

Name:	Studiengang: <input type="checkbox"/> B.A. <input type="checkbox"/> MA. <input type="checkbox"/> Mag.
Vorname:	In FlexNow angemeldet: <input type="checkbox"/> Ja <input type="checkbox"/> Nein
Matrikelnummer:	
Studienfächer:	Fachsemester Informationswissenschaft:

Allgemeine Hinweise:

1. Überprüfen Sie bitte, ob Sie alle Seiten der Klausurangabe vollständig erhalten haben (Gesamtzahl: **11 Seiten**).
2. **Bearbeitungszeit: 90 Minuten**, maximal erreichbare **Punktezahl: 90**. Die jeweils erreichbare Punktezahl ist bei jeder Frage angegeben. Bitte teilen Sie Ihre Arbeitszeit entsprechend ein.
3. Denken Sie daran, die Daten oben einzutragen, **bevor** Sie mit der Bearbeitung beginnen.
4. Treffen Sie bitte die Auswahl Ihrer Antworten bei Multiple-Choice-Fragen **direkt** auf dieser Klausurangabe.
5. Verwenden Sie für die Beantwortung der Freitext-Fragen ebenfalls diese Klausurangabe. Sie können jederzeit auch die Rückseiten beschreiben, falls der Platz auf der Vorderseite nicht ausreichen sollte. Bitte geben Sie in jedem Fall an, auf welche Frage sich die Lösung jeweils bezieht.
6. Benutzen Sie keine Bleistifte, keine rot schreibenden Stifte und kein TippEx, o.ä.
7. Zugelassene Hilfsmittel: Foliensätze zur oben bezeichneten Veranstaltung sowie eigene handschriftliche oder maschinengeschriebene Notizen und Lehrbücher Ihrer Wahl. Sie dürfen diese Hilfsmittel nur auf Ihrem eigenen Arbeitsplatz ablegen.
8. Mobiltelefone sowie Computer am Arbeitsplatz - auch ausgeschaltet - sind **nicht zugelassen**.
9. Geben Sie keine mehrdeutigen (oder mehrere) Lösungen an. In solchen Fällen wird stets die Lösung mit der geringeren Punktzahl gewertet. Eine richtige und eine falsche Lösung ergeben also null Punkte.
10. Wenden Sie sich bei Unklarheiten in den Aufgabenstellungen immer an den Aufsichtsführenden. Hinweise und Hilfestellungen werden dann, falls erforderlich, offiziell für alle Teilnehmer durchgegeben.

Teil I) BENUTZERSTUDIEN (**MAX. 45 PUNKTE INSGESAMT**)

Aufgabe I.1a)

Im Kurs haben wir die Methoden „Card Sorting“ und „Inductive Grounded Theory“ besprochen und ausprobiert. Auf den ersten Blick scheinen sie sehr ähnlich zu sein. Nennen Sie Ähnlichkeiten / Gemeinsamkeiten und Unterschiede zwischen den zwei Methoden im Bezug auf ihre Durchführung und Zweck. (**2 Punkte pro Gemeinsamkeit / Unterschied, Max = 8**)

Gemeinsamkeiten	Unterschiede

Aufgabe I.1b)

Stellen Sie je ein prototypisches Anwendungsbeispiel für beide Methoden dar. (**5 Punkte**)

--

Aufgabe I.2)

Twitter ist ein Internetdienst zur Verbreitung von telegrammartigen Kurznachrichten. Obwohl Twitter viele aktuelle Informationen anbietet, z.B. zur öffentlichen Meinung, Fakten zu aktuellen Themen usw., ist die Suchfunktion auf Twitter bekannt dafür, dass sie schlecht ist. Um das Design von neuen Suchalgorithmen und Suchschnittstellen zu prägen, möchten Sie eine Studie durchführen um herauszufinden wonach Twitter-Nutzer suchen (bzw. suchen würden wenn die Suchfunktion besser wäre), welche Beweggründe Sie für Suchanfragen haben und welche Probleme auftreten.

Aufgabe I.2a)

Interviews: Eine Herangehensweise wäre detaillierte Interviews durchzuführen, in denen Probanden gefragt werden, wonach sie suchen, wie erfolgreich Suchanfragen sind und welche Probleme auftreten. Sie würden Teilnehmer gewinnen, indem Sie Zettel in der Universität, Geschäften usw., aufhängen und Emails über die Uni-E-Maillisten schicken.

Führen Sie einige Vorzüge und Nachteile dieser Methode stichpunktartig an. **(1,5 Punkte pro Grund, Max=9)**

Vorteile	Nachteile

Aufgabe I.2b)

Analyse naturalistischer Suchlogs: Ein Kollege, der bei Twitter arbeitet, bietet Ihnen die Gelegenheit an, eine Stichprobe der Twittersuchlogs zu analysieren. Diese Suchlogs stellen nur eine kleine Datenstichprobe dar (2 Monate Daten von 1000 Nutzer), und beinhalten die eingegebenen Suchanfragen (queries), Ergebnislisten (results lists) und „click-through“ Daten (die Tweets, die Leute aufgrund der Suchergebnisse lesen).

Was wären die Vor- und Nachteile einer Nutzung dieser Daten wenn man Suchverhalten untersuchen möchte? **(1,5 Punkte pro Grund, Max=9).**

Vorteile	Nachteile

Aufgabe I.2c)

Fällt Ihnen eine Methode ein, die Vorteile hätte, die die zwei Methoden in a) und b) nicht haben? Beschreiben Sie diese Methode und wie Sie die Methode durchführen würden (genauso so wie bei den Beispielen oben), und nennen Sie Vor- und Nachteile der Methode im Vergleich mit den Methoden in a) und b).

Aufgabe I.2d)

Bei der Durchführung von Benutzerstudien sind ethische Aspekte sehr wichtig. Besonders bei naturalistischen log-basierten Verfahren, wie dem Twitter Log-Daten Beispiel in b). Nennen Sie wichtige ethische Aspekte, die berücksichtigt werden müssen, wenn der Forscher solche Log-Daten selbst in einer Benutzerstudie sammeln möchte. **(1,5 Punkte pro Aspekt, Max=6)**

TEIL II STATISTISCHE DATENAUSWERTUNG (MAX. 45 PUNKTE INSGESAMT)

Aufgabe II.1) Normalverteilung (8 Punkte)

Kennzeichnen Sie für jede der folgenden Aussagen, ob diese wahr oder falsch ist.

Wahr**Falsch**

- | | | |
|--------------------------|--|--------------------------|
| <input type="checkbox"/> | Für die Standardnormalverteilung gilt $P(-\mu < X < \mu) \approx 0.68$. | <input type="checkbox"/> |
| <input type="checkbox"/> | Die Dichtefunktion ist symmetrisch um $x = \mu$. | <input type="checkbox"/> |
| <input type="checkbox"/> | Für die Standardnormalverteilung gilt $P(X \leq -1) = P(X \geq 1)$. | <input type="checkbox"/> |
| <input type="checkbox"/> | $P(X \leq \mu) = 0.5$ | <input type="checkbox"/> |
| <input type="checkbox"/> | Je größer σ^2 , desto schmaler und höher wird die Dichtefunktion. | <input type="checkbox"/> |
| <input type="checkbox"/> | Der Parameter μ bestimmt nur die Lage, nicht die Gestalt der Dichtefunktion. | <input type="checkbox"/> |
| <input type="checkbox"/> | Falls $Z \sim N(0, 1)$ und $X \sim N(\mu, \sigma^2)$, so gilt für alle möglichen Werte von μ und σ^2 : $P(-1 \leq Z \leq 1) = P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$. | <input type="checkbox"/> |
| <input type="checkbox"/> | Die Dichtefunktion ist für alle μ symmetrisch um $x = 0$. | <input type="checkbox"/> |

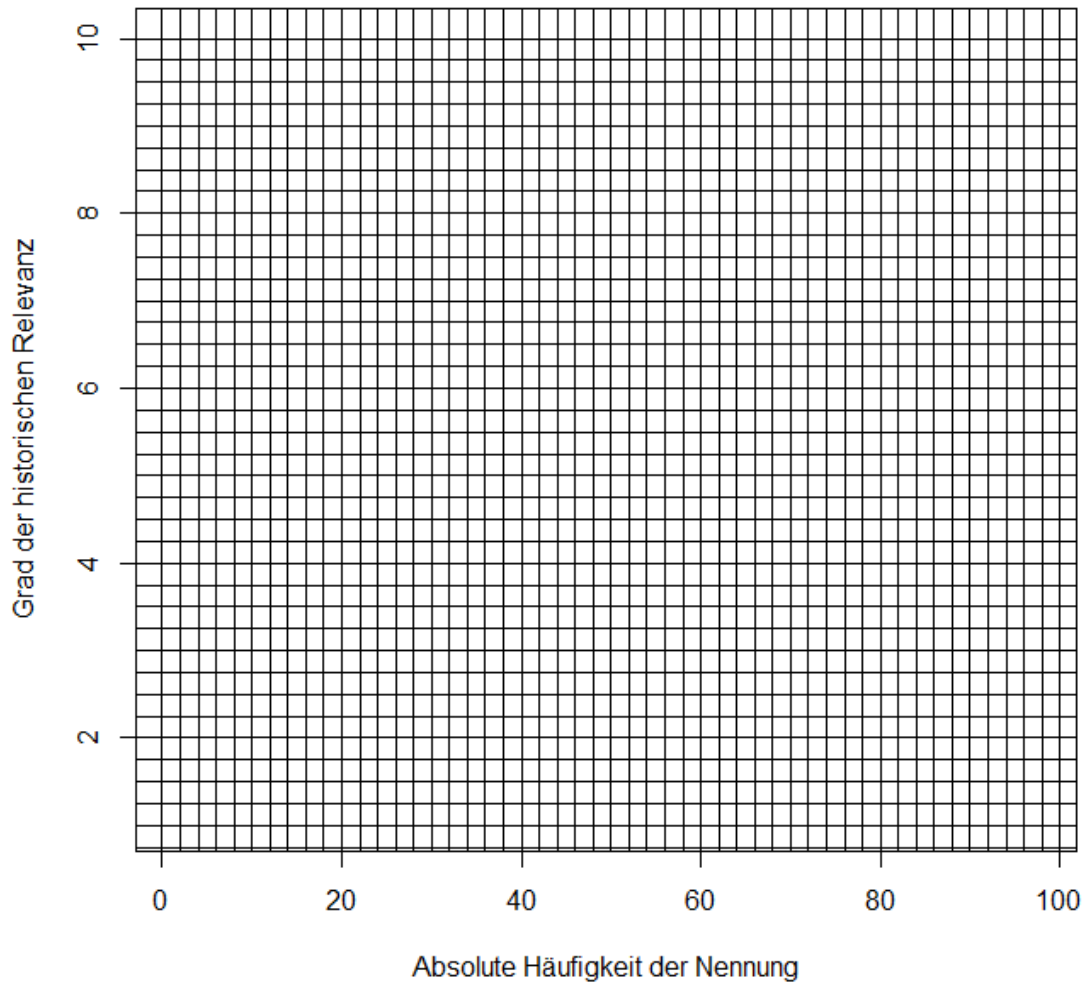
Aufgabe II.2) Lineare Regression (10 Punkte)

Im Rahmen eines Experiments zur Fußgängernavigation werden Touristen in der Stadt Regensburg befragt, welche Objekte Sie zur Orientierung benutzen. Die resultierende Objektliste wird dann Einwohnern von Regensburg vorgelegt; diese müssen nun die historische Relevanz der Gebäude auf einer Skala von 1 (völlig unbedeutend) bis 10 (außerordentlich bedeutend) bewerten. Es ergibt sich folgende Tabelle:

		Dom	Altes Rathaus	Bischofshof	Haus Heuport	Galeria Kaufhof	Maria Läng Kapelle	Dompost
Touristen (n=100)	H	98	87	50	32	70	1	18
Einwohn. (n=100)	Arith. Mittel	9,92	9,98	7,85	8,00	3,98	1,07	7,50

Wünschenswert wäre, aus der historischen Relevanz der Gebäude ableiten zu können, an welchen Gebäuden sich Touristen in Regensburg am häufigsten orientieren.

- a) Zeichnen Sie in das nachstehende Koordinatensystem das Streudiagramm ein, mit dessen Hilfe Sie den interessierenden Zusammenhang analysieren können. (2 Punkte)
- b) Zeichnen Sie die folgenden beiden Geraden in das Koordinatensystem ein. (2 Punkte)
 $y_1 = 0.1x + 4$
 $y_2 = 2x + 4$
- c) Erläutern Sie, wie sie vorgehen würden, um zu bestimmen, welche der beiden Geraden ein besseres lineares Regressionsmodell darstellt. (3 Punkte)
- d) Erläutern Sie, welche Eigenschaft die Skala zur Messung der historischen Relevanz erfüllen muss, damit diese Zusammenhangsanalyse möglich ist? (3 Punkte)



Aufgabe II.3) Hypothesentest I (6 Punkte)

Ein Information-Retrieval-System muss zur Erstellung des klassischen tf-idf-Maßes die Anzahl von Wörtern je Dokument bestimmen. Es werden zwei Algorithmen für das Zählen der Wörter in einem Dokument miteinander verglichen. Dadurch ergibt sich folgende Tabelle:

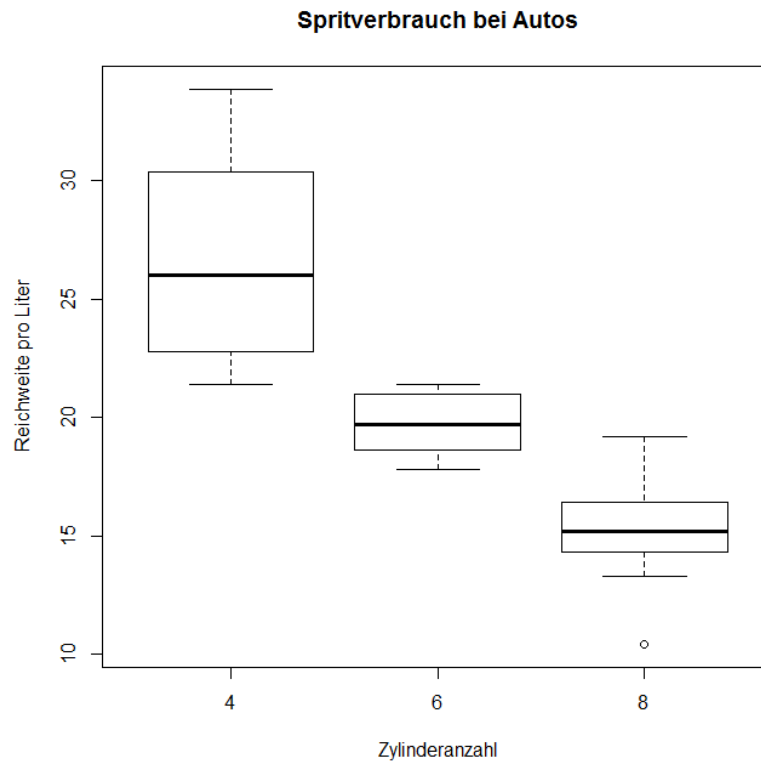
Text	1	2	3	4	5	6	7	8	9	10
Algorithmus A	223	259	248	220	287	191	229	270	245	201
Algorithmus B	220	244	243	211	299	170	210	276	252	189
Differenz A-B	3	15	5	9	-12	21	19	-6	-7	12

Die wissenschaftliche Fragestellung ist nun, ob sich die beiden Algorithmen systematisch unterscheiden.

- a) Inwiefern handelt es sich bei den Ergebnissen der Algorithmen um ein Zufallsexperiment? (2 Punkte)
- b) Handelt es sich um verbundene oder unabhängige Stichproben? Begründen Sie Ihre Einschätzung! (2 Punkte)
- c) Begründen Sie, ob und weshalb sich die Daten dazu eignen, mit Hilfe eines t-Tests zu prüfen, ob die Textlängen von Algorithmus A sich systematisch von denjenigen von Algorithmus B unterscheiden. (2 Punkte)

Aufgabe II.4) Deskriptive Statistik – interpretiert (5 Punkte)

Die nachfolgende Darstellung stellt die Reichweite von Autos und deren Zylinderzahl dar. Es wurden Autos mit 4, 6 und 8 Zylindern untersucht und jeweils deren Reichweite pro Liter gemessen.



Was schließen Sie aus dieser Darstellung

- Bezogen auf die Reichweite pro Liter in Abhängigkeit der Zylinderzahl? (2 Punkte)
- Bezogen auf das Minimum bei 8 Zylindern im Vergleich zum Minimum bei 6 Zylindern? (3 Punkte)

Aufgabe II.5) P-Werte (7 Punkte)

Kennzeichnen Sie für jede der folgenden Aussagen, ob diese wahr oder falsch ist.

Wahr**Falsch**

- | | | |
|--------------------------|--|--------------------------|
| <input type="checkbox"/> | Der P-Wert gibt die Wahrscheinlichkeit an, dass die Prüfgröße in den vorher festgelegten Ablehnungsbereich fällt. | <input type="checkbox"/> |
| <input type="checkbox"/> | Der P-Wert gibt die Wahrscheinlichkeit an, dass die Prüfgröße (PG) in den Ablehnungsbereich fällt, wenn man den gerade für die PG berechneten Wert als kritischen Wert, d.h. als Grenze des Ablehnungsbereiches verwendet. | <input type="checkbox"/> |
| <input type="checkbox"/> | Ist der P-Wert groß, so kann man die Nullhypothese als statistisch abgesichert betrachten. | <input type="checkbox"/> |
| <input type="checkbox"/> | Ist der P-Wert sehr klein, so kann man ziemlich sicher sein, dass die Alternativhypothese gilt. | <input type="checkbox"/> |
| <input type="checkbox"/> | Ist der P-Wert 0.14, so wird die Nullhypothese bei $\alpha = 0.1$ verworfen. | <input type="checkbox"/> |
| <input type="checkbox"/> | Würde man die Nullhypothese bei einem P-Wert von 0.0932 verwerfen, so wäre die Wahrscheinlichkeit für den Fehler erster Art 0.0932. | <input type="checkbox"/> |
| <input type="checkbox"/> | Bei einem klassischen Hypothesentest kann man nur dann eine sichere Aussage machen, wenn der P-Wert sehr groß ist. | <input type="checkbox"/> |

Aufgabe II.6) Unmögliches Ereignis (4 Punkte)

Zeigen Sie mit Hilfe der Axiome von Kolmogorov, dass die Wahrscheinlichkeit für das unmögliche Ereignis gleich Null ist.

Aufgabe II.7) Hypothesentest II (5 Punkte)

Mit Hilfe der Statistik-Software R wurde ein t-Test für unabhängige Stichproben durchgeführt. Den Output von R sehen Sie unten. Beurteilen Sie auf Grundlage dieser Darstellung, ob die nachfolgenden Aussagen wahr oder falsch sind.

```
Two Sample t-test

data: daten1 and daten2

t = -23.9173, df = 1998, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

99.9 percent confidence interval:

 -0.3795436 -0.2876197

sample estimates:
mean of x mean of y
 3.226843  3.560424
```

Wahr**Falsch**

- | | | |
|--------------------------|--|--------------------------|
| <input type="checkbox"/> | Man kann aus dem angegebenen Konfidenzintervall schließen, dass die Nullhypothese bei $\alpha = 0.001$ angesichts der errechneten Prüfgröße abgelehnt wird. (1 Punkte) | <input type="checkbox"/> |
| <input type="checkbox"/> | Die Datensätze enthielten zusammen 2000 Beobachtungen. (1 Punkt) | <input type="checkbox"/> |
| <input type="checkbox"/> | Die Wahrscheinlichkeit, dass die Prüfgröße kleiner als -23.9173 ist, ist beinahe 0. (1 Punkte) | <input type="checkbox"/> |
| <input type="checkbox"/> | Die Differenz der beiden Mittelwerte liegt mit einer Wahrscheinlichkeit von 99.9% in einem Intervall von -0.3795436 bis -0.2876197. (1 Punkte) | <input type="checkbox"/> |
| <input type="checkbox"/> | Es handelte sich um einen einseitigen t-Test (1 Punkte) | <input type="checkbox"/> |