2.3.5 Bestimmung des Stichprobenumfangs

Eine große praktische Bedeutung haben Konfidenzintervalle auch für die Stichprobenplanung. Sie werden herangezogen, um den benötigten Stichprobenumfang zu ermitteln.

- Je genauer und sicherer, desto größer muss der Stichprobenumfang sein
- Genauigkeit: Halbe Länge des Konfidenzintervalls
- Gib Konfidenzniveau (oft 95%) und eine Schranke b_{max} an (inhaltliche Entscheidungen!), also de facto die halbe Breite für die minimal zulässige Genauigkeit
- Bestimme den Mindeststichprobenumfang so, dass γ und b_{max} eingehalten werden

Konkrete Umsetzung für die Anteilsschätzung

γ: Konfidenzniveau

b_{max}: Genauigkeitsschranke

$$z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \quad \leq \quad b_{\text{max}}$$

Auflösen nach n:

$$n \quad \geq \quad \frac{1}{b_{\text{max}}^2} z_{\frac{1+\gamma}{2}}^2 \cdot \bar{X}(1-\bar{X})$$

Da $\bar{X}$ unbekannt ist, verwendet man zur Stichprobenplanung z.B. Schätzwerte früherer Untersuchungen oder als konservative Schranke $\bar{X}(1-\bar{X}) \leq 0.25$.[4]

[4] Dies gilt immer:
betrachte $f(a) = a \cdot (1-a) = -a^2 + a;\ a \in [0,1]$
$\frac{df(a)}{d(a)} = 0 \Leftrightarrow -2a + 1 = 0 \Rightarrow a = \frac{1}{2}$
$\frac{d^2 f(a)}{da^2} = -2$
also Maximum an der Stelle $a_{\text{max}} = \frac{1}{2}$ mit $f(a)_{\text{max}} = (-\frac{1}{2})^2 + \frac{1}{2} = \frac{1}{4}$

Bsp. 2.23.

Gegeben:

- Konfidenzniveau: $\gamma = 95\%$
- Genauigkeit: $b_{\text{max}} = 0.1$

Bestimmung von n:

$$n \geq \frac{1}{b_{\text{max}}^2} z_{\frac{1+\gamma}{2}}^2 \cdot \bar{X}(1-\bar{X}) = \frac{1}{0.1^2} 1.96^2 \cdot 0.25 = 96.04$$

wobei konservativ angesezt wurde $\bar{X}(1-\bar{X}) \leq 0.25$. Also sollten ca. 100 Personen befragt werden.

Bei $\gamma = 5\%$ ergibt sich $n = 385$, bei $\gamma = 1\%$ ergibt sich $n = 9604$

2.3.6 Ausblick

Viele weitere Konfidenzintervalle, auch bezüglich der Unterschiede bei zwei Stichproben

- Differenz von Mittelwerten bei unabhängigen Stichproben
- Differenz von Anteilen bei unabhängigen Stichproben
- Differenz von Mittelwerten bei „verbundenen Stichproben“

Diese gehorchen immer dem gleichen Konstruktionsprinzip (vgl. von Kap. 2.3.3) und sind analog zu den genauer besprochenenen Beispielen zu interpretieren. Man vergleiche auch die Konstruktion von Tests im Kapitel 2.4, die auf einem ähnlichen Ausgangspunkt rekurrieren. In der Tat gibt es einen engen Zusammenhang zwischen Hypothesentests und Konfidenzintervallen (vgl. „Dualität“ am Ende von Kap. 2.4).

2.4 Hypothesentests

2.4.1 Grundprinzipien statistischer Hypothesentests; Motivationsbeispiel

Hypothese: „Behauptung einer Tatsache, deren Überprüfung noch aussteht“ (Leutner in: Endruweit, Trommsdorff: Wörterbuch der Soziologie, 1989).

Statistischer Test: Überprüfung von Hypothesen über die Grundgesamtheit anhand einer Stichprobe

Statistische Testtheorie: Schließe von Stichprobe auf Zutreffen bestimmter Tatsachen in der Grundgesamtheit
Vorgehen:

- inhaltliche Hypothese aufstellen
- Operationalisierung
- inhaltliche Hypothese in statistische Hypothese „übersetzen“
- statistischer Test

Bsp. 2.24.

Studie zur Einstellung der Münchner Bevölkerung zu psychisch Kranken (1989, Methodenkurs Siegfried Lamnek).

Wir betrachten eine Teilstudie: Kooperationsbereitschaft in der Befragung.

1. „Theorie“: Aktive Stellung im öffentlichen Leben beeinflusst Kooperationsbereitschaft positiv.

$$\begin{aligned} Aktiv \quad &\leftrightarrow \quad \text{Altruismus} \\ &\leftrightarrow \quad \text{Interesse an öffentlichen Angelegenheiten} \\ &\qquad\quad \Rightarrow \text{eher bereit, die Rolle des Befragten einzunehmen} \end{aligned}$$

2. Hypothese: „Die Koorperationsbereitschaft der aktiven Personen unterscheidet sich vom Rest der Bevölkerung?“

3. Operationalisierung:

   - Aktiv im öffentlichen Leben

     $\rightarrow$ Verbandsmitgliedschaft ja/nein $=$ Variable X

   - Kooperationsbereitschaft

     $\rightarrow$ Von den überhaupt nicht Kooperierenden hat man keine Daten, deshalb vergleicht man die sofort zur Antwort bereiten Befragten mit den „Primärverweigerern“, die sich erst nach langer Zeit zur Auskunft überreden ließen.
     $\rightarrow$ direkt kooperativ ja/nein $\rightarrow$Variable Y

4. Statistische Hypothesen: „Besteht ein Zusammenhang zwischen X und Y?“

   Statistisches Vorgehen:
   Kann die sog. $Nullhypothese$ „Es besteht <u>kein</u> Zusammenhang zwischen X und Y“ abgelehnt werden?

Herleitung / Motivation eines geeigneten Prüfverfahrens, basierend auf χ^2 (vgl. Statistik I)

Gegebene Daten (relative und absolute Häufigkeiten):

		kooperativ		
		ja	nein	
aktiv	ja	0.27	0.05	0.32
		(95)	(17)	(112)
	nein.	0.53	0.15	0.68
		(186)	(54)	(240)
		0.8	0.2	1
		(281)	(71)	(352)

Vergleiche gegebene Tafel mit der „Unabhängigkeitstafel“

Wie würde denn die Tafel aussehen, wenn kein Zusammenhang bestünde?

Genauer: wie würde das Innere der Tabelle aussehen, wenn Unabhängigkeit (und die gleichen Randverteilungen) herrschen würde, also die Nullhypothese zutreffen würde?

		kooperativ		
		ja	nein	
aktiv	ja	0.256	0.064	0.32
	nein	0.544	0.136	0.68
		0.8	0.2	1

Die Häufigkeiten in der Unabhängigkeitstafel weichen von den tatsächlichen Daten ab. (Vgl. Statistik I: Je stärker die Abweichung, desto stärker ist der Zusammenhang.)

- Wie groß muss die Abweichung sein, um die Nullhypothese abzulehnen?
- Beachte: Die Daten entstammen einer Stichprobe, die mit einem Zufallsfehler behaftet ist. Selbst bei tatsächlich vorliegender Unabhängigkeit ist die Wahrscheinlichkeit, genau die Unabhängigkeitstafel zu beobachten, sehr gering.

Kardinalfrage der Testtheorie:

Weichen die tatsächlichen Daten von der bei Gültigkeit der Nullhypothese zu erwartenden Situation „überzufällig“ stark ab, d.h. so stark, dass man die Abweichung nicht mehr nur der Zufallsstreuung zuschreiben kann?
Nur in diesem Fall ist die Nullhypothese abzulehnen.

Wann ist die Abweichung „groß“, d.h. „überzufällig“?

- Bestimme eine Zufallsvariable T, die in geeigneter Weise den Unterschied einer zufälligen Stichprobe zur Situation der Nullhypothese misst (hier: der χ^2-Abstand zwischen einer Stichprobe und der Unabhängigkeitstafel, vgl. Statistik I).
- Bestimme einen Bereich KR, der “sehr unwahrscheinlich“ ist, falls H_0 gilt („*Ablehnungsbereich*“, „*kritische Region*“)
- Bestimme die Realisation t von T anhand der konkreten Daten (hier: χ^2=2.11)!
  Ist $t \in$ KR, dann Nullhypothese ablehnen, sonst nicht.

Dabei kann man zwei Arten von Fehlern machen:

Wahrheit \ Aktion	H_0 beibehalten	H_0 ablehnen
H_0 wahr	$\surd$	Fehler 1.Art
H_0 falsch	Fehler 2. Art	$\surd$

Fehler 1. Art („α-Fehler “):

Die Nullhypothese wird abgelehnt, obwohl sie in Wirklichkeit richtig ist.

z.B.: Man behauptet, es bestünde ein Zusammenhang, obwohl in Wirklichkeit kein Zusammenhang besteht.

Der Fehler 1. Art soll klein sein (üblich sind 5% oder 10%). Allerdings kann man nicht fordern, dass der Fehler 1. Art bei 0% liegen soll, sonst würde man die Nullhypothese nie ablehnen können.

Fehler 2. Art („β-Fehler “):

Die Nullhypothese wird beibehalten, obwohl sie in Wirklichkeit falsch ist.

Dabei wird so vorgegenagen, dass die **Wahrscheinlichkeit**, einen **Fehler 1.Art** zu begehen, **beschränkt** ist durch eine inhaltliche vorgegebene Schranke α („***Signifikanzniveau***“)
Ein guter statistischer Test garantiert bei einem vergegebenen niedrigen Signifikanzniveau (als Schranke für die Fehlerwahrscheinlichkeit 1. Art) auch eine möglichst geringe Fehlerwahrscheinlichkeit 2. Art.

2.4.2 Präzisierung: Konstruktion eines parametrischen statistischen Tests

1. *Aufstellen der substanzwissenschaftlichen Hypothese / inhaltliche Fragestellung*

   (z.B. Rot/Grün bekommt die absolute Mehrheit, der mittlere Intelligenzscore der Gruppe $\cdots$ beträgt mindestens 130 Einheiten)

2. *Formulieren eines geeigneten statistischen Modells*

   Im Folgenden stets $X_1, \ldots, X_n$ i.i.d. Stichprobe sowie parametrisches Modell mit unbekanntem Parameter ϑ.

   Anteil: $B(1, \pi)$

   Score: $\mathcal{N}(\mu; \sigma^2)$.

3. *Formulierung der statistischen Hypothesen*

- Umformulieren der substanzwissenschaftlichen Hypothesen als Hypothesen über ϑ.
- Verglichen wird immer eine sog. *Nullhypothese* (H_0) mit einer sog. *Alternativhypothese* (H_1).
- Bei parametrischen Fragestellungen unterscheidet man:

a) *einseitige Testprobleme*:

$$H_0 : \vartheta \leq \vartheta_0 \text{ gegen } H_1 : \vartheta > \vartheta_0$$

$$H_0 : \vartheta \geq \vartheta_0 \text{ gegen } H_1 : \vartheta < \vartheta_0$$

b) *zweiseitiges Testproblem*:

$$H_0 : \vartheta = \vartheta_0 \text{ gegen } H_1 : \vartheta \neq \vartheta_0$$

ϑ_0 ist ein fester, vorgegebener Wert, der von inhaltlichem Interesse ist; zu unterscheiden von wahrem Wert ϑ.
Der Begriff einseitig/zweiseitig bezieht sich auf die Alternative, je nachdem ob die Alternative nur aus großen bzw. nur aus kleinen Werten besteht oder ob sowohl große als auch kleine Werte für die Alternative sprechen.

4. *Festlegung des Signifikanzniveaus* α
   Wiederholung aus der Einleitung
   Beim Testen sind folgende Entscheidungen möglich:

$$H_0\text{: ablehnen} \quad \text{oder} \quad H_0\text{: beibehalten}$$

Damit sind zwei verschiedene Arten von Fehlern möglich:

Wahrheit \ Aktion	H_0 beibehalten	H_0 ablehnen
H_0 wahr	$\surd$	Fehler 1.Art
H_0 falsch	Fehler 2. Art	$\surd$

Man kann nicht beide Fehlerwahrscheinlichkeiten gleichzeitig kontrollieren! (Trade-off!)
$\Rightarrow$ asymmetrische Vorgehensweise:

Die Fehlerwahrscheinlichkeit 1. Art wird kontrolliert durch die Angabe einer Oberschranke α („Signifikanzniveau“)

Typische Werte: üblich

$$\alpha = 0.1, \quad \alpha = 0.05, \quad \alpha = 0.01 \quad \alpha = 0.001$$

5. *Festlegen einer Testgröße und einer kritischen Region*

Eine *Testgröße* T ist eine Zufallsgröße $T = g(X_1, \ldots, X_n)$, die „empfindlich gegenüber Abweichungen von H_0 ist“. Die *Kritische Region* KR („Ablehnungsbereich“) besteht aus potentiellen Werten von T, die stark gegen H_0 sprechen.

Liegt t (Realisation von T) in KR, wird man sich gegen H_0 entscheiden.

Damit der Fehler 1. Art durch α beschränkt bleibt, muss die kritische Region KR also so gewählt werden, dass

$$P(T \in KR || H_0) \leq \alpha$$

gilt, d.h. die Wahrscheinlichkeit, dass T in der kritischen Region liegt und damit zur Ablehnung von H_0 führt, darf höchstens α sein, wenn H_0 stimmt.

Umgekehrt soll $P(T \in KR || H_1)$ möglichst groß sein, da dies die Wahrscheinlichkeit ist, die Nullhypothese H_0 abzulehnen, falls sie falsch ist. (Gegenwahrscheinlichkeit zur Wahrscheinlichkeit für den Fehler 2. Art, auch als *Power* oder *Güte* des Tests bezeichnet.)

6. *Auswerten der Stichprobe*

   Berechnung der Realisation t der Testgröße T basierend auf der konkret vorliegenden Stichprobe.

7. *Testentscheidung*

   Ist $t \in KR$, dann H_0 ablehnen, sonst nicht ablehnen.

Bem. 2.25.

- Die wesentlichen Elemente des Tests (Signifikanzniveau, Testgröße, kritische Region) sind unabhängig von den Daten, also vor der Auswertung, zu bestimmen.
- Da nur die Fehlerwahrscheinlichkeit 1. Art kontrolliert werden kann, kann H_0 nicht mit einer a priori kontrollierten Fehlerwahrscheinlichkeit angenommen, sondern nur abgelehnt oder nicht abgelehnt werden.

- Als „gute(r) Forscher(in)“ sollte man deshalb immer das, was man zeigen will, in die Alternativhypothese schreiben.

  z.B. Forscher(in) will zeigen, dass seine Interventionsmaßnahme besser wirkt als ein anderes.
  Nullhypothese: Sie wirkt schlechter oder gleich gut.
  Alternativhypothese: Sie wirkt besser.
  Durch die Kontrolle des Fehlers 1. Art ist gewährleistet, dass die Wahrscheinlichkeit, der Maßnahme irrtümlich eine bessere Wirkung zuzuschreiben, höchstens α ist.

- Allerdings gibt es keineswegs immer einen (einfachen) statistischen Test für jede Nullhypothese.

  z.B. ist es technisch viel einfacher als Nullhypothese $\theta = \theta_0$ zu verwenden, als $\theta \neq \theta_0$. Verwendet man deshalb einen Test mit $H_0 : \theta = \theta_0$, möchte man inhaltlich aber genau dies zeigen, kehren sich die Rollen des Fehlers 1. Art und des Fehlers 2. Art um. Um in diesem Fall einen geringeren Fehler 2. Art zu erzielen, sollte das Signifikanzniveau höher als üblich gewählt werden. Dies ist aber nur ein erster Versuch, diesem Problem beizukommen. Eine saubere technische Behandlung führt auf sogenannte „Äquivalenztests“.