MASTER OF SCIENCE
IN ENGINEERING

Mock Exam Paper For Part I (About 70 Minutes)

Modul / *Module*:	**Advanced Statistical Data Analysis (AdvStDaAn)**
Datum / *Date*:	**xxx**
Dozierende/n / *Teacher/s*:	**Prof. Dr. A. Ruckstuhl**

Name / *Last name*: --

Vorname / *First name*: --

FH / *UAS*: --

Allgemeine Hinweise:

1. Tragen Sie Ihren Namen auf dem Deckblatt und oben auf jedem Blatt ein.
2. Die Fragen können auf Deutsch oder auf Englisch beantwortet werden.
3. Antworten Sie direkt auf dem Aufgabenblatt. Die Rückseiten dürfen verwendet werden.
4. Wenn Sie Beiblätter für die Beantwortung der Fragen benötigen, verwenden Sie für jede Aufgabe ein neues Blatt und tragen Sie Ihren Namen auf jedem Beiblatt ein.
5. Erlaubte Hilfsmittel:

   Taschenrechner

   Skripten und Bücher

   kein Computer

   keine Mobiles, keine iPads, etc.
6. Es werden während der Prüfung keine Fragen zu den Aufgaben beantwortet. Ist Ihnen eine Frage unklar, dann treffen Sie eine Annahme und erklären Sie diese in Ihrem Lösungsweg. Sie wird bei der Korrektur berücksichtigt.
7. Kommunizieren während der Prüfung ist grundsätzlich verboten. Mobiltelefone sind abzuschalten.
8. Keine rote Farbe verwenden, diese ist für die Korrekturen reserviert.
9. Durchgestrichene Passagen werden ignoriert, auch wenn das Durchgestrichene richtig ist.

Viel Erfolg!

General information:

1. *Write your name on the first page and on the top of every page.*
2. *The questions may be answered in German or in English.*
3. *Please answer directly on the question sheet. You may also use the back side.*
4. *If you need supplementary sheets, please use a separate one for every question. Write your name on every supplementary sheet.*
5. *Material allowed during the exam:*

   Calculator

   Manuscripts and books

   no computer

   no cell phone, no iPad, etc.
6. *No question concerning the problems will be answered during the exam. If you don't understand a problem, make an assumption and explain it in your solution. It will be considered by the grader.*
7. *Communication with others during the exam is forbidden. Mobile phones must be turned off.*
8. *Please don't write in red. This color is reserved for grading.*
9. *Portions of answers that have been crossed out won't be considered, even if the deleted part is correct.*

Good luck!

Name *Last name*	--	Vorname *First name*	--

Question 1 (14 Points)

The motivation for collecting this database was the explosion of the USA Space Shuttle Challenger on 28 January, 1986. An investigation ensued into the reliability of the shuttle's propulsion system. The explosion was eventually traced to the failure of one of the three field joints on one of the two solid booster rockets. Each of these six field joints includes two O-rings, designated as primary and secondary, which fail when phenomena called erosion and blowby both occur.

The night before the launch a decision had to be made regarding launch safety. The discussion among engineers and managers leading to this decision included concern that the probability of failure of the O-rings depended on the temperature at launch, which was forecasted to be 31°F. There are strong engineering reasons based on the composition of O-rings to support the judgment that failure probability may rise monotonically as temperature drops. One other variable, the pressure at which safety testing for field join leaks was performed, was available, but its relevance to the failure process was unclear.

The data frame **`chal`** contains the variables number of O-rings at risk on a given flight (**`m`**), number experiencing thermal distress (**`Fails`**), launch temperature (**`temp`** , in °F), and leak-check pressure (**`Pres`** , in psi).

```
R Output:
> chal.glm1 <- glm(cbind(Fails, m-Fails) ~ Pres + Temp,
+                  family=binomial(link=logit), data=chal)
> summary(chal.glm1)

Call:
glm(formula = cbind(Fails, m - Fails) ~ Pres + Temp, family = binomial(link =
logit),
    data = chal)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.05383  -0.65352  -0.56140  -0.03971   2.37171

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.409728   3.178539   1.073   0.2834
Pres         0.007380   0.006447   1.145   0.2523
Temp        -0.107747   0.044648  -2.413   0.0158 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 16.565  on 20  degrees of freedom
AIC: 36.125

Number of Fisher Scoring iterations: 5
```

Continue on next page

```
R Output:
> chal.glm2 <- glm(cbind(Fails, m-Fails) ~ Temp,
                   family=binomial(link=logit), data=chal)
> anova(chal.glm1, chal.glm2, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(Fails, m - Fails) ~ Pres + Temp
Model 2: cbind(Fails, m - Fails) ~ Temp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        20     16.565
2        21     18.086 -1  -1.5212   0.2174

> confint(chal.glm1)
Waiting for profiling to be done...
                   2.5 %      97.5 %
(Intercept) -2.776236540  9.93358512
Pres        -0.004030283  0.02272544
Temp        -0.201164111 -0.02229717

> (h1 <- predict(chal.glm2, newdata=data.frame(Temp=31), type="response", se=T))
$fit
        1
0.8177744

$se.fit
        1
0.2404526

$residual.scale
[1] 1

> h1$fit + c(-1,1) * qnorm(0.975)*h1$se.fit
[1] 0.346496 1.289053
> h2 <- predict(chal.glm2, newdata=data.frame(Temp=31), type="link", se=T)
> h2a <- h2$fit + c(-1,1)*qnorm(0.975)*h2$se.fit
> 1/(1 + exp(-h2a))
[1] 0.1596025 0.9906582
```

a) (4 Points) Formulate a generalized linear model to estimate the probability that an O-ring is defect. (Response and its distribution, linear predictor, and link function.).

Solution: ½ 1

Response **`Fails`** is independent and binomial distributed with expectation E(Fails) = m·μ_i= m·π_i where the size m_i is the number of O-rings at risk and ·μ_i = π_i is the probability of failure of an O-ring.

½

Link = canonical link (i.e., logit) because there is none mentioned explicitly. 1

Linear predictor: $\eta_i=\beta_0 + \beta_1$ Pres + β_2 Temp 1

b) (3 Points) Based on the R output, test the hypothesis (i.e., the main concern) mentioned in the initial text on the 5% significance level. (Discuss how you proceed.)

Solution:
Null hypothesis: $\beta_1 = 0$; Alternative $\beta_1 \neq 0$ 1

Because it is more reliable than the Wald test we use the deviance test (i.e. use anova). 1

Since the p-value of 0.2174 is larger than 5% we cannot reject the null hypothesis; hence might be 0, so Pres has no influence on the probability of failure. 1

c) (2 Points) Based on the R output, report a 95% confidence interval for the coefficient assigned to the explanatory variable **`Temp`**. On which principle (i.e., test statistics) is the calculation done in the R function **`confint()`**.

Solution:
Using the output of confint() the 95% confidence interval is [-0.201, -0.022]. 1

confint () calculates the confidence interval based on the deviance test statistic which is more reliable than the Wald test statistic. 1

d) (3 Points) The expected temperature at launch on January 20, 1986 was 31°F. Based on the R output, report the predicted probability that an O-ring will leak at this temperature. Report its 95% confidence interval. Use the result of the more reliable method and give reasons for your selection.

Solution:
The predicted probability that an O-ring will leak at 31°F is 0.8177744. 1

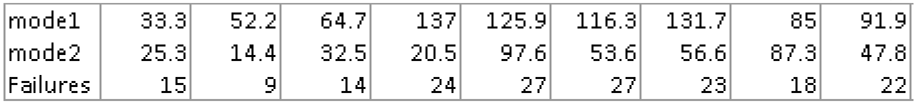
The 95% confidence interval using the more reliable method is [0.1596 0.9907]. 1
It is more reliable because the approach based on the linear predictor guarantees that the confidence interval is within the domain of a probability value. 1

e) (2 Points) The engineering team recognized that they did not have data below 53°F, and decided to look at all cases where there had been signs of O-ring distress. In this case the data was limited only to incidents of O-ring thermal distress, defined as O-ring erosion, blowby, or excessive heating) — exactly the question of interest.
What is wrong with this selection of data for answering the question of interest?

Solution:
It is a mistake to assume that O-rings with no thermal distress do not contain information about the failure of O-rings. 1
Hence it is usually a bad advice to drop data based on the specified criterion because you never know whether this criterion may influence the response (if you need to sample do it randomly!) 1

Question 2 (16 Points)

A machinery is run in two modes and the objective of the analysis is to determine whether the number of failures (Failures) depends on how long the machine is run in mode 1 (mode1) or mode 2 (mode2). The data frame **tm** contains the following data:



mode1	33.3	52.2	64.7	137	125.9	116.3	131.7	85	91.9
mode2	25.3	14.4	32.5	20.5	97.6	53.6	56.6	87.3	47.8
Failures	15	9	14	24	27	27	23	18	22

```
R Output
> tm.glmA <- glm(Failures ~ Mode1 + Mode2, family=poisson(link="identity"),
+                 data=tm)
> summary(tm.glmA)
 ... stripped-down ...

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.99773    3.63545   1.650  0.09899 .
Mode1        0.12081    0.04578   2.639  0.00832 **
Mode2        0.05459    0.06356   0.859  0.39037
---
 (Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16.9964  on 8  degrees of freedom
Residual deviance:  4.1971  on 6  degrees of freedom
AIC: 53.254

Number of Fisher Scoring iterations: 6

> tm$Mode <- tm$Mode1 + tm$Mode2
> tm.glmB <- glm(Failures ~ Mode, family=poisson(link="identity"), data=tm)
> anova(tm.glmA, tm.glmB, test="Chisq")
Analysis of Deviance Table

Model 1: Failures ~ Mode1 + Mode2
Model 2: Failures ~ Mode
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         6     4.1971
2         7     4.6697 -1 -0.47256   0.4918

> summary(tm.glmB)
 ... stripped-down ...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.22215    3.60336   1.727 0.084210 .
Mode         0.09658    0.02626   3.677 0.000236 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16.9964  on 8  degrees of freedom
Residual deviance:  4.6697  on 7  degrees of freedom
AIC: 51.726

Number of Fisher Scoring iterations: 3

1-pchisq(4.6697, 7)
[1] 0.7001979
```

a) (4 Points) Formulate an obvious generalized linear model that could well describe the data. (Response and its distribution, linear predictor, and link function.)

Solution:
The response **`Failures`** is independent and Poisson distributed with expectation $\mu_i=\lambda_i$. 1

The text suggests using the "identity" link, because one rather wants a direct influence of the operating time on the failure rate in each mode. This choice is supported by the fact that both operating times are positive explanatory variables, and thus, with positive parameter values, the linear predictor is also positive. Therefore, the link "identity" guarantees a positive failure rate. -- But the log link is not excluded by these arguments! 2

Linear predictor: $\eta_i=\beta_0 + \beta_1 \text{Mode1} + \beta_2 \text{Mode2}$ 1

b) (2 Points) Based on the R output, report the estimated coefficients. Do the estimated coefficients have a plausible sign?

Solution:
$\beta_0 = 5.997$; $\beta_1 = 0.121$; $\beta_2 = 0.055$ 1

As hoped for, the coefficients have positive signs (-> positive linear predictor). 1

c) (3 Points) Test the hypothesis on the 5% significance level that the influence of both variables Mode1 and Mode2 is equal on the response based on the R output. (Report the null hypothesis, the alternative, how you conclude and the conclusion.)

Solution:
Null hypothesis: $\beta_1 = \beta_2$ Alternative: $\beta_1 \# \beta_2$ 1

We can use a deviance test statistics by introducing a new variable which is the sum of Mode1 and Mode2. Hence we apply **`anova`** to the model of part (a) and the model where the response just depends on the singe variable **`Mode`** (cf. R output). 1
Since the p-value of `0.4918` is larger than 5% we cannot reject the null hypothesis. So β_1 might be equal to β_2. 1

d) (2 Points) Does the residual deviance indicate that the reduced model described in part 2c) is not satisfactory?

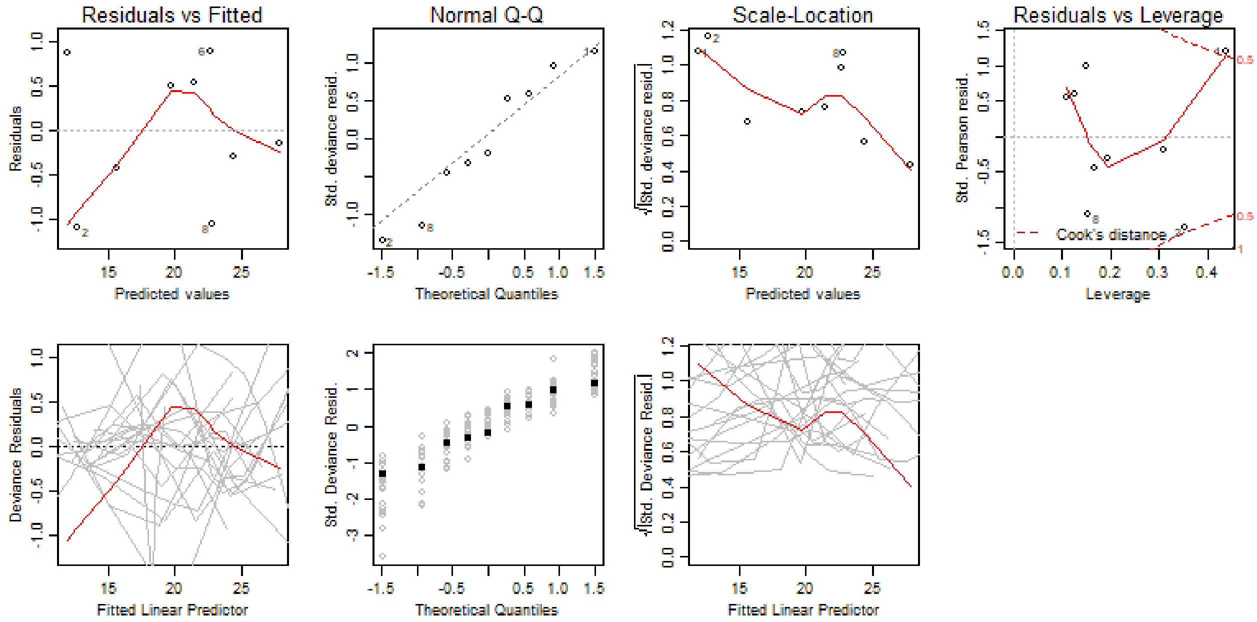
Solution:
Because the response is Poisson distributed we can test on overdispersion:
null hypothesis $\phi=1$ (i.e., no overdispersion); Alternative $\phi\#1$ (i.e., overdispersion) 1

The p-value of the residual deviance is 0.70 which is larger than 5%. Hence there is no evidence that the dispersion parameter ϕ is larger than 1.
That is, according to the dispersion structure the model describes the data well. 1

Name
Last name ....

Vorname
First name ....

e) (5 Points) Run a residual and sensitivity analysis based on these four graphs:



(i) What do you learn from which graph?
(ii) What is your overall conclusion?
(iii) Based on your conclusion are the results in part 2c) and part 2d) statistically sound?

Solution:

- Residual plot: The smoother indicates some non-constant expectation, but it is almost completely within the stochastic fluctuation (i.e. the gray spaghettis). Thus, there is a weak evidence against an adequately specified expectation. 1
- Location-scale plot: The smoother is decreasing but still within the stochastic fluctuation. Hence, there is no evidence against a adequately specified variance. 1
- Normal Q-Q plot: The points scatter more or less around a straight line and all black points are within the stochastic fluctuation. Hence, there is no evidence for outliers or some distortions of the distributional assumption. 1
- Residual against leverage: Since all Cook's distances are smaller than 1, there is no observation with a too large influence. ½
- Conclusion: Model might be adequate although there is a weak hint that the expectation might not be constant. (However, there are just not enough data available.) ½
- Because there is no (serious) evidence against the assumption of the model one can trust the statistical conclusions in part 2c) and part 2d). 1

************************** ENDE *****************************