Probeklausur Frühjahr 2009
Statistik für Human- und Sozialwissenschaftler

Zugelassene Hilfsmittel: Taschenrechner aller Art, Fremdsprachenwörterbücher.

Verlangt und gewertet werden **vier der folgenden fünf** Aufgaben. Lösungsschritte und Teilergebnisse sind ausreichend zu begründen. Eine Angabe des Endergebnisses allein genügt nicht.

Aufgabe 1

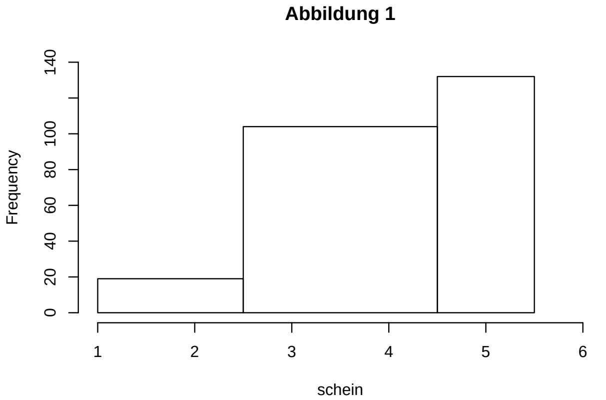
a) Nach einer Studie des Hochschul-Informations-Systems (HIS) brechen an Universitäten ungefähr 30 Prozent der Studenten und an Fachhochschulen ungefähr 22 Prozent der Studenten ihr Studium vorzeitig ab.

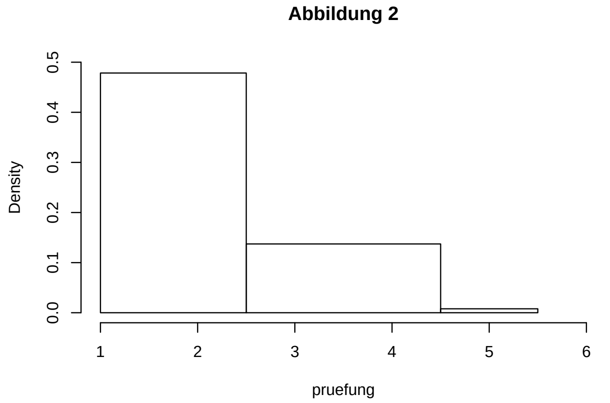
- **a_1)** Warum vergleicht man hier prozentuale Anteile und nicht die absoluten Zahlen der Abbrecher ?
- **a_2)** Was würden Sie vermuten: Handelt es sich bei dieser Studie um eine kontrollierte Studie mit Randomisierung, um eine kontrollierte Studie ohne Randomisierung oder um eine Beobachtungsstudie ? Begründen Sie ihre Antwort.
- **a_3)** An Fachhochschulen ist der Studienverlauf stärker strukturiert und der Praxisbezug intensiver als an Universitäten. Kann man aus der obigen Studie schließen, dass dies in kausalem Zusammenhang steht mit der geringeren Abbrecherquote ?

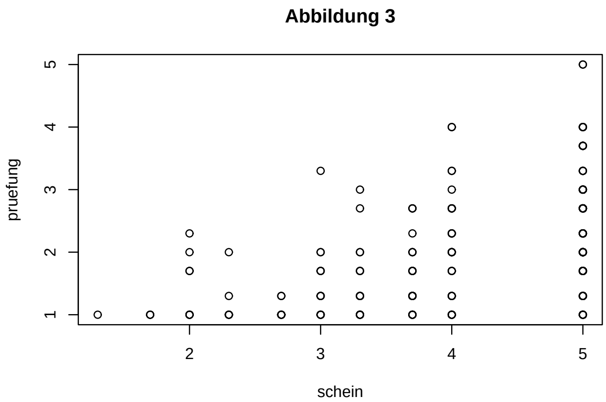
b) Beschreiben Sie **kurz**, was man bei einer Umfrage unter einer Verzerrung durch Auswahl (sampling bias) versteht.

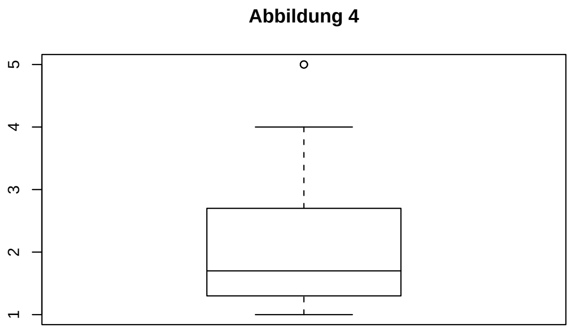
c) Bei Umfragen steht man häufig vor dem Problem, dass ein Teil der Befragten die Antwort verweigert. Warum kann man das Problem nicht dadurch lösen, dass man solange Leute zur Befragung (zufällig) auswählt, bis man eine gewisse Mindestanzahl an Leuten gefunden hat, die bereit sind, an der Umfrage mitzuwirken ?

Aufgabe 2









a) Das Säulendiagramm in Abbildung 1 beschreibt die Noten von $n = 255$ Studenten in der Scheinklausur zur Vorlesung "Statistik I für WiWi".

a$_1$) Inwieweit ist die Darstellung in diesem Säulendiagramm irreführend ?

a$_2$) Stellen Sie die Daten in einem Histogramm so dar, dass die Flächeninhalte der einzelnen Balken (aufgrund von Problemen beim Ablesen der Werte in Abbildung 1 eventuell nur ungefähr) proportional zur Anzahl der Datenpunkte in den zugrundeliegenden Intervallen sind.

b) Das Histogramm in Abbildung 2 beschreibt die Noten von $n = 255$ Studenten in der Diplom-Vorprüfung zur Vorlesung "Statistik I für WiWi". Bestimmen Sie mit Hilfe dieses Histogramms (approximativ) die Anzahl der Studenten, die in der Prüfung eine Note besser als 2.5 hatten (d.h., die als Note eine 1.0, 1.3, 1.7, 2.0 oder 2.3 hatten).

c) Im Streudiagramm in Abbildung 3 sind die Noten in der Scheinklausur gegen die Noten in der Diplom-Vorprüfung abgetragen. Ist die Korrelation zwischen diesen Noten größer oder kleiner als Null ? Begründen Sie ihre Antwort.

d) Der Boxplot in Abbildung 4 beschreibt die Noten in der Diplom-Vorprüfung zur Vorlesung "Statistik I für WiWi". Wie groß ist der Median und wie groß ist der Interquartilabstand dieser Noten ?

Aufgabe 3

Die reelle Zufallsvariable X nehme die Werte 1, 2 und 6 mit den Wahrscheinlichkeiten 1/2, 1/6 bzw. 1/3 an. Die Zufallsvariable Y sei stetig verteilt mit Dichte

$$f : \mathbb{R} \to \mathbb{R}_+, \quad f(x) = \begin{cases} 20 \cdot x^3(1-x) & \text{für} \quad 0 \le x \le 1, \\ 0 & \text{für} \quad x < 0 \text{ oder } x > 1. \end{cases}$$

X und Y seien unabhängig.

a) Bestimmen Sie $\mathbf{E}X$ und $V(X)$.

b) Bestimmen Sie $\mathbf{E}Y$ und $V(Y)$.

c) Bestimmen Sie $\mathbf{E}(X+Y)$ und $V(X+Y)$. An welcher Stelle benötigen Sie hier die Unabhängigkeit von X und Y ?

Aufgabe 4

Ein Glücksrad bleibt nach dem Drehen rein zufällig auf einem von insgesamt 50 Feldern stehen. Bleibt es auf einem der 10 blau gefärbten Felder stehen, so wird ein Gewinn von 5 Euro ausgezahlt. Bleibt es auf einem der 5 grün gefärbten Felder stehen, so wird ein Gewinn von 10 Euro ausgezahlt. Und bleibt es auf dem *einzigen* roten Feld stehen, so wird ein Gewinn von 100 Euro ausgezahlt. Auf den übrigen 34 weiß gefärbten Feldern wird kein Gewinn ausgezahlt.

a) Wie groß ist der Gewinn "im Mittel", und wie groß ist die "mittlere quadratische Abweichung" zwischen dem zufälligen Gewinn und dem Gewinn "im Mittel" ?

b) Für einmaliges Drehen verlangt der Besitzer des Glücksrads einen Einsatz von 5 Euro. Damit beträgt sein Verdienst bei einmaligen Drehen $Y = 5 - X$, wobei X der ausgezahlte Gewinn ist. Wie groß ist sein Verdienst "im Mittel", und wie groß ist die "mittlere quadratische Abweichung" zwischen dem zufälligen Verdienst und dem Verdienst "im Mittel"?

c) Der Besitzer betreibt sein Glücksrad einen Monat lang auf einem Jahrmarkt. In dieser Zeit drehen dabei $n = 6000$ Personen (unbeeinflusst voneinander) am Glücksrad. Bestimmen Sie mit Hilfe des Zentralen Grenzwertsatzes eine untere Schranke für den Verdienst des Besitzers in diesem Monat, die (ungefähr) mit Wahrscheinlichkeit 0.95 überschritten wird.

Hinweise:

Seien $Y_1, \ldots, Y_n$ unabhängige Zufallsvariablen mit $\mathbf{P}_{Y_i} = \mathbf{P}_Y$ $(i = 1, \ldots, n)$, wobei Y die in b) eingeführte ZV ist. Bestimmen Sie mit Hilfe des Zentralen Grenzwertsatzes $x \in \mathbb{R}$ so, dass gilt:

$$\mathbf{P}[\sum_{i=1}^{n} Y_i \ge x] \approx 0.95.$$

Verwenden Sie dabei, dass für die Verteilungsfunktion Φ der $N(0,1)$-Verteilung gilt: $\Phi(1.65) \approx 0.95$ und $\Phi(-1.65) \approx 0.05$.

Nach dem Zentralen Grenzwertsatz gilt

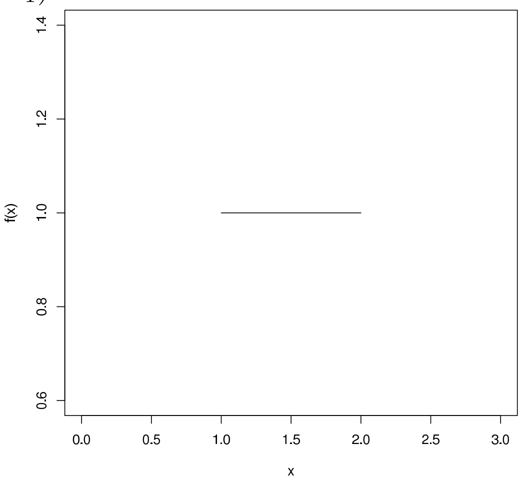
$$\mathbf{P}\left[\frac{1}{\sqrt{n} \cdot \sqrt{V(Y_1)}} \sum_{i=1}^{n} (Y_i - \mathbf{E}Y_1) \le x\right] \approx \Phi(x).$$

Aufgabe 5

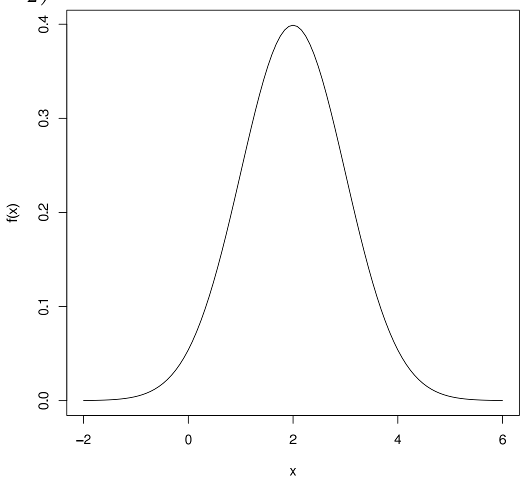
a) Angenommen, die Lotto-Zahlen beim Lotto 6 aus 49 werden unbeeinflusst voneinander jede Woche rein zufällig aus den Zahlen von 1 bis 49 ausgewählt.
Ist es dann wahrscheinlicher, dass nächste Woche die Zahlen $1, 9, 10, 15, 32, 37$ gezogen werden, oder dass nächste Woche die gleichen Zahlen wie in dieser Woche gezogen werden?

b) Welche der beiden Dichten

b$_1$)



b$_2$)



gehört zu einer Normalverteilung? Was können Sie über den Erwartungswert dieser Normalverteilung aussagen?

c) Erläutern Sie jeweils kurz (und evtl. auch anschaulich) die Aussagen des

c$_1$) empirischen Gesetzes der großen Zahlen,

c$_2$) starken Gesetzes der großen Zahlen,

c$_3$) zentralen Grenzwertsatzes.