Prof. Dr. Wolfgang Bischof
TH Augsburg
Wirtschaftspsychologie

1. Übungsblatt zur Vorlesung „Data Analytics II“

Aufgabe 1.1 *(Wilcoxon-Vorzeichen-Rang-Test)*

Betrachten Sie die Situation aus Aufgabe 1 des Übungsblatts zu den Zweistichprobentests.

(i) Testen Sie mit einem Test Ihrer Wahl, ob die Differenzen mit der Normalverteilungsannahme verträglich sind.

(ii) Nehmen Sie nun an, dass die Differenzen nicht als Realisierungen einer Normalverteilung angesehen werden können und prüfen Sie grafisch, ob die Annahme einer symmetrischen Verteilung der Differenzen zum Median gerechtfertigt ist.

(iii) Prüfen Sie auf einem Signifkanznivau von 5%, ob ein positiver Effekt der Therapie nachgewiesen werden kann.

Aufgabe 1.2 *(Wilcoxon-Mann-Whitney-Rangsummentest)*
Sven Red-Selig aus der fünften Klasse behauptet, dass die Jungs seiner Klasse viel schneller laufen als die Mädels. Da ihm die Mädchen widersprechen, bittet Sven seinen Mathelehrer Gerd Genau auf einem Signifikanzniveau von 5% nachzuweisen, dass seine Behauptung stimmt. Versuchen Sie, Svens Behauptung aufgrund folgender Daten zu belegen, in dem Sie annehmen, dass die Zeiten normalverteilt sind.

Wie gehen sie vor, wenn Sie lediglich wissen, dass die Zeiten kontinuierlich verteilt sind?

Jungen	8,4	8,1	10,0	9,0	8,9	8,5	8,6	9,4	9,3	8,7
Mädchen	10,2	9,7	8,2	9,5	10,1	10,8	8,0	8,8	9,6	9,8

Prüfen Sie, ob die Normalverteilungsannahme auf einem Signifikanzniveau von 10% abgelehnt werden kann. Verwenden Sie dazu drei verschiedene Tests auf Normalverteilung.

Aufgabe 1.3 *(Kruskall-Wallis-Test)*

(i) Laden Sie das Paket `DA.students` und betrachten Sie die Datentabelle `dat.studenten`. Weisen Sie signifikant nach, dass die Mathenote abhängig von der Zufriedenheit mit der Mathenote ist. Bestimmen Sie den genauen p-Wert des Testproblems.

(ii) Betrachten Sie die Datentabelle `mpg`. Können Sie auf einem Signifikanzniveau von 5% nachweisen, dass der Hubraum (`displ`) abhängig vom Hersteller ist. Ist das Baujahr (`year`) signifikant abhängig vom Hersteller?

Aufgabe 1.4 *(Chiquadrat-Anpassungstest zur Prüfung von Klassenwsk.en, Ross2006: Bsp. 11.2a, S.439f)*

Es wird vermutet, dass sich berühmte Leute auf ihren Geburtstag freuen (beispielsweise, weil sie da besondere Aufmerksamkeit erhalten). Folglich sind sie vor ihrem Geburtstag in guter mentaler Verfassung, was die Wahrscheinlichkeit dafür senkt, dass sie in den drei Monaten vor Ihrem Geburtstag sterben. Um diese Vermutung zu testen, wurde ein Stichprobe von 1251 Amerikaner:innen zufällig aus dem *Who Was Who in America* ausgewählt und die Monate vor oder nach ihrem Geburtstag gezählt:

Monate, die der Todestag vor (-) oder nach (+) dem Geburtstag liegt	Anzahl der Beobachtungen
-6,-5,-4	277
-3,-2,-1	283
0, 1, 2	358
3, 4, 5	333

(i) Prüfen Sie die Vermutung mit dem Chiquadrat-Anpassungstest auf einem Signifikanzniveau von 1% und berechnen Sie den p-Wert durch die Approximation durch die Chiquadrat-Verteilung.

(ii) Berechnen Sie den p-Wert nun durch 10000-fache Simulation von 1251 Todestagen.

(iii) Wie groß ist die Effektstärke?

(iv) Wie groß ist Teststärke bei der eben berechneten Effektstärke und einem Signifikanzniveau von 5%?

(v) Welche Stichprobengröße hätte ausgereicht, wenn die Effektstärke wie in (iii) berechnet ist, das Signifikanzniveau 5% und die Teststärke 75% beträgt?

Prof. Dr. Wolfgang Bischof
TH Augsburg
Wirtschaftspsychologie

2. Übungsblatt zur Vorlesung „Data Analytics II“

Aufgabe 2.1 *(Chiquadrat-Anpassungstest zur Prüfung einer Verteilungsannahme I)*

Sie nehmen an, dass die Anzahl der Menschen, die täglich Ihre Beratungsstelle aufsuchen, Poisson-verteilt ist mit Erwartungswert 4.2. Prüfen Sie diese Hypothese anhand folgender Häufigkeitstabelle, die sie über die letzten 150 Tage erfasst haben:

Anzahl Besucher	0	1	2	3	4	5	6	7	8	9	10	≥ 11
Anzahl der Tage	0	5	22	23	32	22	19	13	6	4	4	0

Prüfen Sie zusätzlich, ob die täglichen Besucheranzahlen mit einer Poisson-Verteilung mit unbekannter Ankunftsrate verträglich sind. Der Maximum-Likelihood-Schätzer für λ entspricht hier einfach dem arithmetischen Mittel der Ankünfte und beträgt 4.57.

Aufgabe 2.2 *(Chiquadrat-Anpassungstest zur Prüfung einer Verteilungsannahme II)*

150 Studierende schrieben die Mathe-Klausur mit. Die Gesamtpunktzahl lag bei 8 Studierenden unter 20; 18 erreichten zwischen 20 und unter 25 Punkten; 49 lagen zwischen 25 und unter 30; 52 waren zwischen 30 und unter 35; 22 waren zwischen 35 und unter 40; und einer erreichte mehr als 40 Punkte.

(i) Prüfen Sie, ob die Daten mit einer $N_{30,25}$-Verteilung verträglich sind.

(ii) Prüfen Sie nun die Hypothese, dass die Punkte überhaupt mit einer Normalverteilung verträglich sind. Die aus der Häufigkeitstabelle berechneten Maximum-Likelihood-Schätzer betragen 29.67 für μ und 5.36 für σ.

Aufgabe 2.3 *(Chiquadrat-Unabhängigkeitstest)*

In einer Schulklasse wurden die Schüler nach ihrem Lieblingsfach befragt. Dabei ergab sich folgendes Ergebnis:

	Jungen	Mädchen
Deutsch	1	5
Mathematik	4	1
Kunsterziehung	2	5
Sport	18	16

Es soll nun untersucht werden, auf welchem Signifikanzniveau die These, dass die Lieblingsfächer unabhängig vom Geschlecht gewählt werden, abgelehnt werden kann.

(i) Berechnen Sie den approximierten p-Wert des χ^2-Unabhängigkeitstests.

(ii) Prüfen Sie die Approximation durch Simulation.

(iii) Berechnen und simulieren Sie den p-Wert des exakten Tests von Fisher.

Aufgabe 2.4 *(Chiquadrat-Homogenitätstest)*

Am Abend der Bundestagswahl 2017 präsentierten die Fernsehanstalten ARD und ZDF um 18:00 eine letzte Prognose der Wahlergebnisse[1]. Wenn man annimmt, dass die ARD 1200 Menschen und das ZDF 1000 Menschen befragt hat, ergibt sich folgende 2×7-Feldertafel:

[1] Quelle: https://www.wahlrecht.de/news/2017/bundestagswahl-2017.html#prognosen -hochrechnungen

	CDU/CSU	SPD	Linke	Grüne	FDP	AfD	Andere	Sum.
ARD	390	240	108	114	126	162	60	1200
ZDF	335	210	90	90	100	130	45	1000
Sum.	725	450	198	204	226	292	105	2200

Weichen die Ergebnisse signifikant voneinander ab?

Aufgabe 2.5 *(Korrelationstests)*

In der Datei „WeiterbildungOriginalitaet.xlsx“ finden Sie die Umfrageergebnisse von 102 Personen zu den zwei Items:

- Ich bilde mich gerne weiter und versuche, meinen Horizont stets zu erweitern.
- Mir fällt es leicht, originelle Lösungsansätze zu finden.

Die Antworten reichen von 0 („trifft überhaupt nicht zu“) bis 5 („trifft voll und ganz zu“).

(i) Berechnen Sie die Stichprobenkorrelation nach Pearson und die Rangkorrelation nach Spearman.

(ii) Prüfen Sie, ob die Ergebnisse zu jeder Frage mit einer Normalverteilungsannahme verträglich sind. Prüfen Sie anschließend, ob die Ergebnisse zu beiden Fragen mit einer bivariaten Normalverteilungsannahme verträglich sind.

(iii) Prüfen Sie mit den Korrelationstests nach Pearson, Spearman und Kendall, ob eine Korrelation auf einem Signifikanzniveau von 1% nachgewiesen werden kann.

(iv) Führen Sie den Korrelationstest nach Pearson mit den Rängen der Umfrageergebnisse durch und vergleichen Sie das Ergebnis mit dem in (iii) durchgeführten Korrelationstest nach Spearman.

(v) Wie groß ist die Effektstärke beim Korrelationstest nach Pearson? Wie groß ist die Teststärke bei dieser Effektstärke und einem Signifikanzniveau von 5%?

(vi) Wie viele Personen hätten befragt werden müssen, um bei einem Signifikanzniveau von 5% eine Teststärke von 80% zu erreichen?

Prof. Dr. Wolfgang Bischof
TH Augsburg
Wirtschaftspsychologie

3. Übungsblatt zur Vorlesung „Data Analytics II“

Aufgabe 3.1 (Mathe und BWL)

Laden Sie sich die Excel-Datei `Punkte_BWL_Mathe.xlsx` aus unserem Moodle-Kursraum herunter.

Öffnen Sie die Datei und gehen auf Blatt „Ergebnisse“. In der ersten Spalte finden Sie die Matrikelnummern, in den nächsten zwei Spalten stehen jeweils die erreichten Punkte in den Prüfungsfächern BWL und Mathematik.

(i) Suchen Sie nach der Matrikelnummer 880687.

(ii) Sortieren Sie die Liste nach der Matrikelnummer aufsteigend.

(iii) Filtern Sie die Liste so, dass nur noch Kandidaten, die in Mathematik eine Punkteanzahl zwischen 46 und 49 haben, angezeigt werden.

(iv) Heben Sie den Filter wieder auf. Ermitteln Sie das beste Ergebnis in Mathematik mit einer geeigneten Funktion statt zu sortieren.

(v) Erstellen Sie mit Hilfe der Matrixfunktion `Häufigkeit` eine Häufigkeitstabelle für die erreichten Punkte in BWL mit folgender Klasseneinteilung:

86-100
74-85
62-73
50-61
0-49

Hinweis: Machen Sie sich mit der Funktion `HÄUFIGKEIT` *vertraut und beachten Sie, dass man Matrizenfunktionen mit* `Strg + Shift + Enter` *abschließen muss.*

(vi) Erweitern Sie die Tabelle analog für das Fach Mathematik, indem Sie die automatische Ausfüllfunktion geschickt ausnutzen.

(vii) Stellen Sie das Ergebnis in einer geeigneten Grafik dar. Achten Sie auf ansprechende Beschriftungen wie Über- und Unterschriften.

(viii) Stellen Sie mit einer geeigneten Funktion sicher, dass jeder Datensatz eindeutig ist und löschen Sie ggf. doppelte Datensätze.

(ix) Hinterlegen Sie alle Zeilen, in denen mindestens einmal eine Punkteanzahl unter 50 auftritt, mit oranger Farbe.

Hinweise:

- *Verwenden Sie dazu den Regeltyp „Formel zur Ermittlung der zu formatierenden Zellen“ verwenden.*
- *Eine Tabelle kann man dadurch markieren, dass man irgendein Feld in der Tabelle anklickt und dann* `Strg+A` *drückt.*

Aufgabe 3.2 (Mathe und BWL, Teil 2)

Diese Aufgabe setzt Aufgabe 3.1 fort.

(i) Laden Sie sich die Text-Datei `Namen.txt` aus unserem Moodle-Kursraum herunter und lesen Sie die Datei in das leere Excel-Arbeitsblatt `Namen` ein, so dass Sie je eine Spalte mit Überschrift für „Matrikelnummer“, „Anrede“, „Name“ und „Vorname“ bekommen.

(ii) Ergänzen Sie die Matrikelnummern im Blatt `Notenliste` mit Namen, Vornamen, Geschlecht und Punkte für Mathematik und BWL aus den Tabellenblättern `Ergebnisse` und `Namen` mit der Funktion `SVERWEIS`. In der Spalte Geschlecht soll „m“ für männlich und „w“ für weiblich stehen.

(iii) Berechnen Sie aus den Punkten das arithmetische Mittel je Fach über alle Studierenden und nach männlich und weiblich getrennt und den Median über alle.

(iv) Korrelieren die Prüfungsergebnisse der beiden Fächer? Erstellen Sie zunächst einen Scatterplot mit Beschriftungen, interpretieren Sie den Plot und berechnen im Anschluss den Korrelationskoeffizienten.

(v) Aus der Punktzahl ergeben sich folgende Noten:
86-100: Note 1
74-85: Note 2
62-73: Note 3
50-61: Note 4
0-49 : Note 5
Fügen Sie pro Fach eine Spalte mit Note an.

(vi) Stellen Sie die Notenverteilung in einer geeigneten Grafik dar.

Hinweis: Verwenden Sie vor der Erstellung der Grafik die Matrizenfunktion `HÄUFIGKEIT`*, um die absoluten Häufigkeiten der Noten je Fach zählen zu lassen.*

Aufgabe 3.3 (Finanzmathematik und der Solver)

Der Lebenskünstler Gustav Glück möchte mit dem Arbeiten aufhören, wenn er genug Geld verdient hat. Dazu prüft er zu jedem Jahresbeginn, ob das Geld reicht. Nehmen Sie vereinfachend an, dass die Inflationsrate konstant 2% beträgt.

(i) Wie viel Geld benötigt er, wenn er die nächsten 50 Jahre, jeweils am Jahresbeginn, inflationsbereinigt 100.000 Euro für seinen Lebensunterhalt abbuchen möchte (auch im aktuellen Jahr) und der Zinssatz konstant 5% beträgt?

Hinweis: D.h. zu Beginn des zweiten Jahres erhält er Zinsen in Höhe von (Anfangskapital - 100.000)· 0,05 und benötigt 102.000 Euro.

(ii) Wie hoch müsste der Zinssatz sein, damit ein Lottogewinn von 2 Mio. Euro ausreicht?

Hinweis: Die Aufgabe lässt sich gut mit dem Solver lösen. Wenn dieser in dem Excel-Programm auf Ihrem Computer nicht verfügbar ist, können Sie den Solver wie folgt installieren:

Datei → Optionen → Add-Ins → (ganz unten) Excel Add-Ins verwalten
Gehe zu anklicken → Haken bei Solver machen → OK

Die Verwendung des Solvers geschieht dann über:
Daten → Analyse auf Solver klicken und die notwendigen Eingaben vornehmen.

Prof. Dr. Wolfgang Bischof
TH Augsburg
Wirtschaftspsychologie

4. Übungsblatt zur Vorlesung „Data Analytics II“

Aufgabe 4.1 *(Strings, Strings, Strings)*

(i) Lesen Sie die Datei „Kfz_Policen.xlsx“, die Sie in unserem Moodle-Kursraum finden, in den Tibble `Kfz_Policen` ein. Wie viele Policen sind in der Datei aufgelistet?

(ii) Geben Sie alle Namen, die mindestens zweimal den Buchstaben „ü“ enthalten, aus. Was fällt Ihnen auf? Wie viele Namen sind betroffen?

(iii) Die Aktuare Ihrer Versicherung haben herausgefunden, dass Menschen, die mindestens zwei „ü“ im Namen tragen, besonders selten in Unfälle verwickelt sind. Daher möchten die Versicherung all diesen Personen einen Rabatt in Höhe von fünf Prozent gewähren. Passen Sie die Policen im Tibble `Kfz_Policen` entsprechend an.

(iv) Sie können sich nur noch ungefähr an den Namen eines gestrigen Anrufers erinnern, nämlich, dass der Namen „uk“ enthält und mit „ic“ endet. Nun will der Chef aber wissen, wie hoch dessen Prämie ist. Wie finden Sie diese Prämie heraus?

(v) Der Fehlerteufel hat zugeschlagen. Alle Personen mit Nachnamen „Schmidt“ heißen in Wirklichkeit „Schmitt“. Dies betrifft aber nicht Personen mit Doppelnamen wie z.B. „Schmidt-Stiffel“ oder „Lebkuchen-Schmidt“. Wie korrigieren Sie den Fehler mit einem geschickten Befehl?

Hinweis: Die Liste händisch zu durchsuchen, mag bei 1500 Verträgen noch sinnvoll sein, aber denken Sie an Datenbanken, die Millionen Verträge enthalten.

(vi) Lesen Sie die Datei „pi-10-million.txt“ mit dem Befehl `scan` in den Vektor `pi_str` ein, wobei Sie die Option `what = character()` setzen, damit die Zahl π als durchsuchbare Zeichenkette vorliegt. Die Datei enthält die ersten zehn Millionen Nachkommastellen von π. Sie ist in unserem Moodle-Kursraum zu finden. Lernen Sie diese Nachkommastellen bis zur Klausur auswendig.[2] Ab welcher Nachkommastelle wird erstmals eine Folge von sechs gleichen Ziffern vollendet[3]? Geben Sie zusätzlich die Folge der sechs Ziffer aus.

(vii) Bei welcher Nachkommastelle wird erstmals Ihr Geburtsdatum in der Form `TTMMJJ` vollendet, also z.B. die Folge 2,3,0,3,0,1 wenn Sie am 23.3.2001 geboren wurden)?

Aufgabe 4.2 *(Textanalysen für Fortgeschrittene)*

(i) Lesen Sie die Dateien „BewerbungEgoist.txt“ und „BewerbungMuster.txt“ mit dem `scan`-Befehl ein. Berechnen Sie für beide Texte die „Ich-Quote“, das ist die relative Häufigkeit des Wortes „ich“ (klein oder groß geschrieben) an allen Wörtern, wobei auch Zahlen als Worte zählen. Was fällt Ihnen auf? Ist der Unterschied signifikant?

Hinweise: Achten Sie beim Einlesen auf das richtige Encoding. Um die Satzzeichen zu eliminieren, fügen sie die Zeichenketten zunächst mit dem Befehl `str_c` zusammen und zerlegen Sie nachher wieder mit dem Befehl `str_split`, wobei Sie die Option `pattern` geeignet setzen, um die Worte voneinander zu trennen.

(ii) Lesen Sie die Datei „Faust_I.txt“ mit dem `scan`-Befehl ein und erstellen Sie ein Tibble, in dem die absoluten Häufigkeiten der zehn am meisten benutzen Wörter sowie die Häufigkeit der sonstigen Wörter stehen. Ignorieren Sie dabei Groß-/Kleinschreibung.

Hinweis: Sehen Sie sich zuvor den Text von Goethe an. Welche Worte sollten wir bei der Analyse ausschließen?

[2] Das war ein kleiner Scherz. Ich empfehle lediglich zu jedem Geburtstag so viele Nachkommastellen von π auswendig zu lernen, wie man Jahre alt geworden ist. Das hält geistig fit.

[3] Das ist der sogenannte Feynman-Punkt. Wer Genaueres wissen will, der lese: `https://de.wikipedia.org/wiki/Feynman-Punkt`

(iii) Unser Ziel ist es nun, nur aufgrund von Worthäufigkeiten erkennen zu können, ob es sich bei einem fremden Text um einen Auszug aus Faust II handelt oder um einen anderen Text. Lesen Sie dazu die beiden Dateien „Faust_II_Lustgarten_100.txt“ und „HSA-Qualitaetskodex-Novelle-2020.txt“ ein. Ermitteln Sie, wie oft die in (ii) ermittelten Wörter auftreten. Führen Sie anschließend Chi-Quadrat-Anpassungstests durch. Ist ein Unterschied zwischen den beiden Texten ersichtlich?

Hinweis: Sollte ein Wort aus der Liste aus (ii) in einem der Texte gar nicht auftreten, müssen Sie es vor Durchführung des Chi-Quadrattests mit dem Befehl `add_row()` *zum Tibble hinzufügen.*

Quellen zu (ii) & (iii):

- Goethe-Texte: `http://www.digbib.org`
- HSA-Qualitätskodex-Novelle `https://www.hs-augsburg.de/`

Aufgabe 4.3 *(Kategoriale Größen in „General Social Survey“))*

Betrachten Sie den Datensatz `gss_cat` aus der `tidyverse`-Bibliothek.

(i) Untersuchen Sie die Verteilung von `rincome` (= reported income = gemeldetes Einkommen). Warum ist das Standard-Säulendiagramm schwer zu verstehen? Wie können Sie die Darstellung verbessern?

(ii) Welches ist die häufigste Religion in dieser Umfrage? Welches ist die häufigste `partyid`?

(iii) Für welche der Faktoren in `gss_cat` gibt es eine natürliche Reihenfolge der Stufen? Wie können die anderen Faktoren in einem Säulendiagramm sinnvoll sortiert werden?

(iv) Wie haben sich die Anteile der Menschen, die sich als Demokraten, Republikaner und Unabhängige ausgeben, im Laufe der Zeit verändert? Erstellen Sie eine aussagekräftige und gut lesbare Grafik.