Punktschätzer und Konfidenzintervalle

Dr. Mariana Nold

Institut für Soziologie,
Fakultät für Sozial- und Verhaltenswissenschaften,
Lehrstuhl für empirische Sozialforschung und Sozialstrukturanalyse

27. November 2017

Übersicht

Ziel der heutigen Veranstaltung ...

ist es die folgenden Fragen beantworten zu können:

Zielfragen für heute

1. Welche Probleme treten auf, wenn man den Fehler 2. Art nicht berücksichtigt?
2. Wie ist die Power eines Tests definiert?
3. Wie ist die Schätzstatistik definiert?
4. Welche Punktschätzung verwendet man für Anteilswerte?
5. Was ist der Standardfehler?
6. Warum braucht man die Intervallschätzung?
7. Was versteht man unter eine Konfidenzintervall und wie interpretiert man es?
8. Wie sieht das Konfidenzintervall für Anteilswerte aus?

Kritik am klassischen statistischen Testen

(vgl. LMLG S. 136 ff und 166 ff)

- Statistisches Testen spielt in der Praxis der empirischen Sozialforschung einer herausragende Rolle.
- Tatsächlich ist sowohl das Konzept der statistischen Signifikanz als auch die Logik des statistischen Testens überhaupt seit Jahrzehnten der Kritik ausgesetzt.
- Immer wieder haben renommierte Wissenschaftler gefordert, man solle statistische Signifikanztests abschaffen.
- Ich möchte heute einen Kritikpunkt vorstellen und eine Möglichkeit die mit weniger Nachteilen verbunden ist.

Die Füllmenge der Flaschen

- In der Aufgabe 6 des letzten Aufgabenblattes ging es um die Frage, ob eine Maschine die Mineralwasserflaschen befüllt neu eingestellt werden muss.
- Die Maschine soll exakt 500 ml pro Flasche abfüllen. Nehmen wir an, der wahre Mittelwert der Füllmenge ist μ und die wahre Standardabweichung σ. Beide Parameter sind unbekannt. Der Stichprobenumfang ist n.
- Wir wollen uns heute zunächst mit dem Fehler 2. Art beschäftigen. Also mit dem Fehler eine unwahre Nullhypothese nicht zu erkennen und beizubehalten.
- Es kann auch passieren, dass ein relevanter Unterschied nicht signifikant ist. Dass ist dann ein Fehler 2. Art.

Wie falsch darf die Maschine arbeiten?

- Die Toleranz wird auf 20 ml in beide Richtungen festgelegt. Das bedeutet: Eine Akzeptable Füllmenge liegt im Intervall $(480, 520)$.
- Die Instandsetzung der Maschine zur Neuadjustierung der Füllmenge kostet 200 Euro. Daher möchte die Mitarbeiterin diese nur vornehmen, wenn sie auch nötig ist.
- Da eine Abweichung der Füllmenge sowohl nach oben, als auch nach unten zu Problemen führt, arbeiten mir mit der ungerichteten Null-Hypothese $H_0 : \mu = \mu_0 = 500$
- Wenn also die wahre erwartete Füllmenge 490 ml beträgt und der Test diese Abweichung erkennt und H_0 ablehnt, dann entsteht ein Schaden von 200 Euro.

Der Test kennt keinen Toleranzbereich

- Bitte beachten Sie: Die Nullhypothese ist schon bei der geringsten Abweichung nicht mehr wahr. Selbst, wenn die Maschine eine erwartete Füllmenge von 503 ml hat, ist die Nullhypothese tatsächlich falsch.
- Ein Kritikpunkt an der Praxis des statistischen Testens ist daher, dass man unterscheiden sollte zwischen **relevanten** und **signifikanten** Unterschieden.
- Es kann passieren, dass ein nicht relevanter Unterschied als signifikant nachgewiesen wird. Das ist kein Fehler, weder ein Fehler 1. Art, noch ein Fehler 2. Art.

Power oder Stärke eines statistischen Tests

Um zu verstehen, wie man einen statistischen Test richtig anwenden sollte, brauchen wir den Begriff der Power.

Definition: Power oder Stärke eines statistischen Tests

Die Stärke eines statistischen Tests ist dessen Fähigkeit, einen in der Grundgesamtheit vorhandenen Unterschied (oder eine andere Größe, die wir testen wollen) auch tatsächlich zu ermitteln.
Es gilt also: Power$=$ 1-β (Wahrscheinlichkeit des Fehlers 2. Art), eine hohe Power bedeutet also ein kleines Risiko, einen Fehler 2. Art zu begehen.
Eine hoher Power bedeutet aber auch ein hohes Risiko nicht relevante Unterschiede als signifikant nachzuweisen.

Die Power und der Stichprobenumfang

- Desto höher der Stichprobenumfang ist, desto mehr empirischen Information liegt vor. Daher steigt die Power eines Tests mit dem Stichprobenumfang.
- In Aufgabe 6 hatte die Mitarbeiterin 20 Flaschen getestet. Ist das ein gute Anzahl? Wird durch diese Anzahl ein relevanter Unterschied nachweisbar?
- Um diese Frage zu diskutieren, gehen Sie bitte davon aus, dass die Standardabweichung der Maschine 20 ml beträgt. Dieser Wert ist der Mitarbeiterin allerdings nicht bekannt.

Wie hoch ist die Power des einfachen t-Test?

- Die erwartete Füllmenge, die wir testen wollen heißt μ_0
- Die uns unbekannte wirkliche erwarte Füllmenge heißt μ.
- Die wahre Füllmenge ist normalverteilt mit den Parameter μ und σ.
- Die Power dieses Test hängt von der Effektgröße

$$\delta := \frac{\mu - \mu_0}{\sigma}$$

  ab.
- Für unser Beispiel gilt:

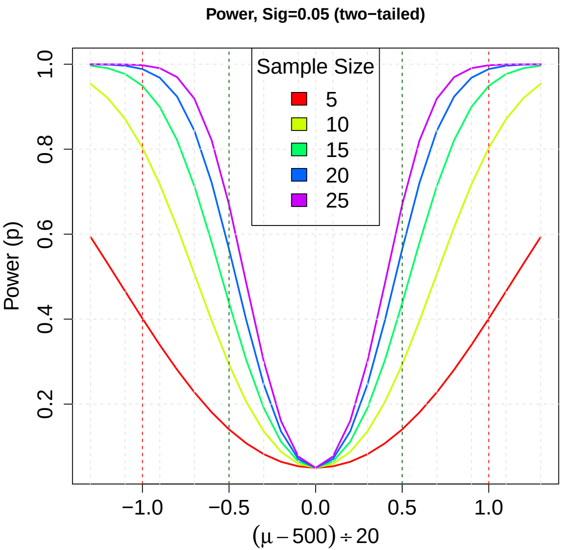
$$\delta := \frac{\mu - 500}{20}$$

- Wie viele Flaschen sollte die Mitarbeiterin für Ihren Test verwenden?
- Wie viele Flaschen sollte Sie verwenden, wenn schon eine Abweichung von 10 ml relevant wäre?

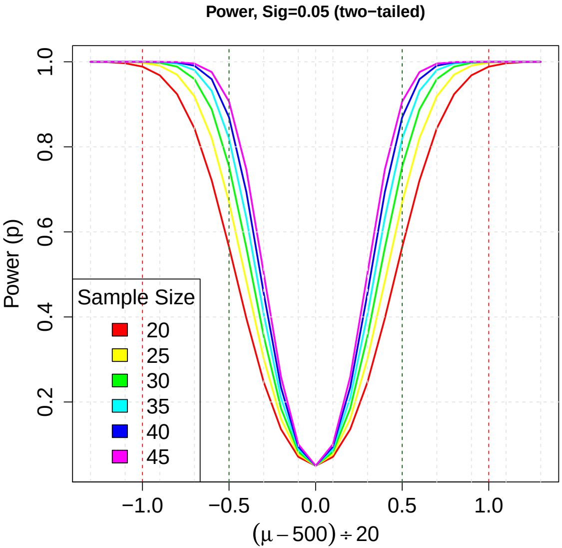
Was ist ein relevanter Unterschied?

- Die erlaubte Toleranz sind 20 ml, also genau eine Standardabweichung. Der Idealfall wäre, wenn der Test nur signifikant ist, wenn die Füllmenge außerhalb des Intervalls $(480, 520)$ liegt.
- Wenn man dieses Intervall auf die Effektgröße umrechnet, erhält man das Intervall $(-1, 1)$.
- Die Grafik auf der nächsten Folie zeigt die Power des einfachen t-Test bei ungerichtete Nullhypothese $\mu = \mu_0 = 500$ für unterschiedliche Stichprobenumfänge.

Power für die Nullhypothese $\mu = \mu_0 = 500$



Ein zu großer Stichprobenumfang?



Statistische Tests unter Berücksichtigung der Teststärke (vgl. LMLG S. 171)

Statistisches Testen, das unter Berücksichtigung der Ideen von Neyman und Pearson auf die Teststärke achtet, muss

1) der Nullhypothese eine eindeutige Alternativhypothese gegenüber stellen.
2) mit Blick auf diese Alternativhypothese (was ist relevant?) den Stichprobenumfang so anpassen,
3) dass eine zufriedenstellende Teststärke erzielt wird.

Es gibt statistische Software, die die Power für viele Tests berechnet. Es ist allerdings die Ausnahme, dass jemand solche Software verwendet.

Statistisches Testen

- Eine vollständige Diskussion mit den Pro und Kontra-Punkten für und gegen das statische Test, können wir hier nicht führen.
- Die bisherigen Folien geben nur einen kleinen Einblick.
- Im Rest der Veranstaltung wollen wir uns mit Punktschätzern und Konfidenzintervallen beschäftigen.
- Die Auswertung der Daten mit Hilfe von Punktschätzern und Konfidenzintervallen ist klar im Vorteil gegenüber statistischen Tests.

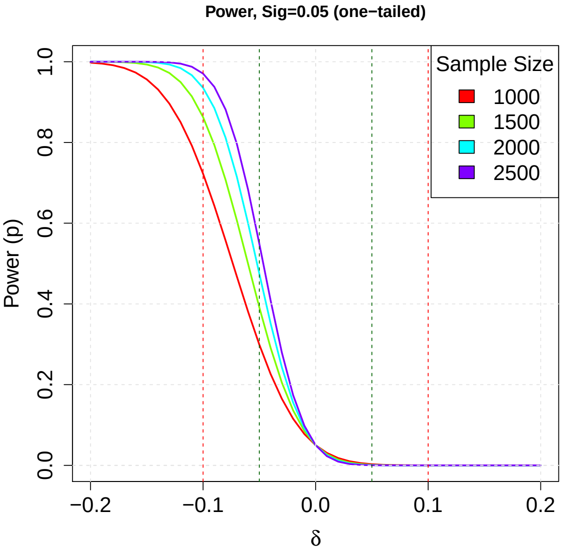
Ausblick Übung: Lesen Mädchen besser als Jungs?

- In Aufgabe 5 des Aufgabenblattes 2, konnten Sie mit dem doppelten t-Test nachweisen, dass Mädchen besser lesen als Jungen.
- Bisher haben wir bei diesem Test die Teststärke nicht berücksichtigt, dass soll in der nächsten Übung erfolgen.
- Ist der signifikante Unterschied auch relevant?
- Das Effektmaß δ des doppelten t-Tests ist:

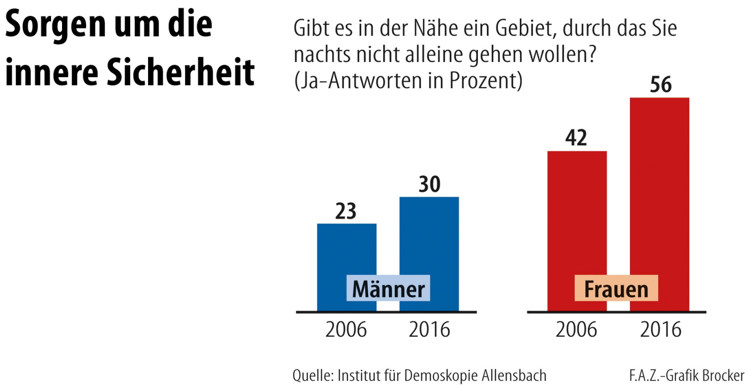
$$\delta := \frac{\mu_X - \mu_Y}{\sigma},$$

wobei μ_X der Erwartungswert der Lesepunkte der Mädchen ist und μ_Y der Erwartungswert der Lesepunkte der Jungen. Die Standradabweichung innerhalb der Gruppen wird als homogen angenommen und mit σ bezeichnet.

Übung: Lesen Mädchen besser als Jungs?



Die diffusen Ängste der Deutschen (17.2.16)



Definition der Zufallsvariablen X_i

Die Grundgesamtheit für die wir uns im Folgenden interessieren ist, die der deutschen Männer im Jahr 2016.

Wie hoch ist der Anteil der besorgten Männer in der Grundgesamtheit?

Wir nennen X_i die Antwort des i ten Mannes in der Befragung.

Es ist

$$X_i := \begin{cases} 0, & \text{“nein,”} \\ 1, & \text{“ja”} \end{cases}$$

je nachdem ob der i-te Mann in der Stichprobe ablehnt oder zustimmt.

Die Stichprobenvariablen

- Ausgangspunkt der Punktschätzung sind n Stichprobenziehungen oder Zufallsexperimente, die durch Zufallsvariablen $X_1, \ldots X_n$ repräsentiert werden.
- $X_1, \ldots X_n$ werden auch als Stichprobenvariablen bezeichnet.
- Häufig fordert man von Stichprobenfunktionen, dass sie unabhängige Wiederholungen von X sind. Durch diese knappe Formulierung wird ausgedrückt, dass
  - die den Zufallsvariablen $X_1, \ldots X_n$ zugrundeliegenden Experimenten unabhängig sind,
  - jedes mal dasselbe Experiment (enthalten in "Wiederholung") durchgeführt wird.
- Man kürzt diese Annahme mit iid-Annahme ab.

Der (Punkt)Schätzer

Der etablierte Schätzer für den Anteil p in der Grundgesamtheit ist

$$\hat{p} = \bar{X} = \sum_{i=1}^{n} x_i.$$

Allgemein gilt:

Definition: Punktschätzer, Schätzfunktion, Schätzstatistik

Ein Punktschätzer für den Grundgesamtheitsparameter θ ist eine Funktion

$$T = \hat{\theta} = g(X_1, \ldots X_n)$$

der Variablen $X_1, \ldots X_n$. Der aus den Realisationen $x_1, \ldots x_n$ resultierende numerische Wert

$$t = \hat{\theta} = g(x_1, \ldots x_n)$$

ist der zugehörige Schätzwert.

Erwartungstreue (vgl. FS. 49)

Die Punktschätzer, die wir in diesem Semester behandeln, sind erwartungstreu.

Definition: Erwartungstreue

Eine Schätzstatistik $\hat{\theta} = g(X_1, \dots X_n)$ heißt erwartungstreu für θ, wenn gilt

$$\mathbb{E}_\theta(\hat{\theta}) = \theta.$$

Der Erwartungswert ist ein Lagemaß für die Verteilung von Zufallsvariablen. Der Erwartungswert einer Zufallsvariable wird gebildet indem man die Werte die die Zufallsvariable annehmen kann mit der entsprechenden Wahrscheinlichkeit gewichtet.
Daher ist die Forderung der Erwartungstreue eine sehr naheliegende und sinnvolle Forderung an einen Schätzer.

Interpretation Erwartungstreue

- Wichtig ist das Verständnis, dass der Erwartungswert eine wichtiges Lagemaß für eine Zufallsvariable ist.
- Wenn die Zufallsvariable eine Schätzstatistik ist, dann gibt ihr Erwartungswert die zentrale Tendenz dieser Schätzstatistik wieder.
- In einfachen Worten bedeutet das: Man erwartet, dass die Realisation der Schätzstatistik dem wahren Parameter entspricht, wenn Erwartungstreue vorliegt.

Punktschätzer und Intervallschätzer für den Anteilswert

- Wir gehen im Folgenden von einer iid Stichprobe aus. X_i iid wie $X \sim Be(p)$.
- Bei der Punktschätzung geht es darum, einen Parameter in der Grundgesamtheit zu schätzen.
- Der entspreche Anteil in der Stichprobe ist 30%.
- Der Schätzer $\hat{p}$ für den Anteil p der besorgten Männer in der Grundgesamtheit ist daher 0.3.
- Wir schätzen also diesen Anteil in der Grundgesamtheit auf 30%. Aber wie präzise ist die Schätzung?
- Kann es sein, dass es in Wirklichkeit nur 10% sind oder gar 50%?

Warum brauchen wir die Intervallschätzung

- Die Punktschätzung liefert einen Schätzer $\hat{\theta}$ im konkreten Beispiel $\hat{p}$.
- Dieser ist im Regelfall nicht mit dem wahren Parameter θ bzw. p identisch.
- In jeder sinnvollen Anwendung ist es daher notwendig, neben dem Schätzwert selbst, die Präzision des Schätzverfahren mitanzugeben.
- Ein Maß für die Präzision ist der Standardfehler, allerdings nur bei erwartungstreuen Schätzern mit symmetrischer Verteilung.
- Ein anderer Weg die Genauigkeit des Schätzverfahren direkt miteinzubeziehen, ist die Intervallschätzung.
- Das Ergebnis der der Intervallschätzung ist nicht ein Wert sondern ein Intervall. Dieses Intervall heißt Konfidenzintervall.

Ein Konfidenzintervall für den Anteilswert

Das approximative $(1-\alpha)$-Konfidenzintervall für den Anteil p in einer Grundgesamtheit ist

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \hat{p} + z_{1-\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}\right)\right], \tag{1}$$

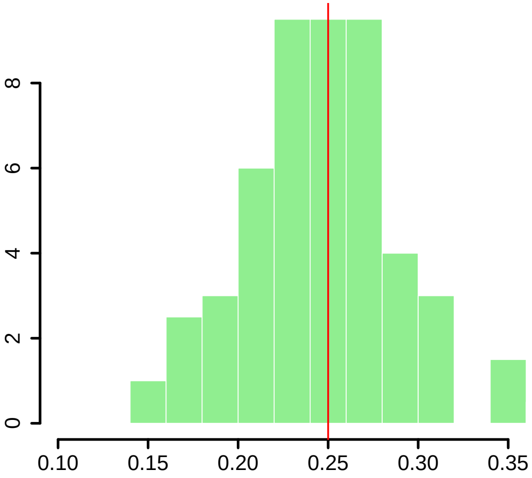
wobei $z_{1-\frac{\alpha}{2}}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung ist.

- Man kann mit `STATA` oder andere Statistik-Software eine Intervallschätzung für eine vorher festgelegtes Konfidenzniveau α berechnen.
- Asymptotisch bzw. approximativ bedeutet, dass man bei großen Stichprobenumfängen dieses Intervall verwendet werden darf.
- Daher steht in der FS S. 52 die Voraussetzung $n \cdot p(1-p) \geq 9$.
- Im Folgenden soll die Bedeutung des Konfidenzintervalls an einer fiktiven Studie demonstriert werden.

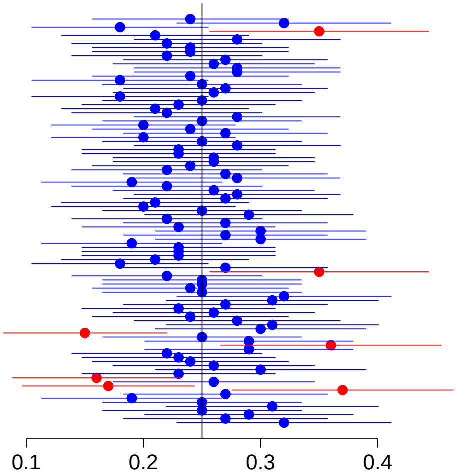
Welche praktische Bedeutung hat das Konfidenzintervall?

- Im Folgenden gehen wir davon aus, dass der wahre (und unbekannte) Anteil der Männer, die der Frage zustimmen, in der Grundgesamtheit bei 25% liegt.
- Sie können sich vorstellen, dass k Personen losgehen und jede Person eine einfach Zufallsstichprobe mit je 100 Personen zieht.
- Wir gehen dabei davon aus, dass die Grundgesamtheit so groß ist, dass keine Person doppelt vorkommt.
- Die Folgenden Grafiken zeigen für $k \in (100, 200, 250)$
  1) das Histogramm der beobachteten Anteilswerte $\hat{p} = \bar{x}$ und
  2) die entsprechenden Konfidenzintervalle wie in (1) (vgl. FS S. 52)

Beobachtete Anteilswerte $\hat{p}$: 100 Befragungen



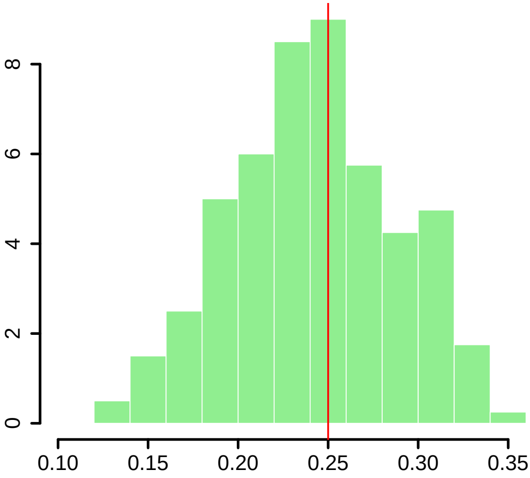
Konfidenzintervalle zu den Anteilswerten: 100 Befragungen



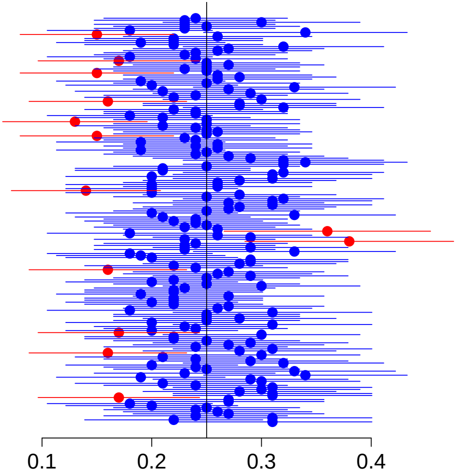
Betrachtung der Ergebnisse

- Von den 100 Personen, die losgegangen sind, um jeweils eine Stichprobe von 100 Personen zu ziehen, haben sieben eine Stichprobe gezogen, deren Konfidenzintervall den wahren Wert nicht enthält.
- Die anderen 93 haben eine Stichprobe gezogen, deren Konfidenzintervall, den wahren Wert enthält.
- Wir wiederholen das Experiment. Diesmal gehen 200 Personen los um jeweils 100 Personen zu befragen.
- Was erwarten Sie? Wie viele Konfidenzintervalle werden den wahren Anteilswert nicht enthalten?

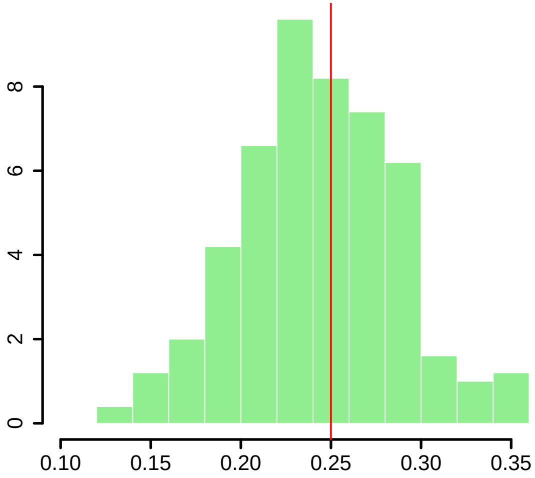
Beobachtete Anteilswerte $\hat{p}$: 200 Befragungen



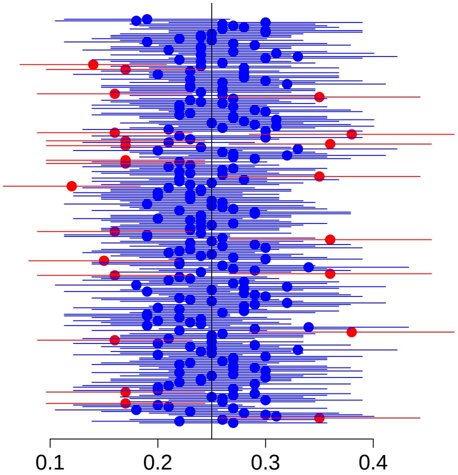
Konfidenzintervalle zu den Anteilswerten: 200 Befragungen



Beobachtete Anteilswerte: 250 Befragungen



Konfidenzintervalle zu den Anteilswerten: 250 Befragungen



Das zweiseitige $(1-\alpha)$- Konfidenzintervall für θ

Das $(1-\alpha)$- Konfidenzintervall für θ

Zu vorgegebener Irrtumswahrscheinlichkeit α liefern die aus den Stichprobenvariablen $X_1, \ldots X_n$ gebildeten Schätzstatistiken $G_u = g_u(X_1, \ldots X_n)$ und $G_o = g_o(X_1, \ldots X_n)$ ein $(1-\alpha)$-Konfidenzintervall (Vertrauensintervall), wenn gilt

$$P(G_u \leq G_o) = 1 \qquad (*)$$

$$P(G_u \leq \theta \leq G_o) = 1 - \alpha. \qquad (**)$$

$1-\alpha$ wird als Sicherheits- oder Konfidenzwahrscheinlichkeit bezeichnet. Das sich aus den Realisationen $x_1, \ldots x_n$ ergebende realisierte Konfidenzintervall besitzt die Form $[g_u, g_0]$, wobei $g_u = g_u(x_1, \ldots x_n)$ und $g_o = g_o(x_1, \ldots x_n)$.

Interpretation des $(1-\alpha)$- Konfidenzintervall

- Die obige Definition besagt, dass die Intervallschätzung gerade so festgelegt ist, dass zwei Dinge erfüllt sind:
  1) Die untere Intervallgrenze liegt mit Sicherheit niedriger ist als die obere Intervallgrenze (*). Dieser Teil der Definition ist aus mathematischen Gründen erforderlich und braucht Sie nicht zu interessieren.
  2) Vor der Berechnung des Konfidenzintervalls ist die Wahrscheinlichkeit dass man ein Konfidenzintervall erhält, dass den wahren Parameter enthält eben gleich $1-\alpha$.(**) Dieser Teil der Definition ist für die Interpretation grundlegend wichtig.
- Fazit: Vor Berechnung des Konfidenzintervalls ist als die Wahrscheinlichkeit ein "blaues" Intervall zu bekommen $(1-\alpha)$ und die Wahrscheinlichkeit für ein "rotes" Intervall α.

Interpretation des Konfidenzintervalls

Was bedeutet diese Interpretation für die Praxis:

- Wenn Sie ein Konfidenzintervall berechnet haben, dann ist es ein "blaues" (enthält θ) oder ein "rotes" (enthält θ nicht).
- Sie wissen nicht, ob ihr Intervall den wahren Parameter enthält oder nicht.
- Sie wissen, dass es eine Eigenschaft des Verfahrens ist ein Konfidenzintervall zu liefern, dass in $(1-\alpha)\%$ der Fälle den wahren Parameter enthält.
- Ist das Intervall breit, ist die Schätzung ungenau, ist das Intervall schmal, ist die Schätzung relativ genau.
- Im allgemeinen gilt, desto größer der Stichprobenumfang desto schmaler das Konfidenzintervall und desto niedriger die Irrtumswahrscheinlichkeit α desto breiter das Intervall.

Der Standardfehler (vgl. LMLG S. 130)

Wir haben jetzt gesehen, wie das Konfidenzintervall interpretiert wird. Wir wollen uns noch anschauen, wie man es herleiten kann. Dazu brauchen wir den Standardfehler.

Definition: Standardfehler

Der Standardfehler, oft mit der englischen Abkürzung S. E. (für *standard error*) bezeichnet, ist die Standartabweichung der Schätz- bzw. Teststatistik. Er kann aus den Stichprobendaten geschätzt werden. In manchen Situationen ist er bekannt und muss nicht geschätzt werden.

Der Standardfehler hängt u. a. von der Stichprobengröße ab:
Große Stichproben produzieren unter sonst gleichen Bedingungen geringe, kleine Stichproben produzieren hingegen große Standardfehler.

Konfidenzintervalls für Anteilswerte und Tests

Oft sind Schätzer (approximativ) normalverteilt, d. h.

$$\hat{\theta} \sim N\big(\theta, \hat{\sigma}_{\hat{\theta}}\big)$$

oder

$$\hat{\theta} \sim N\big(\theta, \sigma_{\hat{\theta}}\big),$$

wenn der S. E. $\sigma_{\hat{\theta}}$ bekannt ist. Wenn die Verteilung des Schätzers zur Familie der Normalverteilungen gehört, sind Konfidenzintervalle nach dem folgenden Muster gebaut:

1) Die untere Grenze G_u wird berechnet als $\hat{\theta} - z_{1-\frac{\alpha}{2}} \cdot S.E.$
2) Die untere Grenze G_o wird berechnet als $\hat{\theta} + z_{1-\frac{\alpha}{2}} \cdot S.E.$

dabei ist $z_{1-\frac{\alpha}{2}}$ das $(1 - \frac{\alpha}{2})$-Quantil der Standardnormalverteilung und S. E. ist der bekannte oder geschätzte Standardfehler des Schätzers $\hat{\theta}$.

Herleitung des Konfidenzintervall I

- Der Schätzer $\hat{p}$ ist asymptotisch normalverteilt.
- Asymptotisch bedeutet, dass man bei großen Stichprobenumfängen von einer Normalverteilung ausgehen darf.
- Daher steht in der FS S. 52 die Voraussetzung $n \cdot p(1-p) \geq 9$.
- Unter dieser Voraussetzung gilt:

$$\hat{p} \sim N(p, \hat{\sigma}_{\hat{p}}),$$

mit

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Herleitung des Konfidenzintervall II

Der Schätzer $\hat{p}$ ist asymptotisch normalverteilt

$$\hat{p} \overset{a}{\sim} N\left(p, \hat{\sigma}_{\hat{p}}\right).$$

Gemäß dem obigen Konstruktionsprinzip ergibt sich dann genau das Konfidenzintervall (1), also

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \hat{p} + z_{1-\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}\right)\right].$$

Herleitung des Konfidenzintervall III

Um das Konstruktionsprinzip anzuwenden, braucht man nur einzusetzen.

1) Die untere Grenze G_u wird berechnet als $\hat{\theta} - z_{1-\frac{\alpha}{2}} \cdot S.E.$
2) Die untere Grenze G_o wird berechnet als $\hat{\theta} + z_{1-\frac{\alpha}{2}} \cdot S.E.$

$$G_u = \hat{p} - z_{1-\frac{\alpha}{2}} \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

und

$$G_o = \hat{p} + z_{1-\frac{\alpha}{2}} \left(\sqrt{\frac{\hat{p}(1-\hat{p}}{n}} \right).$$

Ausblick: Übung

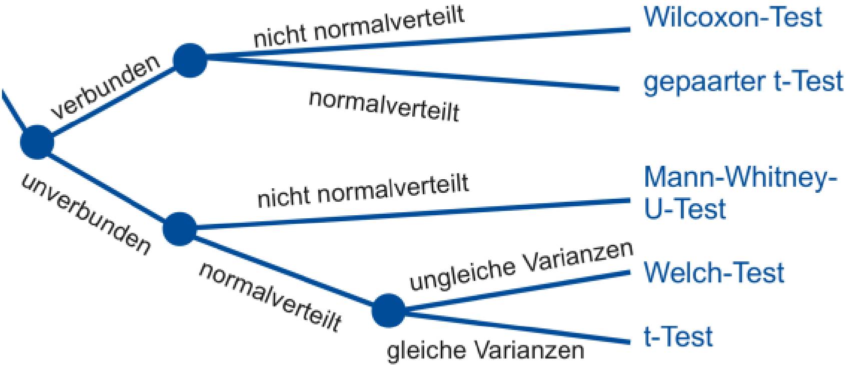


Abbildung: Quelle: `http://www.statistik-und-beratung.de/2013/07/statistischer-vergleich-von-zwei-gruppen/`

Ausblick: Übung

1) Wir schauen uns die Teststärke und das Konfidenzintervall an, um die Frage ob Mädchen besser lesen als Jungs, klären zu können.
2) Der Begriff nicht-parametrischer Test wird erklärt. Der χ^2-Test ist ein nicht-parametrischer Test. Zwei weitere wichtige Tests sind der U-Test und der Wilcoxon-Test. Diese werden vorgestellt. Bei den nicht-parametrischen Tests ist es nicht möglich den Test durch die Berechnung eines Konfidenzintervalls zu umgehen.
3) Wir lernen den Q-Q-Plot kennen, als Methode um zu Entscheiden, ob es sinnvoll ist von einer Normalverteilung auszugehen, oder nicht.
4) Wir sprechen über eine zweite Interpretation von Konfidenzintervallen, die in manchen Fällen möglich ist.