**Klausur: Einführung in die Statistik**
**Sommersemester 2014**

1. Für eine statistische Analyse wird eine Reihe von Merkmalen erfasst- Notieren Sie jeweils mit N, O, I bzw. R, ob es sich um eine Nominal-, Ordinal-, Intervall- oder Ratio-Skala handelt. (2 Punkte)
   __ Die Größe einer Person in Zentimetern
   __ Das Geschlecht einer befragten Person
   __ Die Zufriedenheit eines Befragten mit der Arbeit des Bundesrates auf einer 10er Skala (1: sehr gut, 10: sehr schlecht)
   __ Der Sitzanteil der Christdemokraten im Landtag von Mecklenburg-Vorpommern
   __ Schulnoten (von 1 bis 6)
   __ Körpertemperatur (in Grad Celsius)

2. Welche der folgenden Aussagen zu Lagemaßen ist/sind richtig? (2 Punkte)
   - Der Modus ist der zweithäufigste Wert einer Verteilung
   - Der Median teilt eine Verteilung in zwei gleich große Hälften
   - Modus, Median und arithmetisches Mittel lassen sich nicht für nominalskalierte Daten berechnen
   - Der Median kann ausschließlich für ordinalskalierte Daten berechnet werden

3. Welche der folgenden Aussagen zu Streuungsmaßen ist/sind richtig? (2 Punkte)
   - Streuungsmaße geben Auskunft über die Verteilung der Werte in der Datenmenge
   - Die Varianz entspricht der quadrierten Standardabweichung
   - Die Varianz berechnet sich aus der Different zwischen Minimal- und Maximalwert der Verteilung
   - Je größer die Standardabweichung, desto weiter streuen die Werte um das arithmetische Mittel

4. Mit einem Histogramm (Abbildung 1) wird folgendes dargestellt: (2 Punkte)

4. Mit einem Histogramm (Abbildung 1) wird Folgendes dargestellt (2 Punkte):

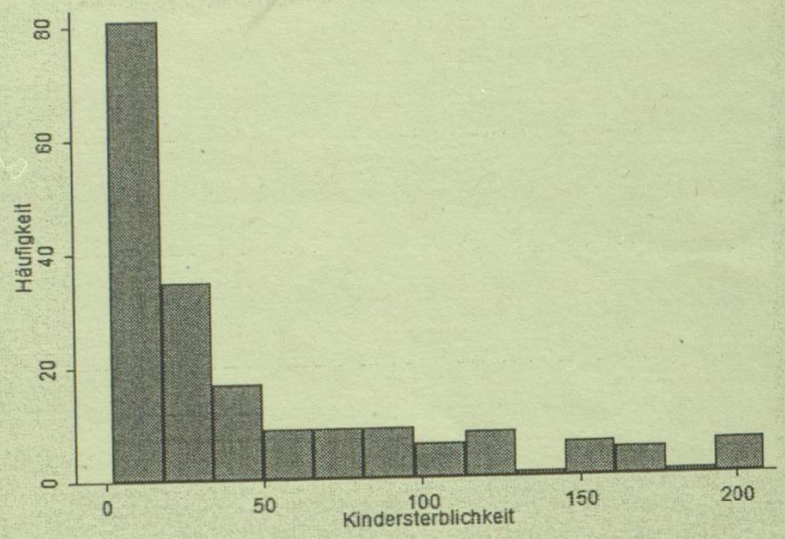



**Abbildung 1: Histogramm – Kindersterblichkeit**
**(pro 1000 Geburten vor dem 5. Lebensjahr verstorbenen Kinder).**

- Die Häufigkeit von ordinalen Merkmalen
- Die bivariate Verteilung von nominalen Merkmalen
- Die Häufigkeit von metrischen Variablen
- Die Signifikanz einer bivariaten Verteilung

5. Welche Aussagen treffen auf die in Abbildung 1 gezeigte Verteilung zu? (2 Punkte)
   - Es handelt sich um eine Normalverteilung
   - Die Verteilung ist rechtsschief
   - Die Verteilung ist rechtssteil
   - Das erste Moment der Verteilung liegt bei deutlich über 150 Todesfälle pro 1000 Geburten

6. Bringen Sie die Schritte eines Hypothesentest in die richtige Reihenfolge, indem sie diese nummerieren. (3 Punkte)
   __ Festlegung der Prüfgröße
   __ Prüfgröße berechnen
   __ Bestimmung der Irrtumswahrscheinlichkeit
   __ Formulierung von Null- und Alternativhypothese
   __ Über Annahme der Hypothese entscheiden

7. Was gilt für statistische Hypothesentests? (3 Punkte)
   - Mittels einer Prüfstatistik kann das Signifikanzniveau zur Annahme der Alternativhypothese H1 bestimmt werden
   - Getestet wird stet die Nullhypothese H0
   - Wenn der p-Wert kleiner als alpha ist ($p < \alpha$), wird die Nullhypothese H0 verworfen
   - Getestet werden können nur Unterschieds- nicht jedoch Zusammenhangshypothesen

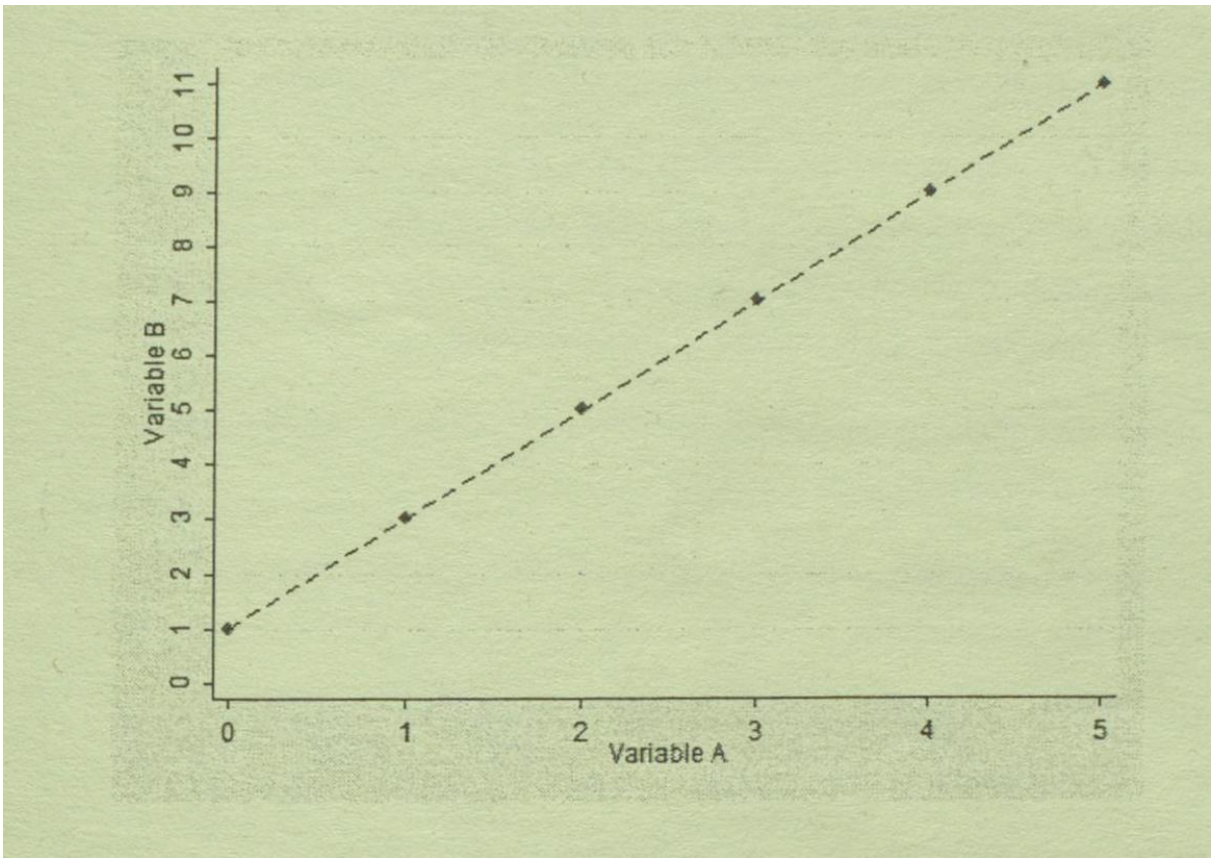
8. Was sind alpha- und beta-Fehler? ( 3 Punkte)
    - Der alpha-Fehler bezeichnet die Irrtumswahrscheinlichkeit, die für die Zuverlässigkeit eines statistischen Tests gerade noch toleriert wird
    - Der beta-Fehler entspricht dem Signifikanzniveau
    - Der beta-Fehler ist die fälschliche Ablehnung der Alternativhypothese H1
    - Der alpha-Fehler ist die fälschliche Ablehnung der Nullhypothese H0

9. Welche der folgenden Aussagen zum Konfidenzintervall ist/sind richtig? (3 Punkte)
    - Ein Konfidenzintervall ist ein Wertebereich, bei dem wir darauf vertrauen können, dass er den wahren Wert der Population mit einer bestimmten Wahrscheinlichkeit überdeckt
    - Das Konfidenzintervall ist dasselbe wie das Vertrauensintervall
    - Das Konfidenzintervall ist ein Signifikanztest für die Nullhypothese H0
    - Konfidenzintervalle sind dazu da, die Güte eines Regressionsmodells zu bewerten

10. Welche der folgenden Aussagen zu Korrelation ist/sind korrekt? (2 Punkte)
    - Spearman's r basiert auf einem Paarvergleich
    - Korrelationen resten, ob ein Zusammenhang zwischen zwei Variablen kausal ist
    - Spearman's r und Pearson's r können Werte von -1 bis +1 annehmen
    - Wenn ein Korrelationskoeffizient einen Wert >0,5 annimmt, kann von einem signifikanten Zusammenhang gesprochen werden

11. Welche der folgenden Aussagen zu linearen Regression ist/sind richtig? (3 Punkte)
    - Die Variablen müssen ein metrischen Messniveau aufweisen
    - Je negativer das R2 einer Regression, desto geringer ist der Zusammenhang zwischen den Variablen
    - Multikollinearität zwischen den unabhängigen Variablen ist bei der Schätzung einer linearen Regression problematisch
    - Heteroskedastizität ist ein Maß für die Güte einer Regression

12. Abbildung 2 zeigt einen fiktiven linearen Zusammenhang zwischen zwei Variablen A und B. Bitte lesen Sie aus der Abbildung die folgenden Regressionskoeffizienten ab. (4 Punkte)




Regressionskonstante: _____
Regressionsgewicht: _____

13. Welcher der folgenden Aussagen zu den Prüfstatistiken ist/sind korrekt? (3 Punkte)
    - Der F-Test wurde von Fischer entwickelt
    - Bei der Regressionsanalyse wird der F-Test genutzt, um die Signifikanz der Regressionskonstante zu testen
    - Der F-Test gibt Auskunft über die Signifikanz des gesamten Regressionsmodells
    - Der t-Test vergleicht die erklärte Varianz eines Regressionsmodells mit der unerklärten Varianz

**In einem (fiktiven) Forschungsprojekt soll untersucht werden, wie sich die Höhe der Kindersterblichkeit in 135 Staaten erklären lässt.**

Untersucht werden insgesamt 135 Staaten. Für diese wird (als abhängige Variable) die Kindersterblichkeit in Todesfällen vor dem 5. Lebensjahr pro 1000 Geburten erhoben.

Die Forscherinnen und Forscher nehmen an, dass die zentrale unabhängige Variable der Urbanisierungsgrad des jeweiligen Staates ist, da in städtischen Gebieten die Gesundheitsversorgung besser ist als auf dem Land. Gemessen wird der Urbanisierungsgrad als prozentualer Anteil der Bevölkerung, die in städtischen Gebieten wohnt.

Zum anderen vermuten die Forscherinnen und Forscher, dass die Kindersterblichkeit zusätzlich durch folgende Kontrollvariablen beeinflusst wird:

2. die Alphabetisierungsrate, gemessen als prozentualer Anteil derjenigen, die Lesen und Schreiben können an der erwachsenen Bevölkerung;

3. die Fertilitätsrate unter sehr jungen Frauen, gemessen als Anzahl der Geburten pro 1000 Frauen im Alter zwischen 15 und 19 Jahren.

Diese drei Einflussfaktoren gehen als unabhängige Variablen in die Analyse ein.

**Zunächst wollen die Forscherinnen und Forscher etwas über die Verteilung ihrer abhängigen Variable wissen und berechnen folgende univariate deskriptive Statistiken.**

| Variable | N | Arithm. Mittel | Median | Standard-abweichung | Minimum | Maximum |
|---|---|---|---|---|---|---|
| **Kinder-sterblichkeit** | 135 | 56,1 | 24 | 54,5 | 3 | 209 |

**Tabelle 1: Univariate Deskription**

14. Welche der folgenden Aussagen zur Tabelle 1: „Univariate Deskription“ ist/sind richtig? (2 Punkte?)
    - Im Durchschnitt aller untersuchten Fälle liegt die Kindersterblichkeit bei 56,1 Fällen pro 1000 Geburten
    - Im Verhältnis zum Median zeigt die Standardabweichung, dass die Variable Kindersterblichkeit ungefähr normalverteilt ist
    - In höchstens 50% der Fälle ist die Kindersterblichkeit niedriger als 54,5 Fälle pro 1000 Geburten
    - Die durchschnittliche Streuung aller Werte um das arithmetische Mittel beträgt 54,5

**Dann untersuchen die Forscherinnen und Forscher zunächst mit Hilfe einer Korrelation, ob zwischen der Kindersterblichkeit als abhängiger und dem Urbanisierungsgrad als unabhängiger Variable ein systematischer Zusammenhang besteht.**

| | Kindersterblichkeit | Urbanisierungsgrad |
|---|---|---|
| Kindersterblichkeit | 1,0 | |
| Urbanisierungsgrad | -0,6 | 1,0 |

**Tabelle 2: Korrelationsmatrix – Kindersterblichkeit und Urbanisierungsgrad**

15. Welche der folgenden Aussagen zu Tabelle 2 ist/sind richtig? (2 Punkte)
    - Es besteht ein negativer Zusammenhang zwischen Kindersterblichkeit und Urbanisierungsgrad
    - Aus Tabelle 2 können wir an Hand der beiden Koeffizienten mit dem Wert „1,0“ ablesen, das Kindersterblichkeit und Urbanisierungsgrad perfekt korrelieren.
    - Aufgrund der Skalenniveaus hätten Kindersterblichkeit und Urbanisierungsgrad nicht gemeinsam in Tabelle 2 dargestellt werden dürfen
    - Der Urbanisierungsgrad erklärt 60% der Varianz der Variable Kindersterblichkeit

**Dann untersuchen die Forscherinnen und Forscher mit Hilfe eines Streudiagramms den Zusammenhang zwischen den unabhängigen Variablen Urbanisierungsgrad und Alphabetisierungsrate.**

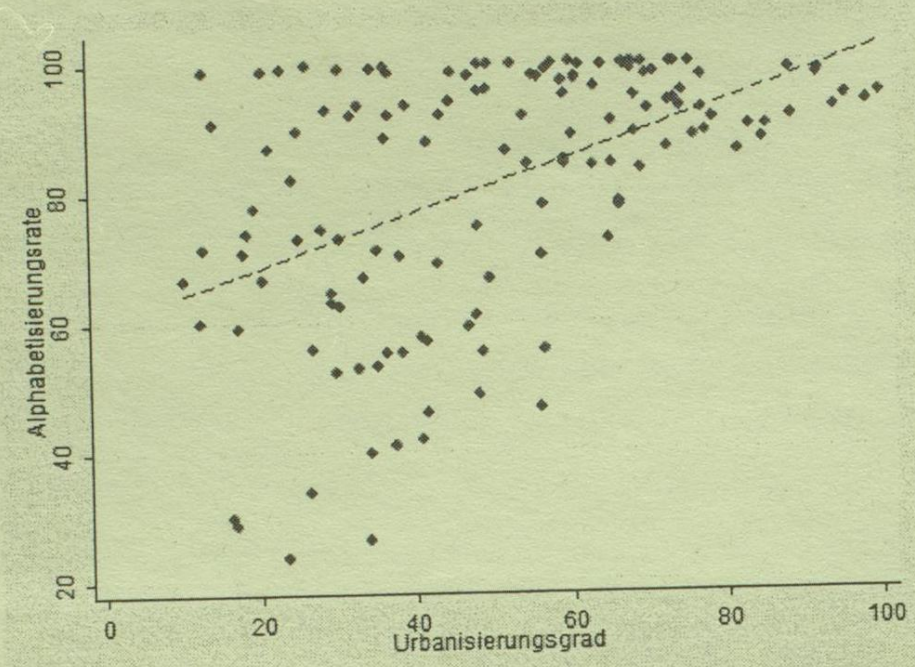



**Abbildung 3: Streudiagramm – Urbanisierungsgrad und Alphabetisierungsrate**

16. Welche der folgenden Aussagen zu dem Streudiagramm in Abbildung 3 ist/sind richtig? (3Punkte)
    - Da zwischen den Urbanisierungsgrad und Alphabetisierungsrate offenkundig ein Zusammenhang besteht, könnte es in Regressionsmodellen, die beide Indikatoren als unabhängige Variablen verwenden, zum Problem der Kollinearität kommen
    - Das Streudiagramm zeigt einen positiven Zusammenhang zwischen Urbanisierungsgrad und Alphabetisierungsrate

- Die Trichterform der Punktewolke weist auf eine homoskedastische Beziehung hin
- Aus Abbildung 3 lässt sich ablesen, dass der Urbanisierungsgrad mindestens 60% der Varianz der Variable Alphabetisierungsrate erklärt

**Anschließend führen die Forscherinnen und Forscher Regressionsanalysen zur Erklärung der Kindersterblichkeit durch. Sie formulieren zwei verschiedene Regressionsmodelle (Modell 1 und Modell 2), die sie vergleichen wollen. Modell 1 ist das minimale Modell mit nur einer unabhängigen Variable, Modell 2 umfasst auch die beiden Kontrollvariablen.**

| | Kindersterblichkeit | |
|---|---|---|
| | Modell 1 | Modell 2 |
| Urbanisierungsgrad | -1.413*** | -0.522*** |
| | (0.177) | (0.119) |
| Alphabetisierungsrate | | -1.111*** |
| | | (0.167) |
| Fertilitätsrate | | 0.508*** |
| | | (0.0670) |
| Konstante | 127.9*** | 143.3*** |
| | (9.789) | (16.16) |
| R-Quadrat | 0.32 | 0.77 |
| Korr. R-Quadrat | 0.32 | 0.77 |
| F-Test | 63,69*** | 147,68*** |
| Beobachtungen | 135 | 135 |

*Anmerkungen*:
*** p < .001, ** p < .01, * p < .05, zweiseitiger Test; Standardfehler in Klammern.

**Tabelle 3: Der Einfluss des Urbanisierungsgrades und zweier Kontrollvariablen auf die Kindersterblichkeit in 135 Staaten.**

17. Die allgemeine Formel für multivariate lineare Regressionsmodelle lautet:

$$Y = a + b_1X_1 + \ldots + b_nX_n + E$$

Stellen Sie für das multivariate Regressionsmodell (Modell 2) die Regressionsgleichung auf, indem Sie aus Tabelle 3 die entsprechenden Werte für die Regressionskoeffizienten (regressionskonstante und Regressionsgesichte) einsetzen: (3 Punkte)

Verwenden Sie dabei für die Bezeichnung der Variablen die folgenden Abkürzungen:
Kindersterblichkeit: KS; Urbanisierungsgrad: URBAN;
Alphabetisierungsrate: ALPHA; Fertilitätsrate: FERTIL.

Modell 2: _____

18. Wie ist das Ergebnis der Regressionsanalyse (vgl. Tabelle 3) zu interpretieren? Welche der folgenden Aussagen zu den Regressionskoeffizienten (Regressionskonstante und Regressionsgewichte) und zur Modellgüte ist/sind richtig? (3 Punkte)

- Wenn alle Menschen auf dem Land leben (Urbanisierungsrate = 0) liegt die Kindersterblichkeit laut Modell 1 bei –1,413 Kindern pro 1000 Geburten

- Modell 2 zeigt: steigt die Fertilitätsrate um eine Einheit an, steigt die Kindersterblichkeit um rund 50% an
- Modell 1 zufolge erklärt der Urbanisierungsgrad 32% der Varianz der Variable Kindersterblichkeit
- Eine Inspektion der R2 und der adjustierten (korrigierten) R2 von Modell 1 und Modell 2 zeigt: Die Hinzunahme der Kontrollvariablen steigert die Modellgüte

19. Welche der folgenden Aussagen zu den Teststatistiken der Regressionsanalyse (vgl. Tabelle 3) ist/sind richtig? (3 Punkte)

- Alle Regressionsgewichte der Modelle 1 und 2 sind statistisch hoch signifikant
- Modell 1 zufolge ist die Regressionskonstante statistisch nicht signifikant
- Der F-Test für Modell 2 zeigt: die Nullhypothese („keine der gewählten unabhängigen Variablen hat einen Einfluss auf die Kindersterblichkeit“) kann mit geringer Irrtumswahrscheinlichkeit zurückgewiesen werden. Das Modell ist statistisch hoch signifikant
- Die Alternativhypothese, die mit der Teststatistik für die einzelnen Regressionskoeffizienten getestet wird, lautet: „Die gewählte unabhängige Variable hat keinen Effekt auf die abhängige Variable.“