

Hypothesentests und Konfidenzintervalle

Von der Evidenz einer Stichprobe zu vorsichtigen Aussagen über die Grundgesamtheit

Ratiomera Statistics

Inhaltsverzeichnis

Zusammenfassung der Kursseite	2
Der grundlegende Denkweg	2
Wähle das Verfahren anhand des Designs	3
Checkliste für die Interpretation	3
Erweiterte Referenz	4
Zweck und Grundlagen	4
Zentrale Ideen	4
Formelleitfaden	5
Die erklärende Abbildung lesen	8
Checkliste zur Interpretation	8
Verbindung zu anderen Themen	9

Zusammenfassung der Kursseite

Inferenz verwendet eine Stichprobe, um etwas über eine Population zu lernen und die Unsicherheit sichtbar zu halten. Beginne mit der Forschungsfrage und dem Design, benenne die Populationsgrösse und wähle erst danach ein Konfidenzintervall oder einen Hypothesentest, der zur Datenstruktur passt.

Der grundlegende Denkweg

Tabelle 1: Verlässliche Inferenz verbindet Forschungsfrage, Design, Berechnung und Interpretation.

Schritt	Leitfrage	Hauptidee
1. Definieren	Welche Populationsgrösse wird untersucht?	Schreibe Parameter sowie Null- und Alternativhypothese vor der Berechnung auf
2. Zuordnen	Sind die Beobachtungen unabhängig, gepaart oder kategoriale Häufigkeiten?	Das Design bestimmt Standardfehler und Referenzverteilung
3. Berechnen	Wie weit liegt die Schätzung in Standardfehlereinheiten vom Nullwert entfernt?	Eine Teststatistik standardisiert die beobachtete Abweichung
4. Beurteilen	Wie ungewöhnlich wäre diese Statistik, wenn das Nullmodell gilt?	Der p-Wert ist eine bedingte Wahrscheinlichkeit unter dem Nullmodell
5. Schlussfolgern	Was stützt die Evidenz beim gewählten Niveau?	Nenne Populationsaussage, Unsicherheit, Richtung und Grenzen der Studie

Für einen Stichprobenmittelwert mit geschätzter Populationsstandardabweichung lautet die Ein-Stichproben-Statistik

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

mit $n - 1$ Freiheitsgraden. Ein Konfidenzintervall verwendet dieselbe Schätzung und denselben Standardfehler:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}.$$

Das Intervall gibt einen Bereich von Populationsmittelwerten an, die beim gewählten Konfidenzniveau mit den Daten und dem Modell vereinbar sind. Seine wiederholungsbezogene Bedeutung betrifft die Methode: Über viele vergleichbare Stichproben würden der angegebene Anteil der Intervalle den festen Populationsparameter enthalten.

Wähle das Verfahren anhand des Designs

Tabelle 2: Ähnlich aussehende Zahlen können unterschiedliche Verfahren erfordern, wenn ihre Abhängigkeitsstruktur verschieden ist.

Datenstruktur	Frage	Hier verwendetes Verfahren
Eine quantitative Stichprobe	Unterscheidet sich ein Populationsmittelwert von einem Referenzwert?	Ein-Stichproben- <i>t</i> -Verfahren
Zwei unabhängige quantitative Gruppen	Unterscheiden sich zwei Populationsmittelwerte?	<i>t</i> -Verfahren für unabhängige Stichproben
Zwei Messungen an denselben Fällen	Unterscheidet sich der Populationsmittelwert der fallweisen Differenzen von null?	Gepaarte <i>t</i> -Verfahren
Zwei kategoriale Variablen	Sind die Variablen in der Population unabhängig?	Chi-Quadrat-Unabhängigkeitstest

Checkliste für die Interpretation

- Ein p-Wert ist die Wahrscheinlichkeit von Daten, die mindestens so wenig mit dem Nullmodell vereinbar sind, unter der Annahme dieses Modells und seiner Bedingungen. Er ist nicht die Wahrscheinlichkeit, dass die Nullhypothese wahr ist.
- Eine Verwerfung ist kein Beweis. Ein Nichtverwerfen ist kein Beweis für Gleichheit.
- Statistische Signifikanz sagt nichts über Effektgrösse, praktische Bedeutung, Messqualität oder Kausalität aus.
- Ein Fehler 1. Art bedeutet, eine wahre Nullhypothese zu verwerfen. Ein Fehler 2. Art bedeutet, eine falsche Nullhypothese nicht zu verwerfen. Power ist die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, wenn die festgelegte Alternative wahr ist.
- Eine grössere Stichprobe kann den Standardfehler verkleinern und die Power erhöhen. Sie behebt aber keine verzerrte Stichprobenziehung, schlechte Messung, ignorierte Abhängigkeit oder ein schwaches Design.

Thema 3 liefert die Inferenzsprache für alle späteren Themen. Korrelation, Regression, partielle Korrelation, multiple Regression und ANOVA ändern Parameter und Modell, behalten aber dieselbe Disziplin: Frage definieren, Design beachten, Unsicherheit quantifizieren und die Schlussfolgerung begrenzen.

Erweiterte Referenz

Zweck und Grundlagen

Die statistische Inferenz nutzt Informationen aus einer Stichprobe, um etwas über eine Grundgesamtheit zu erfahren und dabei die Unsicherheit der Stichprobenziehung zu berücksichtigen. Eine **Grundgesamtheit** ist die vollständige Gruppe von Fällen, auf die sich die Forschungsfrage bezieht. Eine **Stichprobe** ist die beobachtete Teilmenge. Ein **Parameter** ist ein numerisches Merkmal der Grundgesamtheit, zum Beispiel der Populationsmittelwert μ . Ein **Stichprobenkennwert** ist die entsprechende, aus der Stichprobe berechnete Grösse, beispielsweise der Stichprobenmittelwert $\bar{x}$. Der Stichprobenkennwert ist nach der Datenerhebung bekannt. Der Parameter bleibt normalerweise unbekannt.

Würde aus derselben Grundgesamtheit eine neue Zufallsstichprobe gezogen, wäre ihr Stichprobenkennwert gewöhnlich ein anderer. Die hypothetische Verteilung dieses Kennwerts über wiederholte Stichproben ist seine **Stichprobenverteilung**. Ihre Standardabweichung heisst **Standardfehler**. Der Standardfehler misst, wie stark eine Schätzung von Stichprobe zu Stichprobe variiert. Er misst nicht die Streuung einzelner Beobachtungen. Dafür ist die gewöhnliche Standardabweichung zuständig.

Inferenz beruht auf mehr als einer Formel. Die Stichprobe muss glaubwürdig mit der Grundgesamtheit verbunden sein, die Beobachtungen müssen zu den Annahmen des Verfahrens über Abhängigkeiten passen und die Messung muss die beabsichtigte Variable angemessen darstellen. Ein kleiner Standardfehler kann weder Auswahlverzerrung noch unzureichende Messung oder ein ungeeignetes Design korrigieren. Beginne mit Forschungsfrage und Studiendesign, untersuche danach die deskriptiven Kennwerte und wähle erst dann ein Inferenzverfahren.

Element	Sprache der Stichprobe	Sprache der Grundgesamtheit
Lage	Stichprobenmittelwert $\bar{x}$	Populationsmittelwert μ
Anteil	Stichprobenanteil $\hat{p}$	Populationsanteil p
Streuung der Werte	Stichprobenstandardabweichung s	Populationsstandardabweichung σ
Unsicherheit einer Schätzung	Geschätzter Standardfehler	Standardabweichung der Stichprobenverteilung

Zentrale Ideen

Ein Konfidenzintervall gibt einen Bereich von Parameterwerten an, die unter dem Modell mit der Schätzung und ihrer Stichprobenunsicherheit vereinbar sind. Das Konfidenzniveau beschreibt die langfristige Güte des Verfahrens. Würden dasselbe Stichproben- und Intervallverfahren oft wiederholt, enthielte ein festgelegter Anteil der entstehenden Intervalle den festen Populationsparameter. Nachdem ein einzelnes Intervall berechnet wurde, bewegt sich der Parameter nicht zwischen verschiedenen Werten. Das Intervall ist das Ergebnis, das durch die Stichprobenziehung variiert hat.

Ein Hypothesentest beginnt mit einer **Nullhypothese** H_0 , einer genauen Referenzaussage über einen Populationsparameter. Eine **Alternativhypothese** H_1 beschreibt die Richtung oder den Unterschied von inhaltlichem Interesse. Eine Teststatistik misst, wie weit die beobachtete Schätzung in Standardfehlereinheiten vom Nullwert entfernt liegt. Der **p-Wert** ist die Wahrscheinlichkeit, unter Annahme der Nullhypothese und aller Modellbedingungen eine Teststatistik zu erhalten, die mindestens so wenig mit der Nullhypothese vereinbar ist wie die beobachtete. Er ist nicht die Wahrscheinlichkeit dafür, dass die Nullhypothese wahr ist.

Das Signifikanzniveau α ist eine Entscheidungsschwelle, die vor der Betrachtung des Ergebnisses festgelegt wird. Ist der p-Wert höchstens so gross wie α , wird das Ergebnis als statistisch signifikant bezeichnet und H_0 verworfen. Ist der p-Wert grösser als α , wird H_0 nicht verworfen. Das Nichtverwerfen ist kein Beweis dafür, dass es keinen Effekt gibt. Die Daten können ungenau sein, der wahre Effekt kann klein sein oder das Design kann eine begrenzte Teststärke besitzen.

Wirklichkeit und Entscheidung	H_0 nicht verwerfen	H_0 verwerfen
H_0 ist wahr	Richtige Entscheidung, die Nullhypothese beizubehalten	Fehler 1. Art, dessen Wahrscheinlichkeit durch α kontrolliert wird
H_0 ist falsch	Fehler 2. Art, bezeichnet mit β	Richtige Entdeckung, deren Wahrscheinlichkeit Teststärke $1 - \beta$ heisst

Die Teststärke ist die Wahrscheinlichkeit, dass ein Test H_0 verwirft, wenn eine festgelegte Alternative wahr ist. Sie steigt, wenn der wahre Effekt grösser ist, die Werte weniger streuen, die Stichprobe grösser ist oder die Signifikanzregel weniger streng gewählt wird. Diese Einflüsse sind mit Abwägungen verbunden. Eine Planung benötigt deshalb eine inhaltlich bedeutsame Effektgrösse und ein vertretbares Design. Sie darf nicht erst nach der Datenerhebung auf die Suche nach Signifikanz ausgerichtet werden.

Die Wahl des Verfahrens folgt der Struktur der Forschungsfrage. Ein Verfahren für einen Mittelwert in einer Stichprobe vergleicht eine Gruppe mit einem Referenzwert. Ein Verfahren für unabhängige Gruppen vergleicht getrennte Gruppen. Ein Verfahren für verbundene Beobachtungen analysiert zusammengehörige Messungen, etwa dieselben Teilnehmenden vor und nach einer Intervention, indem es die Differenzen innerhalb der Paare verwendet. Ein Chi-Quadrat-Verfahren für eine Kontingenztafel vergleicht beobachtete kategoriale Häufigkeiten mit den unter dem Nullmodell erwarteten Häufigkeiten. In jedem Fall müssen die Analyseeinheit und die Abhängigkeitsstruktur genannt werden.

Formelleitfaden

Bei einer unabhängigen Zufallsstichprobe ist der geschätzte Standardfehler eines Stichprobenmittelwerts gleich der Stichprobenstandardabweichung geteilt durch die Quadratwurzel der Stichprobengrösse:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Die Quadratwurzel erklärt, weshalb die Unsicherheit langsamer abnimmt, als die Stichprobengrösse zunimmt. Wird n mit vier multipliziert, halbiert sich dieser Standardfehler, sofern die Streuung gleich bleibt.

Ein Konfidenzintervall verbindet eine Schätzung $\hat{\theta}$ mit ihrem Standardfehler und einem zum Konfidenzniveau passenden kritischen Wert c :

$$\hat{\theta} \pm c \cdot SE(\hat{\theta})$$

Ist die Populationsstandardabweichung σ bekannt und gilt das angegebene Normalmodell, wird die Standardnormalverteilung als Referenz verwendet:

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Die Alternativhypothese bestimmt die Referenzfläche. Eine zweiseitige Alternative verwendet beide Verteilungsschwänze jenseits von $|z|$ oder $|t|$. Eine einseitige Alternative verwendet den im Voraus festgelegten gerichteten Verteilungsschwanz. Wird die Richtung erst nach der Betrachtung des Ergebnisses gewählt, handelt es sich nicht um einen im Voraus festgelegten einseitigen Test.

Für einen Mittelwert in einer Stichprobe mit geschätzter Populationsstandardabweichung verwendet das Intervall einen kritischen Wert aus einer t -Verteilung mit $n - 1$ Freiheitsgraden:

$$\bar{x} \pm t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

Die zugehörige Teststatistik für eine Stichprobe vergleicht den beobachteten Stichprobenmittelwert mit dem Nullwert μ_0 :

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Der Zähler ist der beobachtete Unterschied zur Nullhypothese. Der Nenner übersetzt diesen Unterschied in Standardfehlereinheiten. Bei verbundenen Messungen berechnest du zuerst eine Differenz d_i für jedes Paar und wendest danach dieselbe Ein-Stichproben-Logik auf die mittlere Differenz $\bar{d}$ an:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$$

Damit bleibt die Paarung erhalten. Würden die Messungen als unverbunden behandelt, ginge die Information darüber verloren, welche zwei Beobachtungen zusammengehören.

Bei zwei unabhängigen Stichproben unter dem in diesem Kurs behandelten Modell gleicher Populationsvarianzen werden zuerst die beiden Stichprobenvarianzen gepoolt:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Danach berechnest du

$$SE(\bar{x}_1 - \bar{x}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}, \quad df = n_1 + n_2 - 2.$$

Das zugehörige zweiseitige Intervall ersetzt den Zähler durch

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Die Annahme gleicher Varianzen, der Name des Verfahrens und die Berechnung müssen zueinander passen. Verbundene Daten benötigen stattdessen das Verfahren für Differenzwerte.

Für eine Planungsfrage mit einer Stichprobe und bekanntem σ definierst du die standardisierte Populationsdifferenz

$$\delta = \frac{\mu - \mu_0}{\sigma}, \quad \text{Power} = 1 - \beta.$$

Im einseitigen z-Planungsmodell des bereitgestellten Materials lautet die für das Signifikanzniveau α und die angestrebte Teststärke $1 - \beta$ benötigte Stichprobengrösse

$$n = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\delta} \right)^2.$$

Runde das Ergebnis auf. Diese Formel gehört zum angegebenen Modell und ist keine universelle Regel für Stichprobengrössen. Die Teststärke nimmt mit der Stichprobengrösse und dem Betrag des Effekts zu. Sie nimmt ab, wenn ein strengeres Signifikanzniveau die Ablehnungsgrenze weiter in den Verteilungsschwanz verschiebt.

Bei zwei kategorialen Variablen lautet die unter Unabhängigkeit erwartete Häufigkeit in Zeile i und Spalte j

$$m_{ij} = \frac{n_{i.}n_{.j}}{n}.$$

Die Chi-Quadrat-Statistik und ihre Freiheitsgrade sind

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \quad df = (k - 1)(l - 1).$$

Bei einer Zwei-mal-zwei-Tabelle ist der Betrag des Phi-Koeffizienten

$$|\phi| = \sqrt{\frac{\chi^2}{n}}.$$

Die in dieser Lernsequenz verwendete Annäherung setzt eine einfache Zufallsstichprobe und erwartete Häufigkeiten über 5 in jeder Zelle voraus. Ein grosses χ^2 spricht gegen Unabhängigkeit und ist keine Evidenz für Unabhängigkeit.

Struktur der Frage	Verfahren im Kurs	Analysierte Grösse
Eine Stichprobe gegen einen Referenzwert	z- oder t-Test für eine Stichprobe	Ein Stichprobenmittelwert
Zwei getrennte Gruppen	Gepoolter t-Test für unabhängige Stichproben	Differenz zwischen Gruppenmittelwerten
Zwei verbundene Messungen	t-Test für verbundene Stichproben	Mittelwert der Differenzen innerhalb der Paare
Unabhängige Gruppen mit einer rangbasierten Frage	Wilcoxon-Rangsummentest	Relative Ränge zwischen den Gruppen
Verbundene Beobachtungen mit einer rangbasierten Frage	Wilcoxon-Vorzeichen-Rang-Verfahren	Vorzeichenränge der Paardifferenzen
Zwei kategoriale Variablen	Chi-Quadrat-Unabhängigkeitstest	Beobachtete gegen erwartete Zellenhäufigkeiten

Die erklärende Abbildung lesen

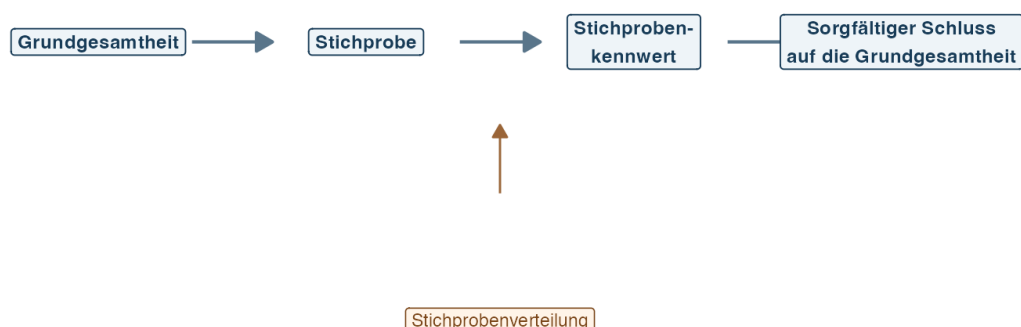


Abbildung 1: Ein horizontaler Ablauf führt von der Grundgesamtheit über die Stichprobe und den Stichprobenkennwert zu einer vorsichtigen Populationsaussage, wobei eine Stichprobenverteilung in den Kennwert einfließt.

Lies die Hauptlinie von links nach rechts. Die Grundgesamtheit ist das Ziel der Forschungsfrage. Die Stichprobe ist der Teil, der beobachtbar wird. Ein Stichprobenkennwert verdichtet die relevante Evidenz, beispielsweise als Stichprobenmittelwert, Differenz, Anteil oder Zusammenhang. Das letzte Feld ist bewusst als vorsichtige Schlussfolgerung bezeichnet, weil der Weg von der Stichprobe zur Grundgesamtheit nie automatisch ist.

Der Pfeil, der von der Stichprobenverteilung nach oben führt, ist die durch Wahrscheinlichkeit bereitgestellte Brücke. Er stellt dar, wie der Stichprobenkennwert unter festgelegten Annahmen über wiederholte Stichproben variieren würde. Ein Konfidenzintervall verwendet diese Streuung, um die Präzision zu zeigen. Ein Test vergleicht den beobachteten Kennwert mit dem unter H_0 erwarteten Stichprobenverhalten. Die Abbildung bedeutet nicht, dass eine grosse Stichprobe automatisch die Verallgemeinerbarkeit garantiert. Stichprobenverfahren, Messung, fehlende Daten, Abhängigkeiten und Studiendesign bestimmen weiterhin, welche Aussage über die Grundgesamtheit zulässig ist.

Die Trennung zwischen «Stichprobenkennwert» und «Schlussfolgerung über die Grundgesamtheit» ist ein hilfreicher Haltepunkt. Frage vor diesem Schritt, ob der Standardfehler das tatsächliche Design abbildet, ob die Annahmen des Verfahrens plausibel sind und ob die Formulierung der Schlussfolgerung dem getesteten Inhalt entspricht. Ein Ergebnis kann einen Zusammenhang oder einen Unterschied stützen, ohne einen kausalen Effekt nachzuweisen.

Checkliste zur Interpretation

Benenne Grundgesamtheit, Stichprobe, Parameter, Stichprobenkennwert und Analyseeinheit. Beschreibe, wie die Fälle in die Stichprobe gelangten. Untersuche die Daten und bestimme fehlende oder ungewöhnliche Beobachtungen. Stimme das Verfahren auf das Skalenniveau der Zielvariable sowie auf unabhängige, verbundene oder kategoriale Daten ab. Formuliere H_0 und H_1 in Worten und Symbolen. Berichte die Schätzung und das Konfidenzintervall zusammen mit Teststatistik, Freiheitsgraden, sofern diese anwendbar sind, p-Wert und einer Interpretation im Kontext.

Forme den p-Wert nicht in eine Wahrscheinlichkeit dafür um, dass eine Hypothese wahr ist. Verwende statistische Signifikanz nicht als Synonym für praktische Bedeutung. Vergleiche Grösse und Unsicherheit der Schätzung mit der Forschungsfrage. Besprich bei einem nicht signifikanten Ergebnis das Intervall und die Präzision, statt die Gleichheit von Gruppen zu behaupten. Werden mehrere Tests durchgeführt, kann die Wahrscheinlichkeit mindestens eines Fehlers 1. Art steigen. Verwende deshalb eine geplante Methode zur Berücksichtigung multipler Tests, wenn sie erforderlich ist.

Verbindung zu anderen Themen

Die Wahrscheinlichkeitsrechnung stellte die Stichprobenverteilungen bereit, die Konfidenzintervalle und Tests ermöglichen. Dasselbe Inferenzmuster begleitet nun jeden späteren Koeffizienten. Eine Korrelation besitzt einen Standardfehler und einen Test. Eine Regressionssteigung besitzt eine Schätzung, ein Intervall und einen p-Wert. Eine partielle Korrelation und jeder multiple Regressionskoeffizient werden bedingt interpretiert. Die Varianzanalyse verwendet eine F -Statistik, um modellbezogene und verbleibende Variation zu vergleichen.

Inferenz ist deshalb kein losgelöstes Ritual am Ende einer Analyse. Sie ist eine geordnete Brücke von der deskriptiven Evidenz einer Stichprobe zu einer begrenzten Aussage über die Grundgesamtheit. Wenn Schätzung, Unsicherheit, Design und inhaltliche Bedeutung gemeinsam betrachtet werden, wird diese Brücke vertrauenswürdig.

Dokument-ID: topic-03-hypothesis-testing-summary-de

Notizen 1 von 3

Hypothesentests und Konfidenzintervalle | Thema 3

Notizen 2 von 3

Hypothesentests und Konfidenzintervalle | Thema 3

Notizen 3 von 3

Hypothesentests und Konfidenzintervalle | Thema 3