Hypothesentests II: Parametrische Tests

Lektion B09 – BSc Wahrscheinlichkeit und Statistik

Digital Finance

Gliederung

1. Recap: Hypothesentest-Ablauf
2. Binomialtest
3. Einstichproben-t-Test
4. Zweistichproben-t-Test
5. Paardifferenzentest
6. Welch-Test vs. gepoolter t-Test
7. Effektgrösse
8. R-Code
9. Zusammenfassung

Lernziele

Am Ende dieser Lektion werden Sie in der Lage sein:

1. Den exakten Binomialtest auf Anteilswerte anzuwenden
2. Den Einstichproben-t-Test bei unbekannter Standardabweichung durchzuführen
3. Zwei unabhängige Stichproben mit dem Zweistichproben-t-Test zu vergleichen
4. Den Paardifferenzentest (gepaarter t-Test) korrekt einzusetzen
5. Welch-Test und gepoolten t-Test zu unterscheiden
6. Effektgrösse (Cohen's d) zu berechnen und zu interpretieren
7. Den passenden parametrischen Test anhand eines Entscheidungsschemas auszuwählen

Diese Ziele leiten, was Sie aus dieser Lektion beherrschen sollten.

Kurze Wiederholung aus B08

5-Schritte-Verfahren (aus Lektion B08):

1. H_0 und H_1 formulieren
2. Signifikanzniveau α festlegen (üblich: 0,05)
3. Teststatistik berechnen
4. p-Wert bestimmen (oder kritischen Wert nachschlagen)
5. Entscheidung: $p < \alpha \Rightarrow H_0$ ablehnen

Heute neu: Wir wenden dieses Verfahren auf **konkrete parametrische Tests** an:

- **Binomialtest** – exakter Test für Anteile
- **t-Tests** – Tests für Mittelwerte bei unbekanntem σ
- **Gepaarter Test** – vorher/nachher Vergleiche

Erinnerung

Fehler 1. Art (α): H_0 fälschlich ablehnen — Fehler 2. Art (β): H_0 fälschlich beibehalten

Das 5-Schritte-Schema bleibt identisch – nur die Teststatistik aendert sich je nach Test.

Binomialtest: Idee

Situation: Wir testen einen Anteilswert p anhand von n unabhängigen Beobachtungen mit Ja/Nein-Ergebnis.

Finanzbeispiel: Die historische Kreditausfallrate beträgt $p_0 = 0{,}02$ (2%). In diesem Jahr beobachten wir bei $n = 100$ Krediten $k = 5$ Ausfälle.

Frage: Hat sich die Ausfallrate signifikant erhöht?

Hypothesen:

$$H_0 : p = 0{,}02 \quad \text{(Ausfallrate ist unverändert)}$$
$$H_1 : p > 0{,}02 \quad \text{(Ausfallrate ist gestiegen)}$$

Teststatistik: Unter H_0 gilt $X \sim B(100;\ 0{,}02)$.
Wir berechnen den **exakten** p-Wert direkt aus der Binomialverteilung – keine Normalapproximation nötig!

Der Binomialtest ist ein exakter Test – er macht keine Approximation.

Binomialtest: Berechnung

Exakter p-Wert (rechtsseitiger Test):

$$p\text{-Wert} = P(X \geq 5 \mid p = 0{,}02) = \sum_{k=5}^{100} \binom{100}{k} \cdot 0{,}02^k \cdot 0{,}98^{100-k}$$

Berechnung:

$$\begin{aligned} P(X \leq 4) &= P(X{=}0) + P(X{=}1) + P(X{=}2) + P(X{=}3) + P(X{=}4) \\ &= 0{,}1326 + 0{,}2707 + 0{,}2734 + 0{,}1823 + 0{,}0902 \\ &= 0{,}9492 \end{aligned}$$

$$\Rightarrow p\text{-Wert} = 1 - 0{,}9492 = \boxed{0{,}0508}$$

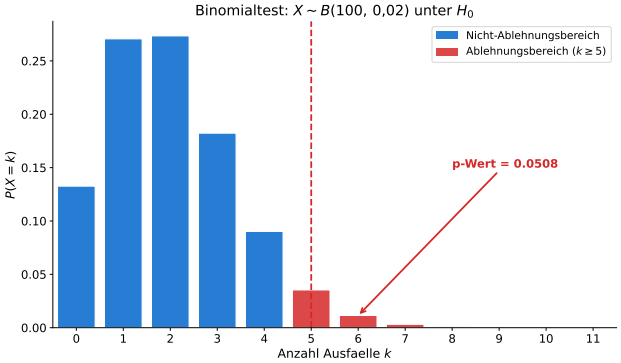
Entscheidung bei $\alpha = 0{,}05$:

$$p = 0{,}0508 > 0{,}05 = \alpha \quad \Longrightarrow \quad H_0 \text{ nicht ablehnen}$$

Interpretation: Trotz 5 statt erwarteter 2 Ausfälle reicht die Evidenz knapp nicht aus, um eine signifikant erhöhte Ausfallrate zu belegen.

Der exakte Test ist konservativ – bei grossen n kann man die Normalapproximation nutzen.

Binomialtest: Visualisierung



Beobachtungen:

- Die blauen Balken zeigen die Wahrscheinlichkeiten unter H_0
- Der rote Bereich markiert den Ablehnungsbereich ($k \geq 5$)
- Die beobachteten $k = 5$ Ausfälle liegen am Rand des Ablehnungsbereichs

Der Binomialtest nutzt die exakte Verteilung – daher "exakter Test".

Einstichproben-t-Test: Idee

Situation: Wir testen, ob der Populationsmittelwert μ einen bestimmten Wert μ_0 hat, wenn die Standardabweichung σ **unbekannt** ist.

Finanzbeispiel: Ist die mittlere Tagesrendite einer Aktie verschieden von 0?
Stichprobe: $n = 250$ Handelstage, $\bar{x} = 0{,}04\%$, $s = 1{,}2\%$.

Hypothesen:

$$H_0 : \mu = 0 \quad \text{(keine systematische Rendite)}$$
$$H_1 : \mu \neq 0 \quad \text{(systematische Rendite vorhanden)}$$

Teststatistik:

$$\boxed{t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}} \quad \sim \quad t_{n-1} \text{ unter } H_0$$

Unterschied zum z-Test: Wir schätzen σ durch die Stichproben-Standardabweichung s und verwenden die t-Verteilung statt der Normalverteilung.

Bei grossen n ($\gtrsim 30$) naehern sich t- und z-Verteilung einander an.

Einstichproben-t-Test: Berechnung

Daten: $n = 250$, $\bar{x} = 0{,}04\%$, $s = 1{,}2\%$, $\mu_0 = 0$.

Schritt 3 – Teststatistik:

$$t = \frac{0{,}04 - 0}{1{,}2/\sqrt{250}} = \frac{0{,}04}{0{,}0759} = 0{,}527$$

Schritt 4 – p-Wert: Zweiseitiger Test mit $\mathrm{df} = 249$:

$$p = 2 \cdot P(T > |0{,}527|) \approx 2 \cdot 0{,}2993 = 0{,}5986$$

Schritt 5 – Entscheidung: Bei $\alpha = 0{,}05$:

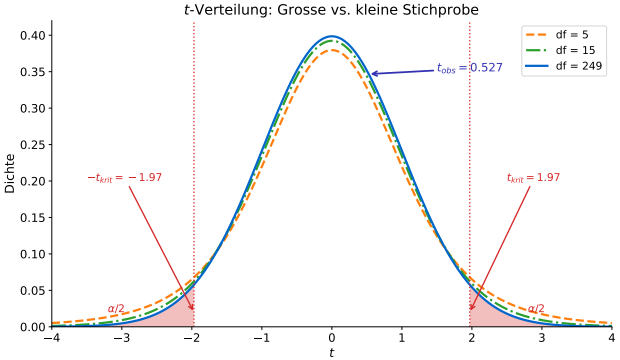
$$p = 0{,}599 \gg 0{,}05 \quad \Longrightarrow \quad H_0 \text{ nicht ablehnen}$$

Interpretation

Die beobachtete mittlere Tagesrendite von 0,04% ist **nicht** statistisch signifikant verschieden von 0. Die Schwankung ist mit zufälliger Variation vereinbar.

Bei $n = 250$ sind die kritischen t-Werte nahe bei $\pm 1{,}96$ (wie beim z-Test).

t-Test: Grosse vs. kleine Stichprobe



Wichtige Unterscheidung:

	Grosse Stichprobe ($n \geq 30$)	**Kleine Stichprobe** ($n < 30$)
Verteilung	$t \approx z$ (Normalvert.)	t_{n-1} (breitere Flanken)
Krit. Wert (α=0,05, 2s)	$\approx$ 1,96	z.B. 2,26 bei n=10
Voraussetzung	robuster	Normalverteilung nötig

Zweistichproben-t-Test: Idee

Situation: Zwei unabhängige Gruppen – sind ihre Mittelwerte verschieden?

Finanzbeispiel: Liefert Fonds A höhere Renditen als Fonds B?

- Fonds A: $n_1 = 30$ Monate, $\bar{x}_1 = 0{,}8\%$, $s_1 = 2{,}5\%$
- Fonds B: $n_2 = 35$ Monate, $\bar{x}_2 = 0{,}3\%$, $s_2 = 3{,}1\%$

Hypothesen:

$$H_0 : \mu_1 = \mu_2 \quad \text{(kein Renditeunterschied)}$$
$$H_1 : \mu_1 \neq \mu_2 \quad \text{(Renditen unterscheiden sich)}$$

Zwei Varianten:

- **Gepoolter t-Test:** $\sigma_1 = \sigma_2$ angenommen (Varianzhomogenität)
- **Welch-Test:** $\sigma_1 \neq \sigma_2$ erlaubt (Standard in der Praxis!)

Im Zweifel: Immer den Welch-Test verwenden – er ist robuster.

Zweistichproben-t-Test: Formeln

Gepoolter t-Test (bei gleichen Varianzen $\sigma_1^2 = \sigma_2^2$):

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Freiheitsgrade: $\mathrm{df} = n_1 + n_2 - 2$

Welch-Test (bei ungleichen Varianzen):

$$\boxed{t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

Freiheitsgrade (Welch-Satterthwaite):

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Die Welch-Formel korrigiert die Freiheitsgrade bei unterschiedlichen Varianzen.

Zweistichproben-t-Test: Berechnung

Daten: $n_1 = 30$, $\bar{x}_1 = 0{,}8\%$, $s_1 = 2{,}5\%$; $n_2 = 35$, $\bar{x}_2 = 0{,}3\%$, $s_2 = 3{,}1\%$.

Welch-Test:

$$t = \frac{0{,}8 - 0{,}3}{\sqrt{\frac{2{,}5^2}{30} + \frac{3{,}1^2}{35}}} = \frac{0{,}5}{\sqrt{0{,}2083 + 0{,}2746}} = \frac{0{,}5}{\sqrt{0{,}4829}} = \frac{0{,}5}{0{,}695} = 0{,}719$$

Freiheitsgrade: $\mathrm{df} \approx 62{,}7 \approx 62$

p-Wert: $p = 2 \cdot P(T > 0{,}719) \approx 0{,}475$

Entscheidung:

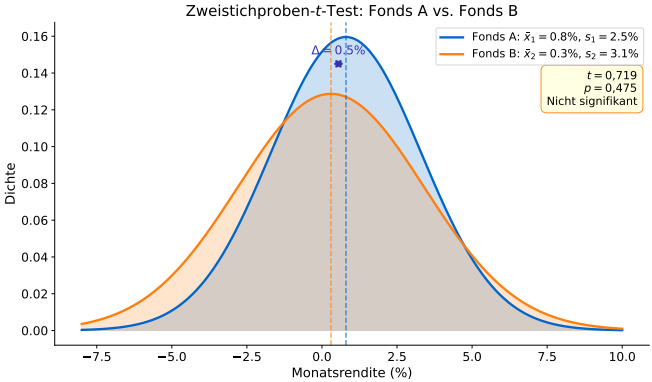
$$p = 0{,}475 \gg 0{,}05 \quad \Longrightarrow \quad H_0 \text{ nicht ablehnen}$$

Interpretation

Der beobachtete Renditeunterschied von 0,5 Prozentpunkten ist **nicht** statistisch signifikant. Es gibt keine genügende Evidenz, dass sich die Fonds in ihrer mittleren Rendite unterscheiden.

Der Welch-Test ist der Standardtest in R (`t.test` nutzt ihn als Default).

Zweistichproben-t-Test: Visualisierung



Beobachtungen:

- Die beiden Verteilungen überlappen sich stark
- Der Unterschied der Mittelwerte (0,5%) ist klein relativ zur Streuung
- Bei grösserer Stichprobe oder geringerer Streuung könnte das Ergebnis signifikant sein

Grafische Darstellung: Starke Ueberlappung deutet auf nicht-signifikanten Unterschied hin.

Paardifferenzentest: Idee

Situation: Messungen an denselben Einheiten zu zwei Zeitpunkten (vorher/nachher).

Finanzbeispiel: Hat ein Risikomanagement-Training den maximalen Drawdown der Trader verbessert?

- $n = 20$ Trader, gemessen **vor** und **nach** dem Training
- Vorher-Mittel: $\bar{x}_{\text{vor}} = -8{,}2\%$, Nachher-Mittel: $\bar{x}_{\text{nach}} = -6{,}5\%$

Trick: Bilde Differenzen $d_i = x_{i,\text{nach}} - x_{i,\text{vor}}$ für jeden Trader.

$\Rightarrow$ Einstichproben-t-Test auf die Differenzen!

Hypothesen:

$$H_0 : \mu_d = 0 \quad \text{(kein Trainingseffekt)}$$
$$H_1 : \mu_d > 0 \quad \text{(Training verbessert Drawdown, d.h. } d_i > 0\text{)}$$

Gepaarte Daten: Derselbe Trader vor und nach dem Training – nicht zwei unabhaengige Gruppen!

Paardifferenzentest: Berechnung

Differenzen: $d_i = x_{i,\text{nach}} - x_{i,\text{vor}}$

Beobachtet: $\bar{d} = 1{,}7\%$, $s_d = 3{,}2\%$, $n = 20$.

Teststatistik:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{1{,}7}{3{,}2/\sqrt{20}} = \frac{1{,}7}{0{,}716} = 2{,}375$$

p-Wert: Rechtsseitig mit $\mathrm{df} = 19$:

$$p = P(T > 2{,}375) \approx 0{,}014$$

Entscheidung: Bei $\alpha = 0{,}05$:

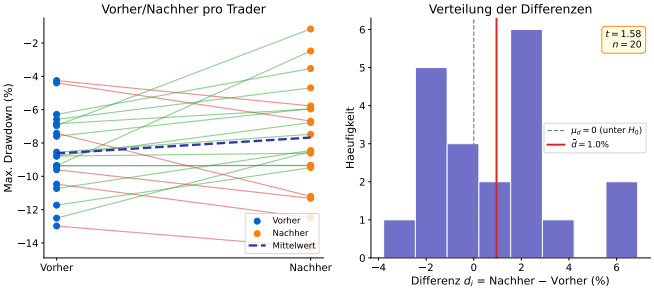
$$p = 0{,}014 < 0{,}05 \quad \Longrightarrow \quad H_0 \text{ ablehnen}$$

Interpretation

Das Risikomanagement-Training hat den maximalen Drawdown der Trader **statistisch signifikant** verbessert (um durchschnittlich 1,7 Prozentpunkte, $p = 0{,}014$).

Der gepaarte Test ist staerker als der Zweistichprobentest, weil Trader-spezifische Varianz eliminiert wird.

Paardifferenzentest: Visualisierung

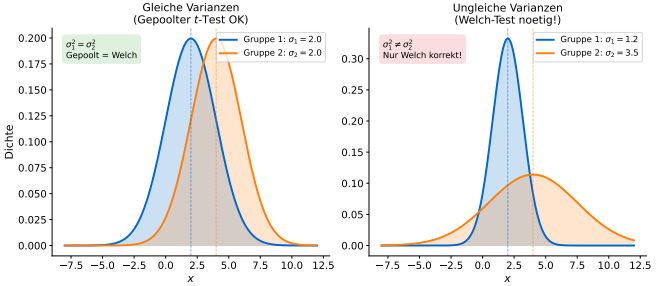


Warum gepaart testen?

- Eliminiert die **zwischen-Trader-Variabilität**
- Jeder Trader ist seine eigene Kontrolle
- Geringere Streuung $\Rightarrow$ höhere Power

Gepaarte Designs sind effizienter: Weniger Daten noetig fuer gleiche statistische Power.

Welch vs. gepoolter t-Test



	Gepoolter t-Test	**Welch-Test**
Annahme	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$ erlaubt
Freiheitsgrade	$n_1 + n_2 - 2$	Welch-Satterthwaite (kleiner)
Power bei $\sigma_1 = \sigma_2$	etwas höher	fast gleich
Robustheit	anfällig	**robuster**

Empfehlung: Im Zweifel **immer den Welch-Test** verwenden.

R nutzt den Welch-Test als Standard (`t.test(..., var.equal=FALSE)`).

Cohen's d: Effektgrösse

Problem: Der p-Wert sagt nichts über die **praktische Bedeutsamkeit** eines Effekts.

Lösung: Cohen's d standardisiert den Unterschied:

Einstichproben-Fall:

$$\boxed{d = \frac{\bar{x} - \mu_0}{s}}$$

Zweistichproben-Fall:

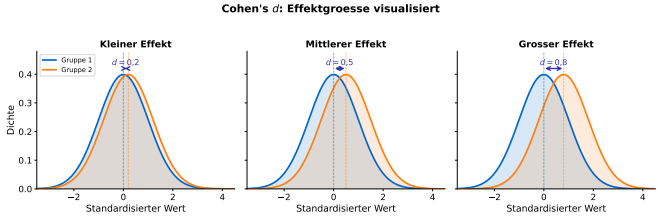
$$\boxed{d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}}$$

Interpretation (Faustregel nach Cohen):

$\lvert d \rvert$	Bewertung	Finanzbeispiel
$\approx 0{,}2$	**kleiner** Effekt	Leichte Renditeänderung
$\approx 0{,}5$	**mittlerer** Effekt	Spürbare Strategieverbesserung
$\approx 0{,}8$	**grosser** Effekt	Fundamentale Marktveränderung

Immer berichten: p-Wert + Konfidenzintervall + Effektgroesse = vollstaendige Analyse.

Effektgrösse: Visualisierung



Rechenbeispiel (Paardifferenzentest):

$$d = \frac{\bar{d}}{s_d} = \frac{1{,}7}{3{,}2} = 0{,}53$$

$\Rightarrow$ **mittlerer Effekt** – das Training hat eine spürbare Wirkung.

Merke: Ein statistisch signifikantes Ergebnis ($p < 0{,}05$) mit kleinem d ($< 0{,}2$) deutet auf einen Effekt hin, der praktisch kaum relevant ist.

Effektgroesse ist stichprobenunabhaengig – im Gegensatz zum p-Wert.

R-Code: Binomialtest

```r
# Binomialtest: Ausfallrate erhoeht?
# H0: p = 0.02, H1: p > 0.02
binom.test(x = 5, n = 100, p = 0.02,
           alternative = "greater")

# Ergebnis:
# number of successes = 5, number of trials = 100
# p-value = 0.0508
# alternative hypothesis: true probability > 0.02
# 95% confidence interval: 0.01819 to 1.00000
```

Interpretation:

- `x = 5`: Anzahl beobachteter Erfolge (Ausfälle)
- `p = 0.02`: Anteil unter H_0
- `alternative = "greater"`: Rechtsseitiger Test
- `p-value = 0.0508`: Knapp nicht signifikant bei $\alpha = 0{,}05$

`binom.test()` **berechnet den exakten p-Wert – keine Normalapproximation.**

R-Code: *t*-Tests

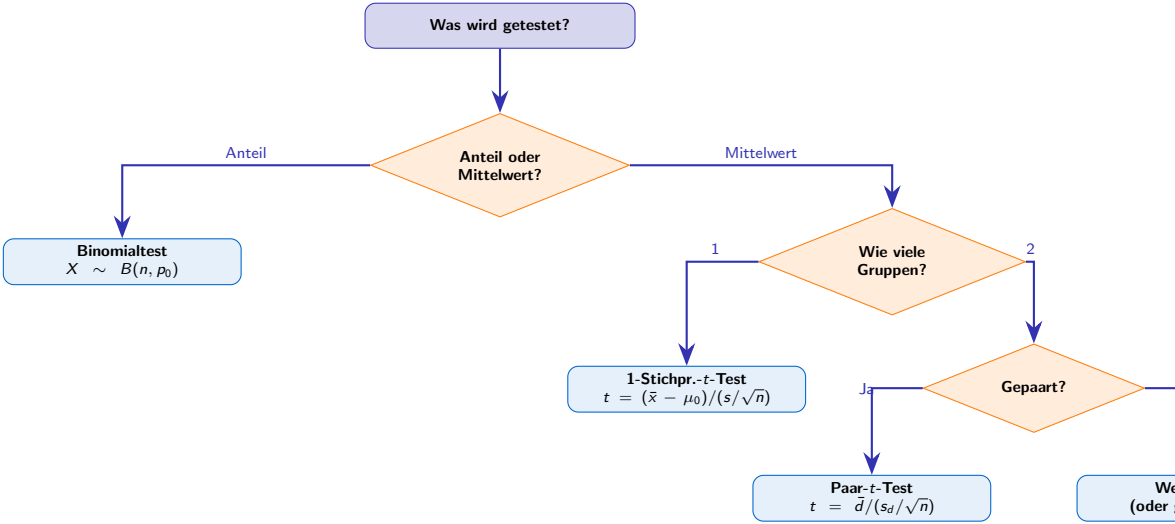
```r
# Einstichproben-t-Test
t.test(renditen, mu = 0)

# Zweistichproben-t-Test (Welch, Standard)
t.test(fonds_a, fonds_b)

# Gepoolter t-Test (gleiche Varianzen)
t.test(fonds_a, fonds_b, var.equal = TRUE)

# Gepaarter t-Test
t.test(nachher, vorher, paired = TRUE,
       alternative = "greater")
```

Wichtige Argumente:

- `mu`: Hypothetischer Mittelwert (Default: 0)
- `var.equal`: `TRUE` = gepoolter Test, `FALSE` = Welch (Default)
- `paired`: `TRUE` = gepaarter Test
- `alternative`: `"two.sided"`, `"greater"`, `"less"`

R liefert automatisch: Teststatistik, df, p-Wert und Konfidenzintervall.

Entscheidungsschema: Welcher Test?



Dieses Schema deckt die vier wichtigsten parametrischen Tests ab.

Übersicht: Parametrische Tests

Test	Stichproben	Parameter	Teststatistik	R-Befehl
Binomialtest	1	Anteil p	exakt (Binomial)	`binom.test()`
1-Stichpr.-t	1	Mittelwert μ	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	`t.test(x, mu=)`
2-Stichpr.-t (Welch)	2 (unabh.)	$\mu_1 - \mu_2$	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	`t.test(x, y)`
Gepaarter t	2 (gepaart)	μ_d	$t = \frac{\bar{d}}{s_d/\sqrt{n}}$	`t.test(paired=T)`

Voraussetzungen:

- t-Tests: Normalverteilung der Daten (bei $n \geq 30$ durch ZGS robust)
- Binomialtest: Unabhängige Bernoulli-Versuche
- Welch-Test: Keine Varianzhomogenität nötig

Die Uebersichtstabelle ist ideal als Formelsammlung fuer die Pruefung.

Lernziele – Checkliste

Prüfen Sie sich selbst:

- ✓ **Binomialtest:** Exakter Test für Anteile, p-Wert aus Binomialverteilung
- ✓ **1-Stichproben-t:** $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, $\mathrm{df} = n - 1$
- ✓ **2-Stichproben-t:** Vergleich zweier unabhängiger Mittelwerte
- ✓ **Welch vs. gepoolt:** Welch ist robuster, Standard in R
- ✓ **Gepaarter t-Test:** Differenzen bilden, dann Einstichprobentest
- ✓ **Effektgrösse:** Cohen's d: klein (0,2), mittel (0,5), gross (0,8)
- ✓ **Testauswahl:** Entscheidungsschema: Anteil/Mittelwert, 1/2 Gruppen, gepaart?
- ✓ **R-Code:** `binom.test()`, `t.test()` mit `paired`, `var.equal`, `alternative`

Wenn Sie alle Punkte verstanden haben, sind Sie bereit fuer die naechsten Lektionen!

Schlüsselformeln

Formel	Name
$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	Einstichproben-t-Teststatistik
$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	Welch-Teststatistik
$t = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	Gepaarter t-Test
$d = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p}$	Cohen's d (Effektgrösse)
p-Wert $= P(X \geq k \mid H_0)$	Exakter Binomialtest

Diese Formeln sind das Fundament der parametrischen Hypothesentests.