Deskriptive Statistik

Ein verständlicher Wegweiser zu Variablen, Verteilungen und numerischen Kennwerten

Ratiomera Statistics

Inhaltsverzeichnis

Zusammenfassung der Kursseite . . . 2
Notation und Entscheidungscheckliste . . . 2
Skalenniveaus . . . 2
Häufigkeitsverteilungen . . . 3
Zentrale Tendenz . . . 3
Variabilität . . . 4
Lineare Transformationen . . . 4
z-Transformation . . . 4
Diagrammwahl nach Skalenniveau . . . 5
Verteilungsformen . . . 5
Irreführende Grafiken . . . 5
Das solltest du jetzt können . . . 5
Erweiterte Referenz . . . 7
Zweck und Grundlagen . . . 7
Zentrale Ideen . . . 7
Formelleitfaden . . . 8
Die erklärende Abbildung lesen . . . 10
Checkliste zur Interpretation . . . 11
Verbindung zu anderen Themen . . . 11

Zusammenfassung der Kursseite

Die deskriptive Statistik verbindet Messung und Analyse. Bestimme zuerst, was jeder Fall, jede Variable und jeder Wert darstellt. Verwende anschliessend das Skalenniveau, um begründbare Häufigkeiten, numerische Kennwerte und Grafiken auszuwählen. Lage, Streuung und Form müssen gemeinsam interpretiert werden, weil keine dieser Dimensionen allein eine vollständige Beschreibung liefert. Diese Konzepte stützen Wahrscheinlichkeit, Inferenz, Korrelation, Regression und alle späteren Themen dieser Lernsequenz.

Notation und Entscheidungscheckliste

Tabelle 1: Zentrale Notation der deskriptiven Statistik.

Symbol	Bedeutung	Prüffrage
x_i	Wert der Beobachtung i	Was stellt ein einzelner aufgezeichneter Wert dar?
n und N	Stichprobengrösse und Grösse der Grundgesamtheit	Beschreibe ich eine Stichprobe oder die vollständige Grundgesamtheit?
$\bar{x}$ und μ	Stichprobenmittelwert und Populationsmittelwert	Passt das Symbol zur Zielgrösse?
s^2 und σ^2	Stichprobenvarianz und Populationsvarianz	Passt der Nenner zur Zielgrösse?
s und σ	Stichproben-SD und Populations-SD	Wie lautet die ursprüngliche Messeinheit?
$\sum$	Den angegebenen Ausdruck über seinen Indexbereich addieren	Was ist der Summand, und wie lauten die Grenzen?

Bestimme vor dem Berichten eines Ergebnisses die Fälle und Variablen, prüfe fehlende oder unzulässige Werte, bestimme das Skalenniveau, untersuche die Verteilung, wähle eine passende Lage und Streuung und benenne, was die Kennwerte über die Grundgesamtheit oder Kausalität nicht belegen können.

Skalenniveaus

Fünf Skalenniveaus bestimmen, welche Operationen und Kennwerte zulässig sind. Jedes Niveau ergänzt Eigenschaften des darüberstehenden Niveaus.

Tabelle 2: Messeigenschaften, zulässige Transformationen und zentrale Kennwerte.

Skala	Definierende Eigenschaft	Zulässige Umcodierung	Zentrale Kennwerte
Nominal	Kategorien ohne Reihenfolge	Eineindeutige Umbenennung	Häufigkeiten, Anteile, Modalwert
Ordinal	Geordnete Kategorien, ungleiche Abstände	Streng monoton steigende Umcodierung	Häufigkeiten, Anteile, Median, Perzentile, Modalwert
Intervall	Gleiche Abstände, festgelegter Nullpunkt	$y = a + bx,\ b > 0$	Mittelwert, SD, Differenzen, Korrelation
Verhältnis	Gleiche Abstände, natürlicher Nullpunkt	$y = bx,\ b > 0$	Kennwerte der Intervallskala und sinnvolle Verhältnisse
Absolut	Natürlicher Nullpunkt und feste natürliche Einheit	$y = x$	Kennwerte der Verhältnisskala und Anzahlen

Intervall-, Verhältnis- und Absolutskalen werden zusammen als **metrische Skalen** bezeichnet.

Häufigkeitsverteilungen

Tabelle 3: Kompakte Übersicht zu Häufigkeitsverteilungen.

Grösse	Formel	Interpretation
Absolute Häufigkeit	n_j	Anzahl Beobachtungen in Kategorie j
Relative Häufigkeit	$f_j = n_j / n$	Anteil in Kategorie j
Kumulative relative Häufigkeit	$F_j = \sum_{k=1}^{j} f_k$	Anteil bis einschliesslich geordneter Kategorie j

Absolute Häufigkeiten summieren sich zu n und relative Häufigkeiten, abgesehen von Rundungsabweichungen, zu 1. Kumulative relative Häufigkeiten setzen eine geordnete Variable voraus und sind für nominale Kategorien nicht sinnvoll.

Zentrale Tendenz

Drei Lagemasse beschreiben einen typischen Wert:

Tabelle 4: Lagemasse und ihre Verwendung.

Kennwert	Formel	Geeignete Verwendung
Mittelwert	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	Symmetrische Verteilungen; metrische Daten
Median	Mittlerer Wert der geordneten Daten	Schiefe Verteilungen; robust gegenüber Ausreissern
Modalwert	Häufigster Wert oder häufigste Kategorie	Jedes Skalenniveau; kann gebunden oder nicht eindeutig sein

Liegt der Mittelwert deutlich über dem Median, kann dies auf Rechtsschiefe hindeuten, insbesondere wenn ein Histogramm zugleich einen langen rechten Randbereich zeigt. Die Reihenfolge der beiden Kennwerte allein belegt die Verteilungsform nicht.

Variabilität

Vier Streuungsmasse beschreiben, wie stark sich Personen unterscheiden:

Tabelle 5: Streuungsmasse und ihre Verwendung.

Kennwert	Formel	Geeignete Verwendung
Spannweite	$x_{\max} - x_{\min}$	Einfacher Überblick; empfindlich gegenüber Extremwerten
IQR	$Q_3 - Q_1$	Robuste Streuung; zusammen mit dem Median
Varianz	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	Grundlage der SD und weiterer Kennwerte
SD	$s = \sqrt{s^2}$	Grössenordnung der Abweichungen vom Mittelwert; zusammen mit dem Mittelwert

Die Ausreissergrenzen eines Boxplots liegen bei $Q_1 - 1.5 \cdot \mathrm{IQR}$ und $Q_3 + 1.5 \cdot \mathrm{IQR}$. Die Whisker reichen bis zu den äussersten beobachteten Werten innerhalb dieser Grenzen. Werte ausserhalb werden als mögliche Ausreisser markiert.

Lineare Transformationen

Eine lineare Transformation $y_i = a + b \cdot x_i$ verändert Kennwerte nach festen Regeln:

Tabelle 6: Wirkung einer linearen Transformation auf deskriptive Kennwerte.

Grösse	Transformationsregel	Wirkung
Mittelwert	$\bar{y} = a + b\bar{x}$	Um a verschoben und mit b skaliert
Varianz	$s_y^2 = b^2 s_x^2$	Mit b^2 skaliert; von a unbeeinflusst
Standardabweichung	$s_y = \lvert b \rvert s_x$	Mit $\lvert b \rvert$ skaliert; von a unbeeinflusst
Rangfolge und Orientierung	Bei $b > 0$ bewahrt	Bei $b < 0$ umgekehrt

Der transformierte Mittelwert folgt $\bar{y} = a + b\bar{x}$. Die Verschiebung a verändert den Mittelwert, aber nicht die Variabilität, während $\lvert b \rvert$ die Abstände und die Standardabweichung skaliert. Ein negatives b spiegelt zudem die Verteilung und kehrt die Rangfolge um.

---

z-Transformation

Für eine nicht konstante Stichprobe mit einem definierten, von null verschiedenen s erzeugt die z-Transformation $z_i = (x_i - \bar{x})/s$ einen Stichprobenmittelwert von 0 und eine Stichprobenstandardabweichung von 1.

Ein z-Wert gibt an, wie viele Standardabweichungen eine Beobachtung über oder unter dem Mittelwert der gewählten Referenzverteilung liegt. Vergleiche zwischen unterschiedlichen Skalen sind nur sinnvoll, wenn dieselbe inhaltliche Eigenschaft erfasst wird, die Messqualität vergleichbar ist und geeignete Referenzgruppen verwendet werden.

Diagrammwahl nach Skalenniveau

Tabelle 7: Diagrammwahl nach Skalenniveau.

Skalenniveau	Geeignetes Diagramm
Nominal	Säulendiagramm, Kreisdiagramm
Ordinal	Geordnetes Säulendiagramm
Metrisch (Intervall, Verhältnis, Absolut)	Histogramm, Boxplot

In Histogrammen ist die **Fläche proportional zur Häufigkeit**. Bei gleichen Klassenbreiten bilden Höhe und Fläche die Häufigkeit korrekt ab. Bei ungleichen Breiten beträgt die Dichtehöhe f_j/w_j. Vergleiche sinnvolle Klassenbreiten, weil sich die sichtbare Form mit der Gruppierung ändern kann.

Verteilungsformen

Tabelle 8: Unabhängige Dimensionen zur Beschreibung der Verteilungsform.

Form	Zentrales Merkmal	Hinweis für die Interpretation
Symmetrisch	Gleichartige Randbereiche auf beiden Seiten	Mittelwert und Median liegen nahe beieinander
Rechtsschief	Langer Randbereich zu hohen Werten	Mittelwert liegt häufig über dem Median; visuell prüfen
Linksschief	Langer Randbereich zu niedrigen Werten	Mittelwert liegt häufig unter dem Median; visuell prüfen
Unimodal	Ein Gipfel	Eine Hauptkonzentration, entsprechend ihrer Form zusammengefasst
Bimodal	Zwei Gipfel	Beide Gipfel zeigen und untersuchen; Teilgruppen nicht voraussetzen
Hohe Kurtosis	Stärkere Gewichtung extremer Abweichungen	Grössere Neigung zu weit entfernten Beobachtungen
Niedrige Kurtosis	Schwächere Gewichtung extremer Abweichungen	Geringere Neigung zu weit entfernten Beobachtungen

Symmetrie, Modalität und Kurtosis können in unterschiedlichen Kombinationen gemeinsam auftreten. Die Gipfelhöhe allein definiert Kurtosis nicht.

Irreführende Grafiken

Grafiken können durch abgeschnittene Achsen, gestauchte oder gestreckte Skalen und ungeeignete Histogrammklassen irreführende Eindrücke erzeugen. Prozentwerte können irreführen, wenn Bezugsgrösse, Stichprobengrösse, Messmethode oder fehlende Daten verschwiegen werden. Grafiken und Zahlen benötigen Kontext und eine gültige Messung.

Das solltest du jetzt können

Du solltest jetzt Fälle, Variablen, Werte und Skalenniveaus bestimmen, die Notation für Mittelwerte, Varianzen und z-Werte lesen, Häufigkeitsdarstellungen erstellen und interpretieren, passende Lage- und Streuungsmasse wählen, Verteilungen mit Tabellen und Grafiken vergleichen, die Wirkung linearer Transformationen vorhersagen, Beobachtungen standardisieren und Darstellungsentscheidungen erkennen können, die eine statistische Botschaft verzerren. Zudem

solltest du die Grenzen einer deskriptiven Schlussfolgerung benennen können, ohne aus einem beobachteten Muster eine Aussage über die Grundgesamtheit oder einen Kausalschluss zu machen.

Erweiterte Referenz

Zweck und Grundlagen

Die deskriptive Statistik bietet dir ein geordnetes Vorgehen, um aus einer Sammlung von Beobachtungen eine verständliche Beschreibung zu entwickeln. Bestimme zuerst die **Fälle**, also die Personen oder Untersuchungseinheiten in den Zeilen eines Datensatzes, und die **Variablen**, also die in den Spalten erfassten Merkmale. Ein Wert ist ein aufgezeichnetes Ergebnis für eine Variable bei einem Fall. Frage vor jeder Berechnung, was die Variable bedeutet, wie sie gemessen wurde und welche Werte möglich sind. So verhinderst du, dass eine mathematisch korrekte Berechnung zu einer irreführenden Beschreibung wird.

Das Skalenniveau bestimmt, welche Auswertungen für eine Variable sinnvoll sind. Eine nominale Variable ordnet Fälle Kategorien zu, ohne diese zu reihen. Eine ordinale Variable besitzt geordnete Kategorien, doch die Abstände zwischen benachbarten Kategorien sind nicht als gleich bekannt. Eine Intervallvariable besitzt bedeutungsvolle gleiche Abstände, aber keinen inhaltlich bedeutsamen absoluten Nullpunkt. Eine Verhältnisvariable besitzt gleiche Abstände und einen bedeutsamen Nullpunkt, sodass auch Verhältnisse interpretiert werden können. Das Skalenniveau ist eine Eigenschaft der Definition und Messung einer Variablen. Es hängt nicht davon ab, wie ihre Werte in einem einzelnen Datensatz aussehen.

Skalenniveau	Was die Werte aussagen	Geeignete erste Kennwerte
Nominal	Ob Fälle derselben oder verschiedenen Kategorien angehören	Häufigkeiten, Anteile, Modus
Ordinal	Kategorienzugehörigkeit und Reihenfolge	Häufigkeiten, Anteile, Median, Quantile
Intervall/Verhältnis	Reihenfolge und bedeutungsvoller numerischer Abstand	Mittelwert, Median, Varianz, Standardabweichung, Quantile

Einige Lehrdatensätze dieses Kurses sind **simulierte Daten**. Das sind am Computer erzeugte Werte, die festgelegten Regeln folgen, und keine Messungen an realen Personen. Eine **Simulation** ist der Vorgang, mit dem diese Werte erzeugt werden. Der Computer verwendet dazu einen **Zufallszahlengenerator**, also einen Algorithmus, der Werte erzeugt, die sich wie Zufallsergebnisse verhalten. Ein **Seed** ist der Startwert für diesen Algorithmus. Wenn derselbe Seed mit denselben Anweisungen erneut verwendet wird, entsteht derselbe Datensatz. Dadurch ist ein Lehrbeispiel reproduzierbar: Du und eine andere lernende Person könnt dieselben Beobachtungen untersuchen und erhaltet dieselben Ergebnisse. Eine Simulation unterstützt das Lernen, doch erzeugte Werte werden dadurch nicht zu Evidenz über eine reale Grundgesamtheit.

Zentrale Ideen

Eine Verteilung beschreibt, wie sich die Werte einer Variablen über ihren möglichen Bereich verteilen. Bei einer kategorialen Variable beginnst du mit einer Häufigkeitstabelle. Die absolute Häufigkeit ist die Anzahl Fälle in einer Kategorie. Die relative Häufigkeit ist diese Anzahl geteilt durch die Gesamtzahl der gültigen Fälle. Relative Häufigkeiten lassen sich als Anteile oder Prozentsätze angeben. Prüfe immer, ob fehlende Werte aus dem Nenner ausgeschlossen wurden. Ein Prozentsatz ist nur dann aussagekräftig, wenn seine Bezugsgrösse bekannt ist.

Bei einer numerischen Variable beschreibst du vier Merkmale gemeinsam: Lage, Streuung, Form und ungewöhnliche Beobachtungen. Der Mittelwert verwendet jeden Wert und bildet den Gleichgewichtspunkt der Verteilung. Der Median ist der mittlere geordnete Wert und teilt die Daten in zwei Hälften. Der Modus ist der häufigste Wert oder die häufigste Kategorie. Die Spannweite

reicht vom kleinsten bis zum grössten Wert. Der Interquartilsabstand umfasst die mittlere Hälfte der geordneten Beobachtungen. Varianz und Standardabweichung fassen zusammen, wie weit die Werte typischerweise vom Mittelwert entfernt liegen.

Frage	Nützliche Evidenz	Gute Lesegewohnheit
Wo liegt das Zentrum der Verteilung?	Mittelwert, Median und manchmal Modus	Mittelwert und Median vergleichen, statt nur eine Zahl anzugeben
Wie stark unterscheiden sich die Beobachtungen?	Spannweite, Interquartilsabstand, Varianz, Standardabweichung	Masseinheit nennen und auf ungewöhnliche Werte achten
Welche Form bilden die Werte?	Histogramm, Boxplot, Häufigkeiten, Schiefe	Auf Symmetrie, Schiefe, Lücken, Häufungen und mehrere Gipfel achten
Sind einzelne Werte überraschend?	Rohwerte, Grafik, Datenprüfung, standardisierte Werte	Zuerst untersuchen und erst danach entscheiden, ob ein Wert fehlerhaft ist

Die Form beeinflusst die Interpretation. In einer annähernd symmetrischen Verteilung sind Mittelwert und Median oft ähnlich. Ein langer rechter Rand zieht den Mittelwert tendenziell nach oben, ein langer linker Rand zieht ihn nach unten. **Modalität** beschreibt die Anzahl und das Muster klarer Gipfel oder Hauptkonzentrationen. Eine Verteilung kann unimodal sein und einen Hauptgipfel besitzen oder multimodal sein und mehrere Häufungen von Beobachtungen zeigen. **Kurtosis** beschreibt, wie leicht Werte weit in den Randbereichen auftreten, verglichen mit einer symmetrischen glockenförmigen Referenzverteilung mit derselben Gesamtstreuung. Die Gipfelhöhe allein definiert die Kurtosis nicht. Eine einzelne Zahl wie der Mittelwert kann diese Merkmale nicht darstellen. Deshalb solltest du eine Grafik und numerische Kennwerte gemeinsam lesen.

Eine ungewöhnliche Beobachtung ist nicht automatisch ein Fehler. Sie kann ein gültiger, aber seltener Fall, ein Codierungsfehler, ein Messproblem oder ein Hinweis darauf sein, dass verschiedene Gruppen zusammengefasst wurden. Prüfe die ursprüngliche Definition und den Erfassungsvorgang, bevor du etwas ausschliesst. Wenn sich eine analytische Entscheidung nach dem Entfernen einer Beobachtung verändert, solltest du diese Empfindlichkeit offen berichten.

Formelleitfaden

Für Kategorie oder Intervall j sei n_j die absolute Häufigkeit und n die Anzahl gültiger Beobachtungen. Die relative Häufigkeit lautet

$$f_j = \frac{n_j}{n}.$$

Bei geordneten Kategorien oder numerischen Intervallen ist die kumulierte relative Häufigkeit bis einschliesslich Kategorie j

$$F_j = \sum_{h=1}^{j} f_h.$$

Die absoluten Häufigkeiten sollten sich zu n und die relativen Häufigkeiten abgesehen von Rundungen zu 1 addieren. Die letzte kumulierte relative Häufigkeit sollte ebenfalls 1 betragen. Kumulierte Häufigkeiten sind nur bei Kategorien mit einer inhaltlich vertretbaren Reihenfolge sinnvoll.

Seien $x_1, x_2, ..., x_n$ die beobachteten Werte einer numerischen Variablen. Der Stichprobenmittelwert addiert alle Werte und teilt die Summe durch die Anzahl Beobachtungen:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Das Zeichen $\sum$ bedeutet, dass die angegebenen Werte addiert werden. Der Index i bezeichnet jeweils eine Beobachtung, von der ersten Beobachtung bis zur Beobachtung n. Die Abweichung $x_i - \bar{x}$ zeigt, wie weit ein Wert über oder unter dem Mittelwert liegt. Positive und negative Abweichungen heben sich beim Addieren gegenseitig auf. Deshalb werden sie für die Varianz zuerst quadriert. Die Stichprobenvarianz verwendet $n - 1$ im Nenner:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Die entsprechende Populationsvarianz verwendet den Populationsmittelwert μ und teilt durch die Populationsgrösse N:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2.$$

Halte die beiden Zielgrössen auseinander. Der Nenner $n-1$ gehört zur korrigierten Stichprobenvarianz, mit der die Variabilität der Grundgesamtheit geschätzt wird. Die Division durch N beschreibt dagegen die vollständigen Werte der Grundgesamtheit selbst.

Die Varianz wird in quadrierten Einheiten angegeben. Mit ihrer Quadratwurzel kehren wir zur ursprünglichen Masseinheit zurück. Die Stichprobenstandardabweichung lautet deshalb:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Ein standardisierter Wert drückt einen einzelnen Wert in Standardabweichungseinheiten aus. Dazu wird der Stichprobenmittelwert abgezogen und durch die Stichprobenstandardabweichung geteilt:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Ein positives z_i bedeutet, dass der Wert über dem Mittelwert liegt, ein negatives z_i bedeutet, dass er darunter liegt. Der Betrag gibt den Abstand in Standardabweichungen an. Die Standardisierung verändert die Einheit und den Bezugspunkt, aber weder die Reihenfolge noch die Form der Beobachtungen.

Median und Quantile beginnen mit den geordneten Beobachtungen. Das erste Quartil Q_1 markiert das untere Viertel, der Median Q_2 die Hälfte und das dritte Quartil Q_3 die unteren drei Viertel. Spannweite und Interquartilsabstand lauten

$$\text{Spannweite} = x_{\max} - x_{\min}, \qquad IQR = Q_3 - Q_1.$$

Konventionen für Stichprobenquantile können unterschiedlich interpolieren. Bei einem kleinen Datensatz kann Software deshalb leicht verschiedene Quartile ausgeben. Eine verbreitete Boxplot-Diagnose verwendet die inneren Grenzen

$$Q_1 - 1.5(IQR) \qquad \text{und} \qquad Q_3 + 1.5(IQR).$$

Werte ausserhalb einer Grenze sind mögliche Ausreisser, die untersucht werden sollten, und keine automatisch zu löschenden Fehler. Die Whisker enden bei den extremsten beobachteten Werten,

die noch innerhalb der Grenzen liegen. Sie enden nicht zwingend genau an den berechneten Grenzen.

Bei einer linearen Transformation $Y = a + bX$ verändern sich Lage und Streuung gemäss

$$\bar{y} = a + b\bar{x}, \qquad s_y^2 = b^2 s_x^2, \qquad s_y = |b| s_x.$$

Die Verschiebung a verändert die Lage, aber nicht die Streuung. Der Faktor b verändert Abstände um $|b|$, weshalb sich die Varianz um b^2 verändert. Die Standardisierung ist der Spezialfall, bei dem der Mittelwert abgezogen und durch die Standardabweichung geteilt wird.

Bei der Höhe von Histogrammbalken ist eine letzte Prüfung nötig. Besitzt Klasse j die relative Häufigkeit f_j und die Breite w_j, lautet ihre Dichtehöhe

$$h_j = \frac{f_j}{w_j}.$$

Die Balkenfläche ist dann $h_j w_j = f_j$. Bei gleich breiten Klassen kann die Höhe die Häufigkeit direkt wiedergeben. Bei ungleichen Klassenbreiten werden Dichtehöhen benötigt, damit weiterhin die Fläche und nicht allein die Höhe die Häufigkeit darstellt.

Die erklärende Abbildung lesen

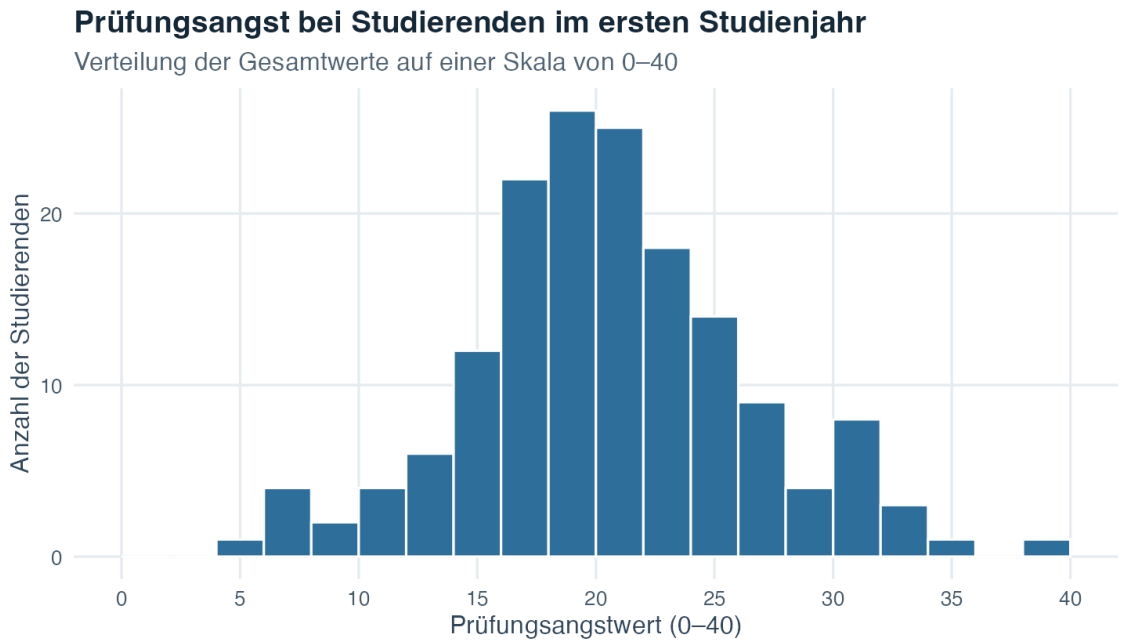

Abbildung 1: Histogramm simulierter Prüfungsangstwerte von null bis vierzig, mit den höchsten Balken nahe der Mitte und schmaler besetzten Rändern an beiden Enden.

Lies zuerst die horizontale Achse. Sie zeigt Prüfungsangstwerte auf einer Skala von 0 bis 40. Die vertikale Achse gibt die Anzahl der Studierenden an. Die Höhe jedes Balkens entspricht somit einer Häufigkeit. Die Balken bei Werten zwischen 18 und 22 sind am höchsten. Die grösste Häufung liegt also nahe der Skalenmitte. An den unteren und oberen Enden finden sich weniger Beobachtungen. Ein Balken nahe bei 40 zeigt, dass mindestens ein hoher Wert vorhanden ist. Aus der Grafik allein lässt sich jedoch nicht erkennen, ob dieser Wert fehlerhaft ist. Bevor du dies beurteilst, müsstest du zur Datendefinition und zum Erfassungsvorgang zurückkehren.

Die Balken berühren sich, weil eine numerische Skala über benachbarte Intervalle hinweg stetig verläuft. Auch ihre Breite ist wichtig: Andere Intervallgrenzen können dieselben Beobachtungen

mehr oder weniger detailliert erscheinen lassen. Die Abbildung ist deshalb als Ansicht einer einzigen Verteilung zu lesen und nicht als Ansammlung getrennter Kategorien. Sie stützt Aussagen über Lage, Streuung, Form und ungewöhnliche Beobachtungen. Sie zeigt weder eine Ursache von Prüfungsangst noch einen Vergleich zwischen Grundgesamtheiten. Ebenso belegt sie nicht, dass dieses simulierte Muster bei realen Studierenden vorkommt.

Checkliste zur Interpretation

Beginne jede deskriptive Beschreibung mit den Fällen, der Variable, ihrem Skalenniveau und dem gültigen Wertebereich. Gib die Anzahl der gültigen und fehlenden Beobachtungen an. Berichte bei einer kategorialen Variable die Häufigkeiten zusammen mit dem Nenner und den Anteilen. Verbinde bei einer numerischen Variable eine Grafik mit Kennwerten der Lage und Streuung. Verwende Mittelwert und Standardabweichung, wenn sie zur Verteilung passen. Untersuche zusätzlich Median und Interquartilsabstand, wenn Schiefe oder ungewöhnliche Beobachtungen eine Rolle spielen.

Halte die Masseinheit sichtbar. Eine Standardabweichung von fünf Punkten bedeutet etwas anderes als fünf Stunden. Bezeichne eine Gruppe nicht ohne einen sinnvollen Vergleich als homogen oder variabel. Prüfe, ob sich eine gerundete Tabelle noch zur erwarteten Gesamtsumme addiert. Kennzeichne simulierte Daten als simuliert. Trenne schliesslich Beschreibung und Erklärung: Ein Muster in den beobachteten Werten zeigt, was der Datensatz enthält. Eine kausale Erklärung erfordert dagegen ein geeignetes Forschungsdesign und Überlegungen, die über die deskriptive Statistik hinausgehen.

Verbindung zu anderen Themen

Die deskriptive Statistik stellt die Sprache bereit, die in der gesamten Lernsequenz verwendet wird. Die Wahrscheinlichkeitsrechnung ergänzt Regeln für das Denken über unsichere Ergebnisse. Die statistische Inferenz verwendet anschliessend eine Stichprobe und ihre Streuung, um vorsichtige Aussagen über eine Grundgesamtheit zu machen. Kovarianz und Korrelation fragen, wie zwei Variablen gemeinsam variieren. Die Regression beschreibt eine Zielvariable als Funktion eines oder mehrerer Prädiktoren, während die partielle Korrelation einen Zusammenhang nach linearer Bereinigung untersucht. Die Varianzanalyse vergleicht Gruppenmittelwerte, indem sie die Gesamtvariation in inhaltlich bedeutsame Bestandteile zerlegt.

Die zentrale Gewohnheit bleibt dabei unverändert: Verstehe die Variablen, untersuche die Verteilung, berechne einen passenden Kennwert und interpretiere ihn im Kontext. Spätere Verfahren ergänzen Unsicherheit und Modelle. Sie ersetzen jedoch nie die verlässliche Beschreibung der Daten, die in die Analyse eingegangen sind.

Dokument-ID: `topic-01-descriptive-statistics-summary-de`

Notizen 1 von 3

Deskriptive Statistik | Thema 1

Notizen 2 von 3

Notizen 3 von 3