

Dr. Quapp: Statistik für Mathematiker mit SPSS

Lösungs-Hinweise 1. Übung

– Beschreibende Statistik & Verteilungsfunktion –

1. Die folgende Tabelle enthält die Pulsfrequenz einer Versuchsgruppe von 39 Personen:

88	76	84	64	60	64	60	64	68	74	68	68	72
76	72	52	72	64	60	56	72	88	80	76	64	72
60	76	88	72	64	60	60	72	92	80	72	64	68

a) Geben Sie eine Tabelle mit den absoluten und relativen Häufigkeiten an und bestimmen Sie statistische Maßzahlen wie Mittelwert, Median, Modalwert, Spannweite, sowie Streuungsmaße wie Varianz, St.Abweichung, Schiefe, Wölbung.

– Eintippen der Daten in einer Spalte des SPSS Datenfensters

auf Var0001 öffnet Fenster, welches Änderung des Variablennamens z.B. in puls gestattet.

öffnet Fenster, welches die deskriptive Statistik ermöglicht. Schiebe puls in rechtes Fenster, und klicke an: Mittelwert, Median, Modalwert, Spannweite, sowie Streuungsmaße wie Varianz, St.Abweichung, Schiefe, Wölbung sind nun nur noch anzuklicken und das Befehlsfenster auszuführen.

Es ist der Erwartungswert

$$E x = \frac{1}{n} \sum_1^n x_i$$

und die Varianz ist die durchschnittliche quadrierte Abweichung vom Erwartungswert:

$$Var x = \sigma^2 = \frac{1}{n} \sum_1^n (x_i - E x)^2$$

wobei aber SPSS nicht diesen Wert als Varianz berechnet, sondern die sogenannte empirische Varianz

$$s^2 = \frac{1}{n-1} \sum_1^n (x_i - E x)^2$$

Also kann $\sigma^2 = (n-1)/n s^2$ berechnet werden.

Die Schiefe und Wölbung werden dann aus den höheren Momenten

$$\mu_k = E(x - Ex)^k$$

berechnet: (Wobei leicht zu sehen ist, daß $\mu_1 = 0$ und $\mu_2 = \sigma^2$ ist.)

$$\text{Schiefe} = \mu_3 / (\sqrt{\mu_2})^3$$

wobei Schiefe < 0 bedeutet, daß die Dichte linksschief ist im Vergleich zur Gaussfunktion mit gleichen Ex und Var x .

$$\text{Woelbung} = \text{Exzentr.} = \text{Kurtosis} = \mu_4 / (\sqrt{\mu_2})^3 - 3$$

wobei Wölbung < 0 bedeutet, daß die Dichte stumpfer ist als die Gaussfunktion mit gleichen Ex und Var x, und > 0 bedeutet dann spitzer.

b) Zeichnen Sie Linien-, Balken-, Kreis- und Fächendiagramme.

Unter stehen die entsprechenden Befehle bereit. Beachten Sie, daß entweder Kategorien oder Einzelwerte über der x-Achse dargestellt werden sollten!

.

Zeichnen Sie das Histogramm: Unter stehen der entsprechende Befehle bereit. Selbstständig wird von SPSS eine Klasseneinteilung vorgenommen.

Soll diese geändert werden, so , und es erscheint neues Fenster:

Es erscheint neues Fenster, in dem der Benutzer mit dem Punkt weitergeführt wird: Die gewünschte Achseneinteilung kann eingestellt werden.

Bei einem etwas "nicht-normalverteiltem" Histogramm sollte man mehrere Varianten der Klasseneinteilung probieren!

c) Gruppieren Sie die Daten in 5 Klassen und zeichnen Sie erneut Histogramm, Linien-, Balken-, Kreis- und Fächendiagramme.

Zuerst muß eine Klassifizierungsvariable erstellt werden; dazu gibt es zwei Möglichkeiten:

-> Klick auf Transformieren -> Klick auf Berechnen

ermöglicht unter Benutzung der **fallweisen Berechnung** eine selektive Besetzung einer Klassifizierungsvariablen.

-> Klick auf Transformieren -> Klick auf Umkodieren -> Klick auf In andere Variablen

ermöglicht die selektive Besetzung einer Klassifizierungsvariablen in einem Fenster: Variable puls ist ins mittlere Fenster zu schieben, und ein Name für die Ausgangsvariable zu wählen. Der Button **Alte und neue Werte** ermöglicht die selektive Zuweisung von Werten zur Klassifizierungsvariablen. Nach einer Zuweisung darf nicht der Ausführungs-Befehl **Hinzufügen** vergessen werden!

(Bemerkung: -> Klick auf Transformieren -> Klick auf Variable kategorisieren

ermöglicht die automatische Besetzung einer Variablen, die Percentile ergibt, also Klassen etwa gleich vieler Werte. Dies ist hier nicht gemeint.)

Unter **-> Klick auf Daten -> Klick auf Aggregieren** kann ein neues Datenfile erzeugt werden, das dann die gruppierten Daten beschreibt. Man schiebt dazu die Klassifizierungsvariable in die Zeile der Break-Variablen, sowie Variable puls ins Klassifizierungsfenster, und klickt an daß die Fallzahl je Break-Gruppe gespeichert werden soll, und daß eine neue Datei erzeugt werden soll. Voreingestellt heißt diese Datei AGGR.sav, der Name kann in einem Fenster geändert werden. (Eine Labelvariable kann extra zur Beschreibung der Klassifizierungsvariablen verwendet werden. Diese erscheint dann z.B. in Grafiken.)

d) Bestimmen Sie für die Ausgangsdaten die empirische Verteilung und zeichnen Sie diese in einem Koordinatensystem.

Die Tabelle mit den relativen Häufigkeiten r_i ist zu verstehen als die diskrete Dichte der Wahrscheinlichkeiten des Auftretens entsprechender Ausprägungen der Puls-Variable X (der Kategorie).

$$P(X = x_i) = r_i$$

Die kumulative Summe dieser relativen Häufigkeiten ist dann die empirische Verteilungsfunktion:

$$empVert(x_j) = \sum_{i=1}^j r_i = P(X \leq x_j)$$

Wir erzeugen nun diese r_i :

-> Klick auf Daten -> Klick auf Sortieren

ermöglicht das Sortieren von puls z.B. aufsteigend (ist voreingestellt). Wir berechnen noch eine Hilfsvariable Hilfe=1/39. Mit

-> Klick auf Transformieren -> Klick auf Zeitreihen

-> Klick auf andere Funktion einstellen -> Klick auf kumulative Summe, und Variable Hilfe in oberes rechtes Fenster schieben: Dies ergibt eine neue Variable, in der die 1/39-Werte nun kumulativ aufaddiert sind. Dies ist die empirische Verteilungsfunktion.

Mit -> Klick auf Graphik -> Klick auf Balkendiagramm kann die empirische Verteilungsfunktion dargestellt werden, wobei die x-Achse die Kategorie des Pulses sein sollte! Als Ordinate soll die berechnete empirische Verteilung dienen. Dies muß im Fenster als spezielle Funktion eingestellt werden.

2. Die Zufallsvariable X sei die Augenzahl eines ausgespielten Würfels mit den möglichen Realisationen der Augen zu $\{0, 1, 2, 3, 4, 5\}$. Der Würfel sei "ehrlich", d.h. es gelte $P(X = k) = 1/6$, $k = 0, \dots, 5$.

Die Zufallsvariable Y sei das Resultat von fünf Münzwürfen, wobei jeweils die "Zahl" gezählt werde. Offenbar hat Y die gleichen Realisationen. Man gehe davon aus, daß Y binomialverteilt ist mit $p=1/2$, also

$$P(Y = k) = \binom{5}{k} \frac{1}{2^5}, \quad k = 0, \dots, 5.$$

Vergleichen Sie die Dichten graphisch.

– Eintippen der Daten: Realisierung, $P(X=k)$, und $P(Y=k)$ in je einer Spalte des SPSS Datenfensters

-> Doppelklick auf Var0001, Var0002 oder Var0003 öffnet je ein Fenster, welches Änderung des Variablennamens z.B. in real, px, py gestattet. Dabei kann beim Eintippen $px=1$, und $py=\{1,5,10,10,5,1\}$ gesetzt werden, und in -> Klick auf Transformieren, -> Klick auf Berechnen kann dann $px/6$ bzw. $py/32$ berechnet werden.

Unter -> Klick auf Graphik und -> Balkendiagramm können beide Dichten gleichzeitig gezeigt werden.

Berechnen Sie Mittelwert und Streuung für X und für Y. (Stimmt die 3 σ -Regel?)

Es ist der Erwartungswert von X

$$E x = \frac{1}{6}(0 + 1 + 2 + 3 + 4 + 5) = 2,5$$

und die Varianz

$$Var x = \sigma_x^2 = \sum_{i=1}^n p_{x_i} (x_i - E x)^2 = \frac{1}{6}(2 \cdot 5^2 + 1 \cdot 5^2 + 0 \cdot 5^2 + 0 \cdot 5^2 + 1 \cdot 5^2 + 2 \cdot 5^2) = 2,92$$

und der Erwartungswert von Y

$$E y = \frac{1}{32}(1 \cdot 0 + 5 \cdot 1 + 10 \cdot 2 + 10 \cdot 3 + 5 \cdot 4 + 1 \cdot 5) = 2,5$$

und die Varianz

$$Var y = \sum_{i=1}^n p_{y_i} (y_i - Ey)^2 = \frac{1}{32} (1 \cdot 2, 5^2 + 5 \cdot 1, 5^2 + 10 \cdot 0, 5^2 + 10 \cdot 0, 5^2 + 5 \cdot 1, 5^2 + 1 \cdot 2, 5^2)$$

also

$$Var y = \sigma_y^2 = 1,25$$

Dann ist $\sigma_x = 1,71$ und $\sigma_y = 1,12$. Je 1σ vom Erwartungswert entfernt liegen bei der Gleichverteilung des Würfels $2/3$ der Werte, und bei der Binomialverteilung von fünf Münzwürfen $20/32$ der Werte. Je 2σ vom Erwartungswert entfernt liegen bei X schon alle Werte, bei Y dann $30/32$ der Werte, und im Intervall 3σ liegen auch bei Y alle Werte. Die 3σ -Regel stimmt also in guter Näherung.

Nimmt man für X je 6 Versuche an, und für Y je 32 Versuche, so kann man mittels SPSS mit Gewichten arbeiten. Vergleichen Sie so berechnete Mittelwerte und Streuungen.

Mit -> Klick auf Analysieren -> deskr.Stat -> Häufigkeiten öffnet sich Fenster, welches die deskriptive Statistik ermöglicht. Schiebe Variable real in rechtes Fenster, und klicke -> Statistik an: jetzt nur Mittelwert und Varianz anklicken.

Das Resultat ist $E x = \text{Mittelwert}_x = 2,5$, aber unter Varianz erscheint der Wert 3,5. !?!

Da wir hier direkt mit 6 Werten der Realisierung arbeiten, ist $n=6$ die Zahl der bearbeiteten Fälle, und SPSS gibt s_x^2 aus, d.h. um σ_x^2 zu erhalten, müssen wir mit $5/6$ multiplizieren: das ergibt das schon bekannte Resultat von 2,92.

Für Y erzeugen wir eine neue Spalte mit "Gewichten", die größer als Eins sein sollen. Also kann die Spalte py mit 32 multipliziert werden. Unter

-> Daten und -> Fälle wichten

kann nun diese Variable als Gewicht für weitere Berechnungen verwendet werden. Dabei wird so verfahren, als ob jeder Wert der Realisierung so oft auftritt, wie sein Gewicht angibt. (Man könnte also auch 32 Zeilen erzeugen, mit den entsprechenden Werten von real: einmal 0, fünfmal 1, usw.)

Mit den analogen Befehlen

-> Klick auf Analysieren -> deskr.Stat -> Häufigkeiten schiebe wieder Variable real in rechtes Fenster, und klicke -> Statistik an: Mittelwert und

Varianz anklicken. Dies gibt dann das Resultat $E y = \text{Mittelwert}_y = 2,5$, also gleich dem von X, aber unter Varianz erscheint nun der Wert 1,29 !

Da wir nun mit 32 Werten als Summe der Gewichte der Realisierung arbeiten, ist $n=32$ die Zahl der bearbeiteten Fälle, und SPSS gibt s_y^2 aus, d.h. um σ_y^2 zu erhalten, müssen wir mit $31/32$ multiplizieren: dies gibt das schon bekannte Resultat von 1.25 .

Also **Merksatz: SPSS erzeugt die sogenannte empirische Varianz !**

(Anhang:

Die Verteilungsfunktion der Binomialverteilung ist in SPSS verfügbar:

`-> Klick auf Transf. -> Klick auf Berechnen` und im Fenster Funktionen auswählen `-> CDF.BINOM(q,n,p)` . Dabei ist hier q der Wert der Realisierung, n der Exponent der Binomialformel, also $n=5$, und p die Wahrscheinlichkeit des einzelnen Bernoulliversuches, also hier $p=0.5$. Im der Berechnungszeile muß dann letztlich stehen: `CDF.BINOM(real,5,0.5)` . (CDF bedeutet cumulative density function)