BA KW | Vorlesung

Einführung in die Statistik

Zufallsgrößen und Konfidenzintervall

Prof. Thomas Hanitzsch

Von Häufigkeiten zu Wahrscheinlichkeiten

- Inferenzstatistische (schließendende) Verfahren basieren auf der Annahme, dass es sich bei den empirisch Beobachtungen um eine zufällige Auswahl von Elementen aus der Grundgesamtheit handelt
- Die entscheidende Frage ist mithin, ob die anhand der Stichprobe erzielten Ergebnisse mit einer bestimmten, im Vorhinein festgelegten Wahrscheinlichkeit auf die Grundgesamtheit übertragen werden können

Wahrscheinlichkeit

- Maß für die Chance, dass bei einem Zufalls-experiment ein bestimmtes Ereignis eintritt
- Denomination: $p(X)$
- Wird angegeben als Zahl zwischen 0 und 1 oder in Prozent (z.B. $p = 0{,}5$ bzw. 50-prozentige Wahrscheinlichkeit)
- Klassische Definition von Pierre-Simon Laplace (1749-1829):

$$p(A) = \frac{Anzahl\ der\ günstigen\ Ereignisse}{Anzahl\ der\ möglichen\ Ereignisse}$$

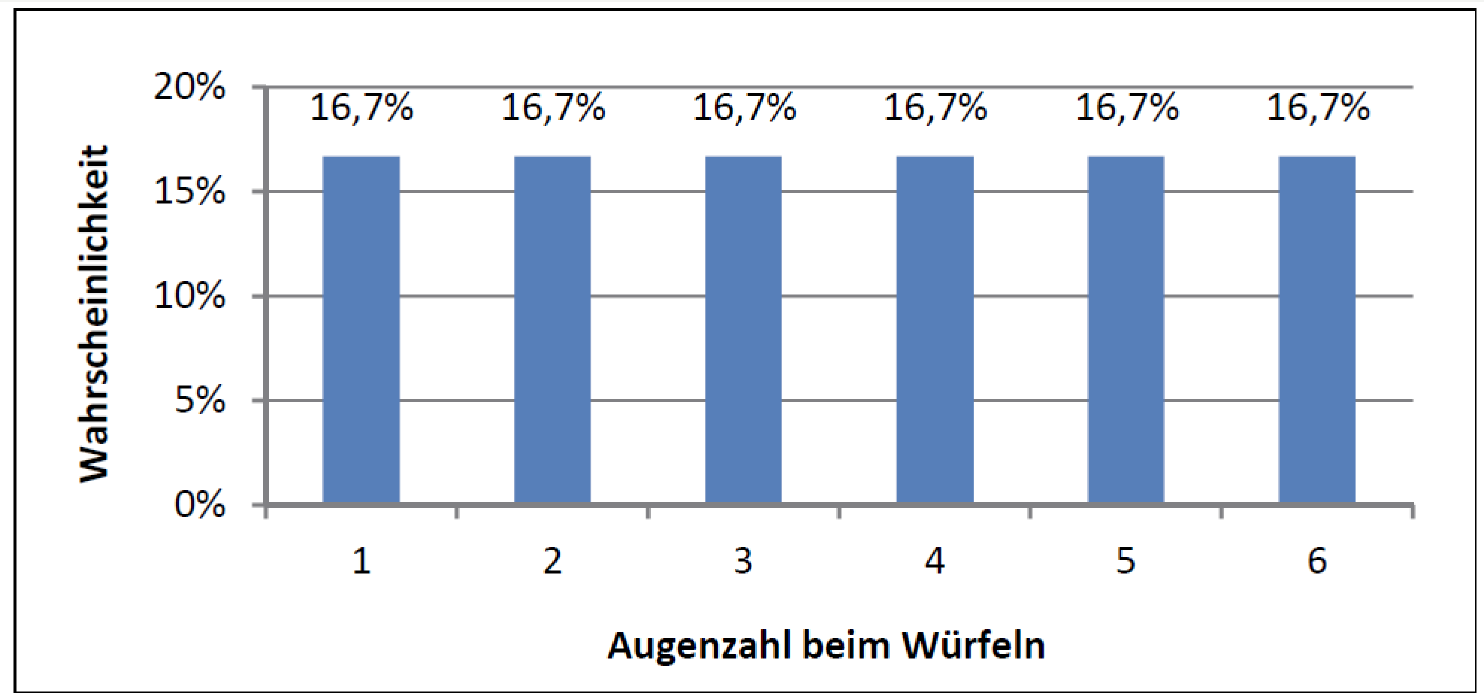
$$\mathrm{P}(5 \text{ oder } 6 \text{ würfeln}) = \frac{\text{⚄ ⚅}}{\text{⚀ ⚁ ⚂ ⚃ ⚄ ⚅}} = \frac{2}{6} = \frac{1}{3} = 0{,}333 = 33{,}3\%$$

Zufallsgrößen

- Variablen, deren Werte vom Zufall abhängig sind
  - z.B. Münz- oder Würfelwurf
  - aber auch empirische Merkmale, die anhand einer Zufallsstichprobe erhoben wurden
- Zufallsgrößen lassen sich formal als Funktion beschreiben, die den Ergebnissen eines Zufallsexperiments Werte zuordnet
- Verteilungsfunktion: $F(X) = P(X \leq x_i)$
  - Verteilungsfunktion $F(X)$ an der Stelle x_i ist damit die Wahrscheinlichkeit, dass die Zufallsgröße X einen Wert kleiner oder gleich x_i annimmt

Wahrscheinlichkeitsverteilungen

- informieren darüber, mit welcher Wahrscheinlichkeit die jeweils möglichen Ereignisse eines Zufallsexperiments auftreten können



Kennwerte und Parameter

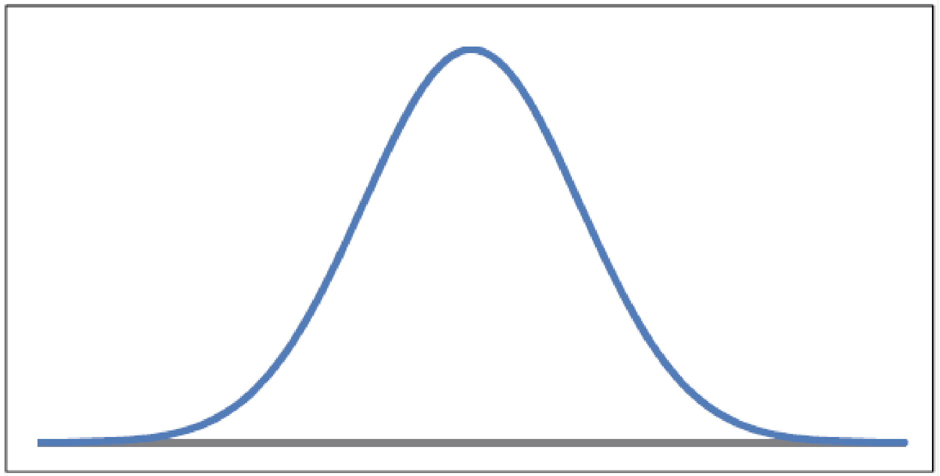
	Stichprobe	**Grundgesamtheit**
	Kennwerte	Parameter
Mittelwert	$\overline{x}$	μ
Standardabweichung	s	σ
Varianz	s^2	σ^2

- Das arithmetische Mittel („Mittelwert“) wird im Kontext von Zufallsvariablen auch als „Erwartungswert“ bezeichnet:

$$E(X) = \mu$$

Die Normalverteilung

- Ist die bekannteste stetige Wahrscheinlichkeitsverteilung
- Ursprünglich von Abraham de Moivre (1667-1754) und Pierre-Simon Laplace (1749-1827) als Annäherung an die Binomialverteilung entwickelt
- Am häufigsten wird sie mit Carl Friedrich Gauß (1777-1855) in Verbindung gebracht, der die Normalverteilung unter anderem in der Astronomie einsetzte
  - „Gauß-Verteilung“
  - „Gaußsche Glockenkurve“

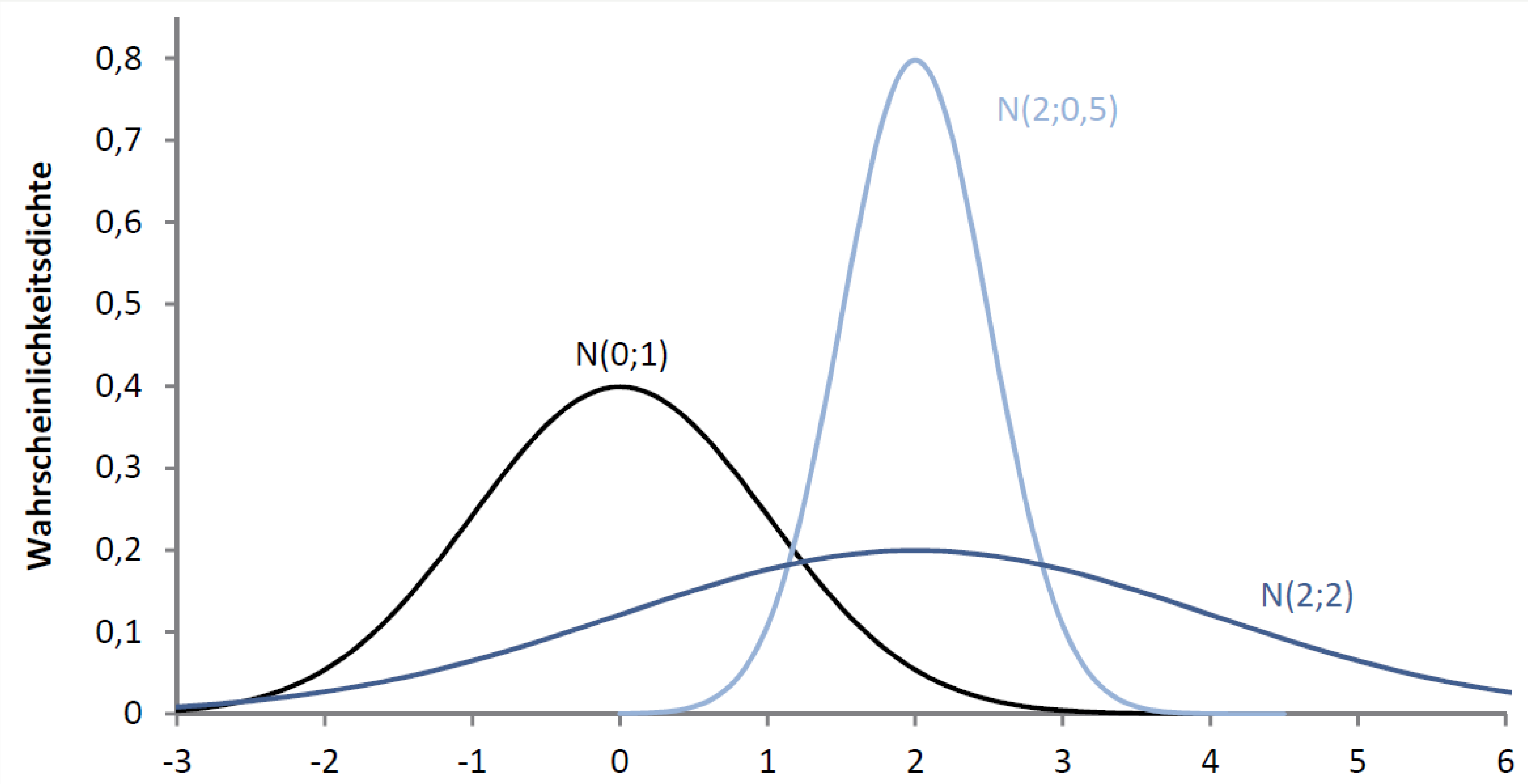


Die Normalverteilung

Eigenschaften:

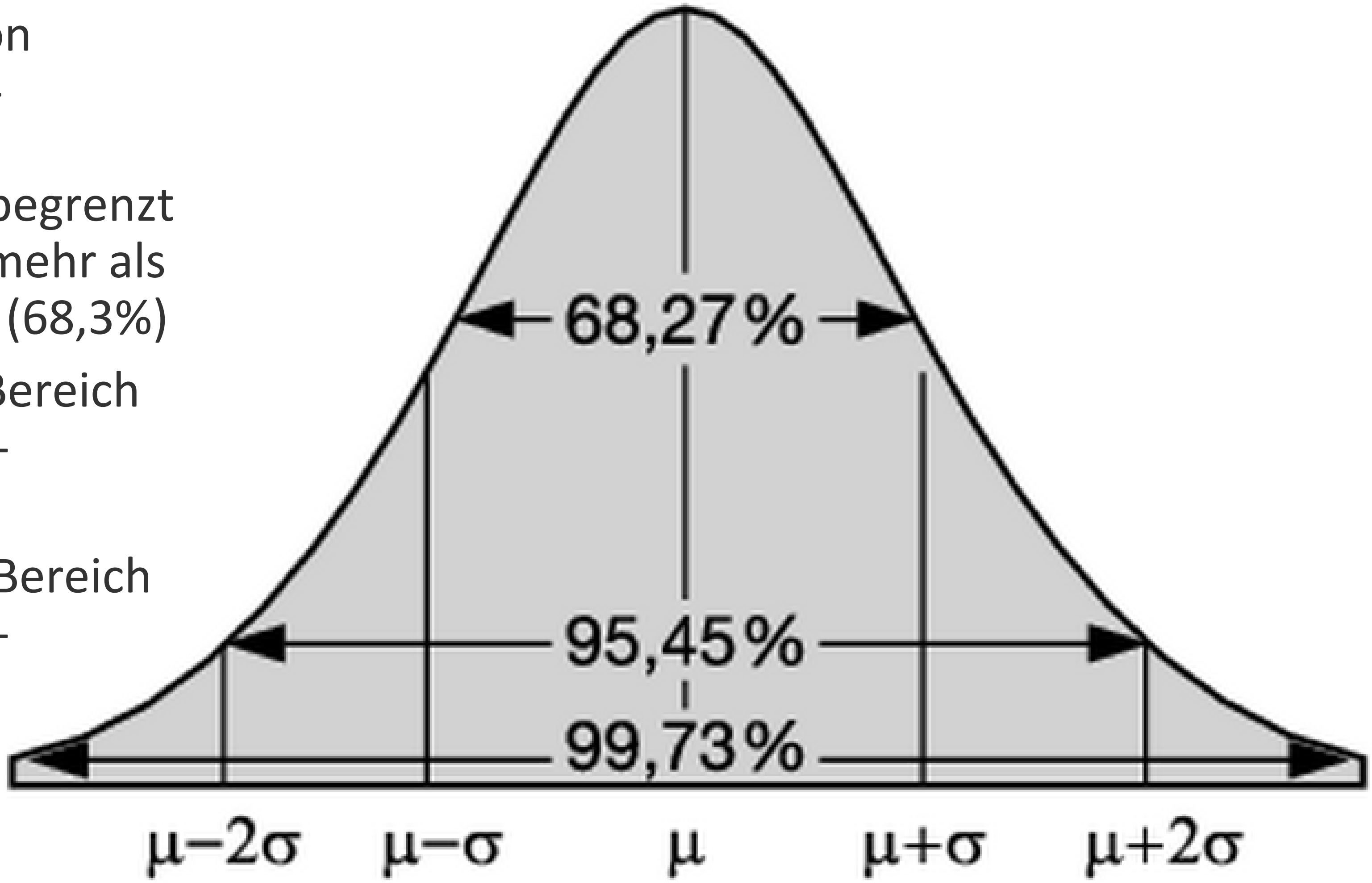
- Normalverteilungen sind symmetrisch und unimodal
- Erwartungswert, Median und Modus sind identisch, liegen genau in der Mitte und teilen die Verteilung exakt in zwei Hälften
- Die meisten Werte streuen nah um den Erwartungswert
- Normalverteilungen nähern sich der X-Achse an, ohne sie zu berühren
- Normalverteilungen sind durch zwei Größen eindeutig bestimmt, nämlich dem Erwartungswert und der Standardabweichung
  → Schreibweise $N(\mu; \sigma)$
- Definitionsbereich (X-Achse) reicht von $-\infty$ bis $+\infty$

Die Normalverteilung: Beispiele



Die Normalverteilung: Dichtefunktion

- Die Fläche, die von ± einer Standardabweichung vom Erwartungswert begrenzt wird, beinhaltet mehr als ca. 2/3 aller Fälle (68,3%)
- 95,4% liegen im Bereich von ± 2 Standardabweichungen
- 99,7 % liegen im Bereich von ± 3 Standardabweichungen

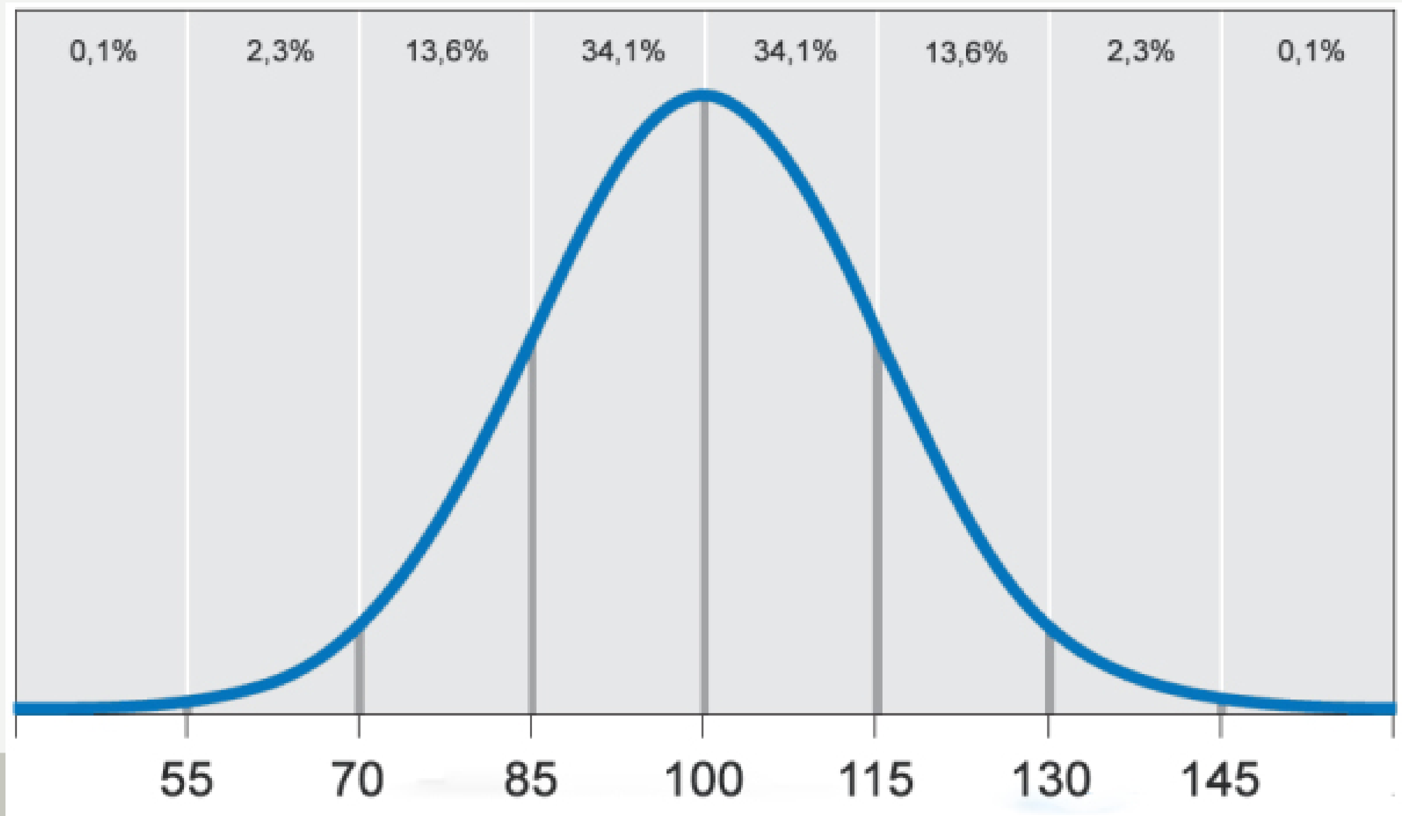


Die Normalverteilung: Dichtefunktion

Beispiel: Intelligenzquotient

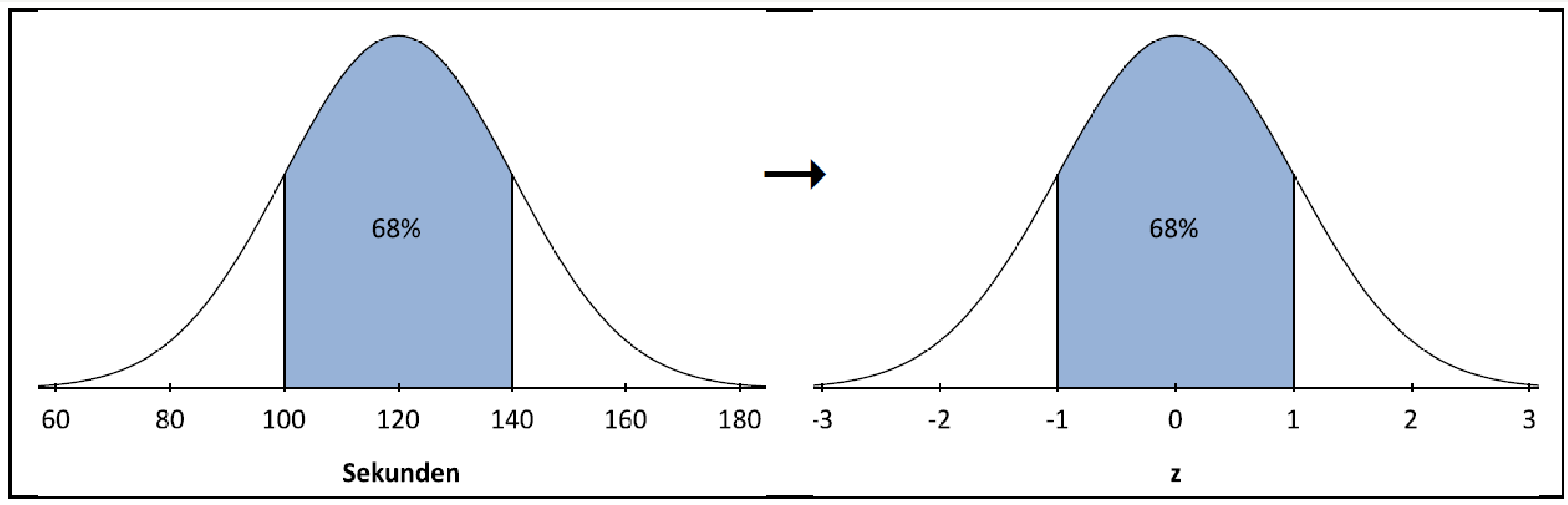
- Normalverteilt: $N(100; 15)$

68% haben einen IQ zwischen 85 und 115 (100±15)
95% haben einen IQ zwischen 70 und 130 (100 ±2·15)



Die Standardnormalverteilung

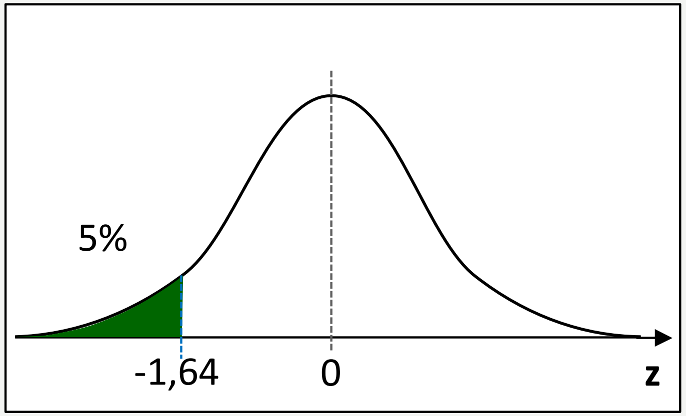
- Sonderfall der Normalverteilung: $N(0; 1)$
- Jede Normalverteilung lässt sich mithilfe der z-Transformation in die Standardnormalverteilung transformieren; dabei werden die Verteilungswerte in z-Werte standardisiert

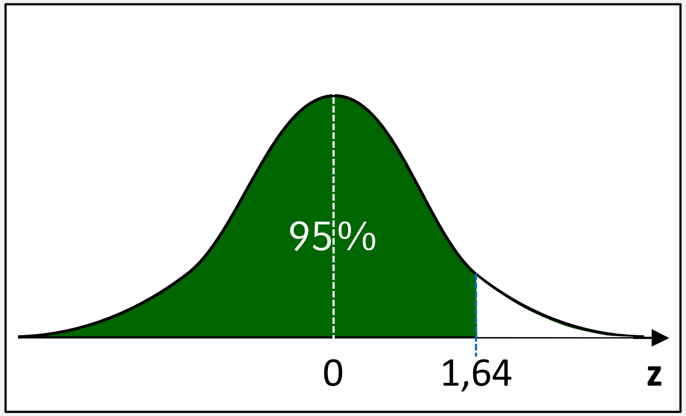


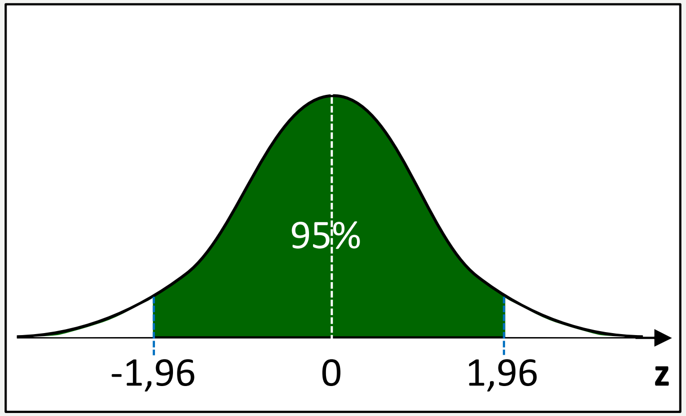
Dichtefunktion

Standardnormalverteilung:

- Wie bei allen stetigen Verteilungen wird die Fläche unterhalb der Kurve auf 100% gesetzt
  → somit entspricht jedem z-Wert (bzw. einer Differenz aus z-Werten) ein bestimmter Flächenanteil







Die Normalverteilung: Bedeutung

- Viele empirische Merkmale folgen einer Normalverteilung: z.B. altersspezifischer IQ, Körpergewicht und -größe, tägliche Rendite von Aktien der Deutschen Bank, jährliche Niederschläge (in mm) am MUC
- Bei genügend großen Stichproben folgt die Verteilung von Mittelwerten (bei multiplen Stichprobenziehungen) einer Normalverteilung
  → Zentraler Grenzwertsatz
  - daher wird die Normalverteilung verwendet, um von Stichproben auf die Verteilung eines Merkmals in der Grundgesamtheit zu schließen
- Viele andere Wahrscheinlichkeitsverteilungen können durch die Normalverteilung angenähert werden (z.B. Binomial- oder t-Verteilung)

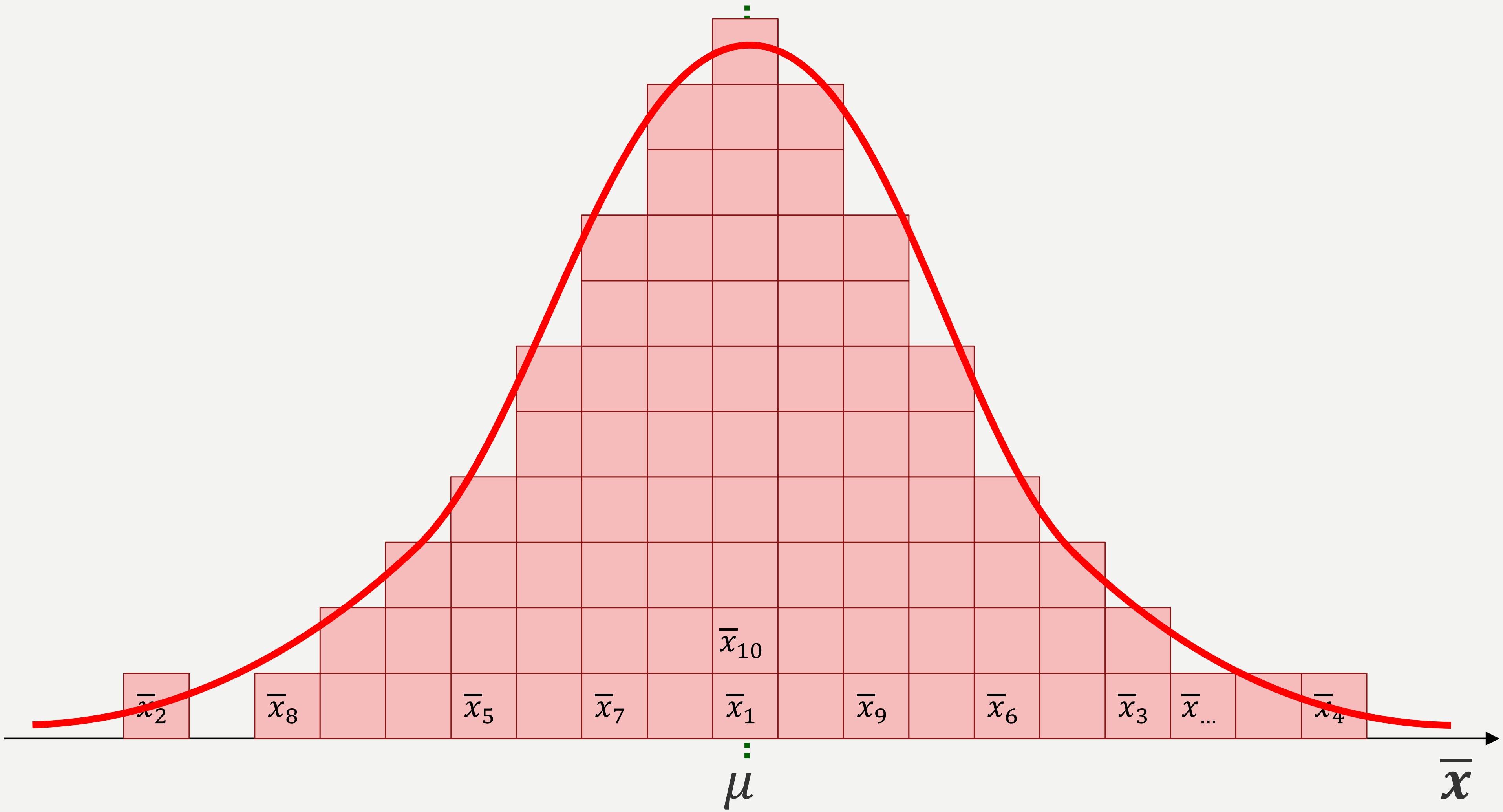
Inferenzstatistik und Stichproben

- Das Ziel der Inferenzstatistik ist es, von den bekannten Kennwerten einer Stichprobe (z.B. Mittelwert, Varianz) auf die unbekannten Parameter einer Grundgesamtheit zu schließen
- Vorteile von Stichproben:
  - Sie sind ökonomischer als Vollerhebungen
  - Sie liefern schneller Ergebnisse
  - Sie führen zu keiner kompletten Kontamination des Merkmalsträgers

Gesetz der großen Zahl (Theorem von Bernoulli)

- Die relative Häufigkeit des Auftretens von einer bestimmten Merkmalsausprägung konvergiert mit wachsendem N gegen die Wahrscheinlichkeit ihres Auftretens
- Beispiel:
  Je häufiger ein Würfel geworfen wird, desto näher kommt die relative Häufigkeit für das Auftreten einer „6" an die Wahrscheinlichkeit für das Auftreten der „6" heran
  $(p = \frac{1}{6})$

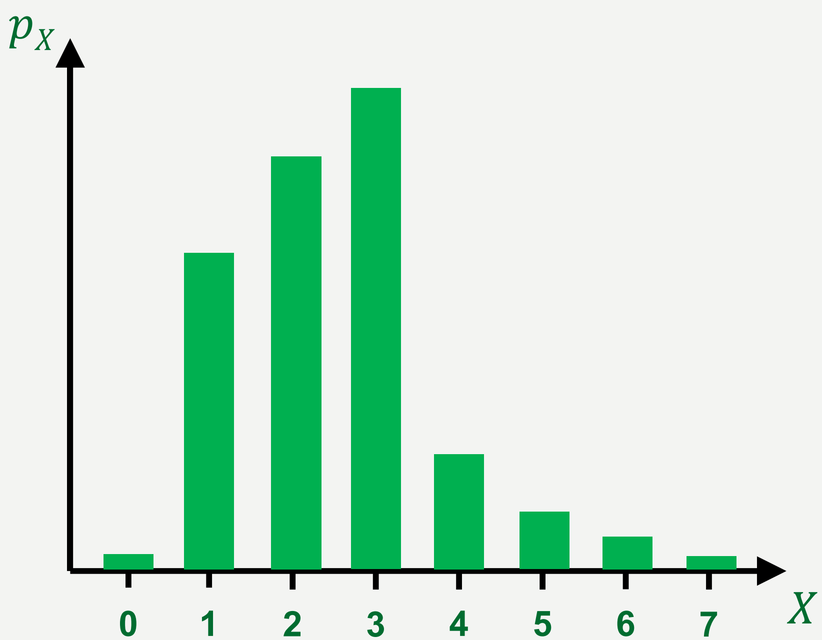
Mittelwertschätzung und Zufall



Zentraler Grenzwertsatz

- Die Verteilung der Mittelwerte aus Stichproben vom Umfang N, die sämtlich derselben Grundgesamtheit entnommen wurden, geht mit wachsendem Stichprobenumfang in eine Normalverteilung über
  - Das gilt bei ausreichend großem Stichprobenumfang (etwa $N \geq 30$) unabhängig von der Verteilung der Ausgangsvariable X in der Grundgesamtheit
  - Bei kleineren Stichprobenumfängen ($N < 30$) gilt die Annahme nur, wenn die Ausgangsvariable in der Grundgesamtheit normalverteilt ist

Die Verteilung von Stichprobenkennwerten



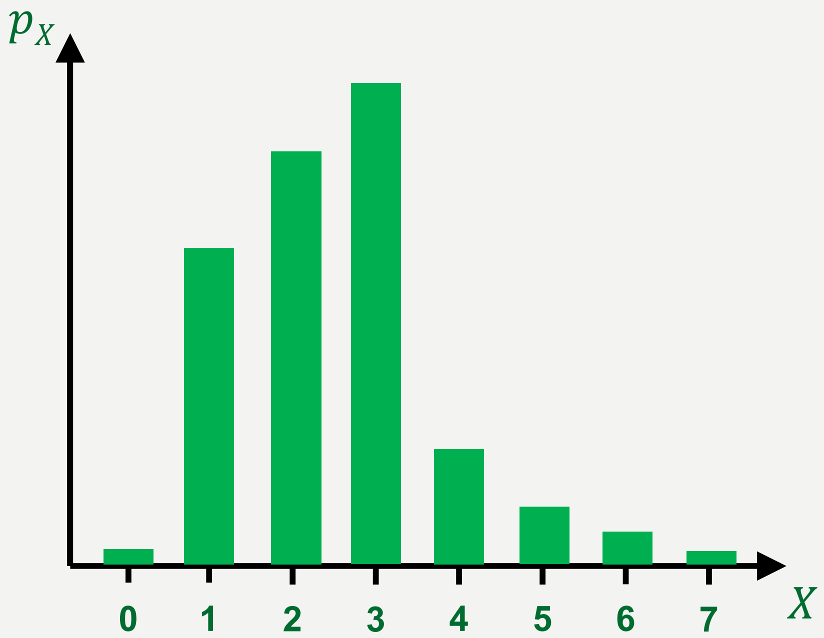
$\mu = 2{,}57;\ \sigma = 1{,}359$

Beispiel:
Gezogene Stichproben (N = 10)

Stichprobe		$\bar{x}$
Stichprobe 1:	*2, 4, 2, 1, 2, 3, 3, 6, 0, 1*	***2,40***
Stichprobe 2:	*1, 2, 2, 1, 7, 2, 4, 1, 1, 1*	***2,20***
Stichprobe 3:	*1, 2, 3, 4, 2, 2, 1, 3, 3, 3*	***2,40***
Stichprobe 4:	*3, 4, 3, 2, 2, 1, 3, 6, 2, 1*	***2,70***
Stichprobe 5:	*2, 3, 2, 1, 2, 3, 4, 6, 5, 2*	***3,00***
Stichprobe 6:	*2, 4, 2, 1, 7, 3, 3, 1, 2, 3*	***2,80***
Stichprobe 7:	*3, 4, 3, 1, 2, 3, 3, 0, 2, 3*	***2,40***
Stichprobe 8:	*2, 4, 2, 1, 2, 3, 3, 3, 5, 1*	***2,60***

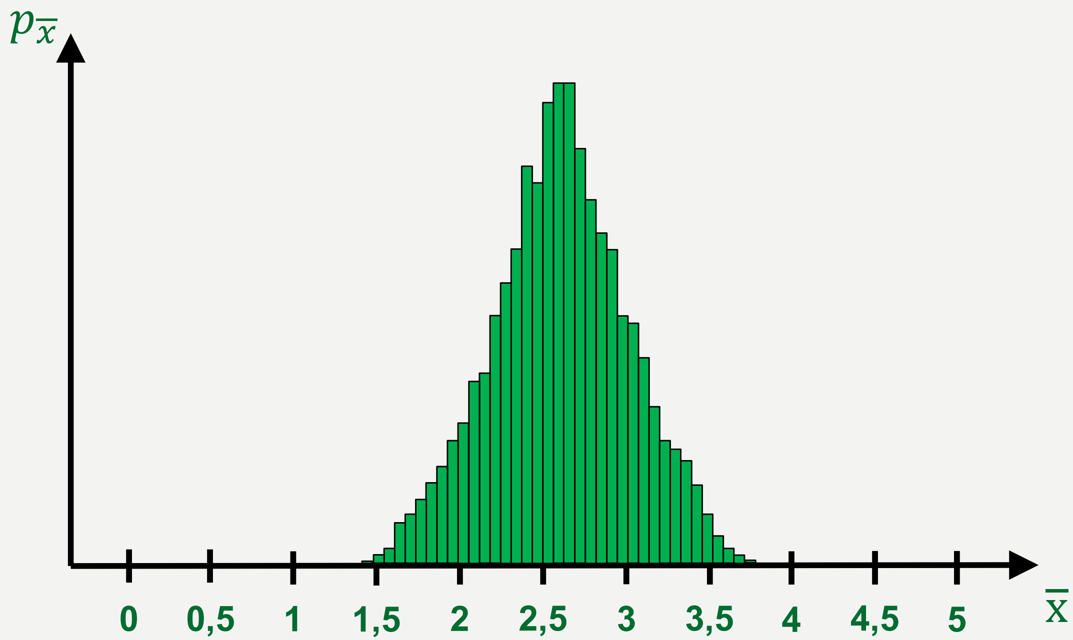
$$\bar{\bar{x}} = \frac{1}{N_{Stichproben}} \sum \bar{x} = \frac{1}{8} \sum \bar{x} = 2{,}56$$

Die Verteilung von Stichprobenkennwerten



Anzahl genutzter Medien

$\mu = 2{,}57; \ \sigma = 1{,}359$



Mittlere Anzahl genutzter Medien bei Stichprobengröße N = 30 (5000 Stichprobenziehungen)

$\mu \approx \bar{\bar{x}} = 2{,}57$

Punktschätzung

- Die relevanten Parameter der Grundgesamtheit sind in der Regel nicht bekannt und müssen daher geschätzt werden:
  - Erwartungswert: $\hat{\mu}_X = \bar{x}$
  - Varianz: $\hat{\sigma}_X^2 = s_X^2$
- Bei der Schätzung muss ein Fehler in Kauf genommen werden, der Standardfehler:

$$SE = \frac{\sigma}{\sqrt{N}}$$ bzw. falls σ nicht bekannt: $$SE = \frac{s}{\sqrt{N}}$$

- Problem: die Genauigkeit der Punktschätzung ist unbekannt

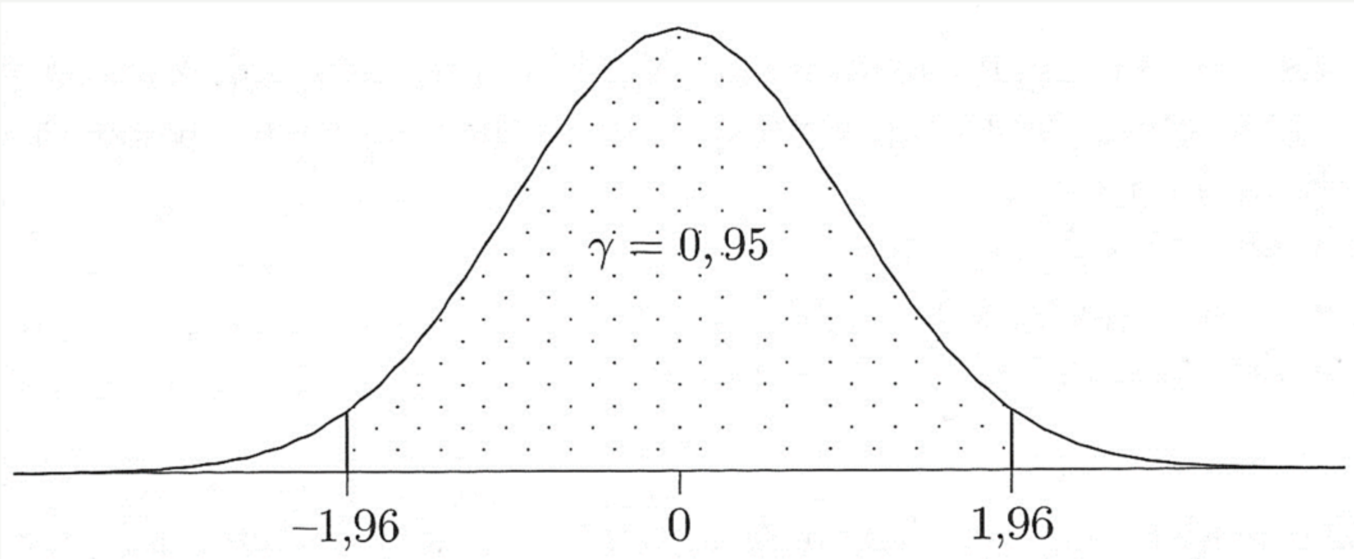
Intervallschätzung

- Schätzung des Bereichs, der bei unendlicher Wiederholung eines Zufallsexperiments mit einer gewissen Wahrscheinlichkeit (→ Konfidenzniveau) die wahre Lage des Parameters einschließt → Konfidenzintervall
- Am häufigsten wird das 95%-Konfidenzintervall berechnet, oft auch das 99%-Intervall
  → Festsetzung des Konfidenzniveaus γ
  (üblicherweise $\gamma = 0{,}95$ oder $\gamma = 0{,}99$)

Konfidenzintervall für μ (σ bekannt)

- Berechnung der Konfidenzintervalls $[K_u; K_o]$:

$$\left[\overline{x} - z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{N}} \; ; \; \overline{x} + z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{N}}\right]$$



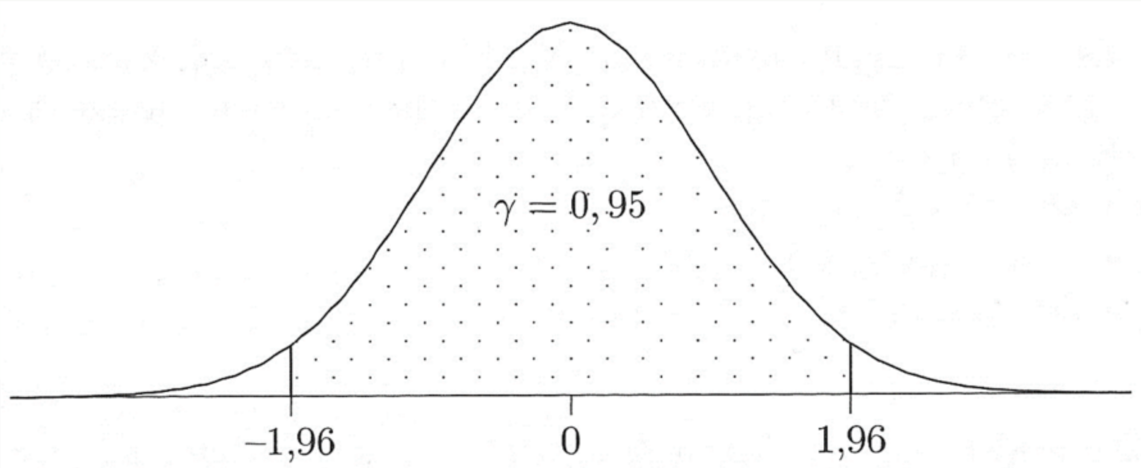
P(x ≤ z)	z
0,5%	–2,58
1,0%	–2,33
2,5%	–1,96
5,0%	–1,64
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

Konfidenzintervall für μ (σ unbekannt; $N \geq 30$)

Approximatives Konfidenzintervall für μ

- Bei genügend großem Stichprobenumfang kann der Zentrale Grenzwertsatz auf den Mittelwert angewendet werden
- Berechnung der Konfidenzintervalls $[K_u; K_o]$:

$$\left[\bar{x} - z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\ ;\ \bar{x} + z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}}\right]$$



$P(x \leq z)$	z
0,5%	−2,58
1,0%	−2,33
2,5%	−1,96
5,0%	−1,64
95,0%	1,64
97,5%	1,96
99,0%	2,33
99,5%	2,58

Approximatives Konfidenzintervall: Beispiel

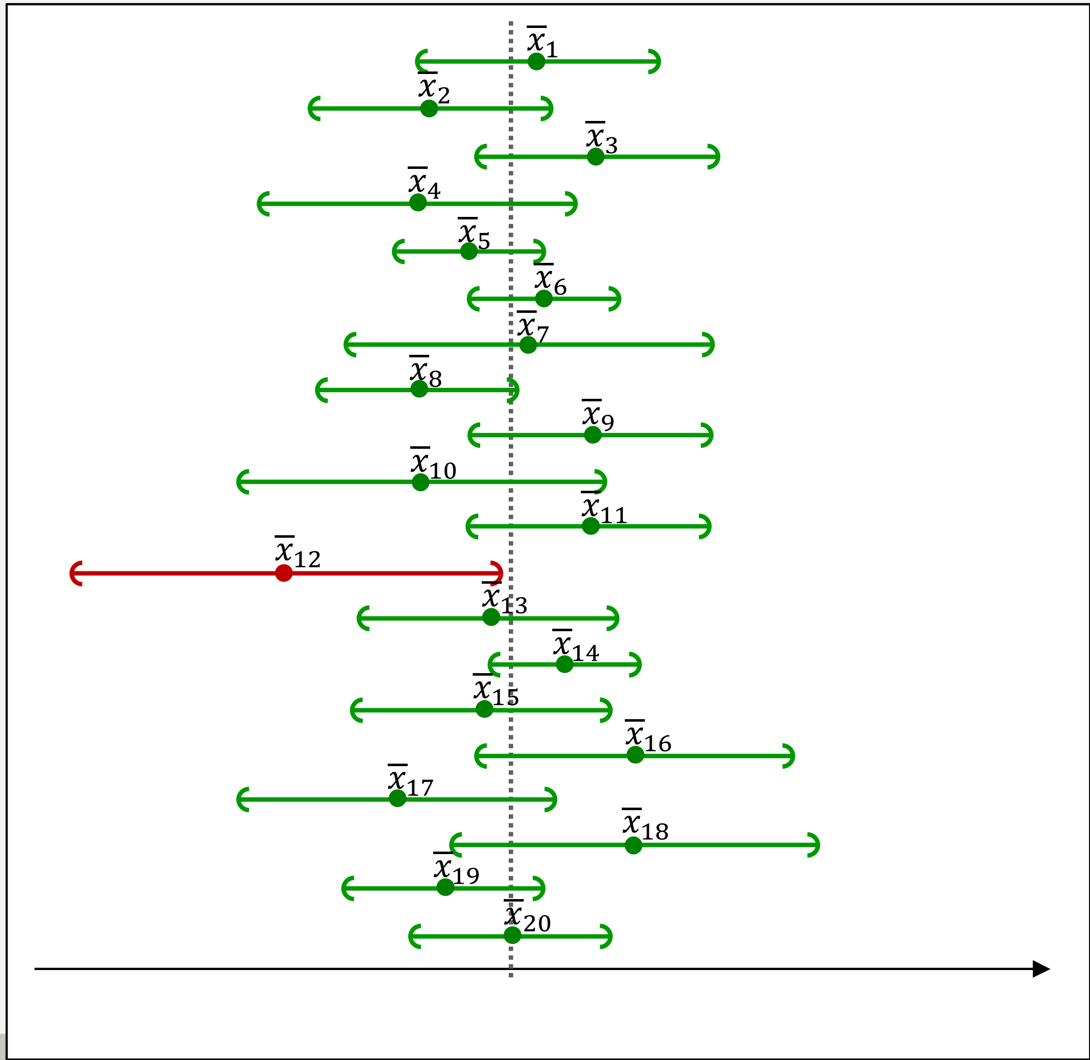
- Problemstellung:
  - Schuldirektor Paul hat in seiner Schule 36 Kinder ($N = 36$) zur Dauer der täglichen Handynutzung befragt und einen Durchschnitt von 60 Minuten ermittelt
    → Punktschätzung: $\bar{x} = 60$; $s = 12$
  - Er möchte nun wissen, in welchem Bereich die durchschnittliche Dauer der Handynutzung aller Kinder in seiner Schule (Grundgesamtheit) mit einer hinreichend großen Wahrscheinlichkeit liegt (Konfidenzintervall: 95%)
    → $\gamma = 0{,}95$; $z = 1{,}96$

$$K_u = \bar{x} - z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}} = 60 - 1{,}96 \cdot \frac{12}{\sqrt{36}} = 60 - 3{,}92 = \mathbf{56,08}$$

$$K_o = \bar{x} + z_{\frac{1+\gamma}{2}} \cdot \frac{s}{\sqrt{N}} = 60 + 1{,}96 \cdot \frac{12}{\sqrt{36}} = 60 + 3{,}92 = \mathbf{63,92}$$

Konfidenzintervalle und Zufall

- Hypothetische Verteilung der Konfidenzintervalle bei einem Konfidenzniveau $\gamma = 0{,}95$



Konfidenzintervall für Anteil $(N \geq 30)$

Approximatives Konfidenzintervall

- Punktschätzung für den Anteil durch die relative Häufigkeit p und Schätzung der Varianz $p \cdot (1-p)$

- Berechnung der Konfidenzintervalls $[K_u; K_o]$:

$$\left[p - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}} \,; p + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}} \right]$$

Konfidenzintervall für Anteil: Beispiel

- Problemstellung:
  - Die Süddeutsche Zeitung möchte in Erfahrung bringen, in wie vielen Münchner Haushalten die SZ abonniert wird. Eine Befragung von 100 Haushalten ($N = 100$) ermittelt einen Anteil von 50 Prozent → Punktschätzung: $p = 0{,}5$
  - Die SZ möchte nun wissen, in welchem Bereich dieser Anteil für alle Münchner Haushalte mit einer hinreichend großen Wahrscheinlichkeit liegt (Konfidenzintervall: 99%) → $\gamma = 0{,}99$; $z = 2{,}58$

$$K_u = p - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}} = 0{,}5 - 2{,}58 \cdot \sqrt{\frac{0{,}5 \cdot (1-0{,}5)}{100}} = 0{,}5 - 0{,}13 = \mathbf{0{,}37}$$

$$K_o = p + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}} = 0{,}5 + 2{,}58 \cdot \sqrt{\frac{0{,}5 \cdot (1-0{,}5)}{100}} = 0{,}5 + 0{,}13 = \mathbf{0{,}63}$$

Approximatives Konfidenzintervall für μ: Aufgabe 1

- Problemstellung:
  - Eine repräsentative Studie an 200 zufällig ausgewählten Personen hat ergeben, dass Paare in Deutschland durchschnittlich 8 Minuten pro Tag miteinander kommunizieren ($s = 2{,}0$).
  - In welchem Bereich können wir anhand dieser Schätzung den tatsächlichen Durchschnittswert für die Paarkommunikation in der deutschen Bevölkerung erwarten (Konfidenzintervall: 99%)?

Approximatives Konfidenzintervall für p: Aufgabe 2

- Problemstellung:
  - Bei der Bundestagswahl 2013 gab es deutschlandweit eine Wahlbeteiligung von 72%. Bei einer Befragung von 120 Zuschauern der Talkshow „Maybrit Illner“ gaben 96 Zuschauer an, 2013 gewählt zu haben.
  - Ist die Wahlbeteiligung unter „Maybrit Illner“-Sehern systematisch höher als in der Gesamtbevölkerung? Zugrunde gelegt wird ein Konfidenzintervall von 95%.

  → Nachdenken ist gefragt!