Basiswissen für Forschende 2

Prof. Dr. Reyn van Ewijk

Lehrstuhl für Statistik und Ökonometrie

JG|U

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Gliederung

1. Typen von Studien
    a. Qualitative Studien
    b. Quantitative Studien
        i. Deskriptive Studien
        ii. Korrelationsstudien
        iii. Kausalstudien
2. Daten: Quellen, Analysemethoden & Variablen
3. Analyse Quantitativer Daten
    a. Deskriptive Statistik
    b. Experimente: Design & Datenanalyse

Literatur zu diesem Foliensatz

1. Planing, P. (2022). *Statistik Grundlagen: das interaktive Lehrbuch mit über 150 YouTube-Videos rund um die Burgerkette FIVE PROFS*. Planing Publishing.

   Kostenlos online verfügbar über: https://statistikgrundlagen.de/ebook

   - **Kapitel 2**
   - **Kapitel 3: nur 3.11**
   - **Kapitel 4 – nicht 4.5, 4.8 und 4.9**
   - **Kapitel 6**

2. Stock JH, Watson MM. (2019). *Introduction to Econometrics.* 4th edition. Pearson

   - Digitale Version über die Universitätsblibliothek
   - Einige Exemplare sind in unseren Bibliotheken verfügbar
   - **Kapitel 3.1, 2.1 und 2.2**

Gliederung: Analyse Quantitativer Daten

1. Deskriptive Statistik

Deskriptive Statistik

- Ziel: Aussagen über eine Grundgesamtheit
- Hierzu: Ziehen einer Stichprobe $Y_1, Y_2, \ldots, Y_n$ mit Y_i i.i.d. (*)
- Stichprobengröße: n
- Für jede Einheit i werden Variablen erhoben
- Einheiten:
  - Individuen, Haushalten, Firmen, …
  - Zeilen im Datensatz
- Variablen: Spalten im Datensatz

(*) i.i.d. = *independently and identically distributed,* unabhängig und identisch verteilt

Deskriptive Statistik

- Ziel: Aussagen über eine Grundgesamtheit
- Hierzu: Ziehen einer Stichprobe $Y_1, Y_2, \ldots, Y_n$ mit Y_i i.i.d. (*)
- Wichtigster Kennwert: Mittelwert

Theoretisch	Empirisch
Erwartungswert: $E(Y) = \mu_Y$	Mittelwert: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

  - Theoretisch = Eigenschaft einer Population, die in ihrer Gesamtheit nicht beobachtbar ist
  - Empirisch = beobachtet, bzw. Eigenschaft einer Stichprobe

(*) i.i.d. = *independently and identically distributed,* unabhängig und identisch verteilt

Der Stichprobenmittelwert als Schätzer des Erwartungswertes

- Stichprobenmittelwert $\bar{Y} = \hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- Ist ein „unverzerrter“ Schätzer
  - Liefert im Schnitt den richtigen Wert (bei unendlicher Anzahl von Stichproben)
  - Verzerrung (*bias*): $E(\hat{\mu}_Y) - \mu_Y$
- Allgemein: Eigenschaften eines guten Schätzers
  1. Er sollte im Durchschnitt den richtigen Wert liefern (**Unverzerrtheit/Erwartungstreue**)
  2. Die Schätzunsicherheit sollte mit der Stichprobengröße abnehmen (**Konsistenz**)
  3. Er sollte eine geringe Varianz besitzen (**Effizienz**)

→ Aussagen über die **Stichprobenverteilung** des Schätzers

Eigenschaft 2: Konsistenz eines Schätzers

- **Konsistenter** Schätzer = liegt in großen Stichproben mit sehr großer Wahrscheinlichkeit in der Nähe des zu schätzenden Parameters liegt
  - Formal: Der Schätzer **konvergiert in Wahrscheinlichkeit** gegen den zu schätzenden Parameter
- Hier:
  - Umso größer die Stichprobe (N), desto sicherer kann man davon sein, dass der Stichprobenmittelwert $\bar{Y}$ sehr Nahe am Mittweltwert der Population μ_Y liegt
  - $\bar{Y}$ ist ein konsistenter Schätzer von μ_Y (Gesetz der Großen Zahlen):

$$\bar{Y} \xrightarrow{p} \mu_Y$$

Eigenschaft 3: Effizienz eines Schätzers

- Effizienz = optimale Nutzung der Daten
- Schätzunsicherheit wird minimiert
- Gemessen an der **Varianz** des Schätzers
  - $var(\bar{Y})$, nicht $var(Y)$ – später mehr
  - $var(\bar{Y}) = \sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$
- Gegenbeispiele:
  - Hälfte der Daten nicht benutzen
  - Die eine Hälfte der Beobachtungen erhält eine Gewichtung von $\frac{1}{2}$, die andere Hälfte von $\frac{3}{2}$

Die 3 Kennwerte der Zentralen Tendenz

1. **Mittelwert**
2. **Median** – der Wert, der eine der Größe nach geordnete Häufigkeitsverteilung, in zwei gleichgroße Hälften teilt
   - Bei n Beobachtungen,n ungerade Zahl: Wert von Person $\frac{n+1}{2}$
   - Bei n Beobachtungen,n gerade Zahl: Mittelwert von Personen $\frac{n}{2}$
   - Bietet Vorteile, wenn es relativ viele extreme Werte (Ausreißer) gibt
3. **Modus (Modalwert)** – der Wert, der in der Verteilung einer diskreten Variable am häufigsten vorkommt

Kennwerte der Zentralen Tendenz

- Ergänze die Tabelle
  - Ja = Kann man hier benutzen
  - Nein = Kann man hier nicht benutzen

Skalenniveau der Variable	Modus	Median	Mittelwert
Nominal	Ja		
Ordinal			
Intervall			
Ratio			Ja

SurveyMonkey:
Kennwerte der zentralen Tendenz

Der Stichprobenmittelwert ist BLUE

- Wieso benutzen wir fast immer den Mittelwert??
  - Schätzer, die auf Mittelwerten basieren haben eine große Bedeutung in der Statistik
  - Grund: Sie haben 6 wünschenswerte Eigenschaften
- Der Stichprobenmittelwert ist ein:
  1. unverzerrter/erwartungstreuer,
  2. konsistenter,
  3. (relativ) effizienter,
  4. bester linearer unverzerrter (BLUE),
  5. Kleinste-Quadrate-

  Schätzer von μ_Y

„Metrische“ oder „Parametrische“ Variablen

- Beachten:
  - Mittelwerte machen nur bei Intervall- & Ratioskalen Sinn
  - Bei Nominal- und Ordinalskalen ergeben Mittelwerte keinen Sinn

- Intervall- & Ratioskala:
  - Unterschied ist in der Praxis meist eher von theoretischem Interesse
  - Werden zusammen als „**metrisch**“ oder „**parametrisch**“ bezeichnet

Der Stichprobenmittelwert ist BLUE

- Der Stichprobenmittelwert ist eine lineare Funktion der Daten:

$$\bar{Y} = \hat{\mu}_Y = \sum_{i=1}^{n} a_i \cdot Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

- z.B. Alternativ: $\tilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \frac{1}{2}Y_3 + \frac{3}{2}Y_4 + \cdots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right)$

  (n = gerade Zahl)

- $\bar{Y}$ besitzt unter allen **linear unverzerrten** Schätzern die minimale Varianz
- $\bar{Y}$ ist der **BLUE** [Bester Linearer Unverzerrter Schätzer, **Best Linear Unbiased Estimator**] von μ_Y

- $\tilde{Y}$ ist ebenfalls unverzerrt, da $E\left(\tilde{Y}\right) = \mu_Y$
- $\tilde{Y}$ ist aber weniger effizient als $\bar{Y}$

Kleinste-Quadrate-Schätzer (KQ-Schätzer)

- $\bar{Y}$ ist der **Kleinste-Quadrate-Schätzer** (*least squares estimator*) des Erwartungswerts μ_Y
  - Der **KQ-Schätzer** ist der Schätzer (m), der die Summe der quadrierten Abweichungen von den beobachteten Daten minimiert: $\sum_{i=1}^{n}(Y_i - m)^2$

- Beweis: First order condition (FOC): $-2\sum_{i=1}^{n}(Y_i - m) = 0$

$$-2 \cdot \sum_{i=1}^{n} Y_i + 2 \cdot \sum_{i=1}^{n} m = 0$$

$$-2 * \sum_{i=1}^{n} Y_i + 2n \cdot m = 0$$

$$m = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Gliederung: Analyse Quantitativer Daten

1. Deskriptive Statistik
    a. Schätzer und ihre Eigenschaften: Mittelwerte
    b. Streuungsmaße

Streuungsmaße

- Maße der zentralen Tendenz (Mittelwerte) sind wichtige deskriptive Statistiken
- Aber: Verteilungen mit dem gleichen Mittelwert (usw.) können unterschiedlich gestreut sein

1. Spannweite
2. Interquartilsabstand
3. Varianz
4. Standardabweichung
5. Schiefe
6. Kurtosis

Streuungsmaße

1. Spannweite

- Differenz zwischen dem größten (Maximum) und dem kleinsten Wert (Minimum)
- Beispiel Instagram-Follower: Was ist hier die Spannweite?

– 22	– 98	– 108
– 40	– 103	– 116
– 53	– 1252	– 121
– 93	– 57	– 125

2. Interquartilsabstand (Inter Quartile Range, IQR)

- Spannweite zwischen Beginn des zweiten und Ende des dritten Quartils
- Umfasst die 50% der Werte in der Mitte der Verteilung
- Beispiel Instagram-Follower: Was ist hier der Interquartilsabstand?

Streuungsmaße

1. Spannweite $P_{100} - P_0$
2. Interquartilsabstand $P_{75} - P_{25}$

- Nachteile von 1. & 2.: berücksichtigen nicht alle Werte, sondern jeweils nur zwei Randwerte
- Spannweite ist stark von Extremwerten abhängig

3. Varianz
4. Standardabweichung
5. Schiefe
6. Kurtosis

Streuungsmaße: Die Varianz

- Durchschnittliche quadrierte Abweichung vom Mittelwert
- Stichprobenvarianz (empirische Varianz, *sample variance*):

$$S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

Bitte beachten: Planing (2022) teilt durch n statt durch $n-1$. Dies ergibt S_{XY} für eine Population statt für eine Stichprobe

Da wir immer nur Stichproben haben, teilen wir immer durch $n-1$

$n-1$ ist eine Korrektur der **Freiheitsgrade** (*degrees of freedom*)

- Quadrieren führt dazu, dass:
  - alle Abweichungen positiv werden $s_Y^2 \geq 0$
  - größere Abweichungen stärker gewichtet werden
- S_Y^2 ist ein unverzerrter und konsistenter Schätzer der Varianz der Grundgesamtheit σ_Y^2
- Varianz – Nachteil: Interpretation ist schwierig

Streuungsmaße: Die Standardabweichung

- S oder SD (*standard deviation*)
- Wurzel aus der Varianz: $S_Y = \sqrt{S_Y^2} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$
- Einfacher zu interpretieren, da in derselben Einheit gemessen wie Y
- Größere SD = größere Streuung
- Bei normalverteilten Daten, z.B. IQ ($\mu = 100; \sigma = 15$):
  - Ca. 1/6 aller Personen (16.87%) haben Werte, die >1 SD höher als der Mittelwert sind

    1 von 6 Personen hat ein IQ über 115 und 1/6 hat ein IQ unter 115
  - 2.28% aller Personen haben Werte, die >2 SD höher als der Mittelwert sind

    2.28% haben ein IQ über 130 und 2.28% ein IQ unter 70
  - 0.13% aller Personen haben Werte, die >3 SD höher als der Mittelwert sind

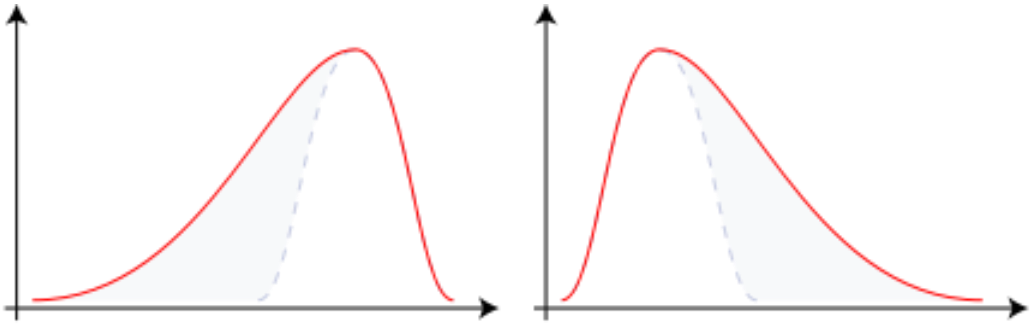
    0.13% haben ein IQ über 145 und 0.13% ein IQ unter 55

Schiefe (skewness)

Populationswert: $\frac{E\left[(Y-\mu)^3\right]}{\sigma_Y^3}$

Stichprobe: $\left(\frac{n}{n-1}\right)^{3/2} \cdot \frac{\frac{1}{n}\sum(Y_i-\bar{Y})^3}{S_Y^3}$

- Maß für die **Symmetrie** der Verteilung
- Symmetrische Verteilung, z.B. Normalverteilung: Schiefe = 0
- Schiefe > 0 $\Longrightarrow$ Verteilung ist **rechtsschief** (z.B. Einkommensverteilung)
- Schiefe < 0 $\Longrightarrow$ Verteilung ist **linksschief**



Linksschief
Schiefe < 0

Rechtsschief
Schiefe > 0

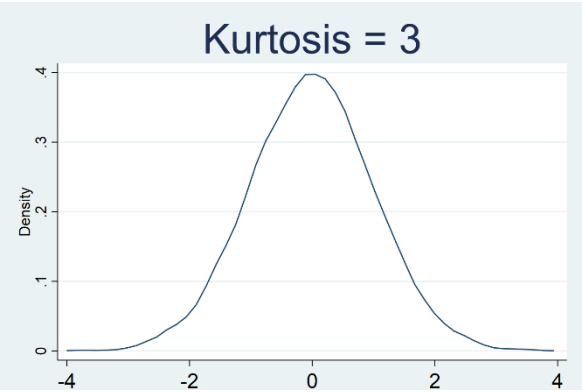
Kurtosis/Wölbung

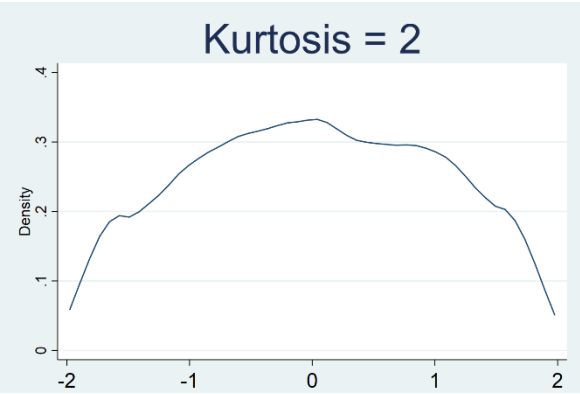
Populationswert: $\frac{E\left[(Y-\mu)^4\right]}{\sigma_Y^4}$

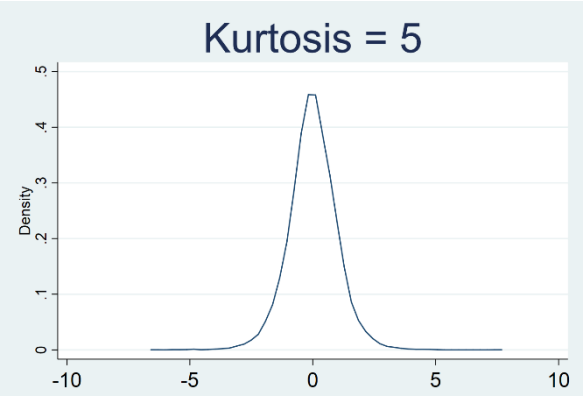
$\Rightarrow$ Immer ≥ 0

Stichprobe: $\left(\frac{n}{n-1}\right)^2 \cdot \frac{\frac{1}{n}\sum(Y_i-\bar{Y})^4}{S_Y^4}$

- Normalverteilung: Kurtosis $= 3$
- Beschreibt Spitzigkeit bzw. Flachheit im Vergleich zur Normalverteilung
- Kurtosis $<\ 3 \Rightarrow$ breiter Gipfel, flache Verteilung
- Kurtosis $>\ 3 \Rightarrow$ schmaler Gipfel, spitze Verteilung

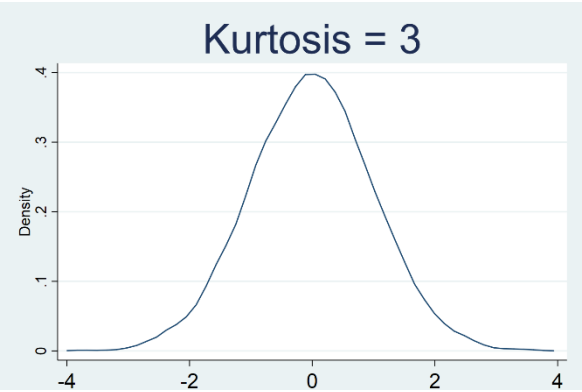


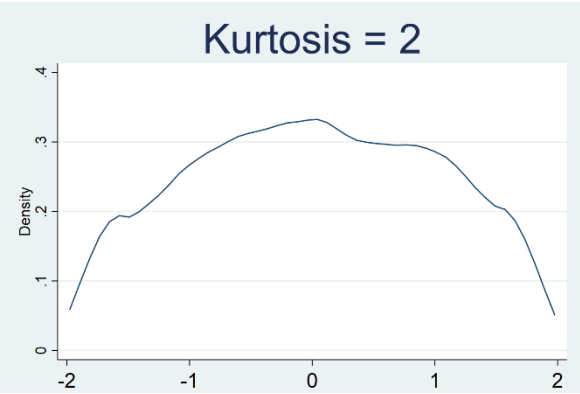


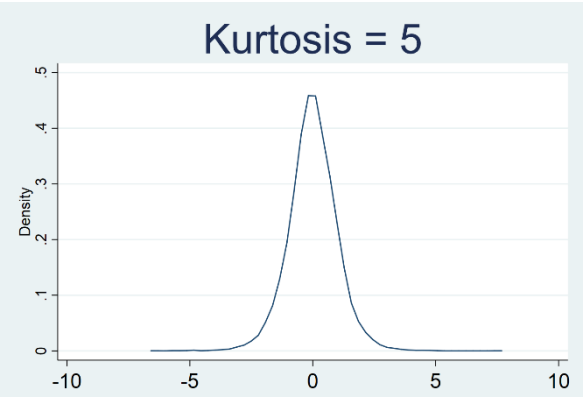


Kurtosis/Wölbung

- Maß für die Masse der Verteilung an den Enden, d.h. für die Wahrscheinlichkeit extremer Ausprägungen („**Ausreißer**“, *outlier*)
- Kurtosis > 3: relativ viele Ausreißer
- Wenn der Wertebereich einer Variable beschränkt ist (z.B. Noten) ist Kurtosis normalerweise < 3







Quiz: Schiefe & Kurtosis der Notenverteilung EmpWi

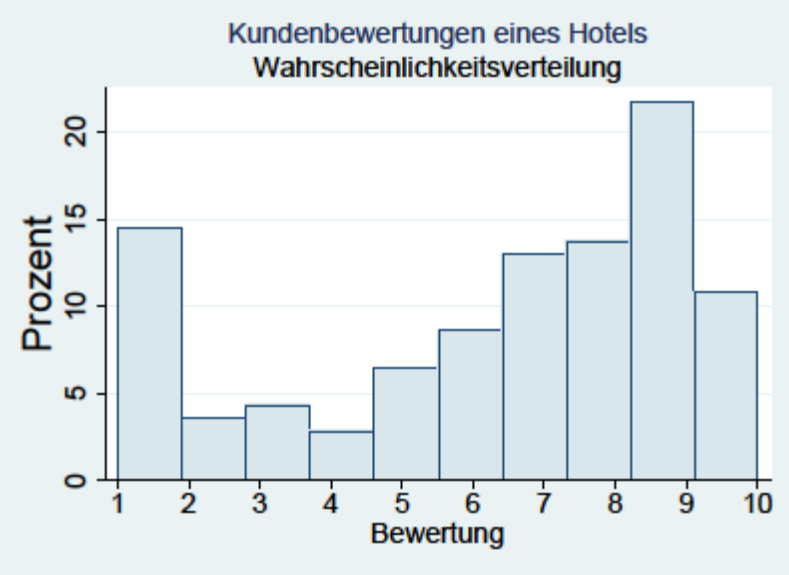
SurveyMonkey:
Notenverteilung Klausur 1
Notenverteilung Klausur 2

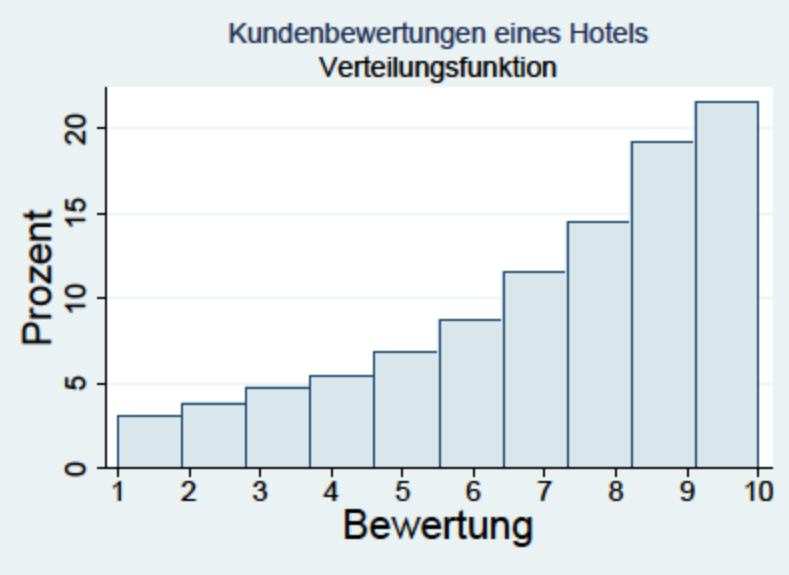
Gliederung: Analyse Quantitativer Daten

1. Deskriptive Statistik
    a. Schätzer und ihre Eigenschaften: Mittelwerte
    b. Streuungsmaße
    c. Verteilungen

Verteilungen – Diskrete Variablen

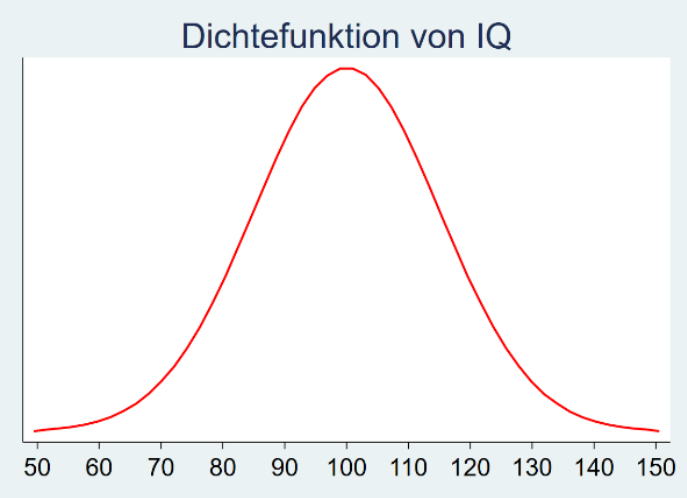
- Diskrete Variablen:
  - **Wahrscheinlichkeitsverteilung** – jedem diskreten Ergebnis ist eine Wahrscheinlichkeit zugeordnet
  - **Verteilungsfunktion** – Wahrscheinlichkeit, dass die Zufallsvariable kleiner oder gleich einem bestimmten Wert y ist (*cumulative distribution function,* ***cdf***)

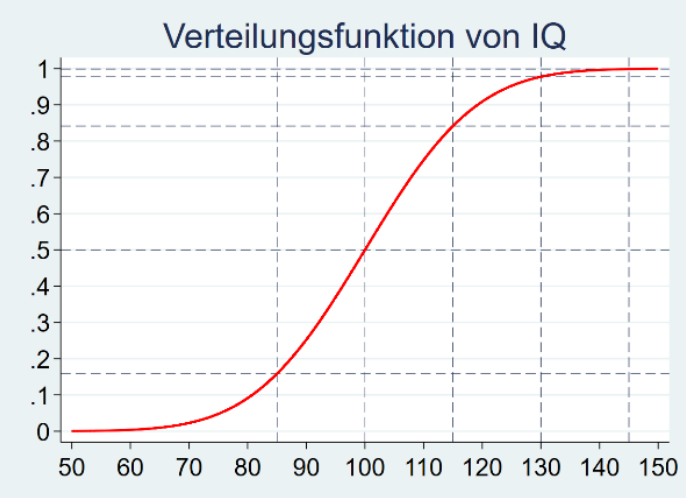




Verteilungen – Stetige Variablen

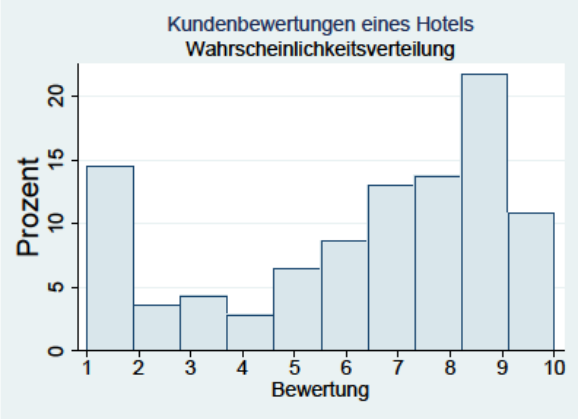
- Stetige Variablen:
  - **Dichtefunktion** – Wahrscheinlichkeiten sind definiert durch die Fläche unter der Dichtefunktion (das *Integral*) – (*probability density function,* ***pdf***)
  - Beachte: $f_Y(y)$ ist keine Wahrscheinlichkeit!
  - **Verteilungsfunktion** (**cdf**): siehe oben

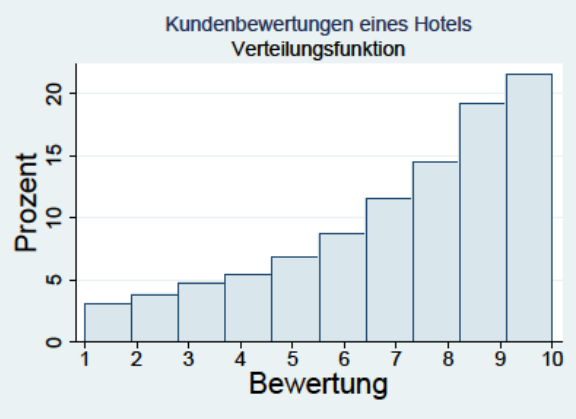




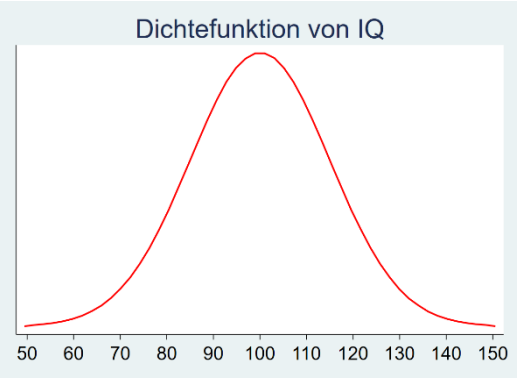
Wahrscheinlichkeitsverteilungen

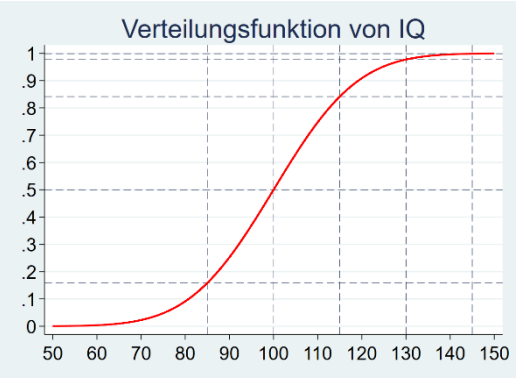
- Diskrete Variablen:
  - Wahrscheinlichkeitsverteilung
  - Verteilungsfunktion (cdf)





- Stetige Variablen:
  - Dichtefunktion (pdf)
  - Verteilungsfunktion (cdf)





- Nominal-, Ordinal-, Intervall-, Ratioskalen
- (Para-)metrische Variablen

Was kann man wo einsetzen?

Normalverteilung

- Eigenschaften der Normalverteilung:
  - Stetige Zufallsvariable
  - $Y \sim N(\mu, \sigma^2)$ mit Erwartungswert μ und Varianz σ^2
  - Schiefe = 0 (symmetrische Verteilung)
  - Kurtosis = 3
  - 95% der Wahrscheinlichkeitsmasse liegt zwischen $\mu - 1.96 \cdot \sigma$ und $\mu + 1.96 \cdot \sigma$
  - 90% liegt zwischen $\mu - 1.645 \cdot \sigma$ und $\mu + 1.645 \cdot \sigma$
  - 99% zwischen $\mu - 2.576 \cdot \sigma$ und $\mu + 2.576 \cdot \sigma$
  - 68.2% zwischen $\mu - 1 \cdot \sigma$ und $\mu + 1 \cdot \sigma$ (15.9% unterhalb von $\mu - 1 \cdot \sigma$)
  - 95.5% zwischen $\mu - 2 \cdot \sigma$ und $\mu + 2 \cdot \sigma$ (2.3% oberhalb von $\mu + 2 \cdot \sigma$)

Standardnormalverteilung

- Eigenschaften der Standardnormalverteilung:
  - $Z \sim N(0,1)$ mit Erwartungswert 0 und Varianz 1
  - Schiefe = 0
  - Kurtosis = 3
  - 95% der Wahrscheinlichkeitsmasse liegen zwischen -1.96 und $+1.96$ (usw.)

Werte auf unterschiedlichen Skalen vergleichen

- Beispiel Bewerbungen auf ein Praktikum in den USA
  - Zwei sehr ähnliche Bewerber
  - Englischtest soll entscheidend sein
  - Wen soll die Firma einstellen?

	Felix	Alexander
Englischtest	TOEFL: 21 Punkte	IELTS: 30 Punkte

- Welche Informationen fehlen uns?

TOEFL = Test of English as a Foreign Language
IELTS = International English Language Testing System

Z-Standardisierung

	Second Decimal Value of z									
z	0	1	2	3	4	5	6	7	8	9
–0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
–0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
–0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
–0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
–0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
–0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
–0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
–0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
–0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621

- Felix: TOEFL 21 Punkte; $\bar{Y} = 18$; $S_Y = 9$ → $Z_i = 0.33$ → 62,93. Perzentil
- Alexander: IELTS 30 Punkte; $\bar{Y} = 25$; $S_Y = 10$ → $Z_i = 0.50$ → 69,15. Perzentil

Werte auf unterschiedlichen Skalen vergleichen

- Jetzt drei sehr ähnliche Bewerbungen
  - Wen soll die Firma einstellen?

	Marie TOEFL	Lena IELTS	Julia TOIEC
Englischtest	22 Punkte	28 Punkte	670 Punkte
Skala	0 – 30	0 – 50	10 – 990
Mittelwert	18	25	650
Standard-abweichung	9	10	80

SurveyMonkey:
Z-Standardisierung

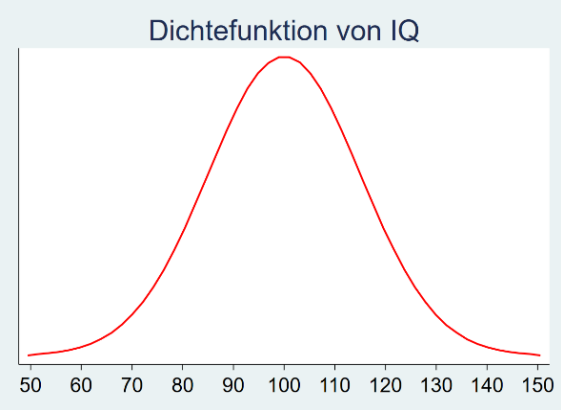
TOEFL = Test of English as a Foreign Language
IELTS = International English Language Testing System
TOIEC = Test of English for International Communication

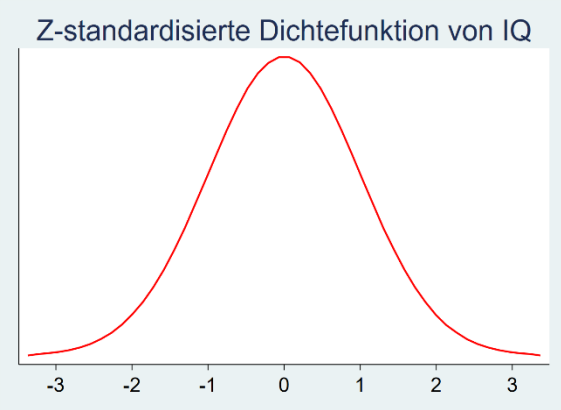
Z-Standardisierung

- Macht Werte auf unterschiedlichen Skalen vergleichbar

$$Z_i = \frac{Y_i - \bar{Y}}{S_Y}$$

- Mit Tabelle der Verteilungsfunktion der Standardnormalverteilung → Position in der Verteilung (Perzentil Scores) berechnen
- Man kann auch ganze Normalverteilungen standardisieren = **Zentrierung** der Daten → Bekommen dadurch Mittelwert 0, Standardabweichung 1





Was ist mit Schiefe & Kurtosis?

Gliederung: Analyse Quantitativer Daten

1. Deskriptive Statistik
    a. Schätzer und ihre Eigenschaften: Mittelwerte
    b. Streuungsmaße
    c. Verteilungen
    d. Z-Standardisierung
    e. Zusammenhänge zwischen Variablen

Zusammenhänge zwischen Variablen: 1. Kovarianz

- Kovarianz zweier Variablen, X und Y
- $S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n-1}$
- Kovarianz einer Variable mit sich selbst = Varianz

Person	Anzeigen gesehen	Produkte gekauft
1	5	10
2	3	7
3	7	10
4	4	9
5	6	14

$S_{XY} = 2.75$

(nachrechnen)

- Bitte beachten: Planing (2022) teilt durch n statt durch $n - 1$. Dies ergibt S_{XY} für eine Population statt für eine Stichprobe
- Da wir immer nur Stichproben haben, teilen wir immer durch $n - 1$

Zusammenhänge zwischen Variablen: 1. Kovarianz

- $S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n-1}$

- Interpretation: schwierig
  - $S_{XY} > 0$: Positiver Zusammenhang zwischen X und Y (X und Y sind gleichzeitig „groß“ und „klein“)
  - $S_{XY} < 0$: Positiver Zusammenhang
  - Wenn X und Y unabhängig sind, ist die Kovarianz gleich Null

Zusammenhänge zwischen Variablen: 2. Korrelation

- Pearson-Korrelationskoeffizient $r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$
- r_{XY} – unverzerrter & konsistenter Schätzer von ρ_{XY} („rho“) = Korrelation in der Grundgesamtheit
- Korrelationen liegen zwischen -1 und $+1$
- Unkorreliertheit: $r_{XY} = 0$
- $r_{XY} > 0$: Y höher wenn X höher
- $r_{XY} < 0$: Y niedriger wenn X höher
- Korrelationen in Absolutbetrag:
  - Ab .10: schwacher Zusammenhang
  - Ab .30: mittlerer Zusammenhang
  - Ab .50: starker Zusammenhang

Zusammenhänge zwischen Variablen: 2. Korrelation

- $r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$

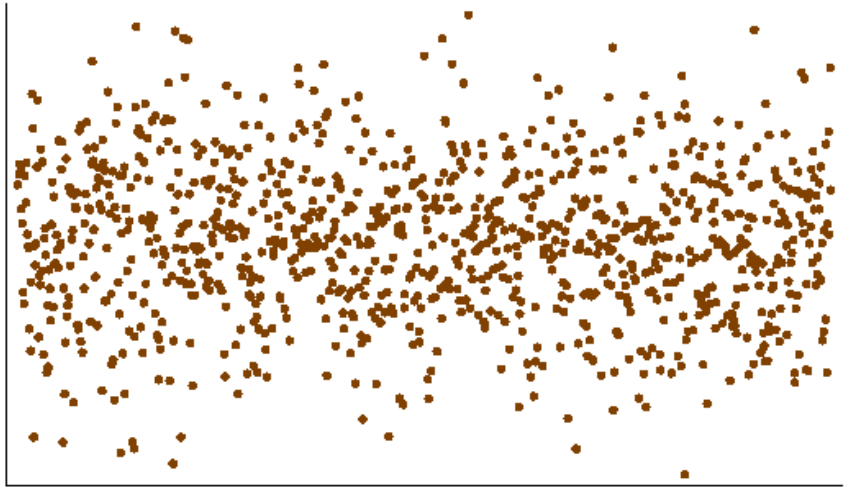
Person	Anzeigen gesehen	Produkte gekauft
1	5	10
2	3	7
3	7	10
4	4	9
5	6	14

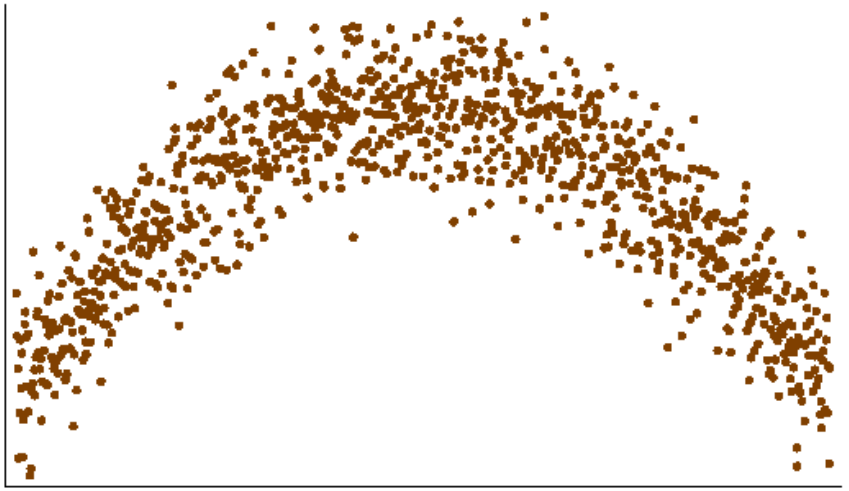
$s_{XY} = 2.75$

- $r_{XY} = 0.62$ (nachrechnen)

Zusammenhänge zweier Zufallsvariablen

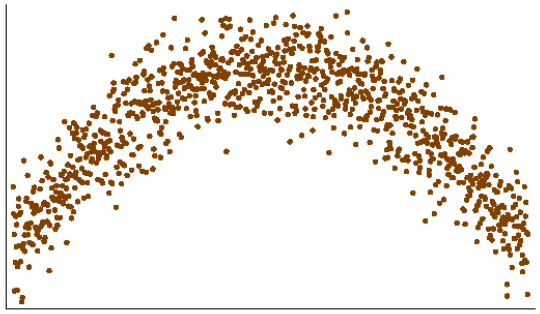
- **Beachte: Kovarianz und Korrelation messen nur lineare Zusammenhänge!**
- $r_{XY} = 0$ und $S_{XY} = 0$ und heißen: kein *linearer* Zusammenhang
- Schokolade Essen & Wohlbefinden: jeweils $r_{XY} = 0$



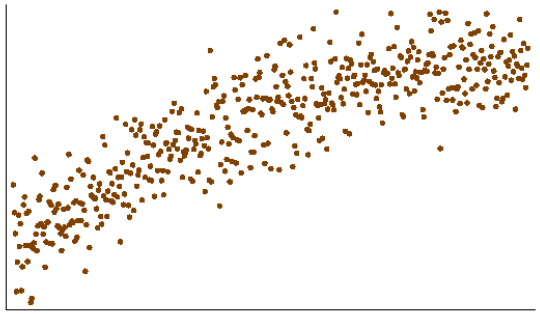


Zusammenhänge zweier Zufallsvariablen

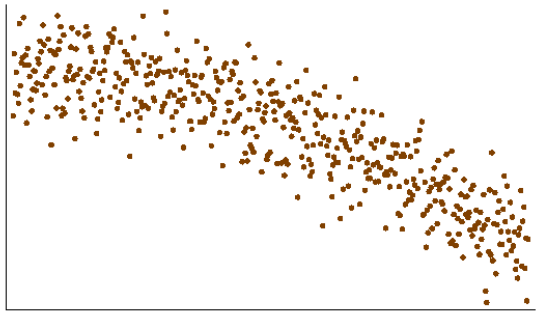
- $r_{XY} = 1 \Longrightarrow$ perfekter positiver, *linearer* Zusammenhang: Alle Punkte liegen auf einer Linie
- $r_{XY} = -1 \Longrightarrow$ perfekter negativer, *linearer* Zusammenhang : Alle Punkte liegen auf einer Linie
- Betrachten wir das rechten Bild der letzten Folie nochmals:



Gesamtbild:
$r_{XY} = 0.00$



Linke Hälfte:
$r_{XY} = 0.85$



Rechte Hälfte
$r_{XY} = 0.85$

- Genau genommen ist der Zusammenhang weder links noch rechts *linear*, sondern quadratisch

Korrelationen & Skalenniveaus

- Bei welchen Skalenniveaus macht der Pearson-Korrelationskoeffizient Sinn?

1. Nominalskala
2. Ordinalskala
3. Intervallskala
4. Ratioskala (Verhältnisskala)

?

Pearson- und Spearman-Korrelationskoeffizient

- Rangkorrelation: Spearman‘s Rho
- Interpretation: siehe Pearson-Korrelationskoeffizient
  - Korrelationen liegen zwischen -1 und $+1$
  - Unkorreliertheit: $r_{XY} = 0$
  - $r_{XY} > 0$: Y höher wenn X höher
  - $r_{XY} < 0$: Y niedriger wenn X höher
- Geeignet wenn eine oder beide Variablen nicht-parametrisch sind
- Berechnung: hier nicht besprochen