1 Beschreibende Statistik

1.1 Auftrag: Grundbegriffe

Lernziele für diesen Auftrag sind:
Ich kann

- folgende Begriffe anhand eines Beispiels erklären: Absolute und relative Häufigkeit, Modus, Median/Zentralwert, Stichprobe, Stichprobenumfang, empirischer Mittelwert, empirische Varianz und empirische Standardabweichung.
- die absolute Häufigkeit eines Wertes bestimmen
- die relative Häufigkeit eines Wertes bestimmen
- den Median/Zentralwert eines reell-wertigen Merkmals bestimmen
- den empirischen Mittelwert eines reell-wertigen Merkmals bestimmen
- empirische Varianz und Standardabweichung eines reell-wertigen Merkmals bestimmen

Endprodukte dieses Auftrags sind: Urliste, Strichliste, absolute Häufigkeiten, relative Häufigkeiten, zwei Histogramme, Median, Modus/Modi, zwei Schätzer für den Mittelwert, eine Notenverteilung, zwei Schätzer für die Varianz respektive die Standardabweichung. Die Begriffe und Aufträge sind im Abschnitt 1.1.3 aufgeführt.
Material: Vorlagenblatt für Urliste, Strichliste und Histogramme und ein Würfel

1.1.1 Einführung

Beim Spiel «Eile mit Weile» würfelt jeder Spieler mit einem Würfel. Die Augenzahl bestimmt, wie viele Felder man mit einer Spielfigur vorrücken darf. Dabei gilt folgende 6er-Regel: Bei einer Sechs darf man noch einmal würfeln. Würfelt man aber dreimal hintereinander eine Sechs, so muss man mit allen Figuren zurück an den Start. Man kann also maximal 6+6+5=17 Felder vorrücken.

1.1.2 Zufallsexperiment

Sie untersuchen nun die Verteilung der Augensumme X. x ist dabei die Anzahl Felder, die man vorrücken kann. Für den Fall von drei Sechsen setzen wir $x = 0$. D.h. der Wertebereich liegt zwischen 0 und 17. Führen Sie das Würfelexperiment 100 Mal durch. Wechseln Sie sich dabei ab. Tragen Sie die Resultate in die Urliste ein.

1.1.3 Begriffe

Absolute Häufigkeit	Die absolute Häufigkeit n_7 des Ergebnisses $x = 7$ ist die Anzahl Striche. Tragen Sie diese in der Spalte n_x ein. Die totale Anzahl Striche, d.h. die Summe $n = \sum_{x=0}^{17} n_x$ sollte natürlich 100 sein.
Relative Häufigkeit	Die relative Häufigkeit ist definiert als die absolute Häufigkeit dividiert durch die Anzahl Experimente. $h(x) = n_x/n$ Berechnen Sie für alle x=0 bis 17 die relative Häufigkeit und tragen diese auf zwei Nachkommastellen gerundet in der Strichliste in die Spalte h(x) ein. In der Spalte «h(x) in %» können Sie die relative Häufigkeit in Prozent eintragen. Die Summe über alle relativen Häufigkeiten ist 1, respektive 100%.

Histogramm

Die Strichliste gibt uns einen guten, graphischen Überblick über die Verteilung der gewürfelten Ergebnisse x. Das ist die Idee eines Histogramms. Zeichnen sie für jeden x-Wert in untenstehendem Diagramm eine Säule der Höhe n_x (Breite zwei Häuschen). Die Fläche jeder Säule ist proportional zur relativen Häufigkeit des entsprechenden Wertes.

Klassenbildung

Nicht immer ist es sinnvoll, für jeden Wert einzeln eine Säule einzuzeichnen. Wenn die interessierende Variable x eine beliebige reelle Zahl sein kann, müssten wir ja unendlich viele Säulen zeichnen. Dann teilt man die Werte in Klassen ein. In unserem Würfelexperiment wäre zum Beispiel folgende Klasseneinteilung sinnvoll:

$$\{0\}, \{1,2,3,4,5\}, \{6\}, \{7,8,9,10,11\}, \{12\}, \{13,14,15,16,17\}$$

Zeichnen Sie für jede Klasse eine Säule in untenstehendes Histogramm. Die Säulenfläche soll dabei proportional zur relativen Häufigkeit sein. Für x in der Klasse {13,14,15,16,17} ist die absolute Häufigkeit gleich $\sum_{x=13}^{17} n_x$. Dividiert man diese Zahl durch die Anzahl Experimente (also 100), so erhält man die relative Häufigkeit. Da die Säule fünf Einheiten breit ist, müssen wir die relative Häufigkeit durch 5 teilen, um die Höhe der Säule zu erhalten. Die Fläche der Säule ist ja Breite mal Höhe.

Median

Der Median oder Zentralwert ist derjenige Wert x_{Med} der an der mittleren (zentralen) Stelle steht, wenn man die Werte aufsteigend sortiert. Für eine ungerade Anzahl Werte ist der Median damit eindeutig festgelegt. Für ein Würfel-Experiment der Urliste {5, 2, 5, 9, 3, 5, 4, 2, 2, 0, 4} ist der Median zum Beispiel $x = 4$: Wieso? Wenn wir die Werte aufsteigend sortieren, so steht in der Mitte eine 4. $0, 2, 2, 2, 3, \mathbf{4}, 4, 5, 5, 5, 9$

Bei einer geraden Anzahl Werte gibt es keinen Wert genau in der Mitte. Man definiert den Median deshalb als das arithmetische Mittel der beiden Werte, welche nahe der Mitte liegen. Für die Urliste {5, 2, 5, 9, 3, 5, 4, 2, 2, 0} ist der Median $\tilde{x} = 3.5$ Wieso? In der sortierten Urliste {0, 2, 2, 2, 3, 4, 4, 5, 5, 5} stehen 3 und 4 in der Mitte; das arithmetische Mittel ist $\frac{3+4}{2} = 3.5$.

Bestimmen Sie den Median für ihre Urliste. Tipp: Anstatt die hundert Zahlen aufsteigend zu sortieren, können Sie die bereits berechneten die absoluten Häufigkeiten berücksichtigen. Welche Werte stehen in der sortierten Urliste an den Stellen 48,49,50,51,52,53 ? Der Median unserer Urliste ist x_{Med}:

Modus

Der Modus oder Modalwert ist derjenige Wert, welcher am häufigsten angenommen wird. Der Modus ist nicht immer eindeutig. Gibt es genau einen häufigsten Wert, so sagt man die Verteilung der Werte sei unimodal. Die Verteilung aus obigem Beispiel ist bimodal, denn die Werte 2 und 5 kommen beide dreimal vor. Das entsprechende Histogramm hat zwei Höcker wie ein Kamel. Verallgemeinernd sprechen die Statistiker auch dann von einer bimodalen Verteilung, wenn das Histogramm zwei unterschiedlich hohe aber deutlich ausgeprägte Höcker hat.

Unsere Verteilung hat folgenden Modus, respektive folgende Modi:

Mittelwert

Der Mittelwert ist definiert als das arithmetische Mittel über die Gesamtheit der Werte. Für n Werte ist der Mittelwert $\bar{x} = \sum_{i=1}^{n} x_i$. Nun ist es äusserst mühsam den Mittelwert über alle hundert Zahlen zu berechnen. Wir schätzen deshalb den Mittelwert, indem wir über eine Stichprobe mitteln. Geschätzte Werte kennzeichnet man mit einem Hut $\widehat{\ }$. So ist eine Schätzung für den Mittelwert als $\widehat{\bar{x}}$ geschrieben. Beachten Sie, dass Schätzer von der Stichprobe abhängig sind. Verschiedene Schätzungen derselben Variable können sich deshalb unterscheiden. Wir werden später untersuchen, mit welcher Unsicherheit Schätzer behaftet sind (Stichwort: Stichprobenverteilung). Der Schätzer für den Mittelwert wird auch als Stichprobenmittel oder als empirischer Mittelwert bezeichnet. «Empirisch» heisst aus Erfahrung, durch Beobachtung gewonnen. Der Hut wird oft weggelassen.

Berechnen Sie den Mittelwert über die ersten zwanzig Werte $x_1, \ldots, x_{20}$, d.h. $\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i$ und den Mittelwert über die zweiten zwanzing Werte $x_{21}, \ldots, x_{40}$, d.h. $\bar{x} = \frac{1}{20} \sum_{i=21}^{40} x_i$

Lagemasse

Median, Modus und Mittelwert sind sogenannte Lagemasse. Sie sagen etwas über die Position des Histogramms auf der x-Achse. Für eine symmetrische Verteilung fallen die drei Lagemasse zusammen. Das ist aber in der Praxis äusserst selten der Fall. Lassen Sie sich deshalb durch die häufig vereinfachten Darstellungen mit einer glockenförmigen Verteilung nicht irreleiten. Median und Mittelwert sind unterschiedliche Lagemasse.

Wenn in einer Firma mit 10 Mitarbeitern der Chef eine Million und alle anderen Fr. 50'000 pro Jahr verdienen, ist der Mittelwert Fr. 145'000. Trotzdem erhalten die meisten Mitarbeiter nicht einmal die Hälfte des Mittelwertes. Ein Klassenschnitt von 4.3 heisst noch lange nicht, dass die meisten Schülerinnen und Schüler eine genügende Note hatten. Konstruieren Sie ein Beispiel mit Mittelwert 4.3 (oder höher) und Median unter 4 !

Varianz und Standardabweichung

Um eine Vorstellung über die Streuung der Werte zu erhalten, berechnen wir für alle Resultate x_i die Differenz $x_i - \bar{x}$ zum Mittelwert. Diese kann positiv (für $x_i > x$) oder negativ (für $x_i < x$) sein. Wir nehmen deshalb das Quadrat und mitteln dann: $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ Die Grösse s^2 heisst Varianz der Grundgesamtheit. Die Wurzel s aus der Varianz ist die sogenannte Standardabweichung der Grundgesamtheit. Da die Berechnung der Varianz über alle hundert Werte viel zu aufwendig wäre, schätzen wir die Varianz anhand einer Stichprobe von zwanzig Werten.

$$\hat{s^2} = \frac{1}{20} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Dazu bräuchten wir aber den Mittelwert $\bar{x}$ welchen wir auch nicht kennen. Wir müssen den Mittelwert $\bar{x}$ deshalb durch unseren Schätzer $\hat{x}$ ersetzen. Es zeigt sich nun, dass bei Verwendung von $\hat{s}^2$ die Varianz systematisch unterschätzt wird und zwar um den Faktor (20-1)/20. Man muss die Formel deshalb entsprechend anpassen:

$$\hat{s^2} = \frac{1}{20 - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Empirische Varianz und Standardabweichung

In der Praxis hat man fast nie Zugriff auf die Grundgesamtheit. Die ganze Grundgesamtheit zu untersuchen ist entweder zu teuer oder schlicht sinnlos, weil die Ware beim Testen zerstört wird. Denken Sie zum Beispiel an die Bestimmung der Lebensdauer von Lampen. In Formelsammlungen finden Sie deshalb immer die sogenannte empirische Varianz

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

respektive die empirische Standardabweichung

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

welche sich auf eine Stichprobe beziehen. n ist dabei der *Stichprobenumfang*, d.h. die Anzahl untersuchter Elemente aus der Grundgesamtheit. Berechnen Sie für die Stichproben $x_1, \cdots, x_{20}$ und $x_{21}, \cdots, x_{40}$ einen Schätzer $\widehat{s^2}$ für die empirische Varianz und $\sqrt{\widehat{s^2}}$ für die empirische Standardabweichung. Seien sie vorsichtig mit der Klammerung, damit der Taschenrechner Sie nicht anlügt!

Streuintervalle

Die Standardabweichung ist deshalb ein wichtiges *Streumass*, weil sich die Werte für viele Verteilungen im Intervall $[\bar{x} - s, \bar{x} + s]$ sammeln. Bestimmen Sie für ihre Verteilung, welcher Anteil (in Prozent) im Intervall $[\bar{x} - s, \bar{x} + s]$, respektive im Intervall $[\bar{x} - 2s, \bar{x} + 2s]$ liegt. Benutzen Sie dazu ihre Schätzer für $\bar{x}$ und s.