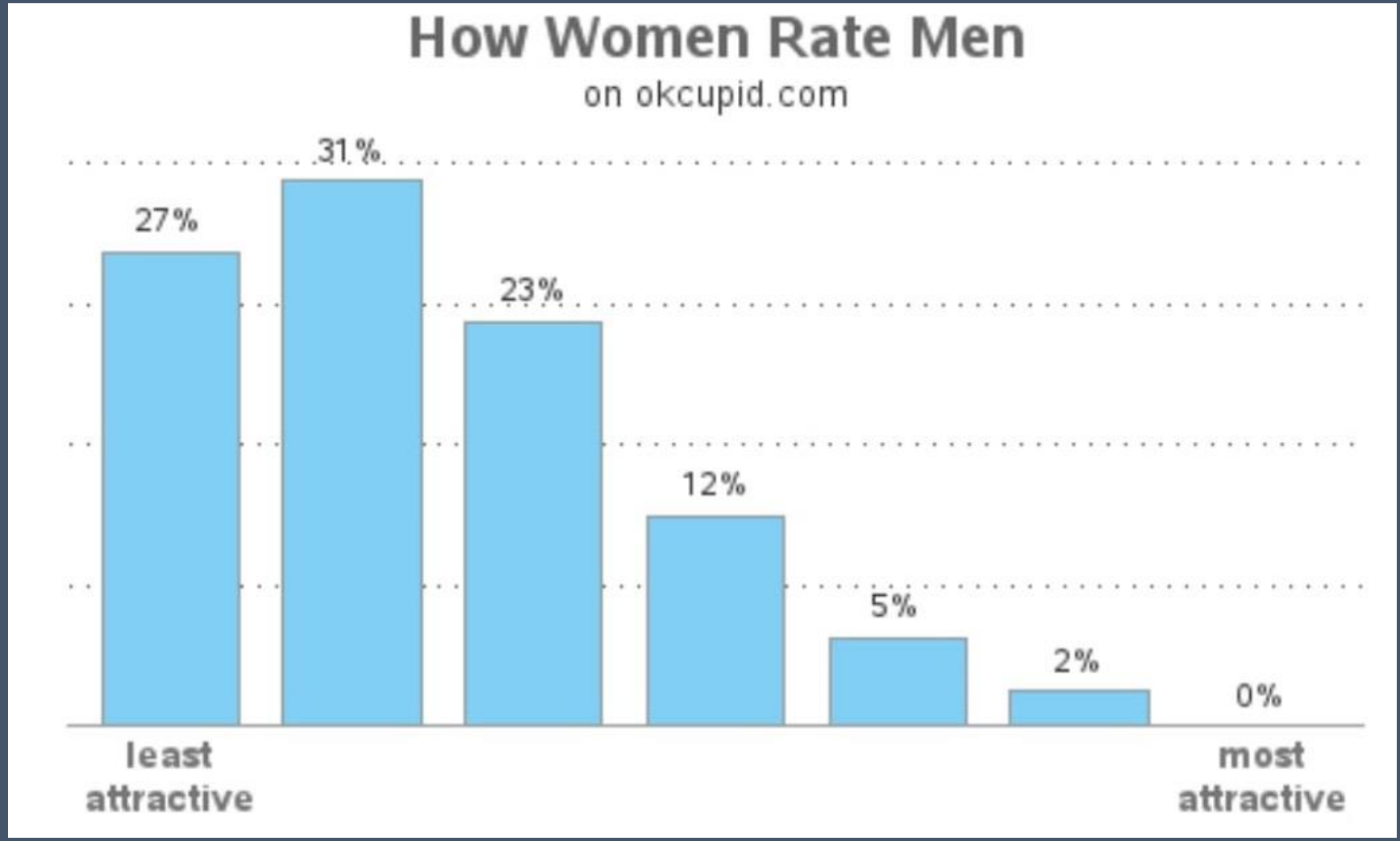
Tutorium Datenauswertung

Schätzen und Testen

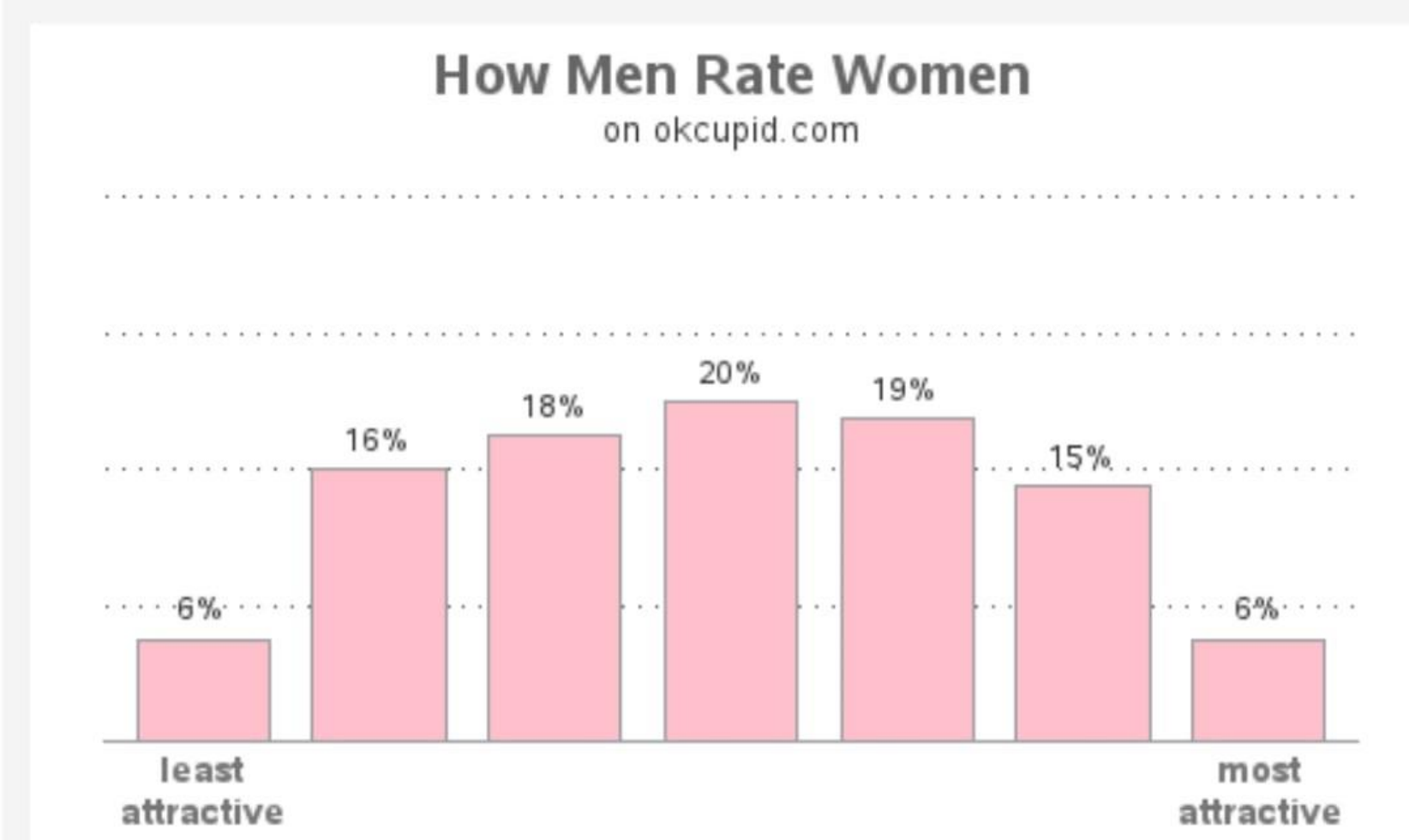
Agenda

- Verteilungen interpretieren
- Die Normalverteilung
- Verteilungen:
  - Stichprobe, Grundgesamtheit, Stichprobenverteilung
  - Standardnormalverteilung
- Schätzen
- Testen

Verteilungen interpretieren I



Verteilungen interpretieren II



Die Normalverteilung I

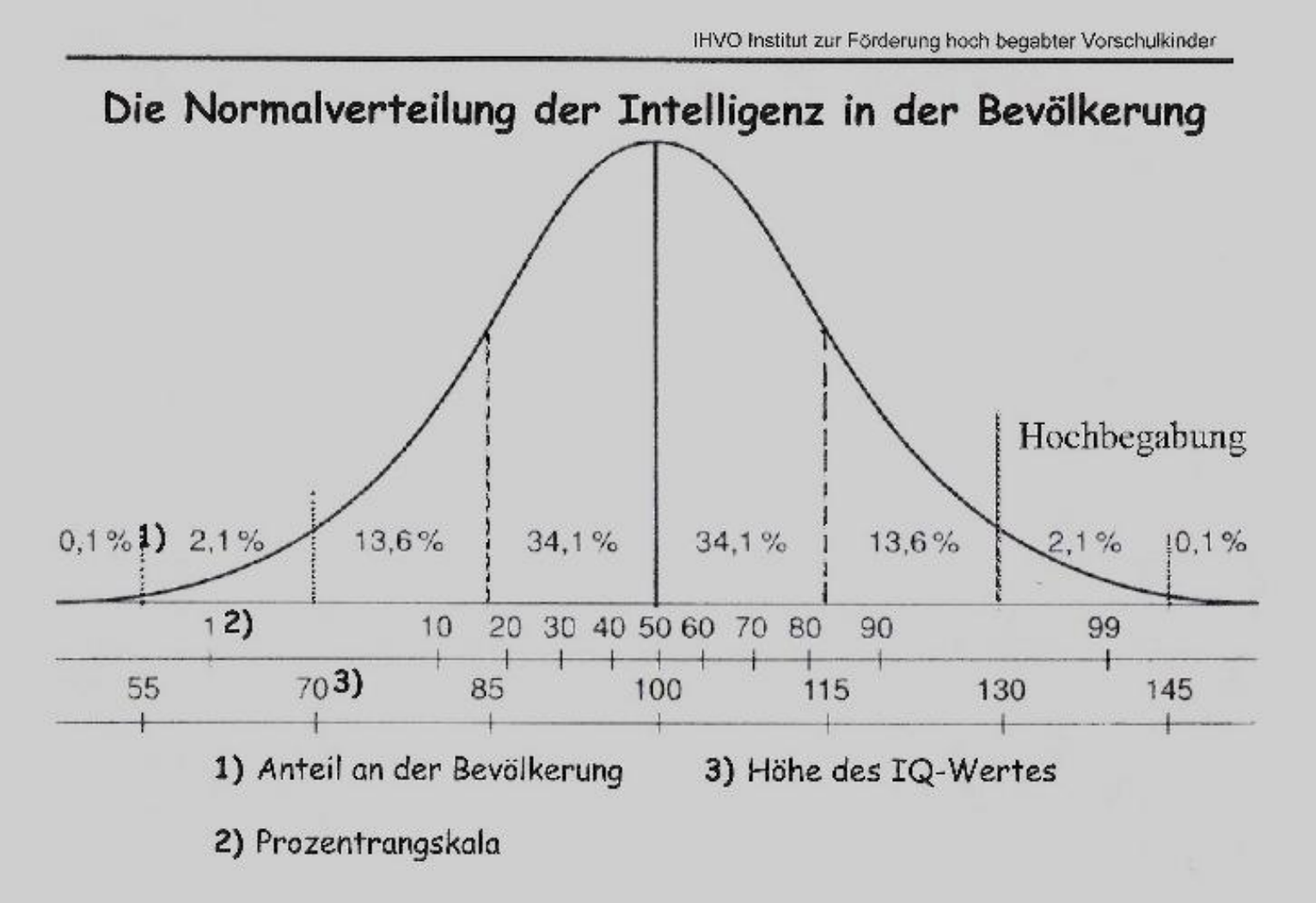
Die Normalverteilung

- Ist die Verteilung, die metrische Variablen oft annehmen
- Wird definiert durch Mittelwert und Varianz
- Hat folgende Eigenschaften:
  - Unimodalität
  - Symmetrie
  - Glockenform

Die Normalverteilung Interpretieren

- 68% der Werte befinden sich zwischen +/- einer Standardabweichung um den Mittelwert
- 95% der Werte befinden sich zwischen +/- 1,96 (≈ 2) Standardabweichungen um den Mittelwert
- 99% der Werte befinden sich zwischen +/- 2.58 Standardabweichungen um den Mittelwert

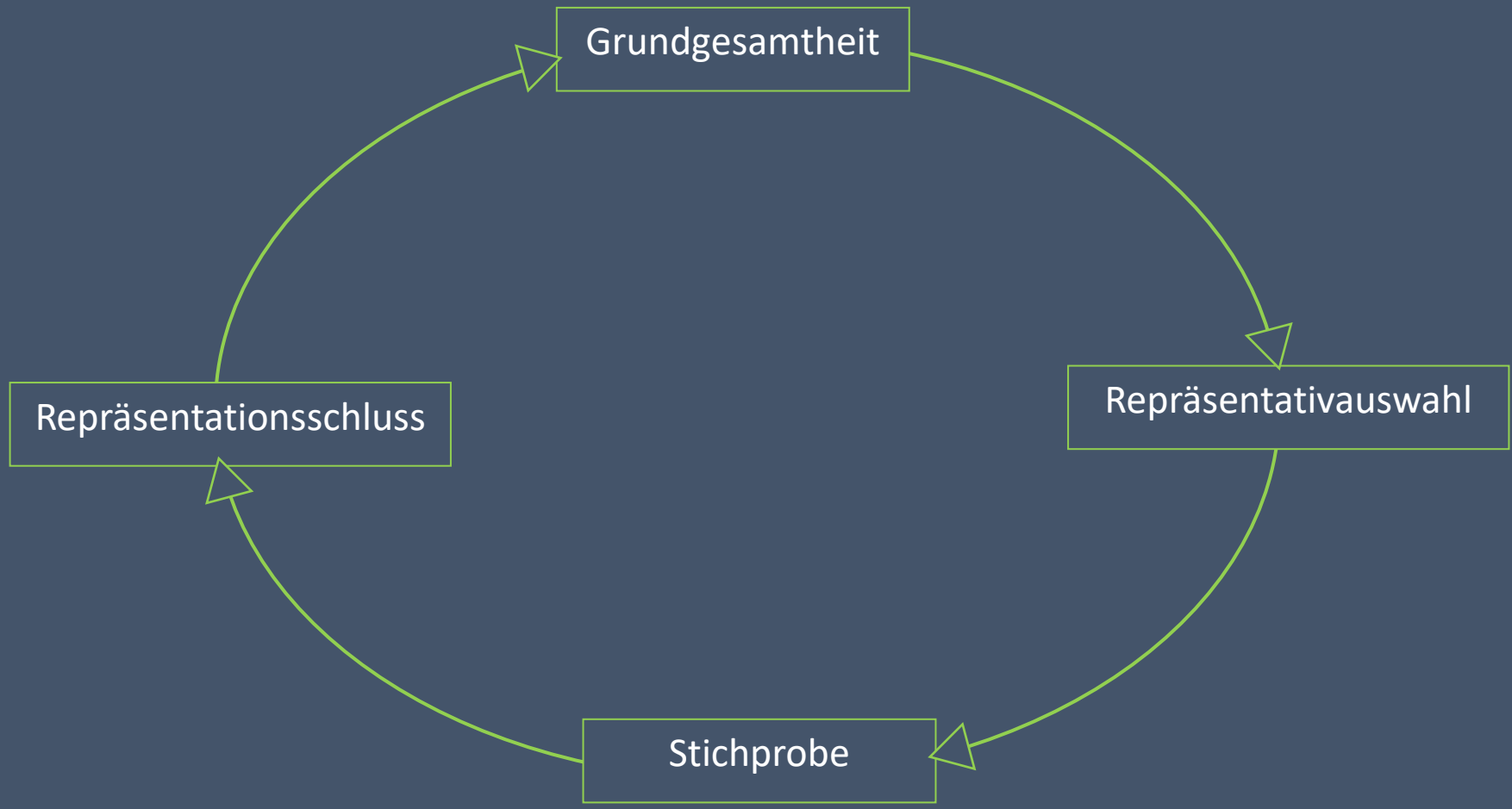
Normalverteilung – Beispiel



Wichtige Verteilungen

- Grundgesamtheit („population“)
  - Ist die Gruppe von Menschen, auf die sich die Forschungshypothese bezieht
  - Ist in der Regel unbekannt
- Stichprobe („sample distribution“)
  - Ist der Teil der Grundgesamtheit, der tatsächlich erhoben wird
  - wird benutzt, um Werte in der Grundgesamtheit zu schätzen
- Stichprobenverteilung („sampling distribution“)
  - Ist die Verteilung der Kennwerte (bspw. Mittelwerte) einer großen Anzahl von Stichproben aus der Grundgesamtheit
  - Ist ab einer Stichprobenanzahl von $n \geq 30$ normalverteilt

Grundgesamtheit und Stichprobe



Der zentrale Grenzwertsatz

Der zentrale Grenzwertsatz („central limit theorem“)

- Setzt voraus, dass die Stichprobe durch Zufallsauswahl entstanden ist (!) und mindestens 30 Fälle enthält
- Besagt, dass die Stichprobenverteilung normalverteilt ist, *unabhängig* von der Verteilung der Grundgesamtheit oder einer einzelnen Stichprobe
- Ist mathematisch bewiesen und Grundlage zum Schätzen von Parametern und Testen von Hypothesen
- Link zum Ausprobieren

Standardfehler

Der Standardfehler (standard error)

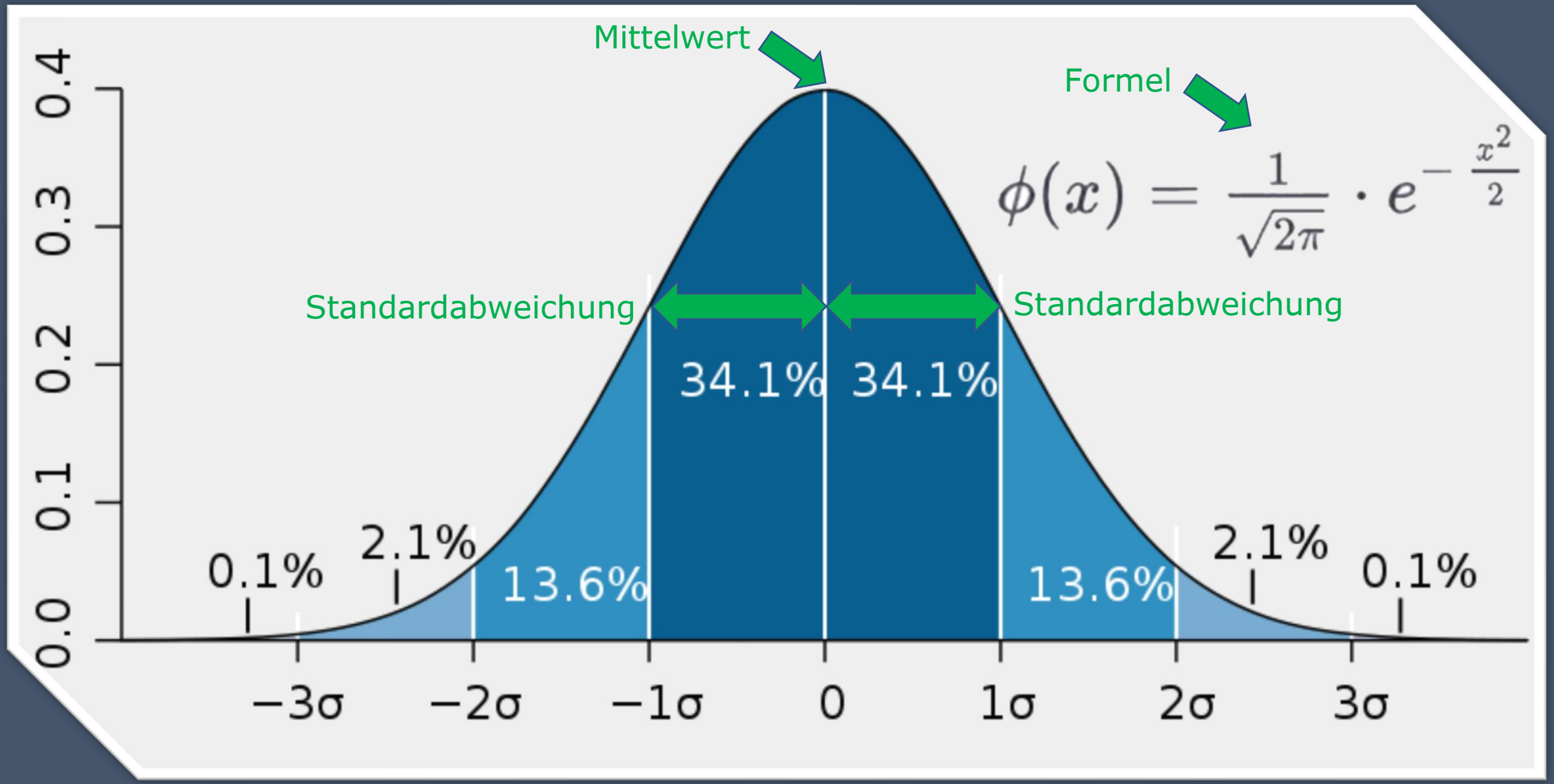
- Ist die Standardabweichung der Stichprobenverteilung eines Parameters
- Wird gebraucht, um Konfidenzintervalle zu berechnen
- Formeln:

	Formal	Nicht-formal
Mittelwerte	$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$	$Standardfehler_{Mittelwert} = \frac{Standardabweichung}{\sqrt{Anzahl\ der\ Fälle}}$
Prozentwerte	$SE_p = \sqrt{\frac{p * (100 - p)}{n}}$	$Standardfehler_{Prozentwert} = \sqrt{\frac{Prozentwert * (100 - Prozentwert)}{Anzahl\ der\ Fälle}}$

Die Standardnormalverteilung

Die Standardnormalverteilung

- Ist die Normalverteilung mit einem Mittelwert von 0, einer Varianz von 1 und einem Flächeninhalt von 1
- Wird auch „z-Verteilung“ genannt
- Wird benutzt, um Skalen zu vereinheitlichen und damit vergleichbar zu machen
- Erhält man durch z-Standardisierung: $z_i = \frac{x_i - \bar{x}}{s}$
- R-Befehl zum Standardisieren: scale(variable)


Mittelwert
Formel
$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$
Standardabweichung
Standardabweichung
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
0.0
0.1
0.2
0.3
0.4
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

Konfidenzintervalle – Basics

Konfidenzintervalle (confidence intervals)

- Schätzen anhand eines Parameters der Stichprobe, in welchem Wertebereich sich der Parameter der Grundgesamtheit befindet
- Gängige Parameter: Mittelwerte, Prozentwerte
- können verschieden sicher sein:
  - Je größer die Irrtumswahrscheinlichkeit, desto geringer der Wertebereich
  - Je geringer die Irrtumswahrscheinlichkeit, desto größer der Wertebereich
- Übliche Irrtumswahrscheinlichkeit: $\alpha = 0.05$, z-Wert ≈ 2

Konfidenzintervalle – Formeln

	Formal	Nicht-formal
Mittelwerte	$CI_{\bar{x}} = \bar{x} \pm z_{\alpha/2} * SE_{\bar{x}}$	Mittelwert $\pm$ z-Wert der Irrtumswahrscheinlichkeit * Standardfehler der Mittelwertverteilung
Prozentwerte	$CI_p = p \pm z_{\alpha/2} * SE_p$	Prozentwert $\pm$ z-Wert der Irrtumswahrscheinlichkeit * Standardfehler der Prozentwertverteilung

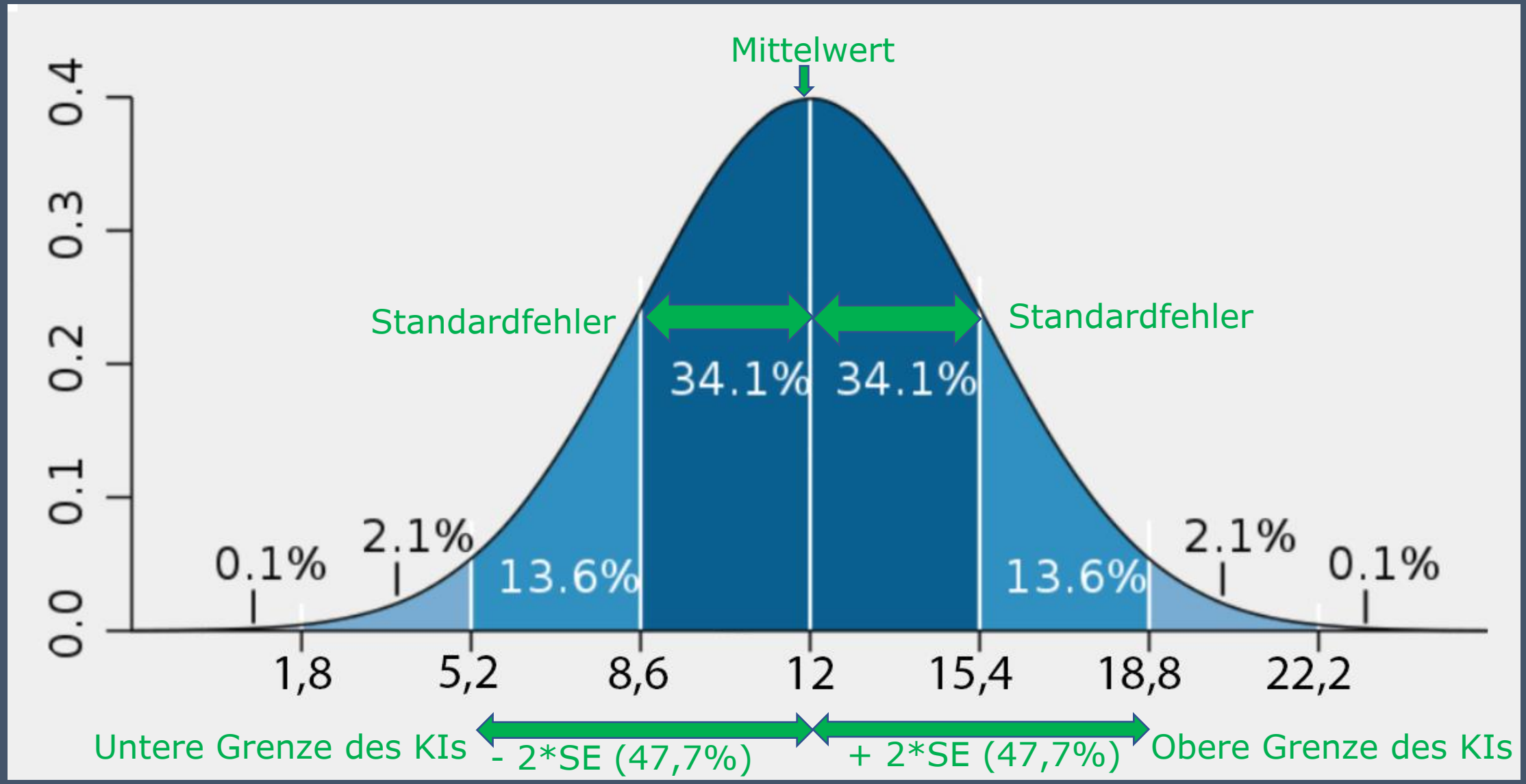
Konfidenzintervalle Mittelwert: Beispiel

Fall	1	2	3	4	5
Ausprägung	15	0	10	20	15

$$\bar{x} = \frac{15 + 0 + 10 + 20 + 15}{5} = 12 \qquad s = \sqrt{\frac{230}{5 - 1}} \approx 7{,}6 \qquad n = 5$$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{7{,}6}{\sqrt{5}} \approx 3{,}4 \qquad CI_{\bar{x}} = \bar{x} \pm z_{\alpha/2} * SE_{\bar{x}} = 12 \pm 2 * 3{,}4 = 12 \pm 6{,}8$$

→ Der Mittelwert der Grundgesamtheit liegt mit einer Irrtumswahrscheinlichkeit von 5% zwischen 5,2 und 18,8. Also: Im Durchschnitt lesen Deutsche ungefähr zwischen 5 und 20 Minuten Zeitung pro Tag.


Mittelwert
Standardfehler
Standardfehler
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
0.0
0.1
0.2
0.3
0.4
1,8
5,2
8,6
12
15,4
18,8
22,2
Untere Grenze des KIs
- 2*SE (47,7%)
+ 2*SE (47,7%)
Obere Grenze des KIs

Konfidenzintervalle Mittelwert: Aufgabe

Aufgabe: Die Variable beispiel$Minuten.Internetkonsum besteht aus 11 Fällen und hat einen Mittelwert von 196 sowie eine Standardabweichung von 97. Schätzt mit einer Irrtumswahrscheinlichkeit von 5%, in welchem Intervall der Mittelwert der Grundgesamtheit liegt. Rundet auf eine Nachkommastelle.

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$CI_{\bar{x}} = \bar{x} \pm z_{\alpha/2} * SE_{\bar{x}}$$

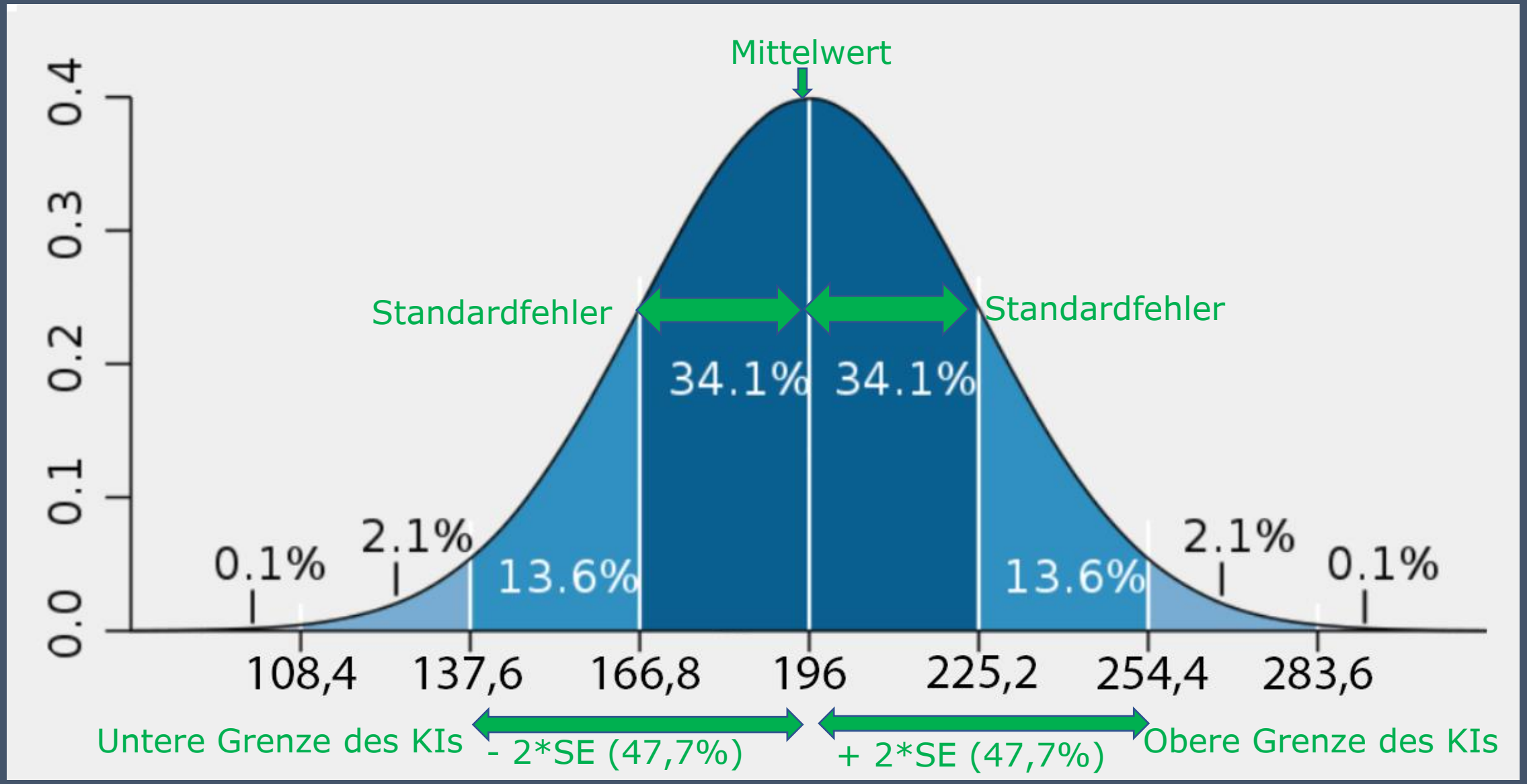
Konfidenzintervalle Mittelwert: Lösung

$\bar{x} = 196,\ s = 97,\ n = 11,\ z_{0.05/2} \approx 2$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{97}{\sqrt{11}} \approx 29.2$$

$$CI_{\bar{x}} = \bar{x} \pm z_{\alpha/2} * SE_{\bar{x}} = 196 \pm 2 * 29.2 = 196 \pm 58{,}4$$

→ Der Mittelwert der Grundgesamtheit liegt mit einer Irrtumswahrscheinlichkeit von 5% zwischen 137,6 und 254,4. Also: Im Durchschnitt surfen Deutsche zwischen 138 und 254 Minuten pro Tag im Internet.


Mittelwert
Standardfehler
Standardfehler
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
0.0
0.1
0.2
0.3
0.4
108,4
137,6
166,8
196
225,2
254,4
283,6
Untere Grenze des KIs
- 2*SE (47,7%)
+ 2*SE (47,7%)
Obere Grenze des KIs

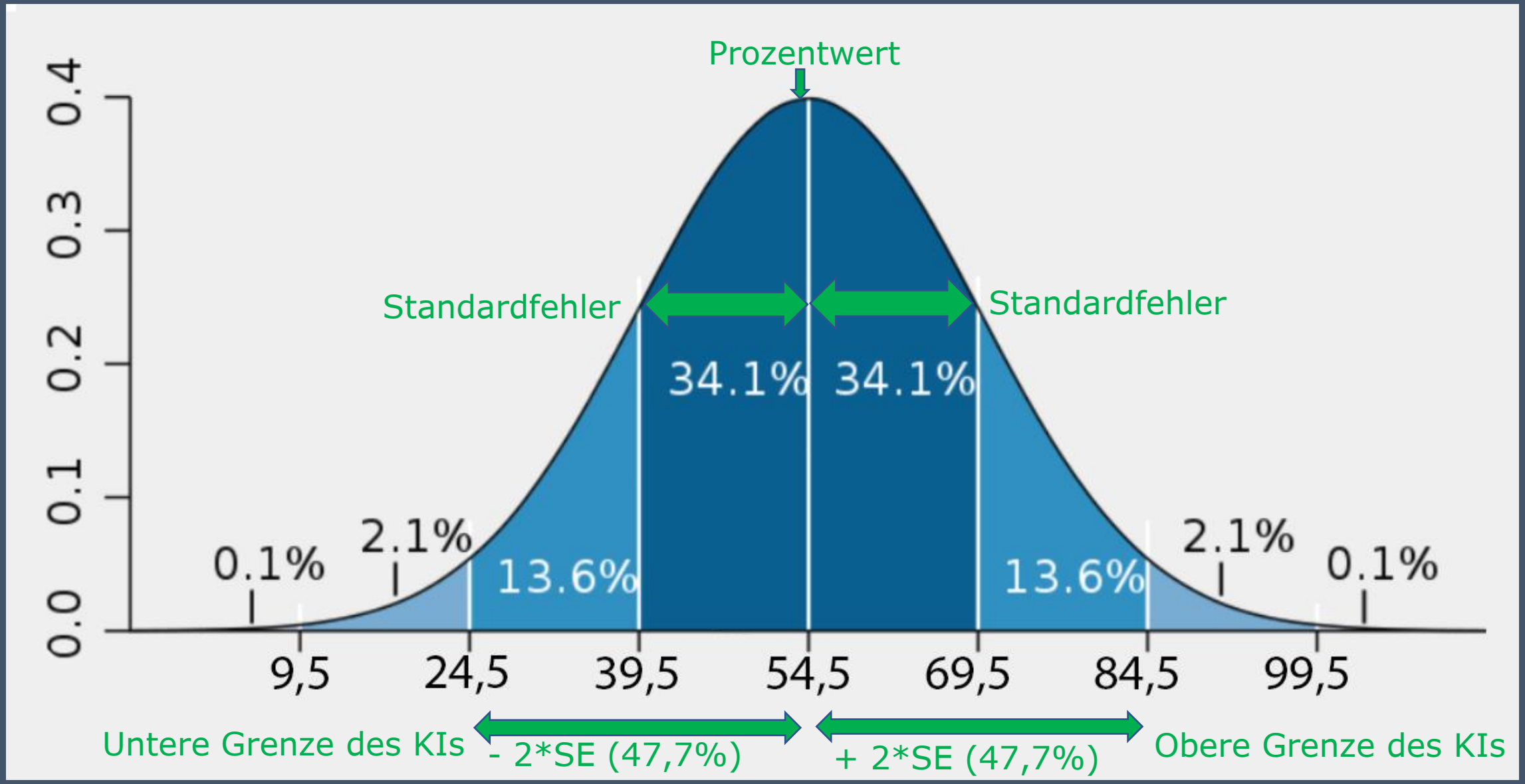
Konfidenzintervalle Prozentwert: Beispiel

Ausprägung	Anzahl	Prozent	Kumulierte Prozent
Single	6	54,5	54,5
Liiert	5	45,5	100

$$SE_p = \sqrt{\frac{p*(1-p)}{n}} = \sqrt{\frac{54{,}5*(100-54{,}5)}{6+5}} = \sqrt{\frac{54{,}5*45.5}{11}} \approx 15$$

$$CI_p = p \pm z_{\alpha/2} * SE_p = 54{,}5 \pm 2 * 15 = 54{,}5 \pm 30$$

→ Der Prozentwert der Grundgesamtheit liegt mit einer Irrtumswahrscheinlichkeit von 5% zwischen 24,5% und 84,5%. Deutsche sind zwischen einer Wahrscheinlichkeit von 24,5% und 84,5% Single.


Prozentwert
Standardfehler
Standardfehler
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
0.0
0.1
0.2
0.3
0.4
9,5
24,5
39,5
54,5
69,5
84,5
99,5
Untere Grenze des KIs
- 2*SE (47,7%)
+ 2*SE (47,7%)
Obere Grenze des KIs

Konfidenzintervall Prozentwerte: Aufgabe

Aufgabe: Im Beispieldatensatz haben 7 von 11 Personen angegeben, weiblich zu sein. Schätzt mit einer Fehlerwahrscheinlichkeit von 5%, wie viele Frauen in der Grundgesamtheit weiblich sind. Rundet jeweils auf eine Nachkommastelle.

$$SE_p = \sqrt{\frac{p * (100 - p)}{n}}$$

$$CI_p = p \pm z_{\alpha/2} * SE_p$$

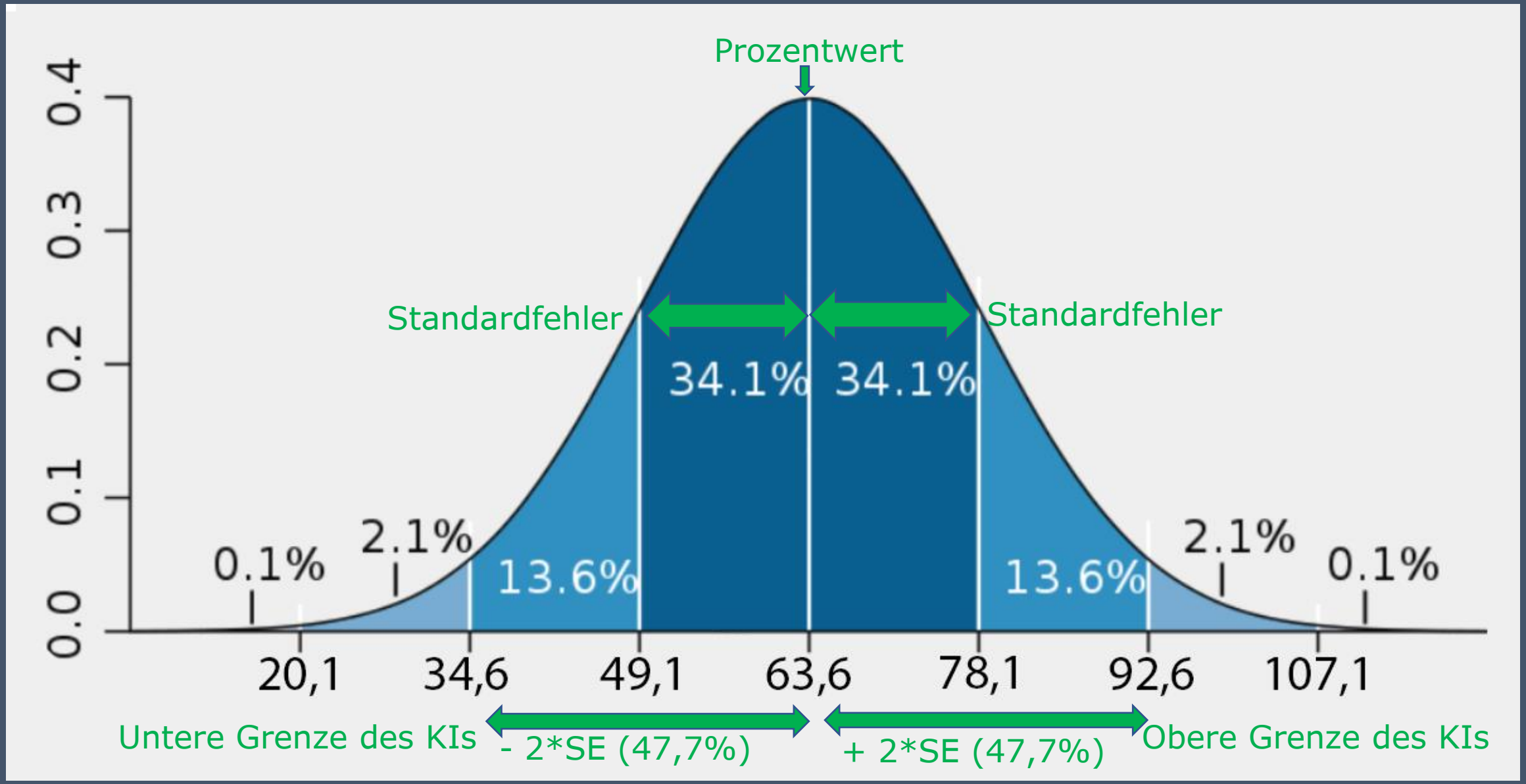
Konfidenzintervalle Prozentwerte: Lösung

$$p = \frac{7}{11} * 100 \approx 63{,}6,\ n = 11,\ z_{0.05/2} \approx 2$$

$$SE_p = \sqrt{\frac{p*(1-p)}{n}} = \sqrt{\frac{63{,}6 * (100-63{,}6)}{11}} = \sqrt{\frac{63{,}6*36{,}4}{11}} \approx 14{,}5$$

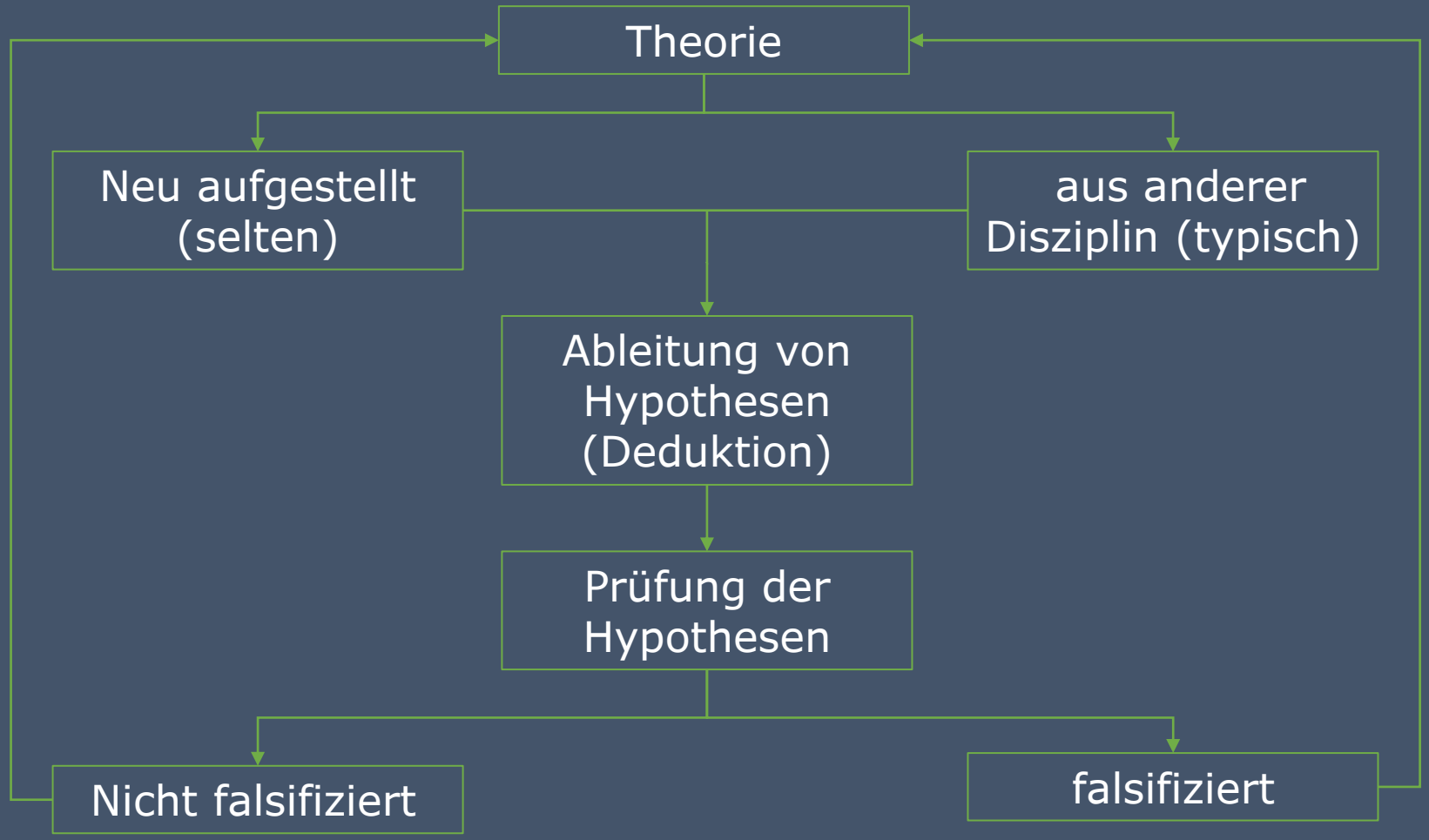
$$CI_p = p \pm z_{\alpha/2} * SE_p = 63{,}6 \pm 2 * 14{,}5 = 63{,}6 \pm 29$$

→ Der Prozentwert der Grundgesamtheit liegt mit einer Irrtumswahrscheinlichkeit von 5% zwischen 34,6% und 92,6%. Also: Deutsche sind zwischen einer Wahrscheinlichkeit von 34,6% und 92,6% weiblich.


Prozentwert
Standardfehler
Standardfehler
34.1%
34.1%
0.1%
2.1%
13.6%
13.6%
2.1%
0.1%
0.0
0.1
0.2
0.3
0.4
20,1
34,6
49,1
63,6
78,1
92,6
107,1
Untere Grenze des KIs
- 2*SE (47,7%)
+ 2*SE (47,7%)
Obere Grenze des KIs

Wiederholung: Kritischer Rationalismus

Wiederholung (ggf. andere Methodik), Forschungsgegenstand erweitern



Verwerfen der Hypothese, Anpassung der Theorie, Methodenkritik

Arten von Hypothesen

Die Forschungshypothese („H_1“)

- Ist die theoretisch fundierte und empirisch überprüfbare Behauptung über einen Zusammenhang/einen Unterschied mindestens zweier Variablen
- Wird auch „Alternativhypothese“ genannt
- Beispiel: Rauchen erhöht das Lungenkrebsrisiko

Die Nullhypothese („H_0“)

- Behauptet, dass es keinen Zusammenhang/Unterschied zwischen den in der Forschungshypothese genannten Variablen gibt
- Ist die Hypothese, von der wir statistisch gesehen ausgehen
- Beispiel: Rauchen erhöht das Lungenkrebsrisiko nicht

Hypothesentests – Basics

Hypothesentests

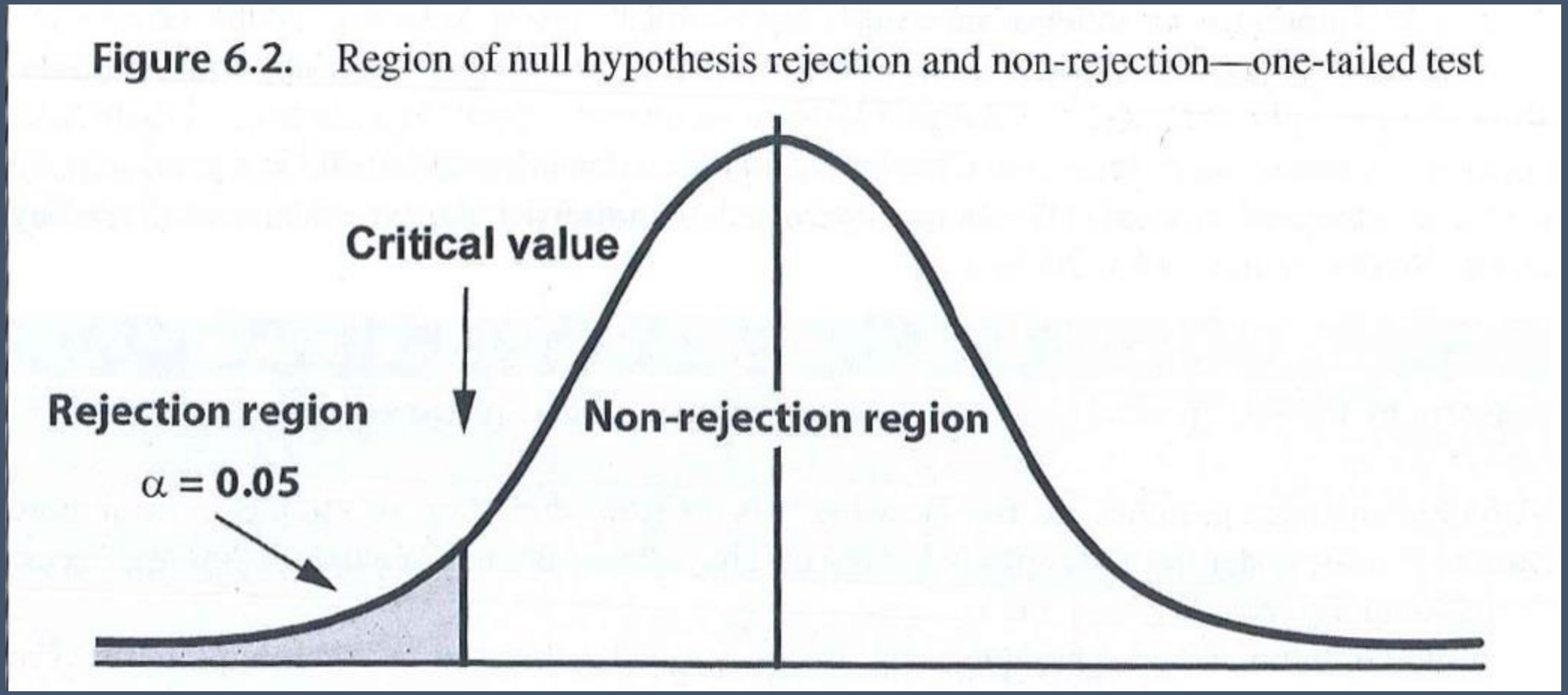
- Versuchen, die Nullhypothese zu falsifizieren, weil die Forschungshypothese nicht verifizierbar ist → kritischer Rationalismus
- Schätzen anhand von Parametern der Stichprobe, wie wahrscheinlich es ist, dass die Nullhypothese zutrifft
- Können fehlerbehaftet sein

Hypothesentests – Basics

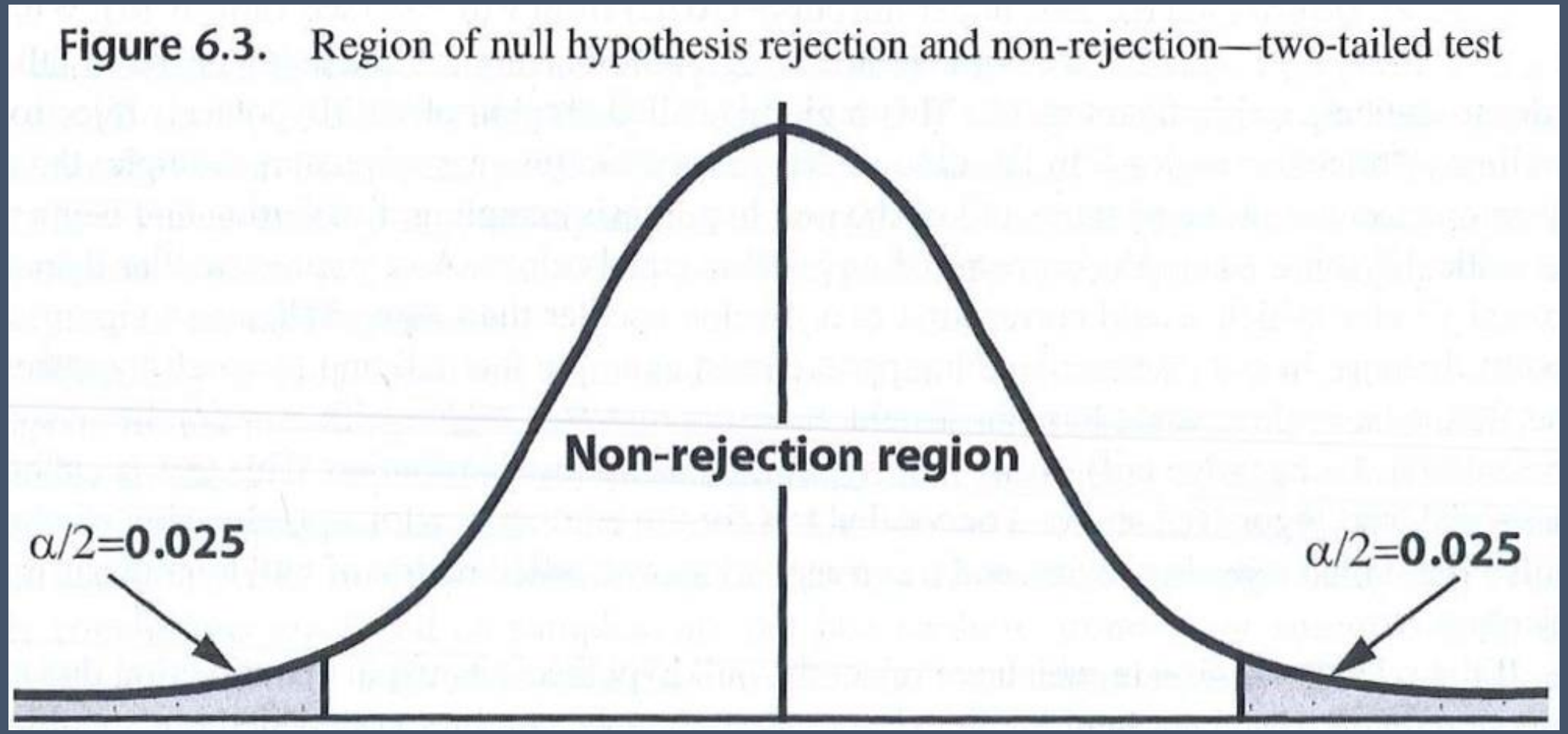
Hypothesentests

- Können auf zwei Arten durchgeführt werden:
  - Um 0 (die Nullhypothese) wird ein Konfidenzintervall gebildet und geschaut, ob der Parameter darin liegt
  - Um den Parameter wird ein Konfidenzintervall gebildet und geschaut, ob der Wert 0 (die Nullhypothese) darin liegt
- Sind abhängig von der Richtung der Hypothese:
  - Gerichtet: einseitiger Test (positiv: rechts, negativ: links)
  - Ungerichtet: beidseitiger Test

Einseitige Hypothesentests



Beidseitiger Hypothesentest



p-Werte

p-Werte

- Geben an, wie wahrscheinlich es ist, dass, wenn die Nullhypothese stimmt, der mit dem Hypothesentest ermittelte Wert vorkommt
- Werden benutzt, um zu ermitteln, ob ein Test signifikant ist
  - Ein Test ist signifikant genau dann, wenn der p-Wert kleiner ist als die Irrtumswahrscheinlichkeit α (i.d.R.: $p<0.05$)
  - Wenn ein Hypothesentest signifikant ist, wird die Forschungshypothese nicht verworfen (ist aber dennoch nicht bestätigt!)

Hypothesentests – Probleme

- α-Fehler:
  - Die Nullhypothese stimmt (es gibt keinen Zusammenhang/Unterschied), aber wird verworfen
  - Beispiel: Ein Medikament hat tatsächlich keinen Effekt, der Test mit der Stichprobe suggeriert ihn aber
- β-Fehler
  - Die Forschungshypothese stimmt, aber wird verworfen
  - Beispiel: Ein Medikament hat einen Effekt, der Test mit der Stichprobe stellt aber keinen fest

Rechnen mit Vektoren

Zeichen	Bedeutung	Zeichen	Bedeutung
+	Addition	-	Subtraktion
*	Multiplikation	/	Division
^	Potenz	sqrt(x)	Wurzel
exp(x)	Exponentialfunktion	log(x)	Natürlicher Logarithmus
sum(x)	Summe	abs(x)	Absoluter Wert

FRAGEN?

Selbststudium

- Vertiefen: Schätzen und Testen (!)

Bis nächste Woche!