Übungsaufgaben zum selber rechnen - Statistik I

Auf den folgenden Seiten finden Sie Übungsaufgaben zu den in der Vorlesung Statistik I behandelten Themenbereichen (die Aufgaben sind entsprechend der Themenbereiche nummeriert).

- Wir empfehlen die Aufgaben zur **Nachbereitung** der Themenbereiche zu nutzen, **um die statistischen Konzepte und Verfahren einzuüben**.
- Die Aufgaben **sollen zu einem tieferen Verständnis beitragen und auf die Prüfung vorbereiten**! (Allerdings liegt - anders als in der Modulprüfung - den Aufgaben keine Formelsammlung bei!)
- Die Aufgaben zu den einzelnen Themenbereichen gliedern sich in
  1. Aufgaben, die händisch zu berechnen sind
  2. Aufgaben, die mit Hilfe des Statistikprogramms „R“ zu berechnen sind

  Hinweise zur den angegebenen Lösungen der Aufgaben, die mit „R“ zu berechnen sind:

  a) Die Lösungen setzen voraus, dass bekannt ist, wie Datensätze geladen (und gespeichert) werden. Einen Überblick wie verschiedene Datensätze geladen und gespeichert werden findet sich im StudIP als PDF-Dokument.
  [Statistik I -> Dateien -> 03_Übungsveranstaltung -> „Tutorium8.pdf“]

  b) Die angegebenen Lösungen setzen voraus, dass der jeweilige Datensatz zuvor mit dem Befehl „attach“ aktiviert wurde. Dieser Befehl aktiviert ein Data Frame dauerhaft, so dass alle nachfolgend eingegebenen Befehle automatisch auf diesen angewendet werden. *Beispiel: attach(datensatzXY).*

  c) Die angegebenen Lösungen stellen lediglich eine Lösungsmöglichkeit dar. Es ist durchaus möglich mit Hilfe einer anderen Funktion / eines anderes Pakets auf dieselbe Lösung zu kommen!

Die Themenbereiche gliedern sich anhand der Prüfungsliteratur wie folgt:

Themenbereich 1: Messtheoretische Grundlagen und Skalenniveaus (Kapitel 5), Arten von Variablen (Kapitel 4.1-4.2)

Themenbereich 2: Deskriptive univariate Auswertung nominalskalierter, ordinalskalierter und kardinalskalierter Variablen und Standardwerte und z-Transformation (Kapitel 6-6.5)

Themenbereich 3: Kovarianz und Korrelation (Kapitel 15.1-15.3)

Themenbereich 4:Einfache lineare Regression (Kapitel 16)

Themenbereich 5:Partial- und Semipartialkorrelation (Kapitel 17.1-17.3)

Themenbereich 6:Multiple Korrelation (Kapitel 18.1-18.6)

Themenbereich 7: Wahrscheinlichkeitsrechnung und Kombinatorik (Kapitel 7)

1. Messtheoretische Grundlagen und Skalenniveaus und Arten von Variablen

1) Definieren Sie die Begriffe Nominalskala, Ordinalskala und Intervallskala

2) Mit Hilfe welches der gerade definierten Skalenniveaus (d.h. Nominalskala, Ordinalskala und Intervallskala) können die folgenden Merkmale sinnvoll gemessen werden?
    a) Einkommen
    b) Schulbildung
    c) Angabe des Berufs
    d) Geschlecht eines Kindes
    e) Alter
    f) Familienstand

3) Bestimmen Sie das Skalenniveau der folgenden Variablen:
    a) Wetter: 1 = Sonne, 2= Wolken, 3 = leicht bewölkt
    b) Reaktionszeit in Sekunden
    c) Zustimmung: 1 = stimme gar nicht zu, 2 = stimme eher nicht zu, 3 = stimme eher zu, 4 = stimme völlig zu
    d) Sexuelle Orientierung: 1 = heterosexuell, 2 = homosexuell, 3 = bisexuell
    e) Schulnoten: 1 = sehr gut, 2 = gut, 3 = befriedigend, 4 = ausreichend, 5 = mangelhaft, 6 = ungenügend
    f) Anzahl gelöster Aufgaben in einem Test: 0 (keine) bis 15 (alle)

4) Was unterscheidet eine diskrete Variable von einer stetigen?

5) Was unterscheidet eine qualitative Variable von einer quantitativen?

2. Deskriptive univariate Auswertung nominalskalierter, ordinalskalierter und kardinalskalierter Variablen

1) Die Befragung von 25 Studierenden hinsichtlich Ihrer Religionszugehörigkeit ergibt folgende Tabelle:

Christen (1)	Atheisten (2)	Buddhisten (3)	Muslime (4)	Hinduisten (5)	Andere (6)
12	5	3	3	1	1

Bestimmen Sie den Modus und erstellen Sie ein Säulendiagramm!

2) Für die Substichprobe der Christen wurde die Schulnote im Fach Religion erfasst. Es ergaben sich folgende Werte:

1 3 5 3 1 3 3 1 4 3 4 2

Bestimmen Sie den Modus und den Median, sowie Q1, Q3 und den Interquartilsabstand.

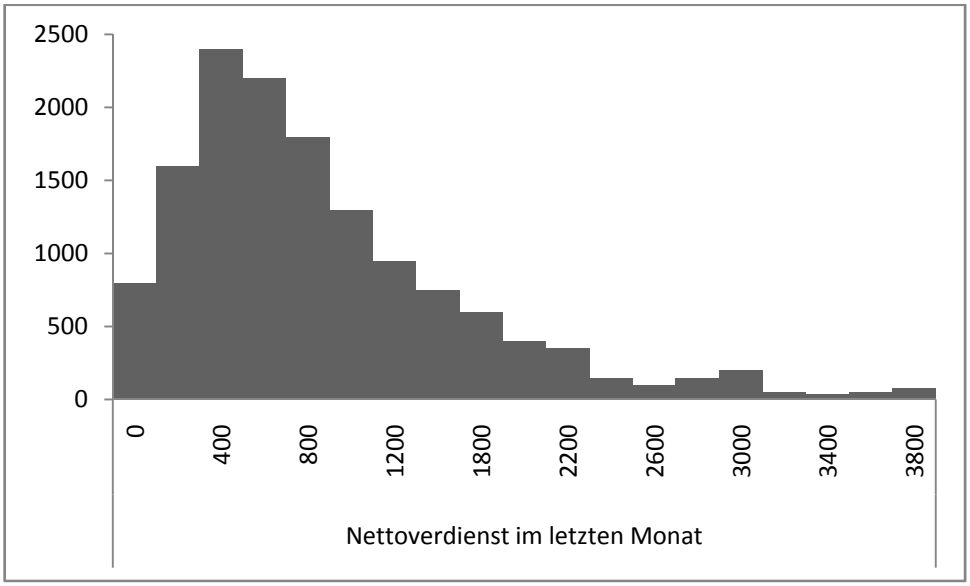
3) Mit Hilfe welcher Diagramme können kardinale Daten (intervallskalierte Daten) visualisiert werden? Nennen Sie 2 Diagrammtypen und charakterisieren Sie diese.

4) Die Befragung von 8 Grundschüler/innen hinsichtlich Ihres wöchentlichen Taschengeldes ergibt folgende Werte (in Euro)

3 5 3 3 4 2 8 4

Bestimmen Sie den Modus, den Median, das arithmetisches Mittel, die Varianz und die Standardabweichung und erstellen Sie ein Box-Whisker-Plot!

5) Im folgenden Diagramm ist der Nettoverdienst von Studierenden eingetragen.



Welche Eigenschaften treffen auf die vorliegende Verteilung zu?

- [ ] Mehrgipflig
- [ ] Symmetrisch
- [ ] Linksschief
- [ ] Rechtsgipflig
- [ ] Linkssteil
- [ ] asymmetrisch
- [ ] eingipflig
- [ ] rechtsschief

6) In Ihrem Studiengang wurde von einer kleinen Stichprobe die Schuhgröße erhoben. Sie bekommen die Daten in einer Datenmatrix:

	Schuhgröße
Proband	
1	36
2	38
3	40
4	38
5	42
6	46

a) Wählen sie ein geeignetes Maß der zentralen Tendenz sowie ein geeignetes Dispersionsmaß und berechnen Sie beides.
b) Ihre neu erworbenen Statistik-Kenntnisse wollen Sie nun für sich in der Praxis nutzen. In Ihrer letzten Klausurphase erzielten Sie in der Klausur der Neuropsychologie 45 Punkte, wohingegen Sie in der Klausur der Sozialpsychologie 70 Punkte erreichten. Sie wissen, dass in der Neuropsychologie im Durchschnitt 50 Punkte erreicht wurden und es eine Streuung von s=10 gibt. In der Sozialpsychologieklausur wiederrum wurden im Schnitt 85 Punkte erreicht und es liegt eine Streuung von s=20 vor.
Wie können Sie ihre beiden Klausurergebnisse vergleichen? Interpretieren Sie das Ergebnis.

NUN ZU  :

7)

a) Legen Sie das Objekt „Schuhgröße“ in R an und lassen Sie sich den Inhalt des Objekts in der Konsole anzeigen.
b) Berechnen Sie nun den Mittelwert, die Standardabweichung, sowie die Varianz des Objekts und vergleichen Sie das Ergebnis mit ihren Ergebnissen aus Aufgabe 6a).
c) Bestimmen Sie zum Abschluss noch (händisch und mit R) Modus und Median und vergleichen Sie die drei Maße der zentralen Tendenz. Was sagt Ihnen der Vergleich der drei Werte über die Verteilung der Daten?

8)

a) Laden Sie nun den Datensatz „bielefeld.erstis.sva“ in R und erschaffen Sie sich einen Überblick über die Struktur der Daten.
b) Sie interessieren sich im Besonderen für die Verteilung der ersten LK’s unter Ihren Kommilitonen und Sie möchten sich mit Hilfe eines Diagramms einen Überblick verschaffen, welcher LK wie häufig vorkommt. Welches Skalenniveau hat die Variable? Entscheiden Sie sich auf dieser Grundlage für ein Diagramm und erstellen Sie es in R. Fügen Sie anschließend noch eine passende Überschrift für das Diagramm ein.
c) Jetzt möchten Sie sich zudem noch anschauen, wie gut Ihre Kommilitonen speziell im Fach Mathematik abgeschnitten haben und Sie schauen sich zu diesem Zweck die Variable „Mathe-Note“ an. Erstellen Sie ein Histogramm dieser Variablen. Mit welchen anderen Diagrammarten könnte diese Variable graphisch dargestellt werden? Erstellen Sie ein weiteres Diagramm einer anderen Art und vergleichen Sie ihre beiden Graphiken?

3. Kovarianz und Korrelation

1) Welche/s Zusammenhangsmaß/e ist/sind für die folgenden Merkmalspaare geeignet:

   a) Studienfach und Geschlecht.

   b) Alter beim Berufseinstieg und Anfangsgehalt im Beruf in Euro.

   c) Platzierung beim Finnbahnmeeting und Platzierung beim 100 Meter Lauf

Bonusfragen:

   a) Studienfach und Anfangsgehalt im Beruf in Euro

   b) Körpergröße und Hobby

   c) Abschlussnote in der Statistikklausur und Religionszugehörigkeit

2) Nach dem Empra Sozialpsychologie (N = 14) ergibt sich folgende Tabelle:

	männlich	weiblich
bestanden	2	4
nicht bestanden	1	1

Berechnen Sie den Zusammenhang von Geschlecht und Ergebnis. Welche(n) Koeffizienten können Sie berechnen und wie interpretieren Sie das Ergebnis? (Hinweis: $\frac{-8}{\sqrt{350}}$ = -0.43; $\frac{-2}{\sqrt{180}}$ = -0.15; $\frac{4}{\sqrt{290}}$ = 0.23)

3) Bei Teilnehmern eines Marathons wurde neben der Platzierung auch das Körpergewicht erfasst:

Person	1	2	3	4	5
Platzierung	4	1	3	2	5
Gewicht	83	82	48	62	45

Berechnen Sie den Zusammenhang von Platzierung und Körpergewicht. Wie interpretieren Sie diesen Zusammenhang?

4) Bei Patienten einer Rehaklinik beurteilten die Bezugstherapeuten Depressivität (1 = gar nicht depressiv bis 5 = sehr depressiv) und Suizidalität (1 = gar nicht suizidal bis 5 = sehr suizidal)

Patient	1	2	3	4
Depressivität	4	3	3	2
Suizidalität	2	3	2	1

Berechnen Sie Kreuzproduktsumme, Kovarianz und Korrelation und interpretieren Sie die Ergebnisse.

NUN ZU  :

5) Betrachten Sie noch einmal die Daten aus Aufgabe 4). Legen Sie in R die beiden Objekte „Depressivität“ und „Suizidalität“ an und berechnen Sie anschließend die Korrelation. Vergleichen Sie ihr Ergebnis mit dem aus Aufgabenteil 4).

6) Rufen Sie nun in R den Datensatz „bielefeldErstis1314.SVA“ auf.

a) Sie fragen sich, ob die Studenten auf höheren Fachsemestern tendenziell älter sind, als die in den niedrigeren Fachsemestern. Lassen Sie sich für einen ersten Überblick einen Scatterplot der beiden Variablen anzeigen.
b) Berechnen Sie nun die Korrelation der beiden Variablen und interpretieren Sie das Ergebnis.
c) In dem Datensatz finden Sie außerdem die beiden Variablen „Studiensicherheit 1“ und „Studiensicherheit 2“. Die Variable „Studiensicherheit 1“ (=STUWA_SI1) gibt an, wie sicher sich die Personen in der Wahl ihres Studienfaches sind, die Variable „Studiensicherheit 2“ (=STUWA_SI2) erfasst, wie gut sich diese Personen vorstellen können, den Studiengang zu wechseln. Überlegen Sie sich, in welchem Zusammenhang die beiden Variablen zueinander stehen könnten und berechnen Sie die Korrelation in R.

4. Einfache Lineare Regression

1) Von 4 Autos sind das Alter und der Bremsweg bei einer Vollbremsung bei 30 km/h gegeben:

Alter (Jahre)	4	5	6	1
Bremsweg (m)	5	6	7	2

a) Erstellen Sie ein Streudiagramm mit dem Alter auf der x- Achse und dem Bremsweg auf der y-Achse.

b) Berechnen Sie die Gleichung der Regressionsgraden und zeichnen Sie die Gerade in das Streudiagramm ein (runden Sie auf 2 Nachkommastellen!).

c) Bestimmen Sie mit Hilfe der Regressionsgraden den Bremsweg bei einem 15 Jahre alten Auto.

2) Bei einer zufällig ausgewählten Gruppe von Zuschauern eines Basketballspiels wurden von jeder Person Körpergröße und Gewicht erfasst. Es ergaben sich folgende Werte:

Körpergrösse	Gewicht
1.60	55
1.60	60
1.70	75
1.80	85
1.80	90

a) Erstellen Sie ein Streudiagramm mit der Körpergröße auf der x- Achse und dem Gewicht auf der y-Achse.

b) Netterweise hat jemand für Sie die Standardabweichungen und Kovarianz berechnet: s_x = .10; s_y = 15; s_{xy} = 1.20. Berechnen Sie die Gleichung der Regressionsgraden.

c) Berechnen Sie den Determinationskoeffizienten und interpretieren Sie ihn.

d) Machen Sie Voraussagen für die folgenden Personen:

Wie schwer bzw. wie groß müssten die folgenden Personen sein?

Person A: Größe: 1,75m

Person B: Größe: 1,35m

Person C: Gewicht: 109kg

NUN ZU  :

3) Betrachten Sie erneut die Daten aus Aufgabenteil 1). Sie wollen mit Hilfe Ihres Statistikprogramms R nun überprüfen, ob Sie die richtige Regressionsgleichung aufgestellt haben. Erstellen Sie dafür die beiden Objekte „Alter“ und „Bremsweg“ in R und berechnen Sie das lineare Modell für die beiden Variablen (der Bremsweg soll durch das Alter vorausgesagt werden).

4) Sie schauen sich noch einmal den Datensatz „bielefeldErstis1314“ an und sie haben die Vermutung, dass die Studierenden, die eine hohe Sympathie für das Fach Mathematik haben, auch mehr Interesse an Statistik haben. Sie fragen sich also, ob man von der Sympathie für das Fach Mathematik (MATHE_1) auf das Interesse für Statistik (INT_STAT) schließen kann.
    a) Berechnen Sie eine passende Regression und stellen Sie die Gleichung der Regressionsgeraden auf.
    b) Wie lautet in diesem Fall der Achsenabschnitt und wie können Sie ihn interpretieren?
    c) Berechnen Sie ebenfalls den Determinationskoeffizienten und interpretieren Sie ihr Ergebnis.

5. Partial- und Semipartialkorrelation

1) In einer lokalen Tageszeitung erscheint folgende Schlagzeile: „Je mehr Allgemeinwissen jemand hat (x) desto mehr Knöllchen bekommt diese Person (y)! Intelligente Menschen halten sich also weniger an allgemeine Verkehrsregeln!“ Beurteilen Sie mit Hilfe der folgenden Daten diese Aussage und berücksichtigen Sie dabei auch das Alter der VP (z)! (Runden Sie auf 2 Nachkommastellen!):
   Die Korrelation zwischen Allgemeinwissen und Anzahl der Knöllchen beträgt $r_{xy} = .90$
   Die Korrelation zwischen Allgemeinwissen und Alter beträgt $r_{xz} = .80$
   Die Korrelation zwischen Alter und Anzahl der Knöllchen beträgt $r_{yz} = .80$
   Prüfen Sie, ob der Zusammenhang zwischen Allgemeinwissen und Anzahl der Knöllchen durch das Alter erklärt werden kann.

2) Die Familienministerin hat festgestellt: „Die Bildung von Kindern (x) ist abhängig vom Einkommen der Eltern (y)!“ und hält Nachhilfekurse für Kinder deren Eltern ein geringes finanzielles Einkommen haben für sinnvoll. Beurteilen Sie diesen Plan, und berücksichtigen Sie dabei die Bildung der Eltern (z)! (Runden Sie auf 2 Nachkommastellen!). Ihre Annahme lautet, dass Eltern mit höherer Bildung besser verdienen. Sie denken also, dass möglicherweise nicht das Einkommen, sondern die Bildung der Eltern, die Unterschiede zwischen den Kindern entstehen lässt.
   Die Korrelation zwischen Bildung der Kinder und Einkommen der Eltern beträgt $r_{xy} = .50$
   Die Korrelation zwischen Bildung der Kinder und Bildung der Eltern beträgt $r_{xz} = .80$
   Die Korrelation zwischen Bildung der Eltern und Einkommen der Eltern beträgt $r_{yz} = .80$

NUN ZU  :

3) Zur Berechnung dieser Aufgabe wird der Datensatz „ExamAnxiety.dat“ benötigt. Finden Sie zunächst heraus, wie Sie den Datensatz „ExamAnxiety.dat“ in R importieren können.
   Der Datensatz enthält drei Variablen: Die Variable „Revise“ steht für die Lernzeit, die Variable „Exam“ für die Note und die Variable „Anxiety“ für die Prüfungsangst.
   a) Die Variablen Prüfungsangst und Note korrelieren mit r=-.44 negativ miteinander – was bedeutet diese Korrelation und welchen Einfluss könnte die Lernzeit auf das Ergebnis haben?
   b) Berechnen Sie die Partialkorrelation von Prüfungsangst (X), Note (Y) und Lernzeit (Z) mit Hilfe der Berechnung von Regressionsresiduen und interpretieren Sie diese.

c) Suchen Sie anschließend nach einem geeigneten Paket, mit welchem die Partialkorrelation berechnet werden kann. Berechnen Sie mit Hilfe des Pakets und der im Paket enthaltenen entsprechenden Funktion erneut die Partialkorrelation und vergleichen Sie ihre Ergebnisse.

6. Multiple Regression

1) Ein Fitnessstudio wirbt mit folgender Aktion: „Besuchen Sie unseren Kurs SUPER-FIT, denn durch die vielen abwechslungsreichen Übungen reduzieren Sie Gewicht UND erhöhen Ihr Lungenvolumen, so dass Sie Ihre Fitness entscheidend steigern können. Bevor Sie teilnehmen, möchten Sie wissen, wie die Fitness (y) durch die beiden Prädiktoren Gewicht (x_1) und Lungenvolumen (x_2) vorhergesagt werden kann. Erstellen Sie die Regressionsgleichung der multiplen Regression (Runden Sie auf 2 Nachkommastellen!). Sie haben die Mittelwerte der Variablen ($\bar{x}_1 = 3$, $\bar{x}_2 = 4$, $\bar{y} = 5$), die Standardabweichungen der Variablen ($s_{x1} = 1$, $s_{x2} = 2$, $s_y = 1$) sowie folgende Korrelationsmatrix gegeben:

	Gewicht (x1)	Lungenvolumen (x2)	Fitness (y)
Gewicht (x1)	1.000		
Lungenvolumen (x2)	-.40	1.000	
Fitness (y)	-.70	.80	1.000

NUN ZU  :

2) Zurück zum Datensatz „bielefeldErstis1314.SVA“. In den Aufgaben zur „Korrelation und Kovarianz“ konnte in Aufgabe h) gezeigt werden, dass die Variablen „STUWASI_1“ (=wie sicher sind sich die Personen in der Wahl ihres Studienfaches) und die Variabel „STUWA_SI2“ (=wie sehr können sich die Personen vorstellen den Studiengang zu wechseln) negativ miteinander korrelieren:

-> Personen, die sich sicherer in ihrer Wahl des Studienfaches sind, können sich weniger gut vorstellen, den Studiengang zu wechseln

a) Mit Ihren neu erworbenen Kenntnissen wollen Sie nun zur Korrelation auch noch die einfache lineare Regression von „STUWASI_1“ auf „STUWASI_2“ in R berechnen. Finden Sie heraus, wie viel der Gesamtvarianz der Studienwahlsicherheit1 durch die Studienwahlsicherheit2 erklärt werden kann.
b) Da Sie auch ihre neuen Kenntnisse zur Multiplen Regression anwenden wollen, fragen Sie sich, ob sich außer der Variablen „STUWA_SI2“ auch noch die Variablen „AND_STUDFA“ (=haben die Personen vorher schon etwas anderes studiert) und „FIU_E“ (=erhalten die Personen finanzielle Unterstützung durch die Eltern) zur Vorhersage der „STUWA_SI1“ eignen.

Berechnen Sie hierzu die multiple Regression, vergleichen Sie den Determinationskoeffizienten mit ihrem Ergebnis aus Aufgabenteil I) und interpretieren Sie ihr Ergebnis.

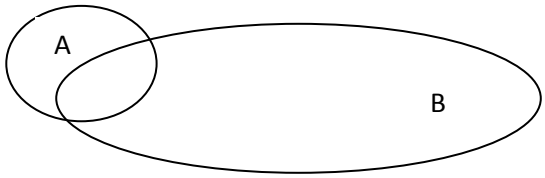
3) Für diese Aufgabe müssen Sie zunächst den Datensatz „recordsales.rda“ laden.

Mithilfe dieses Datensatzes soll ein Modell erstellt werden, das die Anzahl an verkauften Album Platten einer Band („sales“) durch die Variablen „adverts“ (Geld, das in Werbung gesteckt wurde), „airplay“ (Häufigkeit des Spielens eines Songs des Albums im Radio) und „attract“ (Attraktivität der Band) vorhersagt.

a) Stellen Sie zunächst ein einfaches lineares Modell auf mit dem Prädiktor, der die höchste Vorhersagekraft verspricht.
b) Fügen Sie in jedem weiteren (multiplen) Modell genau eine Variable hinzu (Beachtung sinnvoller Reihenfolge!) und berechnen Sie für jedes Modell R^2 und für die multiplen Modelle ΔR^2.

7. Wahrscheinlichkeitsrechnung und Kombinatorik

1) Markieren Sie im nachfolgenden Venndiagramm



a) $A \cap B$

b) $A \cup B$

c) $A \setminus B$

2) 50 Schulkinder werden gefragt, ob sie die Musik-AG, die Kunst-AG oder die Sport-AG besuchen. 10 besuchen nur die Sport-AG, 12 besuchen nur die Kunst-AG, 18 nur die Musik-AG. 6 besuchen die Sport-AG und die Musik-AG aber nicht die Kunst-AG. 4 besuchen die Kunst-AG und die Musik-AG aber nicht die Sport-AG.

Zeichnen Sie ein Venn-Diagramm für diesen Ereignisraum. Wie viele Kinder besuchen insgesamt die Sport-AG? Wie viele die Musik-AG? Wie viele die Kunst-AG?

Schließen sich Belegungen aus? Welche AGs sind erschöpfend (bilden den gesamten Wahrscheinlichkeitsraum)?

3) Vervollständigen Sie den Ergebnisraum folgender Ereignisse:

a) Ziehen einer Zahl von 0-9 $\Omega = \{$ $\}$

b) Zweimaliger Münzwurf $\Omega = \{$ $\}$

c) Ergebnis der Kugel beim französischen Roulette $\Omega = \{$ $\}$

4) Wie wahrscheinlich sind folgende Ereignisse:

a) Beim französischen Roulette bleibt die Kugel auf einer geraden Zahl liegen.

b) Mit einem Würfel wird eine Zahl gewürfelt, die kleiner oder gleich 2 ist.

c) Aus einem Kartenspiel mit 32 Karten wird zufällig eine schwarze Karte gezogen?

d) Aus einem Kartenspiel mit 32 Karten wird zufällig ein Bube gezogen?

e) Unter der Bedingung, dass die Roulette-Kugel bei französischen Roulette in einem schwarzen Feld liegen bleibt: Die Kugel bleibt auf einer gerade Zahl liegen. Achtung! Grade/Ungerade Zahlen sind beim französischen Roulette nicht gleichverteilt auf die Farben rot und schwarz!

5)

a) Wie lautet die Formel des Bayes-Theorems?
b) In einem Restaurant treten 2 Köche mit je einem Gericht gegeneinander an: 80% der Gäste wählen das Gericht von Koch 1, 20% wählen das Gericht von Koch 2. 60% derer, die das Gericht von Koch 1 gewählt haben, waren zufrieden, 40% nicht. 70% derer, die Gericht von Koch 2 gewählt haben, waren zufrieden, 30% nicht. Am Ausgang wählt ein Reporter eine Person zufällig aus und fragt ob es ihr geschmeckt hat. Die Antwort lautet „ja“. Wie groß ist die Wahrscheinlichkeit, dass diese Person das Gericht von Koch 2 gegessen hatte? [Anm: Wie groß ist also die Wahrscheinlichkeit, dass sich jemand für das Gericht von Koch 2 entscheidet UND es ihm/ihr schmeckt?]