# Statistik 1 Tutorium

WS22/23

# Inhaltsverzeichnis

**1 Vorlesung** **1**
1.1 Was ist Statistik . . . 1
1.2 Grundgesamtheit und Stichprobe . . . 1
1.3 Variablen . . . 1
1.3.1 Wiederholung Empirie 1 Hussy-Schreier-Echterhoff: . . . 1
1.4 Skalenniveaus und diskret/stetig . . . 2

**2 Vorlesung** **3**
2.1 Datenmatrix . . . 3
2.2 Häufigkeiten . . . 3
2.3 kumulierte Häufigkeit . . . 3
2.4 Darstellungsarten . . . 4
2.5 Verteilungsformen . . . 4
2.6 Summenzeichen . . . 5

**3 Vorlesung** **5**
3.1 Lagemaße/Maße der zentralen Tendenz . . . 5
3.1.1 Median . . . 5
3.1.2 arithmetische Mittel . . . 5
3.2 Dispersionsmaße/Lagemaße und Verteilungsformen . . . 6
3.2.1 Übung . . . 6

**4 Vorlesung** **6**
4.1 Variationsweite/Spannweite/Range . . . 6
4.2 Interquartilabstand/IQR . . . 6
4.3 Varianz . . . 7
4.4 Standardabweichung . . . 7
4.5 Übung . . . 7

**5 Vorlesung** **8**
5.1 Quantile . . . 8
5.2 Boxplots . . . 8
5.2.1 Ausreißer und Extremwert . . . 8
5.3 Variantionskoeffizient V . . . 8
5.4 Z-Transformation oder Z-Wert . . . 8

**6 Vorlesung** **10**
6.1 Kreuztabelle/ Kontingenztafel . . . 10
6.2 Übung . . . 10

**7 Vorlesung (Zusammenhang zwischen nominal und ordinalskalierten Variablen)** **11**
7.1 Kreuztabellen . . . 11
7.2 Erwartete Häufigkeit . . . 11
7.3 Chi-Quadrat . . . 11
7.4 Normierung Chi-Quadrat mit Phi, Cramers V, C . . . 11
7.4.1 Phi . . . 11
7.4.2 Kontingenzkoeffizient C . . . 12
7.4.3 Cramer's V . . . 12

7.5 Interpretation Phi und Cramer's V . . . . . 12
7.6 Übung . . . . . 13

**8 Vorlesung** **13**
8.1 Spearman's Rho/Rangkorrelationskoeffizient . . . . . 13

# 1 Vorlesung

## 1.1 Was ist Statistik

| Deskriptiv | Inferenz | amtliche Statistik | Explorative Statistik |
|---|---|---|---|
| Merkmale, Zusammenhänge<br>Grafische Darstellung<br>Lage und Streumaße | GG ↔ Stichprobe<br>Stichprobenfehler(sample error) | von Institutionen<br>in Auftrag gegeben | Zusammenhänge in<br>Daten finden, "Big Data",<br>etabliert in Wirtschaft |

## 1.2 Grundgesamtheit und Stichprobe

Grundgesamtheit: Menge der Objekte für die die Aussage der Untersuchung gelten soll
Stichprobe: regelgeleitete Auswahl einer Teilmenge von Elementen aus der Grundgesamtheit
Stichprobenfehler/Sampling Error: Merkmalsausprägung ist in GG und in Stichprobe unterschiedlich
Kleer hatte noch sampling:
Flick Stichproben Kapitel 4

## 1.3 Variablen

### 1.3.1 Wiederholung Empirie 1 Hussy-Schreier-Echterhoff:

1. Beschreiben: Variable B hängt mit Variabale A zusammen A — B

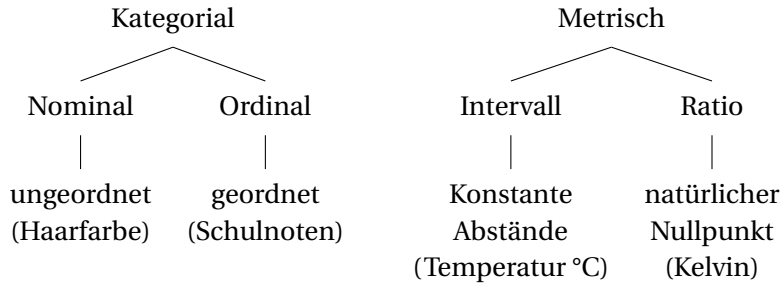
2. Erklären: Variable B ist abhängig von Variable A A→B
Zusammenhang:
positiver Zusammenhang: A↑ B↑
negativer Zusammenhang: A↑ B↓ oder A↓ B↑
Kein Zusammenhang

Kausalrelation:
A⟶B
A⟵B
A⟷B

Beide Zusammen:
A↑ ⟶ B↓

3. erste und zweite Ordnung
1. Ordnung: A ⟶ B
2. Ordnung: A →x →B
x ist intervenierende Variable

### 1.4 Skalenniveaus und diskret/stetig




Pseudometrisch: Ordinal wird oft als Intervall behandelt, wenn genug Ausprägungen vorhanden sind
Welches Skalenniveau haben folgende Variablen?

1. Selbstvertrauen: hoch, mittel, niedrig
   Solution:
2. Jahreszahl (z.B. 1982)
   Solution:
3. Alter
   Solution:
4. Geschlecht
   Solution:
5. Anzahl erreichter Creditpoints
   Solution:
6. Einkommen
   Solution:

diskret: Endlich Viele Werte können angenommen werden 1————————2
stetig: Annahme eines beliebigen Wertes in Intervall 1–1.1–1.127587——-2

Sind die folgenden Variablen diskret oder stetig?

1. Gewicht
   Solution:
2. Jahreszahl (z.B. 1982)
   Solution:
3. Alter
   Solution:
4. Exakte Zeit eines 100m Läufers
   Solution:
5. Geschlecht
   Solution:

6. Einkommen
   Solution:

Kleer hatte noch:
dichotom: 2 Ausprägungen
polytom: Mehr als 2 Ausprägungen

Makro: aggregierte Daten
Mikro: Individualdaten

latent: nicht direkt beobachtbar
manifest: messbar

# 2 Vorlesung

Ziel der Vorlesung: Studierende können univariate Analysen machen

## 2.1 Datenmatrix

- auch *Urliste* genannt
- Spalten → Variablen
- Zeilen → Fälle
- Jeder Fall bekommt eine ID zugewiesen

==> Aufgabe 1 Serdar

## 2.2 Häufigkeiten

**f**requenz und **H**äufigkeit

1. **Absolute Häufigkeit:** $fx_k = Hx_k$
2. **Relative Häufigkeit:** $\frac{fx_k}{n} = hx_k$
3. **prozentuale Häufigkeit:** $hx_k \cdot 100$

## 2.3 kumulierte Häufigkeit

politisches Interesse Allbus:

| Kategorie | $Hx_k$ | $hx_k$ | $hx_k \cdot 100$ | kumulierte prozentuale Häufigkeit |
|---|---|---|---|---|
| sehr stark | 425 | 0,122 | 12,2 | 12,2 |
| stark | 877 | 0,251 | 25,1 | 37,3 |
| mittel | 1437 | 0,412 | 41,2 | 78,5 |
| wenig | 564 | 0,162 | 16,2 | 94,7 |
| überhaupt nicht | 186 | 0,053 | 5,3 | 100 |
| Gesamt | 3490 | 1,000 | 100 | |

1. Wie viele Menschen sind mindestens stark politisch interessiert?
   Solution:

2. Wie hoch ist der prozentuale Anteil der Personen, die weniger als <mittel> interessiert sind?
   Solution:

3. Wie hoch ist der prozentuale Anteil der Personen, <stark, mittel, wenig> angekreuzt haben?
   Solution:

4. Welche Häufigkeitsdarstellung zeigt am besten wie viele Individuen der Stichprobe vor dem 55. Lebensjahr in Rente gegangen sind?
   Solution:

## 2.4 Darstellungsarten

| | Variablenskala | zu beachten |
|---|---|---|
| Piechart | nominal | nur wenig Kategorien |
| Säulendiagramm | nominal, ordinal | Reihenfolge auf X-Achse |
| Histogramm | intervall, ratio | hat Zweck Fläche darzustellen →Polygonzug/Dichteverteilung mit angeben |

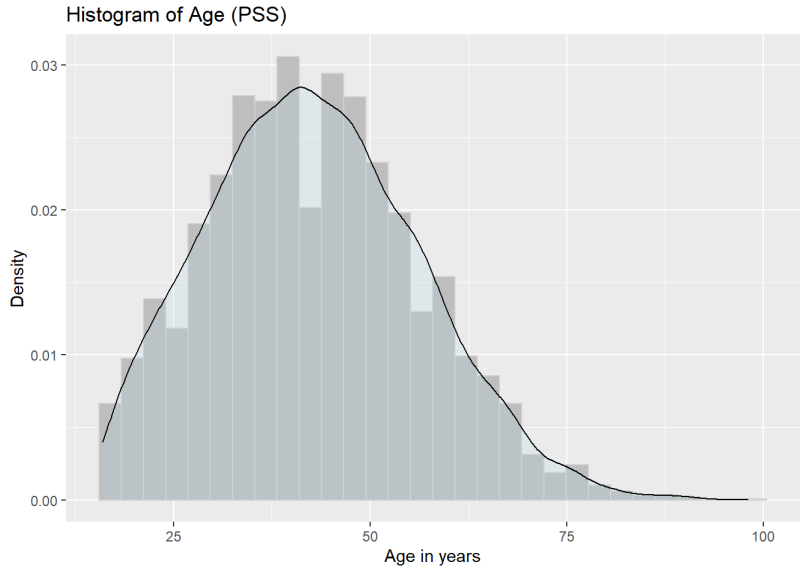



Abbildung 1: Histogramm mit Dichteverteilung

## 2.5 Verteilungsformen

Symetrisch
Bimodal
linkssteil,rechtsschief
rechtssteil,linksschief

### 2.6 Summenzeichen

$\sum_{i=m}^{n} a_i \quad = \quad a_m + a_{m+1} + a_{m+2} + a_{m+3} + \cdots + a_n$

Für alle Aufgaben gilt folgende Tabelle:

| $a_i$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| x | 2 | 1 | 5 | 3 | 3 |

$A = \sum_{i=1}^{n} x_i$ $\quad B = \sum_{i=2}^{4} x_i$ $\quad C = \sum x_i$ $\quad D = \sum {x_i}^2$ $\quad E = (\sum_{i=1}^{n} x_i)^2$

$F = (\sum_{i=1}^{n} x_i) - 1$ $\quad G = \sum_{i=1}^{n} (x_i - 1)$ $\quad H = \sum_{i=1}^{n} (x_i - 1)^2$ $\quad I = (\sum_{i=1}^{n} (x_i - 1))^2$ $\quad J = \frac{\sum_{i=1}^{n} (x_i - 1)}{n}$

$K = \frac{\sum_{i=1}^{n} (x_i - 2)^2}{n}$ $\quad L =$ $\quad M =$ $\quad N =$ $\quad O =$

Lösungen:

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 9 | 14 | 48 | $14^2$ | 13 | 9 | 25 | 81 | $\frac{9}{5} = 1,8$ | $\frac{12}{5}$ | | | | |

# 3 Vorlesung

## 3.1 Lagemaße/Maße der zentralen Tendenz

| | | |
|---|---|---|
| Modus | Wert kommt am häufigsten vor | ratio, intervall, ordinal, nominal |
| Median | Teilt Menge in 2 gleichgroße Teile | ratio, intervall, ordinal |
| arithmetisches Mittel | Durchschnitt | ratio, intervall |

### 3.1.1 Median

| | |
|---|---|
| n - ungerade | $\widetilde{x} = x_{\frac{n+1}{2}}$ |
| n - gerade | $\widetilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$ |

ungerade:

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|
| 3 | 5 | 6 | 8 | 12 |

mittlerer Wert = 6

gerade :

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|
| 3 | 5 | 6 | 8 | 12 | 13 |

Durchschnitt der mittleren beiden Werte = 7

### 3.1.2 arithmetische Mittel

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \leftarrow \text{Datenmatrix (“normale” Formel)}$$

$$\overline{x} = \frac{\sum_{k=1}^{m} (x_k \cdot f_{x_k})}{n} \quad \leftarrow \text{Häufigkeitstabelle (“spezial” Formel)}$$

Erläuterung der Gleichungen:
Um die mittlere Antwort, einen "Durchschnitt", zu berechnen werden zuerst alle Antworten die gegeben wurden aufsummiert und im zweiten Schritt durch die Anzahl der Antworten (n) geteilt.

In der Datenmatrix/Urliste sind alle Antworten als $x_i$'s direkt ablesbar. Diese können einfach aufsummiert werden. Die Anzahl aller Antworten (n) kann an der ID des letzten Falls abgelesen werden (sofern keine Fälle dazwischen herausgefiltert wurden).

In der Häufigkeitstabelle kann man die einzelnen Antworten nicht so direkt ablesen wie in der Datenmatrix. Jedoch wissen wir, dass bspw. 400 Personen Antwortausprägung 1 gegeben haben, 600 Antwortausprägung 2 usw.. Antwort 1 kommt also 500 *mal* in der Datenmatrix vor, Antwort 2 600 *mal* usw.. Wir rechnen also die jeweilige Antwort *MAL* die Anzahl wie oft diese Antwort angegeben wurde. Die Anzahl aller Antworten (n) wird ermittelt indem die Häufigkeiten der einzelnen Antwortausprägungen addiert werden.

### 3.2 Dispersionsmaße/Lagemaße und Verteilungsformen

**Verteilungsübersicht:**

| | |
|---|---|
| Modus < Median < arithmet.Mittel | linkssteil/rechtsschief |
| arithmet.Mittel < Median < Modus | rechtssteil/linksschief |
| 2 Modi, Median = arithmet.Mittel, Modus weicht stark ab | bimodal |
| arithmet.Mittel,Modalwert und Median fast gleich | symetrisch |

#### 3.2.1 Übung

1. Berechne den Median für folgende Werte: 5,2,4,4,3,5,8
   Solution:
2. Was ist der Modus der folgenden Werte? 3,3,4,4,4,5,5,5,5,2,2,1,1,0
   Solution:
3. Um welche Art der Verteilung besitzt folgende Werte: arithmetisches Mittel = 24, Modus = 32, Median = 27
   Solution:

## 4 Vorlesung

### 4.1 Variationsweite/Spannweite/Range

$$V = x_{max} - x_{min}$$

Beispiel: höchster Wert: 10, niedrigster Wert 7
$\rightarrow$10 - 7 = 3 = IQR
Nachteil: Starke Abweichungen einzelner Werte können zu Fehlinterpretationen führen.

### 4.2 Interquartilabstand/IQR

$$IQR = Q_{0.75} - Q_{0.25}$$

## 4.3 Varianz

→durchschnittliche Abweichung, Voraussetzung: Pseudometrisches Skalenniveau

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \quad \longleftarrow \text{Varianz für Grundgesamtheit}$$
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \quad \longleftarrow \text{Varianz für Stichprobe}$$

## 4.4 Standardabweichung

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \quad \longleftarrow \text{Varianz}$$
$$\sqrt{\sigma^2} = \sigma \quad \longleftarrow \text{Standartabweichung}$$

**Beispiel:** (von Statista)
Gefragt wurden 1.000 Personen, wie hoch ihre monatliche Handyrechnung ist. Der Mittelwert liegt bei 40 Euro und die Standardabweichung bei 27. Das heißt, dass die durchschnittliche Entfernung aller Antworten zum Mittelwert 27 Euro beträgt.
Man würde wiefolgt schreiben:

$$\overline{x} = 40€$$
$$\sigma = 27€$$
$$\overline{x} = 40 \pm 27€$$

## 4.5 Übung

Der IQR sind die mittleren 50% der Verteilung.

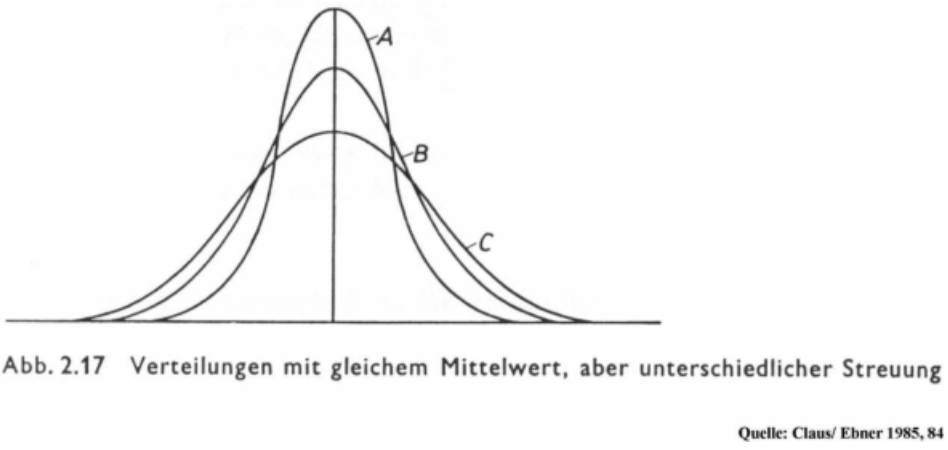



Abb. 2.17 Verteilungen mit gleichem Mittelwert, aber unterschiedlicher Streuung

Quelle: Claus/ Ebner 1985, 84

Abbildung 2: IQR

1. Ordne die IQRs der Verteilungen nach Größe.
   Solution:

2. Wie lautet die Standardabweichung und der Mittelwert für die folgenden Werte 44,67,102,42,52,42
   Solution:

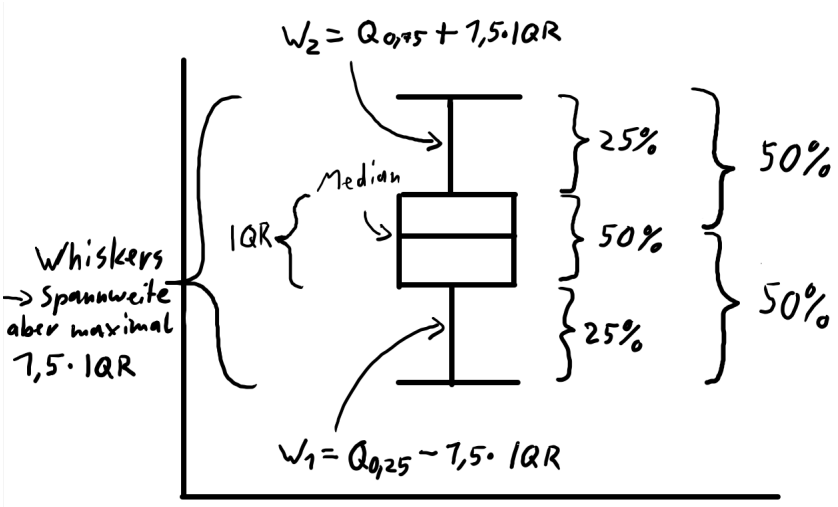
3. Wie lautet die Standardabweichung für folgende Zahlen:
   1,2,3,4,5
   Solution:

# 5 Vorlesung

## 5.1 Quantile

Quartile sind eine Art Quantile. Sie teilen die Anzahl der Werte in gleich große Teile. Quartile teilen mithilfe von 3 Wertegrenzen die Anzahl der Werte in 4 gleichgroße Teile je 25%.

## 5.2 Boxplots




→**(pseudo-) metrisch**

### 5.2.1 Ausreißer und Extremwert

| | |
|---|---|
| Ausreißer | $= 1{,}5 \cdot IQR$ über 3. Quartil/ unter 1. Quartil |
| Extremwert | $= 3 \cdot IQR$ über 3. Quartil/ unter 1. Quartil |

## 5.3 Variantionskoeffizient V

Mit dem Variantionskoeffizienten kann man **Streuungen** verschiedener Verteilungen vergleichen.

$$V = \frac{\sigma}{\overline{x}} = \frac{\text{Standardabweichung}}{\text{arithm. Mittel}} \qquad ; \overline{x} \neq 0$$

## 5.4 Z-Transformation oder Z-Wert

Mit dem Z-Wert kann man **Werte** verschiedener Verteilungen vergleichen. (In Statistik 2 wird die Z-Transformation im Zusammenhang mit der Standardnormalverteilung sehr wichtig.
(Ein interessantes Youtubevideo, es nimmt allerdings schon einige Zusammenhänge vorweg, die später nocheinmal richtig in Statistik 2 vorkommen))

$$z = \frac{x_i - \overline{x}}{\sigma}$$

- Mittelwert aller Z-Werte = 0
- Varianz aller Z-Werte = 1

# 6 Vorlesung

## 6.1 Kreuztabelle/ Kontingenztafel

- für nominale/ordinale Variablen
- Konvention: Zeile-Abhängig/Spalte-unabhängig

Beispieltabelle(fiktional):

x = Studierende besuchen das Tutorium

y = Studierende bestehen die Statistikklausur

Nichtantritt = nicht bestanden

| | Tutorium besucht | Tutorium nicht besucht | *Gesamt* |
|---|---|---|---|
| bestanden | 9 | 59 | 68 |
| nicht bestanden | 2 | 14 | 16 |
| *Gesamt* | 11 | 73 | 84 |

## 6.2 Übung

1. Wie hoch ist der relative Anteil aller eingeschriebenen Studierenden die das Tutorium nicht besucht und die Klausur nicht bestanden haben?
   Solution:
2. Wie hoch ist der relative Anteil aller eingeschriebenen Studierenden die die Klausur bestanden haben?
   Solution:
3. Wie hoch ist der relative Anteil der Tutoriumsbesucher, die die Klausur bestanden haben?
   Solution:
4. Wie viel Prozent aller bestehenden Studierenden haben das Tutorium besucht?
   Solution:

Ergänze folgende Kreuztabelle (Statistisches Bundesamt 2021):

| Geschlecht x Rauchen | Raucher | Nichtraucher | *Gesamt* |
|---|---|---|---|
| männlich | 5059 | A | 22 684 |
| weiblich | B | C | 23 547 |
| *Gesamt* | 8 738 | D | 46 231 |

1. A
   Solution:
2. B
   Solution:
3. C
   Solution:
4. D
   Solution:

# 7 Vorlesung (Zusammenhang zwischen nominal und ordinalskalierten Variablen)

## 7.1 Kreuztabellen

Es gibt 2 Arten von Kreuztabellen:

1. Kontingenztabelle - enthält beobachtete Werte
2. Indifferenztabelle - enthält erwartete Werte

## 7.2 Erwartete Häufigkeit

$$f_{e(ij)} = \frac{Zeilensumme \cdot Spaltensumme}{n}$$

**Erklärung:**

Die Gleichung kann leicht umgestellt werden in: $f_{e(ij)} = Zeilensumme \cdot \frac{Spaltensumme}{n}$. Nun wird deutlich, dass "Spaltensumme durch n" ein Prozentsatz ist (äquivalent ergibt sich $Spaltensumme \cdot \frac{Zeilensumme}{n}$, was im Grunde dasselbe ist).

Dieser Zeilensummenprozentsatz wird nun durch das Malrechnen auf alle Fälle der jeweiligen Spalte der Zelle angewendet. So entsteht der erwartete Wert.

Residuen: Differenz beobachteter und erwarteter Werte
je höher dieser Unterschied ist, desto eher kann man einen Zusammenhang vermuten

## 7.3 Chi-Quadrat

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(f_{b_{ij}} - f_{e_{ij}})^2}{f_{e_{ij}}}$$

**Erklärung:**

Für jede Zelle werden die Abstände von beobachtetem und erwartetem Wert(Residuen) berechnet($f_{b_{ij}} - f_{e_{ij}}$). Da wir nur an den positiven Abständen interessiert sind (wie schon bei der Varianz) wird das Residuum der Zelle quadriert. Nun wird durch den erwarteten Wert geteilt„ weil wir von keinem Zusammenhang, also den erwarteten Werten ausgehen. Alles was wir bisher gerechnet haben wird für alle Zellen gerechnet und zusammengezählt. Das wird in der Gleichung damit erreicht, dass alle Spalten aufsummiert werden ($\sum_{j=1}^{m}$) und diese Ergebnisse für alle Zeilen summiert werden ($\sum_{i=1}^{k}$).

1. $0 < \chi^2 < \infty (nichtstandardisiert)$
2. Je größer der Wert desto größer der Zusammenhang (0: kein Zusammenhang)

**ABER!!!** **Abhängig von n und der Variablenzahl.**

## 7.4 Normierung Chi-Quadrat mit Phi, Cramers V, C

### 7.4.1 Phi

$\phi$ korrigiert die Abhängigkeit von n. Es ist sinvoll in die Analyse mit einzubeziehen, wenn man den Zusammenhang unabhängig von der Gesamtfallzahl interpretieren möchte.

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

- $0 < \phi < 1$, wobei 0: kein Zusammenhang, 1: max. Zusammenhang

### 7.4.2 Kontingenzkoeffizient C

*C* korrigiert die abhängigkeit von der Variablenanzahl. Es ist mega unwichtig, wird nur selten gelehrt. Der Grund dafür ist, dass der Versuch unabhängig von der Kategorienanzahl interpretieren zu können schief geht. Die Kategorienzahl muss wieder bei der Berechnung von $C_{max}$ mit einbezogen werden.

- $0 < C < C_{\text{max.}}$, wobei 0: kein Zusammenhang, $C_{max}$: max. Zusammenhang

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$C_{\text{max}} = \sqrt{\frac{R-1}{R}} \qquad \text{;R = Minimum Zeilen-/Spaltenanzahl}$$

Beispiele für R:

$2 \times 2$ : R = 2

$3 \times 4$ : R = 3

$4 \times 3$ : R = 3

### 7.4.3 Cramer's V

Cramer's V wird für den Vergleich von $\chi^2$ aus verschieden großen Kreuztabellen genutzt.

- 0 < Cramer's < 1

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{\chi^2_{\text{max}}}} = \sqrt{\frac{\chi^2}{n \cdot (R-1)}} = \sqrt{\frac{\chi^2}{n \cdot (\min(k,m)-1)}}$$

## 7.5 Interpretation Phi und Cramer's V

| Cramer's V bzw. Phi | Interpretation |
|---|---|
| $\leq 0{,}05$ | kein Zusammenhang |
| > 0,05 bis 0,10 | sehr schwacher Zusammenhang |
| > 0,10 bis 0,20 | schwacher Zusammenhang |
| > 0,20 bis 0,40 | mittelstarker Zusammenhang |
| > 0,40 bis 0,60 | starker Zusammenhang |
| > 0,60 | sehr starker Zusammenhang |

### 7.6 Übung

| Geschlecht x Rauchen | Raucher | Nichtraucher | *Gesamt* |
|---|---|---|---|
| männlich | 5 059 | 17 625 | 22 684 |
| weiblich | 3 679 | 19 868 | 23 547 |
| *Gesamt* | 8 738 | 37 493 | 46 231 |

Abbildung 3: Statistisches Bundesamt 2021
https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Gesundheitszustand-Relevantes-Verhalten/Tabellen/liste-rauchverhalten.html#119170

1. Berechne und interpretiere $\chi^2$ für die Tabelle.
   Solution:
2. Berechne und interpretiere $\phi$
   Solution:
3. Berechne C max
   Solution:
4. Berechne und interpretiere Camer's V
   Solution:

## 8 Vorlesung

### 8.1 Spearman's Rho/Rangkorrelationskoeffizient

Spearman's $\rho$ (Rangkorrelationskoeffizient) findet Anwendung, wenn beide Variablen ordinal, metrisch oder ordinal und metrisch skaliert sind.

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}$$
$$d_i = R(x_i) - R(y_i) \quad \text{(Differenz der Rangplätze)}$$
$$-1 < \rho < 1$$

**Erklärung:**
Ursprung der Formel: Die Formel ergibt sich (Achtung, vorgriff auf nächste Vorlesung) aus einem Spezialfall von Pearson's r. Es muss lediglich statt den wahren Werten Ränge eingesetzt werden um die Formel zu erhalten.

**Herleitung:** Quelle

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad \text{weil es keine Gleichstände gibt und} x, y \in \{1, 2, ..., n\}$$

Nenner:

$$\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$$
$$= \frac{n(n+1)(2n+1)}{6} - n(\frac{(n+1)}{2})^2 = n(n+1)(\frac{(2n+1)}{6} - \frac{(n+1)}{4})$$
$$= n(n+1)(\frac{(8n+4-6n-6)}{24}) = n(n+1)(\frac{(n-1)}{12})$$
$$= \frac{n(n^2-1)}{12}$$

Zähler:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i(y_i - \bar{y}) - \sum_i \bar{x}(y_i - \bar{y}) = \sum_i x_i y_i - \bar{y}\sum_i x_i - \bar{x}\sum_i y_i + n\bar{x}\bar{y}$$
$$= \sum_i x_i y_i - n\bar{x}\bar{y} = \sum_i x_i y_i - n(\frac{n+1}{2})^2 = \sum_i x_i y_i - \frac{n(n+1)}{12}3(n+1)$$
$$= \frac{n(n+1)}{12}.(-3(n+1)) + \sum_i x_i y_i = \frac{n(n+1)}{12}.[(n-1) - (4n+2)] + \sum_i x_i y_i$$
$$= \frac{n(n+1)(n-1)}{12} - n(n+1)(2n+1)/6 + \sum_i x_i y_i = \frac{n(n+1)(n-1)}{12} - \sum_i x_i^2 + \sum_i x_i y_i$$
$$= \frac{n(n+1)(n-1)}{12} - \sum_i (x_i^2 + y_i^2)/2 + \sum_i x_i y_i = \frac{n(n+1)(n-1)}{12} - \sum_i (x_i^2 - 2x_i y_i + y_i^2)/2$$
$$= \frac{n(n+1)(n-1)}{12} - \sum_i (x_i - y_i)^2/2 = \frac{n(n^2-1)}{12} - \sum d_i^2/2$$

Beide Zusammenfügen:

$$= \frac{n(n+1)(n-1)/12 - \sum d_i^2/2}{n(n^2-1)/12} = \frac{n(n^2-1)/12 - \sum d_i^2/2}{n(n^2-1)/12} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Ich habe die Herleitung hier eingefügt um zu zeigen, dass an diesem Punkt der Vorlesung Schulmathematik vielleicht noch mit Mathe Leistungskurs zum Herleiten ausreicht, allerdings an seine Grenzen stößt. Formeln wie diese können von Nichtmathematikern allerdings durch Kontext und Anwendung versucht werden zu verstehen.

**Anwendung und Beispiel(fiktiv):** Wir gehen von folgender Tabelle aus:

| ID | Anzahl Tutoriumsbesuche | Klausurnote |
|---|---|---|
| 1 | 0 | 8 |
| 2 | 3 | 9 |
| 3 | 0 | 6 |
| 4 | 1 | 5 |
| 5 | 12 | 12 |
| 6 | 10 | 13 |

Nun werden den Werte Ränge zugewiesen, also vom kleinsten Wert bis zum größten Wert durchnummeriert. Sollte ein Wert doppelt (Im Beispiel die 0) vorkommen, wird beiden der Durchschnitt des Ranges zugeordnet. 0 hat Rang 1, weil 0 zweimal vorkommt bilden wir den Durchschnitt der Ränge. Einer Null wird also Rang 1 zugeordnet, der anderen Rang 2 und der Durchschnitt ist 1,5. Die 1,5 wird beiden zugeordnet.

| ID | Ränge der Anzahl Tutoriumsbesuche | Ränge der Klausurnote |
|---|---|---|
| 1 | 1,5 | 3 |
| 2 | 3 | 4 |
| 3 | 1,5 | 2 |
| 4 | 2 | 1 |
| 5 | 5 | 5 |
| 6 | 4 | 6 |

Im Anschluss wird die Differenz der Ränge ($d_i$) berechnet:

| ID | Ränge der Anzahl Tutoriumsbesuche | Ränge der Klausurnote |
|---|---|---|
| $d_i$ | | |
| 1 | 1,5 | 3 |
| 1,5 | | |
| 2 | 3 | 4 |
| 1 | | |
| 3 | 1,5 | 2 |
| 0,5 | | |
| 4 | 2 | 1 |
| 1 | | |
| 5 | 5 | 5 |
| 0 | | |
| 6 | 4 | 6 |
| 2 | | |

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \quad \text{Summe ausformulieren} \rightarrow \text{Ränge einsetzen}$$

$$= 1 - \frac{6 \cdot (1,5^2 + 1^2 + 0,5^2 + 1^2 + 0^2 + 2^2)}{n \cdot (n^2 - 1)} \quad \text{Ränge zusammenrechnen und n einsetzen}$$

$$= 1 - \frac{6 \cdot (2,25 + 1 + 0,25 + 1 + 4)}{n \cdot (6^2 - 1)} \quad \text{weiter zusammenfassen}$$

$$= 1 - \frac{6 \cdot 8,5}{6 \cdot (6^2 - 1)} \quad \text{weiter zusammenfassen}$$

$$= 1 - \frac{51}{6 \cdot (36 - 1)} = 1 - \frac{51}{6 \cdot 35} = 1 - \frac{51}{210} = 1 - 0,243 = 0,757$$