Statistik; 24.10.2022

X^m + s = absoluter Normalbereich

X^m + v% = relativer Normalbereich

-----

Covarianz:

(Abweichung X x Abweichung Y) / n

Hat alleine keinen Sinn, da man Waschmaschinen nicht mit Bananen verrechnen kann

Nur bei kardinalen Skalen sinnvoll

------

4.4.4 Normalverteilung + Stochastische Prozesse

* Zufallsexperiment/stochastische Prozesse : immer gleiches Vorgehen, aber unterschiedliche Ergebnisse (wie in Schule)
  + Min. 50 Sandkörner, Max. 316 Sandkörner
  + Median ist 183, da 183 die Hälfte zwischen 50 und 361 ist
* Da stochastischer Prozess, muss bei Median auch mittlere Anzahl geworfen werden
* Hüllkurve : Spitze bei Median, am Rand weniger 🡪 Nullkurve
* Ballung um die Mitte
  + Daher auch Gipfel am Median, da Symmetrie benötigt wird
  + Arith. Mittelwert nahe Mittelwert, da ziemlich viele Werte und die natürlich Durchschnitt beeinflussen
* Zufallsprozess : modus = mean = median 🡪 werden µ (griechisches m) oder m genannt
  + Erwartungswert : das was man bei unendlich vielen „Würfen“ bekommt
  + µ =E(X)=Wert x Wahrscheinlichkeit + Wert x Wahrscheinlichkeit + ….

µ =E(X) = X1 x P(X=x1) + X2 x P(X=x2)+…

* Knickpunkte der Normalverteilung : Wendepunkte
* Bereich bis zu den Wendepunkten : µ - σ
  + Das was bei anderen Ereignissen Durchschnitt + Standardabweichung ist
* Häufigkeit = wie oft wert X vorkommt
* Gauss : eingipflig, symmetrisch, assymptotisch

Konfidenzintervall:

* Bereich der normalen Abweichung : „bereich in der Mitte“
* *Auch* Vertrauensbereich, Toleranzbereich (v.a. bei Maschinen)

Signifikanzbereich

* Außergewöhnliche Abweichungen : „außerhalb der mittleren Box“
* *Auch* außergewöhnlicher Bereich , Irrtumsbereich, α-Bereich

Sigma-Bereiche :

µ +/- σ = 66,8% aller Werte  
µ +/- 2σ = 95,4 % aller Werte  
µ +/- 3σ = 99,7 % aller Werte

typische definitionen für α-Bereich : 10%, **5%**, 2,5%, 1%, 0,5%

4.5 Cluster

* Einteilung von Daten in Gruppen (Klassifizieren)
  + Beliebig viele Gruppen
  + Beliebig groß
  + „möglichst wenige, aber ausreichend viele“
* Faustregeln im Skript (vor klausur anschauen!)   
  Über die Klassenbildung (Anzahl): **Faustregel**:
  + mindestens 5 und höchstens 20;
  + möglichst Klassen mit gleicher Breite
  + REFA-Vorschlag: Klassenanzahl zwischen n1/2 und n1/3 passend.
  + Papula-Vorschlag: Anzahl der Klassen = n1/2
  + Sturgess-Regel k = 1 + log(2) n
  + Pareto-Regel (80/20-Regel: 80% des Wertes werden von 20% der Menge generiert)

4.5.2 Maschinenlernen:

* Verhalten ändern = lernen ; != Intelligenz
* 2 Gruppen :
  + Cluster : unüberwachtes Lernen
  + Nach Vorgaben : überwachtes Lernen, klassifizierung, braucht erst Übung
* 2 Dimensionen = 2 verschiedene Werte 🡪 X und Y Wert
* KNN :
  + „next neighbor model“ : Punkt wird der Kategorie des nächsten nachbarn zugeordnet
    - Bei k>1 der Kategorie, der die Mehrheit der Nachbarn angehören
  + In Praxis : k immer ungerade, damit es nicht zu „unentschieden“ kommt
* Bei einsortieren in Gruppen : immer begründen können

Distanzmaße:

Mathematische Herangehensweise :

* Satz des Pythagoras / euklidische Differenz
  + a = x – 0
  + b = y – 0
  + c = die Distanz   
    🡺 also a2 + b2 = c2 und dann Wurzel aus c2

Mannhattan-Abstand

* auch Mannheimer Abstand genannt 🡪 Straßen in beiden Städten sind rechtwinklig angelegt
* immer nur gerade oder hoch/runter auf Kästchen gehen 🡪 keine Diagonalen
* D = (x1 – x2 ) + (y1 – y2)

Korrelationen:

* Korrelation = Zusammenhang
* Bei mehreren Kriterien : Graph zum herausfinden von Zusammenhang gut geeignet
* Eine Zeile : ein Datensatz; 2 unterschiedliche Werte, die zu einer „Person“ gehören
* Hüllkurve bei XY-Diagramm in Statistik : um alle Werte außen rum 🡪 Fläche/Ebene auf der alle Punkte liegen
* Alle Punkte graphisch auf einer Linie : totale Korrelation
  + X hängt mit Y zu 100% zusammen (z.B. Körpergröße und Kleidungsgröße)

Verschieden Korrelationen:

* Schmale Linie wie Proportionale Funktion : totale Korrelation
* Dickere „Linie“ : starke Korrelation
* Schwammige „Wolke“ : schwache Korrelation
* Viereck (Werte überall) : keine Korrelation
* Alle Punkte auf einer Linie (Männer u. 20 jährige Frau) : keine Korrelation

Korrelation berechnen:

* Von 1 (starke Korrelation) über 0 (keine Korrelation) bis -1 (starke Korrelation)
* 1 = positive Steigung; -1 negative Steigung der Linie
* Vorzeichen zeigt die Richtung an;
* Faustzahlen nach Zwöfel:  
  <0,2 = sehr geringe  
  0,2 < r < 0,5 = geringe  
  0,5 < r < 0,7 = mittlere  
  0,7 < r < 0,9 = hohe  
  0,9 < r < 1 = sehr hoch
* Ausreißer verfälschen Korrelationskoeffizient :
  + Da sie entweder näher an der imaginären Geraden der totalen Korrelation liegen oder weiter davon entfernt

Korrelationskoeffizient nach Bravais/Pearson:

* Für kardinale Skalen (echte Zahlen, kann man addieren)
* Von 1 (starke Korrelation) über 0 (keine Korrelation) bis -1 (starke Korrelation)
* 1 = positive Steigung; -1 negative Steigung der Linie
* Vorzeichen zeigt die Richtung an;
* Wert zeigt die stärke der Korrelation 🡪 je weiter von 0 weg, desto stärker
* Funktion : r = COVxy/(sx x sy)  
  COV = Kovarianz :   
  COV = Varianz X x Varianz Y (nach Pearson)

Korrelation nach Spearman

* Für ordinale Skalen
* Auch Range von -1 bis 1 (wie Bravais)
* Ordinale Zahlen werden in Rangfolge umgewandelt (1. Platz, 2., 3., usw.)
* Formel : rs = 1 – ( 6 x Differenz2)/n3 – n

1. Differenz aus „Kriterien“ von einzelnen Zeilen gebildet
2. Differenz quadriert
3. Quadratischen Differenzen werden addiert
4. In Formel eingesetzt
   * Formel funktioniert nur, wenn die Summen der Spalten (z.B. Rangfolge 1 und Rangfolge 2) gleichbleiben
   * Bei „Unentschieden“ :
     + Teilsummen müssen gleichbleiben
     + 3x bestes Ergebnis : Rang 1, 2 und 3; also 1 + 2 + 3 = 6, dann durch Anzahl „Plätze“ : 6/3 = 2 🡪 alle sind zweiter und Teilsumme bleibt gleich
   * Wie :
5. Nach „Rang“ ordnen
6. Einzelnen Werten klaren Rang zuordnen
7. Bei mehreren Werten schauen, welche X Plätze sie belegen
8. Wert der X Plätze addieren und durch Anzahl X teilen
9. Die Ränge haben alle diesen Wert und Teilsumme bleibt gleich

* Summen der Spalten müssen gleich bleiben/sein
* Ergebnis = 1; also totale korrelation: einsetzen von 0 in die Gleichung :
  + Wann wird 0 eingesetzt : wenn die Rangfolge in 1 genau gleich ist wie die Rangfolge in 2
* Ergebnis = 0 ; also gar keine Korrelation: einsetzen von 40 in die Gleichung:
  + Wann wird 40 eingesetzt : wenn die rangfolge genau umgedreht wird

Kontigenz-Koeffizient

* Vierfeldertafel/Chi-Quadrat:
* nominale Skalen
* rk = (a x d) – (b x c) / (a+b)(c+d)(a+c)(b+d)
  + *auch* Vierfelderkorrelationskoeffizient genannt
* auch Werte zwischen 1 und -1
* z.B.: Zusammenhang zwischen Geschlecht und Raucher

4.6.1 nicht bearbeitet

4.6.1.5 sonstige Ähnlichkeiten:

Jaccard :

* n ( A ^ B) / n (A u B)
* Anzahl der Daten in A und B / Anzahl der Daten in A oder B

A und B „durch“ A oder B

* Ergebnisse von 0 bis 1
  + 0 : keine der Werte ist in beiden vorhanden; 0 „oben“ 🡪 0 durch irgendwas ist immer 0
  + 1 : genau die gleichen Werte; alles in beiden vorhanden und nichts nur in einem; zwei gleiche Zahlen durcheinander dividiert gibt immer 1

Mengenlehre:

* Zusammenfassung unterschiedlicher Dinge   
  Menge : A B C  
  n = 3
* Bei zwei gleichen :

1. Keine Menge
2. Nur einmal doppeltes Zählen

* A ^ B = Schnittmenge; A geschnitten B
  + Die Daten die sowohl in A als auch in B
* A u B = Vereinigungsmenge
  + Die Daten die mindestens in einem der beiden vorkommen
  + Doppelte werden nur einmal gezählt
* Logische oder : grenzt nicht aus; inklusives oder   
  umgangssprachliches oder : ausgrenzender beigeschmack

Cosinus-Ähnlichkeit

* Punkte als Vektoren angesehen
* Winkel zwischen den Vektoren als Bestimmung der Ähnlichkeit
  + Großer Winkel : unähnlich
  + Kleiner Winkel : ähnlich
  + Bei kleinem Winkel sind sich die Vektoren „näher“
    - „es dauert weniger lange sie aufeinander zu schieben“
  + Cosinus des Winkels : 0 < cos α < 1