Statistik; 21.10.2022

* Addieren nur mit echten Zahlen möglich

Verschieden Erhebungsarten:

* Vollerhebung (N) : alle werden befragt
  + z.B.: dhbw-weite Umfrage über Augenfarbe : alle 34.000 Studenten
* Teilerhebung = Stichprobe
  + Annahme : prozentuale Ergebnisse aus Stichprobe können auf Gesamtheit übertragen werden
  + Wenn selbe Prozentzahlen in Stichprobe teilnehmen, wie in gesamter population vorhanden sind 🡪 repräsentative Stichprobe

Repräsentativ > nicht repräsentativ

* + != Hochrechnung
* Gesetz der großen Zahl :
  + Je mehr Zahlen, desto genauer das Ergebnis
  + Bessere Güte = besseres Ergebnis
* Strichliste : brauch Befragte nur für Befragung dann nicht mehr (großer Vorteil)

-------

Ausreißer:

* Weit von anderen Werten/ vom Durchschnitt entfernt
* Muss immer im Einzelfall entschieden werden
* kann in Analyse begründet weggelassen werden  
  🡪 deshalb : vor Analysebeginn definieren und rauslassen
* normale Ausreißer
  + einzelner Wert nicht gewichtig für den Durchschnitt
    - entweder weil es ziemlich viele Werte gibt
    - oder weil er noch „zu nah“ an anderen Werten liegt
    - z.B.: Körpergröße, Gewicht
* außergewöhnliche Ausreißer
  + einzelner Wert beeinflusst Durchschnitt gewichtig
    - entweder weil es nur wenige Werte gibt
    - oder weil er doch extrem weit weg liegt
    - z.B.: Range von 10 bis 15 und Ausreißer bei 270
* nicht wichtig für median 🡪 Mitte verschiebt sich egal ob Gerade oder nicht nur um ½ Stellen
* nicht wichtig für modus 🡪 Ausreißer nur einzelner Wert und modus gibt häufigsten Wert an
* können begründet aus Analyse weggelassen werden

------

4.1 Häufigkeiten

- absolute Häufigkeit (fi)

* Nummer der Werte für ein bestimmtes Ereignis (X)
* z.B.: Striche in einer „Kategorie“ in Strichliste
* relative Häufigkeit (hi)
  + absolute Häufigkeit/Anzahl d. Befragten
  + h = fi/n = Xi/n
  + kann Unterschiedliche populations vergleichen
  + können als Dezimalzahl/Prozentzahl dargestellt werden
  + werden immer kaufmännisch gerundet
  + wenn Summe != 1, dann einen Wert verfälschen
    - immer größten Wert nehmen, da Auswirkung am kleinsten  
      (z.b.: 56,5 oder 56,6 vs. 4,3 oder 4,4)
* ∑fi = n
* Absolut kumulierte Wahrscheinlichkeit :
  + Anzahl von 0 bis Wert X
  + Fortlaufend addiert
  + Anzahl bis X
* Relative kumulierte Wahrscheinlichkeit
  + Relative Häufigkeit von 0 bis X
  + Fortlaufend addiert
  + Relative Wahrscheinlichkeit bis X
* Bei kumulierten Wahrscheinlichkeiten entspricht Wahrscheinlichkeit des letzten Wert der Summe aller Werte
* Nur Sachen vergleichbar, die vergleichbar sind (selbe Grundlage, nur relative Wahrscheinlichkeiten)

Darstellungen:

* Diagramme : Standarddarstellung von Statistik
* Balkendiagramm : *auch* Histogramm
* Symmetrie : auf beiden Seiten des Maximums gleich „viel“/gleiches Aussehen
  + Nicht symmetrisch : *auch* schief genannt
    - Mehr links : links-schief
    - Mehr rechts : rechts-schief
* Gipfel = Maximum
  + Eingipflig : es gibt ein Maximum
    - Sonderfall : eingipflig, symmetrisch mit symmetrischen Knickpunkten = Normalform
  + Kein Gipfel (alle Werte ungefähr gleich) : nicht gipfeligkeit; Gleichverteilung
  + Auch lokales Maximum als Gipfel bezeichnet
  + Zwei ungleiche Gipfel : Kamelverteilung
* Ordnungsdaten : Struktur um „Balken“ herum  
  Sachdaten : in Diagramm dargestelltes; „Balken“
* Bei 570x Würfeln:   
  Xi=1,2,3,4,5,6  
  fi=0 bis 570
* X = Ereignisse die eintreten können
* Bei Gipfel : Gipfel kann nicht durch Hüllkurve „stechen“  
  ohne Gipfel: egal ob Balken durchsticht oder nicht 🡪 Form verändert sich nicht
* Hüllkurve : Decke über Balken „gelegt“

----

* Optimal = Maximum oder Minimum; nicht eindeutig
* Runden
  + Mathematisch :
    - Aufrunden : immer zur nächst größeren Zahl (z.B.: 2,1 🡪 3)
    - Abrunden : letzte Zahl wird abgeschnitten (1,6 🡪 1)(z.B. Abinote)
  + Kaufmännisch : wie in Schule
    - Aufrunden : von ,5 auf nächste Zahl
    - Abrunden : bis ,4 auf vorherige Zahl
* Durchschnitt = Abweichung nach oben (über Durchschnitt) – Abweichung nach unten = 0;
  + Das was über dem Durchschnitt und was darunter liegt ist gleich
  + Minimum der quadratischen Abweichung
* Median : Verschiebung von einem Wert um X ändert Median nicht, da auf einer Seite X weggenommen wird, aber auf der anderen Seite der Abstand um X größer wird (z.B.: Kindergarten von 5 Dörfern)

-------

Zahlen:

* Absolute Zahl : nur positive Zahlen; bei negativen Zahlen : Beträge
  + Wird u.a. in Lagerhalle angewendet
* Addieren geht nur mit echten Zahlen 🡪 also nur mit Kardinalzahlen
* Arithmetisches Mittel : sozusagen Schwerpunkt; gleiche Abweichung
* Median : kleinste absolute Summe der Differenzen der einzelnen Werte
* Minimum/maximum : keinster Wert X, nicht wie oft X vorkommt
* Mitte der kummulierten Zahlen = Median

----

Abweichung/Streuung :

* Abweichung/Streuung sind synonyme
* Schwellenwert; Value at Risk (VAR): Trennpunkt von außergewöhnlicher Streuung und normaler Abweichung
* Normale Abweichung : bildet die Inter Quantil Range (IQR)
* IQR : immer mindestens 50% aller Werte
* Summe der Abweichung ist immer 0

Inter Quartil Range

* Mindestens 50% der Werte, die bei geordneter Reihe in der Mitte liegen
* Hängt immer vom Median ab
* Bei gerader Anzahl rechts vom Median: IQR zwischen 2 Werten  
  ungerade Anzahl auf einer Seite : IQR genau 1 Wert
* Bei 13 Werten :   
  Wert 6 = Median  
  IQR bei 50%:   
  untere IQR zwischen Wert 3 und 4 🡪 untere Hälfte wird nochmal halbiert  
  obere IQR zwischen Wert 8 und 9 🡪 obere Hälfte wird nochmal halbiert
* Gibt an in welchem Bereich die mittleren 50% aller Werte liegen

Erwartungswert = Durchschnitt bei unendlich vielen Durchführungen

* Unterscheidet sich vom Durchschnitt, da Durchschnitt auf bestimmte Anzahl bezieht
* Bei Zufallsprozess : modus = median = mean

y= E(X) : X1 x (PX=X1) + X2 x P(X=X2)+…

z.B.: Würfel: X=1 bis 6  
E(X): 1 x 1/6 + 2 x 1/6 + 3 x 1/6 + 4 x 1/6 + 5 x 1/6 + 6 x 1/6 = 3,5   
Man rechnet bei unendlich vielen Versuchen mit einem durchschnittlichen Ergebnis von 3,5 Augen

Varianz:

* mittlere Quadratische Abweichung vom Durchschnitt
* mit s bezeichnet
* o2 = (Zahl1-Durchschnitt)2+(Zahl2-Durchschnitt)2+…./Anzahl
* z.B.: 5 Kinder im Alter von 1, 3, 5, 9 und 12 Jahren  
  Durchschnitt : 6 Jahre  
  o2=(1-6)2+(3-6)2+(5-6)2+(9-6)2+(12-6)2/5 = 80/5 = 16  
  🡪 im Durchschnitt weicht das Alter der Kinder um 16 Jahre2 vom Durchschnittsalter ab

Covarianz:

* den Zusammenhang zwischen 2 verschiedenen Merkmalen berechnen

Berechnung COVxy:

* Produkt der Abweichungen in einer „Zeile“ berechnen (die die zusammengehören)
* Abweichungsprodukte mit anderen Abweichungsprodukten addieren
* Summe durch die Anzahl (n) teilen
* COV = [(ZahlX1-DurchschnittX) x (ZahlY1 -DurchschnittY) + (ZahlX2 – Durchschnitt X) x (ZahlY2 – Durchschnitt Y) + ….]/n

Standardabweichung

o=√(Zahl1-Durchschnitt)2+(Zahl2-Durchschnitt)2+(Zahl3-Durchschnitt)2+…/Anzahl

o=√Varianz

* entweder mit o bezeichnet
* gibt die durchschnittliche Abweichung vom Durchschnitt an
* z.B.: familie mit 5 Kindern (siehe Varianz):  
  o=√16=4

Die Alter der Kinder weichen im Durchschnitt um 4 Jahre vom Durchschnittsalter ab

* zwei unterschiedliche Standardabweichung können nicht verglichen werden
* oft bisschen verschwommen, da Quadrierung die Werte „ungenau“ macht

Volatilitätskoeffizient/Variationskoeffizient (prozentuale Abweichung):

* gibt an um viel Prozent die Standardabweichung vom Mittelwert abweicht
* abkürzung : v
* v = s/Durschnitt
* z.B.: Familie mit 5 Kindern von vorher:  
  v = 4/6 = 0,667  
  Das Alter der Kinder weißt im Durchschnitt um 67% vom Durchschnittsalter ab
* v% = (s/X^m) x 100
* kann mit anderen Variationskoeffizienten verglichen werden (z.B.: Kolibri und Elefanten)

mittlere absolute Abweichung

* Abweichung, aber in „absoluten“ Zahlen
* Abgekürzt mit MAD
* I Wert 1 – Durschnitt I + I Wert 2 – Durchschnitt I + I Wert 3 – Durchschnitt I + …. /Anzahl
* z.B.: Familie mit 5 Kindern   
  I 1-6 I + I 3-6I + I 5-6 I + I 9-6 I + I 12-6 I / 5 = 3,6   
  Im Durschnitt weicht das Alter der Kinder um 3,6 Jahre vom Durchschnittsalter ab
* Punktschätzung: genauer Wert wird geschätzt
  + Oft nicht gut, da oft und einfach verfehlbar
* Intervallschätzung : Bereich/Intervall wird geschätzt