

Name:	Vorname:	Matrikel-Nr.:
-------	----------	---------------

## Klausur zur Vorlesung Statistik I und II für Diplom-Studenten, SS 2008

### Aufgabe 1

**(20 Punkte)**

Die deutsche Unfallstatistik listet für das Jahr 2007 alle Unfälle, welche der Polizei gemeldet wurden. Diese sollen bezüglich der an einem Unfall beteiligten Fahrzeuge und Personen, der jeweiligen Zahl der Todesopfer, Schwer- und Leichtverletzten sowie des entstandenen Sachschadens untersucht werden. Dabei wird auch nach Unfallursache (z.B. zu hohe Geschwindigkeit oder Vorfahrtsverletzung), nach Alter und Geschlecht des Haupt-Unfallverursachers, des Verkehrsmittels des Hauptverursachers und der Verletzten und Getöteten (Kfz, Rad, zu Fuß oder ÖPNV) sowie nach Stadtstraßen-, Landstraßen- und Autobahnunfällen unterschieden.

- (a) Geben Sie die statistische Gesamtheit und den Merkmalsträger an. Grenzen Sie die statistische Gesamtheit räumlich, zeitlich und sachlich ab.
- (b) Geben Sie die in der Aufgabenstellung erwähnten Merkmale und ihre Skalierung an. Unterscheiden Sie dabei auch die Verhältnisskalierung von der Absolutskalierung.
- (c) Es wird ein Gesetzesvorschlag eingebracht, nach dem ältere Personen einen regelmäßigen Fahrtauglichkeitstest machen müssen, um ihren Führerschein zu behalten. Dies soll mit Hilfe der Unfallstatistik begründet werden. Welche Kombination der obigen Merkmale würde man dabei zugrunde legen? Welche wichtigen zusätzlichen Angaben fehlen für eine seriöse Begründung?

### Aufgabe 2

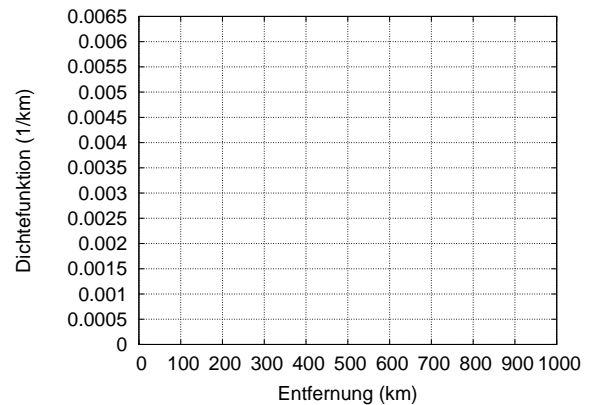
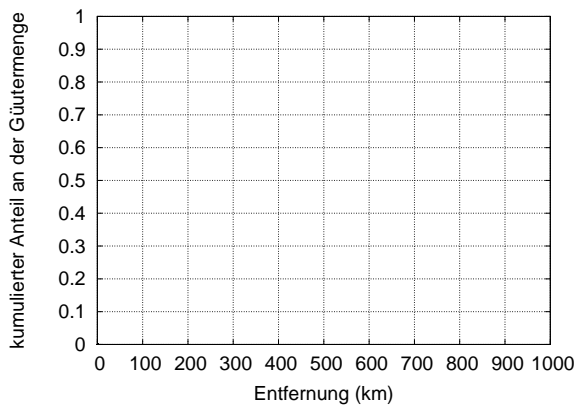
**(60 Punkte)**

Einer Verkehrsstatistik entnimmt man folgende Aufgliederung der transportierten Güter (in Millionen Tonnen) nach verschiedenen Verkehrsmitteln und Entfernungsstufen:

Entfernung (km)	Gütermenge LKW (in Mio t)	Gütermenge Bahn (in Mio t)	Gütermenge Schiff (in Mio t)
0 – 50	1 020	9	0
50 – 100	660	39	10
100 – 200	340	84	95
200 – 500	150	55	40
500 – 1 000	60	21	7

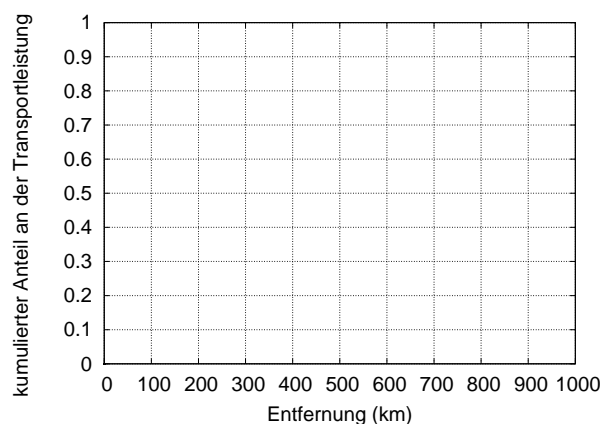
Name:	Vorname:	Matrikel-Nr.:
-------	----------	---------------

- (a) Wieviel Tonnen Güter wurden mit den drei Verkehrsarten jeweils transportiert?
- (b) Wie ist die Aufteilung der Transportmenge auf die drei Verkehrsarten?
- (c) Berechnen Sie nun für die drei Verkehrsarten jeweils die *Transportleistung* in Milliarden tkm (Tonnen-Kilometer). Nehmen Sie dabei an, dass innerhalb einer Entfernungs-kategorie die Transportweiten gleichverteilt sind. Berechnen Sie auch die Anteile der drei Verkehrsarten an der Leistung. Erstellen Sie Tortendiagramme für die transportierte Menge und die Transportleistung.
- (d) Berechnen Sie für die drei Verkehrsmittel jeweils die mittlere Transportweite.
- (e) Konzentrieren Sie sich nun auf den **Schiffstransport**: Berechnen Sie die Verteilungs- und Dichtefunktionen der transportierten Mengen bezüglich der Transportweite und zeichnen Sie das Ergebnis in die beiden Diagramme ein.



Lesen Sie von einem der Diagramme den Median ab.

- (f) Berechnen Sie nun die Verteilungsfunktion der *Transportleistung* im Güterschiffsverkehr bezüglich der Entfernung und tragen Sie das Ergebnis in das folgende Diagramm ein:



- (g) Schiffe werden stichprobenartig auf geschmuggelte bzw. verbotene Ladung untersucht. Wie hoch ist die Wahrscheinlichkeit, dass die überprüften Güter zwischen 500 km und 1000 km weit transportiert werden, wenn die Stichproben (i) am Start- oder Zielpunkt stattfinden, (ii) die Kontrollen irgendwo unterwegs durchgeführt werden?

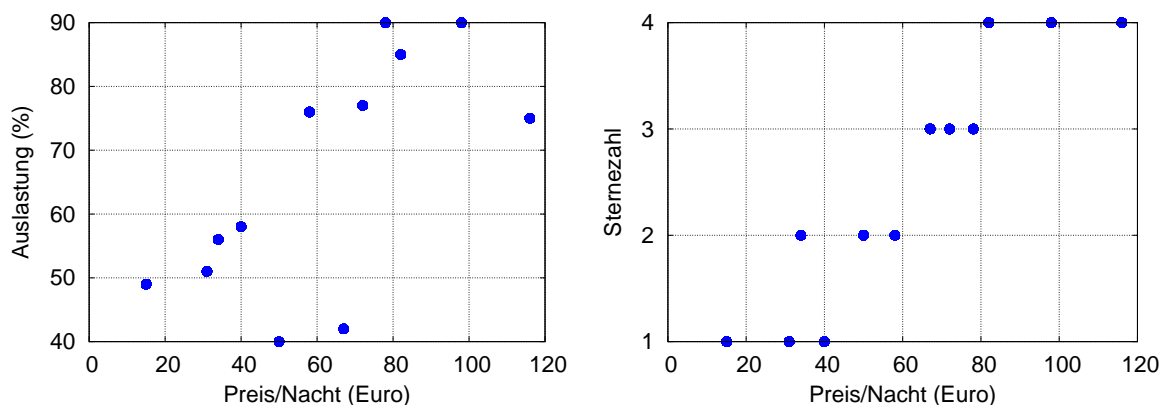
Name:	Vorname:	Matrikel-Nr.:
-------	----------	---------------

**Aufgabe 3****(40 Punkte)**

Gegeben ist folgende Statistik der Übernachtungspreise, Auslastungen und Sternezahlen einiger Hotels in Dresden:

Übernachtungspreis (in €)	72	67	78	40	34	50	98	116	82	31	58	15
Auslastung (%)	77	42	90	58	56	40	90	75	85	51	76	49
Sternezahl	3	3	3	1	2	2	4	4	4	1	2	1

Ihre Statistik-Software erzeugt aus den Daten folgende Streudiagramme:



Außerdem gibt sie folgende Lagemaße, Varianzen und Kovarianzen bezüglich des Preises  $x$  (in €) und der Auslastung  $y$  (in %) aus:

$$\bar{x} = 61.75, \quad \bar{y} = 65.75, \quad s_x^2 = 799.2, \quad s_y^2 = 312.0, \quad s_{xy} = 341.3.$$

- Führen Sie eine lineare Regression der Auslastung als Funktion des Preises durch und zeichnen Sie den Graph der Regressionsfunktion in das linke Streudiagramm ein.
- Was sind für diesen Sachverhalt die offensichtlichen Schwächen der linearen Regression? Zeigen Sie anhand eines frei wählbaren Zahlenbeispiels, dass diese Regressionsfunktion zu unsinnigen Aussagen führen kann.
- Sie erhalten in Aufgabenteil (a) das Ergebnis, dass die Auslastung mit dem Übernachtungspreis tendenziell steigt. Ist dies unsinnig oder plausibel? Begründen Sie ihre Antwort mit dem rechten Streudiagramm und der vom Statistikprogramm gelieferten Korrelation  $r_{xz} = 0.929$  zwischen Übernachtungspreis und Sternezahl  $z$ .
- Geben Sie für die Abhängigkeit der Auslastung vom Hotelpreis die erklärte Varianz, die nicht erklärte Varianz (Residualvarianz), die Gesamtvarianz und das Bestimmtheitsmaß an.

Name:	Vorname:	Matrikel-Nr.:
-------	----------	---------------

**Aufgabe 4****(30 Punkte)**

Bei der Korrektur eines Dokuments auf Rechtschreibfehler entdeckte der erste Überprüfer 5 Fehler und der zweite 4 Fehler. Darunter waren zwei Fehler, welche von beiden entdeckt wurden. Man kann davon ausgehen, dass die beiden Personen unabhängig voneinander korrigiert haben und die Wahrscheinlichkeit für das Aufspüren eines Fehlers bei allen Fehlern dieselbe ist.

- Nutzen Sie die Regeln für bedingte Wahrscheinlichkeiten und Wahrscheinlichkeiten unabhängiger Ereignisse, um die Gesamtzahl der Fehler abzuschätzen. Wieviele unentdeckte Fehler sind danach zu erwarten?
- Könnten es auch mehr oder weniger sein? Nehmen Sie dazu als Nullhypothese  $H_0$  an, dass sich tatsächlich 10 Fehler eingeschlichen haben und die Wahrscheinlichkeit, einen Fehler aufzuspüren, bei der ersten Person bei  $1/2$  und bei der zweiten bei  $2/5$  liegt. Berechnen Sie unter dieser Nullhypothese die Verteilung der von der ersten und zweiten Person entdeckten Fehlerzahl sowie die Verteilung der Gesamtzahl der entdeckten Fehler.

**Aufgabe 5****(25 Punkte)**

Für neue Fahrerassistenzsysteme (z.B. ein Kreuzungsassistent oder eine verbesserte Navigation) ist es wichtig, zu bestimmen, wie stark sich die Fahrtrichtung innerhalb eines bestimmten Zeitraums ändert. Diese Änderung kann prinzipiell auf dreierlei Arten gemessen werden:

- Über den aufgezeichneten Lenkwinkel und die Geschwindigkeit,
- über einen "Gierratensensor", welcher direkt die Drehung des Fahrzeuges misst,
- und über die Auswertung der vom Navigationssystem gelieferten Positionen.

Für eine tatsächliche Änderung der Fahrtrichtung von 30 Grad liefern alle Methoden im Mittel das richtige Ergebnis, aber mit Unsicherheiten (einfachen Standardabweichungen) von  $\sigma_1 = 2$  Grad,  $\sigma_2 = 2$  Grad und  $\sigma_3 = 4$  Grad. Kann man durch Kombinationen der einzelnen Messungen den Fehler verringern?

- Berechnen Sie zunächst die Standardabweichung der Richtungsbestimmung, falls dazu das arithmetische Mittel der drei Messungen verwendet wird.
- Kann man die Genauigkeit optimieren, indem man die "schlechtere" Messmethode 3 ignoriert oder weniger gewichtet? Bestimmen Sie dazu die Varianz eines gewichteten arithmetischen Mittels der drei Messungen, wobei Sie die ersten beiden Messungen gleich gewichten. Minimieren Sie die Varianz durch Variation der Gewichtungen. Mit welchem Anteil trägt die Messmethode 3 im optimalen Fall noch bei? Wie hoch ist die resultierende Varianz?

*Hinweis:* Die Summe der Wichtungen muss 1 betragen!

Name:	Vorname:	Matrikel-Nr.:
-------	----------	---------------

**Aufgabe 6****(15 Punkte)**

Eine bestimmte Bank benötigt für ein Jahr Geld und kann sich dieses nur (von anderen Banken oder Investoren) ausleihen, wenn sie 6% Zinsen bietet. Ausfallsichere Bundesanleihen bieten nur 4% Zinsen. Wie hoch ist die von den Marktteilnehmern angenommene subjektive Wahrscheinlichkeit dafür, dass diese Bank innerhalb des Jahres Pleite geht (und damit das ausgeliehene Geld samt Zinsen verloren ist) ?

**Aufgabe 7****(55 Punkte)**

Bei der Neueinschreibung an eine Universität bewerben sich die Studenten im allgemeinen an mehreren Universitäten gleichzeitig. Die Universitäten nehmen die geeigneten Kandidaten an. Allerdings erhalten viele Studenten von mehreren Universitäten Einladungen und wählen nun ihrerseits aus. Um die Studienplätze eines Studiengangs möglichst optimal zu besetzen, werden daher mehr Studenten akzeptiert als Studienplätze zur Verfügung stehen.

In einem bestimmten Studiengang mit 200 Plätzen ergaben sich in den letzten Jahren bei den ersten 200 Einladungen folgende Absagen von Seiten der Studenten:

Jahr	2005	2006	2007	2008
Zahl der Absagen	83	103	63	111

- Geben Sie Schätzer für den Erwartungswert und die Varianz der Zahl der Absagen bzw. der Absagerquote an.
- Mit welcher Wahrscheinlichkeit ergibt sich höchstens 80 Absagen, wenn Sie eine Gaußverteilung der Absagenquoten mit einem Erwartungswert von 45% und einer Standardabweichung von 10% annehmen?
- Berechnen Sie das Konfidenzintervall der Absagenquote zur Fehlerwahrscheinlichkeit von 5%.
- Es gibt zwei mögliche Ursachen für die Schwankungen: (i) Schwankungen durch die endliche Zahl der Studenten, (ii) Schwankungen der wahrgenommenen Attraktivität des Studiengangs von seiten der Studenten von einem Jahr zum nächsten. Welche Schwankungsursache ist hier größer? Nehmen Sie dazu Unabhängigkeit der beiden Ursachen an und vergleichen Sie die in (a) geschätzte Gesamtvarianz mit der erwarteten Varianz bei konstanter Attraktivität. In letzterem Fall gilt für jeden Studenten die gleiche Absagewahrscheinlichkeit, welche Sie wieder mit 45% annehmen können.

*Hinweis:* Die Zahl der absagenden Studenten ist binomialverteilt.

- Um den Studiengang trotz der Absagen zu füllen, werden an die Studenten 350 Einladungen für die 200 Plätze verschickt. Wie groß ist das Risiko, dass sich mehr als 200 Studenten einschreiben und damit der Studiengang überfüllt ist? Nehmen Sie nun an, dass die Schwankungen von Jahr zu Jahr überwiegen, so dass der erwartete *Anteilswert* der Absagen, nicht aber der Erwartungswert der *Anzahl* der Absagen unabhängig von der Zahl der angenommenen Studenten ist.

# Tabellen

## Werte $\Phi(z)$ der Standardnormalverteilung

$z$	$\Phi(z)$	$z$	$\Phi(z)$	$z$	$\Phi(z)$	$z$	$\Phi(z)$
0.0	0.5000	1.00	0.8413	2.0	0.9772	3.0	0.9987
0.1	0.5398	1.1	0.8643	2.1	0.9821	3.1	0.9990
0.2	0.5793	1.2	0.8849	2.2	0.9861	3.2	0.9993
0.3	0.6179	1.3	0.9032	2.3	0.9893	3.3	0.9995
0.4	0.6554	1.4	0.9192	2.4	0.9918	3.4	0.9997
0.5	0.6915	1.5	0.9332	2.5	0.9938	3.5	0.9998
0.6	0.7257	1.6	0.9452	2.6	0.9953	3.6	0.9998
0.7	0.7580	1.7	0.9554	2.7	0.9965	3.7	0.9999
0.8	0.7881	1.8	0.9641	2.8	0.9974	3.8	0.9999
0.9	0.8159	1.9	0.9713	2.9	0.9981	3.9	1.0000

## Quantile $t_n^{(\alpha)}$ der Studentischen $t$ -Verteilung

$n$	$\alpha = 0.60$	0.70	0.80	0.90	0.95	0.975	0.990	0.995	0.999	0.9995
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
$\infty$	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291