FFHS
Fernfachhochschule
Schweiz
Mitglied der SUPSI

Paper-Prüfung
Fernfachhochschule Schweiz

WS/WS PiBS - Modul Wahrscheinlichkeit und Statistik

Modulprüfung-FS 22

FFHS
Fernfachhochschule
Schweiz
Mitglied der SUPSI

Aufg.	Thema	Anzahl Punkte	
		Max.	Erreicht
1	Deskriptive Statistik I	9	
2	Deskriptive Statistik II	8	
3	Stetige Verteilungen	5	
4	Konfidenzintervalle	6	
5	Lineare Regression	11	
6	Hypothesentest I	4	
7	Hypothesentest II	11	
Total		54	

FFHS
Fernfachhochschule
Schweiz
Mitglied der SUPSI

Aufgabe 1 Deskriptive Statistik I (1+1+3+3+1 Punkte)

910 zufällig ausgewählte, registrierte Abstimmberechtigte in Tampa (Florida, USA) wurden nach ihrer Meinung gefragt, ob illegalen Einwanderern erlaubt sein sollte: i) ihre Anstellung zu behalten und sich für die amerikanische Staatsangehörigkeit zu bewerben ii) ihre Anstellung temporär als Gastarbeiter behalten zu dürfen, aber sich nicht für die amerikanische Staatsangehörigkeit zu bewerben oder iii) ihre Anstellung zu verlieren und des Landes verwiesen werden sollten. Für die Datenanalyse wurden die Ergebnisse des Fragebogens in einer Kontingenztabelle nach der genannten politischen Zugehörigkeit erfasst. Die Befragten wurden nach folgenden politischen Kategorien erfasst: Konservativ, Liberal, Moderat

Antwort	Konservativ	Liberal	Moderat	Gesamt
Antrag amerik. Staatsangeh.	57	101	120	278
Gastarbeiter	121	28	113	262
Landesverweisung	179	45	126	350
Nicht sicher	15	1	4	20
Gesamt	372	175	363	910

a) Wieviel Prozent der eingetragenen Wähler in Tampa (Florida, USA) bezeichnen sich als Konservative? (1 Pkt.)

b) Wieviel Prozent der eingetragenen Wähler in Tampa (Florida, USA) befürworten die Möglichkeit der Einbürgerung? (1 Pkt)

c) Wieviel Prozent der konservativen Wähler sind auch einem Antrag auf amerik. Staatsangehörigkeit gegenüber positiv eingestellt? Wieviel sind es bei Moderaten und Liberalen? (3 Pkt.)

d) Beeinflusst die politische Einstellung/Ausrichtung ihre Einstellung zur Immigration? Begründen und diskutieren Sie anhand der Kontingenztabelle qualitativ, ohne einen χ^2-Unabhängigkeitstest! (3 Pkt.)

e) Könnten Sie sich eine Ergänzung in der Erhebung vorstellen, die ihnen den Zusammenhang aus c) weitergehend erklären könnte? (1 Pkt.)

Aufgabe 2 Deskriptive Statistik II (2+3+1+2 Punkte)

Der Datensatz *iris* enthält Daten zu der Länge und Breite zweier verschiedener Blättertypen (Kelchblättern (engl. *sepal*) und Kronenblätter (engl. *petal*) der drei Irisblumenarten *iris versicolor*, *Iris setosa* und *Iris virginica*. Der Datensatz *iris* ist bereits in R integriert und kann direkt mit `iris` geladen werden. Alternativ können Sie ihn auch über das beigefügte `iris.csv` importieren.

a) Importieren/Laden Sie den Datensatz und benennen Sie ihn `data` und verschaffen Sie sich einen Überblick über die Struktur des Datensatzes. Wieviel Variablen und Observationen enthält er? (2 Pkt.)

b) Bestimmen Sie den arithmetischen Mittelwert, den Median und das 75%-Quantil der Variablen `Sepal.Length`. (3 Pkt.)

c) Überprüfen Sie, dass es sich bei der Variablen `Species` um einen `factor` handelt. (1 Pkt.)

d) Erzeugen Sie einen Boxplot für `Sepal.Length` für die drei Irisarten in einem Diagramm. Was beobachten Sie für die verschiedenen Arten? (2 Pkt.)

Aufgabe 3 Stetige Verteilungen (3+2 Punkte)

Der elektrische Widerstand X von elektronischen Bauteilen ist normalverteilt mit $\mu = 200\Omega$ und $\sigma = 10\Omega$.

a) Wie viel Prozent der Bauteile halten einen Mindestwert von 190Ω ein? (3 Pkt.)

b) Welcher Widerstandswert wird nur von 2% aller Bauteile überschritten? (2 Pkt.)

Anm.: Sie können statt mit einer Tabelle für die Normalverteilung auch mit R arbeiten.

Aufgabe 4 Konfidenzintervalle (2+2+2 Punkte)

Eine allgemeine Sozialumfrage in der Schweiz zu Gesundheitsfragen stellte folgende Frage: "Für wieviel der vergangenen 30 Tage empfanden Sie ihre mentale Gesundheit (dazu zählt Stress, Depression, emotionale Probleme) als nicht gut?- Basierend auf 1151 Schweizer Einwohnern kommt die Umfrage auf ein Konfidenzintervall von 3.40 bis 4.24 Tagen bei einem Konfidenzniveau von 95% im Jahr 2015.

a) Was bedeutet 95% Konfidenzniveau? Erklären Sie! (2 Pkt.)

b) Angenommen, die Forscher erhöhen den Wert von 95% auf 99%. Wird das neue Konfidenzintervall dann kürzer oder länger ausfallen? (2 Pkt.)

c) Würde eine neue Umfrage mit 500 Schweizer Einwohnern durchgeführt werden, wäre dann das Konfidenzintervall kürzer als bei der vorliegenden Umfrage oder länger? Begründen Sie ihre Antwort! (2 Pkt.)

FFHS
Fernfachhochschule
Schweiz
Mitglied der SUPSI

Aufgabe 5 Lineare Regression (3+2+2+2+2 Punkte)

Wie tragen Umweltfaktoren zur Pflanzendiversität auf einer Insel bei? Der Datensatz `Plants.xlsx` liefert Informationen zu einer charakteristischen Anzahl britischer Inseln (jedoch nicht Grossbritannien und Irland) einschliesslich der Angaben zum Längengrad (latitude), Fläche (area) und Entfernung zu Grossbritannien (distance). Wir wollen nun eine lineare Regressionsanalyse durchführen, um zu bestimmen, ob Inseln mit einer grösseren Fläche auch eine grössere Anzahl von Pflanzenarten beheimaten.

a) Importieren Sie das Excel-File `Plants.xlsx` in R. Identifizieren Sie, ob es sich um kategoriale oder quantitative Variablen handelt. Handelt es sich um eine kategorische Variable, dann geben Sie an ob sie ordinal, nominal oder eine id-Variable (*identifier variable*) ist. Wenn die Variable quantitativ ist, geben Sie an, ob sie diskret oder kontinuierlich ist. Anm.: Im Dokument `PlantsDataDictionary.pdf` finden Sie ein Beschreibung der Variablen. (3 Pkt.)

b) Erzeugen Sie ein Streudiagramm, um das Verhältnis von Fläche (`area`) zu Anzahl (`number`) von Pflanzenarten auf den Britischen Inseln darzustellen. (2 Pkt.)

c) Bestimmen Sie die Regressionsgleichung der linearen Einfachregression, um die Anzahl von Pflanzenarten in Abhängigkeit der Fläche zu beschreiben. (2 Pkt.)

d) Erzeugen Sie einen Residuen- (Residual vs Fitted) und eine Q-Q-Plot (Normal-Q-Q) um potentielle Abweichungen von der Annahme einer linearen Regression aufzufinden. (2 Pkt.)

e) Was würden Sie als angehender Informatiker:in/Data Scientist sagen? Ist die lineare Regression die geeignete Methode, um den Zusammenhang zwischen Fläche und Anzahl Pflanzenarten zu beschreiben? Welche Kriterien und Beobachtungen ziehen Sie heran? Begründen Sie ihre Antworten. (2 Pkt.)

FFHS
Fernfachhochschule
Schweiz
Mitglied der SUPSI

Aufgabe 6 Hypothesentest I - Verständnisfragen I (2 + 2 Punkte)

Wahr oder falsch? Begründen Sie ihre Antwort! (Korrekte Antwort+korrekte Begründung 2 Pkt., korrekte Antwort + falsche Begründung 0 Pkt., falsche Antwort -1 Pkt. , keine Antwort 0 Pkt.)

a) Der p-Wert gibt die Wahrscheinlichkeit an, dass die Prüfgrösse (z.B. der Mittelwert) in den vorher festgelegten Ablehnungsbereich fällt. (2 Pkt.)

b) Man begeht keinen Fehler (α oder β (1. oder 2.Art)), wenn man die Nullhypothese verwirft, wenn sie falsch ist. (2 Pkt.)

Aufgabe 7 Hypothesentest II - Statistische Unabhängigkeit(1+2+1+2+2+3 Punkte)

In einer Studie untersuchten Forscher den Zusammenhang zwischen dem Konsum von koffeinhaltigem Kaffee und dem Risiko von Depressionen bei Männern. Sie sammelten Daten von 50.739 Männern, die zu Beginn der Studie im Jahr 1996 keine Depressionssymptome aufwiesen. Sie verfolgten die weitere Entwicklung der Männer bis 2006. Die Forscher verwendeten Fragebögen, um Daten über den Konsum von koffeinhaltigem Kaffee zu erheben. Sie fragten jeden Studienteilnehmer nach ärztlich diagnostizierten Depressionen und nach der Einnahme von Antidepressiva. Die folgende Kontingenztabelle zeigt die Verteilung des Auftretens von Depressionen nach der Menge des koffeinhaltigen Kaffeekonsums.

	Konsum koffeinhaltigen Kaffees					
Diagn. Depression	1 Tasse/ Woche oder weniger	2-6 T./ Woche	1 T./ Tag	2-3 T./ Tag	4 T./ Tag oder mehr	Gesamt
Ja	670	-	905	564	95	2607
Nein	11545	6244	16329	11726	2288	48132
Gesamt	12215	6617	17234	12290	2383	50739

a) Welchen Test führen Sie durch, um die statistische Unabhängigkeit der Merkmale Kaffeekonsum und Auftreten einer Depression zu untersuchen? (1 Pkt.)

b) Formulieren Sie die Nullhypothese H_0 zu diesem Test.(2 Pkt.)

c) Wie gross ist der Anteil von Männern, die insgesamt an Depression leiden? (1 Pkt.)

d) Wie viel Freiheitsgrade berechnen Sie aus der Kontingenztabelle? Wie bestimmen Sie sie? (2 Pkt.)

e) Zeigen Sie durch Rechnung, dass der theoretische Wert des fehlenden Wertes in der Tabelle (siehe Spalte 3) ≈ 340 beträgt. (2 Pkt.)

f) Sie erhalten beim Auswerten der Tabelle eine $\chi^2 = 20.93$. Sie testen für $\alpha = 0.05$, wie lautet der kritische Wert? Verwerfen Sie die Nullhypothese? Begründen Sie! (3 Pkt.)