Statistische Testverfahren in der Schule und in der Praxis

Christoph Richard, FAU Erlangen-Nürnberg
Tag der Mathematik 9. März 2024
Version 12. März 2024

Zusammenfassung

Mit der Rückkehr zu G9 hat sich der Lehrplan für die Oberstufe geändert. Insbesondere kann im Vertiefungsbereich in der 12. Jahrgangsstufe Statistik unterrichtet werden. Wir besprechen das Konzept des statistischen Testens anhand des Binomialtests. Wir geben optimale einseitige und zweiseitige Tests an und besprechen Anwendungsbeispiele. Schließlich gehen wir kurz auf Themen des Vertiefungsbereichs ein.

Fachlehrplan des ISB ab Schuljahr 2024/2025 (Auszug)

- M12 Zufallsgrößen und Binomialverteilung (22 Std)

  Zufallsvariablen auf endlichen Ergebnisräumen, Verteilungsfunktion, Erwartungswert und Varianz, Urnenmodelle, Bernoulli-Kette, Binomialverteilung

- M12 einseitiger Signifikanztest (8 Std)

  Prinzip des Testens, einseitiger Binomialtest, Fehler 1.Art und Fehler 2.Art, Visualisierung

- M13 Normalverteilung (14 Std)

  kontinuierliche Zufallsgrößen, Dichtefunktion, Intervallwahrscheinlichkeiten, Normalverteilung, σ-Regeln

- M12 Vertiefung: Modul 5 Statistik [KWBBH]

  beschreibende Statistik: lineare Regression, schließende Statistik: t-Test oder χ^2-Unabhängigkeitstest, p-Wert, Statistik-Software

Statistik lehren

das didaktische Spannungsfeld

- Schließende Statistik ist konzeptionell und technisch anspruchsvoll.
- Anwender:innen sollten viel statistische Theorie lernen.
- Mathematiker:innen sollten viel "praktische Statistik" lernen.

Lehramts-Ausbildung

- nicht überall Spezialkurs "Stochastik für Lehramt Gymnasium"
- in Bayern derzeit keine Staatsexamensprüfung in Stochastik
- Grundvorlesung Stochastik enthält oft wenig Statistik
- Lehrämtler:innen hören meist keine weiterführende Vorlesung zu angewandter Statistik oder mathematischer Statistik.

Themen dieses Kurses

Konzepte und Methoden des statistischen Testens (5+12+12 Folien)

- Vorhersagekraft statistischer Tests
- einseitige Binomialtests
- zweiseitige Binomialtests

Testen jenseits des Münzwurfs (3 Folien)

- z-Tests, t-Tests
- χ^2-Anpassungstests, χ^2-Unabhängigkeitstest
- exakter Test von Fisher auf Unabhängigkeit

Vorhersagekraft eines medizinischen Tests [H 15.11]

M11: bedingte Wkeit, stochastische Unabhängigkeit

- 4-elementiger Produktraum: Eine getestete Person ist entweder krank oder gesund, ihr Test ist entweder positiv oder negativ.
- guter Test: hohe Sensitivität $\mathbb{P}(P|K)$, hohe Spezifizität $\mathbb{P}(N|G)$
  Schätze diese Werte anhand einer medizinischen Studie.
- Schätze die Wkeit $\mathbb{P}(K)$, dass Person krank ist, mit der Häufigkeit der Krankheit in der Bevölkerung (Prävalenz).
- $\mathbb{P}(K)$ wird als a priori-Wkeit interpretiert, $\mathbb{P}(K|P)$ als a posteriori-Wkeit, dass positiv getestete Person krank ist.
- Chance (odds), dass positiv getestete Person krank ist:

$$\frac{\mathbb{P}(K|P)}{1-\mathbb{P}(K|P)} = \frac{\mathbb{P}(K\cap P)}{\mathbb{P}(G\cap P)} = \frac{\mathbb{P}(P\cap K)}{\mathbb{P}(P\cap G)} = \frac{\mathbb{P}(P|K)}{1-\mathbb{P}(N|G)}\cdot\frac{\mathbb{P}(K)}{1-\mathbb{P}(K)}$$

- Bei kleiner Prävalenz nur geringe positive Vorhersagekraft $\mathbb{P}(K|P)$!
- Effekt kann mit absoluten Häufigkeiten visualisiert werden [RKI]

Beispiel: Corona-Tests

Schnelltests auf SARS-CoV-2 [K]

- Sensitivität $\mathbb{P}(P|K) = 70\%$, Spezifizität $\mathbb{P}(N|G) = 95\%$
- Prävalenz $\mathbb{P}(K) = 80\%$ (Isolier-Abteilung):
  positive Vorhersagekraft $\mathbb{P}(K|P) = 98\%$
- Prävalenz $\mathbb{P}(K) = 3\%$ (Hausarztpraxis):
  positive Vorhersagekraft $\mathbb{P}(K|P) = 30\%$ (!)

"Wer viel testet, findet auch viel." [S]

- Der Landkreis Greiz geriet im Mai 2020 und im April 2021 wegen hoher Inzidenzen in die Schlagzeilen.
- In 2020 wurden in Altenheimen PCR-Tests durchgeführt.
- Anstieg der Inzidenz in 2021 nicht in erster Linie auf neue kostenlose Schnelltests für Symptomlose zurückzuführen

klassische statistische Tests [H 29]

Niveau-α-Binomialtest für H_0 gegen H_1

- zwei mögliche Bereiche für unbekannte Wkeit p einer Münze
- $H_0, H_1 \subset [0,1]$ mit $H_0 \cap H_1 = \varnothing$
- n-facher unabhängiger Münzwurf entscheidet für oder gegen H_0
- Fehler 1.Art: H_0 wird abgelehnt, obwohl H_0 wahr ist.
- Fehler 2.Art: H_0 wird akzeptiert, obwohl H_0 falsch ist.
- Forderung: Wkeit $\alpha(p)$ für Fehler 1.Art höchstens $\alpha \in (0,1)$
- Dann liegt Wkeit $\beta(p)$ für Fehler 2.Art fest, kann sehr groß sein.

Wähle also für H_0 das *Gegenteil* der zu überprüfenden Hypothese!

- Falls man H_0 ablehnen kann, kontrolliert man den Fehler für eine falsche Entscheidung.
- Falls H_0 akzeptiert werden muss, gilt das Prinzip *in dubio pro reo*, obwohl eine falsche Entscheidung sehr wahrscheinlich sein könnte.

Wie gut ist ein statistischer Test?

Binomialtest für H_0 gegen H_1

- Beschreibung der Fehlerwahrscheinlichkeiten mit Gütefunktion

$$p \mapsto G(p) = \mathbb{P}_p(H_0 \text{ ablehnen}) \qquad (p \in H_0 \cup H_1)$$

- $G(p) = \alpha(p)$ auf H_0, $G(p) = 1 - \beta(p)$ auf H_1
- Gütefunktion also idealerweise 01-Sprungfunktion
- Gütefunktion aber meist stetig in p. Falls $p \in H_1$ kleinen Abstand zu H_0 hat, kann $\beta(p)$ dann nahe am Wert $1 - \alpha$ liegen!

Test-Design:

- Mit wachsender Stichprobengröße approximiert die Gütefunktion meist die 01-Sprungfunktion.
- Stelle durch Justieren der Stichprobengröße die gewünschte Genauigkeit des Tests ein.

Vorhersagekraft eines statistischen Tests

statistischer Test $H_0 : p = p_0$ gegen $H_1 : p = p_1$

- Modellierung: H_0 und H_1 treten mit bestimmten Wkeiten auf.
- bestimme positive Vorhersagekraft $\mathbb{P}(H_1 \text{ wahr} \,|\, H_0 \text{ ablehnen})$
- Analogie zum medizinischen Test: $H_0 \mathrel{\hat{=}} G$ gegen $H_1 \mathrel{\hat{=}} K$
- Mit Spezifizität $1 - \alpha$ und Sensitivität $1 - \beta$ gilt

$$\frac{\mathbb{P}(H_1 \text{ wahr} \,|\, H_0 \text{ ablehnen})}{1 - \mathbb{P}(H_1 \text{ wahr} \,|\, H_0 \text{ ablehnen})} = \frac{1-\beta}{\alpha} \cdot \frac{\mathbb{P}(H_1)}{1 - \mathbb{P}(H_1)} \,.$$

Interpretation

- Für $\beta \approx 1 - \alpha$ ist der Test fast wirkungslos. Dies ist der Fall für p_0 nahe p_1 (relativ zur Stichprobengröße).
- Falls H_1 sehr unwahrscheinlich ist, lehnt man H_0 meist fälschlich ab, die positive Vorhersagekraft des Tests ist also klein.
- Diese Situation ist mutmaßlich typisch für statistische Tests [I,N]!

rechtsseitiger Niveau-α-Binomialtest

$H_0 : p \leq p_0$ gegen $H_1 : p > p_0$

- akzeptiere H_0, falls Anzahl der Erfolge S_n klein
- verwerfe H_0, falls S_n größer als kritischer Wert c_α
- Wahrscheinlichkeit für Fehler 1.Art höchstens α, also auf H_0

$$G(p) = \mathbb{P}_p(H_0 \text{ verwerfen}) = \mathbb{P}_p(S_n > c_\alpha) \leq \mathbb{P}_{p_0}(S_n > c_\alpha) \leq \alpha$$

- Für hohe Güte muss Verwerfungsbereich groß sein. Wähle also $c_\alpha \in \{0, \ldots, n\}$ minimal, so dass obige Ungleichung erfüllt ist.

Bemerkung

- c_α ist das kleinste α-Fraktil der Binomialverteilung B_{n,p_0}.
- $G(p) = \beta_{c_\alpha+1, n-c_\alpha}(p)$ stetig in p, mit Verteilungsfunktion $t \mapsto \beta_{r,s}(t)$ der β-Verteilung mit Parametern r, s [G Lem 8.8]

Testentscheidung mit p-Wert

- Nenne $p(k) = \mathbb{P}_{p_0}(S_n \geq k)$ den p-Wert der Beobachtung k.
- Falls $p(k)$ klein, ist die Beobachtung k unwahrscheinlich für H_0.
- Für die Testentscheidung gilt

$$H_0 \text{ ablehnen} \iff p(k) \leq \alpha$$

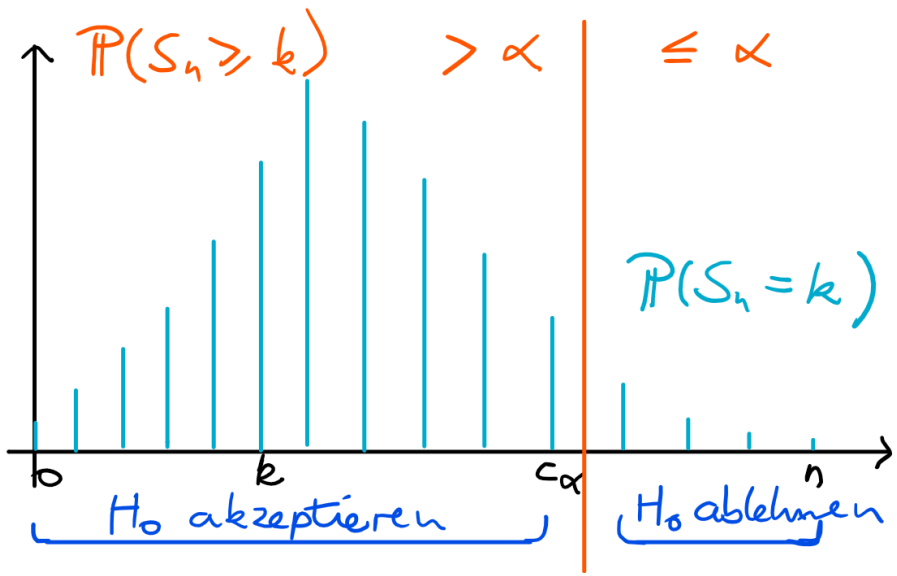
- elegante Entscheidungsmethode, da p-Wert unabhängig von α

p-Werte jenseits des Münzwurfs

- Statistik-Programme geben für eine Reihe von Standard-Tests p-Werte aus.
- Bei zweiseitigen Tests mit p-Werten ist oft nicht der optimale Test implementiert (su).

Testentscheidung mit p-Wert

- Äquivalenz der Entscheidungsregeln:



- kritischer Wert $c_\alpha = \min\{k : \mathbb{P}_{p_0}(S_n > k) \leq \alpha\}$
- Wegen $c_\alpha = \min\{k : p(k) \leq \alpha\} - 1$ gilt

$$H_0 \text{ ablehnen} \iff k \geq c_\alpha + 1 \iff p(k) \leq \alpha$$

Beispiel: Geburtenhäufigkeiten

Werden in Erlangen signifikant mehr Jungen als Mädchen geboren?

- statistisches Modell: Geschlecht wird durch Münzwurf entschieden

  Jeder statistische Test basiert auf einer (oft vereinfachenden) Modellannahme, welche manchmal biologisch oder physikalisch begründet werden kann. Die Passgenauigkeit des Modells sollte stets diskutiert werden!

- Binomialtest $H_0 : p \leq 1/2$ gegen $H_1 : p > 1/2$ zum Niveau $\alpha = 0.05$
- Erlangen 2022: 1059 Geburten, davon 536 Jungen

  (aus Datenbank Genesis des Bayerischen Landesamts für Statistik)

- Testergebnis: H_0 muss akzeptiert werden. ($\beta(0.51) = 0.84$)

  Der Datensatz ist zu klein, um eine eventuelle Abweichung von 1/2 nach oben statistisch absichern zu können. Zufällige Schwankungen überdecken eine kleine Abweichung von 1/2 nach oben oder nach unten.

Diskussion: Geburtenhäufigkeiten

In Erlangen werden aber *jedes Jahr* mehr Jungen geboren!

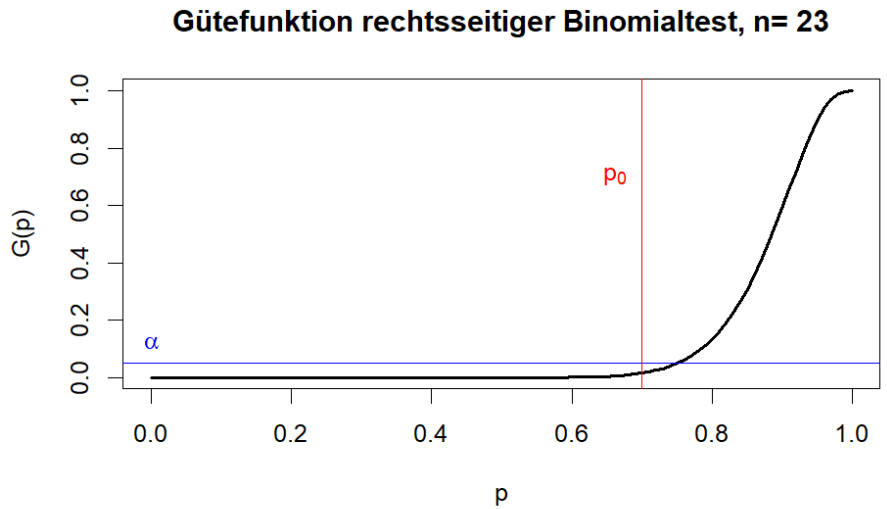
- H_0 kann mit den Geburtenzahlen ab 1972 für Erlangen zum Niveau 0.05 abgelehnt werden. Die Wkeit p ist mit 95%-iger Sicherheit größer als 0.508 (einseitiges 95%-Konfidenzintervall).
- Mit den Geburtenzahlen ab 1972 für Erlangen kann die Modellannahme eines Münzwurfs zum Niveau 0.05 nicht abgelehnt werden (Shapiro-Wilk-Test auf Normalverteilung, zulässig aufgrund großer Stichprobe von ca 1000 Kindern pro Jahr).

Wie könnte es weitergehen?

- Analyse der Geburtenzahlen in anderen Städten und Ländern
- Diskussion mit Biologen über Gründe für die Ungleichheit [H]

rechtsseitiger Binomialtest: Gütefunktion

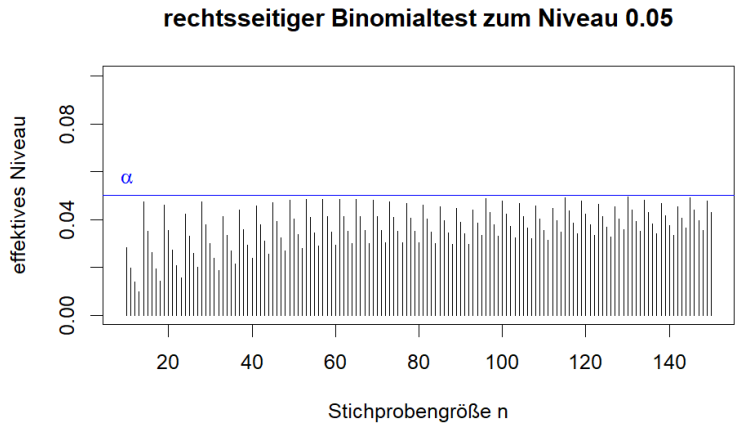
Binomialtest $H_0 : p \leq 0.7$ gegen $H_1 : p > 0.7$ zum Niveau $\alpha = 0.05$



- Der Test schöpft das Niveau α nicht aus.
- Mit zunehmender Stichprobengröße wird die Gütefunktion einer 01-Sprungfunktion bei p_0 immer ähnlicher.
- Man kann also eine geeignete Stichprobengröße wählen, um die Wahrscheinlichkeit für einen Fehler 2.Art zu kontrollieren.

rechtsseitiger Binomialtest: effektives Niveau

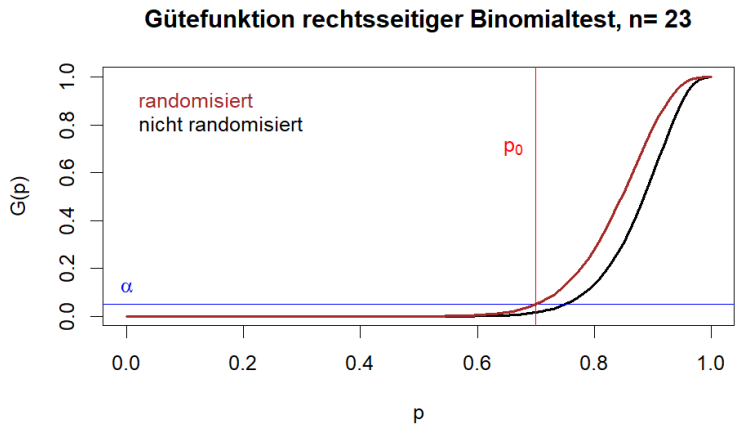
Binomialtest $H_0 : p \leq 0.7$ gegen $H_1 : p > 0.7$ zum Niveau $\alpha = 0.05$



- Das effektive Niveau des Tests ist kleiner als α.
- Damit ist der Test konservativer als gefordert.
- Damit ist die Güte geringer als möglich.
- Dieser Effekt tritt nur bei diskreten Zufallsgrößen auf!

rechtsseitiger Binomialtest: Randomisierung [G 10]

Binomialtest $H_0 : p \leq 0.7$ gegen $H_1 : p > 0.7$ zum Niveau $\alpha = 0.05$



- Lehne auch bei c_α per Münzwurf mit geeigneter Wkeit γ ab:

$$\mathbb{P}_{p_0}(H_0 \text{ ablehnen}) = \mathbb{P}_{p_0}(S_n > c_\alpha) + \gamma \cdot \mathbb{P}_{p_0}(S_n = c_\alpha) \stackrel{!}{=} \alpha$$

- Der randomisierte Test hat höhere Güte, gleichmäßig in p!
- natürliches Verfahren in der Qualitätskontrolle, wird bei medizinischen Tests ungern verwendet
- Gibt es vielleicht noch bessere Niveau-α-Tests?

Optimalität des einseitigen Binomialtests

n-facher Münzwurf für die Hypothesen H_0 gegen H_1:

- Wir untersuchen eine sehr große Klasse von (randomisierten) Tests.
- Ein Test ist eine Abbildung $\phi_n : \{0,1\}^n \to [0,1]$ mit:
  - Im Fall $\phi_n(x) = 0$ akzeptiert man H_0.
  - Im Fall $\phi_n(x) = 1$ akzeptiert man H_1.
  - Allgemein wirft man eine Münze mit Wkeit $\phi_n(x)$ für Erfolg und akzeptiert H_1 bei Erfolg.
- Dann hat ϕ_n die Gütefunktion $G_{\phi_n}(p) = \mathbb{E}_p[\phi_n]$.
- Der Wert $\alpha_{\mathit{eff}} = \sup_{p \in H_0} G_{\phi_n}(p)$ heißt effektives Niveau von ϕ_n.
- ϕ_n heißt Niveau-α-Test, falls $\alpha_{\mathit{eff}} \leq \alpha$.

Wir suchen gleichmäßig beste Niveau-α-Tests. Dies sind Niveau-α-Tests ϕ_n, so dass für jeden anderen Niveau-α-Test ψ_n gilt:

$$G_{\phi_n}(p) \geq G_{\psi_n}(p) \qquad (p \in H_1)$$

Optimalität des einseitigen Binomialtests

Theorem (rechtsseitiger Binomialtest)

- *Obiger randomisierter Niveau-α-Test ist gleichmäßig bester Niveau-α-Test.*
- *Jeder gleichmäßig beste Niveau-α-Test hat die obige Form.*

Bemerkungen

- Entsprechendes gilt für den linksseitigen Binomialtest.
- Neyman-Pearson-Theorie gilt allgemeiner für große Klasse einseitiger Testprobleme (exponentielle Familien) [G Satz 10.10]
- Man verwendet dann oft die nicht randomisierte Variante des optimalen Tests.

Was heißt "signifikant"?

Test $H_0 : p \leq p_0$ gegen $p > p_0$

- Falls man H_0 ablehnen kann, nennt man p signifikant größer als p_0.
- Allerdings könnte p nur wenig größer als p_0 sein. Es wäre also aufschlussreich, auch ein Konfidenzintervall für p angegeben.
- Manchmal gibt man stattdessen eine zulässige Toleranz δ vor und testet $H_0 : p \leq p_0 + \delta$ gegen $H_1 : p > p_0 + \delta$.
- Solche *Überlegenheits-Tests* werden in der Medizin verwendet [W].

"signifikant" oder "bedeutsam"?

Beispiel: FAZ-Artikel vom 30.03.2011 zu Fukushima [KK]

- *"Wir sollten das Risiko [eines schweren Störfalls] aber zumindest kennen und richtig beurteilen können und verstehen, dass es bedeutend größer ist, als theoretische Berechnungen ergeben."*
- statistisches Modell: Münzwurf für GAU eines AKWs
- $p_0 = 4 \cdot 10^{-6}$, 2 GAUs in $442 \cdot 30$ Reaktorjahren, $\alpha = 0.05$
- $H_0 : p \leq p_0$ kann abgelehnt werden. Es gilt sogar $\alpha_{\mathit{eff}} = 0.001$. Die unbekannte Wkeit p ist mit 95%-iger Sicherheit größer als $26 \cdot 10^{-6}$.
- $H_0 : p \leq 10 \cdot p_0$ muss bereits in 2011 beim nicht randomisierten Test akzeptiert werden. ($\alpha_{\mathit{eff}} = 0.017$, $c_\alpha = 2$, $\gamma = 0.4$)

Wie könnte es weitergehen? [P]

- Berücksichtigen theoretische Berechnungen menschliches Versagen?
- Untersuche Beispiele aus dem Brückenbau oder Dammbau.

zweiseitige Binomialtests: Überblick

- Kombination zweier einseitiger Tests
- Reduktion auf einen einseitigen Test
- optimaler zweiseitiger Test
- Konzeptionelles zum zweiseitigen Testen
- Äquivalenz-Tests und Optimalität

TOST-Verfahren: Kombination zweier einseitiger Tests

Testproblem $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$

- Idee: zweiseitiger Test aus "Two One-Sided Tests"
- $H_0 = H_0^{(1)} \cap H_0^{(2)}$ mit $H_0^{(1)} : p \geq p_0$ sowie $H_0^{(2)} : p \leq p_0$
- Lehne H_0 ab genau dann wenn $H_0^{(1)}$ abgelehnt werden kann oder wenn $H_0^{(2)}$ abgelehnt werden kann. Dann gilt

$$\mathbb{P}_{p_0}(H_0 \text{ ablehnen}) \leq \mathbb{P}_{p_0}(H_0^{(1)} \text{ ablehnen}) + \mathbb{P}_{p_0}(H_0^{(2)} \text{ ablehnen})$$

- Dies ist ein Beispiel für die *Kumulierung des α-Fehlers* beim mehrfachen Testen.
- Man wählt meist zwei einseitige Niveau-$\alpha/2$-Tests der obigen Form.
- Dies liefert einen (für $n \to \infty$) asymptotisch optimalen Niveau-α-Test (s.u.).

MPVT-Verfahren: Reduktion auf einseitigen Test

Testproblem $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$

- Idee: Akzeptanzbereich aus wahrscheinlichsten Werten unter H_0
  "Most Probable Values Test"
- Testentscheidung anhand Beobachtung $k \in \{0, \ldots, n\}$:

$$H_0 \text{ verwerfen} \iff \mathbb{P}_{p_0}(S_n = k) < c_\alpha$$

- Der kritische Wert c_α muss $\mathbb{P}_{p_0}(H_0 \text{ verwerfen}) \leq \alpha$ erfüllen.
- Für hohe Güte sollte Verwerfungsbereich groß sein. Wähle also $c_\alpha > 0$ maximal, so dass obige Ungleichung erfüllt ist.
- linksseitiger Test für Teststatistik $k \mapsto T_n(k) = \mathbb{P}_{p_0}(S_n = k)$
- Stimmt für $p_0 = 1/2$ mit TOST überein.
- Andernfalls hat MPVT höhere Güte als TOST, ist also asymptotisch optimal.

optimaler zweiseitiger Binomialtest [LR Bsp 4.2.1]

Ein Niveau-α-Test ϕ_n heißt unverfälscht, falls $G_{\phi_n}(p) \geq \alpha$ für alle $p \in H_1$. Dann wird H_1 mindestens so häufig korrekt wie falsch akzeptiert.

- Der gleichmäßig beste Niveau-α-Test unter allen unverfälschten Niveau-α-Tests ist gegeben durch

$$\phi_n = \begin{cases} 1 & S_n < c_1 \text{ oder } S_n > c_2 \\ \gamma_i & S_n = c_i \text{ für } i = 1, 2 \\ 0 & \text{sonst} \end{cases}$$

- Die Konstanten $c_1 < c_2$ und die Wkeiten γ_1, γ_2 sind hierbei eindeutig festgelegt durch $G_{\phi_n}(p_0) = \mathbb{E}_{p_0}[\phi_n] = \alpha$ und

$$\mathbb{E}_{p_0}[\phi_n \cdot S_n] = \alpha \cdot np_0 \ .$$

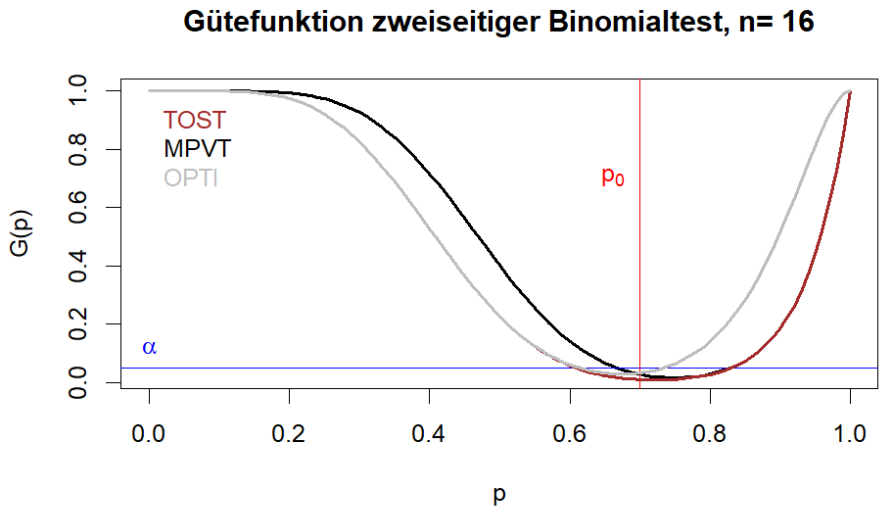
  Dies fixiert den Erwartungswert von S_n im Ablehnungsbereich.

- Für festes c_1, c_2 sind dies lineare Gleichungen für γ_1, γ_2. Die $n(n+3)/2$ Wahlen $0 \leq c_1 < c_2 \leq n$ überprüft man numerisch.

- Man verwendet oft die nicht randomisierte Variante dieses Tests.

Gütefunktionen

Binomialtest $H_0 : p = 0.7$ gegen $H_1 : p \neq 0.7$ zum Niveau $\alpha = 0.05$



- Es sind die nicht randomisierten Versionen der obigen Tests dargestellt.
- Tatsächlich hat der randomisierte optimale Test für Wkeiten nahe $p = 1/2$ eine etwas geringere Güte als der verfälschte MPVT.

Konzeptionelles zum zweiseitigen Testen

(I) Es wird auf $H_1 : p \neq p_0$ getestet. Falls H_0 akzeptiert werden muss, kann man aber nicht erwarten, dass $p = p_0$ gilt.

Eine Interpretation des Testergebnisses beruht auf dem "fränkischen Prinzip": Die Modellierung des Experiments mit einem p_0-Münzwurf widerspricht nicht den Beobachtungen, selbst wenn diese Modellierung das Experiment nicht exakt beschreiben würde.

(II) Es wird auf $H_1 : p \neq p_0$ getestet. Ablehnen von H_0 liefert aber keine Information darüber, wie stark p von p_0 abweicht. Man könnte zusätzlich ein zweiseitiges Konfidenzintervall für p angeben.

(III) Für das umgekehrte Hypothesenpaar $H_0 : p \neq p_0$ gegen $H_1 : p = p_0$ gibt es keinen brauchbaren statistischen Test, weil die Nullhypothese zu groß ist.

Tatsächlich macht es wenig Sinn, eine Hypothese $p = p_0$ statistisch zu erhärten. Meist gibt es keinen "natürlichen" Wert für p_0, selbst wenn die Modellierung das Experiment exakt beschreiben würde.

Konzeptionelles zum zweiseitigen Testen

Beispiel: Wkeit $p_0 = (4 \cdot 365 + 1)^{-1}$ für Geburtstag 29. Februar?

- statistisches Modell: Gleichverteilung der Geburtstage
- Nürnberg 2012: 510.400 Einwohner, 334 Geburtstagskinder
- $H_0 : p = p_0$ muss akzeptiert werden. (349 erwartete, 95%-Konfidenzintervall [299,372] für Zahl der Geburtstagskinder)

allgemeiner:

- In der Statistik werden häufig Tests der Form (I) verwendet. Hierbei wird die Modellierung einer Fragestellung begründet, falls man H_0 akzeptieren muss.
- Beispiele sind Tests auf Vorliegen einer bestimmten Verteilung, Tests auf Unabhängigkeit, Tests auf Gleichheit zweier Erwartungswerte, ...

Wir besprechen im folgenden Tests, welche die Probleme in (II) und (III) angehen, samt Optimalitätsfragen. Solche Tests werden in der Medizin, Pharmazie und Psychologie verwendet.

TOST-Äquivalenztest

Testproblem $H_0 : p \leq p_1$ oder $p_2 \leq p$ gegen $H_1 : p_1 < p < p_2$

- Wähle zum Beispiel $p_1 = p_0 - \delta$ und $p_2 = p_0 + \delta$. Dann testet man

$$H_0 : |p - p_0| \geq \delta \ , \qquad H_1 : |p - p_0| < \delta \ .$$

  Dieser Test auf δ-Äquivalenz von p und p_0 ersetzt das unlösbare Testproblem $H_0 : p \neq p_0$ gegen $H_1 : p = p_0$.
- $H_0 = H_0^{(1)} \cup H_0^{(2)}$ mit $H_0^{(1)} : p \leq p_1$ sowie $H_0^{(2)} : p \geq p_2$
- Lehne H_0 ab genau dann wenn $H_0^{(1)}$ abgelehnt werden kann und wenn $H_0^{(2)}$ abgelehnt werden kann. Dann gilt

$$\mathbb{P}_p(H_0 \text{ ablehnen}) \leq \max\{\mathbb{P}_p(H_0^{(1)} \text{ ablehnen}), \mathbb{P}_p(H_0^{(2)} \text{ ablehnen})\}$$

- Verfahren funktioniert nur bei hinreichend großer Stichprobe!
- Für Niveau-α-Test wählt man obige einseitige Tests zum Niveau α.
- Der so konstruierte Test ist asymptotisch optimal [W Kap 4.4].

TOST-Inäquivalenztest

Testproblem $H_0 : p_1 \leq p \leq p_2$ gegen $H_1 : p < p_1$ oder $p_2 < p$

- Wähle zum Beispiel $p_1 = p_0 - \delta$ und $p_2 = p_0 + \delta$. Dann testet man auf Inäquivalenz

$$H_0 : |p - p_0| \leq \delta \ , \qquad H_1 : |p - p_0| > \delta \ .$$

  Verallgemeinerung des zweiseitigen Binomialtests $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$.

- $H_0 = H_0^{(1)} \cap H_0^{(2)}$ mit $H_0^{(1)} : p \geq p_1$ sowie $H_0^{(2)} : p \leq p_2$
- Lehne H_0 ab genau dann wenn $H_0^{(1)}$ abgelehnt werden kann oder wenn $H_0^{(2)}$ abgelehnt werden kann. Dann gilt

$$\mathbb{P}_p(H_0 \text{ ablehnen}) \leq \mathbb{P}_p(H_0^{(1)} \text{ ablehnen}) + \mathbb{P}_p(H_0^{(2)} \text{ ablehnen})$$

- Zum Beispiel kann man für einen Niveau-α-Test zwei der obigen einseitigen Niveau-$\alpha/2$-Tests verwenden.
- Der so konstruierte Test ist asymptotisch optimal.

Optimale zweiseitige Binomialtests

- Das asymptotisch optimale TOST-Verfahren wird wegen seiner Einfachheit häufig angewendet, nicht nur für Binomialtests.
- Für die TOST-Binomialtests beruht die Testentscheidung auf einer Unterteilung von $E = \{0, \ldots, n\}$ in drei Bereiche

$$E = \{0, \ldots, c_1 - 1\} \cup \{c_1, \ldots, c_2\} \cup \{c_2 + 1, \ldots, n\}$$

  Hiermit soll zB $p < p_1$, $p_1 \leq p \leq p_2$ oder $p > p_2$ begründet werden.
- Tatsächlich ergeben sich *optimale Tests* bei passender Wahl der kritischen Werte c_1 und c_2 und Randomisierung mit Ablehnungswahrscheinlichkeiten γ_1 für c_1 und γ_2 für c_2.
- Den gleichmäßig besten unverfälschten Test für $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$ haben wir bereits oben besprochen.

optimaler Binomialtest auf Äquivalenz

Testproblem $H_0 : p \leq p_1$ oder $p_2 \leq p$ gegen $H_1 : p_1 < p < p_2$

- Die Güte des oben beschriebenen randomisierten Tests ist

$$G(p) = \gamma_1 \mathbb{P}_p(S_n = c_1) + \mathbb{P}_p(c_1 < S_n < c_2) + \gamma_2 \mathbb{P}_p(S_n = c_2) \ .$$

- Wir fordern, dass obiger Test an den Grenzen von H_0 das Niveau α ausschöpfen soll. Es müssen also folgende Bedingungen erfüllt sein:

$$G(p_1) = \alpha \ , \qquad G(p_2) = \alpha \ .$$

- Dies legt die Parameter $0 \leq c_1 < c_2 \leq n$ und $0 \leq \gamma_1, \gamma_2 < 1$ eindeutig fest. Sie können numerisch bestimmt werden [W 4.3].
- Der so erhaltene Test ist der gleichmäßig beste Niveau-α-Test für das obige Testproblem [LR Thm 3.7.1].

Obiges Vorgehen klappt auch für Tests auf Inäquivalenz, und nicht nur für den Münzwurf, sondern für viele exponentielle Familien [LR Kap 4].

Beispiel: Äquivalenztest für Geburtstag am 29. Februar

Beträgt die Wkeit $p_0 = (4 \cdot 365 + 1)^{-1} \approx 6.8 \cdot 10^{-4}$?

- Teste $H_0 : |p - p_0| \geq \delta$ gegen $H_1 : |p - p_0| < \delta$ zum Niveau 0.05
- Wähle $\delta = 0.1 \cdot p_0$, also relative Genauigkeit von 10 Prozent.
- Für den TOST-Äquivalenztest ist die Stichprobe zu klein.
- Für den numerischen Algorithmus aus [W] ist Stichprobe zu groß.
- Allerdings ist eine Normalapproximation eine gute Näherung.
- Verwende also Äquivalenztest für Normalverteilung. Dieser Test ist viel einfacher als für die Binomialverteilung.
- Ergebnis: H_0 muss akzeptiert werden. Nürnberg ist zu klein, um die vermutete Äquivalenz statistisch abzusichern.
- 95%-Konfidenzintervall für p ist $[5.8 \cdot 10^{-4}, 7.3 \cdot 10^{-4}]$

Vertiefung M12 beschreibende und schließende Statistik

- Anhand des Binomialtests wird ein Grundverständnis für statistische Tests erworben.
- Für andere Testprobleme können Tests aus Statistik-Programmen als "black box" verwendet werden, nachdem die Passgenauigkeit der zugrundeliegenden Modelle diskutiert worden ist.
- Dieser Zugang wird im Buch *Mathematik in den Life Sciences* von Gerhard Keller [K] auf universitärem Anfänger-Niveau verfolgt.
- Wir skizzieren die Grundlagen der Testprobleme, welche für die Vertiefung vorgeschlagen wurden:
  - z-Tests, t-Tests
  - χ^2-Anpassungstests, χ^2-Unabhängigkeitstest
  - exakter Test von Fisher auf Unabhängigkeit

  Für die z-Tests und die t-Tests ist ein Grundverständnis von kontinuierlichen Zufallsgrößen notwendig [H 31]!

z-Tests, t-Tests [H 33.9-33.13]

zum Beispiel $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$

- t-Tests auf Erwartungswert μ in Klasse aller Normalverteilungen $N(\mu, \sigma^2)$ bei unbekannter Varianz σ^2
- Teststatistik standardisiertes Stichprobenmittel mit sog. t-Verteilung
- für optimale Tests keine Randomisierung erforderlich
- Der z-Test bei bekannter Varianz ist das didaktische Pendant zum Binomialtest. Kenngrößen werden mit Normalverteilung beschrieben.

auch anwendbar jenseits der Normalverteilung

- asymptotisch exakter Test auf Erwartungswert für *nahezu beliebige* parametrisierte Verteilungsklassen
- Grund: Stichprobenmittel asymptotisch normalverteilt
- Beispiel: Anzahl Erfolge S_n beim n-fachen Münzwurf näherungsweise normalverteilt, falls $np(1-p) \geq 10$, vergleiche [H 27.3-27.4]

χ^2-Anpassungstest, χ^2-Unabhängigkeitstest

- Der χ^2-Unabhängigkeitstest untersucht, ob zwei Zufallsgrößen mit endlichen vielen Werten unabhängig sind. Dieser Test wird aus dem χ^2-Anpassungstest entwickelt.
- Der χ^2-Anpassungstest untersucht, ob eine Zufallsgröße mit endlich vielen Werten bestimmte vermutete Wkeiten besitzt [H 29.7-29.9]. Dies verallgemeinert den zweiseitigen Binomialtest [K 10.2].
- Die diskreten Teststatistiken beider Tests sind asymptotisch χ^2_r-verteilt mit passendem r, d.h. wie die kontinuierliche Summe von r unabhängigen quadrierten $N(0,1)$-Zufallsvariablen.
- Die Theorie zum χ^2-Anpassungstest ist komplex [G 11.2], [LR 14.3.1], die Theorie zum χ^2-Unabhängigkeitstest ist komplexer [G 11.3], [LR 14.3.3].
- Um den Fehler der Normalapproximation zu untersuchen, können die Teststatistiken auf dem Computer simuliert werden. Dies gilt auch für die z-Tests und t-Tests jenseits der Normalverteilung.

exakter Test von Fisher auf Unabhängigkeit

- Der nur asymptotisch exakte χ^2-Unabhängigkeitstest ist bei zu kleiner Stichprobe nicht anwendbar. Diese Situation tritt in der Praxis häufig auf.
- Eine Alternative ist der sogenannte exakte Test von Fisher. Dieser Test wird meist für Zufallsgrößen mit zwei Werten verwendet. Als Teststatistik wird dann ein Eintrag in der Vierfeldertafel verwendet.
- Die Verteilung dieser Teststatistik bei festen Randhäufigkeiten kann mit der hypergeometrischen Verteilung beschrieben werden [G Aufg 11.10, 11.11].
- Für diesen Test existiert eine Optimalitätstheorie [LR 4.6-4.7], welche derjenigen der Binomialtests ähnlich ist, siehe auch [W 6.6].

Literatur zu Statistik und Schuldidaktik

ISB Lehrplan Staatsinstitut für Schulqualität und Bildungsforschung München
`https://www.lehrplanplus.bayern.de`

KWBBH S. Krauss, P. Weber, K. Binder, G. Bruckmaier, S. Hilbert, *Zur Propädeutik des Hypothesentestens in der gymnasialen Oberstufe – Die Diskrepanz zwischen schulischem Stochastikunterricht und tatsächlicher Anwendung*, Preprint (2020)

H N. Henze, Stochastik für Einsteiger, Springer Spektrum (2013)
Einführung auf moderatem mathematischem Niveau, 10. Auflage

K G. Keller, Mathematik in den Life Sciences, Ulmer TB (2011)
Modellbildung und Statistik mit R für Erstsemester

G H.O. Georgii, Stochastik, de Gruyter (2009)
Lehrbuch für Mathematik-Studierende, 5. Auflage

LR E. Lehmann, J. Romano, Testing Statistical Hypotheses, Springer (2022)

W S. Wellek, Testing Statistical Hypotheses of Equivalence and Noninferiority, CRC Press (2010)

CCM C. Christensen, B. Christensen, M. Missong, Statistik klipp & klar, Springer (2019) *beispielorientierter Kompaktkurs für WiWis*

Literatur zu den Beispielen

RKI Corona-Schnelltest-Ergebnisse verstehen, Merkblatt RKI 24.02.2021

K V. Kinne et al, Diagnostische Leistungsfähigkeit von zwei Antigentests, Die Anaesthesiologie 11 (2023), 791–798

S F. Schwilden, "Den Maulkorb brauchen Sie hier nicht", WELT 17.04.2021 mit statistischer Analyse des Effekts von Schnelltests auf Inzidenzen

N R. Nuzzo, Statistical errors, Nature 506 (2014) 150–152
siehe auch die Nature-Merkblätter zur Statistik unter
`www.nature.com/collections/qghhqm/pointsofsignificance`

I J.P.A. Ioannidis, Why most published research findings are false, PLoS Medicine 2 (2005) 696–701, siehe auch ZEIT 14.06.2017

H I.C.W. Hardy, Sex Ratios: Concepts and Research Methods, Cambridge University Press (2003)

KK G. Kauermann, H. Küchenhoff, Nach Fukushima stellt sich die Risikofrage neu, FAZ 30.03.2011

P D. Proske, Ist der Vergleich von Einsturzhäufigkeiten und Versagenswahrscheinlichkeiten sinnvoll?, ce.papers (2019), 48–53