



Lernstoff-Zusammenfassung - Statistik

KI-generierte Zusammenfassung deines eingegebenen Materials (Upload bzw. eingegebene Stichpunkte) - bei Unklarheiten gerne zusätzlich im Original nachschlagen.

Statistik

Merkmale und Skalenniveaus

Jede Untersuchungseinheit (z. B. eine Versuchsperson in einer psychologischen Studie) besitzt Merkmale bzw. Variablen, die unterschiedliche Ausprägungen annehmen können. Bevor man Daten statistisch auswertet, muss man bestimmen, welches Skalenniveau ein Merkmal besitzt, da davon abhängt, welche Rechenoperationen und späteren statistischen Verfahren (Lage- und Streuungsmaße, aber auch Signifikanztests) überhaupt zulässig und sinnvoll sind. Man unterscheidet vier Skalenniveaus mit zunehmendem Informationsgehalt. Die Nominalskala erlaubt nur die Unterscheidung von Kategorien ohne Rangfolge (z. B. Geschlecht, Studienfach, Haarfarbe); hier sind nur Häufigkeitsauszählungen und der Modus sinnvoll. Die Ordinalskala erlaubt zusätzlich eine Rangordnung der Ausprägungen, wobei die Abstände zwischen den Rängen nicht interpretierbar sind (z. B. Schulnoten, Zustimmungsgrade wie 'gar nicht - etwas - stark!'); hier werden zusätzlich Median und Quantile zulässig. Die Intervallskala besitzt gleiche Abstände zwischen den Werten, aber keinen natürlichen Nullpunkt (z. B. Temperatur in Grad Celsius, IQ-Werte); ab diesem Niveau sind arithmetisches Mittel und Varianz zulässig, da Differenzen interpretierbar sind. Die Verhältnisskala (Ratioskala) besitzt zusätzlich einen natürlichen Nullpunkt, wodurch auch Verhältnisaussagen wie 'doppelt so groß' sinnvoll werden (z. B. Reaktionszeit in Millisekunden, Körpergröße, Einkommen). Intervall- und Verhältnisskala werden zusammenfassend als metrische Skalen oder Kardinalskalen bezeichnet, da für sie die vollständige Bandbreite arithmetischer Operationen zulässig ist. Eine zentrale Faustregel lautet: Je höher das Skalenniveau, desto mehr Information steckt in den Daten und desto mehr statistische Verfahren stehen zur Verfügung – ein höheres Skalenniveau kann aber jederzeit künstlich auf ein niedrigeres reduziert werden (z. B. Reaktionszeiten in 'schnell/langsam' einteilen), nicht jedoch umgekehrt.

Schlüsselbegriffe

Skalenniveau · Nominalskala · Ordinalskala · Intervallskala · Verhältnisskala (Ratioskala) · metrische Skala / Kardinalskala · Merkmal/Variable

Häufigkeiten

Um die Verteilung eines Merkmals in einer Stichprobe zu beschreiben, zählt man zunächst, wie oft jede Ausprägung vorkommt. Die absolute Häufigkeit $h(a)$ gibt die Anzahl der Beobachtungen mit der Ausprägung a an. Da absolute Häufigkeiten von der Stichprobengröße n abhängen und somit zwischen Stichproben unterschiedlicher Größe schwer vergleichbar sind, berechnet man die relative Häufigkeit $f(a)$, die den Anteil der Ausprägung a an der Gesamtstichprobe angibt. Die relativen Häufigkeiten aller Ausprägungen eines Merkmals summieren sich stets zu 1 (bzw. 100 %), was eine





wichtige Kontrollrechnung beim Erstellen von Häufigkeitstabellen darstellt. Ab ordinalem Skalenniveau, bei dem eine sinnvolle Rangordnung der Ausprägungen existiert, lassen sich zusätzlich kumulierte Häufigkeiten bilden. Die kumulierte relative Häufigkeit $F(a)$ gibt an, welcher Anteil der Beobachtungen einen Wert kleiner oder gleich a aufweist. Kumulierte Häufigkeiten sind die Grundlage für die spätere Bestimmung von Median und anderen Quantilen, da man über $F(a)$ direkt ablesen kann, an welcher Stelle der Verteilung ein bestimmter Anteil der Daten liegt.

Schlüsselbegriffe

absolute Häufigkeit $h(a)$ · relative Häufigkeit $f(a)$ · kumulierte Häufigkeit $F(a)$ · Stichprobenumfang n

Σ Formeln

$$f(a) = h(a) / n$$

Summe aller $f(a)$ über alle Ausprägungen = 1

$F(a)$ = Anteil der Beobachtungen mit Wert $\leq a$

Lagemaße

Lagemaße (auch Maße der zentralen Tendenz) beschreiben, wo sich das Zentrum einer Verteilung befindet, und fassen damit eine ganze Verteilung in einer einzigen Kennzahl zusammen. Welches Lagemaß zulässig ist, richtet sich nach dem Skalenniveau. Der Modus (Modalwert) ist die am häufigsten auftretende Ausprägung und ist bereits ab Nominalskalenniveau zulässig, da er lediglich Häufigkeiten vergleicht, ohne eine Rangordnung oder Abstände vorauszusetzen. Der Median teilt die der Größe nach geordnete Datenreihe in zwei gleich große Hälften: Bei ungeradem Stichprobenumfang n ist er der mittlere Wert der geordneten Reihe, bei geradem n das arithmetische Mittel der beiden mittleren Werte. Der Median setzt mindestens Ordinalskalenniveau voraus, da eine Rangordnung der Werte nötig ist, und ist wegen seiner reinen Positionsbestimmung robust gegenüber Ausreißern – extreme Einzelwerte verschieben den Median kaum. Das arithmetische Mittel $\bar{x}$ ist die Summe aller Werte geteilt durch deren Anzahl und setzt mindestens Intervallskalenniveau voraus, da es auf der Interpretierbarkeit von Differenzen zwischen Werten beruht. Im Gegensatz zum Median reagiert das arithmetische Mittel empfindlich auf Ausreißer, da jeder einzelne Wert mit seinem vollen numerischen Betrag in die Berechnung eingeht. Dies illustriert das Beispiel von fünf Reaktionszeiten (290, 300, 310, 350, 900 ms): Der Median liegt bei 310 ms und repräsentiert die 'typische' Reaktionszeit gut, während der Mittelwert durch den Ausreißer von 900 ms auf 430 ms nach oben gezogen wird und die zentrale Tendenz der übrigen Werte verzerrt wiedergibt. Dieser Zusammenhang zeigt sich systematisch bei schiefen Verteilungen: In einer rechtsschiefen (linkssteilen) Verteilung, bei der sich wenige sehr große Werte im rechten Ausläufer befinden, gilt typischerweise die Reihenfolge Modus kleiner als Median kleiner als Mittelwert, weil der Mittelwert am stärksten von den extremen Werten im rechten Ausläufer beeinflusst wird. Für die Praxis folgt daraus die wichtige Regel, bei schiefen Verteilungen oder dem Verdacht auf Ausreißer eher den Median zu berichten, bei annähernd symmetrischen metrischen Verteilungen hingegen das arithmetische Mittel, da es alle Informationen der Daten nutzt





und für weiterführende statistische Verfahren benötigt wird.

Schlüsselbegriffe

Modus (Modalwert) · Median · arithmetisches Mittel · Robustheit gegenüber Ausreißern · rechtsschiefe (linkssteile) Verteilung

Σ Formeln

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n$$

Modus < Median < Mittelwert (bei rechtsschiefer Verteilung)

Streuungsmaße

Während Lagemaße das Zentrum einer Verteilung beschreiben, quantifizieren Streuungsmaße, wie stark die einzelnen Werte um dieses Zentrum variieren – zwei Verteilungen können denselben Mittelwert, aber völlig unterschiedliche Streuung besitzen, weshalb Lage- und Streuungsmaße stets gemeinsam berichtet werden sollten. Die Spannweite (Range) ist die einfachste Kennzahl und ergibt sich als Differenz zwischen Maximum und Minimum der Daten; sie ist jedoch extrem anfällig für einzelne Ausreißer, da sie nur auf zwei Extremwerten beruht und die restliche Verteilung ignoriert. Robuster sind die Quartile: Das erste Quartil Q1 trennt die unteren 25 % der Werte von den oberen 75 % ab, das dritte Quartil Q3 trennt die unteren 75 % von den oberen 25 % ab (der Median entspricht dem zweiten Quartil Q2). Der Interquartilsabstand IQR ergibt sich als Differenz Q3 minus Q1 und umfasst damit die mittleren 50 % der Daten; da er auf Rangpositionen beruht, ist er robust gegenüber Ausreißern und ab Ordinalskalenniveau sinnvoll einsetzbar. Für metrische Daten ist die Varianz das zentrale Streuungsmaß: Sie berechnet sich als mittlere quadrierte Abweichung der Einzelwerte vom arithmetischen Mittel. Dabei muss zwischen zwei Varianten unterschieden werden: In der rein deskriptiven Statistik wird durch n geteilt (s^2 = Summe der quadrierten Abweichungen durch n), während man bei der Schätzung der Populationsvarianz aus einer Stichprobe durch n minus 1 teilt (korrigierte oder 'erwartungstreue' Stichprobenvarianz), um eine systematische Unterschätzung der wahren Populationsvarianz zu vermeiden. Da diese beiden Varianten unterschiedliche Zahlenwerte liefern, muss in jeder Berechnung explizit angegeben werden, welcher Divisor verwendet wurde. Die Standardabweichung s ist die Quadratwurzel der Varianz und hat gegenüber der Varianz den Vorteil, dieselbe Maßeinheit wie die Ursprungsdaten zu besitzen (z. B. Millisekunden statt Millisekunden zum Quadrat), wodurch sie inhaltlich leichter interpretierbar ist. Das Rechenbeispiel mit den Werten 2, 4, 6 (Mittelwert 4, quadrierte Abweichungen 4, 0 und 4, Summe 8) verdeutlicht den Unterschied: Bei Division durch n = 3 ergibt sich eine Varianz von etwa 2,67 und eine Standardabweichung von etwa 1,63; bei Division durch n – 1 = 2 ergibt sich eine Varianz von 4 und eine Standardabweichung von 2 – der Unterschied zwischen beiden Varianten wird bei kleinen Stichproben besonders deutlich.

Schlüsselbegriffe

Spannweite (Range) · Quartil (Q1, Q3) · Interquartilsabstand (IQR) · Varianz s^2 · Standardabweichung s · korrigierte Stichprobenvarianz (n–1)





Σ Formeln

$$R = \text{Maximum} - \text{Minimum}$$

$$IQR = Q3 - Q1$$

$$s^2 (\text{deskriptiv}) = \sum (x_i - \bar{x})^2 / n$$

$$s^2 (\text{Schätzer, korrigiert}) = \sum (x_i - \bar{x})^2 / (n - 1)$$

$$s = \sqrt{s^2}$$

z-Standardisierung

Werte aus unterschiedlichen Verteilungen – etwa Testergebnisse aus zwei verschiedenen psychologischen Tests mit unterschiedlichem Mittelwert und unterschiedlicher Streuung – lassen sich nicht direkt miteinander vergleichen, da ein Rohwert von z. B. 60 Punkten je nach Schwierigkeit und Streuung des Tests ganz unterschiedlich zu bewerten ist. Um solche Werte vergleichbar zu machen, führt man eine z-Transformation (Standardisierung) durch: Von jedem Einzelwert x wird der Mittelwert der zugehörigen Verteilung subtrahiert, und diese Differenz wird durch die Standardabweichung der Verteilung geteilt. Der resultierende z-Wert gibt an, um wie viele Standardabweichungen ein Wert über (bei positivem z) oder unter (bei negativem z) dem Mittelwert seiner Verteilung liegt, und ist damit unabhängig von der ursprünglichen Maßeinheit und Skala. Eine wichtige Eigenschaft standardisierter Werte ist, dass die gesamte Menge der z-Werte einer Stichprobe selbst wieder einen Mittelwert von 0 und eine Standardabweichung von 1 besitzt. Das Beispiel von Anna verdeutlicht den praktischen Nutzen: In Test A erzielt sie 60 Punkte bei einem Gruppenmittelwert von 50 und einer Standardabweichung von 5, was einem z-Wert von 2 entspricht (sie liegt zwei Standardabweichungen über dem Durchschnitt). In Test B erzielt sie 75 Punkte bei einem Mittelwert von 70 und einer Standardabweichung von 10, was einem z-Wert von 0,5 entspricht. Obwohl der Rohwert in Test B höher ist (75 gegenüber 60 Punkten), zeigt der Vergleich der z-Werte, dass Anna relativ zu ihrer jeweiligen Vergleichsgruppe in Test A deutlich besser abgeschnitten hat. Die z-Standardisierung ist damit ein grundlegendes Werkzeug, um Leistungen, Testwerte oder Messungen aus unterschiedlichen Skalen und Verteilungen fair miteinander zu vergleichen, und bildet zugleich die Basis für die spätere Arbeit mit der Standardnormalverteilung in der Inferenzstatistik.

Schlüsselbegriffe

z-Standardisierung · z-Wert · Standardnormalverteilung (Grundlage) · Vergleichbarkeit von Verteilungen

Σ Formeln

$$z = (x - \bar{x}) / s$$

Grafische Darstellung von Verteilungen

Neben Kennzahlen wie Lage- und Streuungsmaßen werden Verteilungen häufig grafisch dargestellt, da sich so Form, Schiefe und Ausreißer einer Verteilung oft schneller erfassen lassen als anhand reiner Zahlen. Die Wahl der geeigneten Grafik hängt wiederum vom Skalenniveau des Merkmals ab. Für nominale und ordinale Daten, bei





denen diskrete Kategorien vorliegen, eignen sich Balkendiagramme (zur Darstellung absoluter oder relativer Häufigkeiten pro Kategorie) und Kreisdiagramme (zur Darstellung der prozentualen Anteile am Gesamten). Für metrische Daten, die in der Regel viele unterschiedliche Einzelwerte annehmen, werden die Werte zunächst in Klassen (Intervalle) eingeteilt und mit einem Histogramm dargestellt; dabei ist wichtig, dass nicht die Höhe, sondern die Fläche eines Balkens der Häufigkeit (bzw. relativen Häufigkeit) der jeweiligen Klasse entspricht, was insbesondere bei unterschiedlich breiten Klassen relevant wird. Eine besonders informationsreiche und kompakte Darstellung metrischer Daten ist der Boxplot, der auf den Lage- und Streuungsmaßen des vorangegangenen Kapitels beruht: Er zeigt den Median als Mittellinie, die Box wird durch die Quartile Q1 und Q3 begrenzt (die Boxlänge entspricht damit dem Interquartilsabstand IQR), und sogenannte Whisker erstrecken sich bis zu den am weitesten entfernten Werten innerhalb eines definierten Bereichs, während darüber hinausgehende Werte einzeln als Ausreißer eingezeichnet werden. Der Boxplot eignet sich besonders gut, um mehrere Gruppen (z. B. Versuchsgruppen in einem Experiment) nebeneinander hinsichtlich Lage, Streuung und Schiefe ihrer Verteilungen zu vergleichen. Als übergreifende Merkregel für die Klausur gilt: Zunächst muss das Skalenniveau des Merkmals bestimmt werden, da dieses festlegt, welche Lage- und Streuungsmaße überhaupt zulässig sind; liegen Ausreißer oder schiefe Verteilungen vor, sollten robuste Kennwerte wie Median und IQR berichtet werden, während bei annähernd symmetrischen metrischen Daten ohne Ausreißer Mittelwert und Standardabweichung die aussagekräftigeren und für weiterführende Verfahren notwendigen Kennwerte darstellen.

Schlüsselbegriffe

Balkendiagramm · Kreisdiagramm · Histogramm · Boxplot · Whisker · Flächenproportionalität im Histogramm

