Prüfungstutorat «Einführung in die Statistik»

Nick Glättli

FS23

Fachverein Polito

Organisatorisches

Tutorat in deutscher Sprache (Questions in english welcome!)

Dauer: 13:00 bis 17:00 Uhr

Pause: Nach Bedarf

- Bitte melden

Fragen jederzeit stellen!!

Aufbau zur Orientierung nach Vorlesungen, Fokus auf Konzepten

R-Code auf Github

Wiederholung: Forschungslogik

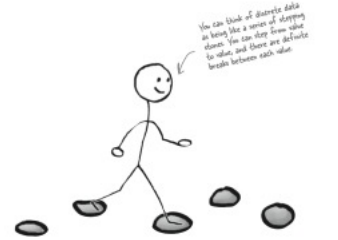
Variablen: Diskret vs Kontinuierlich

Diskret: Endliche Anzahl Kategorien

- Beispiel: Regimetyp, Parteien, Anzahl Kinder

Kontinuierlich: Werte können jeden belibigen Wert annahmen

- Beispiel: Körpergrösse, Alter, Einkommen



Skalenniveaus

Nominal

Ordinal

Intervall

Ratio

Nominalskala

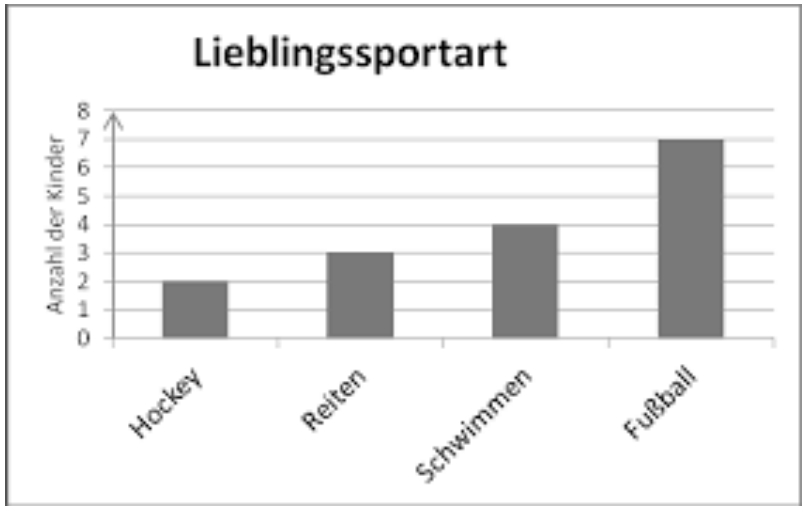
Relation: Keine Rangfolge

Werte: Dient ausschliesslich der Identifikation

Transformationen: Werte müssen noch immer eindeutig sein

Statistische Auswertung: Häufigkeit, Modus (häufigster Wert)

Beispiel: Geschlecht, Religion



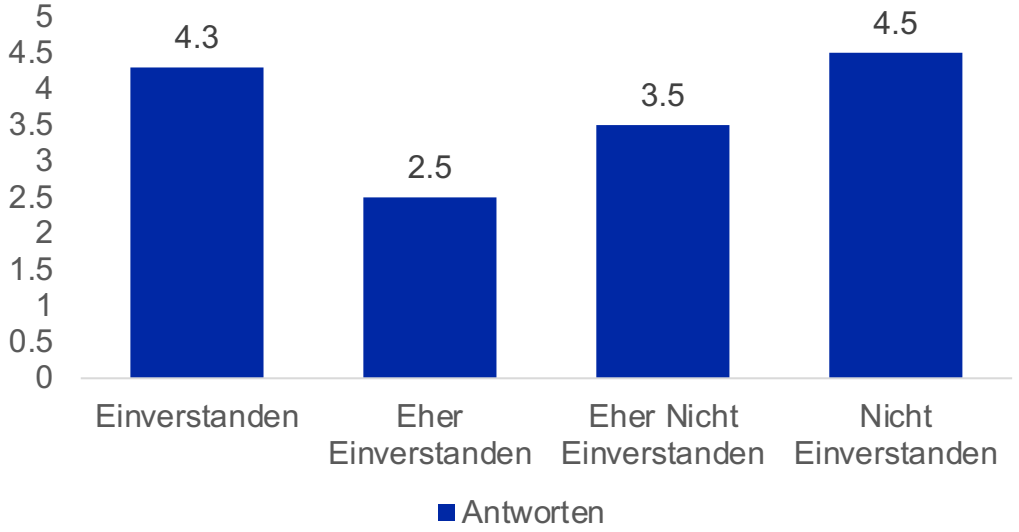
Ordinalskala

Relation: Natürliche Rangfolge, Abstände nicht quantifizierbar

Transformationen: Rangerhaltend → monoton steigend (zBsp Multiplikation)

Statistische Auswertung: Nominalskala + Median

Beispiel: Schulnoten, Skalen bei Befragungen



Intervallskala

Relation: Natürliche Rangfolge, Abstände quantifizierbar, kein natürlicher Nullpunkt

Transformationen: Linear

Statistische Auswertung: Ordinalskala + arithmetischer Mittelwert

Beispiel: Grad Celsius

Ratioskala

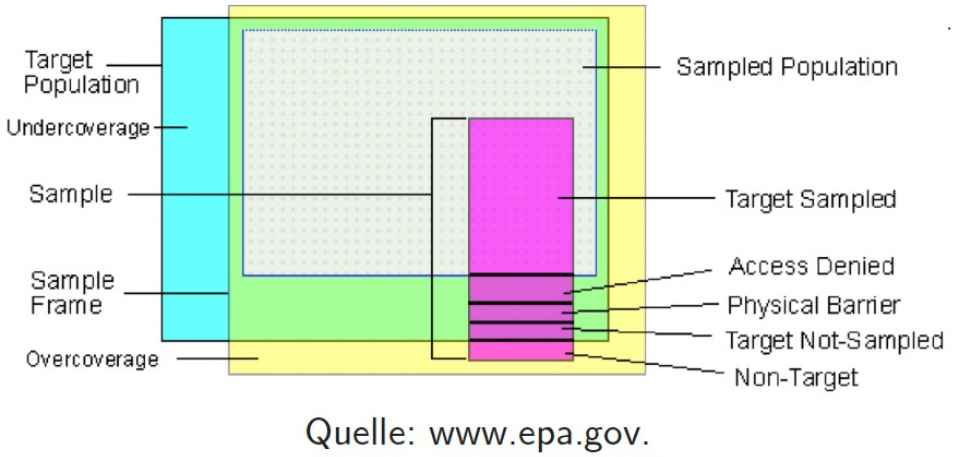
Relation: Natürliche Rangfolge, Abstände quantifizierbar, Natürlicher Nullpunkt

Transformationen: Rangerhaltend → Ähnlichkeitstransformationen

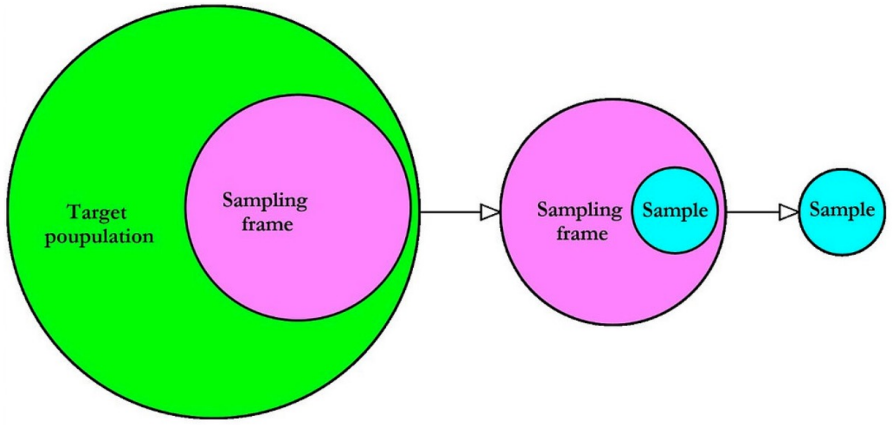
Statistische Auswertung: Intervallskala + geometrisches Mittel

Beispiel: Grad Kelvin

Sampling



Quelle: www.epa.gov.



Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 3

Häufigkeiten

Absolut: n

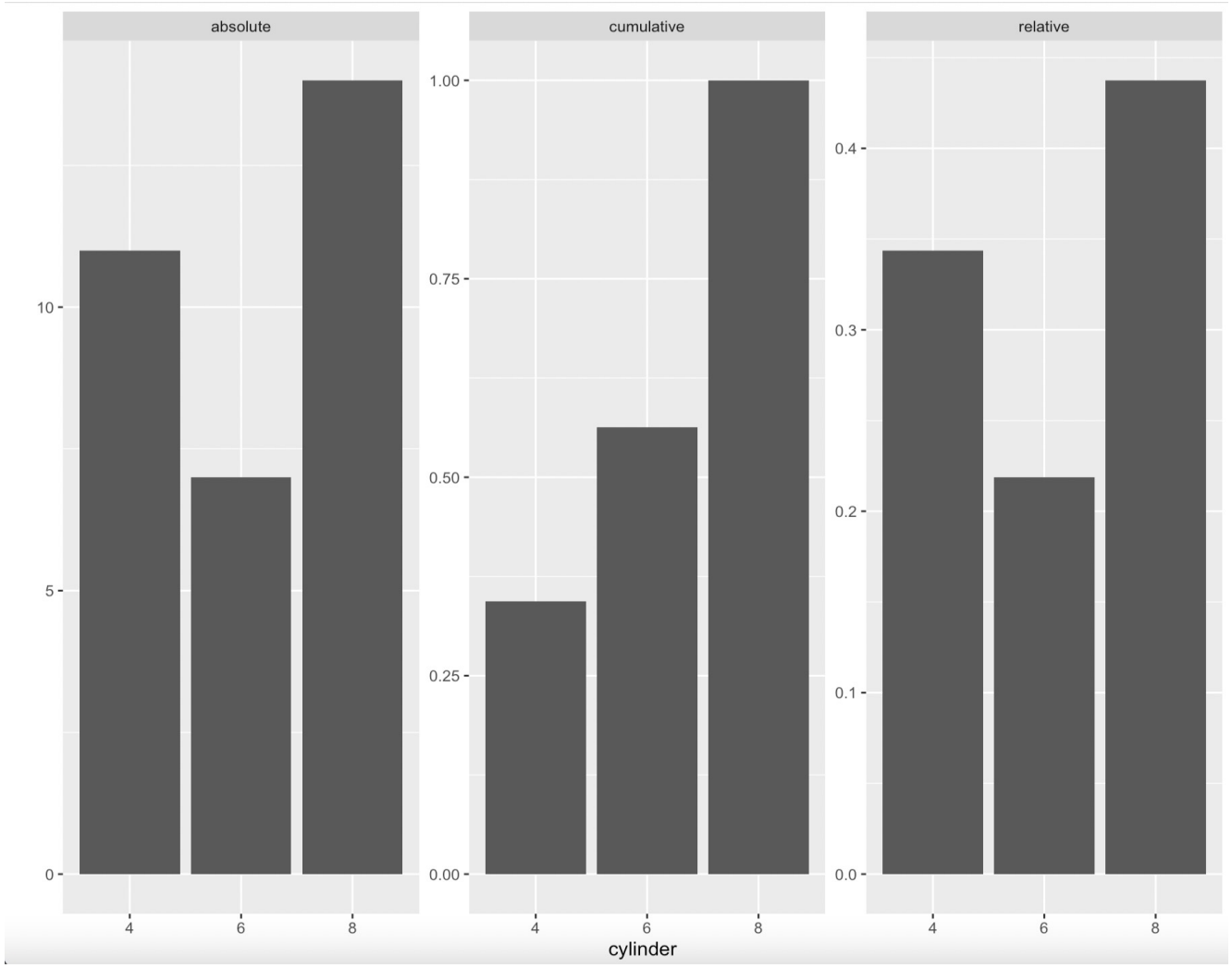
Relativ: n/N

Absolut: $\sum n/N$

Ausprägung	f	f'	cum.freq.
sehr einverstanden	440	0.409	40.9
eher einverstanden	268	0.249	65.9
eher nicht einverstanden	145	0.134	79.4
überhaupt nicht einverstanden	221	0.206	100
Total	1074	1.000	–

Quelle: Vox 114

Häufigkeiten



Messkategorien (Sturges-Regel)

Problem: Bei metrisch skalierten Variablen können Häufigkeiten schwer ermittelt werden bzw. Machen häufig kein Sinn.

Beispiel Alter: Wollen wir wissen wie viele 18.34-Jährige in unserem Sample sind?

Lösung: Messkategorien!

Sturges Regel:

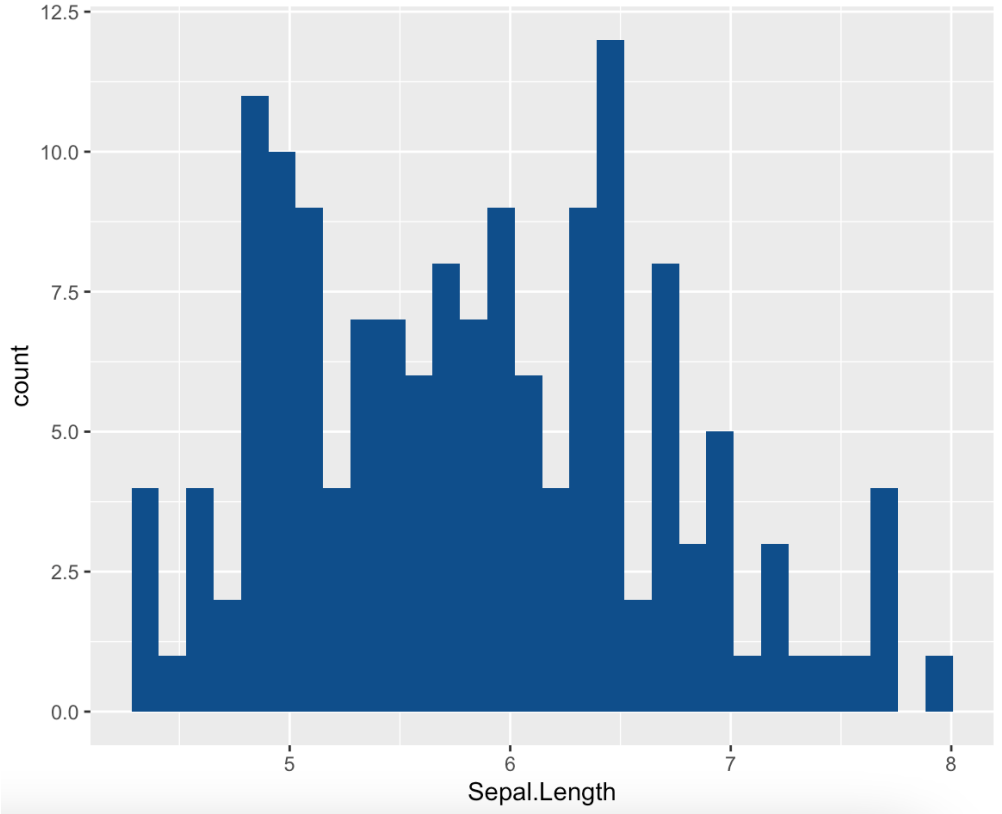
Anzahl Katgeorien: $K = 1 + \lceil log_2 n \rceil$

Messintervall: $\frac{x_{max} - x_{min}}{K}$

Histogramm

Das Histogramm wird häufig genutzt um die Verteilung von metrischen Variablen darzustellen. Klassischerweise sind die Messklassenbreiten gleich.

→ Häufigkeitsdichte = Häufigkeit



Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 4

Masse der zentralen Tendenz

Wichtigste Masse

- **Modus**
- **Median**
- **Arithmetischer Mittelwert**
- Geometrischer Mittelwert
- Harmonischer Mittelwert

Modus

Häufigste Wert im Sample

Besonders wichtig bei Nominalskalen

Auenfarbe	Anzahl
Braun	3
Grün	**5**
Blau	2

Median

Mittlerer Wert im Sample (genau in der Hälfte)

- Formel (n = ungerade): $x_{\frac{n+1}{2}}$
- Formel (n = gerade): $\frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n+1}{2}})$

Median ist eines der wichtigsten Masse, aufgrund der **hohen Robustheit**.

Was ist der Median-Stundenlohn?

Alter	Stundenlohn
15	30
15	44
17	45
18	50
30	

Median

Mittlerer Wert im Sample (genau in der Hälfte)

- Formel (n = ungerade): $x_{\frac{n+1}{2}}$
- Formel (n = gerade): $\frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n+1}{2}})$

Median ist eines der wichtigsten Masse, aufgrund der **hohen Robustheit**.

Was ist der Median-Stundenlohn?

44.5

Alter	Stundenlohn
15	30
15	44
17	45
18	50
30	

Arithmetischer Mittelwert

Summe der Werte durch deren Anzahl

Formel: $\frac{\sum x}{N}$

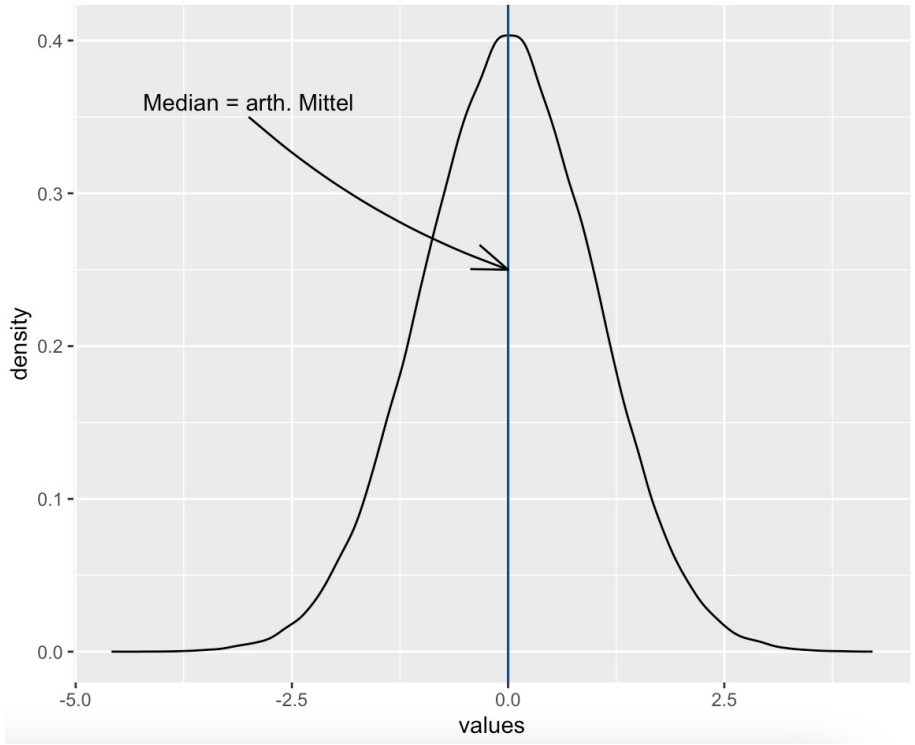
Vorteile

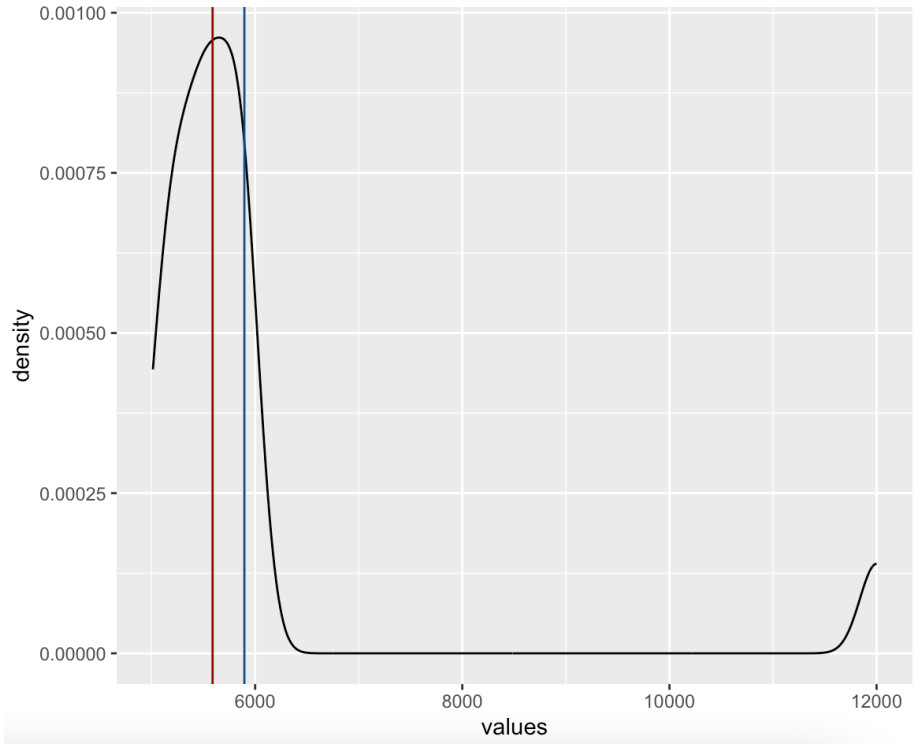
- Die Informationen werden vollständig ausgeschöpft
- Schwerpunkteigenschaft: Die Summe aller Abweichungen = 0
- Optimalitätseigenschaft: Die Summe der quadrierten Abweichungen der Einzelwerte vom arithmetischen Mittel ist minimal

Nachteil

- Nicht Robust gegenüber Ausreissern

Median oder arithmetisches Mittel?





Median oder arithmetisches Mittel?

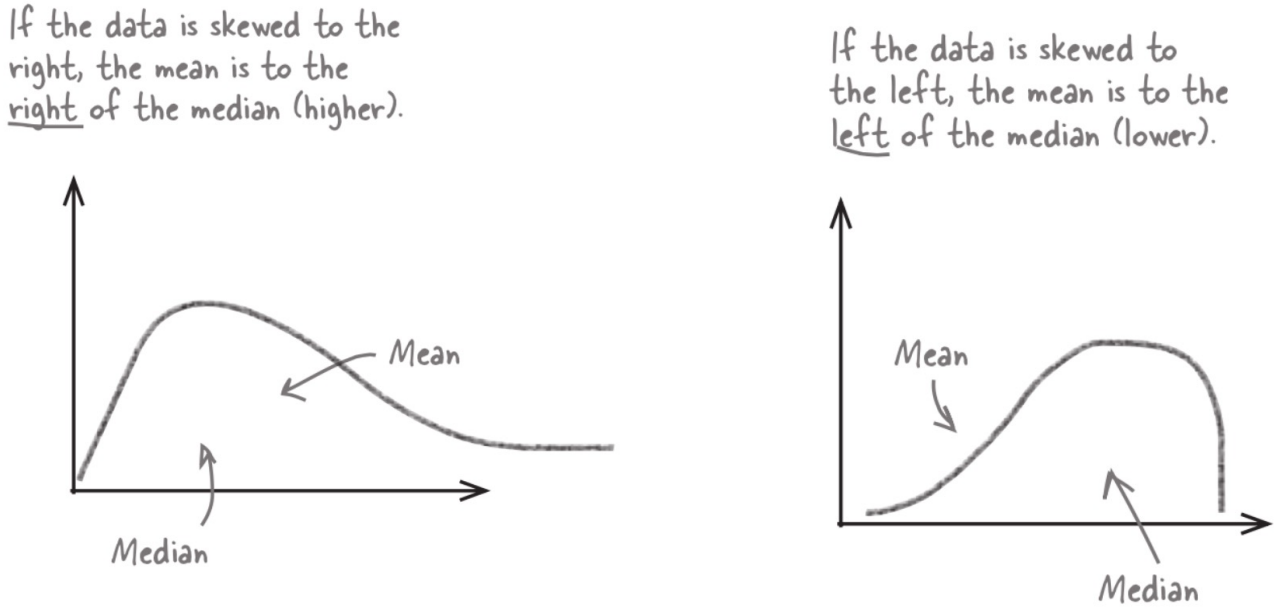
Median: Mitte des Samples

Arithmetischer Mittelwert: Mitte der Spannweite

Schiefe Verteilung

Rechtsschiefe Verteilung: Mehr Werte links des Medians

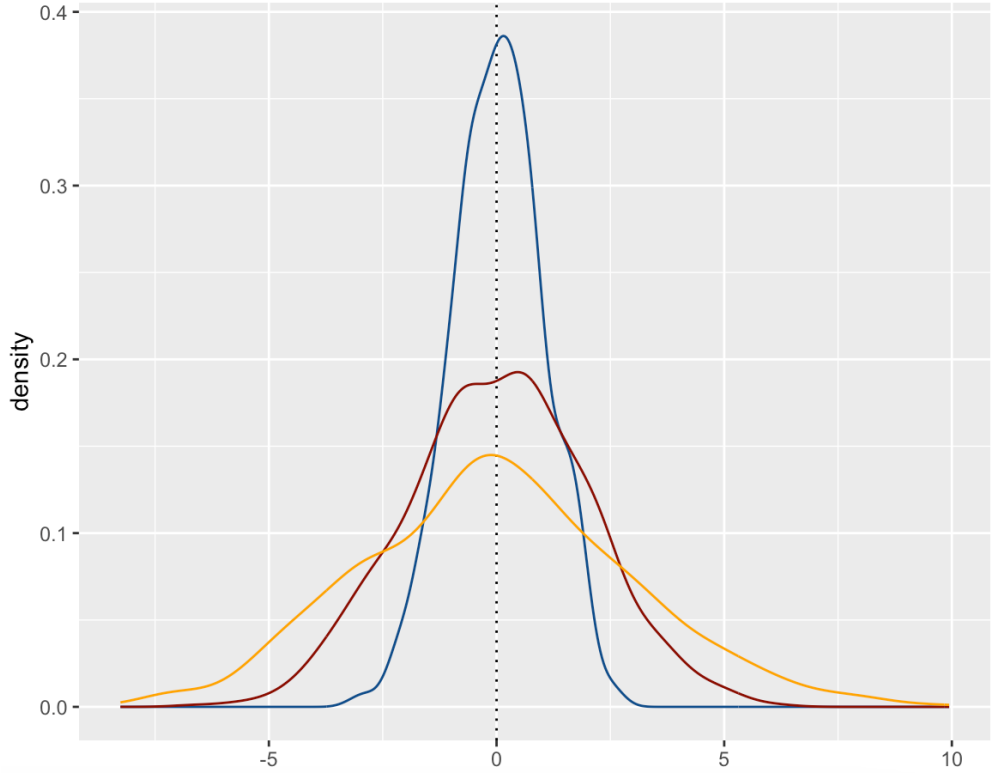
Linksschiefe Verteilung: Mehr Werte rechts des Medians



Streuung

Was ist Streuung: Gesamte Abweichung der Sample-Werte vom Mittelwert

Verteilungen unterscheiden sich trotz identischer Masse der zentralen Tendenz!
Mittelwert ist bei allen 3 Kurven 0.



Streeungsmasse

- **Variationsratio**
- Spannweite
- **Perzentil**
- **IQR**
- **Varianz**
- **Standardabweichung**

Variationsratio

Formel: $\mathbf{1 - \frac{f_m}{n}}$

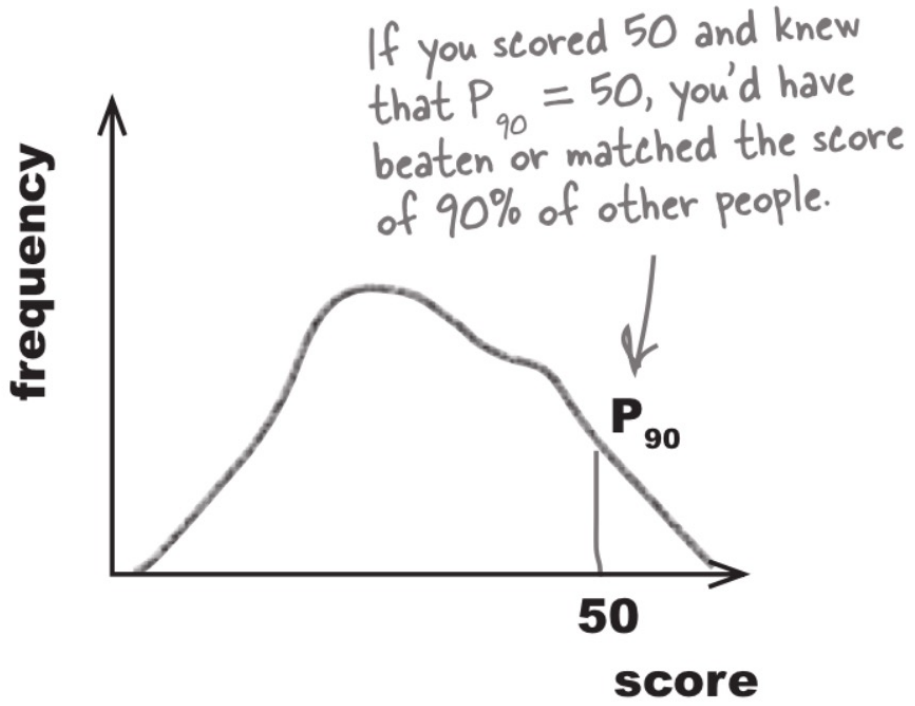
f_m= *Modus*

Was heisst das? Wie viel Prozent der Werte entsprechen nicht dem Modus.

Eignet sich also für Nominalskalen.

Perzentil

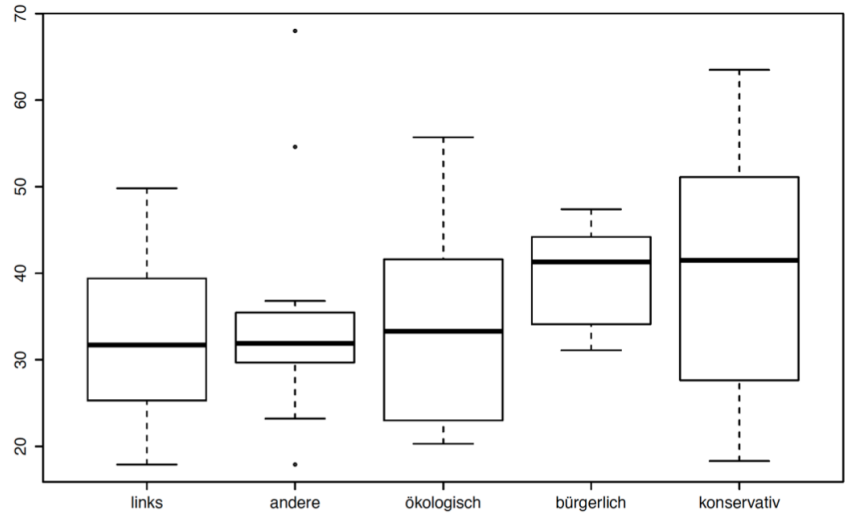
Das Perzentil gibt an, wie viel Prozent der Werte <= X sind.

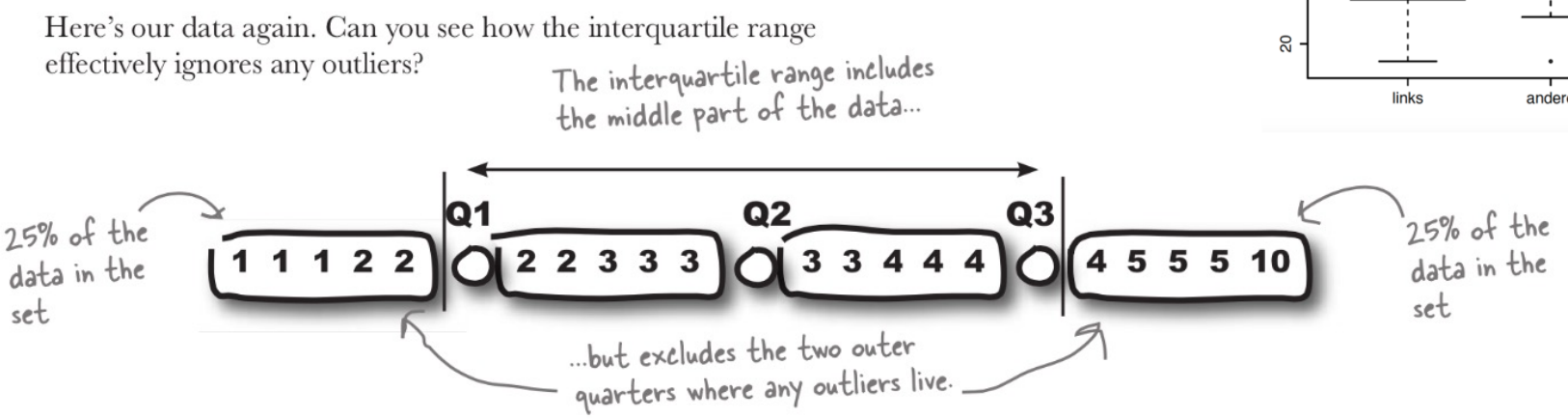


Interquartilsabstand

Beschränkung des Samples uf die mittleren 50% der Werte

– Besonders Robust gegen Ausreisser





Varianz

Die Varianz ist eines der wichtigsten Masse in der Statistik

Formel Populationsvarianz: $\sigma^2 = \frac{\sum x_i - \overline{x}}{N}$

Formel Stichprobenvarianz: $s^2 = \frac{\sum x_i - \overline{x}}{n-1}$

Varianz ist die durchschnittliche quadrierte Abweichung vom artithmetischen Mittel.

Standardabweichung

Die Standardabweichung ist nicht anderes als die Wurzel der Varianz.

Dadurch verliert man die Potenz → einfachere interpretation

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 5

Relatives Risiko, Odds und Odds ratio

Das **relative Risiko** drückt das Verhältnis des Riskikos zweier Subgruppe aus.

Formel: $\frac{a/(a+c)}{b/(b+d)}$

Die **Odds** geben das Verhältnis des Auftretens eine dichotomen Variable an.

Die **Odds Ratio** vergleicht die Odds zweier Gruppen.

Formel: $\frac{a/c}{b/d}$

Relatives Risiko, Odds und Odds ratio

Das **relative Risiko** drückt das Verhältnis des Riskikos zweier Subgruppe aus.

Formel: $\frac{a/(a+c)}{b/(b+d)}$

Die **Odds** geben das Verhältnis des Auftretens eine dichotomen Variable an.

Die **Odds Ratio** vergleicht die Odds zweier Gruppen.

Formel: $\frac{a/c}{b/d}$

	Männer	Frauen
Ja	30	40
Nein	70	60

Berechne das RR, die Odds (Ja) und die Odds Ratio (Ja nach Geschlecht).

RR = 0.75, Odds (Ja) = 0.54, Odds Ratio = 1.125

Chi-Quadrat

Chi-Quadrat ist ein Zusammenhangsmass zweier diskreter Variablen.

$$\chi^2_{(R-1)(C-1)} \sim \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

Beispiel aus der Vorlesung:

	Parteiidentifikation								
	SVP	SP	FDP	CVP	GPS	glp	andere	keine	
Nein	68	86	119	84	15	25	36	119	552
Nein erw.	81.4	102.6	98.8	66.8	34.2	23.3	33.7	111.3	
Ja	82	103	63	39	48	18	26	86	465
Ja erw.	68.6	86.4	83.2	56.2	28.8	19.7	28.3	93.7	
	150	189	182	123	63	43	62	205	1017

Quelle: VOTO (ungewichtete Werte).

Chi-Quadrat

Chi-Quadrat ist ein Zusammenhangsmass zweier diskreter Variablen.

$$\chi^2_{(R-1)(C-1)} \sim \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

	Parteiidentifikation							
	SVP	SP	FDP	CVP	GPS	glp	andere	keine
Nein	–13.4	–16.6	20.2	17.2	–19.2	1.7	2.3	7.7
Ja	13.4	16.6	–20.2	–17.2	19.2	–1.7	–2.3	–7.7

Quelle: VOTO (ungewichtete Werte).

Beispiel aus der Vorlesung:

	Parteiidentifikation								
	SVP	SP	FDP	CVP	GPS	glp	andere	keine	
Nein	68	86	119	84	15	25	36	119	552
Nein erw.	81.4	102.6	98.8	66.8	34.2	23.3	33.7	111.3	
Ja	82	103	63	39	48	18	26	86	465
Ja erw.	68.6	86.4	83.2	56.2	28.8	19.7	28.3	93.7	
	150	189	182	123	63	43	62	205	1017

Quelle: VOTO (ungewichtete Werte).

Wann brauche ich welches Zusammenhangsmass?

Nominale Skalen

- Chi Quadrat
- 2 Dichotome Variablen: Phi
- Cramérs V
- Lambda (Reduktion des Schätzfehlers in %)

Ordinale Skalen

- Gamma
- Spearman’s Roh

Gamma

Goodman und Kruskal's Gamma ist ein Zusammenhangsmass für ordinal skalierte Variablen.

Konkordantes Paar?

Formel

$$\gamma = \frac{N_s - N_d}{N_s + N_d}$$

wobei:

N_s ist die Anzahl konkordanter Paare
N_d ist die Anzahl diskordanter Paare

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

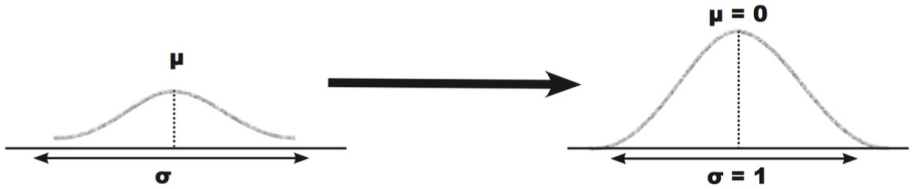
Vorlesung 6

z-Score

Streuungen bei unterschiedlicher Skalierung können nicht direkt miteinander verglichen werden.

Deshalb: Standardisierung!

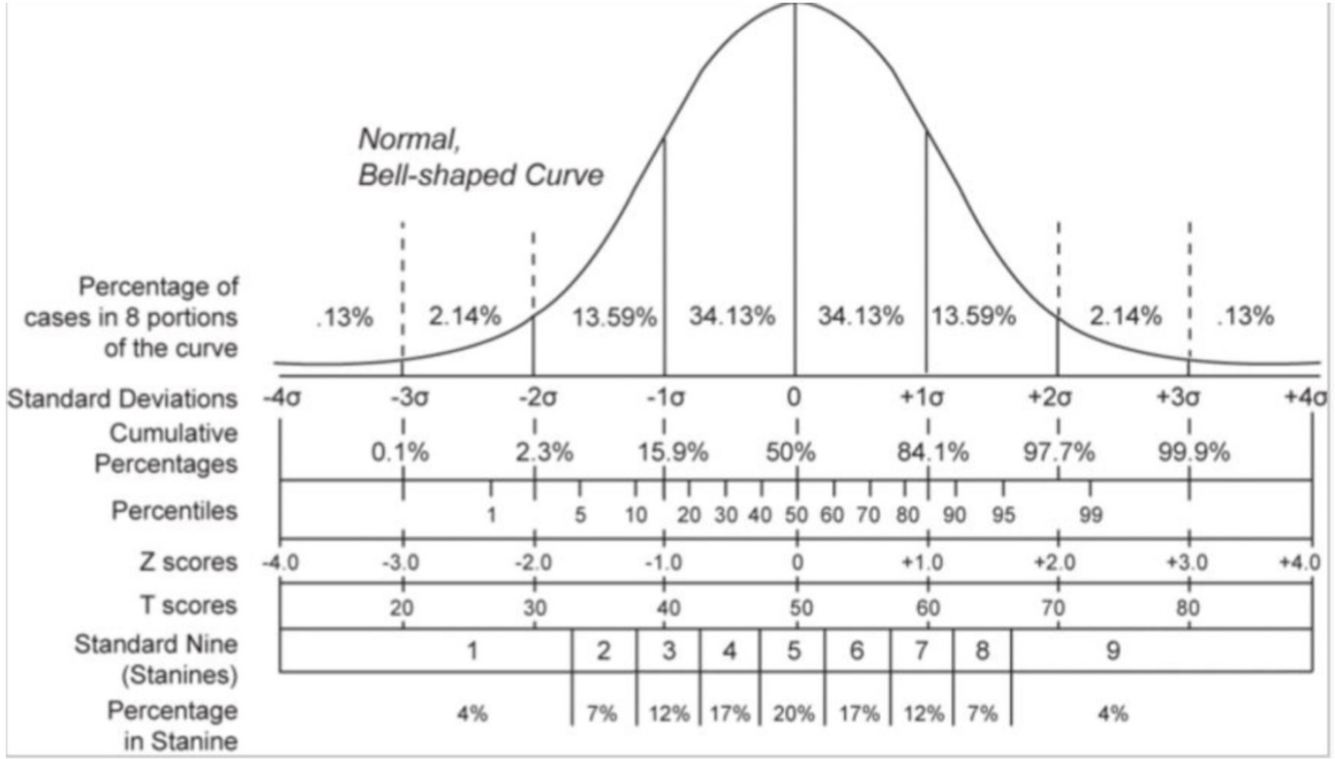
Z-Transformation: $\frac{x-\overline{x}}{s}$



→ Verteilung mit Mittelwert = 0 und s = 1

→ Einheitliche Verteilung = Interpretation und Vergleichbarkeit

Z-Score (wichtigste Folie des Studium)



Covarianz

Die Covarianz leigt zwischen –inf und +inf. Sie zeigt die Korralation zwischen zwei metrischen Variablen an.

Formel: $\frac{1}{n-1}\sum_{i=1}^{1}(x_i - \overline{x}) * (y_i - \overline{y})$

Was heisst das?

Cov = 0 → Kein Zusammenhang

Cov > 0 → positiver Zusammenang

Cov < 0 → negativer Zusammenhang

Korrelationskoeffizient r

Der Koeffizient r liegt zwischen -1 und 1. Er misst lineare zusammenhänge zwischen zwei metrischen variablen.

Formel: $\frac{cov(X,Y)}{\sigma_X * \sigma_Y}$

Was bedeutet das?

r = 0 → kein linearer Zusammenhang

r = -1 → perfekt negativ linearer Zusammenhang

r = 1 → perfekt positiv linearer Zusammenhang

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 7

Ereignis- und Ergebnisraum

Jede Zufallsvariable hat ein Ergebnisraum.

Ergebnisraum

- **Menge aller möglicher Ausprägungen**
- **Disjunkt!**

Beispiel Würfel: 1, 2, 3, 4, 5, 6

Eine einzelne Auprägung ist ein Ergebnis (oder Elementarergebnis).

Beispiel Würfel: 4

Eine Teilmenge des Ergebnisraums ist ein Ereignis.

Beispiel Würfel: 2,5,6

Ereignis- und Ergebnisraum

Ergebnisse haben KEINE Wahrscheinlichkeit!

(Sie sind schlicht Elemente der Zufallsvariable)

Ereignisse haben aber eine Wahrscheinlichkeit.

Beispiel:

P(Elementarergebnis 1) = {}

P(1) = 1/6

P(1, 5) = 2/6

Kolmogorov-Axiome

Axiome müssen erfüllt sein, um mit Wahrscheinlichkeiten rechnen zu können.

Positivität

Wahrscheinlichkeit für das Eintreten eines Elementarereignisses >= 0

Additivität

Wahrscheinlichkeit eines Ergebnisses (Teilmenge) entspricht der Summe der einzelnen Wahrscheinlichkeiten.

Normiertheit

Summe der Wahrscheinlichkeiten aller möglichen Ausprägungen = 1

Gesetz der grossen Zahlen

Wird ein Zufallsexperiment wiederholt, so nähert sich die relative Häufigkeit der Wahrscheinlichkeit an.

→ **Je grösser n, desto genauer die Schätzung!**

→ **Bei hohem n ist die relative Häufigkeit eine gute Schätzung für P(X)**

```
> mean(samp10)
[1] 1.4
> mean(samp100)
[1] 1.44
> mean(samp1000)
[1] 1.498
```

Mengenoperationen

Komplement

- Definition: Gegenereignis
- Beispiel: Komplement zu 6 = 1-5
- Wahrscheinlichkeit: $P(A') = 1 - P(A)$

Vereinigung

- Definition: Outome A oder B
- Beispiel: 2 oder 4
- Wahrscheinlichkeit: $P(A \cup B) = P(A) + P(B)$

Durchschnitt

Teilmenge

Bedingte Wahrscheinlichkeiten I

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \rightarrow P(A \cap B) = P(A|B) * P(B)$$

Additionsregel 2

Wir erinnern uns: $P(A \cup B) = P(A) + P(B)$

Das gilt jedoch nur, wenn sich A und B ausschliessen, also $P(A \cap B) = \{\}$

Allgemein für diskunkte und nicht disjunkte Ereignisse gilt: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bayes-Theorem

Bayes-Theorem ist die **WICHTIGSTE** Regel der bedingten Wahrscheinlichkeiten.

$$\boldsymbol{P(A|B) = \frac{P(A) * P(B|A)}{P(B)}}$$

Bayes-Theorem

Übung aus der Prüfung 2022:

Drei Fabriken stellen Glühbirnen her, um den Markt zu beliefern. **Fabrik A produziert 20%, Fabrik B 50% und Fabrik C 30% aller Glühbirnen.**

2% der in Fabrik A produzierten Glühbirnen, 1% der in Fabrik B produzierten Glühbirnen und 3% der in Fabrik C produzierten Glühbirnen sind defekt.

Sie kaufen zufällig eine Glühbirne im Supermarkt (ohne zu wissen, wer der Hersteller ist) und stellen fest, dass sie defekt ist.

Wie gross ist die Wahrscheinlichkeit (in Prozent!), dass diese Glühbirne von Fabrik B produziert wurde?

Bayes-Theorem

Übung aus der Prüfung 2022:

Drei Fabriken stellen Glühbirnen her, um den Markt zu beliefern. **Fabrik A produziert 20%, Fabrik B 50% und Fabrik C 30% aller Glühbirnen.**

2% der in Fabrik A produzierten Glühbirnen, 1% der in Fabrik B produzierten Glühbirnen und 3% der in Fabrik C produzierten Glühbirnen sind defekt.

Sie kaufen zufällig eine Glühbirne im Supermarkt (ohne zu wissen, wer der Hersteller ist) und stellen fest, dass sie defekt ist.

Wie gross ist die Wahrscheinlichkeit (in Prozent!), dass diese Glühbirne von Fabrik B produziert wurde?

Lösung: $\frac{0.5 * 0.01}{0.2*0.02+0.5*0.01+0.3*0.03} \approx 0.8$

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 8

Wahrscheinlichkeitsmassenfunktion

Die WMF ordnet jeder Ausprägung einer Zusaffalsvariable eine Wahrscheinlichkeit zu.

(Falls Ausprägung kein Element der Zufallsvariable ist, ist die Wahrscheinlichkeit 0)

Haben alle Ausprägungen die selbe Wahrscheinlichkeit, so spricht man von einer **Gleichverteilung**.

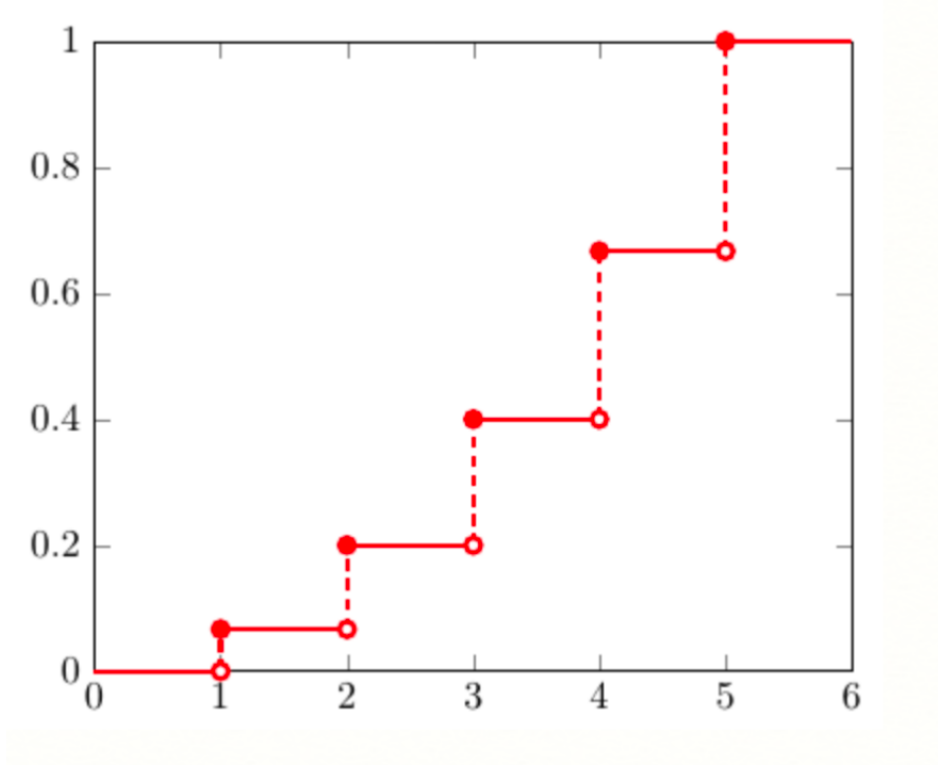
(Bspw. Würfel)

Kummulative Verteilungsfunktion

Simpel: Kummulation der Wahrscheinlichkeiten

Bei diskkreten Zufallsvariablen steigt die Funktion Ruckartig an.

Beispiel Würfel: P(X<3) = 1/3



Erwartungswert diskreter Zufallsvariable

Erwartungswert: Mittelwert auf lange Sicht

→ Gesetz der grossen Zahlen!

Vereinfacht: Wert * Wahrscheinlichkeit aufsummiert

Übung:

Was ist der Erwartungswert eines Würfelwurfes?

Erwartungswert diskreter Zufallsvariable

Erwartungswert: Mittelwert auf lange Sicht

→ Gesetz der grossen Zahlen!

Vereinfacht: Wert * Wahrscheinlichkeit aufsummiert

Übung:

Was ist der Erwartungswert eines Würfelwurfes?

3.5 → bei Gleichverteilung = Median

Wahrscheinlichkeitsdichtefunktion

Metrische Variablen haben **KEINE** Massenfunktion. Stattdessen haben sie eine **Dichtefunktion**.

Die Funktion zeigt die **Dichte** der Wahrscheinlichkeit an, nicht die Wahrscheinlichkeit selbst!

→ Analog Histogramm; Fläche = Wahrscheinlichkeit

Bei *Gleichverteilung* ist die Fläche einfach zu errechnen. Sonst muss das Integral berechent werden.

Statistische Unabhängigkeit

Statistische Unabhängigkeit gilt, wenn A und B nicht miteinander korrelieren, bzw sich gegenseitig beeinflussen.

Daraus lässt sich ableiten:

$P(A|B) = P(A)$

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 9

Varianz bei Zufallsvariablen

Bei Zufallsvariablen ändert sich die Formel der Varianz leicht, denn die Wahrscheinlichkeiten müssen berücksichtigt werden.

$$\sigma^2 = \sum_j (x_j - \mu)^2 f(x_j)$$

$$\sigma^2 = \int (x_j - \mu)^2 f(x) dx$$

Rechnen mit Erwartungswerten

Lineartransformation

- X lässt sich mit E[X] ersetzen
- 3 - 9 * X = 3 – 9 * E[X]

Multiplikation (unabhängig)

- X und Y lässt sich mit E[X] und E[Y] ersetzen
- X * Y = E[X] * E[Y]

Multiplikation (abhängig)

- Covarianz muss berücksichtigt werden
- X * Y = E[X] * E[Y] + Cov(X,Y)

Rechnen mit Erwartungswerten

Varianz

- X lässt sich mit E[X] ersetzen
- Var = $E[X^2] – (E[X])^2$

$$E[X^2] = \sum_{x=1}^{6} x^2 P_X(x) = 1^2(\frac{1}{6}) + 2^2(\frac{1}{6}) + ... + 6^2(\frac{1}{6}) = 15.167$$

Verteilungen: Ein Überblick

Diskrete Zufallsvariablen

- Bernouilli (2 Ausprägungen; n = 1)
- Binomial (2 Ausprägungen; n > 1)
- Poisson (Mehr als 2 Ausprägungen)

Metrische Zufallsvariablen

- Chi-Quadrat
- Normalverteilung
- t-Verteilung
- f-Verteilung

Normalverteilung

Notation: $X \sim \mathcal{N}(\mu, \sigma)$

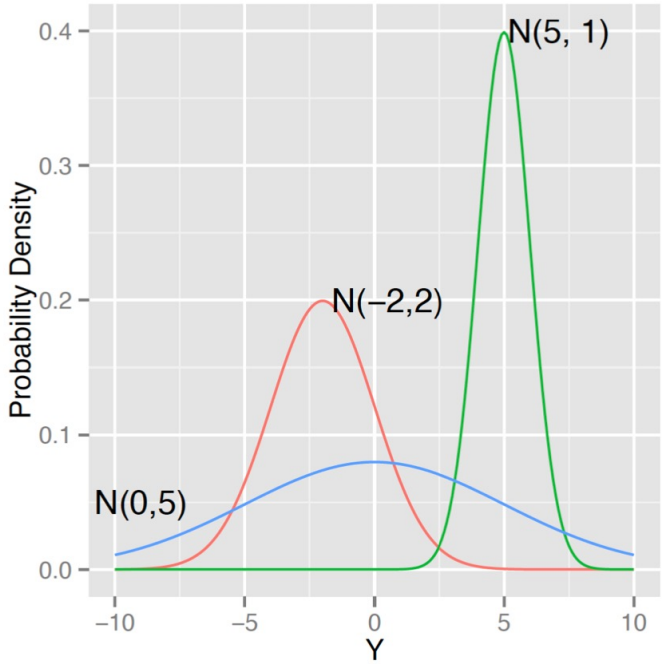
→ Die Normalverteilung hat zwei Parameter: Mittelwert und Standardabweichung

Eigenschaften

- Stetig und symetrisch
- Glocken-Form
- Modus = Median = artithmetischer Mittelwert

Weshab interssiert alle diese Verteilung?

- Viele Eigenschaften sind natürlicherweise Normalverteilt.
- Vergleichbarkeit mit Standardnormalverteilung



Standardnormalverteilung: Die Wahrscheinlichkeit

Remember: Bei stetigen Variablen ist die Wahrscheinlichkeit das Integral.

Aber: Das ist echt mühsam zu berechnen.

Lösung: **z-Statistik Tabelle**

Z-Transformation haben wir schon mal gemacht: $$Z = \frac{X - \bar{\mu}}{\sigma}$$

Übung:

Was ist der z-Score für X = 21.5 bei N(30, 4.3)?

Standardnormalverteilung: Die Wahrscheinlichkeit

Remember: Bei stetigen Variablen ist die Wahrscheinlichkeit das Integral.

Aber: Das ist echt mühsam zu berechnen.

Lösung: **z-Statistik Tabelle**

Z-Transformation haben wir schon mal gemacht: $$Z = \frac{X - \mu}{\sigma}$$

Übung:

Was ist der z-Score für X = 21.5 bei N(30, 4.3)?

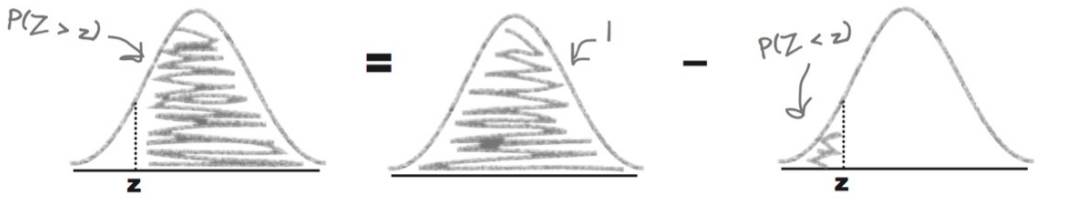
(21.5 - 30) / 2.07 = - 4.106

Standardnormalverteilung: Die Wahrscheinlichkeit

Die Wahrscheinlichkeit muss je nach Fragestellung anders berechnet werden.

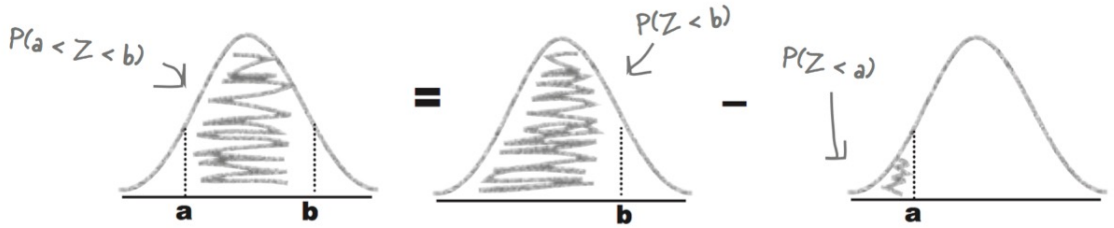
P(Z > z) = 1 - P(Z < z) ← taller than her date.

In other words, take the area where Z < z away from the total probability.



P(a < Z < b) = P(Z < b) - P(Z < a) ← probability that a date is within a particular range.

In other words, calculate P(Z < b), and take away the area for P(Z < a).

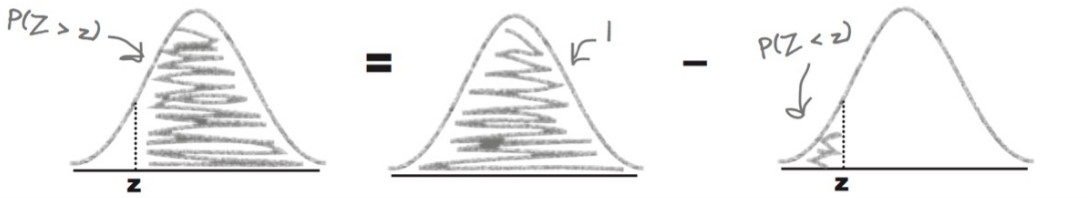


Standardnormalverteilung: Die Wahrscheinlichkeit

Die Wahrscheinlichkeit muss je nach Fragestellung anders berechnet werden.

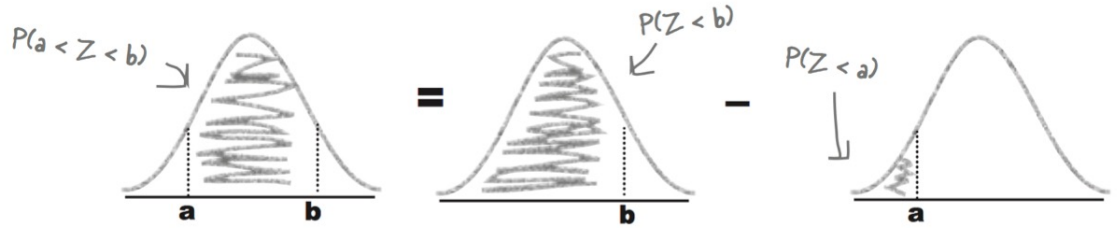
P(Z > z) = 1 - P(Z < z) ← taller than her date.

In other words, take the area where Z < z away from the total probability.



P(a < Z < b) = P(Z < b) - P(Z < a) ← probability that a date is within a particular range.

In other words, calculate P(Z < b), and take away the area for P(Z < a).



Wichtigster z-Score zum Merken: **1.96**!!

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 10

Zufallsauswahl

Jedes Element der Grundgesamtheit hat die gleiche oder bekannte Chance (>0) erhoben zu werden!

Vorteil Zufallsauswahl:
Stichproben variieren immer. Zufallsstichproben variieren zufällig → wir können berechnen

Zentraler Grenzwertsatz

Jede Stichprobenstatistik nähert sich einer Normalverteilung an.

Arten der Stichprobeninferenz

Stichproben Mittelwerte

Beispiel: Link-rechts Selbsteinschätzung

Stichproben Anteile

Beispiel: Ja-Anteil Klimaschutzgesetz

Arten der Stichprobeninferenz

Stichproben Mittelwerte

Beispiel: Link-rechts Selbsteinschätzung

Stichproben Anteile

Beispiel: Ja-Anteil Klimaschutzgesetz

Vorgehen bei Prüfungsfragen

Schritt 1: Um welche Art handelt es sich?

Schritt 2: Entsprechende Formel heraussuchen.

Stichproben: Mittelwert

Betrachtet man Mittelwerte, so haben auch die eine Verteilung.

$$Var(\bar{X}) = \frac{\sigma_X^2}{n}$$

$$\sqrt{\frac{\sigma_x^2}{n}} = \frac{\sigma_x}{\sqrt{n}}$$ = Standardfehler (se)

Wenn $X \sim N(\mu, \sigma^2)$, dann gilt: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Stichproben: Anteil

Der Standardfehler bei Anteilen errechnet sich relativ simpel:

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$$

Kurze Übung

Wie lautet die Stichprobenverteilung?

Übung 1: $X \sim N(26, 7); n = 12$

Übung 2: $Var(X) = 10,\ sd = 3, n = 46$

Kurze Übung

Wie lautet die Stichprobenverteilung?

Übung 1: $X \sim N(26, 7); n = 12$

Lösung: $N(26, \frac{7}{\sqrt{12}})$

Übung 2: $Mean(X) = 10,\ Var(X) = 3, n = 46$

Lösung: $N(10, \frac{3}{\sqrt{46}})$

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 11

Konfidenzintervall

Was ist das Konfidenzintervall?

Konfidenzintervall

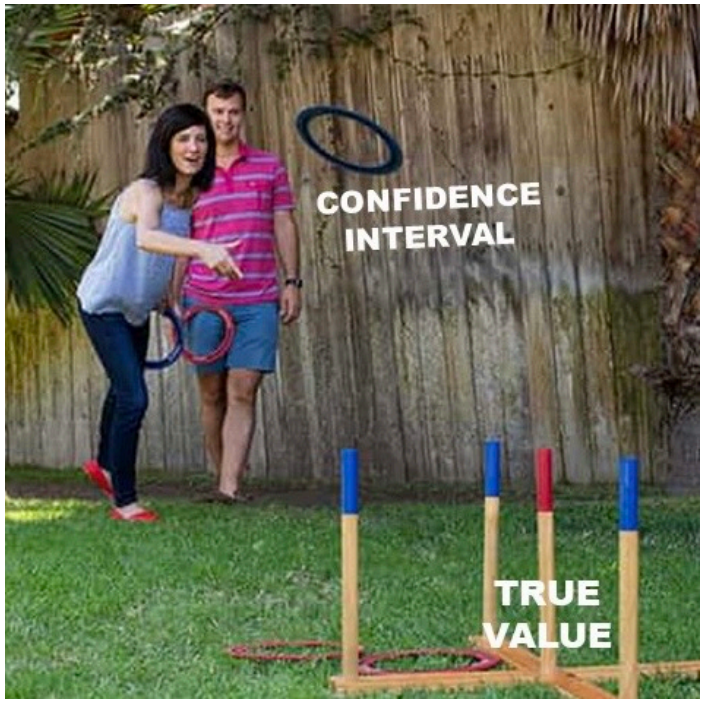
Was ist das Konfidenzintervall?

Konfidenzintervall

Was ist das Konfidenzintervall?

Konfidenzintervall

Was ist das Konfidenzintervall?



Twitter Thread:
https://twitter.com/kareem_carr/status/1662102939144687617?s=20

Konfidenzintervall: Niveau

Das Konfidenzniveau gibt an wie sicher wir uns sein wollen, dass der wahre Wert im KI beinhaltet ist.

Je grösser das Konfidenzniveau, desto schwerer werden Signifikanzaussagen.

Forschungsstandard: 95%

Es gibt aber auch Forschende, die mit Konfidenz zwischen 90% und 99% arbeiten.

Konfidenzintervall: Berchenen

Allgemeine Formel: $\mu \pm z * se$

z-Wert für das jeweilige Konfidenzintervall. Also bei 90% suchen wir Wert für 0.05 und 0.95.

Konfidenzintervall: Berchenen

Allgemeine Formel: $\mu \pm z * se$

z-Wert für das jeweilige Konfidenzintervall. Also bei 90% suchen wir Wert für 0.05 und 0.95.

Aufgabe aus der letzten Prüfung:

Wir führen eine Vorwahlbefragung zur Initiative «Gegen den F-35 (Stopp F-35») durch. Der Stichprobenumfang unserer Befragung beträgt 900 Befragte. In unserer Stichprobe geben 45 Prozent der Befragten an, sie würden der Initiative zustimmen.

Wie viel beträgt der Stichprobenfehler (margin of error) dieses Anteilwertes (in Prozentpunkten) bei einem vorgegebenen Konfidenzniveau von 95%?

$$1.96 * \sqrt{\frac{0.45 * 0.55}{900}} \approx 0.0325$$

Universität Zürich[UZH]

Institut für Politikwissenschaft Zürich IPZ

Vorlesung 12

Nullhypothese

Als Nullhypothese bezeichnet man die **Grundannahme** → Kein Zusammenhang

Das Ziel von Hypothesentest ist es, zu überprüfen ob die Nullhypothese standhält.

Nullhypothese wird erst verworfen, wenn die Alternativhypothese mit genügend Konfidenz standhält.

Arten des Hypothesentest

Einseitig

- Test kann links- oder rechtsseitig ein.
- Linksseitig → H_A ist kleiner als H_0
- Rechtsseitig → H_A ist grösser als H_0

Beidseitig

- Der häufigste Test, welcher angewandt wird.
- H_A ist ungleich H_0 → grösser oder kleiner

Hypothesentest anwenden

Wir errechnen den z-Wert folgendermassen

Mittelwert der Grundgesamtheit (μ), wobei Nullhypothese $\mu = \mu_0$:

$$z_\mu = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Anteilswert der Grundgesamtheit (p), wobei Nullhypothese $p = p_0$:

$$z_p = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}$$

Wichtig: 95% Intervall (zweiseitig) hat z = 1.96; 95% Intervall (einseitig) hat z = 1.645

p-Wert

Aus dem z-Wert lässt sich der p-Wert errechenen.

Der p-Wert gibt die Wahrscheinlichkeit an, dass sich H_A und H_0 überlappen.

Doch brauchen wir den p-Wert eigentlich? Nein!

→ Zeit sparen und nur z-Wert kennen.

Entscheidungsfehler

Fehler der 1. Art

– Fälschliche Ablehnung der Nullhypothese
– Wahrscheinlichkeit Typ 1 Fehler = alpha

Fehler der 2. Art

– Fälschliche beibehaltung der Nullhypothese

Da wir in der Forschung immer **konservativ schätzen**, ist Typ 1 Fehler schlimmer als Typ 2.

t-Test vs z-Test

Wichtig für die Prüfung zu wissen:

Ab n >= 30 kann man auch z-Verteilung herbei ziehen. Darunter immer t-Test.

Ein paar Prüfungsfragen

```
        One Sample t-test

data:  anes08$obamafeel
t = 24.136, df = 2292, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 63.11257 65.43170
sample estimates:
mean of x
 64.27213
```

Frage 1

In der amerikanischen Wahlumfrage von 2008 wurden 2293 Teilnehmer gebeten, eine Bewertung von Barack Obama auf einem 0-100 Gefühlsthermometer abzugeben. Die übliche Auslegung dieser Skala ist, dass Werte unter 50 negative Gefühle repräsentieren und diejenigen über 50 positive Gefühle. Wir führen einen t-Test durch und erhalten folgende Resultate:

Geben sie an welche der folgenden Aussagen richtig und welche falsch sind.

- Im Durchschnitt hat das Sample negative Gefühle gegenüber Obama.
- Der Standardfehler beträgt etwa 0.58.
- Die Nullhypothese, die getestet wird, ist $\mu \leq 50$.
- Um den p-Wert zu erhalten, hätten wir in diesem Fall anstelle einer t-Verteilung auch die Standardnormalverteilung verwenden können, da bei hinreichend hoher Fallzahl t- und Z-Verteilungen fast identisch sind.

```
        One Sample t-test

data:  anes08$obamafeel
t = 24.136, df = 2292, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 63.11257 65.43170
sample estimates:
mean of x
 64.27213
```

Frage 1

In der amerikanischen Wahlumfrage von 2008 wurden 2293 Teilnehmer gebeten, eine Bewertung von Barack Obama auf einem 0-100 Gefühlsthermometer abzugeben. Die übliche Auslegung dieser Skala ist, dass Werte unter 50 negative Gefühle repräsentieren und diejenigen über 50 positive Gefühle. Wir führen einen t-Test durch und erhalten folgende Resultate:

Geben sie an welche der folgenden Aussagen richtig und welche falsch sind.

- Im Durchschnitt hat das Sample negative Gefühle gegenüber Obama.
- Der Standardfehler beträgt etwa 0.58.
- Die Nullhypothese, die getestet wird, ist $\mu \leq 50$.
- Um den p-Wert zu erhalten, hätten wir in diesem Fall anstelle einer t-Verteilung auch die Standardnormalverteilung verwenden können, da bei hinreichend hoher Fallzahl t- und Z-Verteilungen fast identisch sind.

Frage 2

Sie lesen in einer Auswertung von Befragungsdaten, dass das Variationsverhältnis (variation ratio) der Variablen «Konfession» .63 beträgt. Was heisst das? Welche der folgenden Aussagen sind in diesem Zusammenhang richtig und welche falsch?

- Die Abweichung von der Modalkategorie beträgt im Schnitt .63 Standardabweichungen.
- Die Modalkategorie der Variablen Konfession enthält 37 Prozent aller Fälle
- Würden wir in einem Gedankenspiel für alle Befragten den Modalwert der Konfession voraussagen, so würden wir in 63 Prozent aller Fälle richtig liegen.
- 63 Prozent aller Fälle sind ausserhalb der Modalkategorie.

Frage 2

Sie lesen in einer Auswertung von Befragungsdaten, dass das Variationsverhältnis (variation ratio) der Variablen «Konfession» .63 beträgt. Was heisst das? Welche der folgenden Aussagen sind in diesem Zusammenhang richtig und welche falsch?

- Die Abweichung von der Modalkategorie beträgt im Schnitt .63 Standardabweichungen.
- Die Modalkategorie der Variablen Konfession enthält 37 Prozent aller Fälle
- Würden wir in einem Gedankenspiel für alle Befragten den Modalwert der Konfession voraussagen, so würden wir in 63 Prozent aller Fälle richtig liegen.
- 63 Prozent aller Fälle sind ausserhalb der Modalkategorie.

Frage 3

An einem Sportwettbewerb (z.B. Kunstturnen) nahmen 5 Männer und 7 Frauen teil. Die Männer erzielten einen Durchschnittswert (=arithmetischer Mittelwert) von 16.42, während die Summe aller Einzelwertungen für die Frauen 121 betrug. Wie lautet der arithmetische Mittelwert aller Teilnehmenden, das heisst von Frauen und Männer zusammen?

– 14.925
– 16.925
– 13.925
– 15.925

Frage 3

An einem Sportwettbewerb (z.B. Kunstturnen) nahmen 5 Männer und 7 Frauen teil. Die Männer erzielten einen Durchschnittswert (=arithmetischer Mittelwert) von 16.42, während die Summe aller Einzelwertungen für die Frauen 121 betrug. Wie lautet der arithmetische Mittelwert aller Teilnehmenden, das heisst von Frauen und Männer zusammen?

– 14.925
– 16.925
– 13.925
– 15.925

Frage 4

Betrachten Sie folgende Tabelle, welche die gemeinsame Verteilung der beiden Variablen Parteizugehörigkeit («Gelbe» und «Grüne») und die Haltung zu irgendeiner beliebigen Massnahme x (Zustimmung vs. Ablehnung dieser Massnahme) zeigt.

Welche der folgenden Aussagen im Zusammenhang mit diesen Tabellenwerten ist korrekt?

– Die «Gewinnchancen» oder Odds, der Massnahme x zuzustimmen, sind für die Gelben rund 42 Mal höher als für die Grünen.

– Das relative Risiko zwischen Grünen und Gelben der Massnahme x zuzustimmen, beträgt rund 42%.

– Die «Gewinnchance» oder Odds, der Massnahme x zuzustimmen, sind für die Gelben rund 7 Mal höher als für die Grünen.

	"Gelbe"	"Grüne"
Zustimmung	187	28
Ablehnung	31	193

Frage 4

Betrachten Sie folgende Tabelle, welche die gemeinsame Verteilung der beiden Variablen Parteizugehörigkeit («Gelbe» und «Grüne») und die Haltung zu irgendeiner beliebigen Massnahme x (Zustimmung vs. Ablehnung dieser Massnahme) zeigt.

Welche der folgenden Aussagen im Zusammenhang mit diesen Tabellenwerten ist korrekt?

– Die «Gewinnchancen» oder Odds, der Massnahme x zuzustimmen, sind für die Gelben rund 42 Mal höher als für die Grünen.

– Das relative Risiko zwischen Grünen und Gelben der Massnahme x zuzustimmen, beträgt rund 42%.

– Die «Gewinnchance» oder Odds, der Massnahme x zuzustimmen, sind für die Gelben rund 7 Mal höher als für die Grünen.

	"Gelbe"	"Grüne"
Zustimmung	187	28
Ablehnung	31	193

Letzte Fragen?

Viel Erfolg!