FernUniversität in Hagen

Fakultät für
Kultur- und Sozialwissenschaften

Prüfungsamt

Übungsklausur zum Modul 2.1

"Methoden und Analyseverfahren"

im Bachelor-Studiengang "Politik und Organisation"

Klausur zur Modulfassung des WS 2007/08 mit den Kursen 03607, 33207 und 33208.
Die zu dieser Modulfassung folgende "echte" **Klausur am 10. März 2008** wird bezüglich Aufbau und Schwierigkeitsgrad vergleichbar sein.

Name: **Matrikel-Nr. :** | | | | | | | |

Unterschrift:

Block 1 (Kurs 03607 „Empirische Sozialforschung")						
Aufgabe	1-1	1-2	1-3	1-4	1-5	**Σ**
maximale Punktzahl	**6**	**6**	**6**	**4**	**3**	**25**
erreichte Punktzahl						

Block 2 (zu den Kursen 33207 „Statistik und wissenschaftstheoretische Grundlagen" und 33208 „SPSS")										
Aufgabe	2-1	2-2	2-3	2-4	2-5	2-6	2-7	2-8	2-9	**Σ**
maximale Punktzahl	**6**	**10**	**3**	**13**	**7**	**9**	**13**	**10**	**4**	**75**
erreichte Punktzahl										

Gesamtpunktzahl: **Note:**

Prüfer: **Unterschrift des Prüfers:**

Bearbeitungshinweise und Hinweise:

Es sind *beide* Aufgabenblöcke zu bearbeiten. Für Berechnungen in dieser Klausur können eine separat verteilte **Formelsammlung** sowie ein **Taschenrechner** herangezogen werden.

Bitte beantworten Sie alle Fragen, insbesondere die in Block 1, kurz, *nicht* in Form von Aufsätzen.

Wenn Sie 50 Punkte erreichen, also 50 % der maximal erreichbaren Punkte, haben Sie die Klausur auf jeden Fall bestanden.

Hinweise zu dieser Übungsklausur:

Sie finden ausführliche **Lösungen bis Ende November im Studienportal**. Sehen Sie also bis dahin von Fragen zur Lösung der Aufgaben ab. Diese Übungsklausur wird auch im Rahmen der in Hagen stattfindenden modulbezogenen Präsenzveranstaltung am 22./23. 2. 2008 besprochen. Die Formelsammlung ist im Prinzip für die Bearbeitung nicht unbedingt erforderlich, weil Sie ja zu Hause auch den Kurs 33207 heranziehen können. In der Klausur im März 2008 können Sie die Studienbriefe des Moduls aber *nicht* verwenden. Sie haben dann aber eine Formelsammlung, die alle benötigten Formeln abdeckt. Um die Übungsklausur unter klausurähnlichen Bedingungen zu bearbeiten, empfiehlt es sich zu versuchen, zunächst nur Formelsammlung und Taschenrechner als Hilfsmittel bei der Löusng der Aufgaben einzusetzen und nur bei Bedarf noch in den Kurs 33207 zu schauen.

Block 1: Da der Kurs 03607 bei der Änderung der Modulstruktur zum WS 2007/08 nicht verändert wurde (abgesehen von einer Streichung der Kapitel 8-9), sind die Aufgaben zu Block 1 bei der Musterklausur durchweg Aufgaben aus alten Klausuren. Auch das Gewicht des Blocks bleibt mit 25 P. von 100 P. unverändert.

Block 2: Dieser Block hat mit 75 P. dasselbe Gewicht, das bei früheren Klausuren insgesamt den Aufgabenblöcken zu den ehemaligen Kursen 03607 und 33208 zukam (25 P. und 50 P. waren hier bisher veranschlagt). Diese beiden Kurse bilden ja nun den Kurs 33207. Neu ist aber, dass der Block zum früheren Statistikkurs 33208 ein Wahlblock war und diese Wahloption nun nicht mehr besteht, weil der frühere Kurs 33211 jetzt im Rahmen des Moduls 2.2 geprüft wird. Die Klausur *im März 2008* wird in Block 2 – wie diese Übungsklausur – noch *keine Fragen* enthalten, *die sich auf SPSS beziehen*.

Noch ein Tipp: Nutzen Sie die bei der Prüfungsvorbereitung auch die in der Lernumgebung „Moodle“ bereitgestellten Arbeitsmaterialien (optional), die Ihnen den Zugang zu einzelnen Themenbereichen der Statistik erleichtern können (Experimente und Animationen mit Kommentaren) – den Zugang zu „Moodle“ finden Sie im Studienportal zum Modul 2.1.

Block 1: Empirische Sozialforschung 25 P.

Aufgabe 1-1 (Hypothesenbildung) 6 P.

In der Wahlforschung wird üblicherweise davon ausgegangen, dass die Parteiidentifikation, die Kandidatensympathie und die Einstellung zu Sachthemen des Wahlkampfes die Parteiwahlentscheidung beeinflussen. Formulieren Sie zum Zusammenhang zwischen den hier benannten Variablen eine Hypothese, benennen Sie die unabhängigen und die abhängigen Variablen und stellen Sie den Zusammenhang in Form eines Hempel-Oppenheim-Schemas dar.

Lösung:

Unabhängige Variable(n):

Abhängige Variable(n):

Zusammenhangshypothese:

Hempel-Oppenheim-Schema:

Aufgabe 1-2 (experimentelles Design) 6 P.

Welche Merkmale kennzeichnen ein klassisches experimentelles Design, das in sozialwissenschaftlichen Experimenten typischerweise mit Menschen durchgeführt wird?

Lösung:

Ziel eines experimentellen Designs:

Merkmale eines sozialwissenschaftlichen Experiments (nur kurze Antworten):

Aufgabe 1-3 (Auswahlverfahren) **6 P.**

Definieren Sie kurz die Begriffe *Grundgesamtheit*, *Vollerhebung*, *Auswahleinheit* und *Erhebungseinheit* und erläutern sie den unterschiedlichen Bedeutungsumfang an einem Beispiel:

Lösung:

Grundgesamtheit:

Vollerhebung:

Auswahleinheit:

Erhebungseinheit:

Beispiel zur Illustration aller vier Begriffe (kein Aufsatz, nur knapp antworten!):

Aufgabe 1-4 (Datenerhebung) **4 P.**

Beschreiben Sie, was ein reaktives oder um ein nicht-reaktives Verfahren der Datenerhebung kennzeichnet. Wie sind Befragung und Beobachtung in diesem Zusammenhang zu bewerten? Geben Sei nur kurze Antworten.

Charakterisierung von reaktiven bzw. nicht-reaktiven Verfahren der Datenerhebung:

Einordnung der Erhebungsverfahren *Befragung* sowie *Beobachtung* :

Aufgabe 1-5 (Fragebogen) **3 P.**

Was versteht man bei der Mikroplanung (Gestaltung einzelner Themenblöcke) des Fragebogens unter *Kontexteffekten*? Was kann man tun, um diese zu herauszufinden? Geben Sie auch hier kurz gefasste Antworten.

Lösung:

Kontexteffekt:

Identifikation von Kontexteffekten:

Block 2: Statistik und wissenschaftstheoretische Grundlagen 75 P.

Aufgabe 2-1 (Skalenarten) **6 P.**

In etlichen Städten werden Mietspiegel erstellt, die für Mieter und Vermieter eine Marktübersicht zu Miethöhen liefern. In solchen Mietspiegeln werden u. a. die nachstehenden 6 Merkmale erfasst. Geben Sie bei jedem Merkmal an, welcher Skalentyp zutrifft, indem Sie in jeder der 6 Tabellenzeilen genau ein Kreuz (x) setzen. Der Begriff „Metrische Skala“ wird als Oberbegriff für die Skalentypen Intervallskala, Verhältnisskala und Absolutskala verwendet. (In der letzten Spalte ist also genau dann ein Kreuz zu setzen, wenn das betreffende Merkmal einem der drei genannten Skalentypen zuzuordnen ist.)

Merkmal	**Zutreffender Skalentyp**		
	Nominalskala	**Ordinalskala**	**Metrische Skala**
Größe einer Wohnung in m^2			
Baujahr des Gebäudes			
Anzahl der Zimmer			
Art der Heizung (Gas, Elektro, Öl)			
Nettomiete für die Wohnung in €			
Lage der Wohnung (6 Kategorien von „Spitzenlage = 1“ bis „Problemlage = 6)			

Aufgabe 2-2 (Aussagenlogik) **10 P.**

a) Nachstehend sind jeweils zwei Prämissen (P1 und P2) und ein aus diesen abgeleiteter logischer Schluss (Konklusion K) aufgelistet, die zusammen jeweils ein Argument darstellen. Nicht alle Schlüsse K sind unbedingt korrekt. Tragen Sie unter „Antworten" das Kürzel „w" für korrekte Konklusionen und „f" bei Fehlschlüssen ein.

P1	$a \rightarrow \neg b$	$\neg a \rightarrow \neg b$	$b \rightarrow \neg a$
P2	b	a	b
K	$\neg a$	b	$\neg a$

Antworten: ( ) ( ) ( ) (3 P.)

b) Verfahren Sie ebenso bei den nachstehend angegeben Prämissen P1 und P2 und der angegeben Konklusion K, d. h. tragen Sie auch hier „w" oder „f" ein. Füllen Sie zuvor die vorgegebene Wahrheitstabelle aus, aus der sich die korrekte Antwort („w" oder „f") ergibt:

P1: $a \vee \neg b$

P2: $\neg a$

K: $\neg a \wedge \neg b$

Antwort: ( )

a	b	P1: $a \vee \neg b$	P2: $\neg a$	K: $\neg a \wedge \neg b$
w	w			
w	f			
f	w			
f	f			

(7 P.)

Bepunktung bei Aufgabenteil a: Je 1 P. pro zutreffender Antwort. Bei unzutreffenden Antworten wird zur Vermeidung von Ratestrategien je ein Punkt abgezogen. Die Gesamtpunktzahl für Aufgabenteil a darf dabei aber 0 P. nicht unterschreiten.

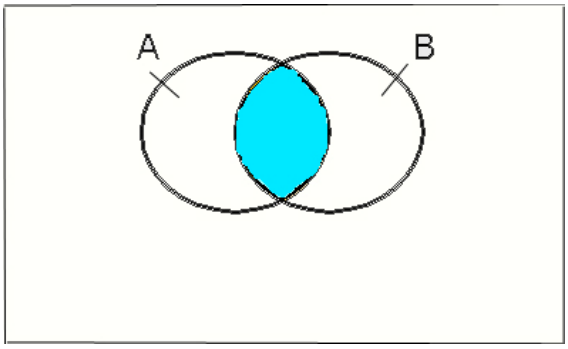
Bepunktung bei Aufgabenteil b: Für jede korrekt ergänzte Spalte der Wahrheitstabelle 2 P., also maximal 6 P. (keine Abzüge bei Fehlantworten). Für die korrekte Antwort („w" oder „f") gibt es nur dann einen weiteren Punkt, wenn die Antwort nicht im Widerspruch zu den Inhalten der letzten Spalte der Wahrheitstabelle steht.

Aufgabe 2-3 (Venn-Diagramme / Mengenoperationen) **3 P.**

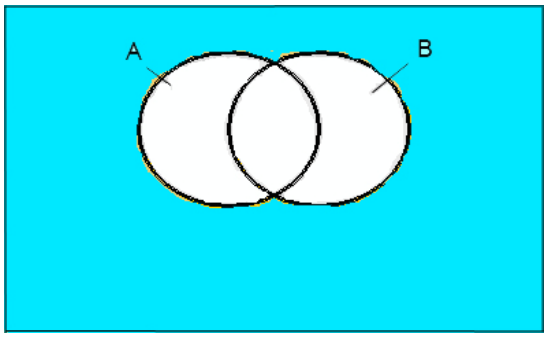
Venn-Diagramme werden zur Veranschaulichung von Mengenoperationen verwendet. Diese bestehen aus einem Rechteck, in der Mengen A, B, C, … als Kreise oder Ellipsen dargestellt sind. Das Rechteck repräsentiert eine Grundgesamtheit, von der die eingezeichneten Mengen Teilmengen sind.

Nachstehend sind vier Venn-Diagramme abgebildet, die sich auf die Verknüpfung von zwei Mengen A und B oder drei Mengen A, B und C beziehen. Die Verknüpfungen erfolgen über die Symbole „$\cup$" (Vereinigung von Mengen) und „$\cap$" (Schnittmengenbildung) und „$\neg$" (Komplementärmenge einer Menge, auch durch das Zeichen „$\overline{\ }$" über der betreffenden Menge darstellbar).

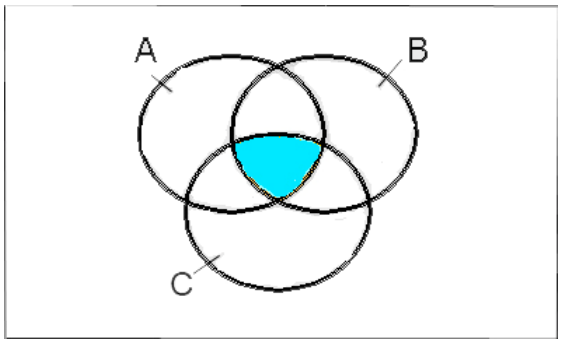
Geben Sie unter Verwendung der genannten Symbole wieder, was in den letzten drei Venn-Diagrammen durch die markierten Flächen dargestellt ist. Beim ersten Venn-Diagramm ist die Lösung schon vorgegeben.



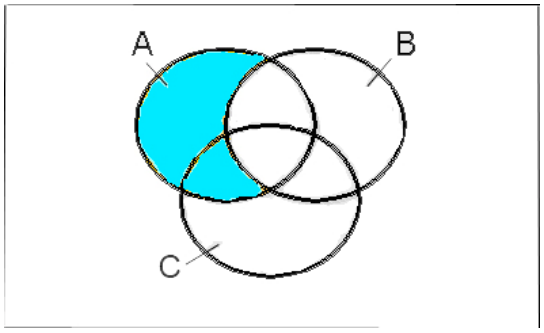
Lösung: $A \cap B$



Lösung:



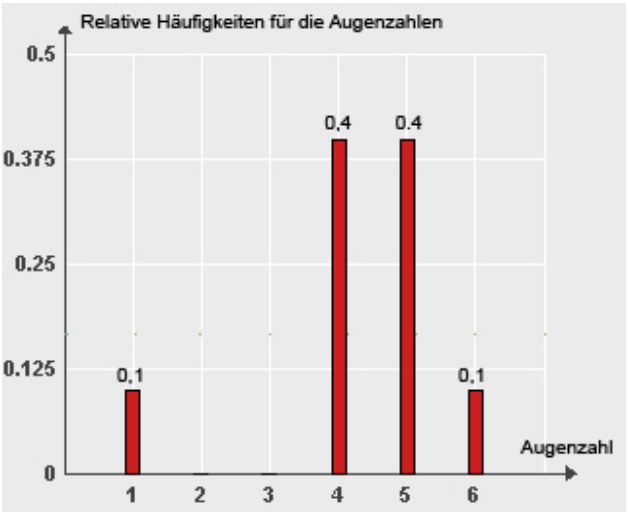
Lösung:



Lösung:

Aufgabe 2-4 (Lage- und Streuungsparameter) **13 P.**

Nachstehend ist das Ergebnis eines Würfelexperiments visualisiert, bei dem 10 Mal nacheinander mit einem Würfel gewürfelt wurde. Dargestellt sind die relativen Häufigkeiten für die sechs möglichen Ausgänge eines Wurfs.



a) Vervollständigen Sie die mittlere Zeile in der folgenden Tabelle: (1 P.)

	Beobachtete Augenzahl					
	1	2	3	4	5	6
Absolute Häufigkeit						
Relative Häufigkeit	0,1	0	0	0,4	0,4	0,1

b) Geben Sie auch die alle 10 Beobachtungen umfassende Urliste an, wobei die Werte der Urliste nach zunehmender Größe zu ordnen sind: (1 P.)

Geordnete Urliste:

c) Berechnen Sie dann auf der Basis der beim Experiment beobachteten Werte den Median x_{med} und den Mittelwert $\bar{x}$. (2 P.)

Median:

Mittelwert:

d) Berechnen Sie für den per Experiment generierten Datensatz auch die Spannweite und die Standardabweichung. (3 P.)

Spannweite:

Standardabweichung:

e) Die Augenzahl beim einmaligen Wurf mit einem Würfel ist eine Zufallsvariable X. Berechnen Sie den Erwartungswert *E(X)* und die Varianz *Var(X*) dieser Zufallsvariablen.

(4 P.)

Erwartungswert:

Varianz:

f) Die Abbildung zu Beginn dieser Aufgabe zeigte das Ergebnis eines Würfelexperiments, das aus n = 10 aufeinanderfolgenden Würfen eines Würfels bestand. Wie würde sich der aus den n beobachteten Augenzahlen errechnete Mittelwert $\bar{x}$ verhalten, wenn man den Wert n bei diesem Experiment immer weiter erhöhte?

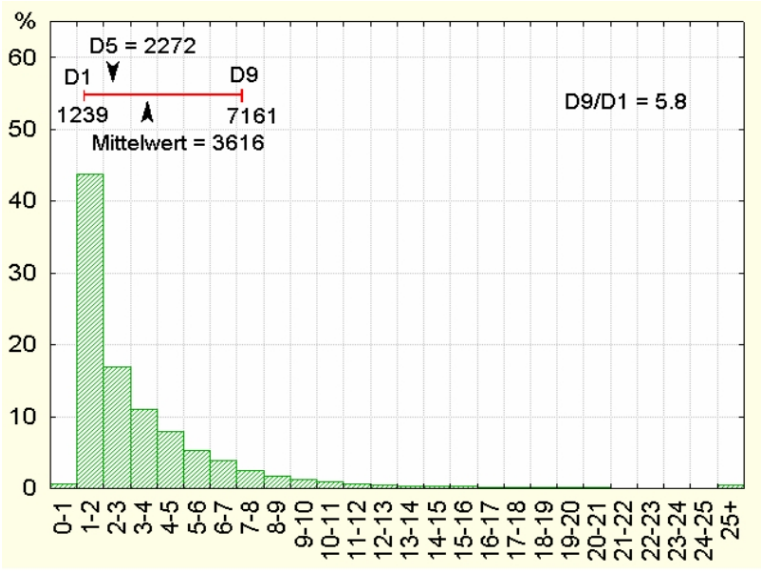
(2 P.)

Antwort (nur ein Satz!):

Aufgabe 2-5 (Darstellung und Beschreibung von Verteilungen) **7 P.**

Die folgende Abbildung zeigt für das Referenzjahr 2002 Jahresbruttoverdienste von Arbeitnehmern in Lettland (über eine Million Arbeitnehmer im Bereich „Industrie und Dienstleistungen, offizielle Daten von Eurostat). Auf der horizontalen Achse sind Einkommensklassen wiedergegeben (jeweils in 1000 €), auf der vertikalen Achse die Besetzungshäufigkeiten für die einzelnen Einkommensklassen (relative Häufigkeiten in Prozent). Man erkennt z. B., dass fast 45 % der Arbeitnehmer Jahresbruttoverdienste hatten, die zwischen 1000 € und 2000 € lagen.

In der Abbildung sind auch drei aus den Originaldaten errechneten Quantile wiedergegeben, nämlich das 10%-Quantil (hier abgekürzt mit D1), das 90%-Quantil (mit D9 abgekürzt) und das 50%-Quantil, also der Median (hier mit D5 angesprochen). Außerdem ist der - ebenfalls aus den Originaldaten errechnete – Mittelwert $\bar{x}$ ausgewiesen.



a) Wie heißt das hier zur Visualisierung der Einkommensverteilung verwendete grafische Instrument? (Die über der Grafik wiedergegebenen numerischen Informationen können Sie bei der Beantwortung dieser Frage ignorieren.) (1 P.)

Antwort (nur ein Wort oder ein Satz!):

b) Charakterisieren Sie die Form der Einkommensverteilung mit einem geeigneten Begriff.

(1 P.)

Antwort:

c) Das durch das 10% - und 90%-Quantil definierte Intervall gibt zusammen mit dem Median bereits wesentliche Charakteristika des umfangreichen Originaldatensatzes wieder. Man könnte anstelle dieses Intervalls auch einen Boxplot zur Beschreibung wesentlicher Eigenschaften des Datensatzes verwenden. Welche Größen werden benötigt, um einen Boxplot zeichnen zu können? (2 P.)

Antwort (nur Stichworte):

d) Grundsätzlich könnte man Jahresbruttoverdienste auch auf Euro und Cent genau ausweisen (ohne jede Rundung oder Zusammenfassung von Daten zu Einkommensklassen) und die Daten anhand eines Stabdiagramms darstellen. Welcher Nachteil wäre damit im Vergleich zur oben gewählten Visualisierungsform verbunden? (1 P.)

Antwort (nur ein Satz):

e) Neben dem Median oder Mittelwert könnten man zur Beschreibung von Einkommensdaten auch den Modus verwenden. Wie ist dieser definiert? Ist der Modus auch im Falle ordinal- und nominalskalierter Merkmale erklärt? (2 P.)

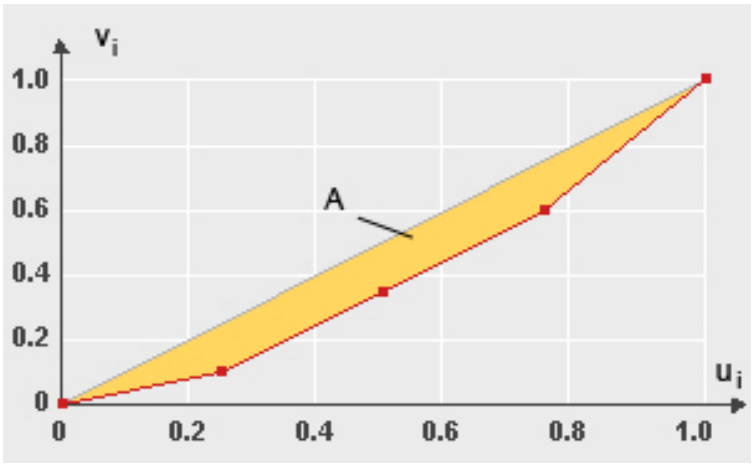
Definition des Begriffs „Modus“ (nur ein Satz!):

Der Modus ist auch für ordinal- und nominalskalierte Merkmale erklärt: ja ( ) nein ( )

Aufgabe 2-6 (Konzentrationsmessung) **9 P.**

In einem EU-Staat konkurrieren vier Firmen im Bereich der Telekommunikation. Seien $x_1 = 20$, $x_2 = 50$, $x_3 = 50$ und $x_4 = 80$ die Umsätze dieser Firmen im letzten Geschäftsjahr (Umsätze jeweils in Millionen €). Die nachstehende Abbildung zeigt die auf der Basis dieser Umsatzdaten errechnete Lorenzkurve (Polygonzug). Die Stützpunkte $(u_i; v_i)$ der Lorenzkurve sind auf der Lorenzkurve betont und ihre Koordinaten neben der Grafik wiedergegeben.

i	u_i	v_i
0	0	0
1	0,25	0,10
2	0,50	0,35
3	0,75	0,60
4	1	1



a) Berechnen Sie den Gini-Koffizienten und den normierten Gini-Koeffizienten. Bezeichnen Sie ersteren mit G und letzteren mit G*. (5 P.)

Berechnung von G:

Berechnung von G*:

b) In der Grafik ist neben der Lorenzkurve noch eine durchgezogene Kurve eingezeichnet, die die Punkte $(0;0)$ und $(1;1)$ direkt verbindet. Wie hängt die mit A bezeichnete Fläche zwischen Lorenzkurve und durchgezogener Kurve mit dem Gini-Koeffizienten zusammen?

(1 P.)

Antwort (nur ein kurzer Satz oder eine Gleichung):

c) Wie müsste man die Umsätze x_1, x_2, x_3 und x_4 der vier Firmen definieren, wenn man erreichen will, dass die Lorenzkurve mit der durchgezogenen Linie übereinstimmt?

(2 P.)

Antwort (nur ein Satz) :

c) Welchen Wert nimmt der Gini-Koeffizient bei der in Aufgabenteil c beschriebenen Situation an?

(1P.)

Antwort (nur ein Satz oder eine Gleichung):

Aufgabe 2-7 (Zusammenhangsmessung bei nominalskalierten Merkmalen) **13 P.**

Bei einer Kommunalwahl in einer Großstadt geht es um die Besetzung des Oberbürgermeisteramts. Zwei Kandidaten A und B stellen sich zur Wahl. Vor der Wahl werden n = 642 wahlberechtigte Männer und Frauen nach ihrer Präferenz befragt. Die Ergebnisse sind in der folgenden Vierfeldertafel zusammengefasst:

	Kandidat A	Kandidat B	Zeilensumme
Männer	144	153	297
Frauen	200	145	345
Spaltensumme	344	298	642

a) Berechnen Sie auf 3 Dezimalstellen genau den χ^2-Koeffizienten (*Chi*^{2}-Koeffizienten).

(6 P.)

Lösung:

b) Berechnen Sie dann auch auf 3 Dezimalstellen genau den Kontingenzkoeffizienten *C* sowie den korrigierten Kontingenzkoeffizienten C_{korr}. (4 P.)

Berechnung von *C*:

Berechnung von C_{korr} :

c) Welche Unter- und Obergrenze gibt es für den korrigierten Kontingenzkoeffizienten, d. h. welche Schranken kann C_{korr} generell nicht unter- bzw. überschreiten? (2 P.)

Antwort: Die Untergrenze von C_{korr} ist gegeben durch ….. , die Obergrenze ist durch ……

d) Was lässt sich aus dem in Aufgabenteil b errechneten Ergebnis zum Zusammenhang zwischen den Merkmalen „Geschlecht der befragten Person“ und „Kandidatenpräferenz“ aussagen? (1 P.)

Antwort:

Aufgabe 2-8 (Zusammenhangsmessung bei ordinal- und intervallskalierten Merkmalen)

10 P.

a) Für zwei intervallskalierte Variablen X und Y liegen Beobachtungen in Form von Wertepaaren (x_1, y_1), (x_2, y_2), …, (x_n, y_n), vor, die in den folgenden zwei Grafiken anhand je eines Streudiagramms veranschaulicht seien. Geben Sie an, welche Größenordnung der Korrelationskoeffizient r nach Bravais-Pearson in den beiden Abbildungen hat. Verwenden Sie dabei Codes für die folgenden Bereiche:

1 = Korrelationskoeffizient r ist genau – 1

2 = r liegt im Bereich von -1 bis -0,5

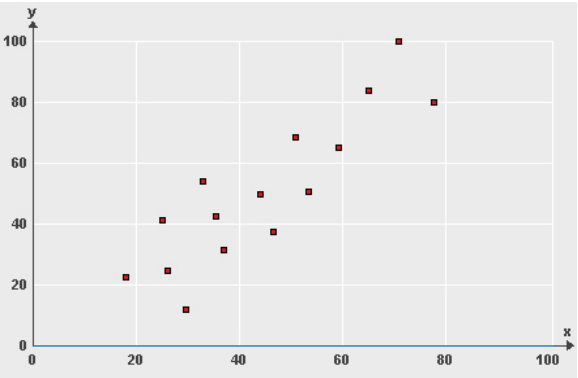
3 = r liegt im Bereich von -0,5 bis 0

4 = r liegt im Bereich von 0 bis 0,5

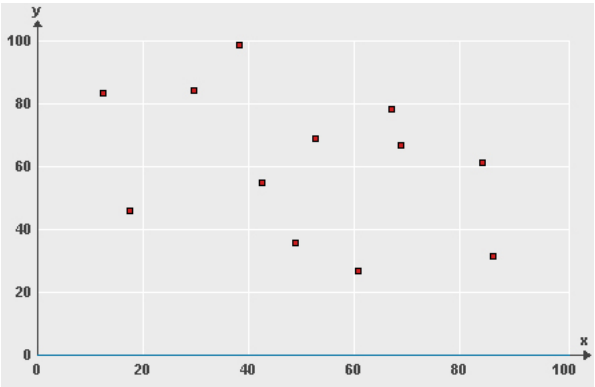
5 = r ist größer als 0,5

6 = Korrelationskoeffizient r ist genau 1

(4 P.)



Code:



Code:

b) Welche der drei Aussagen sind wahr, welche falsch. Tragen Sie „w“ bzw. „f“ ein. (3 P.)

Der Korrelationskoeffizient r nach Bravais-Pearson misst bei zwei intervallskalierten Merkmalen die Stärke eines linearen Zusammenhangs. ( )

Wenn r für zwei intervallskalierte Merkmalen den Wert Null annimmt, kann durchaus ein nicht-linearer Zusammenhang vorliegen. ( )

Wenn r nahe bei 1 liegt, ist zwischen den Merkmalen stets ein starker kausaler Zusammenhang gegeben. ( )

Bepunktung bei Aufgabenteil b: Je 1 P. pro zutreffender Antwort. Bei unzutreffenden Antworten wird zur Vermeidung von Ratestrategien je ein Punkt abgezogen. Die Gesamtpunktzahl für Aufgabenteil b darf dabei aber 0 P. nicht unterschreiten.

d) Es sei angenommen, dass zwei unabhängige Kreditsachbearbeiter die Kreditwürdigkeit von fünf Sparkassenkunden – hier mit 1 bis 5 nummeriert - anhand einer 10-stufigen Ratingskala bewerten, bei der die Punktzahl 1 sehr schlechte und die Punktzahl 10 sehr gute Bonität bezeichne. Die Ergebnisse sind nachstehend ausgewiesen:

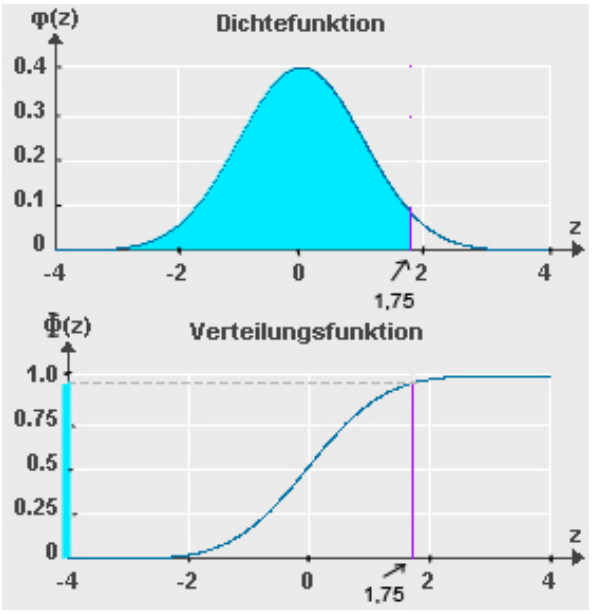
Nr des Kunden	Beurteilung durch Sachbearbeiter A	Beurteilung durch Sachbearbeiter B
1	5	6
2	8	9
3	9	7
4	2	4
5	6	5

Berechnen Sie den Rangkorrelationskoeffizenten r_S nach Spearman. (3 P.)

Lösung:

Aufgabe 2-9 (Dichte- und Verteilungsfunktion der Standardnormalverteilung) **4 P.**

Nachstehend sind die Dichtefunktion der Standardnormalverteilung und die Verteilungsfunktion der Standardnormalverteilung wiedergegeben. Erstere ist hier mit $\varphi(z)$, letztere mit $\Phi(z)$ bezeichnet. Auf der z-Achse ist jeweils der Punkt z = 1,75 markiert. Einer hier nicht wiedergegebenen Tabelle der Standardnormalverteilung kann man entnehmen, dass $\Phi(1,75) = 0,9599$ gilt. (Dieser Wert entspricht im oberen Teil der Abbildung dem Inhalt der markierten Fläche unter der Dichtekurve und im unteren Teil der Länge des auf der vertikalen Achse markierten Abschnitts, der von 0 bis zur gestrichelten Linie reicht.)



a) Welchen Wert nimmt die Verteilungsfunktion $\Phi(z)$ an der Stelle z = - 1,75 an?

(2 P.)

Lösung:

b) Berechnen Sie die Wahrscheinlichkeit dafür, dass eine standardnormalverteilte Variable Z Werte zwischen z = -1,75 und z = 1,75 annimmt. (2 P.)

Lösung:

Formelsammlung zur Klausur zum Modul 2.1

"Methoden und Analyseverfahren"

im Bachelor-Studiengang "Politik und Organisation"

(**Diese Formelsammlung wird bei Klausurende nicht eingesammelt**.)

Für Berechnungen in dieser Klausur kann diese kleine *Formelsammlung eingesetzt werden* und zwar sowohl **für die Klausur zur alten Modulfassung** vom SS 2007 **wie auch für die Klausur zur neuen Modulstruktur** vom WS 2007/08. Es werden für die Klausurbearbeitung nicht unbedingt alle hier aufgeführten Formeln wirklich benötigt.

Lage- und Streuungsparameter empirischer Verteilungen:

Median: Wenn man einen Datensatz x_1, x_2, …, x_n nach aufsteigender Größe ordnet, so ist der Median x_{med} bei ungeradem n der eindeutig bestimmte mittlere Wert der geordneten Folge. Bei geradem n gibt es zwei mittlere Werte und der Median ist dann der Mittelwert aus diesen beiden Werten.

Mittelwert: $\bar{x} = \sum_{i=1}^{n} x_i$ (auch: arithmetisches Mittel genannt)

Spannweite: Differenz r aus größtem und kleinstem Wert der geordneten Folge.

Varianz: $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ (genauer: empirische Varianz, weil aus Daten ermittelt)

Standardabweichung: $s = \sqrt{s^2}$ (genauer: empirische Standardabweichung)

Wenn mehr als ein Merkmal im Spiel ist, verwendet man zur Unterschiedung i. a. tiefgestellte Indizes, z.B. s_x für die Standardabweichung eines Merkmals *X* oder s_y für die Standardabweichung eines Merkmals *Y*.

Konzentrationsmessung:

Auf der Basis eines Datensatzes $x_{1,}\ x_{2,}\ \ldots,\ x_n$ (ungruppierte Daten) sei eine Lorenzkurve errechnet worden, also ein Polygonzug, der durch *n+1* Stützpunkte $(u_0; v_0), (u_1; v_1), \ldots (u_n; v_n)$ definiert ist mit $(u_0; v_0) = (0;0)$ und $(u_n; v_n) = (1;1)$. Der (nicht-normierte) Gini-Koeffizient *G* errechnet sich dann gemäß

$$G = \sum_{i=1}^{n} v_i u_{i-1} - \sum_{i=1}^{n} v_{i-1} u_i .$$

Dessen Obergrenze hängt von *n* ab. Durch Multiplikation mit $\frac{n}{n-1}$ erhält man den normierten Gini-Koeffizienten *G**, dessen Obergrenze nicht mehr von *n* abhängt (die Obergrenze ist hier 1).

Zusammenhangsmessung bei nominalskalierten Merkmalen

Gegeben seien ein nominalskaliertes Merkmal *X* mit den Ausprägungen a_1, a_2, …, a_k und ein nominalskaliertes Merkmal *Y* mit den Ausprägungen b_1, b_2, …, b_m. Die beobachteten absoluten oder relativen Häufigkeiten h_{ij} bzw. f_{ij} für die k·m möglichen Ausprägungskombinationen (a_i,b_j) lassen sich in einer Tabelle zusammenstellen, die Kontingenztabelle oder Kontingenztafel heißt. Im Spezialfall k = m = 2 spricht man von einer Vierfeldertafel. Eine Kontingenztafel (nachstehend eine für absolute Häufigkeiten) wird häufig noch durch die Randverteilungen von X und Y ergänzt:

	b_1	b_2	...	b_j	...	b_m	Zeilensummen (Randverteilung von X)
a_1	h_{11}	h_{12}	...	h_{1j}	...	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2m}	$h_{2\cdot}$
:	:	:		:		:	:
a_i	h_{i1}	h_{i2}	...	h_{ij}	...	h_{im}	$h_{i\cdot}$
:	:	:		:		:	:
a_k	h_{k1}	h_{k2}	...	h_{kj}	...	h_{km}	$h_{k\cdot}$
Spaltensummen (Randverteilung von Y)	$h_{\cdot 1}$	$h_{\cdot 2}$	...	$h_{\cdot j}$	...	$h_{\cdot m}$	n

Aus einer solchen Tabelle lässt sich der χ^2-Koeffizient (*Chi²*-Koeffizient) gemäß

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(h_{ij}-e_{ij})^2}{e_{ij}}$$

bestimmen mit $e_{ij} = \frac{1}{n}\cdot$ (Produkt aus i-ter Zeilensumme und j-ter Spaltensumme) $= \frac{1}{n}\, h_{i\cdot}\, h_{\cdot j}$

Der Kontingenzkoeffizient *C* und der korrigierte Kontingenzkoeffizient C_{korr} ergeben sich als

$$C = \sqrt{\frac{\chi^2}{\chi^2+n}}\,, \qquad C_{korr} = \frac{C}{C_{\max}}$$

Dabei ist $C_{\max} = \sqrt{\frac{q-1}{q}}$ mit q = min (k,m).

Erwartungswert, Varianz und Standardabweichung einer diskreten Zufallsvariablen

Sei X eine Zufallsvariable, die *n* Ausprägungen x_1, x_2, …, x_n annehmen kann. Es seien $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$, …, $p_n = P(X = x_n)$ die zugehörigen Eintrittswahrscheinlichkeiten. Diese definieren die Wahrscheinlichkeitsfunktion der diskreten Variablen *X.* Der mit μ *oder E(X)* bezeichnete Erwartungswert von *X* ist gegeben durch

$$\mu = E(X) = \sum_{i=1}^{n} p_i x_i$$

und die mit σ^2 oder *Var(X)* abgekürzte Varianz sowie die *Standardabweichung* σ *(genauer: theoretische Varianz* und *theoretische Standardabweichung,* weil nicht aus Daten abgeleitet*)* durch

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} p_i (x_i - \mu)^2, \quad \sigma = \sqrt{\sigma^2} .$$

Zusammenhangsmessung bei ordinalskalierten ubnd intervallskalierten Merkmalen

Gegeben seien intervallskalierte Merkmale *X* und *Y*, für die n Beobachtungen $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ vorliegen. Die Kovarianz der Merkmale (genauer: empirische Kovarianz) ist dann gegeben durch

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

und der Korrelationskoeffizient für diese Merkmale nach Bravais-Pearson durch

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} .$$

Die letzte Gleichung ist äquivalent mit der Darstellung

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} .$$

Für ordinalskalierte Merkmale *X* und *Y* ist das Zusammenhangsmaß r_{xy} nicht anwendbar, weil $\bar{x}$ und $\bar{y}$ nicht definiert ist. Man ordnet daher die Beobachtungswerte x_1, x_2, …, x_n sowie y_1, y_2, …, y_n nach aufsteigender Größe und ordnet jedem Wert x_i bzw. y_i einen Rangplatz $r(x_i)$ resp. $r(y_i)$ zu. Als Zusammenhangsmaß ist dann der Rangkorrelationskoeffizient r_S nach Spearman anwendbar. Dieser ist wie folgt definiert, wenn $d_i = r(x_i) - r(y_i)$ die Rangplatzdifferenzen bezeichnen:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}.$$

Grundbegriffe der Kombinatorik

Wenn man aus einer Grundgesamtheit von *N* Elementen eine Stichprobe des Umfangs n zieht, so ist die Anzahl *K* der möglichen Stichproben zum einen davon abhängig, ob die Ziehung der einzelnen Elemente mit oder ohne Zurücklegen erfolgt, und zum anderen auch davon, ob die Reihenfolge, mit der die einzelnen Elemente gezogen werden, berücksichtigt wird oder nicht. Die in den vier Fällen resultierenden Werte *K* sind nachstehend ausgewiesen:

Art der Stichprobenziehung	Ziehen ohne Zurücklegen	Ziehen mit Zurücklegen
Ziehen mit Berücksichtigung der Reihenfolge	$\frac{N!}{(N-n)!}$	N^n
Ziehen ohne Berücksichtigung der Reihenfolge	$\binom{N}{n}$	$\binom{N+n-1}{n}$

Dabei ist z. B. $\binom{N}{n}$ (lies: Binomialkoeffizient N über n) definiert gemäß

$$\binom{N}{n} = \frac{N!}{(N-n)!n!}$$

mit *N! = 1 · 2 · … · N* (lies: N-Fakultät) und analog definierten Termen *n!* und *(N-n)!*

Bernoulli-Experimente und Binomialverteilung

Ein Experiment mit zwei möglichen Ausgängen A und $\overline{A}$ heißt auch Bernoulli-Experiment. Das Ergebnis eines Bernouilli-Experiments ist also die Realisation einer Zufallsvariablen mit zwei Ausprägungen. Ist *p = P(A)* die Wahrscheinlichkeit für den Eintritt des Ereignisses A, so ist die Wahrscheinlichkeit *P($\overline{A}$)* für den Eintritt des Komplementärereignisses durch *1-p* gegeben.

Wenn man ein solches Bernoulli-Experiment *n*-mal durchführt, spricht man von einer Bernoulli-Kette. Zählt man bei einer aus *n* Einzelexperimenten bestehenden Bernoulli-Kette die Anzahl *X* der Experimente, bei denen A als Ergebnis auftritt, so ist *X* ebenfalls eine Zufallsvariable. Die Verteilung der Zählvariablen *X* heißt Binomialverteilung. Sie kann offenbar die Werte 0, 1, 2, … n annehmen.

Da es bei Durchführung von *n* Bernoulli-Experimenten $\binom{n}{k}$ Möglichkeiten gibt, dass insgesamt *k*-mal das Ereignis A eintritt und die Eintrittswahrscheinlichkeit für jeden der $\binom{n}{k}$ Fälle $p^k(1-p)^{n-k}$ ist, gilt: Die Wahrscheinlichkeit *P(X = k)* dafür, dass für *X* die Ausprägung *X = k* beobachtet wird, ist

$$P(X = k) = \binom{n}{k} p^k(1-p)^{n-k} \qquad k = 0,1, 2, ..,n$$

Diese Gleichung, in die die Parameter *n* und *p* eingehen, ist die Wahrscheinlichkeitsfunktion einer binomialverteilten Zufallsvariablen *X*. Für eine mit Parametern *n* und *p* binomialverteilte Variable *X* sagt man auch, dass sie *B(n, p)*-verteilt sei und anstelle von *P(X = k)* schreibt man dann oft *B(X = k | n, p)*. Für den Erwartungswert $\mu = E(X)$ und die Varianz $\sigma^2 = Var(X)$ einer *B(n, p)*-verteilten Zufallvariablen *X* hat man

$$\mu = E(X) = np, \qquad \sigma^2 = Var(X) = np \cdot(1 - p).$$