Abschlussprojekt
von
Pragya Kaundal

Thema:	Statistik mit R
Dozentin:	Julianne Wawerda
Projektzeitraum:	15.02.2021 – 12.03.2021
Partner:	Franziska Wilhelm, Karl Martin Henn, Mitja Wogatzky, Torben Erichsen

Inhaltsverzeichnis

Aufgabe 1: Grundlagen

SAP vorher	2	5	2	7	5	6	1	3	7	3
SAP nachher	10	10	8	6	4	9	4	8	7	5

1) Berechne die Mittelwerte, Modus/Modi und die Mediane (SAPvorher und SAPnachher)

- Mittelwert:
  Das arithmetische Mittel (Mittelwert, engl. "mean") ist das gebräuchlichste Maß der zentralen Tendenz. Es ist gleich dem mathematischen Durchschnitt.

  $$\bar{x} = \frac{1}{10} \cdot (2+5+2+7+5+6+1+3+7+3) = 4.10$$

  Der Mittelwert für SAPvorher ist 4.1.

  $$\bar{x} = \frac{1}{10} \cdot (10+10+8+6+4+9+4+8+7+5) = 7.10$$

  Der Mittelwert für SAPnachher ist 7.1.

- Modus/Modi:
  Der Modus (Modalwert, engl. "mode") ist der häufigste Wert einer Verteilung.
  Werte ordnen: 1, 2, 2, 3, 3, 5, 5, 6,7,7.
  Die Modi für SAPvorher ist: 2, 3, 5 und 7.

  Werte ordnen: 4, 4, 5, 6, 7, 8, 8, 9, 10, 10
  Die Modi für SAPnachher ist: 4, 8 und 10.

- Median:
  Der Median (Zentralwert, engl. "median") teilt eine Stichprobe in zwei gleich große Hälften. Er ist damit das 50%-Quantil der Verteilung einer Variablen. Es liegen genauso viele Werte unter wie über diesem Wert.

$$\mathrm{Med}(X) = \begin{cases} X[\frac{n}{2}] & \text{if n is even} \\ \frac{(X[\frac{n-1}{2}]+X[\frac{n+1}{2}])}{2} & \text{if n is odd} \end{cases}$$

X = ordered list of values in data set

n = number of values in data set

Es liegt eine gerade Anzahl an Werten vor, also ist der Median gleich dem arithmetischen Mittel der beiden mittigen Werte.

median = (3+5) / 2 = 4

Der Median für SAP vorher ist 4.

median = (7+8) / 2 = 7.5

Der Median für SAP nachher ist 7.5.

2) Berechne die Varianzen und Standardabweichungen (SAPvorher und SAPnachher)

$$\text{Standardabweichung der Bevölkerung} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

$$\text{Standardabweichung der Stichprobe} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}}$$

$$SD = \sqrt{Varianz}$$

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{9}(4.1-2)^2 + (4.1-5)^2 + \ldots + (4.1-3)^2} = 2.18$$

Der Standardabweichung für SAPvorher ist 2.18.

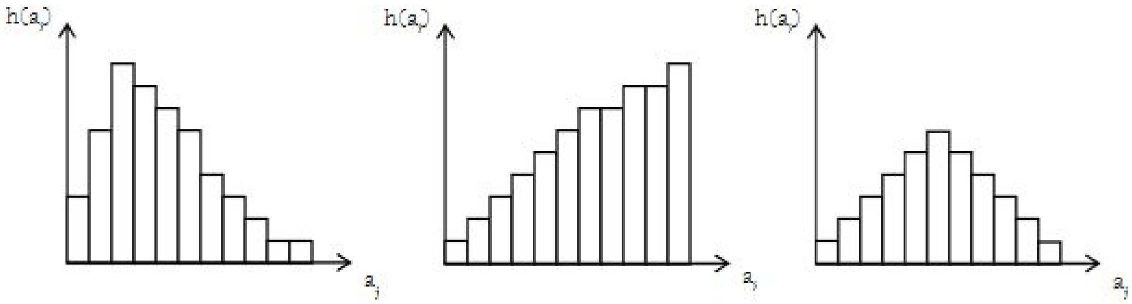
Der Varianz für SAPvorher ist 4.77.

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{9}(7.1-10)^2 + (7.1-10)^2 + \ldots + (7.1-4)^2} = 2.28$$

Der Standardabweichung für SAPnachher ist 2.28.

Der Varianz für SAPnachher ist 5.21.

3) Ist der Graph recht-, linksschief und symmetrisch?



__________________ __________________ __________________

rechtsschief linksschief symmetrisch

4) Ordne der Daten das Skalenniveau zu: (Nominal, Ordinal, Intervall, Ratio, Absolut), welcher Operation ist erlaubt.

Art der Variable	Skalenniveau	Operation
Militärdienstgrad	Ordinal	=/≠ </ >
Alter (in Jahren)	Ratio	=/≠ </ > +/ - ÷ / *
Verkehrsdichte (Auto pro min)	Ratio	=/≠ </ > +/ - ÷ / *
Geschlecht (w/m/d)	Nominal	=/≠
Fahrpreise (in Euro)	Ratio	=/≠ </ > +/ - ÷ / *
Nationalität	Nominal	=/≠
Schulbildung (Gymnasium, Realschule,...)	Ordinal	=/≠ </ >
Intelligenzquotient	Intervall	=/≠

		</ > +/ -
Studienfach (z.B. Mathe, Physik, Maschinenbau)	Nominal	=/≠
Semesterzahl (nur ganze Semester sind möglich)	Absolut	=/≠ </ > +/ - ÷ / *
Klausurpunkte	Ratio	=/≠ </ > +/ - ÷ / *
Tarifklassen bei der Kfz-Haftpflicht (Vollkasko, Teilkasko, Haftpflicht)	Ordinal	=/≠ </ >

5) Ordne den Daten die folgenden Variablen das Variablenniveau zu (stetig vs. diskret).

Nr.	Wert	Variable	
		diskret	stetig
1	Steuerklasse	◉	○
2	Geschlecht	◉	○
3	soziale Schicht	◉	○
4	Einkommenssteuer	○	◉
5	Temperatur in Kelvin	○	◉
6	Windstärke in Meter/Sekunde	○	◉
7	Körpergewicht	○	◉
8	Schulnote (1-6)	◉	○
9	Klausurpunkte	○	◉
10	Einwohnerzahl	◉	○

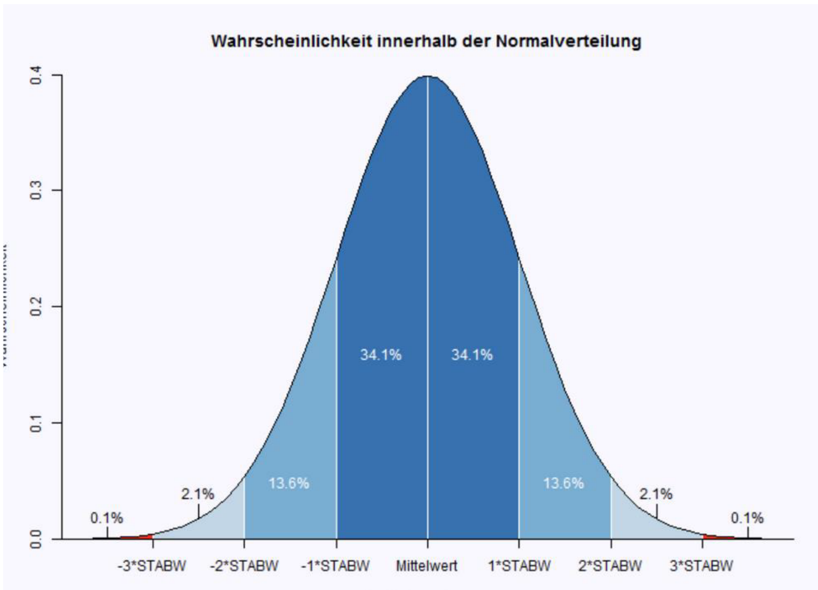
11	Semesterzahl	◉	○
12	Handelsklasse (Obst)	◉	○

6) Beschreibe in Sätzen, was der Unterschied und Gemeinsamkeiten zwischen Standardnormalverteilung und der Normalverteilung ist. Verwenden Sie die Formeln.

<u>Normalverteilung</u>

- Die Normal- oder Gauß-Verteilung (nach Carl Friedrich Gauß) ist ein wichtiger Typ stetiger Wahrscheinlichkeitsverteilungen .Sie wird auch als Gaußsche Glockenkurve bezeichnet.
- Ausprägungen der Variablen in sehr guter Näherung durch eine Normalverteilung beschrieben. Das Aussehen einer Normalverteilung ähnelt sehr einer Glocke, wobei die Funktionswerte der Kurve gegen 0 streben, wenn man die x-Werte gegen Unendlich gehen lässt.
- Normalverteilungen sind symmetrisch um den Mittelwert verteilt (d.h. die Schiefe beträgt 0).
- Beim Mittelwert besitzt die Verteilung ihr Maximum.
- Die Fläche zwischen der Kurve und der x-Achse beträgt 1.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



<u>Standardnormalverteilung:</u>

- Eine besondere Form der Normalverteilung ist die Standardnormalverteilung.
- Für sie gilt, dass der Mittelwert bei 0 liegt und die Standardabweichung bei 1, also $\mu=0$ und $\sigma=1$. Damit nimmt die Funktionsgleichung folgende Form an:

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

- Durch Standardisierung bzw. z-Transformation kann eine Normalverteilung in eine Standardnormalverteilung überführt werden. Auf diese Weise können unterschiedliche Verteilungen besser miteinander verglichen werden. Dazu setzt man als neue Variable $z = (x-\mu)/\sigma$.

Aufgabe 2: Multiple Choice

Der Sonderpunkt gilt nur innerhalb dieser Aufgabe. Maximal können Sie 10 Punkte erreichen.

1) Ein Bravais-Pearson-Korrelationskoeffizient von 0,85 deutet auf eine schwache lineare Korrelation hin.	
Richtig	**Falsch**
2) Der Interquartilsabstand (IQR) ist der doppelte Abstand zwischen Median und Modus.	
Richtig	**Falsch**
3) Die Modi lassen sich nur bestimmen, wenn eine unimodale Verteilung vorliegt.	
Richtig	**Falsch**
4) Nominalskalierte Daten können in eine natürliche Reihenfolge gebracht werden.	
Richtig	**Falsch**
5) Ausreißer wirken sich auf die Ergebnisse nicht robuster Analyseverfahren besonders stark aus.	
Richtig	Falsch
6) Die Standardabweichung berechnet sich nicht als positive Wurzel aus der Varianz.	
Richtig	**Falsch**
7) Die Kurtosis ist ein Maß für die Wölbung einer Verteilung.	
Richtig	Falsch
8) Die Berechnung der Varianz setzt mindestens metrisch skalierte Daten voraus.	
Richtig	Falsch
9) Die Spannweite ist der absolute Abstand zwischen dem kleinsten und dem größten Wert.	
Richtig	Falsch
10) Der Bravais-Pearson-Korrelationskoeffizient kann nur Werte zwischen 0 und 1 annehmen.	

Richtig	Falsch
11) Der statistische Ersatz fehlender Werte setzt mindestens metrisch skalierte Daten voraus.	
Richtig	Falsch

Aufgabe 3: Zusammenhangshypothese

Datensatz:

Var 1 =

Var 2 =

Aufgabenstellung

1) Hypothese
2) Voraussetzungen
3) Grundlegende Konzepte: Was ist Pearson?
4) Grafische Veranschaulichung des Zusammenhangs
5) Deskriptive Statistik
6) Ergebnisse der Korrelationsanalyse
7) Berechnung des Bestimmtheitsmasses
8) Berechnung der Effektstärke
9) Eine Aussage

Aufgabe 4: Unterschiedshypothese

Datensatz:

Var 1 =

Var 2 =

<u>Aufgabenstellung</u>

1) Hypothese
2) Voraussetzungen des t-Tests für unabhängige Stichproben
3) Grundlegende Konzepte: Was ist t-Test für unabhängige Stichproben?
4) Deskriptive Statistiken
5) Test auf Varianzhomogenität (Levene-Test)
6) Ergebnisse des t-Tests für unabhängige Stichproben
7) Berechnung der Effektstärke
8) Eine Aussage

Aufgabe 5: Unterschiedshypothese

Datensatz:

Var 1 =

Var 2 =

Aufgabenstellung

1) Hypothese
2) Voraussetzungen des t-Tests für abhängige Stichproben
3) Grundlegende Konzepte: Was ist t-Test für abhängige Stichproben?
4) Deskriptive Statistiken und Korrelation
5) Ergebnisse des t-Tests für abhängige Stichproben
6) Berechnung der Effektstärke
7) Eine Aussage

Aufgabe 6: Unterschiedshypothese

Datensatz:

Var 1 =

Var 2 =

<u>Aufgabenstellung</u>

1) Hypothese
2) Voraussetzungen für die einfaktoriellen Varianzanalyse ohne Messwiederholung
3) Grundlegende Konzepte: Was ist die einfaktoriellen Varianzanalyse ohne Messwiederholung
4) Deskriptive Statistiken
5) Prüfung der Varianzhomogenität (Levene-Test)
6) Ergebnisse der einfaktoriellen Varianzanalyse ohne Messwiederholung
7) Post-hoc-Tests
8) Profildiagramm
9) Berechnung der Effektstärke
10) Eine Aussage