Abschlussprojekt

von

ALI IBRAHIM YILMAZ

Thema:	Statistik mit R
Dozentin:	Julianne Wawerda
Projektzeitraum:	09.12.2020 – 11.12.2020
Partner:	Abdelrahman Barakat, Teklehaimanot Zere Aman

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)

M(SAPvorher) = (2+5+2+7+5+6+1+3+7+3) / 10 = **4.1**

Modi(SAPvorher) = (1, **2, 2, 3, 3, 5, 5,** 6, **7, 7**) → **2, 3, 5, 7**

Mediane(SAPvorher) = (1, 2, 2, 3, **3, 5**, 5, 6, 7, 7) → 3 +5 /2 = **4**

M(SAPnachher) = (10+10+8+6+4+9+4+8+7+5) / 10 = **7.1**

Modi(SAPnachher) = (4, 4, 5, 6, 7, 8, 8, 9, 10, 10) → **4, 8, 10**

Mediane(SAPnachher) = (4, 4, 5, 6, 7, 8, 8, 9, 10, 10) → 7+8 /2 = **7.5**

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

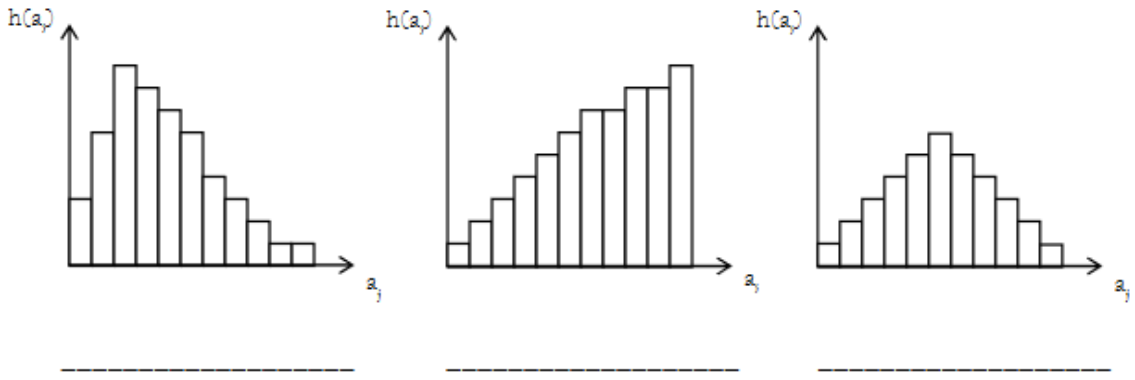
S^2 (vorher) = $[(1-4.1)^2+(2-4.1)^2+(2-4.1)^2+(3-4.1)^2+(3-4.1)^2+(5-4.1)^2+(5-4.1)^2+(6-4.1)^2+(7-4.1)^2+(7-4.1)^2]$ / (10-1) = **4.29 (varianz)**

S (vorher) = sqrt($[(1-4.1)^2+(2-4.1)^2+(2-4.1)^2+(3-4.1)^2+(3-4.1)^2+(5-4.1)^2+(5-4.1)^2+(6-4.1)^2+(7-4.1)^2+(7-4.1)^2]$ / (10-1)) = **2.07 (SD)**

S^2 (nachher) = $[(4-7.1)^2+(4-7.1)^2+(5-7.1)^2+(6-7.1)^2+(7-7.1)^2+(8-7.1)^2+(8-7.1)^2+(9-7.1)^2+(10-7.1)^2+(10-7.1)^2]$ / (10-1) = **4.689 (varianz)**

S (nachher) = sqrt($[(4-7.1)^2+(4-7.1)^2+(5-7.1)^2+(6-7.1)^2+(7-7.1)^2+(8-7.1)^2+(8-7.1)^2+(9-7.1)^2+(10-7.1)^2+(10-7.1)^2]$ / (10-1)) = **2.16 (SD)**

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



a) rechtschief b) linksschief c) symmetrisch

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

Art der Variable	Skalenniveau	Operation
Militärdienstgrad	Ordinal	(=/!=, </>)
Alter	Ratio	(=/!=, </>, +-*/)
Verkehrsdichte	Ratio	(=/!=, </>, +-*/)
Geschlecht	Nominal	(=/!=)
Fahrpreise	Ratio	(=/!=, </>, +-*/)
Nationalität	Nominal	(=/!=)
Schulbildung(Gymnasium-Real-Haupt)	Ordinal	(=/!=, </>)
Intelligenzquotient	Intervall	(=/!=, </>, +-)
Studienfach	Nominal	(=/!=)
Semesterzahl(1-8)	Absolut	(=/!=, </>, +-*/)
Klausurpunkte(0-15)	Ratio	(=/!=, </>, +-*/)
Tarifklassen bei der Kfz-Haftpflicht	Ordinal	(=/!=, </>)

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

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

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

Eine besondere Form der Normalverteilung ist die Standardnormalverteilung.

Das Aussehen der beiden Normalverteilungen ähnelt sehr einer Glocke, wobei die Funktionswerte der Kurve gegen 0 streben, wenn man die x-Werte gegen Unendlich gehen lässt. Beide sind symmetrisch um den Mittelwert verteilt. Beim Mittelwert besitzt die Verteilung ihr Maximum.

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

Für sie gilt, dass der Mittelwert bei 0 liegt und die Standardabweichung bei 1, also μ=0 und σ=1. Damit nimmt die Funktionsgleichung folgende Form an:

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

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

Aufgabe 3: Zusammenhangshypothese

Datensatz: diamonds.csv

Var 1 = "price" in US Dollar

Var 2 = "carat" (Gewicht des Diamanten)

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: insurance.csv

Var 1 (AV) = "charges" (das Geld in US Dollars, das von Krankenkasse für Behandlungen bezahlt werden muss)

Var 2 (UV)= "smoker" (Raucher, nicht-Raucher)

Aufgabenstellung

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: verbunden2.xlsx

Var 1 = "Zufriedenheit"

Var 2 = "Messzeitorten (Land, Stadt)"

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: insurance.csv

Var 1 (AV) = "charges" (das Geld in US Dollars, das von Krankenkasse für Behandlungen bezahlt werden muss)

Var 2 (UV) = "children" (die Anzahl der Kinder)

<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