**Handout Korrelation & Skalentransformation**

Variablen Definition:

* Eindeutige Zuordnung von Zahlen (=Realisationen) zu Merkmalen
* Extensionale Definition= zählt alle Realisationen der Variable (kann jede beliebige Zahl sein)
* Intensionale Definition = gibt Vorschrift an, die Variable eindeutig spezifiziert
* Diskret = endlich & feste Variable

-> dichotom: 2 diskrete Werte

-> polytom: mehr als 2 diskrete werte

* Stetig = kann unendlich viele beliebige Werte annehmen (reelle Zahlen)

Skala

* Festlegung einer Einheit, in der das Merkmal gemessen wird
* Qualitativ: Nominalskala & Ordinalskala
* Quantitativ: Intervall-, Verhältnis-& Absolutskala

**Skalentransformation und Z-Standardisierung**

* Umwandlung einer Skala in eine andere (bzw. die Werte)
* **Ziel:** Merkmalsverteilungen mit unterschiedlichen Mittelwerten und Streuungen vergleichbar machen Umwandlung von einer Skala in eine Andere
* Beurteilung der Werte bezüglich ihrer relativen Lage in der Verteilung

Schritt 1: z-Standardisierung jedes Datenpunktes **𝑧 =𝑥 − μ𝑥 \* σ𝑥**

Schritt 2: (lineare) Transformation jedes Datenpunktes in die neue Skala **xneu = z ⋅ σneu + μneu**

**Eigenschaften**

* Für normalverteilte z-Werte gilt: **µ = 0 σ= 1**
* (Bzw. Festlegung eines neuen Mittelwerten & Standardabweichung)
* Der Wert z gibt an, wie viele Standardabweichungen und in welche Richtung ein Messwert xi vom Mittelwert entfernt ist
* Durch die z-Transformation wird die Form der Verteilung nicht beeinflusst!

**Korrelationen**

Kovarianz

* Bivariate Intervalldaten
* Positiver oder negativer Zusammenhang zwischen 2 Datenreihen
* Positiv wenn gleichsinniger Zusammenhang/ negativ wenn gegensinnig
* Erfüllt nicht Forderung der Invarianz
* Multiplikation verändert auf äquivariante Weise -> Kovarianz verändert sich mit -> Äquivarianz keine gute Eigenschaft

Produkt-Moment Korrelation

* X & Y z-standardisieren -> befreit von Äquivarianz

=korrel(Spalte1; Spalte2)

* Korrelationskoeffizient Eigenschaften:
* Ab Intervallskala
* Zwischen -1 & 1
* r= 0 -> kein Zusammenhang
* neg. r = gegensinniger Zusammenhang / pos. R = gleichsinnig
* Ausreißer abhängig
* Lineare Transformationen keine Auswirkung

Faustregeln Cohen 1988



* Vorsicht bei Interpretation -> hohe Korrelation nur wegen Ausreißer? -> Scatterplot betrachten
* In experimentellen Studien erst r=.75 hoch
* Zufallskorrelation wegen zu kleiner Stichproben
* Voraussetzung für Kausalität: 1.Korrelation ungleich 0 2.Ursache vor Wirkung 3. Andere Erklärung für Kovariarion ausgeschlossen 4. Raum-zeitlich indifferent

KORRELATION IST NICHT GLEICH KAUSALITÄT

Grafische Beschreibung

* Scatterplot:
* Zusammenhang von Messwertpaar in Punktwolke abgebildet
* Einfach interpretierbar

Punktbiseriale Korrelation

=(MWx1 – MWx2)/SDgesamt\*Wurzel(n1\*n2)/(n1+n2)

* X = dichotom & nominalskaliert / Y = intervallskaliert
* wie 2 intervallskalierte Variablen betrachten
* Dichotomisieren von Variablen führt zu beliebigen Ergebnissen obwohl eigentlicher Zusammenhang anders ist

X‘->X <-> Y

* gleiche Eigenschaften wie PMK

Biseriale Korrelation

* Korrektur der kriteriumsabhängigen Veränderung (dichotomisieren)
* gleiche Eigenschaften wie PMK
* Normalverteilungsannahme der stetigen Variable
* rpbis vorzuziehen, da keine Normalverteilungsannahme

Tetrachorische Korrelation

=cos(PI()/(1+ Wurzel(n11\*n22/(n12\*n21))))

* 2 künstlich dichotomisierte Variablen
* 2x2 Kontingenztabelle
* Überschätzt Korrelation, wenn Randverteilung stark asymmetrisch oder nxy < 5
* Selten in Praxis genutzt

Rangkorrelation nach Spearman

=rang.mittelw (x-Wert; alle Daten;1)

* Bivariante Ordinaldaten
* Abstände nicht interpretierbar -> Rangordnung nutzen
* Rangbildung
* Ties bilden bei mehreren gleichen Werten von X
* PMK der Ränge berechnen
* Spearmans r: Wertebereich von -1 & 1; Robust bezüglich Ausreißern, invariant bei streng monotonen Transformationen

Phi Koeffizient

* Bivariante Nominaldaten
* Unabhängigkeit in Kontingenztabellen: Variable X sagt nichts über Y aus; Randhäufigkeiten bleiben gleich
* Abhängigkeit in Kontingenztabellen: Variable X sagt etwas über Y aus; Verbundhäufigkeiten betrachten
* Berechnung: 1. Weg: 1) Variablen numerisch beschreiben 2) Datentabellen erstellen 3) PMK berechnen

 2. Weg: 1) Phi-Koeffizient-Formel anhand 2x2 Kontingenztabelle

* Gleiches Maß wie r
* Positives Phi: Kombination auf Hauptdiagonale hoch
* Negatives Phi: Kombination auf Nebendiagonale hoch
* Selbe Eigenschaften wie PMK
* Nur interpretierbar in Bezug auf Kontingenztabelle & Vorzeichen
* Wegen schiefen Randhäufigkeiten Phi = -1 & +1 nicht erreichbar
* Phi Max + & - berechnen
* Hauptdiagonale / Nebendiagonale auf 0 setzen
*  Phi an maximal mögliche Korrelation normieren = Phi norm
* Interpretation:
* Phi & Phi norm sehr unterschiedlich -> Phi max sehr klein
* 2 Gruppen in Daten: Für Mehrheit stimmt schwacher Zusammenhag

Chi2 Koeffizient

* Beobachtete Kontingenztabelle mit erwarteter (fiktiver) vergleichen
* Indifferenztabelle berechnen

= Verbundhäufigkeiten unter Unabhängigkeit

* Chi2 = 0 bei perfekter Unabhängigkeit

=Summenprodukt((beobachtete Werte- erwartete Werte)^2/ erwartete Werte

* Zellhäufigkeit > 5
* beliebig große Werte
* Normieren um zu interpretieren

Cramers V

=Wurzel(Chi2/n\*(min(k,m)-1)))

k= Anzahl der Zeilen

m= Anzahl der Spalten

* Als Korrelationskoeffizient interpretierbar
* V = 0 bei perfekter Unabhängigkeit
* Zwischen 0 & 1
* Wo genau der Zusammenhang liegt kann nicht gesagt werden