ZUSAMMENFASSUNG STATISTIKEN UND GRAFISCHE DARSTELLUNGEN

Zusammenfassung zu Mathematik-Prüfung über Statistiken und grafische Darstellungen.

Exposee

Zusammenfassung zu FACH-Prüfung vom DATUM über THEMA.

RaviAnand Mohabir
ravianand.mohabir@stud.altekanti.ch
https://dan6erbond.github.io

Inhalt

Status: ☒ in Bearbeitung ☐ Beendet

1 Beschreibende Statistik

1.1 Einleitung

Entscheidungen werden oft auf der Grundlage statistischer Aussagen getroffen. Krankenkassen berechnen bspw. ihre Prämien anhand von Statistiken über die Häufigkeiten von Krankheitsfällen, Arztbesuchen etc.

In einer statistischen Datenerhebung geht es darum, Informationen über Personen oder Dinge zu sammeln. Im Idealfall werden die entsprechenden Daten von allen Personen oder Dingen erfasst, die für die jeweilige Untersuchung interessant sind. Man spricht dann von einer Grundgesamtheit oder Population. Wenn der Aufwand dafür zu gross ist, verlässt man sich stattdessen auf eine Stichprobe, d.h. eine repräsentative Teilmenge der Grundgesamtheit. Die Auswahl der Stichprobe muss repräsentativ sein, d.h. dass sie die wesentlichen Eigenschaften der Grundgesamtheit wiedergeben soll.

1.2 Grundbegriffe

In einer statistischen Erhebung wird aus einer bestimmten Grundgesamtheit eine Stichprobe von Personen oder Dingen ausgewählt und hinsichtlich bestimmter Merkmale (Variablen) untersucht. Jedes Merkmal kann bestimmte Merkmalsausprägungen (Variablenwert) annehmen). Man kann dabei folgende Grundtypen von Merkmalen unterscheiden:

Quantitative (metrische) Merkmale besitzen einen natürlichen Zahlenwert, der direkt durch eine Messung bestimmt werden kann. Metrische Merkmale heissen stetig, wenn sie innerhalb gewisser Grenzen jeden Zahlenwert annehmen können, andernfalls heissen sie diskret.

Qualitativ Merkmale besitzen keinen natürlichen Zahlenwert und können deshalb nur verbal beschrieben oder zahlenmässige codiert werden. Qualitativ Merkmale heissen ordinal, wenn man sie ordnen kann, ansonsten heissen sie nominal.

Beispiele:

Merkmal	***Merkmalsausprägung***	***Grundtyp***
Augenfarbe	blau, braun, grün, grau	qualitativ, nominal
Schulnoten (USA)	A, B, C, D, E, F	qualitativ, ordinal
Anzahl Geschwister	1, 2, 3, 4	metrisch, diskret
Körpergrösse (in cm)	150, 160, 170…	metrisch, stetig

Als Umfang der Stichprobe bezeichnet man die Anzahl (n) von Personen bzw. Dingen, welche in der Datenerhebung berücksichtigt werden. Nach erfolgter Datenerhebung hat man eine Liste (Urliste) von Merkmalsausprägungen in Form von n Daten und kann zählen, wie oft die verschiedenen Merkmalsausprägungen vorkommen. Man spricht dann von der Häufigkeit einer Merkmalsausprägung:

- Die absolute Häufigkeit H der Merkmalsausprägung x, gibt an, wie oft die Merkmalsausprägung x vorkommt.
- Die relative Häufigkeit h, einer Merkmalsausprägung x ist der prozentuale Anteil der Merkmalsausprägung x an der gesamten Stichprobe.

1.3 Darstellung von Daten

Statistische Daten können auf unterschiedliche Weise dargestellt werden. Die simpelste Darstellung ist eine Datentabelle. Die Tabelle ist aber sehr abstrakt, weshalb meistens Diagramme verwendet werden: Kreisdiagramme, Tortendiagramme, Histogramme und Prozentstreifen.

1.4 Einteilung in Klassen

Bei der Ermittlung der Häufigkeit von Merkmalen ist es oft nicht sinnvoll, jede Merkmalsausprägung einzeln zu betrachten. Insbesondere bei stetigen metrischen Merkmalen – wie zum Beispiel der Körpergrösse – teilt man deshalb die Merkmalsausprägungen in so genannte Klassen ein.

144	150	154	154	160	160	162	162	163	164
164	164	164	165	167	167	168	169	170	171
171	172	172	172	173	174	175	176	176	176
177	178	179	179	182	182	182	182	184	185
186	187	187	188	189	190	190	191	193	205

Da in dieser Urliste viele Werte nur ein- oder zweimal vorkommen, macht es viel mehr Sinn, die beinahe liegenden Werte in Klassen zusammen zu tragen. Es werden 7 Klassen erstellt, mit jeweils 10cm Abstand (Klassenbreite).

Körpergrösse (in cm)	Anzahl
140-149	1
150-159	3
160-169	14
170-179	16
180-189	11
190-199	4
200-209	1

Es ist eine Frage der Übersicht und Ästhetik, in wie viele Klassen eine Datenmenge unterteilt werden soll. Üblicherweise wird folgende Faustregel angewendet:

Die Einteilung einer Stichprobe vom Umfang n in Klassen soll in der Regel so erfolgen, dass für die Anzahl k der Klassen gilt:

$$k \approx \sqrt{n}\ aber\ k \leq 20$$

1.5 Das Summensymbol

In der Mathematik und im Speziellen in der Statistik kommt es oft vor, dass man viele ähnliche Terme addieren muss. Das Summensymbol fasst solche Terme zusammen. Möchten wir n Summanden addieren, so lässt sich die Summe mit dem Summensymbol abgekürzt schreiben als:

$$\sum_{i=1}^{15} i = 1+2+3+4+5+6+7+8+9+10+11+12+13+14+15$$

1.6 Zentralmasse

Für statistische Untersuchungen ist das «Zentrum» der Urliste von besonderer Bedeutung. Solche «Zentren» werden als Zentralmasse bezeichnet. Das wohl bekannteste Zentralmass ist das arithmetische Mittel, das uns zum Beispiel als Notendurchschnitt in der Schule begegnet:

Das arithmetische Mittel x (Mittelwert, Durchschnitt) einer Urliste, bestehend aus den n Zahlen x, ist definiert durch

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n} * \sum_{i=1}^{n} x_i$$

Die ist aber nicht die einzige Möglichkeit, ein «Zentrum» der Urliste zu definieren. Stattdessen können wir auch den häufigsten Wert oder denjenigen Wert, der genau in der Mitte der sortierten Urliste liegt, als «Zentrum» definieren. Der Wert, der am häufigsten vorkommt, heisst Modus (oder Modalwert) und der Wert, der genau in der Mitte liegt, heisst Median (oder Zentralwert).

Der Modus ist vor allem bei qualitativen Merkmalen sinnvoll. Der Modus ist aber nicht eindeutig: Es ist durchaus möglich, dass es zwei häufigste Werte gibt.

Sind die Zahlen einer Zahlenliste der Grösse nach geordnet und ist n ungerade, so heisst der Wert in der Mitte der Liste Media. Ist n gerade, so ist der Median das arithmetische Mittel der beiden Werte in der Mitte der Liste.

1.7 Streuungsmasse

Die Zentralmasse reicht nicht aus, um die Verteilung bzw. die Streuung der Daten um den Mittelwert zu beschreiben. Ein sehr einfaches Streuungsmass ist die so genannte Spannweite:

Ist x_{min} der kleinste und x_{max} der grösste Wert einer Urliste, so bezeichnet man die Grösse $x_{max} - x_{min}$ als Spannweite der Urliste.

Die Spannweite gibt also die Länge des gesamten Bereichs an, über den sich die Urliste erstreckt. Allerdings sagt die Spannweite nichts darüber aus, ob der Grossteil der Datenwert um den Mittelwert konzentriert ist oder ob die Werte gleichmässig über die ganze Spannweite verstreut sind.

Eine Möglichkeit wäre das arithmetische Mittel aller Abweichungen vom Mittelwert zu ermitteln. Jedoch erhalten wir hierbei ein sehr unbefriedigendes Ergebnis: 0

Der Grund für dieses Ergebnis ist, dass die Abweichungen sowohl positiv als auch negativ werden können und dass sich die gesamten positiven und die gesamten negativen Beträge gegenseitig aufheben. Dies könnten wir vermeiden, wenn wir stattdessen die Beträge der Abweichungen mitteln, sodass alle Beträge positiv sind und sich deshalb nicht gegenseitig aufheben können. Das Rechnen mit Beträgen ist aber sehr umständlich, Stattdessen werden die (ebenfalls positiven) Quadrate der Abweichungen gemittelt. Die Zahl σ^2 heisst (theoretische) Varianz. Weil hier die Abweichungen der Datenwerte vom Mittelwert quadriert werden, hat die Varianz nicht die Einheit der Datenwerte: Sind die Datenwerte bspw. Körpergrössen, angegeben in cm, so hat die Varianz die Einheit cm^3. Deshalb wird in der Regel nicht die Varianz, sondern die so genannte Standardabweichung betrachtet: $\sqrt{\sigma^2} = \sigma$.

1.8 Normalverteilungen

Es scheint in der Natur zahlreicher Merkmale zu liegen, dass die Merkmalsausprägungen einer angemessenen Stichprobe mehr oder weniger symmetrisch und glockenförmig um den Mittelwert verteilt werden.

Dazu gehören beispielsweise die Körpergrösse und das Gewicht von Probanden, der I+, die Füllmenge von Mineralwasserfalschen, Messfehler in Labors etc. Eine derartige Verteilung wird jeweils dort beobachtet, wo aus Erfahrung eine Art Norm erwartet wird. Bei der Körpergrösse werden wir bestimmt nicht erwarten, einen 10cm oder 5m hohen Menschen zu finden. Stattdessen erwarten wir auf Grund unserer Erfahrung, dass die meisten Werte von Erwachsenen in einem Normbereich um 160-190 cm liegen und dass die Häufigkeit nach aussen hin abnimmt. Derartige Verteilungen werden Normalverteilungen (oder auch Gaussverteilungen) genannt. Sie werden näherungsweise durch eine nach Carl Friedrich Gauss benannte Gaussfunktion beschrieben, deren Graph eine glockenförmige Kurve ist.

Eine praktische Besonderheit der Normalverteilung ist der enge Zusammenhang zwischen der Gaussfunktion und der Standardabweichung:

Entspricht die Urliste einer Normalverteilung, so gilt Näherungsweise:

68.3% aller Werte liegen zwischen $x = \mu - \sigma$ und $x = \mu + \sigma$

95.5% aller Werte liegen zwischen $x = \mu - 2\sigma$ und $x = \mu + 2\sigma$

99.7% aller Werte liegen zwischen $x = \mu - 3\sigma$ und $x = \mu + 3\sigma$

Dabei ist μ das arithmetische Mittel der Grundgesamtheit und σ die (theoretische) Standardabweichung.

1.9 Quartile

Die Zentral- und Streuungsmasse geben nur wenig Einblick in die Verteilung der Merkmalsausprägungen einer Urliste. Insbesondere dann, wenn diese nicht normalverteilt sind, ist es nützlich weitere Kenngrössen zur Hand zu haben, die etwas mehr Aufschluss über die Verteilung geben. Eine sehr einfache und schnelle Möglichkeit bietet die Verwendung von sogenannten Quartilen:

Die Quartile teilen eine sortierte Urliste in vier Abschnitte, sodass in jedem Abschnitt nahezu 25% der Daten enthalten sind. Das zweite Quartil ist der Median der gesamten Urliste, das erste Quartil ist der Median in der ersten Hälfte und das dritte Quartil ist der Median der zweiten Hälfte. Die Quartile können gut mit Hilfe von Boxplots dargestellt werden.

2 Zentral- und Streumasse

2.1 Zentralmasse

2.1.1 Modus

Das arithmetische Mittel befasst sich auf eine mögliche Ausprägung, sprich, einem Wert. Mit dem Modus wird der am häufigsten vorkommenden Wert beschrieben, der Modus kann bereits für nominal skalierte Merkmale bestimmt werden.

2.1.2 Median

Bei Werten mit ordinalem Merkmal, nicht metrischem Merkmal, kann nicht gerechnet werden. Die Urliste kann jedoch der Grösse nach geordnet werden. Der Mittelwert oder auch Median ist der Wert, welcher in der sortierten Urliste genau in der Mitte liegt.

2.1.3 Arithmetisches Mittel

Das arithmetische Mittel berechnet sich als Summe aller Werte dividiert durch die Anzahl der Werte. Das arithmetische Mittel kann nur von metrisch skalierten Daten berechnet werden.

2.1.4 Nutzung

Das Skalenniveau schränkt die Möglichkeit, einen Mittelwert zu wählen ein. Trotzdem kann die Frage auftauchen welchen Mittelwert man benutzen soll. Speziell bei metrisch skalierten Daten ist die Frage, ob das arithmetische Mittel oder ob der Median verwendet werden sollte, nicht immer eindeutig zu beantworten. Beim bspw. der Anzahl Personen sollte es klar sein, dass «Bruchteile von Personen» sachlich sinnlos sind. Das spricht bei solchen Fällen für den Median, denn dieser liefert tatsächliche Werte. Das heisst, dass der Median die Menge der Messwerte nicht verletzt. Mit dem arithmetischen Mittel kann jedoch besser zurückgerechnet werden als mit dem Median. Möchte man aus einer Stichprobe Hinweise auf die Grundgesamtheit gewinnen, dann ist meistens das arithmetische Mittel besser, weil es in gewissem Sinn die Informationen der Stichprobe besser ausnützt. Wichtig ist auch dass das arithmetische Mittel stärker auf «Ausreisser» reagiert. Der Median weniger. Ist bei einem «Ausreisser» klar, dass es sich um einen Messfehler handelt, kann man ihn gut weglassen, jedoch ist es bei Urlisten mit unklaren «Ausreissern» besser den Median zu verwenden.

Manchmal trifft man aber auch eine Mischung aus Median und arithmetischem Mittel an. Beim Sport werden oft bei teilweise subjektiven Bewertungen zuerst die beste und schlechteste Note gestrichten und erst dann das arithmetische Mittel berechnet. Hier geht man davon aus, dass alle korrekten Bewertungen, relativ nahe beieinander liegen müssen. Fehler fallen mit diesem System weniger ins Gewicht. Manchmal sind aber auch Daten «natürlich» stark ungleich verteilt: Der Lohn ist ein gutes Beispiel dafür. Ein paar wenige Manger verdienen ein Vielfaches von dem, was die grosse Mehrheit verdient. Das Mittel des Lohnes ist dann oft viel höher als das Einkommen der überwiegenden Mehrheit der Angestellten.

2.1.5 Andere Mittelwerte

Manchmal ist aber keiner der vorgekommenen Mittelwerte geeignet. Bspw. bei Zinssätzen, der durchschnittliche Zinssatz kann über die Jahre 2% betragen, obwohl der Endwert bei 2% nicht stimmen würde. In solchen Situationen können andere Mittelwerte geeigneter sein, bspw. das geometrische Mittel und das harmonische Mittel.

2.1.6 Gewichtetes arithmetisches Mittel

Kommen bei der Berechnung des arithmetischen Mittels gewisse Werte mehrfach vor und kennt man die absoluten Häufigkeiten, dann lässt sich das arithmetische Mittel bequemer wie folgt bestimmen. Kommen in einer Urliste die verschiedenen metrischen Ausprägungen mit den jeweiligen der absoluten Häufigkeiten vor, dann gilt:

$$\bar{x} = \frac{H_1 * X_1 + H_2 * X_2 + \cdots + H_k * x_k}{n}$$

Wobei $n = H_1 + H_2 + \cdots + H_k$ der Umfang der Stichprobe ist. Dieser Durchschnitt, der heisst gewichtetes arithmetisches Mittel, da jeder Wert gemäss seiner absoluten Häufigkeit gewichtet wird. Diese Art der Berechnung ist besonders praktisch nach einer Klasseneinteilung.

2.2 Streumasse

Durchschnitte gibt es, um viele Werte möglichst gut durch einen einzelnen Wert darzustellen. Dass der Durchschnitt allein nicht immer die gewünschte Aussagekraft hat, wird den meisten bewusst sein. Offensichtlich sagt der Durchschnitt nichts über die Schwankungen innerhalb der Messwerte aus. Zusätzlich zum Durchschnitt wäre noch eine weitere Kennzahl praktisch. Eine, die uns Vorstellung davon gibt, wie gross die Schwankungen sind. Eine solche Kennzahl nennt man Streumass.

2.2.1 Mittlere absolute Abweichung

2.2.2 Standardabweichung

3 Grafische Darstellungen

3.1 Skalenniveaus und Klassen

3.2 Tabellen und Grafiken

3.3 Säulen und Balken

3.4 Kreise und Torten

3.5 Punkte

3.6 Linien und Kurven

3.7 Sonstiges