b) The ICC equals 0.4967 which tells us that there is quite moderate agreement between the two treatments.

(c) The EBLUP for patient 9 equals -2.0071.

(d) We are not specifically interested in serial randomness thus perform a conditional randomness tests.

(e) The number of runs equals 12 corresponding to a p-value of 0.8141239, according to which we have no reason to doubt about the null-hypothesis of randomness of the EBLUP's sequence. Thus we can also evaluate whether the EBLUP's of a_i are generated independently from SAS. Furthermore, we can check the independence of the residuals using the conditional runs test. Hint: use

```
ods output SolutionR=EBLUP;
PROC MIXED DATA=LONGCOAG METHOD=TYPE3;
```

.

(f) With a similar motivation as in (d), from the conditional randomness test we get a p-value of 0.745 ($r = 21$) therefore we have no reason to doubt about the randomness of the residuals.

Answer to Question 4.6

(a) The ANOVA model is $y_{ij} = \mu + a_i + e_{ij}$, with μ the overall intercept, $a_i \sim N(0, \sigma_S^2)$ the random effect (intercept) for subject i (it expresses the between-subject variability) and $e_{ij} \sim N(0, \sigma_E^2)$ the residual term, accounting for the within-subject variability.

(b) The EBLUP of subject 1 is -0.01764 and the EBLUP of subject 2 is -0.5176.

(c) The conditional runs test report a p-value of 0.7725 which gives us no reason to doubt about the null hypothesis of randomness.

Answer to Question 4.9

(a) $\hat{r}_1 = 0.991$

(b) The (conditional) runs test is appropriate here since we are not interested in serial correlation, $r = 8$, $\frac{r-\mu(r)}{\sigma(r)} = -16.95709$ and $p < 0.001$.

This is an interesting result since the model seems to fit the data pretty well, see Figure 4. However, in Figure 5 it is shown that the residuals are clearly clustered over time. In practice this shows that the model proposed here overfits the data and will poorly perform for future prediction.

Answer to Question 4.10

(a) Unconditional runs test on $(x_{i+1} > x_i)$.

(b) Conditional runs test on ($x_i >$ median)

(c) Autocorrelation with a time-lag of 3 months.

(d) The cross correlation with a time-lag of 1 weeks (not discussed during lecture, but the generalization of the autocorrelation to two time series).

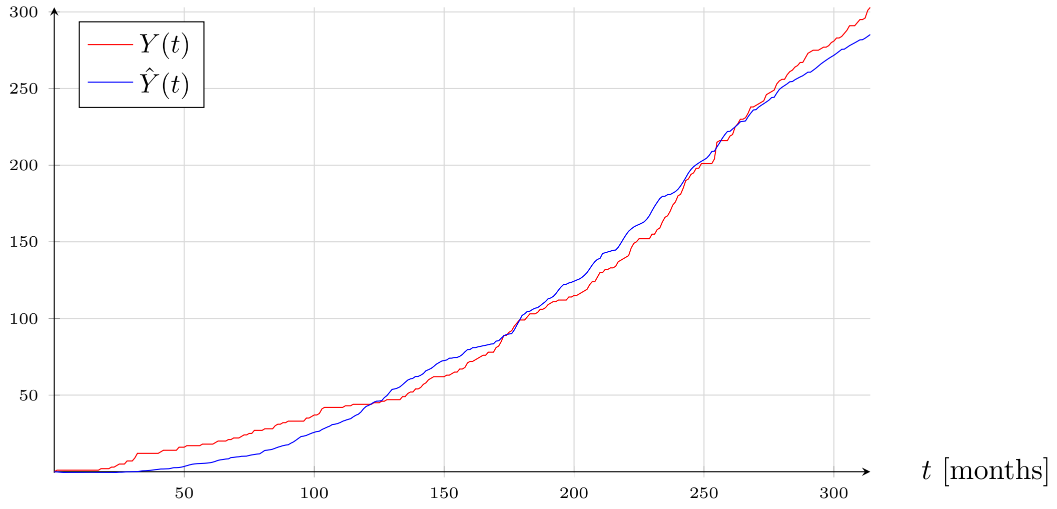


Figure 4: Accumulated quake count $Y(t)$ and its hypothesized model $\hat{Y}(t)$

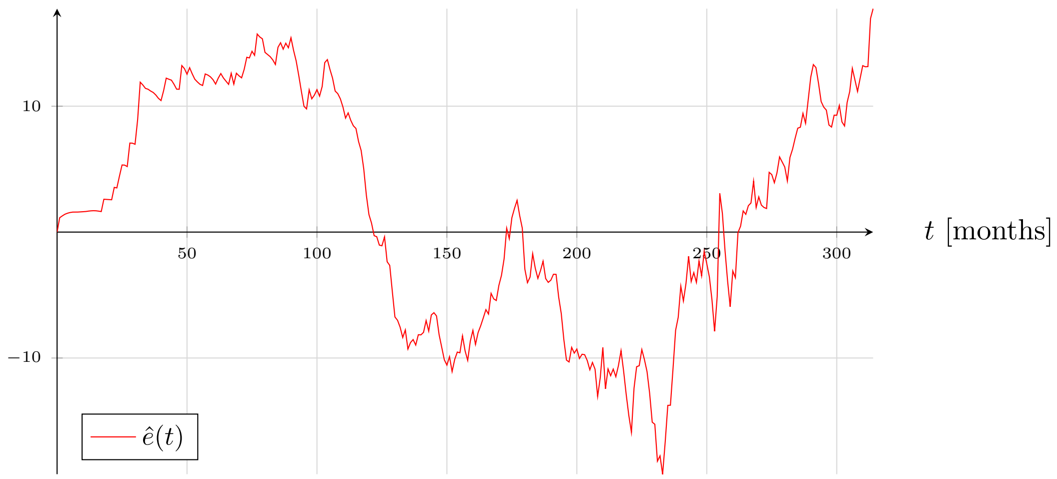


Figure 5: Residual $\hat{e}(t)$ versus t

Week 5:
Normality and outlier tests

Questions

For the questions of this week we will continue to use the IVF dataset and the RCT dataset (both available on Canvas).

Question 5.1

In Question 1.2 from Week 1, you assumed that the age of the mother (`AGEM`) is normally distributed. In this question you will verify these assumptions.

(a) Test if the age of the mother is normally distributed using a Shapiro-Wilk test. Report the p-value, the value of the test statistic and your conclusion.

(b) Test if the age of the mother is normally distributed using a Kolmogorov-Smirnov test. Report the p-value, the value of the test statistic and your conclusion.

(c) Test if the age of the mother is normally distributed using a Cramer-von Mises test. Report the p-value, the value of the test statistic and your conclusion.

(d) Test if the age of the mother is normally distributed using an Anderson-Darling test. Report the p-value, the value of the test statistic and your conclusion.

(e) Which of the tests you performed in (a)-(d) is the most appropriate to test if the age of the mother is normally distributed?

(f) Test if the age of the mother is normally distributed using a test based on skewness. Report the p-value, the value of the test statistic and your conclusion.

(g) Test if the age of the mother is normally distributed using a test based on kurtosis. Report the p-value, the value of the test statistic and your conclusion.

(h) What is the advantage of the tests you performed in (f) and (g) compared to the tests performed in (a)-(d)? In which situations would you prefer the kurtosis or skewness test?

Question 5.2

In Question 1.3 from week 1, you assumed that the birth (`BW`) is not normally distributed, but that using an appropriate Box-Cox transformation the data becomes (approximately) normally distributed. In this question you will verify these assumptions.

(a) Use the Grubbs test to remove all outliers of `BW` from the dataset, and then performing an adequate test for normality on the remaining data. Report the patient numbers of the outliers, and report p-value, the value of the test statistics and your conclusion for the normality test.

(b) Use Tukey's method to remove all outliers of `BW` from the dataset, and then performing an adequate test for normality on the remaining data. Report the patient numbers of the outliers, and report p-value, the value of the test statistics and your conclusion for the normality test.

(c) Answer (a) and (b) on the Box-Cox transformed birth weight, use $\lambda = 2$. Can you conclude that the assumptions in Question 1.6 from week 1 are valid.

Question 5.3

In Question 1.6 from week 1, you assumed that the gestational age (`GA`) is not normally distributed, but that the transformation `LOG(44 - GA)` should (approximately) have a normal distribution. In this question you will verify these assumptions.

(a) Use the Doornbos test to remove all outliers of `GA` from the dataset, and then performing an adequate test for normality on the remaining data. Report the patient numbers of the outliers, and report p-value, the value of the test statistics and your conclusion for the normality test.

(b) Use Hampel's rule to remove all outliers of `GA` from the dataset, and then performing an adequate test for normality on the remaining data. Report the patient numbers of the outliers, and report p-value, the value of the test statistics and your conclusion for the normality test.

(c) Answer (a) and (b) on the transformed variable `LOG(44 - GA)`. Can you conclude that the assumptions in Question 1.6 from week 1 are valid.

Question 5.4

In exercise 4.5 from last week you fitted an ANOVA model. In this exercise you will verify if the ANOVA assumptions of normality and homoscedasticity ("equal" variances) of the residuals hold. Fit the ANOVA model from exercise 4.5 and export the residuals.

(a) Check if the residuals you exported are normally distributed using an appropriate test. Report the p-value, the value of the test statistic and your conclusion.

(b) Use the residuals you exported to test if the they have equal variances. Report the null hypothesis, the test statistic, and the p-value and your conclusion.

(c) Based on your answers in (a) and (b), would you prefer to use an ANOVA model or use the Kruskal-Wallis test in exercise 4.5?

Question 5.5

In exercise 4.6 from last week you fitted an ANOVA model. In this exercise you will verify if the ANOVA assumptions of normality and homoscedasticity ("equal" variances) of the residuals hold. Fit the ANOVA model from exercise 4.6 and export the residuals.

(a) Check if the residuals you exported are normally distributed using an appropriate test. Report the p-value, the value of the test statistic and your conclusion.

(b) Based on your answers in (a) would you prefer to use an ANOVA model or use the Kruskal-Wallis test in exercise 4.6?

Question 5.6

In the table below are the measured values of 12 observations (the same as in Question 1.4 from week 1). Create a new dataset containing the table below, and use this dataset to answer the following questions.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
Value	25.0	27.4	17.1	22.1	20.8	21.3	22.5	29.2	27.9	25.7	24.7	18.8

(a) Use an appropriate test to check if the measured values are normally distributed. Report p-value, the value of the test statistics and your conclusion.

(b) Perform the Doornbos test, the Grubbs test, Hampel's rule, and Tukey's method to check if any of the measured values can be classified as an outlier.

Question 5.7

During the lecture you saw two equivalence relations for the leave-one-out mean $\bar{y}_k$ and leave-one-out variance s_k^2. In this exercise you will proof these relations.

(a) Proof the following equivalence relation for the leave-one-out mean

$$y_k - \bar{y}_k = \frac{n}{n-1}\left(y_k - \bar{y}\right),$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $\bar{y}_k = \sum_{i=1, i\neq k}^{n} y_i/(n-1)$.

(b) Proof the following equivalence relation for the leave-one-out variance

$$s_k^2 = \frac{n-1}{n-2}s^2 - \frac{n}{(n-1)(n-2)}\left(y_k - \bar{y}\right)^2,$$

where $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$ and $s_k^2 = \sum_{i=1, i\neq k}^{n}(y_i - \bar{y}_k)^2/(n-2)$, and $\bar{y}$ and $\bar{y}_k$ are as given in part (a).

Question 5.8

In this exercise you will create SAS macro's to accelerate the detection of outliers in the future. Each macro should work on a given variable on an arbitrary dataset.

(a) **[CSE]** Create a SAS macro that performs the two-sided Grubbs test to detect a single outlier.

```
%macro grubbs(dataset,var);
    --- YOUR CODE ---
%mend;
```

Use your macro to answer Question 5.2 part (a), does your answer agree with the result you obtained previously?

(b) **[CSE]** Create a SAS macro that performs the two-sided Doornbos test to detect a single outlier.

```
%macro doornbos(dataset,var);
    --- YOUR CODE ---
%mend;
```

Use your macro to answer Question 5.3 part (a), does your answer agree with the result you obtained previously?

(c) **[CSE]** Create a SAS macro that uses Tukey's method to find all outliers in a dataset.

```
%macro tukey(dataset,var);
    --- YOUR CODE ---
%mend;
```

Use your macro to answer Question 5.2 part (b), does your answer agree with the result you obtained previously?

(d) **[CSE]** Create a SAS macro that uses Hampel's rule to find all outliers in a dataset.

```
%macro hampel(dataset,var);
    --- YOUR CODE ---
%mend;
```

Use your macro to answer Question 5.3 part (b), does your answer agree with the result you obtained previously?

(e) **[CSE]** Adapt your macro in parts (a) and (b) to also perform both one-sided tests.

Question 5.9

In this question you will analyze the fraction of samples that Tukey's method classifies as an outlier.

(a) Suppose that data is sampled from a random variable X that has a standard normal distribution (i.e. $X \sim \mathrm{N}(0, 1)$). Compute the fraction of samples that will be classified as an outlier by Tukey's method.

(b) How does the fraction you obtained in (a) change if instead X had an exponential distribution with parameter 1 (i.e. $X \sim \mathrm{Exp}(1)$).

Question 5.10

In this question you will analyze the fraction of samples that Hampel's rule classifies as an outlier.

(a) Suppose that data is sampled from a random variable X that has a standard normal distribution (i.e. $X \sim \mathrm{N}(0, 1)$). Compute the fraction of samples that will be classified as an outlier by Hampel's rule. Hint: you may use that the median of a chi-squared distribution with 1 degree of freedom is approximately equal to $\sqrt{(1 - 2/9)^3}$.

(b) How does the fraction you obtained in (a) change if instead X had an exponential distribution with parameter 1 (i.e. $X \sim \mathrm{Exp}(1)$).

Answers

Answer to Question 5.1

	Test	Statistic	p-value
(a)	Shapiro-Wilk	0.9934	0.3284
(b)	Kolmogorov-Smirnov	0.0400	> 0.15
(c)	Cramer-von Mises	0.0593	> 0.25
(d)	Anderson-Darling	0.4219	> 0.25

(e) The data has many ties, therefore the Anderson-Darling test from (d) is the most appropriate.

(f) The skewness is estimated as -0.178018, which results in a normalized test statistic of -1.173 with a corresponding p-value of 0.241.

(g) The kurtosis is estimated as -0.1849355, which results in a normalized test statistic of -0.546 with a corresponding p-value of 0.585.

(h) The first four tests are omnibus tests, i.e. in case of rejection we don't know why normality is violated. If the model or test statistic you have in mind relies heavily on the skewness or kurtosis of your data having a specific form, then you could specifically test for this.

Answer to Question 5.2

(a) There are no outliers. Because there are some ties we use the Anderson-Darling test, the statistic is 1.5307 and the p-value is less than 0.0050.

(b) There are two outliers with `ID=98` and `ID=294`. Again there are some ties, so we use the Anderson-Darling test, the statistic is 1.3141 and the p-value is less than 0.0050.

(c) There are no outliers using either the Grubbs test or Tukey's method. Because there are some ties we use the Anderson-Darling test, the statistic is 0.3724 and the p-value is more than 0.25.

Answer to Question 5.3

(a) There is one outlier (with `ID=174`) which we exclude from the dataset. Next we execute another Doornbos test on this new data (thus performing the Holm-Bonferroni procedure), now no more outliers are detected since $p = 0.36836$ (otherwise we would continue this procedure). Because there are many ties we use the Anderson-Darling test, the statistic is 17.4270 and the p-value is less than 0.0050.

(b) The outliers are: `174`, `294`, `295`, `222`, `223`, `309`, `310`, `97`, `9`, `303`, `98`, `42`, `141`, `142`, `139`, `261`, `262`, `115`, `120`, `121`, `276`, `277`, `78`, and `79`. Again there are many ties, so we use the Anderson-Darling test, the statistic is 1.8951 and the p-value is less than 0.0050.

(c) There are no outliers using the Doornbos test. Because there are some ties we use the Anderson-Darling test, the statistic is 0.5166 and the p-value is more than 0.1971.

Using Hampel's rule we get four outliers: `47`, `174`, `166`, `210`. Because there are some ties we use the Anderson-Darling test, the statistic is 0.5932 and the p-value is more than 0.1257.

Answer to Question 5.4

(a) We use the Anderson-Darling test, the statistic is 0.9203 and the p-value is 0.0203.

(b) Using Levene's test we find the statistic $F = 0.15$ and p-value equal to 0.8636.

(c) Based on the answer in (a) we reject normality, so strictly speaking one cannot use ANOVA. However, we have a large sample size and based on (a) we see that our data is still relatively close to normal. Because ANOVA is quite robust against departures from normality, using ANOVA would still be a valid option in this case.

Answer to Question 5.5

(a) We use the Anderson-Darling test, the statistic is 1.8084 and the p-value is less than 0.0050.

(b) We reject normality, so the Kruskal-Wallis test is preferred.

Answer to Question 5.8

(a) The following SAS macro performs the two-sided Grubbs test.

```
%macro grubbs(dataset,var,id,alpha=0.05);
ods select none;
proc means data=&dataset mean var n;
    var &var;
    output out=out mean=mean var=var n=n;
run;

data outliers;
    set &dataset(keep=&id &var);
    if _n_=1 then set out;
    /* statistic */
    u = abs((&var - mean) / sqrt(var));
    /* critical value */
    t = quantile("t", &alpha / (2*n), n-2);
    c = (n-1) * sqrt(t**2 / (n * (t**2 + n - 2)));
    /* check if this is an outlier */
    if(u > c) then
        outlier = "yes";
    else
        outlier = "no";
    /* p-value */
    u_inv = u*sqrt((n-2)*n) / sqrt(1 - (u**2-(n-2))*n);
    p_value = min(2*n*(1-cdf("t", u_inv, n-2)), 1);
    keep &id p_value u c outlier;
run;

proc sort data=outliers;
    by descending u;
run;
ods select all;

title1 "Grubbs test";
title2 "Variable: &var";
proc print data=outliers(obs=1) label noobs;
    var &id p_value u c outlier;
    label &id="id"
          p_value="p-value"
          u="statistic (|u|)"
          c="critical value"
          outlier="is outlier?";
run;
title;
%mend;
```

(b) The following SAS macro performs the two-sided Doornbos test.

```
%macro doornbos(dataset,var,id,alpha=0.05);
ods select none;
proc means data=&dataset mean var n;
    var &var;
    output out=out mean=mean var=var n=n;
run;

data outliers;
    set &dataset(keep=&id &var);
    if _n_=1 then set out;
    /* leave-one-out variance */
    var_loo = ((n - 1) / (n - 2)) * var
      - (n / ((n - 1)*(n - 2)))*(&var - mean)**2;
    /* statistics */
    w = abs((&var - mean) / sqrt(var_loo * (n - 1) / n));
    /* critical value */
    c = quantile("t", 1 - &alpha / (2*n), n-2);
    /* check if this is an outlier */
    if(w > c) then
        outlier = "yes";
    else
        outlier = "no";
    /* p-value */
    p_value = min(2*n*(1-cdf("t", w, n-2)), 1);
    keep &id p_value w c outlier;
run;

proc sort data=outliers;
    by descending w;
run;
ods select all;

title1 "Doornbos test";
title2 "Variable: &var";
proc print data=outliers(obs=1) label noobs;
    var &id p_value w c outlier;
    label &id="id"
          p_value="p-value"
          w="statistic (|W|)"
          c="critical value"
          outlier="is outlier?";
run;
title;
%mend;
```

(c) The following SAS macro uses Tukey’s method to find all outliers.

```
%macro tukey(dataset,var,id);
ods select none;
proc means data=&dataset median p25 p75;
    var &var;
    output out=quartiles p25=p25 p75=p75;
run;

data tukey;
    set &dataset(keep=&id &var);
    if _n_=1 then set quartiles;
    iqr = p75 - p25;
    lower = p25 - 1.5*iqr;
    upper = p75 + 1.5*iqr;
    if &var >= lower and &var <= upper then delete;
run;
ods select all;

title1 "Outliers found using tukey's method.";
title2 "Variable: &var";
proc print data=tukey noobs label;
    var &id &var lower upper iqr;
    label lower="Lower limit"
          upper="Upper limit"
          iqr="Interquartile range (IQR)";
run;
title;
%mend;
```

(d) The following SAS macro uses Hampel's rule to find all outliers.

```
%macro hampel(dataset,var,id);
ods select none;
proc means data=&dataset median;
    var &var;
    output out=median median=median;
run;

data hampel;
    set &dataset(keep=&id &var);
    if _n_=1 then set median;
    abs_dev = abs(&var - median);
run;

proc means data=hampel median;
    var abs_dev;
    output out=abs_dev_median median=abs_dev_median;
run;

data hampel;
    set hampel;
    if _n_=1 then set abs_dev_median;
    abs_norm_val = abs_dev / abs_dev_median;
    if abs_norm_val <= 3.5 then delete;
run;

proc sort data=hampel;
    by descending abs_norm_val;
run;
ods select all;

title1 "Outliers found using hampel's rule.";
title2 "Variable: &var";
proc print data=hampel noobs label;
    var &id &var abs_norm_val;
    label abs_norm_val="absolute normalized value (z_k)";
run;
title;
%mend;
```

Answer to Question 5.9

(a) Let $X \sim \mathrm{N}(0,1)$, then the median of X is given by $m := \mathrm{median}(X) = 0$, and the quartiles are $z_{0.75} = -z_{0.25} = 0.6745$. Therefore, the theoretical IQR equals 1.3490. Using this, the fraction of samples that will be classified as an outlier is

$$\mathbb{P}(|X| < 1.5 \cdot 1.3490) = 2\mathbb{P}(X < -1.5 \cdot 1.3490) = 0.0430\,.$$

(b) Let $X \sim \mathrm{Exp}(1)$, then the median of X is given by $m := \mathrm{median}(X) = \log(2)$, the 25% quartile equals $q_{0.25} = 0.2877$, and the 75% quantiles equals $q_{0.75} = 1.3863$, hence the theoretical IQR equals 1.061. Using this, the fraction of samples that will be classified as an outlier is

$$\begin{aligned}\mathbb{P}(|X - \log(2)| < 1.5 \cdot 1.061) &= \mathbb{P}(X < \log(2) - 1.5 \cdot 1.061) + \mathbb{P}(X > \log(2) + 1.5 \cdot 1.061)\\ &= 0.0661\,.\end{aligned}$$

Answer to Question 5.10

(a) Let $X \sim \mathrm{N}(0,1)$, then the median of X is given by $m := \mathrm{median}(X) = 0$. Using that $(X-m)^2 = X^2 \sim \chi^2_1$, we find that the median absolute deviation is given by

$$\begin{aligned}m_d &:= \mathrm{median}(|X-m|) = \mathrm{median}(|X|) = \mathrm{median}\Big(\sqrt{X^2}\Big) = \sqrt{\mathrm{median}(X^2)}\\ &\approx \sqrt{(1-2/9)^3} \approx 0.6859\,.\end{aligned}$$

Then the fraction that will be classified as an outlier is given by

$$\begin{aligned}\mathbb{P}\Big(\frac{|X-m|}{m_d} > 3.5\Big) &= \mathbb{P}\Big(|X| > 3.5\, m_d\Big) = \mathbb{P}\Big(|X| > 3.5\, m_d\Big)\\ &= \mathbb{P}\Big(X > 3.5 \cdot 0.6859\Big) + \mathbb{P}\Big(X < 3.5 \cdot 0.6859\Big) \approx 0.0164\,.\end{aligned}$$

We conclude that if the data is sampled from a standard normal random variable, about 1.64% of these samples will be classified as an outlier.

(b) Let $X \sim \mathrm{Exp}(1)$, then the median of X is given by $m := \mathrm{median}(X) = \log(2)$. To obtain the median absolute deviation we find the value for m_d such that $\mathbb{P}(|X - m| > m_d) = 0.5$.

$$\begin{aligned}0.5 &= \mathbb{P}\Big(|X - \log(2)| > m_d\Big)\\ &= \mathbb{P}\Big(X - \log(2) > m_d\Big) + \mathbb{P}\Big(-(X - \log(2)) > m_d\Big)\\ &= \mathbb{P}\Big(X > m_d + \log(2)\Big) + \mathbb{P}\Big(X < -m_d + \log(2)\Big)\\ &= 1 - \frac{\mathrm{e}^{-m_d} + \mathrm{e}^{m_d}}{2} = 1 - \sinh(m_d)\,.\end{aligned}$$

Therefore, the value of m_d is

$$m_d = \sinh^{-1}(1/2) \approx 0.4812\,.$$

Then the fraction that will be classified as an outlier is given by

$$\begin{aligned}\mathbb{P}\Big(|X-m| > 3.5\,m_d\Big) &= \mathbb{P}\Big(X-m > 3.5\,m_d\Big) + \mathbb{P}\Big(-(X-m) > 3.5\,m_d\Big)\\ &= \mathbb{P}\Big(X > 3.5\,m_d + m\Big) + \mathbb{P}\Big(X < -3.5\,m_d + m\Big)\\ &= \mathbb{P}\Big(X > 3.5\,m_d + m\Big) + 0\\ &= \exp\Big(-3.5\cdot 0.4812 - \log(2)\Big) \approx 0.0928\,.\end{aligned}$$

We conclude that if the data is sampled from an exponential random variable, about 9.28% of these samples will be classified as an outlier.

Week 6: Multiple testing and two-way ANOVA

Questions

For the questions of this week we will continue to use the IVF dataset and the RCT dataset (both available on Canvas).

Question 6.1

Let us start this week by reconsidering some of the questions before using the knowledge gained in the lecture covering the issues of multiple testing.

(a) Redo the Chi-squared tests in Question 2.7 (a) and apply an appropriate correction to test simultaneously the H$_0$: $p_{\texttt{FIS,TRT=i}} = p_{\texttt{FIS,TRT=j}}$, for $i, j \in \{0, 1, 2\}$ and $i \neq j$. Does this correction change your conclusion?

(b) Use Tables 1 and 2 presented in Question 3.11 to test for a treatment effect of the two drugs using Bonferroni's correction on the results of an appropriate test. Report the null-hypothesis, the test results, and the conclusion of the hypothesis. How does this correction change your conclusions? Hint: you should not use SAS for this question.

(c) Use the full dataset (including `suppC`) from Question 4.1. Fit an one-way ANOVA model to test whether the effect of the new dietary supplement is significantly different than one of the two existing ones. That is, test H$_0$: $\mu_A = \mu_C$ and H$_0$: $\mu_B = \mu_C$ simultaneously at a FWER of 5%.

Question 6.2

In this exercise you need the IVF dataset. You can assume that the infant motor profile (`IMP`) can be modeled with an ANOVA model. We will investigate the treatment effect (`TRT`) taking into account the subject to subject variability.

(a) Write down the introduced ANOVA model mathematically, so you should specify

$$y_{ij} = \dots ,$$

and explain all the parameters used.

(b) Fit the ANOVA model, and report the ICC for the subject to subject variance component. What does this ICC tell you?

(c) Formulate the traditional null and alternative hypotheses for the effect of treatment. Report the value of the test statistic, the p-value and your conclusion.

(d) We would like to compare all pairs of treatment and test differences simultaneously at a FWER of $\alpha = 0.05$. Use an appropriate multiple testing approach and report the corrected p-values and conclusions.

(e) Imagine the hypothetical setting were we performed Scheffé's method method in (d) and one confidence interval for the difference of two means did *not* contain zero. What do we know about this CI if we would have applied Tukey's studentized range test?

(f) Imagine the hypothetical setting were we performed Tukey's studentized range test in (d) and one confidence interval for the difference of two means did *not* contain zero. What do we know about this CI if we would have applied Scheffé's method method?

(g) Can you answer question (c) using a non parametric test? If so, report the value of the test statistic, the p-value and your conclusion.

Question 6.3

In this exercise you need the RCT dataset, use only the data of the first day (`TIME=1`). You can assume that the hemoglobin values (`RESP`) can be modeled with an ANOVA model. We will investigate the effect of the medical center `CENTER` in the RCT dataset taking into account the effect of `TRT`.

(a) Write down the introduced ANOVA model mathematically, so you should specify

$$y_{ij} = \dots ,$$

and explain all the parameters used.

(b) Fit the ANOVA model, and report the estimates for the effects of `CENTER` and `TRT`.

(c) Formulate the traditional null and alternative hypotheses for the effect of medical center. Report the value of the test statistic, the p-value, and your conclusion.

(d) Fit a one-way ANOVA model with treatment as a fixed effect. Compare the estimates of the treatment effect in this one-way ANOVA model with the estimates you obtained from your two-way ANOVA model fitted in Question (b). Can you explain the results?

(e) Use Tukey's studentized range test and test for contrast of the means of medical centers. Could you intuitively explain the differences found?

(f) Use Dunnett's many-to-one test verify for each additional treatment (`TRT = 1,2,3`) whether they result in significantly different `RESP` outcomes. Report the null hypotheses and the corrected p-values.

(g) Can you answer Question (c) using a non-parametric test? If so, report the value of the test statistic, the p-value and your conclusion.

Question 6.4

In this question we will discuss several scenarios in which one has multiple H_0's to test. Argue which multiple testing methods could be used and which is most appropriate in your opinion.

(a) You did 3 Wilcoxon-rank-sum tests to compare 3 groups.

(b) You performed an ANOVA analysis to check if 4 groups are comparable. Only if you find evidence for a deviation you are interested in which group deviates.

(c) You want to fit an ANOVA model to 8 groups, where you expect that only one group level is different from the rest. This deviation is practically relevant if the effect size is greater than two.

(d) You performed several McNemar test's to compare 5 groups to one control group. You are mainly worrying about correctness of the results for the groups that in reality differ the most from the control group.

(e) You fit an ANOVA model to compare 8 groups with a control group and you expect that only one group level is different from the rest.

(f) You fit an ANOVA model to compare several groups and want to be as conservative as possible in rejecting the pair-wise H_0's.

Question 6.5

Consider the final measurements in the RCT dataset (`TIME = 6`). How should the total number of participant be divided over the 4 groups for a most efficient design according to Dunnett's theory.

Answers

Answer to Question 6.1

(a) We can use Bonferroni's or Sidak's correction here.
Using Bonferroni correction, the adjusted α level is: $\frac{0.05}{3} \approx 0.01667$. Using this correction, no null hypothesis can be rejected.
Using Sidak's correction, the adjusted α level is: $1 - (1 - 0.05)^{\frac{1}{3}} \approx 0.01695$. Using this correction, no null hypothesis can be rejected.

(b) The null hypothesis in both cases is: $H_0 : p_{high,pre} = p_{high,post}$. Even without multiplicity correction, the null was not rejected. Therefore, correcting the p-values will not change the conclusions, as this only increases p-values. For drug B, the p value is $p = 0.049$. Therefore, before the correction, we can reject H_0 since $p < 0.05$. However, after the correction, we cannot, since $p \not< 0.025$.

(c) Since we compare supplement effects against a control supplement (as opposed to comparing all pairs of supplements) we use Dunnett's correction. We obtain a p value of $p = 0.0016$ for $H_0 : \mu_A = \mu_C$ and $p = 0.1295$ for $H_0 : \mu_B = \mu_C$. Therefore, we conclude that only supplement C is significantly different from supplement A.

Answer to Question 6.2

(a) The ANOVA model is given by:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where y_{ij} is the infant motor profile of subject i with treatment j. On the right side, μ is the overall mean, $\alpha_i \sim \mathrm{N}(0, \sigma_G)$ is the random effect of subject i, β_j is the fixed effect of treatment j and $e_{ij} \sim \mathrm{N}(0, \sigma_E)$ is the residual.

(b) The type-3 estimate of σ_G is negative, and we therefore set it to 0. Consequently, the ICC is equal to 0. This tells us that the variance between subjects is negligible compared to the variance within subjects.

(c) The competing hypotheses for the fixed effect of treatment are: $H_0 : \forall_j : \beta_j = 0$ versus $H_1 : \exists_j : \beta_j \neq 0$. The test statistic is $F = 5.47$ with p value $p <= 0.0045$. We therefore reject the null hypothesis in favor of the alternative hypothesis; at least one treatment has a fixed effect unequal to 0.

(d) Since we compare all pairs of treatment effects (as opposed to testing treatment effects against a control treatment), we can use Tukey's studentized range test:

- Treatment 0 vs Treatment 1: $p = 0.9999$;
- Treatment 0 vs Treatment 2: $p = 0.0092$;
- Treatment 1 vs Treatment 2: $p = 0.0270$.

We prefer Tukey's method to e.g. Bonferroni's or Sidak's method, as the correction takes the dependence of the estimates into account.

(e) Since Scheffé's method is more conservative than Tukey's studentized range test, it's confidence intervals are wider. Therefore, if a confidence interval does not contain 0 with Scheffé's method, then the corresponding interval with Tukey's method surely does not contain 0 as well.

(f) We know the corresponding interval using Scheffé's method will be wider, and therefore, we cannot be certain that we would reject H_0 using Scheffé method too.

(g) Be careful not to turn to Kruskal-Wallis' test here. Krusal-Wallis can be interpreted as a non-parametric one-way ANOVA with a fixed effect. However, in this setting, we have both a fixed and a random effect, and Krusal-Wallis is not applicable. In week 7, you will learn about Friedmann's test, which is applicable in this setting.

Answer to Question 6.3

(a) The ANOVA model is given by:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where y_{ij} is the hemoglobin value of center i using treatment j. On the right side, μ is the overall mean, α_i is the fixed effect of center i, β_j is the fixed effect of treatment j and $e_{ij} \sim \mathrm{N}(0, \sigma)$ is the residual.

(b) The estimates for the fixed effects are given in the table below:

Fixed effect	*Estimate*	*Fixed effect*	*Estimate*
α_1	-0.1610	μ	7.6754
α_2	0.09697	β_0	-0.1456
α_3	0.1733	β_1	0.008808
α_4	-0.05662	β_2	-0.06295
α_5	0.3876	β_3	0
α_7	0		

(c) The competing hypotheses for the fixed effect of center are: $H_0 : \forall_i : \alpha_i = 0$ versus $H_1 : \exists_i : \alpha_i \neq 0$. The test statistic is $F = 6.31$ with p value $p < 0.0001$. We therefore reject the null hypothesis in favor of the alternative hypothesis; at least one center has a fixed effect unequal to 0.

(d) The estimates for the fixed treatment effect are (slightly) different in the two models. Since the two-way ANOVA also fits the fixed effect of center (and explains some variance through this effect) the effect of treatment is changed.

(e) Using Tukey's test, we can only reject $H_0 : \mu_1 = \mu_5$ ($T = -4.97$, $p < 0.0001$) and $H_0 : \mu_1 = \mu_3$ ($T = -3.28$, $p = 0.0139$). Medical center 1 could have been an academic hospital. If this is not the case, then it seems that hospital 1 is the place to go since the blood loss in this hospital is significantly lower.

(f) The three sets of hypotheses are, for $i = 1, 2, 3$: $H_0 : \beta_i - \beta_0 = 0$ vs $H_1 : \beta_i - \beta_0 \neq 0$. The corrected p-values are $p = 0.2503$, $p = 0.7070$ and $p = 0.2935$ respectively.

(g) Compare this question with Question 6.2(g). Again, we cannot use Krusal-Wallis here, as we have not one, but two fixed effects. Unfortunately, Friedmann's test (which you will learn about in week 7) is not applicable either.

Answer to Question 6.4

(a) A Bonferroni or Sidak experiment wise correction could be applied.

(b) Least square difference.

(c) Tukey's studentized range test > Bonferroni Experiment-wise correction > F-test.

(d) Experiment wise Bonferroni-Holm correction.

(e) Dunnett's many to one adjustment.

(f) Scheffé's method.

Answer to Question 6.5

Since $n = 716$ and $m = 4$ we find

$$\begin{aligned} n_2 &= \frac{716}{3+\sqrt{3}} = 151.3 \\ n_1 &= 151.3\sqrt{3} = 262.03. \end{aligned}$$

Since this optimal value does not equal integer values we should compare the total variance of the pair-wise comparisons (see slide 39 W6-1):

n_2	n_1	Total variance
151	263	$\sigma_E^2 \cdot 0.031274$
152	260	$\sigma_E^2 \cdot 0.031275$

Table 4: Dunnett's most efficient design

Week 7: Higher way ANOVA

Questions

For the questions of this week we will continue to use the IVF dataset and the RCT dataset (both available on Canvas).

Question 7.1

In this question we will analyze a subset of the RCT dataset, limit your analysis to the observations where `ID` ≤ 75.

(a) Fit a mixed 2-way ANOVA to test for an effect of treatment (`TRT`) taking into account the possible variability introduced by subject (`ID`). Compute the confidence interval of $\overline{Y}_{\cdot 3}$ using Satterthwaite's approach.

(b) Compare the results in (a) with the fit of a 1-way ANOVA ignoring the possible variability introduced by subject. Discuss your findings.

We continue this question by considering a different model where we focus on a possible effect of `TIME`.

(c) Fit the 2-way ANOVA model involving `TIME` and `ID` to test for an effect of the `TIME` variable.

(d) Alternatively, perform Friedman's test to test for an effect of `TIME` taking into account the patient to patient variability.

(e) Would you prefer the test performed in question (c) or the one performed in question (d). Motivate your answer.

Question 7.2

In this exercise you need the IVF dataset. You can assume that the infant motor profile (`IMP`), that is observed at three different dates, can be modeled with a mixed effects ANOVA model. We will consider the fixed effects treatment (`TRT`) and period (`PER`), their interaction, and the random effect `ID`.

(a) Write down the introduced ANOVA model mathematically, so you should specify

$$Y_{ijk} = \dots ,$$

and explain all the parameters used. Hint: patients are nested within treatment.

(b) Formulate the traditional null and alternative hypotheses for the effect of treatment. Report the value of the test statistic, the p-value and the conclusion.

(c) Formulate the traditional null and alternative hypotheses for the interaction effect of treatment and period. Report the value of the test statistic, the p-value and the conclusion.

(d) Fit the ANOVA model, and report the variance component estimates, the ICC, and the EBLUP for patients 9, 11, and 33.

(e) Validate the ANOVA assumptions on the residuals. That is, check if the distributions of the residuals are normal and homoscedastic (i.e. "equal" in variance).

(f) Compute the treatment effects of treatment-M (`TRT = 1`) and treatment-C (`TRT = 2`) with respect to the control group (`TRT = 0`) with the 95% confidence intervals? Hint: do not forget to use the appropriate degrees of freedom.

(g) Answer the question in (f) for only the measurements after 10 months (`PER` $=$ 10). Explain why you cannot simply remove the data of the other periods and apply your answer in (f) on the reduced dataset.

Question 7.3

In this exercise you need the RCT dataset. You can assume that the hemoglobin values (`RESP`), that is observed at six different days, can be modeled with a mixed effects ANOVA model. We will consider the fixed effects treatment (`TRT`), day (`TIME`), the hospital (`CENTER`), all possible interactions, and the random effects `ID`.

(a) Write down the introduced ANOVA model mathematically, so you should specify

$$Y_{ijkl} = \dots ,$$

and explain all the parameters used. Hint: don't forget the nested structure of patients.

(b) Formulate the traditional null and alternative hypotheses for the effects of treatment, center, and their interaction. For each of these, report the value of the test statistic, the p-value, and your conclusion. What do these conclusions mean in practice?

(c) Compute the estimates of the means over time, for the 'Device' treatment (`TRT = 2`) and centers 1, 2, and 3.

(d) Report the variance component estimates, the ICC, and the EBLUP for patients 1, 2, and 3.

(e) Compute the treatment effects of the 'Filter' treatment (`TRT = 1`), the 'Device' treatment (`TRT = 2`), and the 'Device+Filter' treatment (`TRT = 3`) with respect to the control group (`TRT = 0`) with the 95% confidence intervals? Hint: do not forget to use the appropriate degrees of freedom.

(f) Answer the question in (e) for time times 2, 4, and 6 separately (i.e. give three separate answers for `TIME` equal to `2`, `4`, and `6`). Explain why you cannot simply remove the data of the other periods and apply your answer in (e) on the reduced dataset.

From the real dataset a sensor variable, which indicates which device has been used for the measurement, was omitted. Use the SAS code below to recover this column.

```
data RCT_DATASET_NEW;
    set RCT_DATASET;
    where CENTER=1;
    if TIME < 4 then SENSOR = 1;
    else SENSOR = 2;
    keep ID RESP SENSOR;
run;
```

We continue with a sub-analysis of the first medical center `CENTER = 1` and will ignore the `TIME` and `TRT` variables. **Don't forget to consider, when it's possible, the interaction terms.** *The following computations will take a while (up to 30 minutes).*

(g) Assume that the sensors are disposable tests that can only be used for 3 measurements on one specific subject. Fit an ANOVA model to investigate whether the sensors introduce variability? If so, report the test statistic and the p-value.

(h) Assume that the medical center bought two sustainable sensors (1 and 2) to get more reliable measurements. Fit a FULL ANOVA model (so with interaction `ID*SENSOR`) to investigate whether the sensors introduce variability and discuss the new results.

(i) Finally, consider the setting with sustainable sensors. How would you change your model from the previous question to investigate whether *these* two sensors give rise to significantly different measurements?

Question 7.4

Consider the model discussed in Question 7.3, but now consider `CENTER` as a random effect and ignore the factor `ID` (and the interactions with `ID`).

(a) What could be a reason to include the medical center as a random effect?

(b) Derive the mean squares variance components table, as discussed during the final lecture for this full model.

(c) Provide an unbiased estimates for $\sigma^2_{\texttt{CENTER*TIME}}$.

(d) Provide a test statistic to test H$_0$: $\sigma^2_{\texttt{CENTER*TIME}} = 0$.

Question 7.5

Consider the 'volume of glands' interaction model example presented during the lecture. Derive the distribution of

(a) $\dfrac{MS_S}{MS_{SO}}$, if $\sigma^2_S = 0$.

(b) $\dfrac{MS_S}{MS_E}$, if $\sigma^2_S = 0$.

(c) $\dfrac{MS_O}{MS_{SO}}$, if $\beta_1 = ... = \beta_J = 0$.

(d) $\dfrac{MS_{SO}}{MS_E}$, if $\sigma^2_{SO} = 0$.

Question 7.6

Illustrate how Satterthwaite's approach could be used to derive an approximate $1-\alpha$ % confidence interval for the parameter $\theta = \mathbb{E}[T]$ of interest.

Answers

Answer to Question 7.1

(a) In this exercises the default approach of SAS does not give very different results than the Satterthwaite's approach due to the large DF. For smaller DF as was the case in the lecture (oncologists) example the difference will be significant.

Method	DF	LCL	UCL
SAS default	70	5.8550	6.3784
Satterthwaite's	70.8	5.8551	6.3783

Table 5: Confidence interval $\bar{Y}_{.3}$

(b) The results of the one-way ANOVA are $F = 3.59$ and $p = 0.0139$, while te results for treatment in the two-way ANOVA where $F = 2.07$ and $p = 0.1125$. By ignoring the variability introduced by subject we would conclude a significant effect of TRT. If one would investigate normality of the residuals and equality of variance within the TRT groups it becomes clear that both models do not meet the ANOVA assumptions.

(c) $F = 102.08$ and $p < .0001$.

(d) $Q = 197.7761$ and $p < .0001$.

(e) In this case the conclusion of the two tests are both very significant. However, normality of the residuals is clearly rejected (Shapiro-Wilk, $p = 0.0008$) and equality of variances within the TIME groups is rejected (Bartlett's test, $p = 0.0005$). In practice we would only rely on the result of the Friedman's test.

Answer to Question 7.2

(a)

$$Y_{ijk} = \mu + a_i + b_j + c_{ij} + d_{k(i)} + e_{ijk},$$

where

- Y_{ijk} is a future IMP value of an observation k with TRT=i at PER=j.
- μ equals the overall mean.
- a_i is the fixed effect of treatment TRT, such that $\sum_{i=0}^{2} a_i = 0$.
- b_j is the fixed effect of the age of the child PER, such that $b_4 + b_{10} + b_{18} = 0$.
- c_{ij} is the (fixed) interaction effect between treatment and the age of the child (TRT*PER), such that $\sum_{i=0}^{2}(c_{i4} + c_{i10} + c_{i18}) = 0$.
- $d_{k(i)}$ is the random (nested) effect of patient ID, within treatment TRT, such that $d_{k(i)} \sim \mathcal{N}(0, \sigma^2_{\mathtt{ID}})$.
- e_{ijk} is the random residual, such that $e_{ijk} \sim \mathcal{N}(0, \sigma^2_E)$.

(b) $H_0 : a_0 = a_1 = a_2 = 0$ versus $H_1 : \exists i \in \{0, 1, 2\} : a_i \neq 0$. The F-value equals 5.91 (Table *"Type 3 Tests of Fixed Effects"*) which gives rise to a $p = 0.0031$ based on which we reject H_0. We conclude that treatment seems to have an effect on the IMP score, on average TRT $= 1$ is expected to increase the IMP with $0.97 - 0.57 = 0.40$ compared to no treatment ($\texttt{TRT} = 0$), while $\texttt{TRT} = 2$ is expected to decrease the IMP with $0.57 - 0 = 0.57$ compared to no treatment.

(c) $H_0 : \forall i, j c_{ij} = 0$ versus $H_1 : \exists i, j : c_{ij} \neq 0$ for $i \in \{0, 1, 2\}$ and $j \in \{4, 10, 18\}$. The F-value equals 1.84 (Table *"Type 3 Tests of Fixed Effects"*) which gives rise to a $p = 0.1205$ based on which we cannot reject H_0 at a significance level of $\alpha = 0.05$. Based on this dataset we cannot conclude that the age of the child has an effect on the performance of the treatment.

(d) The estimates of the variance components equal $\hat{\sigma^2_{\texttt{ID}}} = 1.5449$ and $\hat{\sigma^2_E} = 7.0802$. Therefore the estimate for the ICC equals $\frac{1.5449}{1.5449+7.0802} = 0.179$ representing the correlation coefficient between two measurements on the same subject. The $EBLUP$s equal -3.2104, 0.08713 and -2.3317 for patients 9, 11 and 33, respectively.

(e) Based on the histogram of the residuals normality of the residuals seems to be a reasonable assumption. A formal Anderson-Darling (ties in residual data) results in a p-value of 0.0363 and would reject normality at $alpha = 0.05$, this was to be expected for such a large sample $n = 728$. The residuals seem to be approximatly normal distributed, so the Chi-square distributional assumptions on the MS will also be approximately correct. To use the results of our analysis we should also test for homogeneity of variance, we compare all groups based on the fixed effects (TRT*PER), Bartlett's test results in $p = 0.13$ and fails to reject homoscedasticity. The model seems to be valid to be used in practice.

(f) You should use Satterthwaite's approach for the degrees of freedom, please notice that the effect of using DDFM=SAT is very small in the case of a large dataset. The CI of the effect of the treatment-M ($\texttt{TRT} = 1$) compared to the control group equals $(-0.6650, 0.6614)$[4] and for treatment-C $(-1.4565, -0.3190)$.

(g) At PER=10 The CI of the effect of the treatment-M ($\texttt{TRT} = 1$) compared to the control group equals $(-1.5193, 0.4396)$ and for treatment-C $(-1.4761, 0.2027)$. Data from the other periods is used to estimate the variance components which influence the width of the CI.

Answer to Question 7.3

(a) $y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + a_{l(ik)} + e_{ijkl}$ where

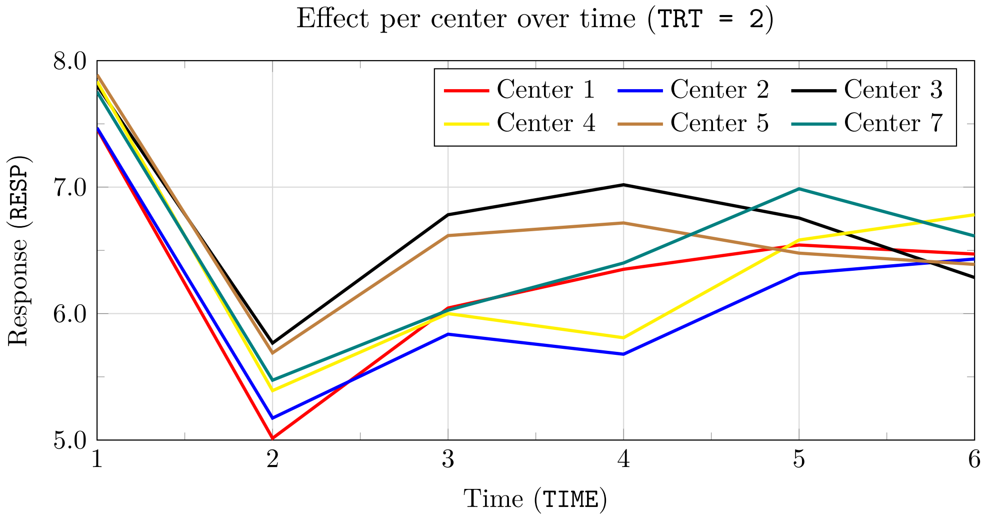
- y_{ijkl} is the response value of a unit l within CENTER=k treated with TRT=i at time point T=j.
- μ equals the overall mean.
- α_i is the fixed effect of treatment TRT, such that $\sum_{i=0}^4 \alpha_i = 0$.
- β_j is the fixed effect of the time TIME, such that $\sum_{j=1}^7 \beta_j = 0$.
- γ_k is the fixed effect of the center CENTER, such that $\sum_{k=1}^7 \gamma_k = 0$.
- $\alpha\beta_{ij}$ is the (fixed) interaction effect between treatment and time (TRT*TIME), such that $\sum_i \sum_j \alpha\beta_{ij} = 0$.
- $\alpha\gamma_{ik}$ is the (fixed) interaction effect between treatment and center (TRT*CENTER), such that $\sum_i \sum_k \alpha\gamma_{ik} = 0$.

[4] $(-0.6650, 0.6614) = (-0.6614 - 2 \cdot 0.001770, 0.6650 - 2 \cdot 0.001770)$ by subtracting the estimate once I get a CI centered around zero by subtracting again I get the CI for $\mu_1 - \mu_0$ instead of the CI of $\mu_1 - \mu_0$

- $\beta\gamma_{jk}$ is the (fixed) interaction effect between time and center (TIME*CENTER), such that $\sum_j \sum_k \beta\gamma_{jk} = 0$.
- $\alpha\beta\gamma_{ijk}$ is the (fixed) interaction effect between treatment, time and center (TRT*TIME*CENTER), such that $\sum_i \sum_j \sum_k \alpha\beta\gamma_{ijk} = 0$.
- $a_{l(ik)}$ is the random (nested) effect of patient ID, within treatment and center TRT*CENTER, such that $a_{l(ik)} \sim \mathcal{N}(0, \sigma^2_{\texttt{ID}})$.
- e_{ijkl} is the random residual, such that $e_{ijkl} \sim \mathcal{N}(0, \sigma^2_E)$.

(b) $H_0 : \alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = 0$ versus $H_1 : \exists i \in \{0, 1, 2, 3\} : \alpha_i \neq 0$. The F-value equals 6.79 which gives rise to a $p = 0.0002$ based on which we reject H_0. We conclude that treatment seems to have an effect on the response variable. For center it holds that: $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = 0$ versus $H_1 : \exists k \in \{1, 2, 3, 4, 5, 6\} : \gamma_k \neq 0$. The F-value equals 14.99 which gives rise to a $p < 0.0001$ based on which we reject H_0. The average profile of the response variable differs from center to center. For the interaction term we have: $H_0 : \alpha\gamma_{ik} = 0, \forall i, k$ vs $H_a : \exists (i, k) \in \{0, 1, 2, 3\} \otimes \{1, 2, 3, 4, 5, 7\} : \alpha\gamma_{ik} \neq 0$. The F-value equals 0.60 which gives rise to a $p = 0.8729$ based on which we can't reject H_0. This means that there is no difference in the response variable between people who got the same treatment in different medical centers.

(c) We give as an example the average profile of people treated with TRT=2 in medical center 1 at the first time point. $y_{211} = 6.4451 + 0.1683 - 0.1645 + 0.02184 + 1.1016 + 0.03840 + 0.06280 - 0.2129 = 7.4606$. For the complete summary of the average profiles see the image below.



(d) The estimates of the variance components equal $\hat{\sigma^2_{\texttt{ID}}} = 0.3299$ and $\hat{\sigma^2_E} = 0.3882$. Therefore the estimate for the ICC equals $\frac{0.3299}{0.3299+0.3882} = 0.4594$. The $EBLUP$s equal -0.3783, -0.02829 and -0.1926 for patients 1, 2 and 3, respectively.

(e) You should use Satterthwaite's approach for the degrees of freedom. The CI for the effect of treatment 1 (TRT $= 1$) compared to the control group equals $(-0.1032, 0.2461)$, the effect of treatment 2 (TRT $= 2$) compared to the control group equals $(0.1209, 0.4590)$ and the effect of treatment 3 (TRT $= 3$) compared to the control group equals $(0.1589, 0.5174)$.

(f) At TIME=2 the CI of the effect of treatment 1 (TRT $= 1$) compared to the control group equals $(-0.2879, 0.1830)$, for treatment 2 equals $(-0.1572, 0.2982)$ and for treatment 3 is $(-0.1026, 0.3803)$. A similar reasoning applies for the other two time periods. Please note that data from the other periods is used to estimate the variance components, and therefore influence the width of the CI: separate time analysis would change the results.

(g) We fit an ANOVA model with a nested factor ($SENSOR(ID)$). The variance component related to the sensor is negatively estimated (-0.1929), the F-test statistics equals 0.50 and the p-value is 1.0000, which is expected based on the negative estimation. In these cases it is recommended to set the variance component to zero.

(h) We fit an ANOVA model with $SENSOR$ as random effect. The variance component related to $SENSOR$ is 0.003670 with a related statistic of 0.72 which gives a p-value of 0.4724. Therefore we can claim that there is no variability associated to sensor.

(i) We fit an ANOVA model with $SENSOR$ as fixed effect. The p-value of the F test is 0.0016. We can therefore claim that there is a difference in the average behaviour of the two sensors.

Answer to Question 7.4

(a) In some studies we arbitrarily (random) select only a few medical centers, for convenience, although our population of interest are all medical centers.

(b) The mean square variance component table is as follows:

	σ_{TRT}	σ_T	σ_{CTR}	σ_{TRT*T}	σ_{T*CTR}	$\sigma_{TRT*CTR}$	$\sigma_{TRT*CTR*T}$	σ_E
MS_{TRT}	0			0		JL	L	1
MS_T		0		0	IL		L	1
MS_{CTR}			IJL		IL	JL	L	1
MS_{TRT*T}				0			L	1
MS_{T*CTR}					IL		L	1
$MS_{TRT*CTR}$						JL	L	1
$MS_{TRT*CTR*T}$							L	1
MS_E								1

(c) An estimate for $\sigma^2_{\texttt{CENTER*TIME}}$ is $(MS_{T*CTR} - MS_{TRT*CTR*T})/IL$.

(d) A test statistic to test H$_0$: $\sigma^2_{\texttt{CENTER*TIME}} = 0$ is $\frac{MS_{T*CTR}}{MS_{TRT*CTR*T}}$.

Answer to Question 7.5

(a)

$$\frac{MS_S}{MS_{SO}} = \frac{\frac{(I-1)MS_S}{I-1}}{\frac{(I-1)(J-1)MS_{SO}}{(I-1)(J-1)}}$$

$$= \frac{\frac{(I-1)MS_S}{I-1}}{\frac{(I-1)(J-1)MS_{SO}}{(I-1)(J-1)}} \cdot \frac{\sigma_E^2 + K\sigma_{SO}^2}{\sigma_E^2 + K\sigma_{SO}^2}$$

If $\sigma_S^2 = 0$, then

$$\frac{MS_S}{MS_{SO}} = \frac{\frac{\frac{(I-1)MS_S}{\sigma_E^2+K\sigma_{SO}^2+JK\sigma_S^2}}{I-1}}{\frac{\frac{(I-1)(J-1)MS_{SO}}{\sigma_E^2+K\sigma_{SO}^2}}{(I-1)(J-1)}}.$$

Since $\frac{(I-1)MS_S}{\sigma_E^2+K\sigma_{SO}^2+JK\sigma_S^2} \sim \chi^2_{(I-1)}$ and $\frac{(I-1)(J-1)MS_{SO}}{\sigma_E^2+K\sigma_{SO}^2} \sim \chi^2_{(I-1)(J-1)}$, we conclude that

$$\frac{MS_S}{MS_{SO}} \sim F_{(I-1),(I-1)(J-1)}.$$

(b) We apply the same reasoning as in question (a). However, the distribution of $\frac{IJ(K-1)MS_E}{\sigma_E^2+K\sigma_{SO}^2}$ is unknown and we get stuck.

(c) We apply the same reasoning as in question (a) to conclude

$$\frac{MS_O}{MS_{SO}} \sim F_{(J-1),(I-1)(J-1)}.$$

(d) We apply the same reasoning as in question (a) to conclude

$$\frac{MS_{SO}}{MS_E} \sim F_{(I-1)(J-1),IJ(K-1)}.$$

Answer to Question 7.6

Using Satterthwaite's approach we estimate *df* such that $df\frac{T}{\mathbb{E}[T]}$ is Chi-squared distributed with *df* degrees of freedom. After estimating *df* with $\hat{df}$, we use

$$\mathbb{P}\left(\chi^{-2}_{\hat{df}}(\tfrac{\alpha}{2}) \leq \hat{df}\frac{T}{\mathbb{E}[T]} \leq \chi^{-2}_{\hat{df}}(1-\tfrac{\alpha}{2})\right) \approx 1-\alpha,$$

where$\chi^{-2}_{df}(x)$ equals the x^{th} quantile of a Chi-square distribution with *df* degrees of freedom. Equivalently,

$$\mathbb{P}\left(\frac{\hat{df}\cdot T}{\chi^{-2}_{\hat{df}}(1-\frac{\alpha}{2})} \leq \mathbb{E}[T] \leq \frac{\hat{df}\cdot T}{\chi^{-2}_{\hat{df}}(\frac{\alpha}{2})}\right) \approx 1-\alpha,$$

Reference

SAS Installation guide

In this course we will use the statistical software SAS University Edition for our analysis. The student version of SAS studio runs on a virtual machine.

Preparation

Create an account on the SAS website below.

`https://login.sas.com/opensso/UI/Login?realm=/extweb&goto=http://www.sas.com/en_us/partnernet/home.html`

Installation

After you logged in on the SAS website visit the url below and follow the four-step installation.

`https://www.sas.com/en_us/software/university-edition/download-software.html`

(1) Download the virtualization software VirtualBox, and and create a "myfolders" folder on your computer. You can download VirtualBox using the following link

`https://www.virtualbox.org/wiki/Downloads`

(2) Download the SAS University Edition vApp.

(3) Import SAS University Edition into VirtualBox, and share your "myfolders" folder with VirtualBox.

(4) Start SAS University Edition.

Common errors

Over the past years we became familiar with some common errors. Fortunately, these can be fixed rather easy.

(1) *Failed to open a session for virtual machine SAS University Edition. Details: VT-x is disables in the BIOS for all CPU modes*

   (i) Enter your `BIOS` settings by restarting your computer while pressing `F10`.

   (ii) Go to `System configuration`.

   (iii) Go to `Device configuration`.

   (iv) Enable `VT-x`.

   (v) Finally, save your changes and leave the `BIOS` settings.

(2) *This kernel requires an x86-64 CPU, but only detected an i686 CPU. Unable to boot please use a kernel appropriate for your CPU.*

   (i) Launch your `VirtualBox` and open the settings.

   (ii) Go to `General`.

   (iii) Switch the `Version` to `Red Hat`.

(3) *The native API dll was not found (C:\WINDOWS\system32\WinHvPlatform.dll) (VERR_NEM_NOT_AVAILABLE).*

(i) Launch `Control panel`.

(ii) Go to `Programs`.

(iii) Go to `Programs and Features`.

(iv) Go to `Turn Windows features on or off`.

(v) Uncheck *'Hyper-V'* box.

If you encounter different errors try to Google this yourself before consulting one of the teaching assistants.

SAS Tutorial

SAS codes generally consist of either a `data` step to create or update a dataset, or a `proc` step to analyze your data. It is crucial to note that the end of a statement line is detected by `;`. Text comments can be made using `/* ... */`.

Data import and export

There are different ways to import data into the SAS environment. You can enter data manually by specifying the name of the dataset (e.g. `COLOR`) and the names of the variables (e.g. `REGION`, `EYES`, `HAIR`, and `COUNT`) after an `input` statement. Furthermore, you should specify if your variables are numerical (default) or character valued (use `$`). The operator `@@` allows you to enter multiple observations from a single record of raw data. Start your raw data after a `datalines;` statement and make sure you end this piece with a `;`. Below we present the code to create a dataset with 6 observations.

```
data COLOR;
    input REGION EYES$ HAIR$ COUNT@@;
    datalines;
    1  blue   fair    23 1 blue   red   7
    1  blue   medium  24 1 blue   dark  11
    1  green  fair    19 1 green  red   7
    ;
run;
```

If you want to export your data (e.g. to import the dataset using a different software package) you can make use of `proc export`. The output file type can be passed on to SAS via the `dbms=` command, for the possible types see the online documentation.

```
proc export data=COLOR
    outfile="/folders/myfolders/color.csv"
    dbms=csv
    replace;
run;
```

Please run the code presented above to save the `COLOR` as a .csv file. We proceed by importing the dataset as `COLOR2` using `proc import`.

```
proc import out=WORK.COLOR2
    datafile="/folders/myfolders/color.csv"
    dbms=csv
    replace;
    getnames = yes;
run;
```

We briefly discuss the different parts of the code:

- `WORK` leads to storing data in the WORK library.
- `"/folders/myfolders/"` is the pathname to access the shared folder.
- `dbms` specifies the type of file to import.
- `replace` overwrites the dataset if a dataset with the same name already exists in the WORK library.
- `getnames` allows to use the column names in a file as variable names in SAS.

Since you will be continuisly using the folder “myfolders” it is efficient to store this location as `SASDATA` using a `libname` statement.

```
libname SASDATA "/folders/myfolders";
```

During this course we will mainly work with *.sas7bdat* files. Use the code below to save the COLOR dataset at the file location SASDATA as *savedata.sas7bdat* (please check that this file is created in your “myfolders” folder).

```
data SASDATA.savedata;
    set COLOR;
run;
```

Importing a *.sas7bdat* file that is saved in “myfolders” is the most convenient way to import a dataset. Using the code below you can create the dataset COLOR3 by reading in the *savedata.sas7bdat* file.

```
data COLOR3;
    set SASDATA.savedata;
run;
```

Basic procedures

Let us continue by analyzing and processing the COLOR dataset using basic procedures. One of the most useful procedures is proc means. This procedure can be used to obtain summary statistics like the mean, median, standard deviation, skewness, kurtosis, and percentiles. Run the code below to get these summary statistics for the COUNT variable from the COLOR dataset.

```
proc means data=COLOR mean median std kurt skew p10;
    var COUNT;
run;
```

SAS can also present these statistics per subgroup (e.g. per eye-color group), using the by statement.

```
proc means data=COLOR mean median std kurt skew p10;
    var COUNT;
    by EYES;
run;
```

Furthermore, the output out = statement can be used to save the results of the analysis in a new dataset named OUTPUT.

```
proc means data=COLOR mean median std kurt skew p10;
    var COUNT;
    by EYES;
    output out = WORK.OUTPUT;
run;
```

Summary statistics can also be reported using proc univariate. Using the code below we create a new dataset name OUTDATA containing the 10% percentile of the COLOR dataset. See the online documentation for all arguments that can be saved as output.

```
proc univariate data=COLOR;
    var COUNT;
    output out=OUTDATA pctlpts=10 pctlpre=percent;
run;
```

Furthermore, proc univariate can be used to draw histograms with fitted probability density curves (e.g. a normal density), or to draw quantile-quantile plots.

```
proc univariate data=COLOR;
    var COUNT;
    histogram COUNT / normal;
    qqplot    COUNT / normal;
RUN;
```

Subsets and merging

We will finish this tutorial by discussing subsets of your data as well as merging datasets. To create the subset `SUBDATA` we use a `data` step and start by setting the old dataset `COLOR`. The `where` statement can be used to put constrains on the old dataset. You can use the online documentation as a reference for all comparison operators.

```
data SUBDATA;
    set COLOR;
    where HAIR ^= 'red';
run;
```

Using the code above we created a subset with all observations that did not have red hair. Within this `data` step it is possible to process your data. Next to selecting the records that did not have red hair, we delete participants with a `COUNT` value above 30 as well as those with green eyes. Finally, we create a new variable `COUNTSQ` which equals the initial `COUNT` variable squared.

```
data SUBDATA;
    set COLOR;
    where HAIR ^= 'red';
    if COUNT > 30 or EYES = 'green' then delete;
    else COUNTSQ = COUNT**2;
run;
```

To illustrate how to combine two datasets, we start with creating two subsets `SUBSET1` and `SUBSET2` using the code below. The `keep` statement can be used to select only a subset of the variables, and thus deleting the other variables. The `drop` statement can be used similarly, but defines which columns of the dataset should be left out.

```
data SUBSET1;
    set COLOR;
    keep REGION COUNT;
run;

data SUBSET2;
    set COLOR;
    drop HAIR;
run;
```

In the case you would be working with two (comparable) datasets from different sources it is easy to paste one under the other by using the `set` statement.

```
data COMBINED;
    set SUBSET1 SUBSET2;
run;
```

Sometimes you are dealing with two (or more) datasets that all contain information about the same subjects. Let us create such an example using the code below. Try to understand why the `@@` is only used while creating the second dataset.

```
data CLASS;
    input NAME $ YEAR $  MAJOR $ ;
    datalines;
    Jennifer    first
    Tom         third       Theater
    Elissa      fourth      Math
    Rachel      first       Math
    ;
run;

data TIME;
    input Name$ DATE$  TIME ROOM@@;
    datalines;
    Jennifer    14sep2000   10  103
    Rachel      14sep2000   10  103
    Tom         14sep2000   11  207
    Elissa      15sep2000   10  105
    ;
run;
```

You can create a new merged (by the NAME variable) dataset CL_TIME using the code below.

```
data CL_TIME;
    merge CLASS TIME;
    by NAME;
run;
```

However, if you run this code SAS will return an error. While programming in SAS it can be crucial to sort you dataset in the order of the variables you are using. To do so one can rely on proc sort, which is also useful if you want to quickly see the maximum value of a variable. We will now sort the dataset CLASS and TIME by NAME since we want to merge based on the value of this later.

```
proc sort data=CLASS;
    by NAME;
run;

proc sort data=TIME;
    by NAME;
run;
```

Rerunning the merging code that previously failed now results in creating the combined dataset CL_TIME. While merging it is possible to process the sub datasets using keep and drop within brackets behind the names of the subsets. This way it is also possible to use the rename statement to change variable names.

```
data CL_TIME;
    merge CLASS (keep=NAME YEAR)
          TIME  (drop=DATE rename=(ROOM=CLASS_NUMBER));
    by NAME;
run;
```

Question 0

On Canvas you can find the datafile *iq.sas7bdat* dataset file, including the following variables:

- LANG: language test score.

- `IQ`: IQ score.
- `CLASS`: class id.
- `GS`: number of students in the class.
- `SES`: social-economic status of family.
- `COMB`: a binary variable indicating whether the students taught in a multi-grade class (`1`=yes and `0`=no).

Practice with the basic SAS commands by answering the questions below.

(a) Save the datafile in your "myfolders" and import the `IQ` dataset into SAS.

(b) Compute all the summary statistics for the variable `IQ` using `proc means`.

(c) Compute these summary statistics for the variable `IQ` in each `CLASS` variable adding

```
    by CLASS;
```

(d) Draw a quantile-quantile plot and a histogram for the variable `IQ` using `proc univariate`.

SAS Procedures

For the questions in this exercise bundle you will need to use the software SAS. Below you can find a list of the most common SAS procedures together with links to their online documentation.

`proc copula`	overview	syntax
`proc corr`	overview	syntax
`proc export`	overview	syntax
`proc freq`	overview	syntax
`proc glm`	overview	syntax
`proc import`	overview	syntax
`proc means`	overview	syntax
`proc mixed`	overview	syntax
`proc npar1way`	overview	syntax
`proc print`	overview	syntax
`proc rank`	overview	syntax
`proc reg`	overview	syntax
`proc sort`	overview	syntax
`proc transpose`	overview	syntax
`proc ttest`	overview	syntax
`proc univariate`	overview	syntax

Dataset descriptions

For the questions in this exercise bundle you will need to following two datasets (available on Canvas). Below you can find a detailed description of these datasets and the variables contained within them.

To access these datasets from SAS, download them and place them in the `myfolders` directory you created during the installation of your SAS University Edition. They should now be available under "My Folders" in SAS. Note: you might need to click the refresh button.

IVF dataset:
A longitudinal data set on the neurological performance of children who are conceived with in-vitro fertilization

Women who have difficulties in getting pregnant can apply for an in-vitro fertilization treatment. These treatments do increase the probability of getting pregnant, but little is known about its long-term effect on the offsprings. A longitudinal follow-up study of mothers participating in a randomized controlled trial on fertilization was conducted. This longitudinal data set contains information on the neurological performance of $n = 253$ children observed at three different time points (at ages of approximately 4, 10, and 18 months). The neurological performance in children is quantified with the infant motor profile (IMP) and higher scores means better performance. The data set contains the following variables.

- Design variables.
  - `ID`: A unique patient number.
  - `TRT`: Treatment indicator (Control = 0, TRT-M = 1, TRT-C = 2).
  - `PER`: Intented age of child for neurological tests (4, 10, or 18 months).
- Characteristics of the parents.
  - `TTP`: Time between treatment and pregnancy (in years).
  - `AGEM`: Age of mother at conception.
- Characteristics of delivery.
  - `FIS`: Stress indicator for child during delivery (No = 0, Yes = 1).
  - `GA`: Duration of pregnancy measured from last menstrual period (in weeks).
- Characteristics of the child.
  - `AGE`: Age of child at the moment of neurological assessments (in weeks).
  - `SEX`: Gender of the offspring (Girl = 0, Boy = 1).
  - `BW`: The weight of the child at birth (in grams).
  - `IMP`: The neurological response ($\in [0, 100]$).

RCT dataset:
A randomized clinical trial on blood process devices for open hart surgery

A randomized clinical trial on blood process devices for open heart surgery was conducted in the Netherlands to investigate how different blood process devices (treatment) help the clinical outcome of patients who undergo open heart surgery. One of the clinical outcomes or responses of the trial was hemoglobin values (Hb) of patients measured at six sequential days, starting at the first day after the surgery. The data (presented in long form) included several variables, and the following variables were selected for our purpose.

- Trial design variables.
    - `ID`: A unique patient number.
    - `CENTER`: A number indicating the hospital.
    - `TIME`: the day at which Hb is measured.
    - `TRT`: Treatment indicator with four levels (Control = 0; Filter = 1; Device = 2; Device + Filter = 3).
- Patient baseline characteristics (characteristics which were measured before trial starts).
    - `AGE`: the age of the patient at the time of the surgery (in years).
    - `SEX`: the gender of the patients (Female = 0; Male = 1).
    - `ANEMIA`: a variable that indicates whether the patient suffers from reduced number of red blood cells (No = 0; Yes = 1). It is defined on the basis of the hemoglobin value. Less than 13 g/dL for males and 12 g/dL for females indicates anemia.
- Patient operation characteristics
    - `HEPA`: the level of heparin that is administered during the operation (in international units).
    - `RESP`: the Hb value (in g/dL).