

Wiederholung

Statistik I

Sommersemester 2009

Rest: Anwendungsbeispiele

Wiederholung

Deskriptive Statistik

- Daten/graphische Darstellungen

- Lage- und Streuungsmaße

- Zusammenhangsmaße

- Lineare Regression

Inferenzstatistik

- Wahrscheinlichkeitsrechnung

- Zentraler Grenzwertsatz

- Konfidenzintervalle

- Hypothesentests

Zusammenfassung

Rohdaten etc.

- ▶ Skalenniveau
- ▶ Rohdatenmatrix & Permutationen
- ▶ Häufigkeitsauszählungen (eindimensionale Tabelle), absolute & relative Häufigkeiten
- ▶ Mehrdimensionale Tabellen (Untertabellen)
- ▶ Prozentuierungsarten
- ▶ Grundlage für viele graphische Darstellungen

Graphische Darstellungen

- ▶ Dreidimensionale Darstellungen vermeiden
- ▶ Histogramm/Dichteschätzung vs. Balkendiagramm
- ▶ Tortendiagramme
- ▶ Zwei- und mehrdimensionale Darstellungen
- ▶ Streudiagramme
- ▶ Zeitreihen und Kartogramme

Mittelwerte, Streuungs- und Lagemaße

1. Mittelwerte

- ▶ Modus
- ▶ Median und Quartile
- ▶ Arithmetisches Mittel

2. Streuungsmaße

- ▶ Spannweite (range), Interquartilsabstand
- ▶ Standardabweichung/Varianz
- ▶ (in der Stichprobe/in der Grundgesamtheit)

3. Schiefe (skewness)

4. Wölbung (kurtosis)

5. z-Standardisierung

Zwei nominalskalierte Merkmale

- ▶ Kreuztabelle, zwei nominale Merkmale z. B. Konfession \times Wahlverhalten
- ▶ Spalten- vs. Zeilen- vs. Gesamtprozent
- ▶ Zusammenhang = Abweichung von zufälliger Verteilung

1. PRE-Maß λ

2. Maße auf Basis von χ^2

- ▶ χ^2 : Empirische vs. Indifferenztablelle
- ▶ „Standardisierung“ von χ^2
- ▶ ϕ , C , V

Zwei ordinalskalierte Variablen

- ▶ Tabelle muß sortiert sein (Werte aufsteigend von links nach rechts, oben nach unten)
- ▶ positiver Zusammenhang: „mehr“ von einer Variable – „mehr“ von der anderen
- ▶ Logik Paarvergleich: konkordante Paare, diskordante Paare, ties
- ▶ γ
- ▶ τ_b (berücksichtigt ties)
- ▶ Somers' D (asymmetrische Variante von τ_b)

Eine nominalskalierte, eine intervallskalierte Variable

- ▶ Nominalskalierte Variable definiert Gruppen
- ▶ Zusammenhang = Homogenität, Streuung innerhalb Gruppen kleiner als Gesamtstreuung
- ▶ η^2 : SAQ_{gesamt} vs. $SAQ_{Gruppe1} + SAQ_{Gruppe2} + \dots$
- ▶ Perfekter Zusammenhang: keine Streuung innerhalb der Gruppen

Zwei intervallskalierte Variablen

- ▶ Zusammenhang: gemeinsame Abweichung vom jeweiligen Mittelwert
- ▶ Positiv: mehr von x , mehr von y ; negativ: mehr von x , weniger von y
- ▶ Berechnung: Summe der Abweichungsprodukte (SAP, vergleichbar mit χ^2)
- ▶ Standardisierung von SAP:
 - ▶ Division durch Zahl der Fälle \rightarrow Kovarianz
 - ▶ Division durch Produkt der beiden Standardabweichungen
- ▶ Korrelationskoeffizient r

Warum Regression?

- ▶ Statistisches Modell
- ▶ Abhängige Variable y als Funktion von
 - ▶ Unabhängigen Variablen
 - ▶ Zufälligen Einflüssen
- ▶ Regression \neq Kausalität
- ▶ Einfachstes Modell: lineare Einfachregression
 - ▶ Eine unabhängige Variable
 - ▶ Lineare (konstante) Wirkung
- ▶ Ausgangspunkt für komplexere Modelle

Lineare Einfachregression: Erweiterungen

- ▶ „Scheinbeziehungen“ → statistische Kontrolle
- ▶ Mehrere unabhängige Variablen
- ▶ Nicht-lineare, z. B. kurvilineare (U-förmige) Zusammenhänge
- ▶ Non-parametrische Regression
- ▶ Interaktion: Wirkung von x_1 hängt ab vom *Niveau* von x_2 und umgekehrt (Bsp.: Wahl Front National = Konstante + Arbeitslosenquote + Ausländeranteil + Arbeitslosenquote \times Ausländeranteil)

Regression: Qualität

- ▶ R^2 : Anteil der systematischen Einflüsse an der Gesamtvarianz von y
- ▶ Root Mean Squared Error: Wurzel aus der mittleren quadrierten Differenz zwischen empirischem y und prognostiziertem Wert

Inferenzstatistik

- ▶ Übertragbarkeit Stichprobenergebnisse – Grundgesamtheit?
- ▶ Basiert auf Wahrscheinlichkeitsrechnung
- ▶ Setzt Zufallsstichproben voraus (Stichprobenergebnisse als Zufallsvariablen mit *bekannter* Verteilung)
- ▶ Zwei Hauptgebiete
 1. Konfidenzintervalle
 2. Hypothesentests

Wahrscheinlichkeitsrechnung

- ▶ Zufallsvariablen, Zufallsexperiment, Wahrscheinlichkeiten als relative Häufigkeiten von Elementarereignissen
- ▶ $0 \leq P \leq 1$
- ▶ Verteilungen \rightarrow Wahrscheinlichkeitsverteilungen
- ▶ Totale Wahrscheinlichkeit: Summe der Wahrscheinlichkeiten aller möglichen Elementarereignisse = 1
- ▶ Vereinigungsmenge von Ereignissen ($A \cup B$): A oder B oder A und B
- ▶ Schnittmenge von Ereignissen ($A \cap B$): A und B
- ▶ Konditionale Wahrscheinlichkeit: Wie wahrscheinlich ist Ereignis A , wenn B bereits eingetreten ist: $P(A|B)$

Zentraler Grenzwertsatz

- ▶ Grundgesamtheit vs. Stichprobe
- ▶ Einfache Zufallsvariable mit Verteilung in GG und ähnlicher Verteilung in Stichproben
- ▶ Komplexe Zufallsvariablen: additives Zusammenwirken von einfachen Zufallsvariablen (z. B. durch Mittelwertbildung)
- ▶ Stichprobenkennwertverteilung: Verteilung der komplexen Zufallsvariablen über sehr viele Stichproben mit identischem Umfang hinweg
- ▶ Zentraler Grenzwertsatz: Für $n > 30$ Normalverteilung der Stichprobenmittelwerte, unabhängig davon wie Ausgangsvariablen verteilt sind
- ▶ Normalverteilung mit
 - ▶ Mittelwert = Mittelwert Ausgangsvariable in GG
 - ▶ Standardabweichung = „Standardfehler“ (des Mittelwertes ...) = Funktion von Streuung in GG und Wurzel aus n

Konfidenzintervalle

- ▶ Wie stark kann Stichprobenmittelwert von wahren Mittelwert in GG abweichen?
- ▶ Probabilistische Aussage
- ▶ Baut auf ZGWS auf
- ▶ Für alle Normalverteilungen 95% der Fläche/Wahrscheinlichkeit ± 1.96 Standardabweichungen vom Mittelwert entfernt (99% $\rightarrow \pm 2.58$)
- ▶ 95% aller Stichprobenmittelwerte nicht weiter als ± 1.96 Standardfehler vom wahren Mittelwert entfernt
- ▶ Umkehrschluß: Intervall ± 1.96 Standardfehler um Stichprobenmittelwert wird in 95% aller Anwendungen wahren Mittelwert in GG mit einschließen

Logik Hypothesentests

- ▶ Gilt in GG „Alternativhypothese“ (H_A)?
 - ▶ Gerichtet vs. ungerichtet
 - ▶ Spezifisch (fast nie) vs. unspezifisch
- ▶ „Nullhypothese“ (kein Effekt/Zusammenhang/Unterschied)
 - ▶ (Idealerweise) etablierter Forschungsstand
 - ▶ Konservativ (vs. Induktion)
 - ▶ Soll nur aufgegeben, wenn Risiko gering ist
 - ▶ „Irrtumswahrscheinlichkeit“, Fehler 1. Ordnung, α -Fehler \rightarrow $< 1/5/10\%$
- ▶ Statistisch „signifikantes“ Ergebnis
- ▶ Stichproben-/Testergebnis als Zufallsvariable, bekannte Verteilung (z. B. wg. Grenzwertsatz)
- ▶ Hier behandelte Tests: z -, χ^2 -, t -Test(s)

z-Test

- ▶ Wie wahrscheinlich ist es, daß Stichprobe mit Mittelwert \bar{x} aus bekannter GG mit Mittelwert μ stammt?
- ▶ Nullhypothese: Stichprobe aus bekannter GG
- ▶ Alternativhypothese (gerichtet/ungerichtet): Stichprobe aus neuer GG
- ▶ μ bekannt, Standardfehler berechnen
- ▶ Differenz $\bar{x} - \mu$ durch Standardfehler teilen (Standardisierung)
- ▶ Vergleich mit Flächenanteilen aus tabellierter Standardnormalverteilung (z-Verteilung) → Entscheidung über Signifikanz

t -Test

- ▶ Vergleich zweier Stichprobenmittelwerte – Differenz zwischen Mittelwerten μ_1 und μ_2 in den Grundgesamtheiten real?
- ▶ (Nullhypothese: kein Unterschied)
- ▶ Alternativhypothese: Differenz, ggf. mit Richtung der Abweichung
- ▶ Abhängige vs. unabhängige Stichproben
- ▶ Annahme: Varianzen in beiden Grundgesamtheiten identisch (sonst Korrektur) → Schätzung aus den Stichproben
- ▶ Differenz zwischen zwei Stichprobenmittelwerten (über sehr viele Paare von Stichproben hinweg) t -verteilt
- ▶ Zahl der Freiheitsgrade beachten

χ^2 -Test

- ▶ Kreuztabellen für zwei nominalskalierte Variablen
- ▶ Weicht Verteilung in GG von Indifferenztabelle ab (Unabhängigkeit)?
- ▶ χ^2 wie gewohnt berechnen
- ▶ Unter relativ allgemeinen Bedingungen folgen empirische χ^2 -Werte über viele Stichprobenziehungen hinweg (aus einer GG, in der Nullhypothese gilt) theoretischer χ^2 -Verteilung
- ▶ Freiheitsgrade beachten
- ▶ Vergleich empirischer Wert mit theoretischer Verteilung → Flächenanteil → Wahrscheinlichkeit unter Nullhypothese
- ▶ $\leq \alpha$? → „signifikant“

Zusammenfassung

- ▶ Vorbereitung auf Klausur
 - ▶ Folien
 - ▶ Mitschriften
 - ▶ **Aufgaben rechnen** (aus Agresti, Gehring/Weins oder Schumann)

Zusammenfassung

- ▶ Vorbereitung auf Klausur
 - ▶ Folien
 - ▶ Mitschriften
 - ▶ **Aufgaben rechnen** (aus Agresti, Gehring/Weins oder Schumann)
- ▶ Mit einfachen Grundkenntnissen erschließt sich relativ viel

Zusammenfassung

- ▶ Vorbereitung auf Klausur
 - ▶ Folien
 - ▶ Mitschriften
 - ▶ **Aufgaben rechnen** (aus Agresti, Gehring/Weins oder Schumann)
- ▶ Mit einfachen Grundkenntnissen erschließt sich relativ viel
- ▶ Statistik II (BA Kernfach)
 - ▶ Mehr Inferenz
 - ▶ Mehr Modelle
 - ▶ Mehr Anwendungen