Differenz von Anteilen bei abhängigen Stichproben

Voraussetzungen:

- X und Y sind zwei Bernoulli-Größen mit

$$p_X = P(X = 1)$$
$$p_Y = P(Y = 1)$$

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ **abhängige, verbundene** Stichproben
- Absolute Häufigkeiten werden in einer Kontingenztafel festgehalten

	Y=0	Y=1
X=0	n_{11}	n_{12}
X=1	n_{21}	n_{22}

Der Test von McNemar

Voraussetzungen:

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ **abhängige, verbundene** binäre Zufallsvariable
- Absolute Häufigkeiten werden in einer Kontingenztafel festgehalten

	Y=0	Y=1
X=0	n_{11}	n_{12}
X=1	n_{21}	n_{22}

Nullhypothese H_0 : $P(X = 1) = P(Y = 1)$
Die beiden Zellen $n_1 1$ und $n_2 2$ liefern keine Information zum Unterschied der beiden Wahrscheinlichkeiten P (X=1) und P(Y=1). Betrachte die beiden Zellen der Wechsler: n_{12} und n_{21}
Unter der Nullhypothese sollte $n_{12} = n_{21}$ gelten.
Daher ist unter H_0:

$$n_{12} \sim B(n_{12} + n_{21}, 0.5)$$

Man führt also einen Binomialtest mit $\pi = 0.5$ und $n = n_{12} + n_{21}$ durch.

Nichtparametrische Tests

- Bis auf den χ^2-Unabhängigkeits-Test bauen alle Tests auf der (zumindestens approximativen Gültigkeit der) Normalverteilungsannahme auf.
- Problematisch, z.B.
  - bei kleinen Stichprobenumfängen (z.B. in der Vorbereitung von strukturierten Beobachtungen, bei nicht reaktiven Verfahren oder in der Psychologie und Medizin)
  - oder bei ordinalen Daten mit wenigen unterschiedlichen Ausprägungen.
- Hier kann die unreflektierte Anwendung der Standardtests zu Fehlergebnissen führen.
- Ein wichtiger Ausweg: nichtparametrische Tests = Verteilungsfreie Verfahren
- Hier wird die Information in den Beobachtungen auf Ränge, bzw. größer/kleiner Vergleiche reduziert.
- Bekannteste Beispiele: Wilcoxon-Test, Vorzeichentest.

Der Wilcoxon Test für unabhängige Stichproben

Voraussetzungen:

- X und Y sind zwei Größen mit Medianen med_X und med_Y
- $X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ **unabhängige** Stichproben
- $H_0 : \text{med}_X = med_Y$ vs. $H_1 : \text{med}_X \neq \text{med}_Y$

Grundidee: Betrachte die Ränge aus allen Beobachtungen X_i und Y_j und bezeichne diese mit $\text{rg}(X_i)$ und $\text{rg}(Y_j)$, z.B.
$X_1 = 3,\ X_2 = 5,\ Y_1 = 6,\ Y_2 = 1,\ Y_3 = 4 \ \Rightarrow$
$\text{rg}(X_1) = 2,\ \text{rg}(X_2) = 4,\ \text{rg}(Y_1) = 5,\ \text{rg}(Y_2) = 1\ ,\ \text{rg}(Y_3) = 3$

Testgröße:

$$T = \sum_{i=1}^{m} \text{rg}(X_i)$$

Die exakte Verteilung von T kann berechnet werden. Für hinreichend große n und m kann sie durch eine NV approximiert werden. Ablehnung von H_0 falls große und kleine T.

Multiple Testprobleme

- Gegeben sei ein rein zufälliger Datensatz mit 50 Variablen ohne irgendeinen Zusammenhang.
- Testen aller Variablenpaare auf einen Zusammenhang (1225 Tests)

  Bei vorgegebener Irrtumswahrscheinlichkeit von 5% gilt für die Anzahl fälschlich verworfener Nullhypothesen $X \sim B(1225, 0.05)$ und somit $E(X) = 61.25$.

  Im Durchschnitt wird also mehr als 61 mal die Nullhypothese, dass kein Zusammenhang besteht, verworfen.

$\Rightarrow$ wenige, sinnvolle Hypothesen *vorher inhaltlich* überlegen. In den Daten entdeckte „Zusammenhänge" als statistisch signifikant nachzuweisen, ist (fast) zirkulär.

- Es gibt Ansätze, wie man bei großen Hypothesensystemen diesem Problem entkommt.

$\Rightarrow$ Theorie des multiplen Testens

Z.B. Adjustierung des Irrtumswahrscheinlichkeit: Statt α betrachte man α/(Anzahl der Tests). Diese spezielle Korrektur ist aber meist überkonservativ und kann durch bessere Korrekturen ersetzt werden.

Bestimmung des Stichprobenumfangs

- Bestimmung mit Obergrenze für Fehler 2. Art (β-Fehler)
- β = P(Nullhypothese wird nicht abgelehnt | H_1 wahr)
- $1 - \beta$ wird auch als Power bezeichnet
- Üblicher Wert: Power = 0.8 bzw. $\beta = 0.2$

Fehler 2. Art im Binomialtest

Interessante Größen:

- H_0: $\pi = 0.5$ versus $H_1 : \pi > 0.5$
- β: $P(H_0$ wird beibehalten $| H_1)$
- π_w: wahre Erfolgswahrscheinlichkeit

Numerische Werte für n=10 und c=7
$\pi_w = 0.6$ und $\beta = 0.95$
$\pi_w = 0.7$ und $\beta = 0.85$
$\pi_w = 0.8$ und $\beta = 0.62$
Numerische Werte für n=20 und c=13
$\pi_w = 0.6$ und $\beta = 0.87$
$\pi_w = 0.7$ und $\beta = 0.58$
$\pi_w = 0.8$ und $\beta = 0.02$
Numerische Werte für n=100 und c=57
$\pi_w = 0.6$ und $\beta = 0.37$
$\pi_w = 0.7$ und $\beta = 0.007$
$\pi_w = 0.8$ und $\beta < 0.0001$

Fehler 2. Art im Beispiel des Binomialtests

Fehler 2. Art sinkt hier mit:

- Wachsender wahrer Erfolgswahrscheinlichkeit ($\pi > 0.5$)
- Wachsendem Stichprobenumfang

Stichprobenumfangsbestimmung

- **Festlegen von:**
  α, β, Effektgröße π_w
- **Beispiel:**
  $\pi_w = 0.7$
  $\Rightarrow n = 20 \quad \Rightarrow \beta = 0.58$ zu klein
  $\Rightarrow n = 100 \quad \Rightarrow \beta = 0.07$ zu groß
- **Wähle kleinstes n mit:**
  $\beta \leq 0.2$
- **Hier:**
  $n = 37$, $\beta = 0.19$
  $n = 36$, $\beta = 0.26$

Bestimmung des Stichprobenumfangs beim t-Test

Vergleich zweier Gruppen

- Zielgröße: Y
- Relevanter Unterschied: d
- Annahme zur Streuung: σ
- Signifikanzniveau: $\alpha = 0.05$
- Fehler 2.Art: $\beta = 0.2$

Bestimmung des Stichprobenumfangs beim t-Test

Einige Werte für den Stichprobenumfang im Zwei-Stichproben t-Test

- unter $\alpha = 0.05, \beta = 0.2$
- zweiseitige Fragestellung

Beachte: Es kommt nur auf das Verhältnis d/σ (Effektstärke) an.

d/σ	$n_1 = n_2$
2	5
1.33	10
1	17
0.66	37
0.58	50
0.29	100

Zusammenfassung Tests

Statistischer Test

Ein statistischer Test ist eine Untersuchung, ob man eine Hypothese über die Grundgesamtheit mit Hilfe einer Stichprobe widerlegen kann

Wichtige Tests

- einseitiger Test $\longleftrightarrow$ zweiseitiger Test
- Tests auf Anteile, Erwartungswerte und deren Differenzen
  - eine Stichprobe
  - zwei Stichproben $\rightarrow$ verbunden oder unverbunden
- Test auf Unabhängigkeit: χ^2-Unabhängigkeitstest

Interpretation eines Tests

- Statistische Signifikanz bedeutet: Das Ergebnis ist nicht durch Zufall erklärbar.
- Statistische Signifikanz bedeutet nicht unbedingt, dass der Unterschied relevant ist.
  $\rightarrow$ Daher immer die Größe des Effekts (Schätzung) angeben!
- Ein statistischer Test liefert aber allein noch keinen kausalen Zusammenhang.
- Bei nicht signifikantem Ergebnis immer Konfidenzintervalle der entsprechenden Parameter angeben. Sie geben eine Idee, wie groß der nicht nachgewiesene Effekt sein könnte.