Institut für Mathematik
Prof. Dr. P. Ruckdeschel
M.Sc. T. Werner

WiSe 18/19
03. April 2019

Statistik I: Einführung in die Angewandte Statistik

Wiederholungsklausur (120 Minuten)

Name, Vorname: .... **Matr.-Nr.:** ....

Studiengang: .... **Fachsemester:** ....

a) Tragen Sie zuerst Ihre persönlichen Angaben ein, unterschreiben Sie und legen Sie Ihren Personalausweis/Lichtbildausweis bereit.

b) Prüfen Sie nach, ob Ihr Klausurexemplar vollständig ist und aus 44 Seiten besteht.

c) Um Ihnen die Übersicht zu erleichtern, sind am Anfang der Klausur auf den Seiten 2 bis 9 alle Aufgaben einmal abgedruckt.

d) Hilfsmittel sind nicht zugelassen. Notationen und Hinweise finden Sie auf Seite 8.

e) Schreiben Sie Ihre Lösungen **ab Seite 10** jeweils direkt unter den Aufgabentext und auf die nächstfolgende Seite. Bei Bedarf stehen Ihnen am Ende der Klausur ab Seite 42 zusätzliche Blätter zur Verfügung. Vermerken Sie bei der entsprechenden Aufgabe die Seitenzahl der Seite, die die Fortsetzung Ihrer Lösung enthält. **Die gehefteten Blätter dürfen nicht voneinander getrennt werden; die Benutzung eigenen Papiers ist nicht zulässig.**

f) Benutzen Sie einen dokumentenechten Stift (keinen Bleistift) und schreiben Sie nicht mit Rot. Mit Bleistift Geschriebenes und lose Blätter werden bei der Korrektur ignoriert.

g) Achten Sie darauf, den Lösungsweg zu jeder bearbeiteten Aufgabe nachvollziehbar aufzuschreiben und notwendige Zwischenschritte anzugeben.

h) Täuschungsversuche, insbesondere mitgebrachte Zusatzmaterialien, wie z.B. Smartphone, führen zum Nichtbestehen der Klausur.

i) Die durch die Übungen erworbenen Bonuspunkte werden nur bei Bestehen der Klausur berücksichtigt.

Zur Kenntnis genommen: ....
Unterschrift

Bewertung:

	Aufg. 1	Aufg. 2	Aufg. 3	Aufg. 4	Aufg. 5	Aufg. 6	$\sum$ 1–6	Bonus
Punkte	30	10	7	9	8	9	73	7
erreicht								
Korrektur								

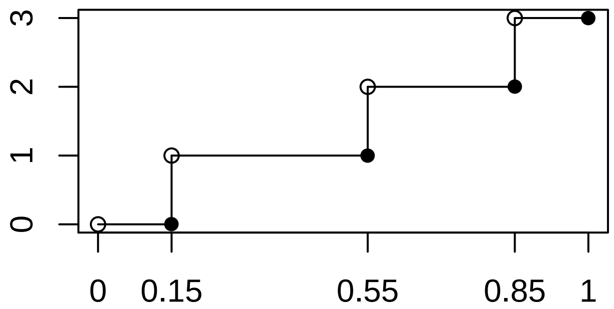
	Aufg. 7	Aufg. 8	Aufg. 9	Aufg. 10	Aufg. 11	Aufg. 12	$\sum$ 7–12	Gesamt	Note
Punkte	14	16	9	8	20	10	77	150 (+7)	
erreicht									
Korrektur									

Aufgabe 1 Wahr oder falsch? (30 Punkte)

Geben Sie jeweils ein Gegenbeispiel oder einen Beweis / eine Begründung. Dabei dürfen Sie sich gegebenenfalls auf Sätze der Vorlesung berufen.

a) Sei F^- die Pseudoinverse einer diskreten Verteilung. Ist $F(t) = \alpha$ für ein $\alpha \in]0,1[$, so ist $F^-(\alpha) = t$.

b) Die folgend graphisch dargestellte Funktion ist eine Quantilsfunktion einer diskreten Verteilung.



c) Sei $f(k)$ eine Wahrscheinlichkeitsfunktion für $k \in \mathbb{N}_0$. Dann ist $f(k) > 0$ für alle $k \in \mathbb{N}_0$ möglich.

d) Stimmen in einem Boxplot der Median und der obere Rand der Box überein, so können Ausreißer nach oben und die untere Antenne (mit Länge > 0) nicht gleichzeitig auftreten.

e) Der Spearmansche Rangkorrelationskoeffizient ist immer im Intervall $[-1, 1]$.

f) Das in der Vorlesung eingeführte $(1-\alpha)-$Konfidenzintervall I_n für eine unbekannte Binomialwahrscheinlichkeit p ist immer eine Teilmenge von $[0, 1]$.
Erinnerung: Sei q das $(1-\alpha/2)$-Quantil von $\mathcal{N}(0,1)$, π_n die relative Häufigkeit zum Stichprobenumfang n und dann war $I_n = [\hat{\pi}_n - s_n, \hat{\pi}_n + s_n]$ für $s_n^2 = q^2(\hat{\pi}_n - \hat{\pi}_n^2)/n$.

g) Sei $\mathcal{P} = \{P_\lambda = \text{Pois}(\lambda) \mid \lambda > 0\}$ und seien X_i u.i.v. P_λ. Dann kann man einen unverzerrten Schätzer für θ mit beliebig kleiner Varianz finden.

h) Die Prüfgröße eines Varianztests (also ein χ^2- oder ein F-Test) nimmt stets Werte größer als Null an.

i) Sind die erklärende und die abhängige Variable im einfachen linearen Regressionsmodell negativ korreliert, so ist der durch den Kleinste-Quadrate-Schätzer geschätzte Steigungsparameter von Null verschieden.

j) Beim multiplen linearen Regressionsmodell ist das volle Modell, welches alle Prädiktoren berücksichtigt, immer optimal hinsichtlich der erklärten Varianz auf den Trainingsdaten.

Aufgabe 2 Deskriptive Kennzahlen (2+3+5=10 Punkte)

Sie dürfen die Formeln auf Seite 8 verwenden.

a) Sie haben die Beobachtungen

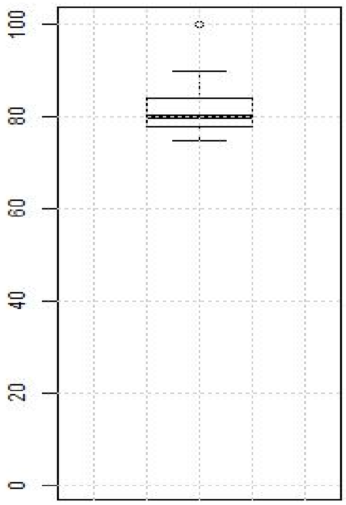
$$x = (-12, 5, 8, 1, -1, -25, 0, -2)$$

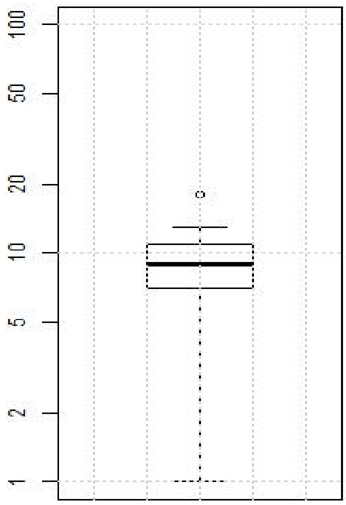
vorliegen. Berechnen Sie daraus das $\alpha-$getrimmte und das $\alpha-$winsorisierte Mittel für $\alpha = 0.25$.

b) Berechnen Sie die Kurtosis der Stichprobe.

HINWEIS: Für Beobachtungen $X = (X_1, ..., X_n)$ seien $Y_i := (X_i - \bar{X}_n)/\hat{\sigma}^2_X$. Die Kurtosis ist dann $\bar{(}Y_n^4) - 3$.

c) Interpretieren Sie die folgenden Boxplots. Beschreiben Sie dabei exakt, was jede horizontale Linie in den Boxplots aussagt. Vergleichen Sie die beiden Verteilungen hinsichtlich Symmetrie und Streuung.





HINWEIS: Beachten Sie die unterschiedlichen y-Achsen-Skalierungen!

Aufgabe 3 Konzentrationen (4+2+1=7 Punkte)

Gegeben seien vier Gruppen von Personen mit einem kumulierten Vermögen von 2000 (Angabe in Tausend Euro). Niemand von den 60 Personen der ersten Gruppe, die sich ein Vermögen von 60 teilen, hat Schulden, jedoch besitzen alle weniger als die 100 Personen, die eine weitere Gruppe bilden. Sie kommen insgesamt auf 900, was 400 mehr ist als das Gesamtvermögen der Gruppe von 25 Personen, die aber individuell jeweils mehr besitzen als jede der besagten 100 Personen, aber weniger als die bisher ungenannten übrigen Personen.

a) Bestimmen Sie für je eine Person aus jeder Gruppe jeweils eine Ober- und Untergrenze für das individuelle Vermögen.

b) Bestimmen Sie weiterhin die maximale Anzahl an Personen aus der dritten Gruppe (die 25 Personen), die mindestens 75 besitzen können.

c) Wie groß ist die maximale Anzahl an Personen, die gar nichts besitzen?

HINWEIS: Ist a) nicht gelungen, so rechnen Sie in b) und c) mit den falschen Grenzen $\{1,2\}$ für Gruppe 1, $\{3,4\},\{5,6\},\{7,8\}$ für Gruppen 2, 3 bzw. 4.

Aufgabe 4 Definitionen, Sätze und Beweise (3+4+2=9 Punkte)

a) Formulieren Sie (ohne Beweis) den Satz von der totalen Wahrscheinlichkeit und beweisen Sie damit den Satz von Bayes.

b) Beweisen Sie mit der Tschebyschev-Ungleichung das schwache Gesetz der Großen Zahlen für reellwertige Zufallsvariablen $X_i \sim P$ i.id. mit Erwartungswert μ und Varianz σ.

c) Geben Sie eine statistische Anwendung beim Schätzen, beim Testen oder bei der Regression an, wo man dieses Gesetz verwendet und skizzieren Sie, wozu man es braucht.

Aufgabe 5 Abhängigkeit und Unabhängigkeit (4+4=8 Punkte)

Gegeben seien zwei Zufallsvariablen X und Y, die jeweils die Ausprägungen 1 und (-1) annehmen können. Wir haben Realisationen des Paares (X, Y) erhoben, allerdings sind in der Kontingenztabelle die inneren Zellen versehentlich gelöscht worden.

$Y \setminus X$	-1	1	$\sum$
-1			7
1			4
$\sum$	8	3	

a) Befüllen Sie die Tabelle mit den Variablen $N_{-1,-1}$, $N_{1,-1}$, $N_{-1,1}$, $N_{1,1}$. Bestimmen Sie eine Wertebelegung von $N_{i,j}$ so, dass maximale Abhängigkeit bzw. bestmögliche Unabhängigkeit zwischen X und Y herrscht. Verwenden Sie das Abstandsmaß χ, das Sie beim Unabhängigkeitstest verwenden.

b) Sie haben reellwertige Zufallsvariablen T_X und T_Y, die die Temperatur an zwei ausgewählten Orten X und Y beschreiben. Sie haben nun n Messungen in Grad Celsius durchgeführt und herausgefunden, dass X und Y positiv korreliert sind, einmal gemessen durch den Pearson-, einmal durch den Spearman-Korrelationskoeffizienten.

Welche Eigenschaft sichert Ihnen, dass diese Abhängigkeit nicht verloren geht, wenn Sie auf die Fahrenheit-Skala wechseln? Beweisen Sie für den Spearman-Korrelationskoeffizienten, dass diese Eigenschaft hier vorliegt.

HINWEIS: Ist t die Temperatur in Grad Celsius, so ist $32+1.8t$ die Temperatur in Fahrenheit.

Aufgabe 6 Diskrete Verteilungen (4+5=9 Punkte)

Sei X eine $\{0,2\}-$, Y eine $\{-2,-1\}-$ und Z eine $\{-1,0\}-$wertige Zufallsvariable. Lesen Sie sorgfältig, was in den folgenden beiden Tabellen gegeben ist:

$P(X=i,Y=j,Z=1)$				$P(X=i,Y=j\|Z=0)$		
$\downarrow j, i \rightarrow$	0	2		$\downarrow j, i \rightarrow$	0	2
-1	0.2	0.3		-1	0	0.15
-2	0.05	0.25		-2	0.05	0

a) Berechnen Sie $\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[Z]$.

b) Berechnen Sie $\mathbb{E}[Y^2]$ und $\mathbb{E}[XYZ]$.

Aufgabe 7 Qualität von Schätzern (4+6+4=14 Punkte)

a) Definieren Sie den MSE eines Schätzers. Beweisen Sie überdies dessen Bias-Varianz-Zerlegung.

b) Seien $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.id.. Herr B.Est meint, Varianzen könne man besser schätzen als mit den in der Vorlesung gegebenen Schätzern

$$S_{n,m} := \frac{1}{m}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \quad m \in \{n-1, n\}.$$

Sei allgemein $S_n^\alpha := \alpha S_{n,n}$. Bestimmen Sie

$$\hat{\alpha} := \underset{\alpha}{\operatorname{argmin}}(MSE(S_n^\alpha)).$$

Hinweis: Beachten Sie: Ein unverzerrter Schätzer für die Varianz ist

$$\frac{1}{n-1}\sum_i (x_i - \bar{x}_n)^2.$$

Ferner ist

$$\mathcal{L}(\chi_n^2) = \sum_{i=1}^{n} X_i^2 \text{ mit } \; X_i \sim \chi_1^2 \; \text{ i.id.},$$
$$\operatorname{Var}(\chi_1^2) = 2\sigma^4.$$

c) Seien $X_1, ..., X_n$ unabhängig identisch $\mathcal{N}(\mu, \sigma^2)-$verteilt mit $\sigma^2 := 4\mu \in [1,5]$. Wie groß muss n sein, damit in jedem Falle die Wahrscheinlichkeit, dass der Abstand von $\bar{X}_n$ und $\mathrm{E}[\bar{X}_n]$ weniger als 10 Prozent von $\mathrm{E}[\bar{X}_n]$ ausmacht, mindestens 97.5 Prozent ist?

Aufgabe 8 Schätzen und Konfidenzintervalle (4+2+3+3+4=16 Punkte)

Umberto und Kunigunde haben einen Würfel mit inhomogener Gewichtsverteilung, was dazu führt, dass es beim Würfeln nicht mehr gleichwahrscheinlich ist, auf welcher Seite der Würfel liegen bleibt. Sie würfeln so lange, bis sie zum dritten Mal eine sechs sehen. Für die Wartezeit der Fehlversuche bis zum ersten Treffer in diesem Kontext betrachten wir $X \sim \mathrm{Geom}(p)$, d.h., die diskrete Verteilung mit Wahrscheinlichkeitsfunktion $P(X = k) = p(1-p)^k, k \in \mathbb{N}_0$. Es gilt für $q := (1-p)$, dass

$$\mathbb{E}(X) = \frac{q}{p}, \mathrm{Var}(X) = \frac{q}{p^2}.$$

Ferner ist für $X_1, \ldots, X_m \overset{\text{u.i.v.}}{\sim} \mathrm{Geom}(p)$ die Summe

$$S_m = \sum_{i=1}^{m} X_i$$

NegBinom$(m, p)-$verteilt.

a) Konstruieren Sie allein auf Basis der Zahl der Würfelwürfe bis zur jeweils nächsten 6 einen Momentenschätzer für die unbekannte Wahrscheinlichkeit p_6.

b) Skizzieren Sie, wie Sie den ML-Schätzer für p_6 berechnen würden (nur beschreiben, d.h., die Rechnungen führen Sie nicht durch).

c) Umberto und Kunigunde haben nun den Momentenschätzer $\hat{p}_6^{mom}$ aus a). Undine und Sieghold stöbern heimlich in ihren Aufzeichnungen, finden jedoch nur den berechneten Wert für diesen Schätzer, wissen aber nicht, bis zur wievielten gesehenen sechs gewürfelt wurde. Bestimmen Sie den Momentenschätzer für die gesuchte Anzahl in Abhängigkeit von $\hat{p}_6^{mom}$.

d) Untersuchen Sie $\hat{p}_6^{mom}$ auf Konsistenz.

e) Seien nun für alle 6 mögliche Augenzahlen jeweils p_j die wahren Wahrscheinlichkeiten und r_{jn} die relativen Häufigkeiten basierend auf n Würfen. Sie haben Konfidenzintervalle

$$I_j := [r_{jn} - \delta_{jn}^-, r_{jn} + \delta_{jn}^+]$$

für alle j.

Erläutern Sie, wie man die δ_{jn} mit dem Bootstrapverfahren gewinnen kann.

Aufgabe 9 Hypothesentests (3+2+4=9 Punkte)

a) Darf man bei einem statistischen Test H_0 und H_1 immer vertauschen? Begründen Sie dies.

b) Wann heißt ein statistischer Test unverzerrt? Was bedeutet dies anschaulich?

c) Jemand würfelt so lange, bis zum 100. Mal eine Sechs erscheint. Da 1000 Würfe benötigt wurden, glaubt dieser Mensch, dass der Würfel unfair ist und will einen Gaußtest durchführen. Ist dies zulässig? Betrachten Sie dazu die (beobachtete) durchschnittliche Zahl der Würfelwürfe, bis jeweils zum 100. Mal eine 6 eintritt. Plausibilisieren Sie, ob dies auch effizient ist, um mit möglichst wenig Würfen eine Aussage zu vorgegebener Signifikanz treffen zu können.

Aufgabe 10 Rätsel lösen... (2+2+4=8 Punkte)

Angenommen, es gibt eine besonders mürrische Sphinx, die vorbeikommende Reisende nicht ziehen lässt, wenn sie ein Rätsel lösen konnten, sondern die Lösung dreier Rätsel verlangt. Die einzige Person, die das geschafft hat, hat Nachforschungen angestellt und stellt sein Ergebnis in folgendem Rätsel dar:

„*40 der Personen, die das erste Rätsel lösen konnten, scheiterten an dem zweiten, während sechs derer, die auch dies bewältigten, das letzte Rätsel nicht überlebten. 222 Personen sprachen insgesamt mit der Sphinx.*"

Sie betrachten die Verteilung für die Zahl S der richtigen Antworten, also eine Verteilung auf $\{0, 1, 2, 3\}$. Nach Betrachtung der relativen Häufigkeiten meinen Sie, dass eine $Bin(3, 0.1)-$Verteilung passen könnte.

a) Stellen Sie explizit Ihre Hypothese auf und nennen Sie den benötigten statistischen Test.

b) Berechnen Sie aus den Daten die richtigen relativen Häufigkeiten.

c) Führen Sie den statistischen Test zum Niveau $\alpha = 0.05$ durch.

Aufgabe 11 Lineare Regression (3+9+2+3+3=20 Punkte)

a) Seien X_1, X_2 erklärende Variablen in einem linearen Regressionsmodell. Unsere beobachteten Werte für X_1 bzw. X_2 sind $(2, 3, 8)$ bzw. $(-2, -1, 6)$. Stellen Sie die Regressormatrix für folgende Szenarien explizit auf:
   - X_1, X_2 als erklärende Variablen, kein Achsenabschnitt
   - Das einfache lineare Regressionsmodell bzgl. X_2
   - Polynomiale Regression bis Grad 3 für X_1

b) Betrachten Sie folgenden R−Output:

```
Call:
lm(formula = Y ~., data = D)

Residuals:
Min      1Q     Median   3Q      Max
-6.347   -1.700  0.451   1.590   6.881

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.9278    0.8640      3.389    0.001268 **
X1           0.6593    0.3993      1.651    0.104094
X2           -0.3702   0.3279      -1.129   0.263517
X3           -1.4822   0.3673      -4.035   0.000162 ***
X4           -0.9753   0.3680      -2.650   0.010353 *
X5B          4.1965    1.0142      4.138    0.000115 ***
X5C          -5.9071   0.9972      -5.924   1.82e-07 ***
```

```
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.689 on 58 degrees of freedom
Multiple R-squared: 0.8003, Adjusted R-squared: 0.7796
F-statistic: 38.73 on 6 and 58 DF, p-value: < 2.2e-16
```

(i)) Welche Koeffizienten sind signifikant? Begründung? Was bedeutet anschaulich der Koeffizient von X2?

(ii) Zeigen Sie anhand der anderen vorliegenden Informationen, wie Sie auf den t-Wert für X4 kommen. Welche Hypothese steckt dahinter? Wie bekommen Sie für die Variable X3 den angegebenen p-Wert?

(iii) Wie interpretieren Sie die letzten beiden Koeffizienten in der Tabelle? Interpretieren Sie die vorletzte Zeile des Outputs.

(iv) Wie viele Beobachtungen lagen eigentlich vor?

c) Sie haben zwei Modelle und erhalten für das erste Modell einen AIC von 2.4 und einen BIC von 3.2, während Sie für das zweite Modell einen AIC von 3.2 und einen BIC von 2.9 bekommen. Wann würden Sie sich eher für das erste, wann eher für das zweite entscheiden? Begründen Sie Ihre Antwort.

d) Motivieren Sie, was der Vorteil sparsamer Modellwahl (viele Koeffizienten sind Null) ist. Schreiben Sie das Optimierungsproblem der linearen Regression mit Ridge- und Lasso-Strafterm auf.

e) Visualisieren Sie Ihre Strafterme für $p = 1$. Welcher Strafterm wäre eigentlich optimal? Welche Vorteile haben der Ridge- und Lasso-Strafterm hinsichtlich numerischer Optimierung, welcher durch den eigentlich optimalen Strafterm verloren ginge?

Aufgabe 12 Generalisierte lineare Modelle (2+2+1+2+3=10 Punkte)

a) Definieren Sie ein allgemeines generalisiertes lineares Modell. Was versteht man unter einer Linkfunktion? Wie lautet diese für das gewöhnliche lineare Regressionsmodell (Aufgabe 11)?

b) Nehmen Sie an, Ihre abhängige Variable Y nimmt die Werte ,,Vampir" und ,,Banshee" an. Sie haben p erklärende Variablen $X_1, ..., X_p, p \geq 2$, zur Verfügung. Stellen Sie für diesen konkreten Fall Ihr Probit-Modell mit allen erklärenden Variablen auf.

c) Sie haben nun die Koeffizienten für Ihr Modell aus b) vorliegen. Der Koeffizient β_p von X_p ist 7.22. Wie interpretieren Sie diesen Koeffizienten auf der Prädiktorskala, also auf Ebene $X\beta$? Welchen Effekt hat eine Vergrößerung von β_p bei sonst gleichen Koeffizienten und gleichen Prädiktoren X_i auf Ebene der Wahrscheinlichkeiten?

d) Angenommen, für drei neue Beobachtungen (d.h., alle X-Werte sind bekannt, der Y-Wert jeweils unbekannt) sind die berechneten Werte von $X\beta$ 0.26, 0.61 und 0.98. Für welche der Beobachtungen sagen Sie ,,Vampir" bzw. ,,Banshee" voraus?

e) Was ist der Unterschied des Logit-Modells zum Probit-Modell? Was ist beim Ersteren die Link-Funktion? Kann es noch andere geeignete Link-Funktionen für unser Problem aus b) geben?

Hinweise (für alle Aufgaben):

Tabelle einiger Quantile der Standardnormalverteilung:

α	0.005	0.01	0.025	0.05
	-2.58	-2.33	-1.96	-1.64

Tabelle von $(\chi^2_d)^{-1}(\alpha)-$Werten:

d \ α	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995
1	0.000039	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.1	5.99	7.38	9.21	10.6
3	0.072	0.11	0.22	0.35	7.81	9.35	11.34	12.84
6	0.678	0.87	1.24	1.64	12.59	14.45	16.81	18.55
149	108.29	111.8	117.1	121.79	178.49	184.69	192.07	197.21
150	109.14	112.67	117.98	122.69	179.58	185.8	193.21	198.36
151	109.99	113.53	118.87	123.6	180.68	186.91	194.34	199.51

Tabelle von $t_\nu^{-1}-$Werten:

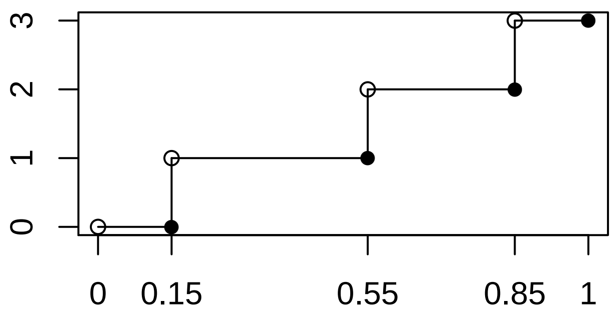
ν \ α	0.005	0.01	0.025	0.05
99	-2.6264	-2.3646	-1.9842	-1.6604
100	-2.6259	-2.3642	-1.9840	-1.6602
101	-2.6254	-2.3638	-1.9837	-1.66
102	-2.6249	-2.3635	-1.9835	-1.6599
149	-2.6092	-2.3516	-1.976	-1.6551
150	-2.609	-2.3514	-1.9759	-1.6551
151	-2.6088	-2.3513	-1.9758	-1.655
190	-2.6019	-2.3461	-1.9725	-1.6529
191	-2.6018	-2.346	-1.9725	-1.6529
192	-2.6017	-2.3459	-1.9724	-1.6528

Aufgabe 1 Wahr oder falsch? (30 Punkte)

Geben Sie jeweils ein Gegenbeispiel oder einen Beweis / eine Begründung. Dabei dürfen Sie sich gegebenenfalls auf Sätze der Vorlesung berufen.

a) Sei F^- die Pseudoinverse einer diskreten Verteilung. Ist $F(t) = \alpha$ für ein $\alpha \in]0,1[$, so ist $F^-(\alpha) = t$.

b) Die folgend graphisch dargestellte Funktion ist eine Quantilsfunktion einer diskreten Verteilung.



c) Sei $f(k)$ eine Wahrscheinlichkeitsfunktion für $k \in \mathbb{N}_0$. Dann ist $f(k) > 0$ für alle $k \in \mathbb{N}_0$ möglich.

d) Stimmen in einem Boxplot der Median und der obere Rand der Box überein, so können Ausreißer nach oben und die untere Antenne (mit Länge > 0) nicht gleichzeitig auftreten.

e) Der Spearmansche Rangkorrelationskoeffizient ist immer im Intervall $[-1, 1]$.

f) Das in der Vorlesung eingeführte $(1-\alpha)-$Konfidenzintervall I_n für eine unbekannte Binomialwahrscheinlichkeit p ist immer eine Teilmenge von $[0, 1]$.
Erinnerung: Sei q das $(1-\alpha/2)$-Quantil von $\mathcal{N}(0,1)$, π_n die relative Häufigkeit zum Stichprobenumfang n und dann war $I_n = [\hat{\pi}_n - s_n, \hat{\pi}_n + s_n]$ für $s_n^2 = q^2(\hat{\pi}_n - \hat{\pi}_n^2)/n$.

g) Sei $\mathcal{P} = \{P_\lambda = \mathrm{Pois}(\lambda) \mid \lambda > 0\}$ und seien X_i u.i.v. P_λ. Dann kann man einen unverzerrten Schätzer für θ mit beliebig kleiner Varianz finden.

h) Die Prüfgröße eines Varianztests (also ein χ^2- oder ein F-Test) nimmt stets Werte größer als Null an.

i) Sind die erklärende und die abhängige Variable im einfachen linearen Regressionsmodell negativ korreliert, so ist der durch den Kleinste-Quadrate-Schätzer geschätzte Steigungsparameter von Null verschieden.

j) Beim multiplen linearen Regressionsmodell ist das volle Modell, welches alle Prädiktoren berücksichtigt, immer optimal hinsichtlich der erklärten Varianz auf den Trainingsdaten.

Aufgabe 2 Deskriptive Kennzahlen **(2+3+5=10 Punkte)**

Sie dürfen die Formeln auf Seite 8 verwenden.

a) Sie haben die Beobachtungen

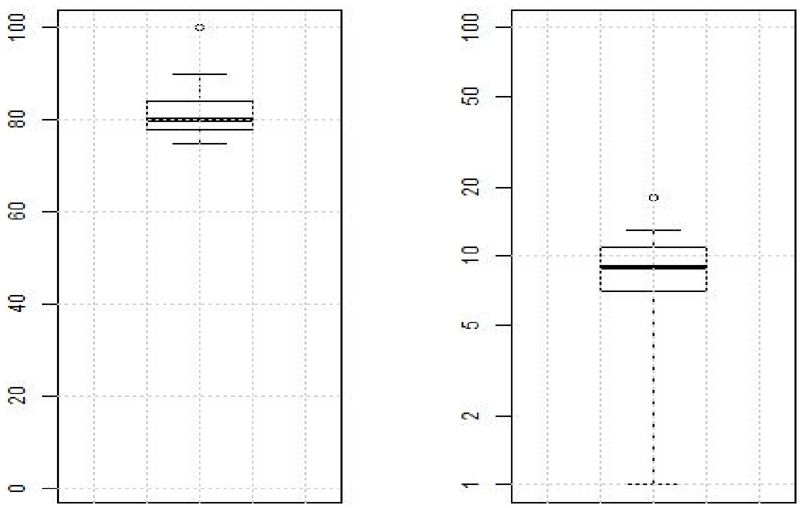
$$x = (-12, 5, 8, 1, -1, -25, 0, -2)$$

vorliegen. Berechnen Sie daraus das $\alpha-$getrimmte und das $\alpha-$winsorisierte Mittel für $\alpha = 0.25$.

b) Berechnen Sie die Kurtosis der Stichprobe.

HINWEIS: Für Beobachtungen $X = (X_1, ..., X_n)$ seien $Y_i := (X_i - \bar{X}_n)/\hat{\sigma}_X^2$. Die Kurtosis ist dann $\bar{(}Y_n^4) - 3$.

c) Interpretieren Sie die folgenden Boxplots. Beschreiben Sie dabei exakt, was jede horizontale Linie in den Boxplots aussagt. Vergleichen Sie die beiden Verteilungen hinsichtlich Symmetrie und Streuung.



HINWEIS: Beachten Sie die unterschiedlichen y-Achsen-Skalierungen!

Aufgabe 3 Konzentrationen **(4+2+1=7 Punkte)**

Gegeben seien vier Gruppen von Personen mit einem kumulierten Vermögen von 2000 (Angabe in Tausend Euro). Niemand von den 60 Personen der ersten Gruppe, die sich ein Vermögen von 60 teilen, hat Schulden, jedoch besitzen alle weniger als die 100 Personen, die eine weitere Gruppe bilden. Sie kommen insgesamt auf 900, was 400 mehr ist als das Gesamtvermögen der Gruppe von 25 Personen, die aber individuell jeweils mehr besitzen als jede der besagten 100 Personen, aber weniger als die bisher ungenannten übrigen Personen.

a) Bestimmen Sie für je eine Person aus jeder Gruppe jeweils eine Ober- und Untergrenze für das individuelle Vermögen.

b) Bestimmen Sie weiterhin die maximale Anzahl an Personen aus der dritten Gruppe (die 25 Personen), die mindestens 75 besitzen können.

c) Wie groß ist die maximale Anzahl an Personen, die gar nichts besitzen?

Hinweis: Ist a) nicht gelungen, so rechnen Sie in b) und c) mit den falschen Grenzen $\{1, 2\}$ für Gruppe 1, $\{3, 4\}, \{5, 6\}, \{7, 8\}$ für Gruppen 2, 3 bzw. 4.

Aufgabe 4 Definitionen, Sätze und Beweise **(3+4+2=9 Punkte)**

a) Formulieren Sie (ohne Beweis) den Satz von der totalen Wahrscheinlichkeit und beweisen Sie damit den Satz von Bayes.

b) Beweisen Sie mit der Tschebyschev-Ungleichung das schwache Gesetz der Großen Zahlen für reellwertige Zufallsvariablen $X_i \sim P$ i.id. mit Erwartungswert μ und Varianz σ.

c) Geben Sie eine statistische Anwendung beim Schätzen, beim Testen oder bei der Regression an, wo man dieses Gesetz verwendet und skizzieren Sie, wozu man es braucht.

Aufgabe 5 Abhängigkeit und Unabhängigkeit (4+4=8 Punkte)

Gegeben seien zwei Zufallsvariablen X und Y, die jeweils die Ausprägungen 1 und (-1) annehmen können. Wir haben Realisationen des Paares (X, Y) erhoben, allerdings sind in der Kontingenztabelle die inneren Zellen versehentlich gelöscht worden.

$Y \setminus X$	-1	1	$\sum$
-1			7
1			4
$\sum$	8	3	

a) Befüllen Sie die Tabelle mit den Variablen $N_{-1,-1}$, $N_{1,-1}$, $N_{-1,1}$, $N_{1,1}$. Bestimmen Sie eine Wertebelegung von $N_{i,j}$ so, dass maximale Abhängigkeit bzw. bestmögliche Unabhängigkeit zwischen X und Y herrscht. Verwenden Sie das Abstandsmaß χ, das Sie beim Unabhängigkeitstest verwenden.

b) Sie haben reellwertige Zufallsvariablen T_X und T_Y, die die Temperatur an zwei ausgewählten Orten X und Y beschreiben. Sie haben nun n Messungen in Grad Celsius durchgeführt und herausgefunden, dass X und Y positiv korreliert sind, einmal gemessen durch den Pearson-, einmal durch den Spearman-Korrelationskoeffizienten.

Welche Eigenschaft sichert Ihnen, dass diese Abhängigkeit nicht verloren geht, wenn Sie auf die Fahrenheit-Skala wechseln? Beweisen Sie für den Spearman-Korrelationskoeffizienten, dass diese Eigenschaft hier vorliegt.

Hinweis: Ist t die Temperatur in Grad Celsius, so ist $32 + 1.8t$ die Temperatur in Fahrenheit.

Aufgabe 6 Diskrete Verteilungen (4+5=9 Punkte)

Sei X eine $\{0,2\}-$, Y eine $\{-2,-1\}-$ und Z eine $\{-1,0\}-$wertige Zufallsvariable. Lesen Sie sorgfältig, was in den folgenden beiden Tabellen gegeben ist:

$P(X=i,Y=j,Z=1)$				$P(X=i,Y=j \mid Z=0)$		
$\downarrow j, i \rightarrow$	0	2		$\downarrow j, i \rightarrow$	0	2
-1	0.2	0.3		-1	0	0.15
-2	0.05	0.25		-2	0.05	0

a) Berechnen Sie $\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[Z]$.

b) Berechnen Sie $\mathbb{E}[Y^2]$ und $\mathbb{E}[XYZ]$.

Aufgabe 7 Qualität von Schätzern **(4+6+4=14 Punkte)**

a) Definieren Sie den MSE eines Schätzers. Beweisen Sie überdies dessen Bias-Varianz-Zerlegung.

b) Seien $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.id.. Herr B.Est meint, Varianzen könne man besser schätzen als mit den in der Vorlesung gegebenen Schätzern

$$S_{n,m} := \frac{1}{m} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2, \quad m \in \{n-1, n\}.$$

Sei allgemein $S_n^\alpha := \alpha S_{n,n}$. Bestimmen Sie

$$\hat{\alpha} := \underset{\alpha}{\operatorname{argmin}}(MSE(S_n^\alpha)).$$

Hinweis: Beachten Sie: Ein unverzerrter Schätzer für die Varianz ist

$$\frac{1}{n-1} \sum_i (x_i - \bar{x}_n)^2.$$

Ferner ist

$$\mathcal{L}(\chi_n^2) = \sum_{i=1}^{n} X_i^2 \text{ mit } \; X_i \sim \chi_1^2 \; \text{ i.id.},$$

$$\operatorname{Var}(\chi_1^2) = 2\sigma^4.$$

c) Seien $X_1, ..., X_n$ unabhängig identisch $\mathcal{N}(\mu, \sigma^2)$−verteilt mit $\sigma^2 := 4\mu \in [1, 5]$. Wie groß muss n sein, damit in jedem Falle die Wahrscheinlichkeit, dass der Abstand von $\bar{X}_n$ und $\mathrm{E}[\bar{X}_n]$ weniger als 10 Prozent von $\mathrm{E}[\bar{X}_n]$ ausmacht, mindestens 97.5 Prozent ist?

Aufgabe 8 Schätzen und Konfidenzintervalle (4+2+3+3+4=16 Punkte)

Umberto und Kunigunde haben einen Würfel mit inhomogener Gewichtsverteilung, was dazu führt, dass es beim Würfeln nicht mehr gleichwahrscheinlich ist, auf welcher Seite der Würfel liegen bleibt. Sie würfeln so lange, bis sie zum dritten Mal eine sechs sehen. Für die Wartezeit der Fehlversuche bis zum ersten Treffer in diesem Kontext betrachten wir $X \sim \text{Geom}(p)$, d.h., die diskrete Verteilung mit Wahrscheinlichkeitsfunktion $P(X = k) = p(1-p)^k, k \in \mathbb{N}_0$. Es gilt für $q := (1-p)$, dass

$$\mathbb{E}(X) = \frac{q}{p}, \text{Var}(X) = \frac{q}{p^2}.$$

Ferner ist für $X_1, \ldots, X_m \overset{\text{u.i.v.}}{\sim} \text{Geom}(p)$ die Summe

$$S_m = \sum_{i=1}^{m} X_i$$

NegBinom$(m, p)-$verteilt.

a) Konstruieren Sie allein auf Basis der Zahl der Würfelwürfe bis zur jeweils nächsten 6 einen Momentenschätzer für die unbekannte Wahrscheinlichkeit p_6.

b) Skizzieren Sie, wie Sie den ML-Schätzer für p_6 berechnen würden (nur beschreiben, d.h., die Rechnungen führen Sie nicht durch).

c) Umberto und Kunigunde haben nun den Momentenschätzer $\hat{p}_6^{mom}$ aus a). Undine und Sieghold stöbern heimlich in ihren Aufzeichnungen, finden jedoch nur den berechneten Wert für diesen Schätzer, wissen aber nicht, bis zur wievielten gesehenen sechs gewürfelt wurde. Bestimmen Sie den Momentenschätzer für die gesuchte Anzahl in Abhängigkeit von $\hat{p}_6^{mom}$.

d) Untersuchen Sie $\hat{p}_6^{mom}$ auf Konsistenz.

e) Seien nun für alle 6 mögliche Augenzahlen jeweils p_j die wahren Wahrscheinlichkeiten und r_{jn} die relativen Häufigkeiten basierend auf n Würfen. Sie haben Konfidenzintervalle

$$I_j := [r_{jn} - \delta_{jn}^-, r_{jn} + \delta_{jn}^+]$$

für alle j.

Erläutern Sie, wie man die δ_{jn} mit dem Bootstrapverfahren gewinnen kann.

Aufgabe 9 Hypothesentests **(3+2+4=9 Punkte)**

a) Darf man bei einem statistischen Test H_0 und H_1 immer vertauschen? Begründen Sie dies.

b) Wann heißt ein statistischer Test unverzerrt? Was bedeutet dies anschaulich?

c) Jemand würfelt so lange, bis zum 100. Mal eine Sechs erscheint. Da 1000 Würfe benötigt wurden, glaubt dieser Mensch, dass der Würfel unfair ist und will einen Gaußtest durchführen. Ist dies zulässig? Betrachten Sie dazu die (beobachtete) durchschnittliche Zahl der Würfelwürfe, bis jeweils zum 100. Mal eine 6 eintritt. Plausibilisieren Sie, ob dies auch effizient ist, um mit möglichst wenig Würfen eine Aussage zu vorgegebener Signifikanz treffen zu können.

Aufgabe 10 Rätsel lösen... **(2+2+4=8 Punkte)**

Angenommen, es gibt eine besonders mürrische Sphinx, die vorbeikommende Reisende nicht ziehen lässt, wenn sie ein Rätsel lösen konnten, sondern die Lösung dreier Rätsel verlangt. Die einzige Person, die das geschafft hat, hat Nachforschungen angestellt und stellt sein Ergebnis in folgendem Rätsel dar:

„*40 der Personen, die das erste Rätsel lösen konnten, scheiterten an dem zweiten, während sechs derer, die auch dies bewältigten, das letzte Rätsel nicht überlebten. 222 Personen sprachen insgesamt mit der Sphinx.*"

Sie betrachten die Verteilung für die Zahl S der richtigen Antworten, also eine Verteilung auf $\{0, 1, 2, 3\}$. Nach Betrachtung der relativen Häufigkeiten meinen Sie, dass eine $Bin(3, 0.1)-$Verteilung passen könnte.

a) Stellen Sie explizit Ihre Hypothese auf und nennen Sie den benötigten statistischen Test.

b) Berechnen Sie aus den Daten die richtigen relativen Häufigkeiten.

c) Führen Sie den statistischen Test zum Niveau $\alpha = 0.05$ durch.

Aufgabe 11 Lineare Regression **(3+9+2+3+3=20 Punkte)**

a) Seien X_1, X_2 erklärende Variablen in einem linearen Regressionsmodell. Unsere beobachteten Werte für X_1 bzw. X_2 sind $(2, 3, 8)$ bzw. $(-2, -1, 6)$. Stellen Sie die Regressormatrix für folgende Szenarien explizit auf:

- X_1, X_2 als erklärende Variablen, kein Achsenabschnitt
- Das einfache lineare Regressionsmodell bzgl. X_2
- Polynomiale Regression bis Grad 3 für X_1

b) Betrachten Sie folgenden R−Output:

```
Call:
lm(formula = Y ∼., data = D)

Residuals:
Min       1Q    Median    3Q     Max
-6.347    -1.700   0.451    1.590    6.881

Coefficients:
             Estimate   Std. Error   t value    Pr(> |t|)
(Intercept)   2.9278     0.8640       3.389     0.001268 **
X1            0.6593     0.3993       1.651     0.104094
X2           -0.3702     0.3279      -1.129     0.263517
X3           -1.4822     0.3673      -4.035     0.000162 ***
X4           -0.9753     0.3680      -2.650     0.010353 *
X5B           4.1965     1.0142       4.138     0.000115 ***
X5C          -5.9071     0.9972      -5.924     1.82e-07 ***
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.689 on 58 degrees of freedom
Multiple R-squared: 0.8003, Adjusted R-squared: 0.7796
F-statistic: 38.73 on 6 and 58 DF, p-value: < 2.2e-16
```

(i)) Welche Koeffizienten sind signifikant? Begründung? Was bedeutet anschaulich der Koeffizient von X2?

(ii) Zeigen Sie anhand der anderen vorliegenden Informationen, wie Sie auf den t-Wert für X4 kommen. Welche Hypothese steckt dahinter? Wie bekommen Sie für die Variable X3 den angegebenen p-Wert?

(iii) Wie interpretieren Sie die letzten beiden Koeffizienten in der Tabelle? Interpretieren Sie die vorletzte Zeile des Outputs.

(iv) Wie viele Beobachtungen lagen eigentlich vor?

c) Sie haben zwei Modelle und erhalten für das erste Modell einen AIC von 2.4 und einen BIC von 3.2, während Sie für das zweite Modell einen AIC von 3.2 und einen BIC von 2.9 bekommen. Wann würden Sie sich eher für das erste, wann eher für das zweite entscheiden? Begründen Sie Ihre Antwort.

d) Motivieren Sie, was der Vorteil sparsamer Modellwahl (viele Koeffizienten sind Null) ist. Schreiben Sie das Optimierungsproblem der linearen Regression mit Ridge- und Lasso-Strafterm auf.

e) Visualisieren Sie Ihre Strafterme für $p = 1$. Welcher Strafterm wäre eigentlich optimal? Welche Vorteile haben der Ridge- und Lasso-Strafterm hinsichtlich numerischer Optimierung, welcher durch den eigentlich optimalen Strafterm verloren ginge?

Aufgabe 12 Generalisierte lineare Modelle (2+2+1+2+3=10 Punkte)

a) Definieren Sie ein allgemeines generalisiertes lineares Modell. Was versteht man unter einer Linkfunktion? Wie lautet diese für das gewöhnliche lineare Regressionsmodell (Aufgabe 11)?

b) Nehmen Sie an, Ihre abhängige Variable Y nimmt die Werte ,,Vampir" und ,,Banshee" an. Sie haben p erklärende Variablen $X_1, ..., X_p, p \geq 2$, zur Verfügung. Stellen Sie für diesen konkreten Fall Ihr Probit-Modell mit allen erklärenden Variablen auf.

c) Sie haben nun die Koeffizienten für Ihr Modell aus b) vorliegen. Der Koeffizient β_p von X_p ist 7.22. Wie interpretieren Sie diesen Koeffizienten auf der Prädiktorskala, also auf Ebene $X\beta$? Welchen Effekt hat eine Vergrößerung von β_p bei sonst gleichen Koeffizienten und gleichen Prädiktoren X_i auf Ebene der Wahrscheinlichkeiten?

d) Angenommen, für drei neue Beobachtungen (d.h., alle X-Werte sind bekannt, der Y-Wert jeweils unbekannt) sind die berechneten Werte von $X\beta$ 0.26, 0.61 und 0.98. Für welche der Beobachtungen sagen Sie ,,Vampir" bzw. ,,Banshee" voraus?

e) Was ist der Unterschied des Logit-Modells zum Probit-Modell? Was ist beim Ersteren die Link-Funktion? Kann es noch andere geeignete Link-Funktionen für unser Problem aus b) geben?