ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Bemerkungen bzgl statistichen Tests

(basierend auf Slides von Lukas Meier)

Einseitige vs. zweiseitige Tests

- Die Entscheidung für eine einseitige oder zweiseitige Alternative H_A hängt von der **Fragestellung** ab.
- Eine einseitige Alternative ist dann angebracht, wenn nur ein Unterschied in eine **bestimmte Richtung** von Bedeutung / Interesse ist (Bsp. **Überschreitung** Grenzwert).
- Der einseitige Test ist auf der einen (irrelevanten) Seite «blind», dafür verwirft er auf der anderen (relevanten) Seite früher als der zweiseitige Test (da der Verwerfungsbereich früher beginnt).
- Man sagt auch, dass er eine **grössere Macht** hat in diesem Bereich (siehe später).

p-Wert

- Zur Erinnerung: Test mittels Verwerfungsbereich:
  - Wir setzen das Signifikanzniveau α im voraus fest.
  - Aus α und der Verteilung der Teststatistik unter H_0 berechnen wir den Verwerfungsbereich. Je kleiner (grösser) α, desto kleiner (grösser) ist der Verwerfungsbereich.
  - Beachte: Das Signifikanzniveau α und der Verwerfungsbereich sind fix und hängen nicht von den Daten ab. Die Teststatistik hängt von den Daten ab und ist eine Zufallsvariable.
  - Wir verwerfen H_0 falls der realisierte Wert der Teststatistik im Verwerfungsbereich liegt.
- Alternativ: wir benutzen den p-Wert anstelle vom Verwerfungsbereich
- Definition des **p-Werts:** Der p-Wert eines Tests ist die W'keit, unter der Nullhypothese ein **mindestens so extremen Wert der Teststatistik** (bzgl der Alternative) zu beobachten wie das aktuell beobachtete.

Illustration p-Wert beim einseitigen T-Test («nach oben»)

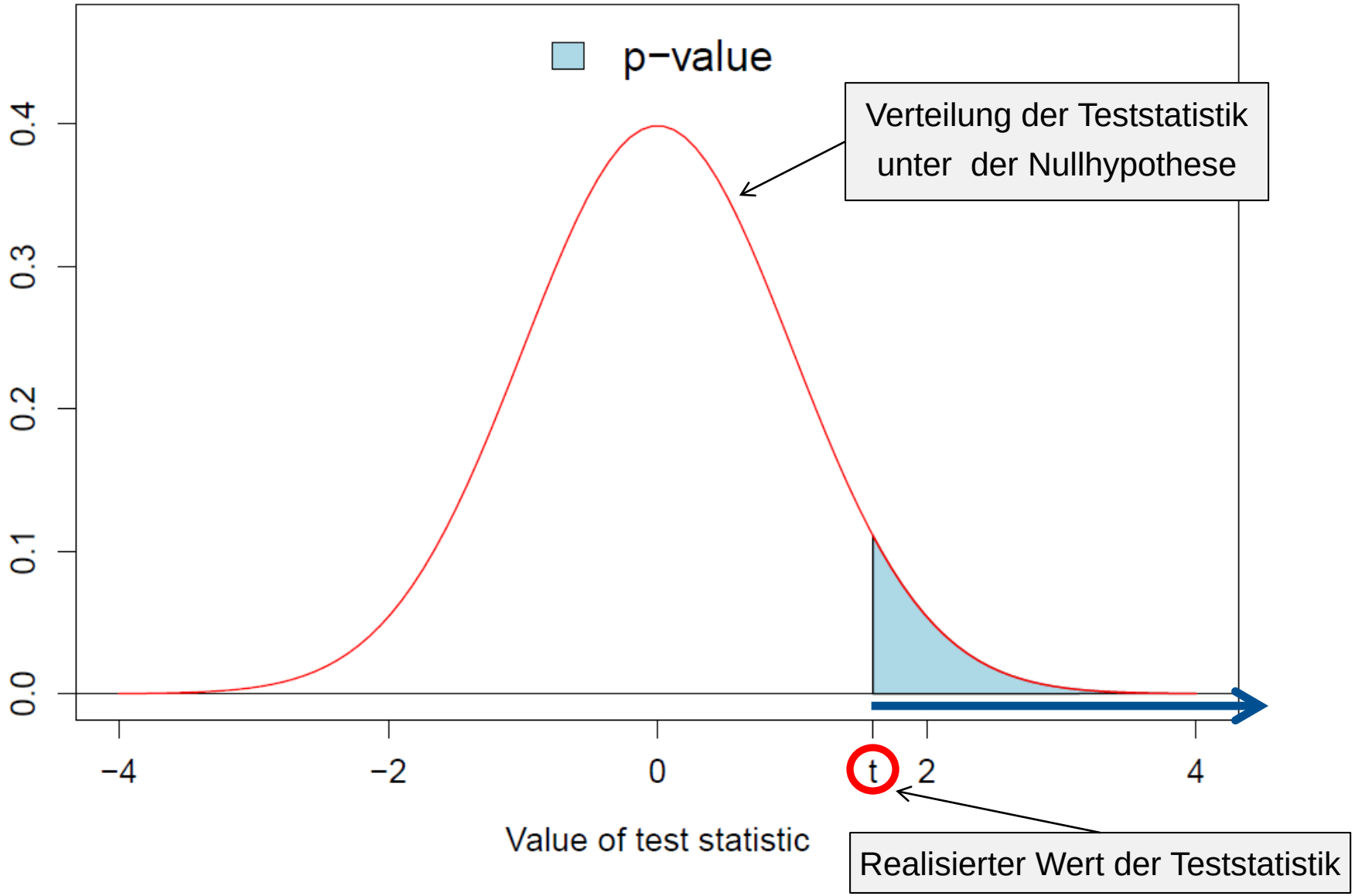


Illustration p-Wert beim einseitigen T-Test («nach unten»)

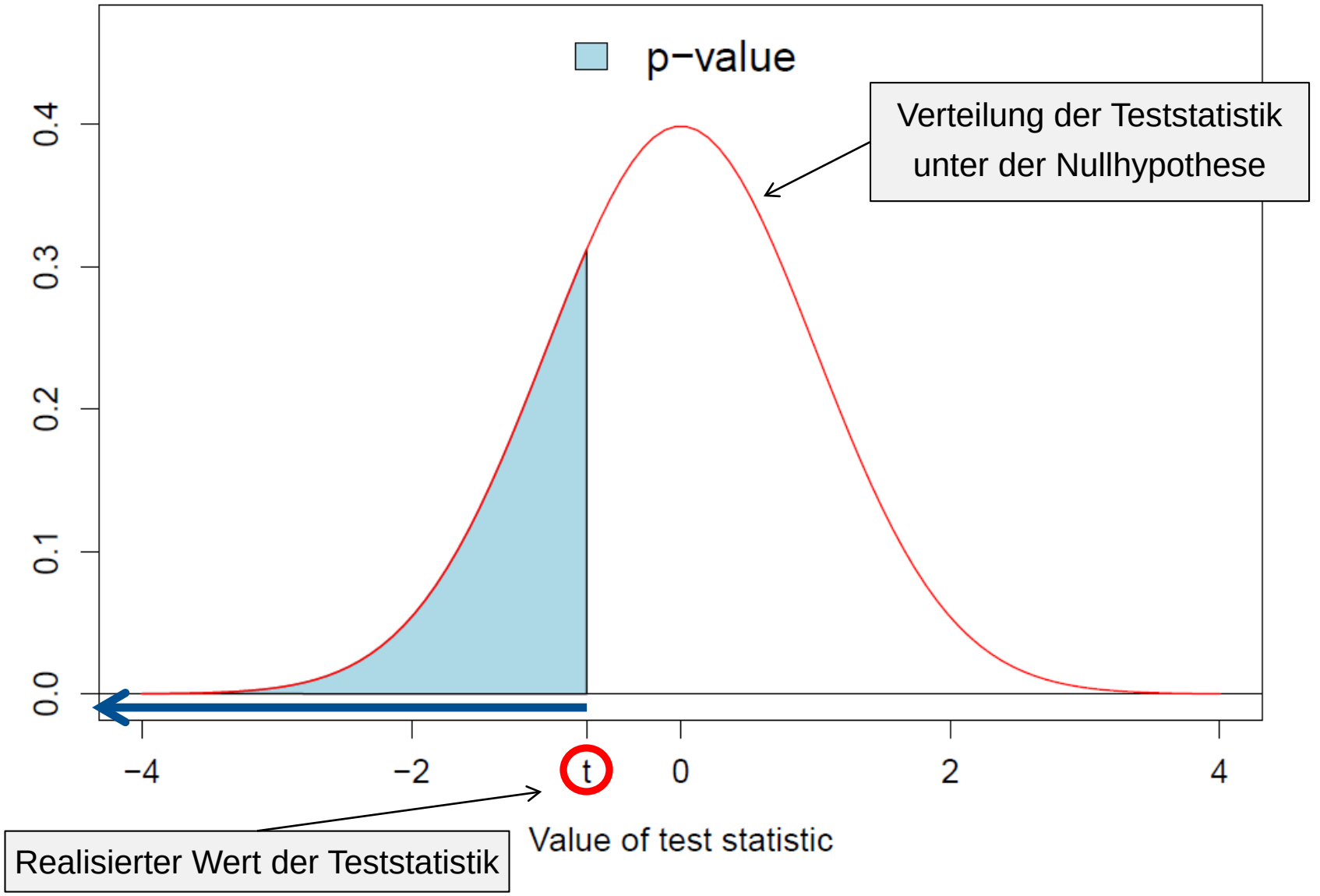
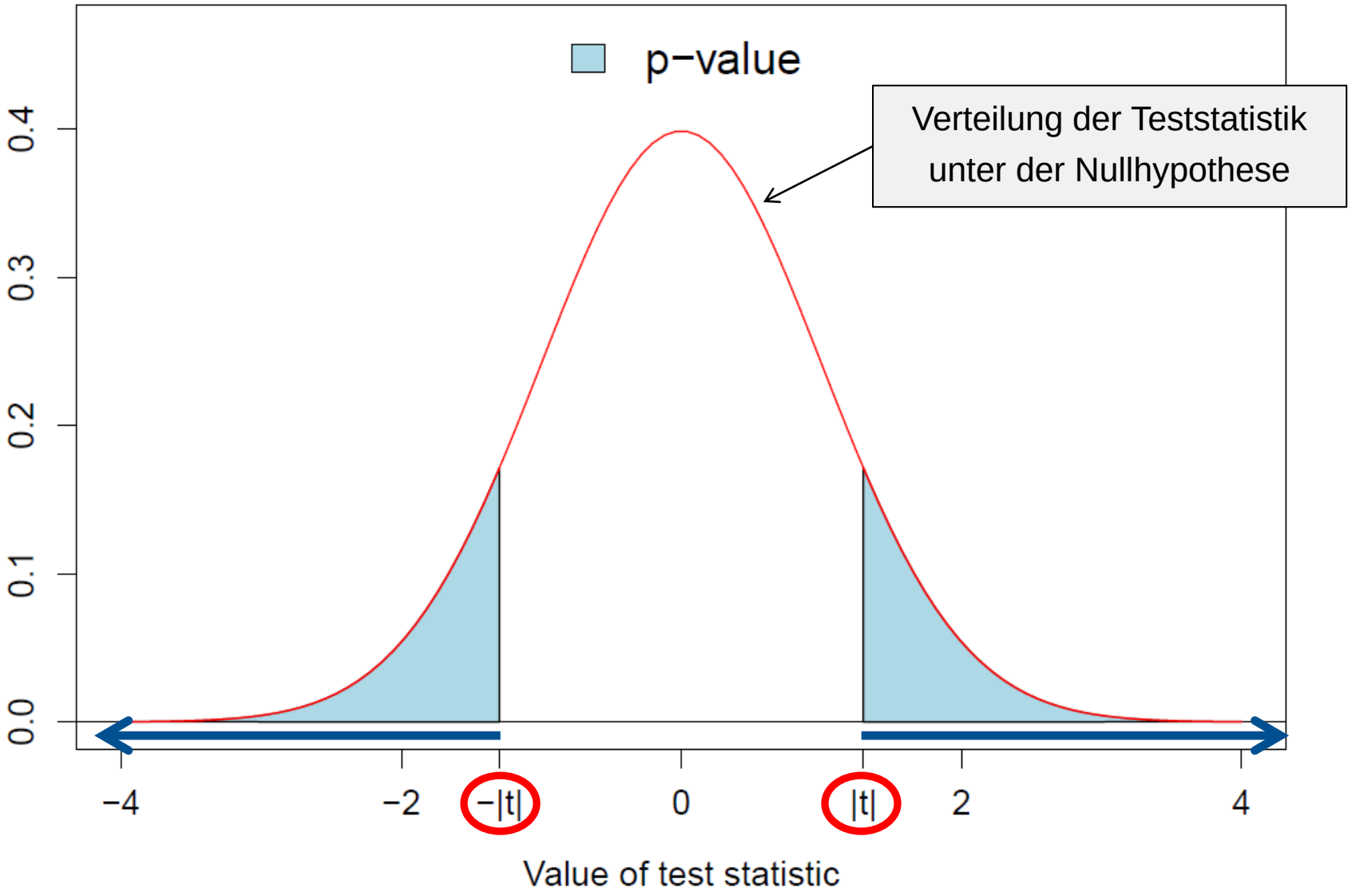


Illustration p-Wert beim zweiseitigen T-Test



p-Wert

- Es gilt (siehe Wandtafel): $pWert \leq \alpha \Leftrightarrow Teststatistik\ im\ Verwerfungsbereich$

- Test mittels p-Wert:
  - Wir setzen das Signifikanzniveau α im voraus fest.
  - Wir berechnen den p-Wert.
  - Beachte: Das Signifikanzniveau α ist fix und hängt nicht von den Daten ab. Der p-Wert hängt von den Daten ab und ist also eine Zufallsvariable.
  - Wir verwerfen H_0 falls $pWert \leq \alpha$.

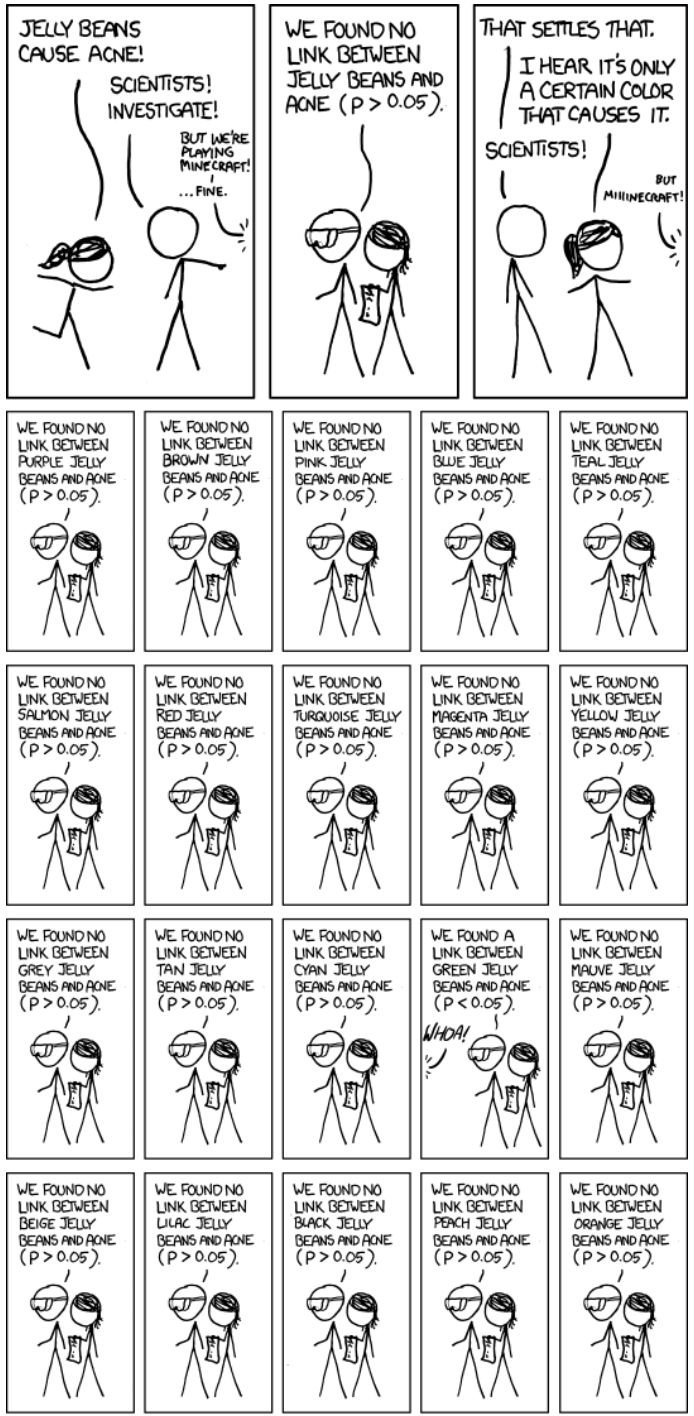
- Clickerfrage p-Wert

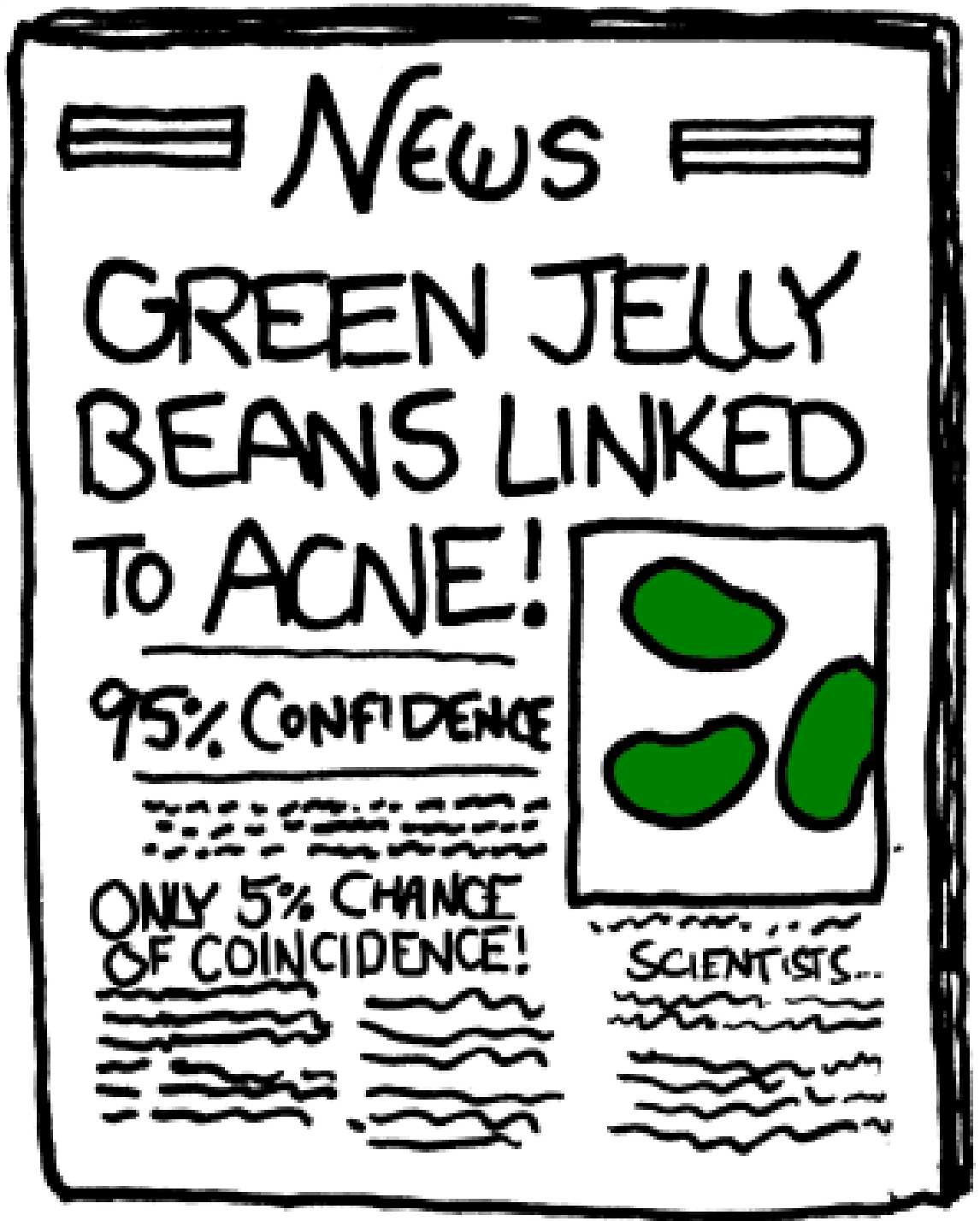
p-Wert

- Beachte: der p-Wert ist eine Wahrscheinlichkeit, **berechnet unter der Annahme, dass H_0 stimmt**. Er sagt also nichts über die Wahrscheinlichkeit **ob** H_0 oder H_A stimmt. Insbesondere:
  - $pWert \neq P(H_0\ stimmt)$
  - $pWert \not\geq P(Fehler\ 1.\ Art)$

p-Wert: Nutzen / Gefahren

- Der p-Wert kann als «**standardisierte Teststatistik**» verwendet werden. Wir können am p-Wert **direkt** ablesen, ob die Nullhypothese verworfen wird.
- Einige «Gefahren» des p-Werts:
  - Ein kleiner p-Wert ist nicht automatisch fachlich relevant, denn der **p-Wert sagt nichts über die Effektgrösse**.
    ⇒ Berechne auch das Vertrauensintervall.
  - Multiples Testing / p-value Hacking: Falls H_0 gilt, dann erwartet man in $\alpha \times 100\%$ der Tests einen signifikanten p-Wert (i.e., $pWert \leq \alpha$). **Falls man also genügend viele Tests macht, dann findet man immer einen signifikanten p-Wert.** Die Garantie $P(Fehler\ 1.Art) \leq \alpha$ gilt nur für einen einzelnen Test!
    ⇒ Mache nur einen im voraus genau beschriebenen Test. Oder beschreibe wieviele Tests gemacht wurden, und benutze multiple testing correction.
- Interessante Artikel:
  - http://www.nature.com/news/scientific-method-statistical-errors-1.14700
  - http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503





Source:

https://xkcd.com/882/

Multiple testing correction

- Einfachste Methode: Bonferroni correction:
  - Wenn man K tests macht, und man möchte

$$P(Fehler\ 1.\ Art\ in\ mindestens\ einem\ Test) \leq \alpha$$

dann kann man jeden einzelnen Test zum Niveau α/K machen.

  - Beweis:

$$P(Fehler\ 1.\ Art\ in\ mindestens\ einem\ der\ K\ Tests)$$

$$= P(\{Fehler\ 1.\ Art\ in\ Test\ 1\}\ OR\ \ldots OR\ \{Fehler\ 1.\ Art\ in\ Test\ K\})$$

$$\leq \sum_{j=1}^{K} P(Fehler\ 1.\ Art\ in\ Test\ j) \leq \sum_{j=1}^{K} \frac{\alpha}{K} = \alpha$$

Macht

- Sei $\boldsymbol{\theta} \in \boldsymbol{H_A}$ und sei

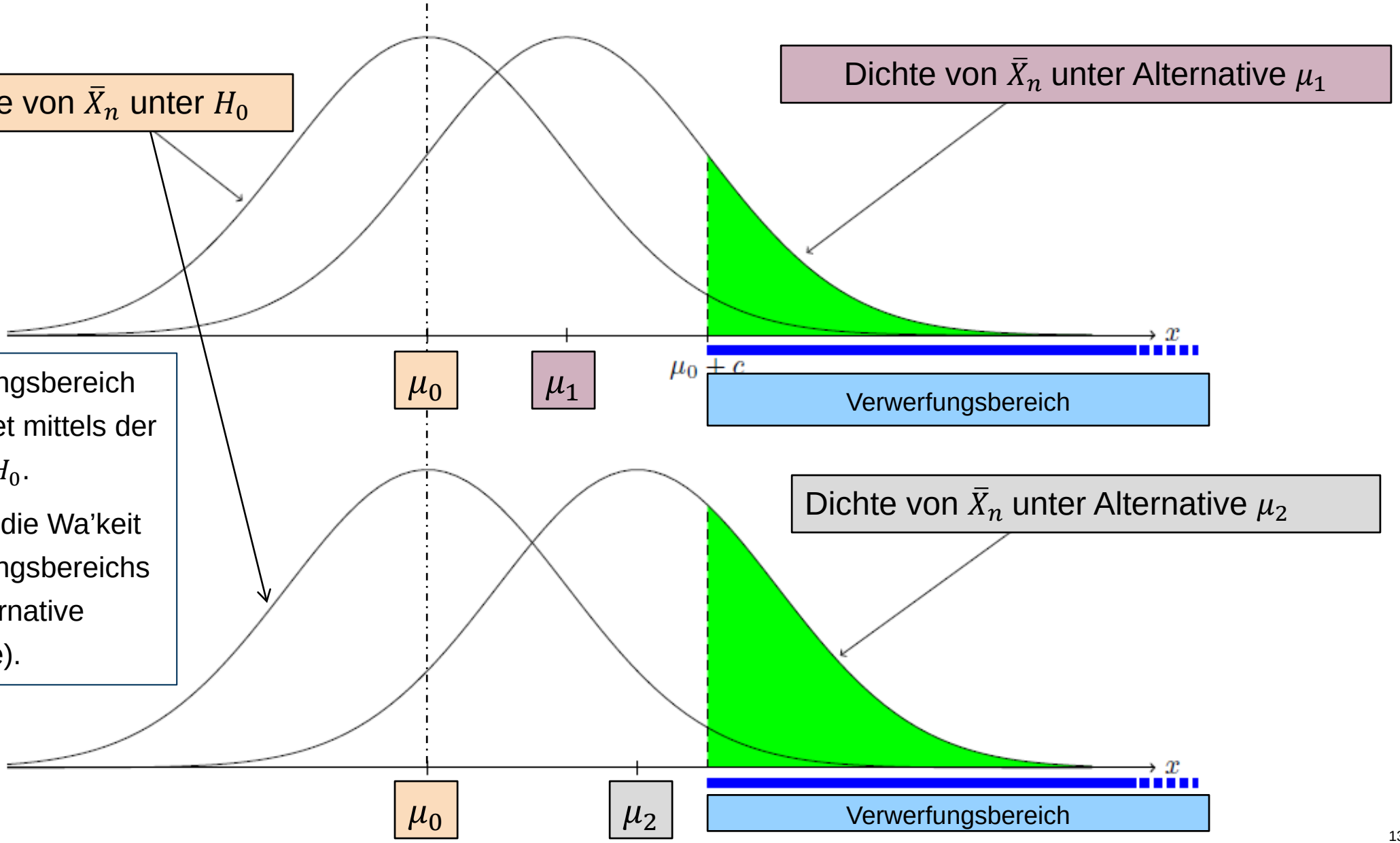
$$\beta(\theta) = P_\theta(Fehler\ 2.Art) = P_\theta(Test\ verwirft\ H_0\ nicht).$$

Die Macht eines Tests ist dann:

$$\mathrm{P}_\theta(Test\ verwirft\ H_0) \;=\; 1 - P_\theta(Test\ verwirft\ H_0\ nicht) \;=\; 1 - \beta(\theta).$$

- Die Macht hängt also von θ ab.

Macht beim einseitigen Z-Test $(H_0: \mu = \mu_0, H_A: \mu > \mu_0)$



- Der Verwerfungsbereich wird berechnet mittels der Dichte unter H_0.
- Die Macht ist die Wa'keit des Verwerfungsbereichs unter der Alternative (grüne Fläche).

Bemerkungen zur Macht

- Je grösser der Unterschied zwischen μ_0 und μ_A, desto grösser wird die Macht.
- Je grösser die Stichprobe, desto grösser wird die Macht. Begründung: Weil $Var(\bar{X}_n) = \sigma^2/n$, konzentrieren die Dichten sich mehr um μ_0 und μ_A.

- Die Macht ist wichtig zur Ermittlung der nötigen Stichprobengrösse.
  - Sie vermuten z.B. eine **bestimmte Abweichung** von der Nullhypothese (z.B. $\mu = 1$ statt $\mu = \mu_0 = 0$).
  - Sie planen ein Experiment und wollen mit einer Wahrscheinlichkeit von 80% die Nullhypothese verwerfen können (= Macht).
  - Man kann dann die **nötige Stichprobengrösse** n berechnen.

Statistische Signifikanz vs. Relevanz

- Statistische Tests werden in der Praxis oft «missbraucht» und falsch angewendet → schlechter Ruf der Statistik.
- Das Problem ist: Je grösser unsere Stichprobe ist, desto eher werden wir signifikante Effekte finden, denn die Nullhypothese stimmt in der Regel nie **exakt**.(Zur Erinnerung: Je grösser die Stichprobe, desto grösser die Macht.)
- Wenn wir z.B. $H_0: \mu = 400$ testen und in Tat und Wahrheit gilt aber $\mu = 401$, so werden wir bei genügend grosser Stichprobe n mit hoher Wahrscheinlichkeit ein signifikantes Testresultat erhalten.
- Ob etwas signifikant ist, ist also unter anderem eine Frage des **Aufwands** ($).

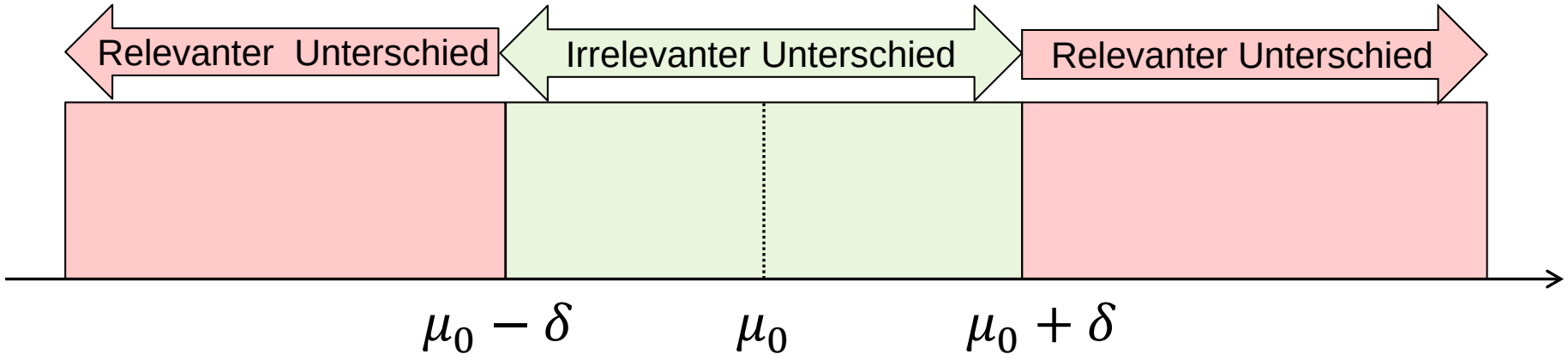
Statistische Signifikanz vs. Relevanz

- Die (wichtigere) Frage ist: Wann haben wir ein **relevantes** Resultat?
- Wir müssen vorher definieren, was «Relevanz» bedeutet.
- Was ein **relevanter Unterschied** ist, hängt ab vom Fachgebiet / Fachwissen. Die Statistik hat hier **keine Antwort**!
- Bsp: Durchmesser von Zylinderscheiben:

  Mit was für Abweichungen vom Sollwert kann man leben?

Statistische Signifikanz vs. Relevanz

- Wir müssen also eine **Differenz** δ angeben, ab der man sagt, dass ein Unterschied **relevant** ist für eine entsprechende Anwendung.

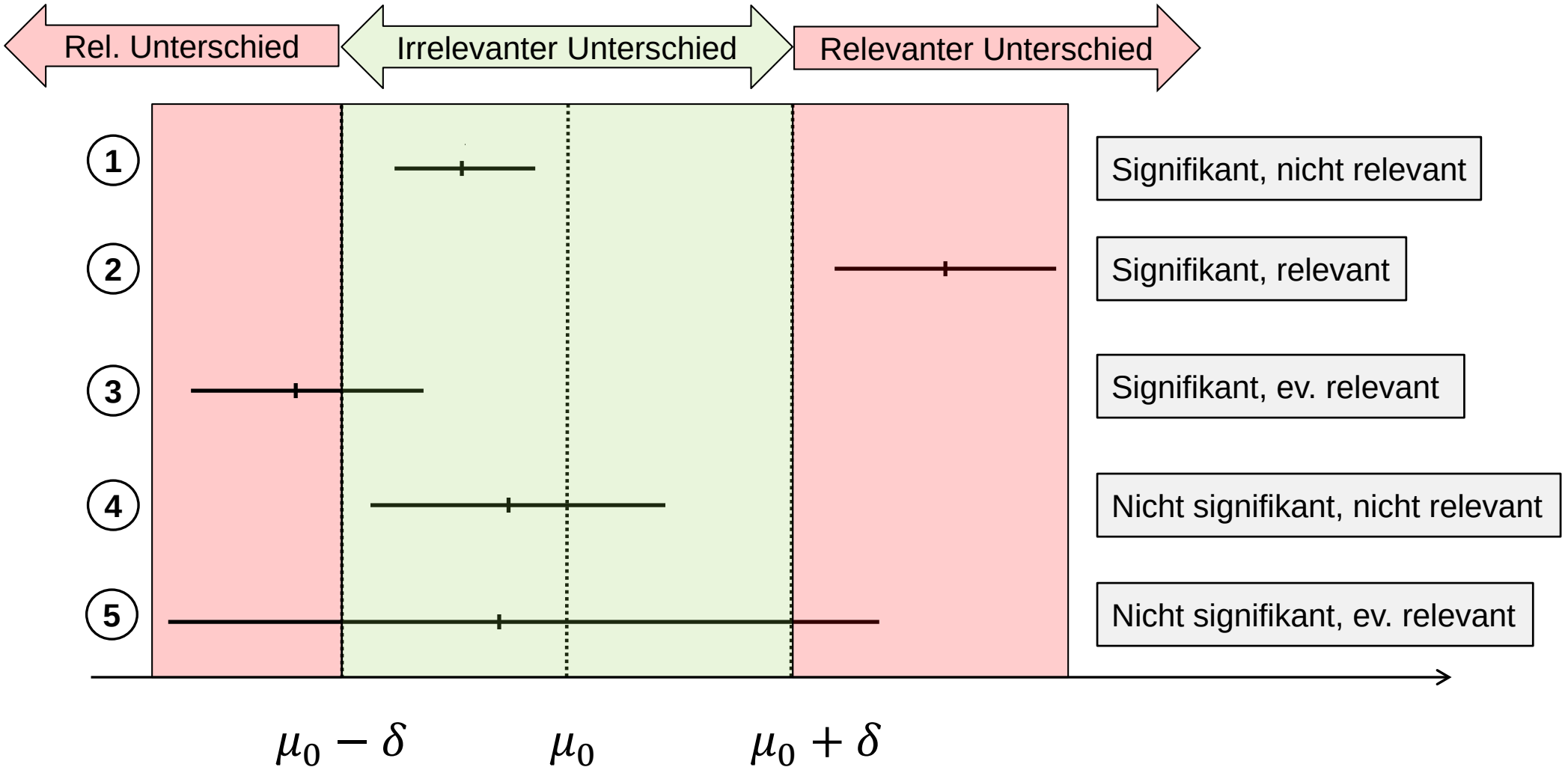


- Basierend auf unseren Daten berechnen wir dann ein Vertrauensintervall für den Parameter von Interesse.
- Die Idee besteht nun darin, dass man schaut, wo das **Vertrauensintervall** bzgl. obigen Bereichen liegt.

Statistische Signifikanz vs. Relevanz

- Liegt das Vertrauensintervall ganz im «relevanten Bereich», so spricht man von einem relevanten Effekt.
- Ist zwar der Test signifikant (d.h. VI enthält μ_0 nicht) aber das VI liegt ganz im «irrelevanten Bereich», so hat man zwar ein signifikantes, aber **kein** relevantes Resultat.
- Siehe auch Bsp. nächste Slide.

Statistische Signifikanz vs. Relevanz



Statistische Signifikanz vs. Relevanz

- Man kombiniert also «das Beste aus beiden Welten»:

  Das Fachwissen und die Statistik, die einem hilft, die Unsicherheit zu quantifizieren (durch das VI).

- Es reicht in der Regel also nicht, sich nur «blind» auf die statistische Signifikanz zu verlassen (obwohl dies vielerorts so gemacht wird).

- Wir müssen uns zusätzlich auch immer fragen: «Ist das auch ein **relevantes** Resultat?».