Einführung in die Statistik
LMU Sommersemester 2014
Dozenten: Ewerdwalbesloh & Schlagbauer

Janosch Maier

2. Juli 2014

Inhaltsverzeichnis

1 Messen & Skalenniveaus

- Erleben & Versuche versuchen zu Beschreiben, Erklären & Vorherzusagen
- Empirische Wissenschaft: Theorie → Hypothese; Konfrontation von Hypothese mit Realität

1.1 Messen

- Zuordnen von Zahlen zu Objekten nach bestimmten Regeln
- Messung hat Einfluss auf statistische Auswertbarkeit

1.1.1 Empirische Forschung

- Fragestellung (Allgemein Formuliert)
- Hypothese
- Operationalisierung (UV + AV)
- Datenerhebung
- Datenauswertung
- Interpretation

1.1.2 Variablen

- Interindividuell: Zwischen Versuchspersonen
- Intraindividuell: Innerhalb einer Versuchsperson
- Stetig: Beliebig viele Ausprägungen (z.B. Gewicht, Größe)
- Diskret: Nur Abstufungen (z.B. Note, Anzahl Geschwister)

1.1.3 Hypothesen

- Kausal: Wenn – Dann
- Falsifizierbarkeit (Popper: Kritischer Rationalismus)
- Spezifisch (Nicht nur Einzelfall / Erlaubt Vorhersagen)

1.2 Skalenniveaus

- Zuordnungsregel: Zahl repräsentiert Objekt (empirisch) korrekt
- Festlegung abhängig von: Objekteigenschaften, Abbildungsart durch Messinstrument

1.2.1 Nominalskala

- Zuordnung genau einer Ziffer pro Merkmalsausprägung
- Exklusivität, Exhaustivität
- Kein Mittelwert möglich (da keine Reihenfolge)

1.2.2 Ordinalskala

- Reihenfolge der Merkmalsausprägungen
- Exklusivität, Exhaustivität, Totale Ordnung
- Unterschiede in Abstand nicht möglich

1.2.3 Intervallskala

- Gleich Große Abstände zwischen Merkmalsausprägungen
- Exklusivität, Exhaustivität, Totale Ordnung, Äquidistanz
- Kein Natürlicher Nullpunkt
- Erlaubt nur noch Lineare Transformationen

1.2.4 Verhältnisskala

- Anfangspunkt ist natürlicher Nullpunkt
- Exklusivität, Exhaustivität, Totale Ordnung, Äquidistanz, Natürlicher Nullpunkt
- Nur Relationale Tranformationen erlaubt

2 Deskriptive Statistik & Graphise Darstellung

2.1 Maße der zentralen Tendenz

2.1.1 Modalwert

- Am häufigsten Vorkommender Wert
- Stabil gegenüber Extremwerten
- Auf allen Skalenniveaus möglich

2.1.2 Median

- Hälfte der Werte über / unter Median
- Stabil gegenüber Extremwerten
- Erfordert Ordinalskala
- Ungerade: Mittlerer Wert, Gerade: Mittelwert der beiden mittleren Zahlen

2.1.3 Arithmetisches Mittel (Mittelwert)

- Empfindlich gegenüber Extremwerten
- Erfordert Intervallskala
- $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

2.1.4 Gewogenes Arithmetisches Mittel (GAM)

- Mittelwert von Mittelwerten
- $\overline{\overline{x}} = \frac{\sum_{i=1}^{p}(n_i \cdot \overline{x_i})}{\sum_{i=1}^{p} n_i}$

2.2 Maße der Streuung (Dispersionsmaße)

2.2.1 Variationsbreite (Range)

- Maximum - Minimum
- Sehr empflindlich gegenüber Extremwerten
- Mindestens Ordinalskala
- Geringer Informationsgehalt

2.2.2 Varianz

- Streuung um Mittelwert
- Mindestens Intervallskala
- $\overset{\wedge}{\sigma}{}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
- Teilen durch $n - 1$ statt n, da nur Schätzung. In der Gesamtpopulation ist Varianz i.A. etwas größer.

2.2.3 Population & Stichproben

- Population (= Grundgesamtheit): Griechische Zeichen
- Stichproben (= Teilmenge der Population): Lateinische Zeichen

2.2.4 Standardabweichung

- $\hat{\sigma}_x = \sqrt{\hat{\sigma}_x^2}$

2.2.5 Quartile

- Q_1: Prozentrang von 25. 25% liegen unterhalb des unteren Quartils.
- Q_2: Prozentrang von 50. Median
- Q_3: Prozentrang von 75. 75% liegen unterhalb des oberen Quartils.
- Interquartilsabstand $IQA = Q_3 - Q_1$

2.3 Verteilungen

- Schiefe: Linkssteil = Rechtsschief vs. Rechtssteil = Linksschief
- Breite (Exzess): Leptokurtisch (Schmalgipflich) vs. Platykurtisch (Breitgipflig)
- Symmetrisch vs. Asymmetrisch
- Unimodal (Ein Maximum) vs. Bimodal (Mehrere Maxima)

2.4 Darstellung von Daten

2.4.1 SPSS-Output

- Absolute Häufigkeit f
- Relative Häufigkeit $f_{rel} = \frac{f}{n}$
- Prozente + Gültige Prozente
- Kummulierte Prozente

2.4.2 Grafische Darstellungen

vgl. Zusammenfassung Empirische Forschungsmethoden II

- Histogramm
- Balken- / Kreisdiagramm
- Box-Plot ($1,5Q_3 <$ Ausreiser $< 3Q_3 <$ Extremwerte)
- Scatter-Plot (= Streudiagramm)
- Fehlerbalken (i.A. Mittelwetr $\pm$ Standardabweichung)

3 Standardisierung von Daten

- Absolutkennwerte können nicht verglichen werden, da unterschiedliche Verteilungen
- Lösung: Verteilung normieren (z-Normierung nur möglich bei Normalverteilung)

3.1 Z-Transformation

$$z_i = \frac{x_i - \bar{x}}{s_x} \tag{1}$$

3.2 Normalverteilung

- Normalverteilungsannahme: Biologische Werte sind i.A. normalverteilt
- Symmetrisch, Asymptotische Annäherung an X-Achse
- Modus = Median = Mittelwert
- Histogramm mit Balken diskret
- Kontinuierliche Verteilung kann nur Aussagen über Abschnitte geben, da "einzelne Wahrscheinlichkeiten" unendlich klein sind.
- Fläche unter der Kurve ist Wahrscheinlichkeit. Gesamtfläche = 1
-
- Bsp: IQ-Wert = $100 + 15 \cdot z_x$

3.3 Standardnormalverteilung

- Entsteht durch Z-Transformation
- Mittelwert 0, Standardabweichung 1 ⇒ Vergleichbar

3.4 Repräsentativität / Genauigkeit

3.4.1 Inferenzstatistik

Schluss von einer Teilmenge auf Gesamtheit ⇒ Stichprobe soll Population wiederspiegeln

3.4.2 Repräsentativität

- Definition über Auswahl der Stichprobe: Zufallsauswahl
- Einfache Zufallsstichprobe: Globale Repräsentativität
- Geschichtete Zufallsstichprobe: Aufteilung in Gruppen anhand von Zufallsstichproben ⇒ Repräsentativität hinsichtlich des Schichtungsmerkmals

- Klumpenstichprobe: Zufallsauswahl in Klumpen $\Rightarrow$ Eingeschränkte Repräsentativität
- Ad-hoc-Stichprobe: Auswahl nach Verfügbarkeit $\Rightarrow$ Nicht repräsentativität

3.4.3 Stichprobenkennwertverteilung

- Wie gut ist Schätzung?
- Unendlich viele Zufallsstichproben der Größe n mit Kennwert k, dann ergibt sich eine Häufigkeitsverteilung der Kennwerte
- Eigenschaften der Verteilung kennzeichnen Güte der Schätzung
- Streuung Maß, wie gut Stichprobenkennwert Populationswert schätzt

3.4.4 Standardfehler des Mittelwerts

$$\sigma_m = \sqrt{\frac{\sigma^2}{n}} \tag{2}$$

$$\overset{\wedge}{\sigma}_m = \sqrt{\frac{s^2}{n-1}} \tag{3}$$

- Je größer Populationstreuung, desto höher der Fehler
- Je größer die Stichprobe, desto kleiner der Fehler
- Wahrscheinlichkeit von 68,26% liegt wahrer Populationswert zwischen $\pm$ einem Standardfehler

3.4.5 Konfidenzintervall

- Intervall in dem z.B. 90% der Stichprobenmittelwerte liegen
- Gesamtheit folgt einer Normalverteilung mit Populationsmittelwert μ und Streuung σ

$$MI = m \pm z_{x\%} \cdot \overset{\wedge}{\sigma}_m \tag{4}$$

- $95,0\%$: z = 1,96
- $99,0\%$: z = 2,58
- $99,5\%$: z = 2,81

4 Hypothesentesten und t-Test

4.1 Hypothesentesten

- Eigenschaften einer Population als Hypothese postuliert
- Überprüfung ob Eigenschaften durch Stichprobenergebnisse bestätigt

4.1.1 Hypothesearten

- Unterschiedshypothesen (Mittelwertunterschiede, ...)
- Zusammenhangshypothesen (Korrelationen)
- Ungerichtete Hypothesen vs. Gerichtete Hypothesen
- Unspezifische Hypothesen vs. Spezifische Hypothesen (Größe des Zusammenhangs)

4.1.2 Nullhypothese & Alternativhypothese

- H_0: Kein Unterschied zwischen Bedingungen (Kein Zusammenhang zwischen Variablen). Mittelwertunterschiede nur Standardfehler
- H_1: Unterschied (Zusammenhang). Mittelwertunterschiede systematisch

	H_1	H_0
ungerichtet - unspezifisch	$\mu_A \neq \mu_B$	$\mu_A = \mu_B$
gerichtet - unspezifisch	$\mu_A > \mu_B$	$\mu_A \leq \mu_B$
gerichtet - spezifisch	$\mu_A > \mu_B + x$	$\mu_A \leq \mu_B$ + x

4.1.3 Vorgehen

- Übersetzen von inhaltlicher in statistische Hypothese
- Berechnung des empirischen Kennwertes
- Bestimmung der Kennwertverteilung
- Statistische Entscheidung für H_0/H_1
- Übersetzen in inhaltliche Bedeutung

4.1.4 Fehlerarten

	H_0 gilt	H_1 gilt
Entscheidung für H_0		β-Fehler
Entscheidung für H_1	α-Fehler	

- α-Niveau vorher festgelegt: $p(H_1|H_0)$

4.2 t-Test für unabhängige Stichproben

- Entscheidungsregel, ob sich Mittelwerte systematisch unterscheiden
- Bei intervallskalierten Daten
- Parametrisches Verfahren (Setzt Verteilung voraus und Signifikanzprüfung auf Grund dieser Verteilung)
- Wie wahrscheinlich ist empirisch gefunden Differenz, wenn H_0 gilt? $\Rightarrow$ Stichprobenwerteverteilung von Mittelwertsdifferenzen

4.2.1 t-Verteilung

$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \tag{5}$$

$\mu_1 - \mu_2 = 0$, wenn H_0 gilt

4.2.2 Freiheitsgrade

- Wie viele Werte dürfen variiert werden, damit es genau zu einem bestimmten Ergebnis kommt
- Bei t-Test: $df = n_1 + n_2 - 2$

4.2.3 Signifikanzniveau

- t-Wert klein $\Rightarrow$ H_0 wird abgelehnt.
- Grenzwert ist α-Fehler-Niveau / Signifikanzniveau

4.2.4 Signifikanzprüfung

- Bestimmung der Wahrscheinlichkeit des empirischen t-Wertes und Vergleich mit festgelegtem Signifikanziveau
- Vergleich von empirischem und kritischem t-Wert (Ablesen aus Tabelle in Abhängigkeit von Freiheitsgraden)

4.2.5 Voraussetzungen

- Intervallskalaniveau
- Normalverteilt
- Varianzhomogenität
- t-Test robust, wenn Gruppen annähernd gleich groß und > 30 VPN

4.3 t-Test für abhängige Stichproben

- $d_i = x_{i1} - x_{i2}$
- Stichprobenkennwert ist Mittelwert der Differenzen $\bar{x}_d$
- Streuung: $\overset{\wedge}{\sigma}_{\bar{x}_d} = \frac{\overset{\wedge}{\sigma}_d}{\sqrt{n}}$, $\overset{\wedge}{\sigma}_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{x}_d)^2}{n-1}}$
- $t_{abhngig} = \frac{\bar{x}_d - \mu_d}{\overset{\wedge}{\sigma}_{\bar{x}_d}}$
- $df = n - 1$

5 Effektgröße & Teststärke

5.1 Effektgröße

- Inhaltliche Bewertung eines Effekts
- Standardisierung zum verschiedene Untersuchungen vergleichen zu können $\Rightarrow$ Distanz von Mittelwerten
- Effektstärkenmaß d zur Standardisierung (Gleiche Streuung der Stichproben angenommen: Varianzhomogenität)
- Bester Schätzer: Gepoolte Streuung $d = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_x} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}^2_{x_1} + \hat{\sigma}^2_{x_2}}{2}}}$
- Beurteilung abhängig von inhaltlichen Überlegungen. Erste Orientierung: 0,2 klein, 0,5 mittel, 0,8 groß

5.2 Teststärke

- Nicht-signifikantes Ergebnis. β-Fehler möglich, also H_1 könnte trotzdem gelten
- Teststärke ist wahrscheinlichkeit H_1 anzonehmen, wenn sie in Wirklichkeit gilt: $1 - \beta$
- β-Fehler & Teststärke abhängig von: Signifikanzniveau, Stichprobengröße, angenommenem Effekt
- Bei kleiner Effektstärke hat gutes Signifikanzniveau einen großen β-Fehler zu Folge
- Je größer die Stichproben, desto schmaler Stichprobenkennwertverteilungen (weniger Überschneidung der Kurven) $\Rightarrow$ Kleinerer Standardfehler, jede Differenz kann Signifikant werden

5.2.1 Determinanten des t-Tests

Jede Determinante kann aus den anderen 3 berechnet werden.

- Signifikanzniveau
- Stichprobengröße
- Angenommener Effekt
- β-Fehler

5.2.2 Stichprobenumfangsplanung

- Stichprobengröße nicht zu klein (kleine α & β-Fehler)
- Stichprobengröße nicht zu groß (Signifikanz nur bei bedeutsamem Effekt)
- Bestimmung der Stichprobengröße auf Grund der anderen 3 Determinanten.

6 Nich-parametrische Tests

- Nominaldaten / Ordinaldaten
- Keine Normalverteilung (z.B. zu kleine Stichprobe)
- Keine Varianzhomogenität
- Grundprinzip:
  - Nur ordinale Information der Daten genutzt: Künstliche Äquidistanz mit Abstand 1 $\Rightarrow$ Mittelwertsberechnung
  - Rohdaten Rangplätzen zuweisen, Jede Person von Gruppe 1 mit jeder Person von Gruppe 2 vergleichen, Ermittlung von Rangplatzüber/-unterschreitungen, Berechnen der Prüfgröße U, Bestimmung der statistischen Signifikanz

6.1 Mann-Whitney U-Test (Unabhängige Stichproben)

- Rangplätze vergeben. Kleinster Wert entspricht kleinster Rang
- Rangplatzüberschreitungen: Erste Person aus Gruppe 1 hat Rang x. Wieviele Personen aus Gruppe 2 haben einen höheren Rangplatz? Für alle Personen aus Gruppe 1 wiederholen $U = n_1 \cdot n_2 + \frac{n_1 \cdot n_1 + 1}{2} - T_1$ (T entspsricht Rangsumme)
- Rangplatzunterschreitungen: $U' = n_1 \cdot n_2 + \frac{n_2 \cdot n_2 + 1}{2} - T_2$
- $U = n_1 \times n_2 - U'$
- Große Stichproben ($n_1, n_2 > 20$) $\Rightarrow$ Kennwerteverteilung nähert sich Normalverteilung an $\Rightarrow$ Signifikanztest mit Hilfe der z-Verteilung; U, U' symmetrisch zum Mittelwert
  - $U = U'$
  - $\mu_U = \frac{n_1 \cdot n_2}{2}$
  - $\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$
  - $z = \frac{U - \mu_U}{\sigma_U}$
  - Entscheidungsregel: Zweiseitig: $|Z_{emp}| > Z_{krit_{(1-\frac{\alpha}{2})}}$, Einseitig: $|Z_{emp}| > Z_{krit_{(1-\alpha)}}$
- Kleine Stichproben: Nicht normalverteilt, Vergleich des kleineren Wertes mit kritischem U-Wert.
  - U_{krit} ist maximaler U-Wert, der noch zu signifikantem Ergebnis führt (Anders, als sonst) $\Rightarrow$ Lehne H_0 ab, wenn $U_{emp} \leq U_{krit}$.
- Bei gleichen Messwerten: Verbundene Ränge. Bsp: $\frac{5+6}{2} = 5.5$
- Korrektur der Streuung: $\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}} \cdot \sqrt{\frac{N^3 - N}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12}}$

6.2 Wilcoxon-Test (Abhängige Stichproben)

- Differenzen bilden, Ränge vergeben & Rangnummern berechnen
  - Rangreihenfolge anhand des Betrags
  - Paare mit Nulldifferenz nicht berücksichtigt (N verringern)
  - Rangsumme für positive & negative Differenzen berechnen (Aufsummieren), W ist Rangsumme mit kleinerem Betrag
- Prüfgröße für N berechnen – Nachschlagen in Tabelle
- Entscheidnugsregel anwenden $W_{emp} < W_{krit} \Rightarrow$ Ablehnen von H_0.

7 Kovariation und Korrelation

- Zusammenhang von Variablen untersuchen
- Gibt es einen / Wie groß ist der Zusammenhang zwischen ...
- Zusammenhang: Variablen variieren systematisch miteinander (Varianzen hängen zusammen)
- Zusammenhang $\neq$ Kausalität (Kausale Interpretation nur bei: Zeitlicher Ordnung, Theoretisch plausibler Annahme, Systematische Variation einer Variable)

7.1 Streudiagramm (Scatterplot)

- Graphische Darstellung des Zusammenhangs
- X-Achse $\rightarrow$ Prädiktor, Y-Achse $\rightarrow$ Kriterium
- Art des Zusammenhangs: Linear, Quadratisch, Kein Zusammenhang

7.2 Kovarianz

- Beobachtungsabhängig
- Maß für gemeinsame Variation zweier Variablen
- $cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$
- Positive Kovarianz: positiver Zusammenhang, Negativ: Negativer Zusammenhang, 0: Kein (linearer) Zusammenhang
- Wertebereich: Maximum ist Produkt beider Variablen-Streuungen $\Rightarrow$ Abhängig von Maßeinheit, Streuung der Variablen

7.3 Korrelation

- Beobachtungsunabhängig
- Produkt-Moment-Korrelation / Pearson Korrelation $\Rightarrow$ Standardisierte Kovarianz mit Wertebereich $[-1, 1]$.
- $r_{xy} = \frac{cov_{emp}}{cov_{max}} = \frac{cov(x,y)}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$
- $+1$ ist perfekt positiver Zusammenhang, -1 perfekt negativ, 0 kein linearer Zusammenhang
- Skalenniveau des Korrelationskoeffizienten: Ordinalskalenniveau
- Interpretierbarkeit der Korrelationsstärke abhängig von Forschungsgebiet & Situation (Labor vs. Feldexperiment)
- Signifikanz und Effekstärke für die Korrelation
- Cohens Konventionen: .10 klein, .30 mittel, .50 groß

7.3.1 Signifikanztest von Korrelationen

- "Ist die Korrelation von Null verschieden?"
- Mit welcher Wahrscheinlichkeit stammt empirische Korrelation r aus einer Population mit einer Korrelation $\rho = 0$.
- $\Rightarrow$ t-Test
- $t_{df} = \frac{r \cdot \sqrt{N-2}}{\sqrt{1-r^2}}$

7.3.2 Effektstärke

- Determinationskoeffizient r^2
- Vorzeichen geht dur Quadrierung erklären
- Intervallskaliert, Prozentmaß
- Wertebereich zwischen 0 (Kein Effekt) und 1 (100% Effektaufklärung)

7.3.3 Fisschers Z-Transformation

- Bilden des Mittelwerts von Korrelationen kann nicht direkt ermittelt werden. Deshalb umrechnen in z-Werte und berechnen des Mittelwertes. Dann zurückrechnen in Korrelation
- $Z = \frac{1}{2} \cdot \ln(\frac{1+r}{1-r})$
- $\bar{r} = \frac{e^{2 \cdot \bar{z}} - 1}{e^{2 \cdot \bar{z}} + 1}$
- Eignung für Produkt-Moment Korrelation, Rangkorrelation und punktbiseriale Korrelation

7.3.4 Rangkorrelation

- Erfasst, wie Rangreihen systematisch variieren
- Ordinaldatenniveau / nicht normalverteilte Daten
- $r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{N \cdot (N^2 - 1)}$
- d ist Differenz der Rangplätze zwischen Variable x und y
- Wertebreich: -1, +1
- Prüfung auf Signifikanz über t-Test wenn $n > 30$

7.3.5 Punktbasierte Korrelation

- Zusammenhang zwischen Intervallskalierten und dichotom nominalskalierten Variable
- $r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{\hat{\sigma}_y} \cdot \sqrt{\frac{n_0 \cdot n_1}{N^2}}$
- Wertebereich: -1 bis +1
- Positive Korrelation, wenn y-Werte in x_0 im Durchschnitt kleiner sind, als die y-Werte in x_1
- Konzeptuelle Entsprechung mit t-Test (Zusammenhang vs. Unterschied)

7.3.6 Partialkorrelation

- Dritte Variable verantwortlich für Zusammenhang? (Scheinkorrelation)
- $r_{xy|z} = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{(1 - r_{yz}^2) \cdot (1 - r_{xz}^2)}}$
- $t_{df} = r_{xy|z} \cdot \sqrt{\frac{N-2}{1 - r_{xy|z}^2}}$

7.3.7 Suppressorvariable

- Drittvariable Z (Suppressor) unterdrückt Zusammenhang
- Z ist mit X unkorreliert, aber mit Y korreliert sie hoch ⇒ Unterdrückt Varianz von Y, die für Zusammenhang von X und Y nicht relevant ist.

8 Regressionsanalyse

- Regredieren (Zurückführen) von einer unbekannten Variable auf eine Bekannte $\Rightarrow$ Vorhersagen der unbekannten Variablen
- Bei einer Korrelation: Darstellen einer Linie, die den besten Schätzer darstellt, um von x auf y zu schließen
- Regression liefert Schätzwerte mit gewisser Unschärfe
- Regressionsgerade: Gerade, zu der die Punkte innerhalb der Punktewolke maximal abweichen (Besser, bei stärkerer Korrelation)
- x ist unabhängige Variable: Prädiktor
- y ist abhängige Variable: Kriterium
- Einfache lineare Regression: Nur ein Prädiktor, Linearer Zusammenhang
- Regressionsgleichung: $\hat{y} = b \cdot x + a$
- b: Regressionsgewicht ist "Vorhersagekraft" des Prediktors: Um wie viel ändert sich das Kriterium, wenn sich Prädiktor um eine Einheit verändert.
- a: Höhenlage
- $\hat{y}$ ist mit bestimmtem Fehler behaftet. Vorhersagefehler $y_i - \hat{y}$ ist Residuum
- Kriterium der kleinsten Quadrate: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min$
- $b_{yx} = \frac{cov(x,y)}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}$
- $a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$
- Kovarianz von 0: $b = 0 \Rightarrow \hat{y}_i = \bar{y}$
- Standardisierung: $\beta_{yx} = b \cdot \frac{\sigma_x}{\sigma_y}$ ist Standardisiertes Regressionsgewicht. Gibt an, um wieviele Standardabweichungen sich Kriterium verändert, wenn Prädiktor sich um eine Standardabweichungen verändert.

8.1 Einfache, Lineare Regression

- $\beta_{yx} = cor(x, y)$: Standardisierte Regressionsgewicht entspricht Produkt-Moment-Korrelation
- Einfache lineare Regression macht Annahmen über Kausalität (unabhängig, ob diese Zutreffen)
- x, y z-Standardisiert: $x_{zy} = 0 \Rightarrow$ Regressionsgerade verläuft durch Ursrpung. Steigung entspricht Korrelation

8.2 Vorhersagekraft des Prädiktors

8.2.1 Prüfung der Regressionsgewichte

- Signifikanztest, ob Regressionsgewicht bedeutsam ist.
- $t = \frac{b}{s_b}$: b unstandardisiertes Regressionsgewicht, s_b: Standardfehler des Regressionsgewicht
- Signifikanz für b berechnet, gilt auch für β

8.2.2 Residualvarianz

- Abweichung zwischen Empirischen & vorhergesagten Werten $\Rightarrow$ Residualvarianz
- Zusammenhang zwischen Empirischen & vorhergesagten Werten $\Rightarrow$ Kovarianz
- Gesamtvarianz: Wie weit weicht jeder empirische y-Wert vom Mittelwert ab
- Regressionsvarianz (Aufgeklärte Varianz): Wie weit weicht der geschätzte y-Wert vom Mittelwert ab?
- Residualvarianz: Abweichung des Schatzers vom tatsächlichen Messwert (Vorhersagefehler)
- $\overset{\wedge}{\sigma}{}^2_y = \overset{\wedge}{\sigma}{}^2_{\hat{y}} + \overset{\wedge}{\sigma}{}^2_{[\frac{x}{y}]}$ Gesamtvarianz = Regressionsvarianz + Residualvarianz
- Inhaltlich: Warum variiert ein Merkmal. Prädiktor erklärt mit Hilfe der Regressionsgeraden so viel Varianz wie möglich.
- Residuum kann theoretisch durch andere Prädiktoren erklärt werden.
- Extremfälle
  - $r = 1 \rightarrow \overset{\wedge}{\sigma}{}^2_y = \overset{\wedge}{\sigma}{}^2_{\hat{y}}$
  - $r = 0 \rightarrow \overset{\wedge}{\sigma}{}^2_y = \overset{\wedge}{\sigma}{}^2_{[\frac{x}{y}]}$

8.2.3 Standardschätzfehler

- Wurzel der Residualvarianz
- $\sigma_{\frac{x}{y}]} = \sqrt{\overset{\wedge}{\sigma}{}^2_{[\frac{x}{y}]}}$

8.2.4 Determinationskoeffizient r^2

- Anteil der Regressionsvarianz an Gesamtvarianz
- $r^2 = \frac{\overset{\wedge}{\sigma}{}^2_{\hat{y}}}{\overset{\wedge}{\sigma}{}^2_y}$
- Einfache lineare Regression: $r^2 = \beta^2$
- Daumenregel: $r^2 > 0,02$ klein; > 0.13 moderat; > 0.26 stark

8.3 Multiple lineare Regression

- Multivariates Verfahren: Mehrere abhängige/unabhängige Variablen
- Vorhersage des Kriteriums durch mehrere Prädiktoren (Inhaltlich Interessant, Störvariablen)
- $\hat{y} = a + b_1x_1 + b_2x_2 + ... + b_kx_k$
- Multiple Regression berücksichtigt Interkorrelationen zwischen Prädiktoren (Auspartialisieren, von geteilter Varianz)

8.3.1 Multiple Regressionsgewichte

- $b_{yx_1} = r_{yx_1|x_2} \cdot \frac{\sqrt{s_y^2 \cdot (1 - r_{x_2y}^2)}}{\sqrt{s_{x_1}^2 \cdot (1 - r_{x_2x_1}^2)}}$
- Eigenständiger Beitrag eines Prädiktors zur Kriteriumsvorhersage
- KEINE Entsprechung der Korrelation (wie bei Einfacher Linearer Regression)
- Geringe Korrelation zwischen Prädiktoren $\Rightarrow$ Hohe individuelle Varianzaufklärung
- Suppressorvariablen möglich (negatives Regressionsgewicht)
- Signifikanztest für einzelne Regressionsgewichte möglich
- Verständnis als Regressionsgewicht der Residuen möglich: x_1 klärt einen Teil der Varianz von y auf. Nicht aufgeklärt wird das Residuum. Varianzaufklärung durch Residuum von x_2 am Residuum von y ist multiples Regressionsgewicht.

8.3.2 Semi-Partialkorrelation

- Anteil der Variabilität in X oder Y, der auf Z zurück geht wird entfernt
- $r_y(x, z) = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{(1 - r_{xz}^2)}}$
- vgl. Partialkorrelation: Anteil der Variablität in X & Y, der auf Z zurück geht wird entfernt

8.3.3 Determinationskoeffizient

- Multiples R^2
- Anteil der Varianz, die alle Prädiktoren gemeinsam aufklären
- Mit 100 Multipliziert: Prozent der aufgeklärten Varianz
- Summe der Semipartialdeterminationen (quadrierte Semipartialkorrelationen)
- Semipartialdeterminationen abhängig von der Reihenfolge des Einschlusses. R^2 unabhängig davon immer gleich.

8.3.4 Parsimonität

- Welches Regressionsmodell ist das beste? (Bedeutsame Prädiktoren, Beste Vorhersagekraft)
- Methoden: Einschluss (Alles gleichzeitig betrachten), Hierarchisch (Änderung der Prädiktoren iterativ)

⇒ Parimonitätsprinzip: Bestes & sparsamstes Modell finden

8.3.5 Voraussetzungen

- Linearer Zusammenhang zwischen allen Prädiktoren & Kriterium
- Normalverteilung der Fehler
- Unabhängigkeit der y-Werte
- Multivariate Normalverteilung
- Homoskedastizität (Streuung des Kriterums unabhängig vom Prädiktor)
- Keine Multikollinearität (Diese tritt auf, wenn Prädiktoren stark miteinander korrelieren, dann verzerrte Regressionsgewichte)

9 Varianzanalyse (ANOVA – Analysis of Variance)

- Unterschied im Gruppen-Mittelwert bei mehr als zwei Gruppen
- “Gibt es einen Unterschied zwischen ... und ... und ...?
- Immer unspezifisch / zweiseitig
- H_0: Kein Unterschied zwischen den Gruppen
- H_1: Ein Unterschied zwischen mindestens zwei Gruppen
- Problem multipler t-Tests: α-Fehlerkummulierung ($\alpha_{gesamt} = 1 - (1 - \alpha_{Test})^m$), Verringerte Teststärke (Testen von Teilgruppen)

9.1 Einfaktorielle ANOVA ohne Messwiederholung

- Wie viel Stichprobenvarianz ist auf unabhängige Variable zurückzuführen? (Und wie viel ist ”Fehler“)
- Varianzzerlegung: Gesamtvarianz = Systematische Varianz + Residualvarianz $\sigma_{gesamt} = \sigma^2_{sys} + \sigma^2_{res}$ / $\sigma_{gesamt} = \sigma^2_\alpha + \sigma^2_\epsilon$
- Systematisch Varianz: zwischen den Gruppenmittelwerten
- Residualvarianz: Innerhalb der Gruppen
- Ist $\sigma^2_{sys} >> \sigma^2_{res}$? $\Rightarrow$ Signifikanztest

9.1.1 Begriffe

- Faktor: Unabhängige Variable, die in Gruppen aufteilt
- Faktorstufen: Anzahl der Bedingungen eines Faktors (Treatmentfaktoren: Randomisierung, Kassifikationsfaktoren: Probandeneigenschaften)
- Haupteffekt bei Underschied von mindestens zwei Stufen eines Faktors

9.1.2 Zerlegung der Gesamtvarianz

- Quadratsumme: $QS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- Freiheitsgrade $df_x = n - 1$
- $\overset{\wedge}{\sigma}^2 = \frac{QS_x}{df_x}$
- $QS_{ges} = QS_{sys} + QS_{res}$, $df_{ges} = df_{sys} + df_{res}$

9.1.3 Gesamtvarianz

$$\overset{\wedge}{\sigma}^2_{ges} = \frac{QS_{ges}}{df_{ges}} = \frac{\sum_{i=1}^{p}\sum_{m=1}^{n}(x_{mi} - \bar{G})^2}{N-1} \tag{6}$$

$\bar{G}$ Gesamtmittelwert

m VPN-Nummer in Gruppen

i Gruppennummer

p Anzahl der Gruppen

n Anzahl VPN pro Gruppe

N Gesamtgröße der Stichprobe

9.1.4 Residualvarianz

$$\overset{\wedge}{\sigma}^2_{res} = \frac{QS_{res}}{df_{res}} = \frac{\sum_{i=1}^{p}\overset{\wedge}{\sigma}^2_i}{p} \tag{7}$$

Mit der Varianz innerhalb einer Gruppe: $\overset{\wedge}{\sigma}^2_i = \frac{\sum_{i=1}^{n}(x_{mi} - \bar{A}_i)^2}{n-1}$

9.1.5 Systematisch Varianz

$$\overset{\wedge}{\sigma}^2_{sys} = \frac{QS_{sys}}{df_{sys}} = \frac{n\sum_{i=1}^{p}(\bar{A}_i - \bar{G})^2}{p-1} \tag{8}$$

9.1.6 Ungleiche Gruppengrößen

- Jeweils n_i statt n
- Bei der Residualvarianz muss das n_1 in die Summe gezogen werden

9.1.7 Signifikanzprüfung

$$F_{(df_{sys};df_{res})} = \frac{\overset{\wedge}{\sigma}^2_{sys}}{\overset{\wedge}{\sigma}^2_{res}} \tag{9}$$

- F-Wert 1, wenn systematische Varianz 0
- F-Wert $>$ 1, wenn systematischer Einfluss des Faktors
- Vergleich mit kritischem F-Wert / Vergleich der Wahrscheinlichkeit des F-Werts mit α-Niveau
- $F_{krit} < F_{emp} \Rightarrow$ Signifikant, H_0 wird verworfen

9.1.8 Effekstärke

- Omega-Quadrat schätzt Effekt in Population (Prozent aufgeklärter Varianz)
  - Nur bei Einfaktorieller Varianzanalyse bei unabhängigen Gruppen mit gleicher Größe
  - $\omega^2 = \frac{QS_{sys} - df_{sys} \cdot \hat{\sigma}^2_{res}}{QS_{ges} + \hat{\sigma}^2_{res}}$
- Eta-Quadrat beschreibt Effekt in der Stichprobe (Prozent aufgeklärter Varianz, nicht vergleichbar)
  - $\eta^2 = \frac{QS_{sys}}{QS_{ges}}$
  - .01 = kleiner, .06 = mittlerer, .14 = starker Effekt
- Partielles-Eta-Quadrat bei einfaktorieller ANOVA identisch mit η^2
  - $\eta^2_{partiell} = \frac{QS_{sys}(Faktor)}{QS_{sys}(Faktor) + QS_{res}}$
- η^2 überschätzt Populationseffekt, deshalb ω^2 bevorzugen

9.1.9 Post-Hoc-Tests

- Tukey HSD (Honest Significant Difference)
  - Differenz zweier Gruppen, das kumuliertes α-Niveau, festgelegtes nicht überschreitet
  - Unterschied > HSD $\Rightarrow$ Grupenunterschied
  - α-Fehlerkumulierung: Einzelne α_i so gewählt, das nach Kumulierung α nicht überschritten
  - Teststärke: Mindestens Haupteffekt der Varianzanalyse, da Streuung der gesamten Stichprobe zu Grunde liegt
  - $HSD = q_{krit(\alpha,r,df_{res})} \sqrt{\frac{\hat{\sigma}^2_{res}}{n}}$
- Bonferroni-Korrektur
  - Herabsetzen des α-Niveaus für einzelne t-Tests
  - Anzahl Einzelvergleiche $m = \binom{p}{2}$
  - Adjustiertes Alpha $\alpha_{adj.} = \frac{\alpha}{m}$
  - Problem: α-Niveau sing exponentiell.

9.1.10 Voraussetzungen für die Varianzanalyse

- Intervalldatenniveau der abhängigen Variable
- Normalverteilung der abhängigen Variable in Population (Robust, Probleme bei kleinen Stichproben)
- Varianzhomogenität: Gleiche Varianzen der Gruppen in Population (Robust, Probleme bei kleinen Stichproben)
- Unabhängigkeit der Messwerte (Sonst Varianzanalyse mit Messwiederholung)

9.2 Zweifaktorielle Varianzanalyse

- Zwei unabhängige Variablen; Jeder Proband gehört zwei Gruppenarten an
- Anzahl der Versuchsbedingungen entpsricht Produkt der Anzahl der Faktorstufen (Bsp. 3 $\times$ 2 faktoriellen Varianzanalyse)
- Idealerweise gleich viele Probanden pro Bedingung
- Tabelle mit Mittelwerten aller Faktorstufenkombination
- Zerlegung in systematische- & Residualvarianz, Zerlegung der Systematischen Varianz in Varianz zu A, B und Interaktion
- $\overset{\wedge}{\sigma}^2_{sys} = \overset{\wedge}{\sigma}^2_{\alpha} + \overset{\wedge}{\sigma}^2_{\beta} + \overset{\wedge}{\sigma}^2_{\alpha \cdot \beta}$
- $QS_{ges} = QS_A + QS_B + XS_{A \times B} + XS_{res}$
- $df_{ges} = df_A + df_B + df_{A \times B} + df_{res}$
- $\overset{\wedge}{\sigma}^2_A = \frac{QS_A}{df_A} = \frac{\sum_{i=1}^{p} nq(\bar{A}_i - \bar{G})^2}{p-1}$, mit $p = \#$Faktorstufen von A, $q = \#$Faktorstufen von B, $n =$ VPN einer Zelle ($A \times B$-Kombination)

9.2.1 Haupteffekte

- Haupteffekte A, B + Interaktion/Wechselwirkung; Jeweils Signifikanzprüfung
- Untersuchung der Haupteffekte unabhängig von der jeweils anderen Variable analog zu Einfaktoriellen Varianzanalyse

9.2.2 Interaktionseffekt

- Schätzer für die Varianz: Vorhersage auf Grund der Haupteffekte.
- Schätzer für Varianz der Wechselwirkung ist Abweichung der Gruppenmittelwerte von den vorhergesagten Werten auf Grund der Haupteffekte
- Erwartete Varianz auf Grund der Haupteffekte: $\bar{AB}_{y(erwartet)} = \bar{A}_i + \bar{B}_i - \bar{G}$
- Abweichung ist Schätzer für Interaktioneffekt:
  $\overset{\wedge}{\sigma}^2_{A \times B} = \frac{QS_{A \times B}}{df_{A \times B}} = \frac{\sum_{j=1}^{q} \sum_{i=1}^{p} n(\bar{AB}_{ij} - \bar{AB}_{ij(erwartet)})^2}{(p-1)(q-1)}$
- Schätzer für Residualvarianz: *TODO*
- Interaktionsdiagramm: Vgl. Empirische Forschungsmethoden I
  - Keine Wechselwirkung: Parallele Linien
  - Ordinale Wechselwirkung, Richtung der Haupteffekte bleibt gleich, Beide Haupteffekte eindeutig interpretierbar
  - Semiordinale/hybride Wechselwirkung, Richtung eines Haupteffekts bleibt gleich, Nur ein Haupteffekt interpretierbar
  - Disordinale Wechselwirkung, Keine Richtung des Haupteffekts bleibt gleich, Kein Haupteffekt interpretierbar

9.2.3 Effektstärke

- Anteil der Aufgeklärten Varianz durch Faktor (Anteil der Gesamtvarianz): η^2 – Effekstärken abhängig, Effekstärken addieren sich zur gesamten aufgekläen Varianz
- Varianz durch Faktor im Verhältnis zur Residualvarianz: $\eta^2_{partiell}$ – Effektstärken unabhängig, Effektstärken addieren sich nicht zur gesamten aufgekläen Varianz, (Wird in SPSS angegeben)

9.2.4 Voraussetzungen

- Gleichen Voraussetzungen wie Einfaktorielle Varianzanalyse ohne Messwiederholung
- Intervallskalierung der abhängigen Variablen
- Normalverteilung des Merkmasl (Robust)
- Varianzhomogenität in allen Bedingungskombinationen (Robust bei großen Stichproben)
- Unabhängigkeit der Messwerte (Keine Beeinflussung von Personen untereinander)