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. Ziel 3
4. Ziel 4
5. Ziel 5
6. Ziel 6

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:

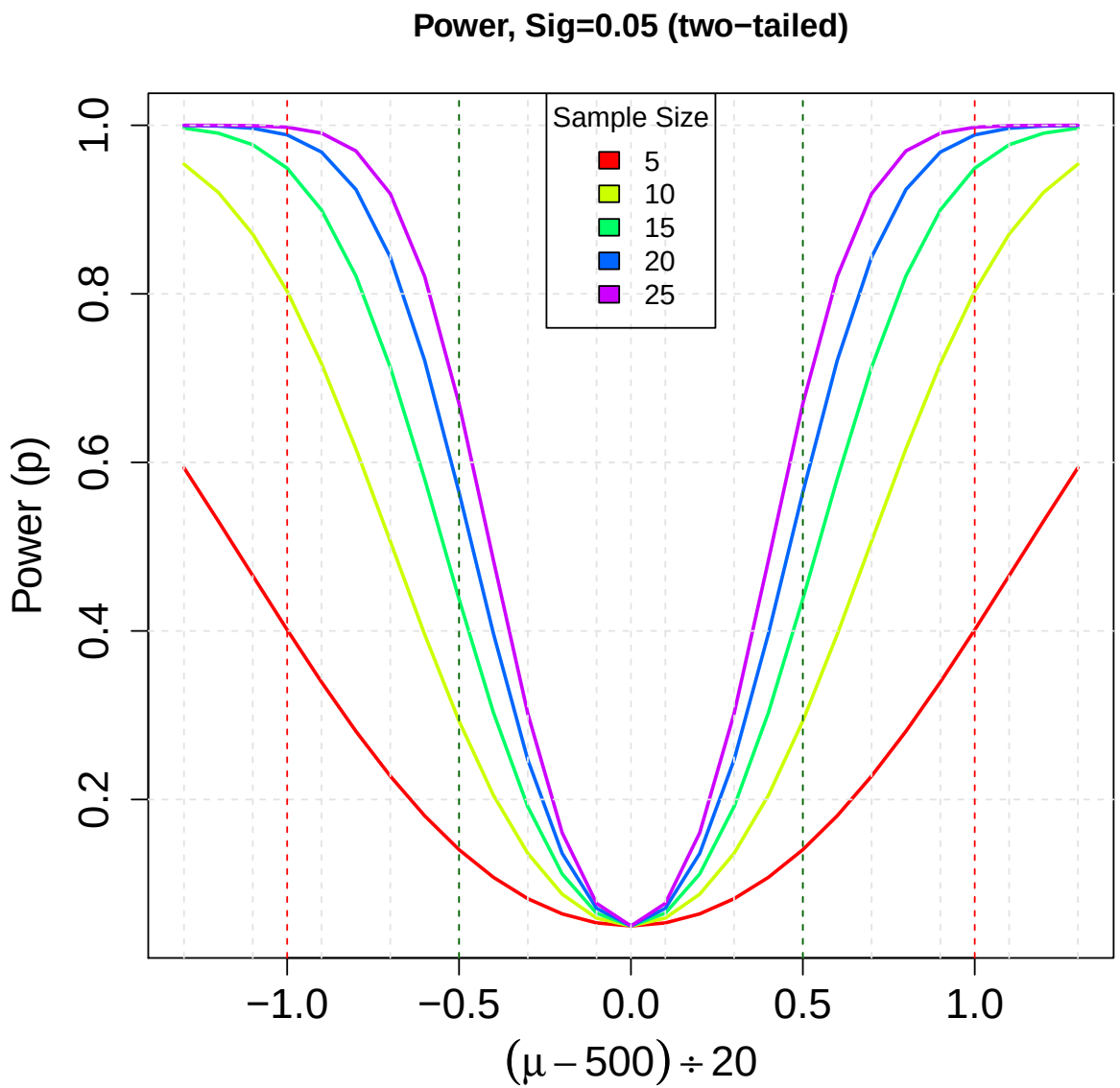
$$\delta := \frac{\mu - 500}{20}$$

- 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?

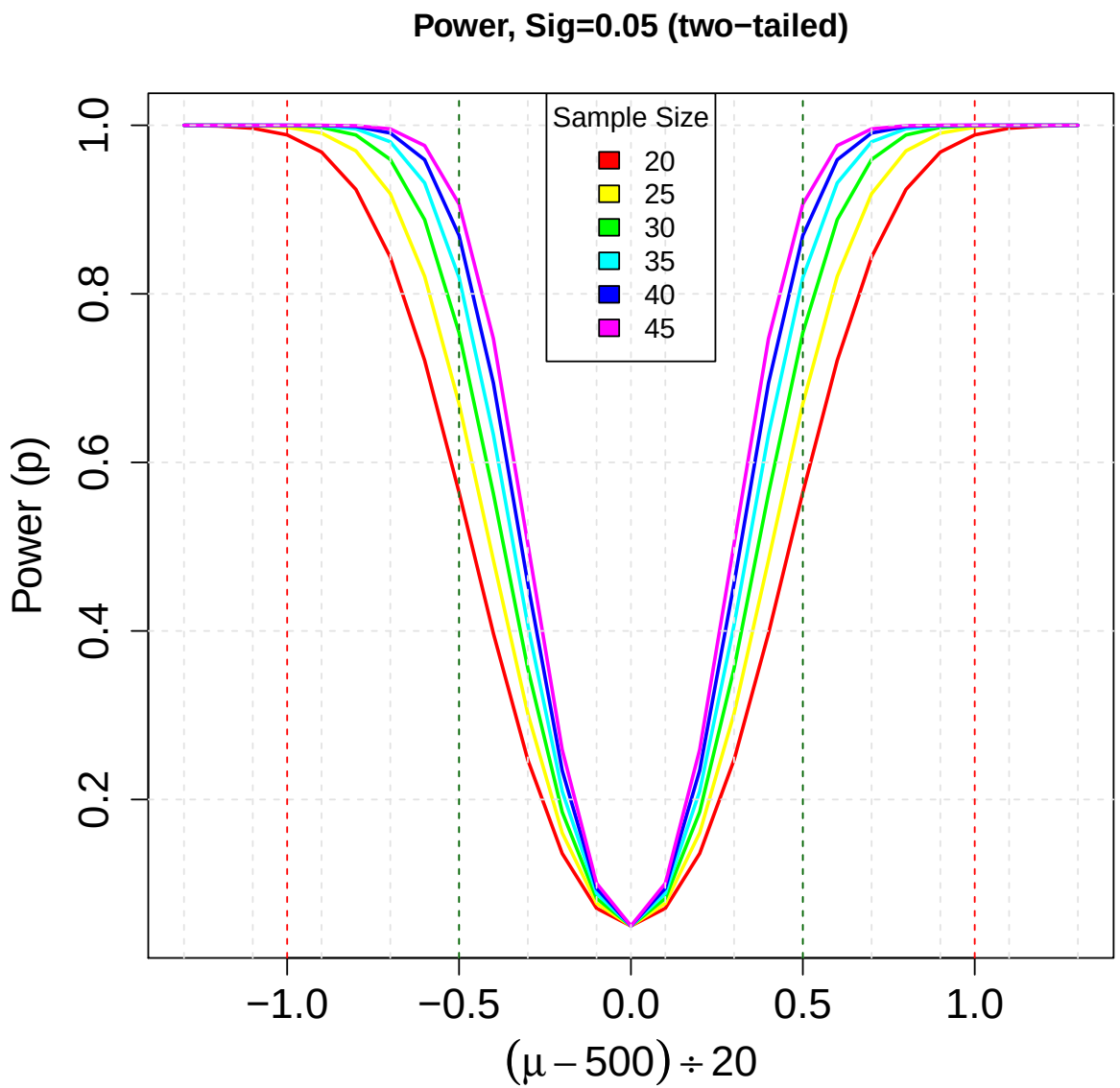
Was ist ein relevanter Unterschied?

- 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

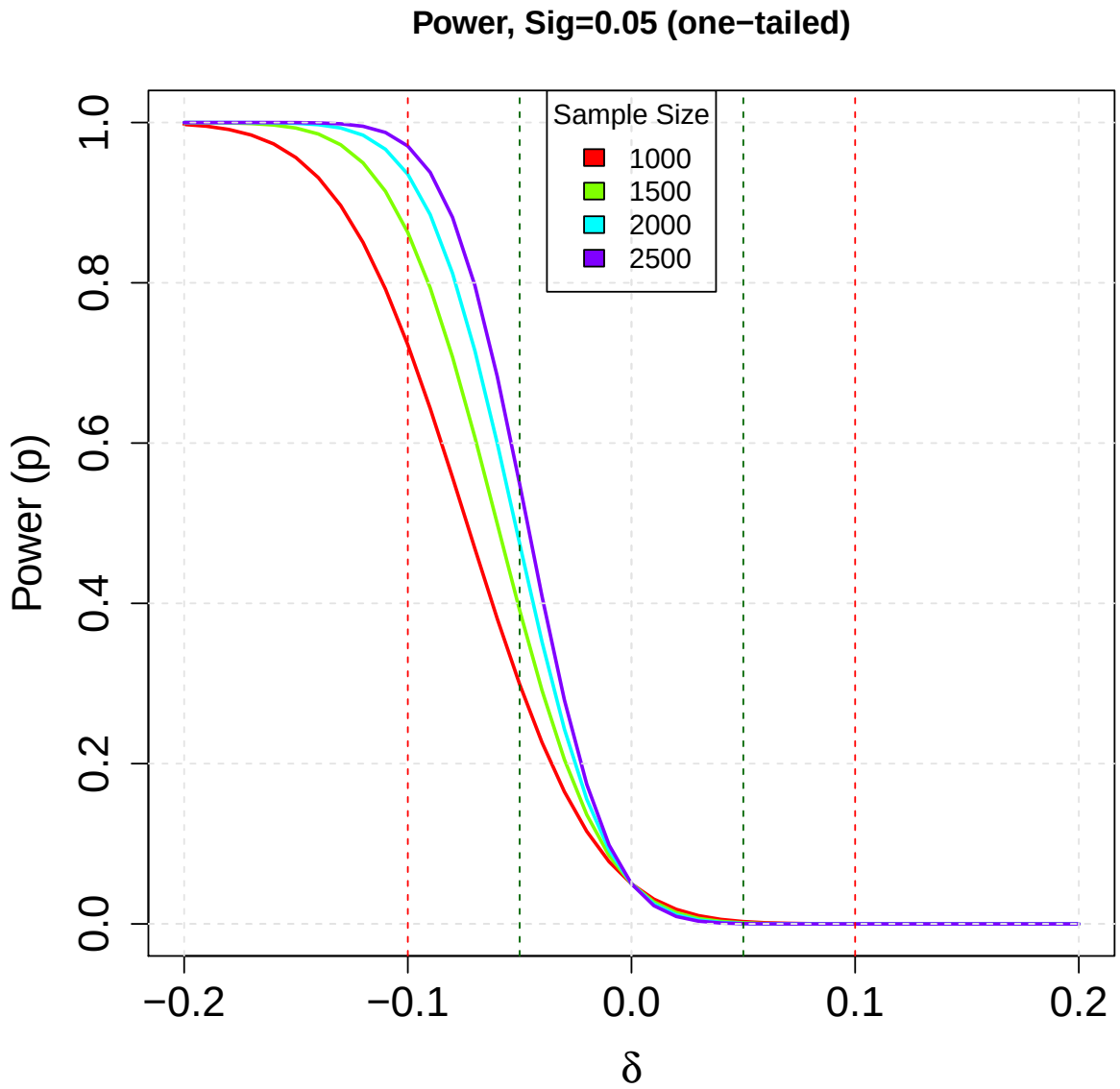
1) der Nullhypothese eine eindeutige Alternativhypothese gegenüber stellen.
2) mit Blick auf diese Alternativhypothese (was ist relevant?) den Stichprobenumfang so anpassen,
3) dass eine zufriedenstellende Teststärke erzielt wird.

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

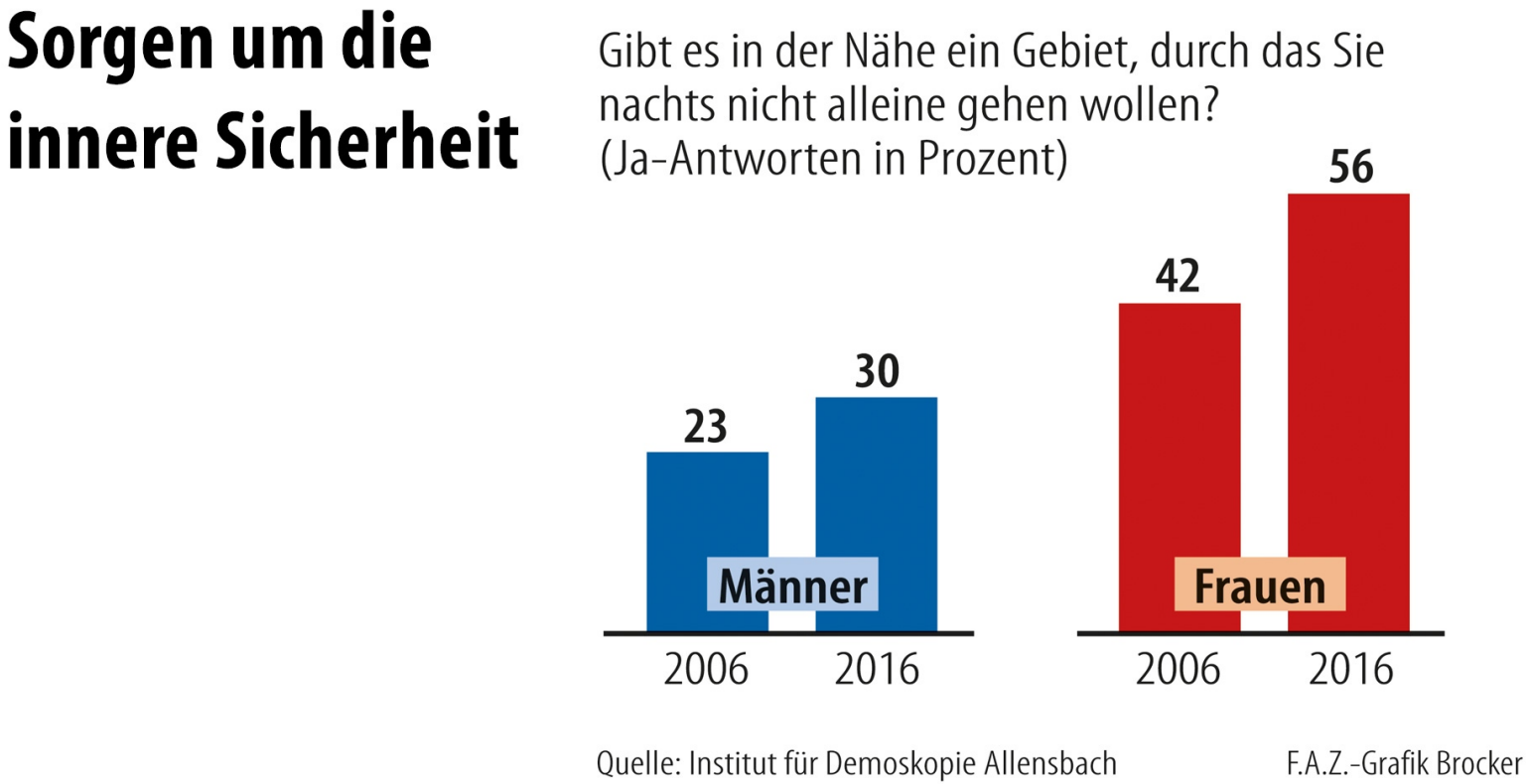
Statistisches Testen

- 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.

Übung: Lesen Mädchen besser als Jungs?



Die diffusen Ängste der Deutschen (17.2.16)



Untersuchungsdaten: Allensbach-Umfrage

- Befragter Personenkreis: Deutsche Wohnbevölkerung ab 16 Jahre in der Bundesrepublik Deutschland
- Anzahl der Befragten: 1521 Befragungszeitraum: 01. Februar bis 11. Februar 2016
- Methode: Repräsentative Quotenauswahl Art der Interviews: Mündlich-persönliche Interviews

Punktschätzer für den Anteilswert

- Bei der Punktschätzung geht es darum, einen Parameter in der Grundgesamtheit zu schätzen. Mit Bezug auf obige Grafik können wir uns z. B. für den Anteil der Männer interessieren, die 2016 ein Gebiet in ihrer Nähe kennen, durch das Sie nachts nicht alleine gehen wollen.
- Der entspreche Anteil in der Stichprobe ist 30%.
- Es ist plausibel, den in der Stichprobe beobachteten Anteilswert als Schätzer für die Grundgesamtheit zu verwenden.
- Wie schätzen also, dass in Deutschland 30% der deutschen ab 16 Jahren diese Frage mit ja beantworten. Aber wie sicher ist die Schätzung?

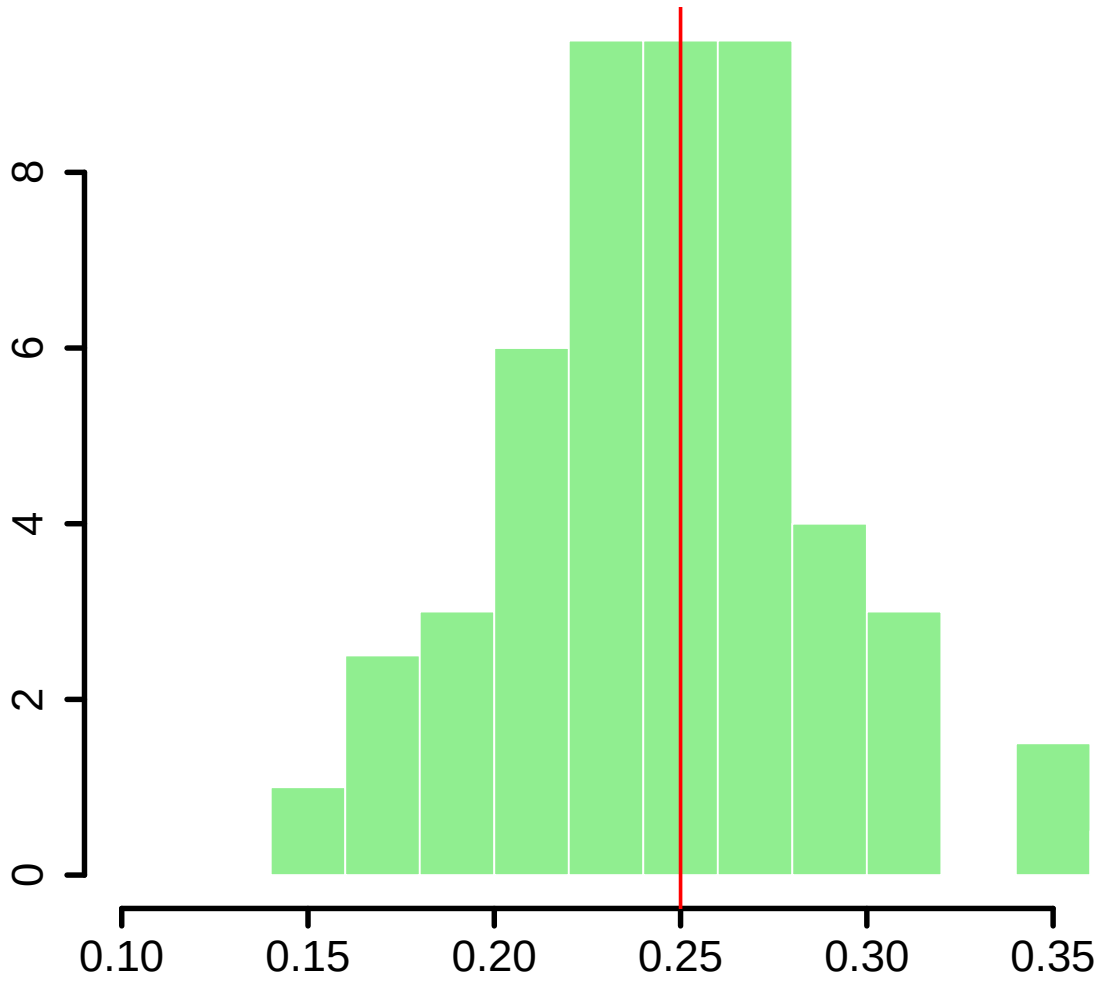
Ein Konfidenzintervall für den Anteilswert: Menge aller beibehaltenen Nullhypothesen

- Auf S. 52 ihrer Formelsammlung ist das Konfidenzintervall für den Anteilswert definiert.
- Sie können dieses Konfidenzintervall interpretieren, als ein Intervall, dass alle Nullhypothesen enthält, die der approximative Binomialtest beibehalten würde.
- Wenn man eine Stichprobe von 100 Männern hat, von denen 30 die Frage bejahen, ergibt sich $[0.22, 0.40]$
- Wenn man eine Stichprobe von 1000 Männern hat, von denen 300 die Frage bejahen, ergibt sich $[0.27, 0.33]$

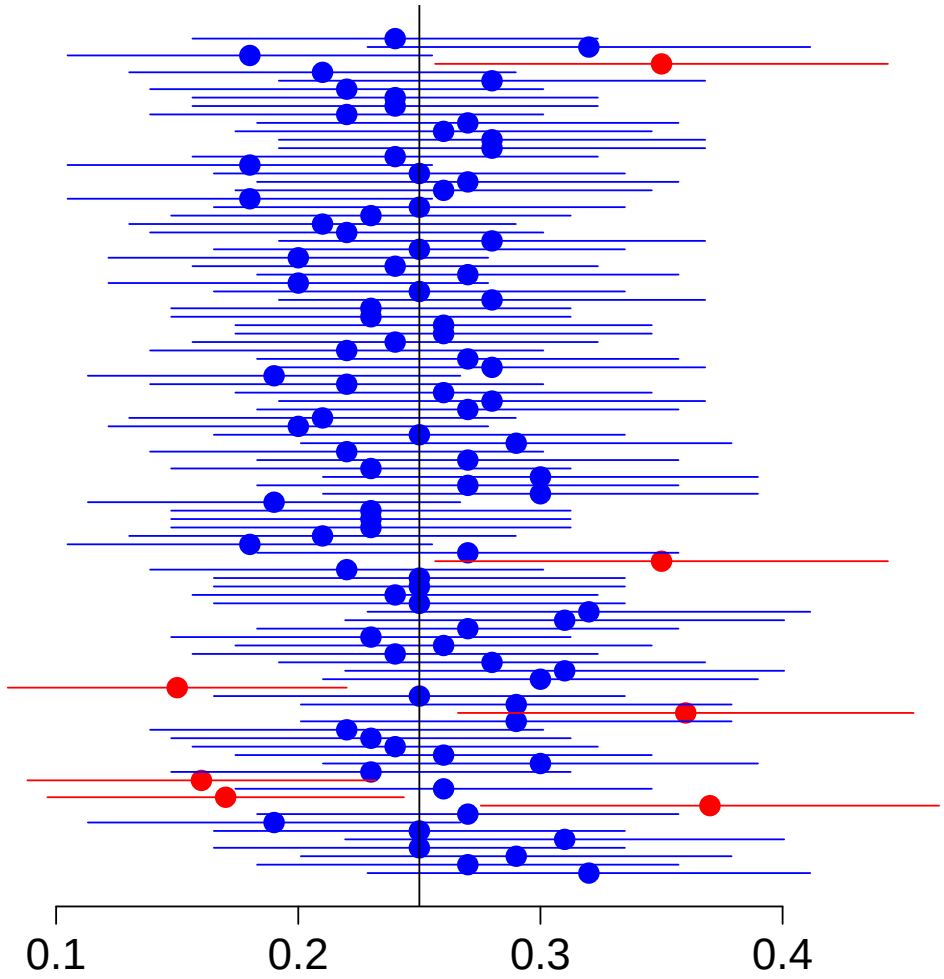
Eine andere Interpretation des Konfidenzintervall

- Im folgenden gehen wir davon aus, dass der wahre (und unbekannte) Anteil der männlichen Personen ab 16 Jahren, die der Frage zustimmen, in der Grundgesamtheit bei 25% liegt.
- Stellen Sie sich vor wir erheben 100 Stichproben mit je 100 Personen aus der Grundgesamtheit.
- Wir gehen dabei davon aus, dass die Grundgesamtheit so groß ist, dass keine Person doppelt vorkommt.
- Sie können sich vorstellen, dass *Personen* losgehen und jede Person eine Stichprobe mit je 100 Personen zieht.
- Die Folgenden Grafiken zeigen
  1) das Histogramm der 100 berechneten Anteilswerte und
  2) die entsprechenden Konfidenzintervalle

Beobachtete Anteilswerte: 100 Befragungen



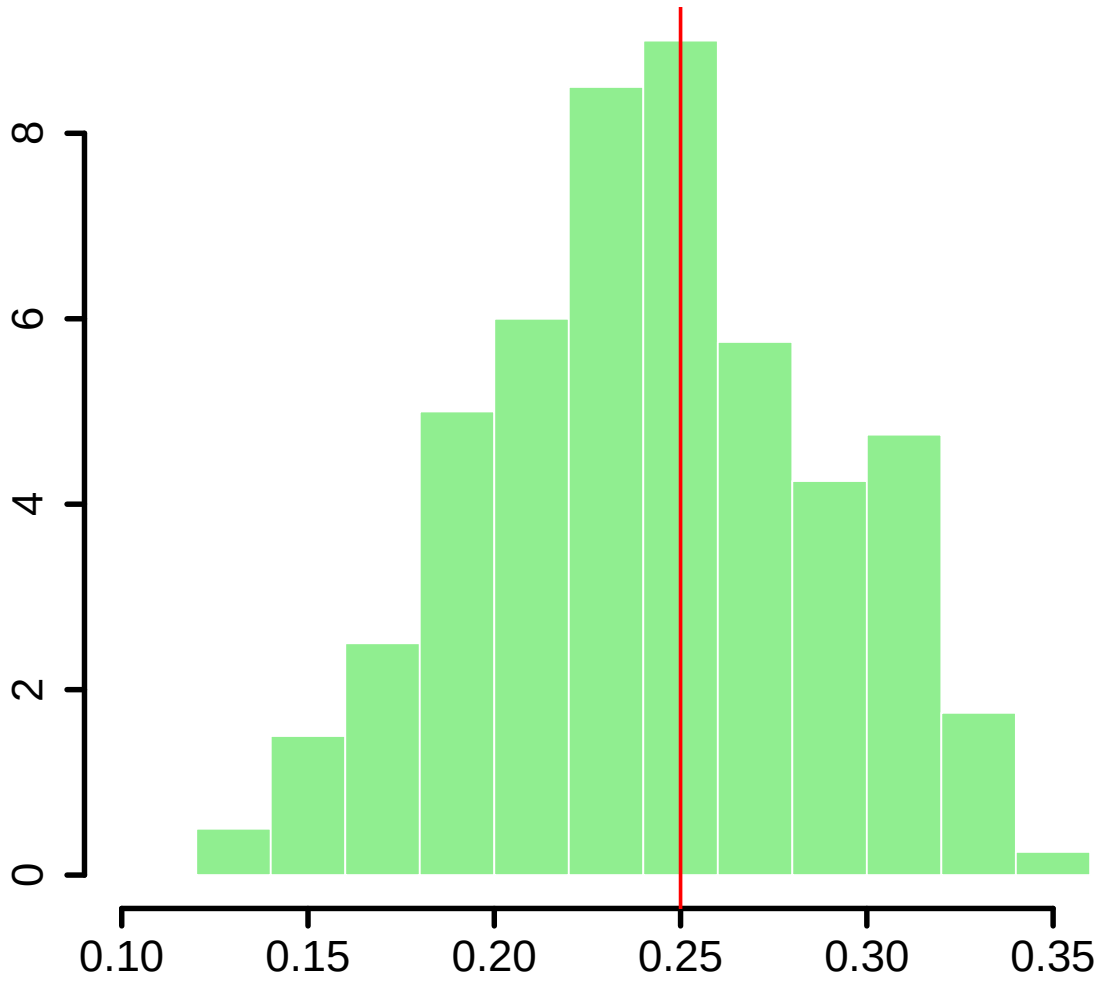
Konfidenzintervalle zu den Anteilswerten: 100 Befragungen



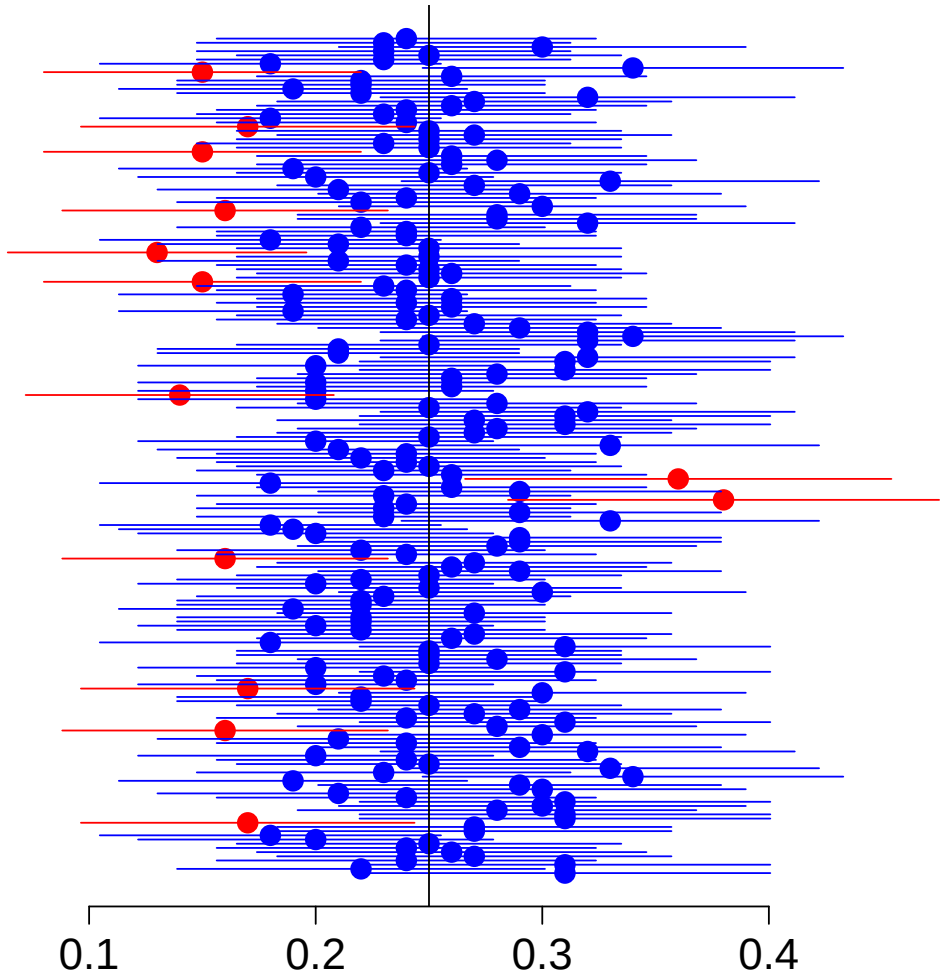
Betrachtung der Ergebnisse

- Von den 100 Personen, die losgegangen sind, um jeweils eine Stichprobe von 100 Personen zu ziehen, haben fünf eine Stichprobe gezogen, deren Konfidenzintervall den wahren Wert nicht enthält.
- Die anderen 95 haben eine Stichprobe gezogen, deren Konfidenzintervall, den wahren Wert enthält.
- Wir wiederholen das Experiment. Diesmal gehen 500 Personen los um jeweils 100 Personen zu befragen.
- Was erwarten Sie? Wie viele Konfidenzintervalle werden den wahren Anteilswert nicht enthalten?

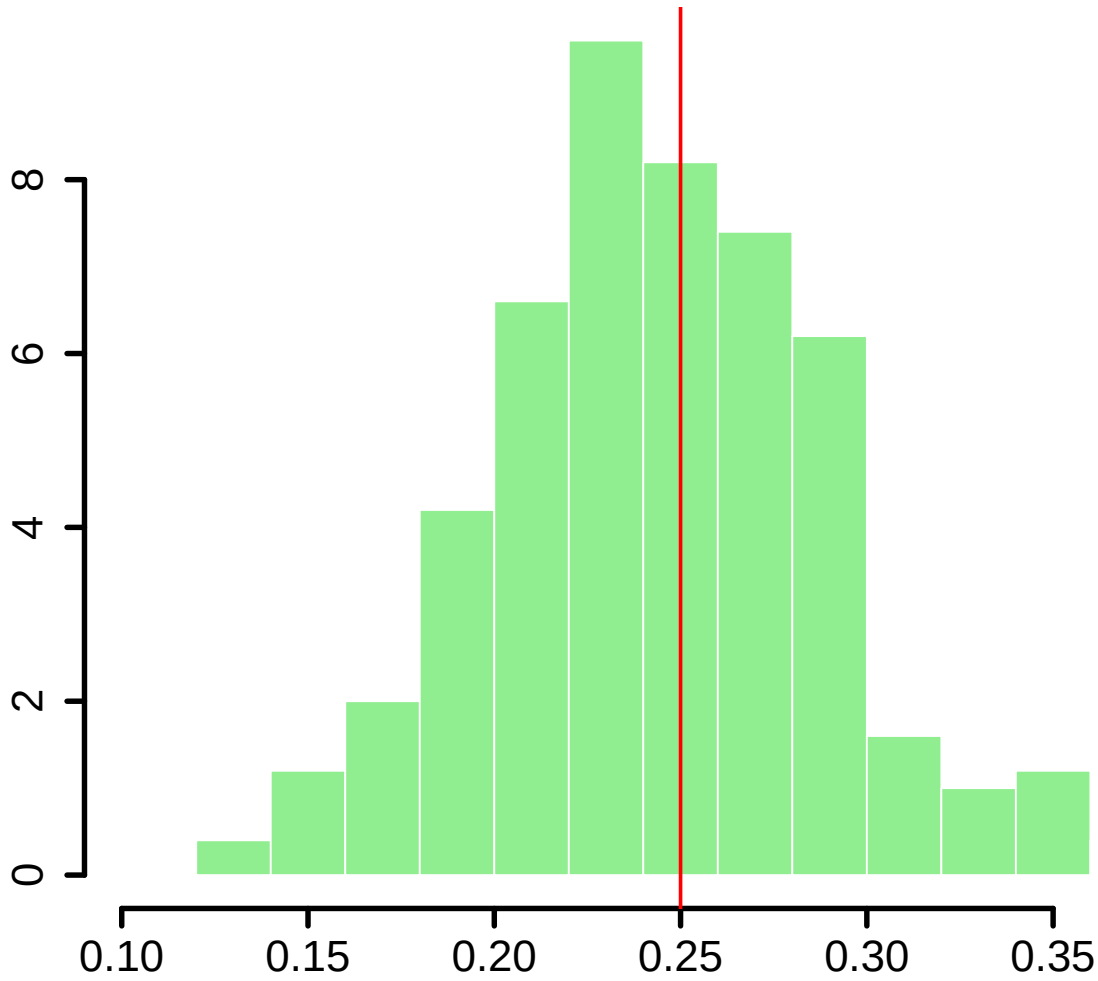
Beobachtete Anteilswerte: 200 Befragungen



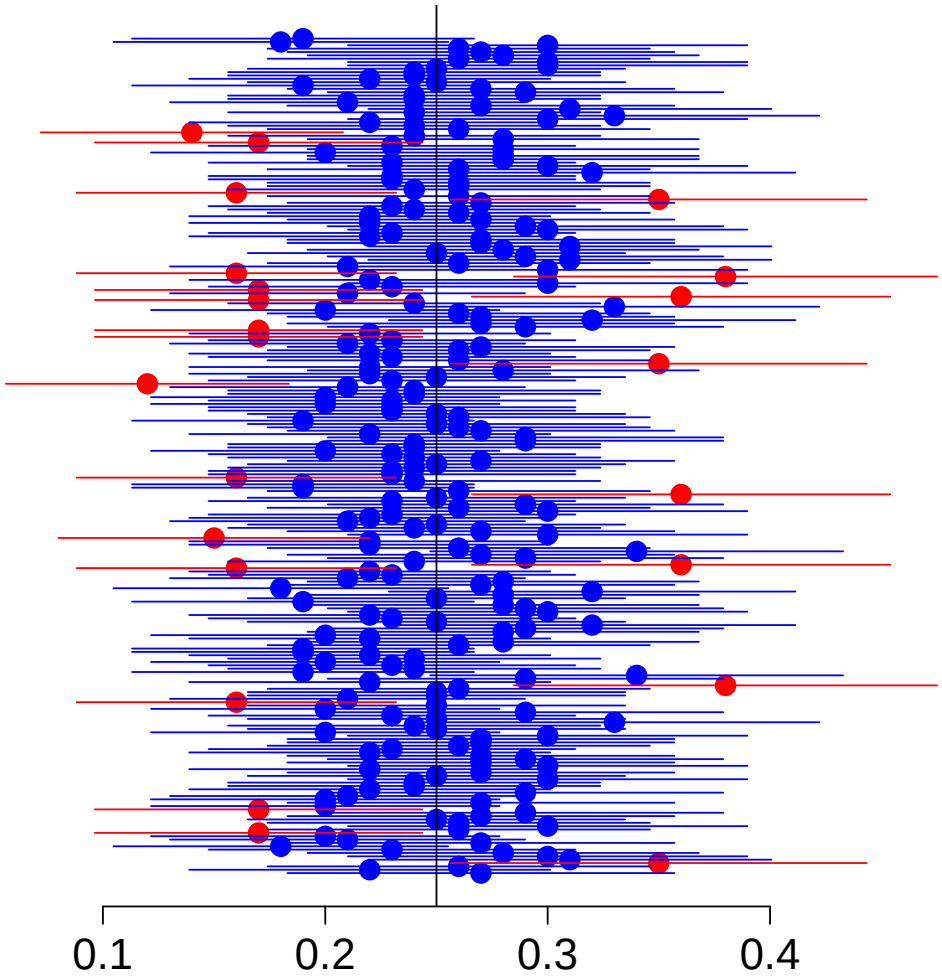
Konfidenzintervalle zu den Anteilswerten: 200 Befragungen



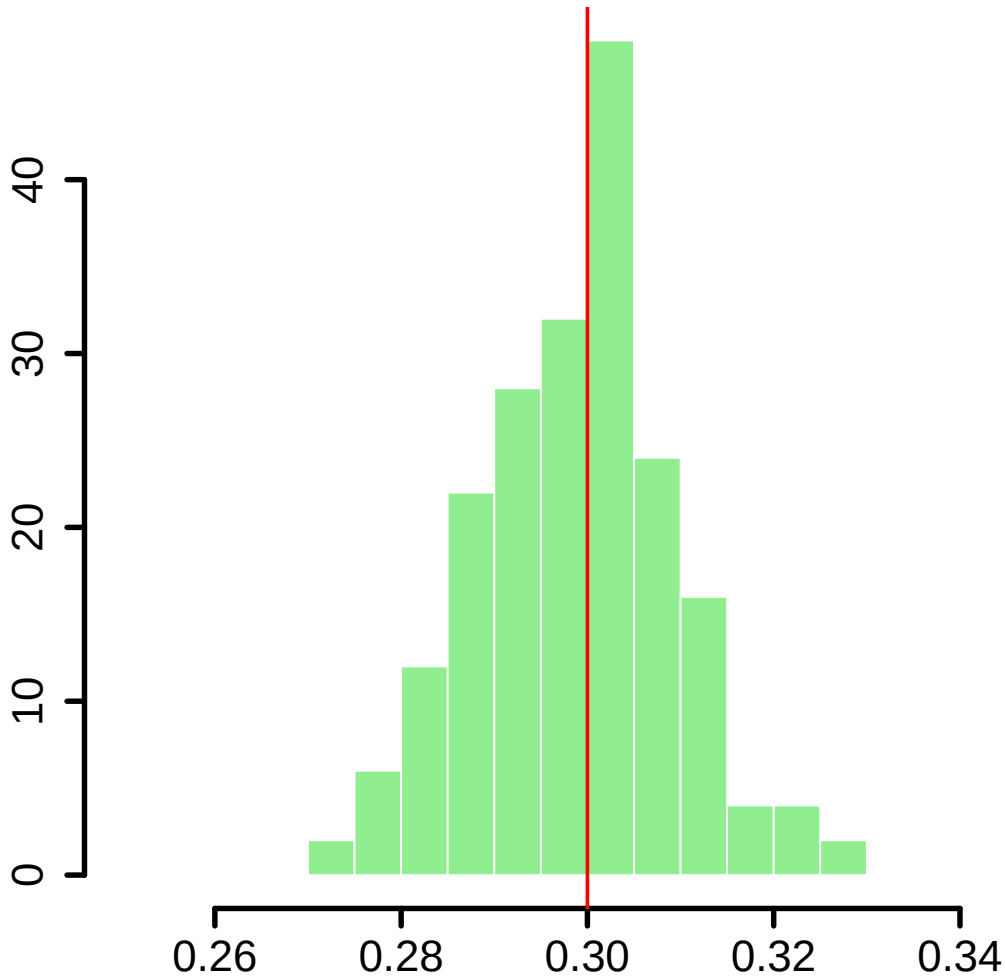
Beobachtete Anteilswerte: 250 Befragungen



Konfidenzintervalle zu den Anteilswerten: 250 Befragungen



Allensbach-Umfrage : 100 Befragungen



Allensbach-Umfrage: 100 Befragungen