Interpretation von Testergebnissen I

- Durch die Asymmetrie in den Fehlerwahrscheinlichkeiten 1. und 2. Art ist Vorsicht bei Interpretation von Testergebnissen geboten!
- Es besteht ein großer Unterschied zwischen dem Aussagegehalt einer **Ablehnung** von H_0 und dem Aussagegehalt einer **Annahme** von H_0:
  - Fällt die Testentscheidung **gegen** H_0 aus, so hat man — sollte H_0 tatsächlich **erfüllt** sein — wegen der Beschränkung der Fehlerwahrscheinlichkeit 1. Art durch das Signifikanzniveau α nur mit einer typischerweise geringen Wahrscheinlichkeit $\leq \alpha$ eine Stichprobenrealisation erhalten, die **fälschlicherweise** zur **Ablehnung von** H_0 geführt hat.
    Aber: Vorsicht vor „Über"interpretation als Evidenz für Gültigkeit von H_1: Aussagen der Form *„Wenn H_0 abgelehnt wird, dann gilt H_1 mit Wahrscheinlichkeit von mindestens* $1-\alpha$*"* sind unsinnig!
  - Fällt die Testentscheidung jedoch **für** H_0 aus, so ist dies ein vergleichsweise meist schwächeres „Indiz" für die Gültigkeit von H_0, da die Fehlerwahrscheinlichkeit 2. Art nicht kontrolliert ist und typischerweise große Werte (bis $1-\alpha$) annehmen kann. Gilt also tatsächlich H_1, ist es dennoch mit einer sehr großen Wahrscheinlichkeit möglich, eine Stichprobenrealisation zu erhalten, die **fälschlicherweise nicht** zur **Ablehnung von** H_0 führt.

Aus diesem Grund sagt man auch häufig statt „H_0 wird angenommen" eher „H_0 kann nicht verworfen werden".

Interpretation von Testergebnissen II

- Die Ablehnung von H_0 als Ergebnis eines statistischen Tests wird häufig als
  - **signifikante Veränderung** (zweiseitiger Test),
  - **signifikante Verringerung** (linksseitiger Test) oder
  - **signifikante Erhöhung** (rechtsseitiger Test)

  einer Größe bezeichnet. Konstruktionsbedingt kann das Ergebnis einer statistischen Untersuchung — auch im Fall einer Ablehnung von H_0 — aber **niemals** als zweifelsfreier Beweis für die Veränderung/Verringerung/Erhöhung einer Größe dienen!
- Weiteres Problem: Aussagen über die Fehlerwahrscheinlichkeiten 1. und 2. Art gelten nur perfekt, wenn alle Voraussetzungen erfüllt sind, also wenn
  - Verteilungsannahmen erfüllt sind (Vorsicht bei „approximativen" Tests) und
  - tatsächlich eine **einfache Stichprobe** vorliegt!
- Vorsicht vor „Publication Bias":
  - Bei einem Signifikanzniveau von $\alpha = 0.05$ resultiert im Mittel 1 von 20 statistischen Untersuchungen, bei denen H_0 wahr ist, konstruktionsbedingt in einer Ablehnung von H_0.
  - Gefahr von Fehlinterpretationen, wenn die Untersuchungen, bei denen H_0 nicht verworfen wurde, verschwiegen bzw. nicht publiziert werden!

Interpretation von Testergebnissen III

„signifikant" vs. „deutlich"

- Ein „signifikanter" Unterschied ist noch lange kein „deutlicher" Unterschied!
- Problem: „Fluch des großen Stichprobenumfangs"
- Beispiel: Abfüllmaschine soll Flaschen mit 1000 ml Inhalt abfüllen.
  - Abfüllmenge schwankt zufällig, Verteilung sei Normalverteilung mit bekannter Standardabweichung $\sigma = 0.5$ ml, d.h. in ca. 95% der Fälle liegt Abfüllmenge im Bereich ± 1 ml um den (tatsächlichen) Mittelwert.
  - Statistischer Test zum Niveau $\alpha = 0.05$ zur Überprüfung, ob mittlere Abfüllmenge (Erwartungswert) von 1000 ml abweicht.
- Tatsächlicher Mittelwert sei 1000.1 ml, Test auf Grundlage von 500 Flaschen.
- Wahrscheinlichkeit, die Abweichung von 0.1 ml zu erkennen (Berechnung mit Gütefunktion, siehe Folie 103): 99.4%
- Systematische Abweichung der Abfüllmenge von 0.1 ml zwar mit hoher Wahrscheinlichkeit (99.4%) signifikant, im Vergleich zur (ohnehin vorhandenen) zufälligen Schwankung mit $\sigma = 0.5$ ml aber keinesfalls deutlich!

Fazit: „Durch wissenschaftliche Studien belegte signifikante Verbesserungen" können vernachlässigbar klein sein ($\rightsquigarrow$ Werbung...)

Der p-Wert

- Hypothesentests „komprimieren" Stichprobeninformation zur Entscheidung zwischen H_0 und H_1 zu einem vorgegebenen Signifikanzniveau α.
- Testentscheidung hängt von α **ausschließlich** über kritischen Bereich K ab!
- Genauere Betrachtung offenbart: Abhängigkeit zwischen α und K ist **monoton** im Sinne der Teilmengenbeziehung.
  - Gilt $\widetilde{\alpha} < \alpha$ und bezeichnen $K_{\widetilde{\alpha}}$ und K_α die zugehörigen kritischen Bereiche, so gilt für alle bisher betrachteten Gauß-Tests $K_{\widetilde{\alpha}} \subsetneq K_\alpha$.
  - Unmittelbare Folge ist, dass Ablehnung von H_0 zum Signifikanzniveau $\widetilde{\alpha}$ mit $\widetilde{\alpha} < \alpha$ automatisch eine Ablehnung von H_0 zum Niveau α zur Folge hat (auf Basis derselben Stichprobeninformation)!
  - Außerdem wird K_α für $\alpha \to 0$ beliebig klein und für $\alpha \to 1$ beliebig groß, so dass man für jede Realisation T der Teststatistik sowohl Signifikanzniveaus α mit $T \in K_\alpha$ wählen kann, als auch solche mit $T \notin K_\alpha$.
- Zusammenfassend kann man also zu jeder Realisation T der Teststatistik das kleinste Signifikanzniveau α mit $T \in K_\alpha$ bestimmen (bzw. das größte Signifikanzniveau α mit $T \notin K_\alpha$). Dieses Signifikanzniveau heißt ***p*-Wert** oder **empirisches (marginales) Signifikanzniveau**.
- Mit der Information des p-Werts kann der Test also für **jedes beliebige Signifikanzniveau** α entschieden werden!

p-Wert bei Gauß-Tests

auf den Mittelwert bei bekannter Varianz

- Der Wechsel zwischen „$N \in K_\alpha$“ und „$N \notin K_\alpha$“ findet bei den diskutierten Gauß-Tests offensichtlich dort statt, wo die realisierte Teststatistik N gerade mit (einer) der Grenze(n) des kritischen Bereichs übereinstimmt, d.h.
  - bei rechtsseitigen Tests mit $K_\alpha = (N_{1-\alpha}, \infty)$ für $N = N_{1-\alpha}$,
  - bei linksseitigen Tests mit $K_\alpha = (-\infty, -N_{1-\alpha})$ für $N = -N_{1-\alpha}$,
  - bei zweiseitigen Tests mit $K_\alpha = (-\infty, -N_{1-\frac{\alpha}{2}}) \cup (N_{1-\frac{\alpha}{2}}, \infty)$ für

$$N = \begin{cases} -N_{1-\frac{\alpha}{2}} & \text{falls } N < 0 \\ N_{1-\frac{\alpha}{2}} & \text{falls } N \geq 0 \end{cases} .$$

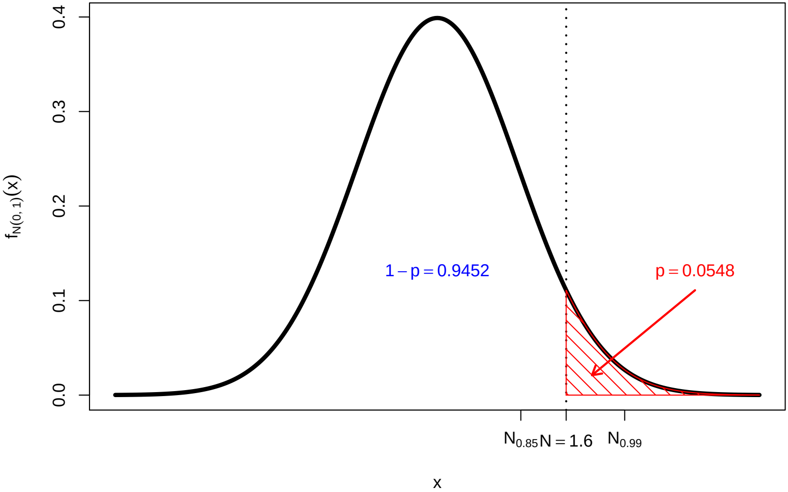
- Durch Auflösen nach α erhält man
  - für rechtsseitige Tests den p-Wert $1 - \Phi(N)$,
  - für linksseitige Tests den p-Wert $\Phi(N)$,
  - für zweiseitige Tests den p-Wert

$$\left.\begin{array}{ll} 2 \cdot \Phi(N) = 2 \cdot (1 - \Phi(-N)) & \text{falls } N < 0 \\ 2 \cdot (1 - \Phi(N)) & \text{falls } N \geq 0 \end{array}\right\} = 2 \cdot (1 - \Phi(|N|))$$

  sowie die alternative Darstellung $2 \cdot \min\{\Phi(N), 1 - \Phi(N)\}$.

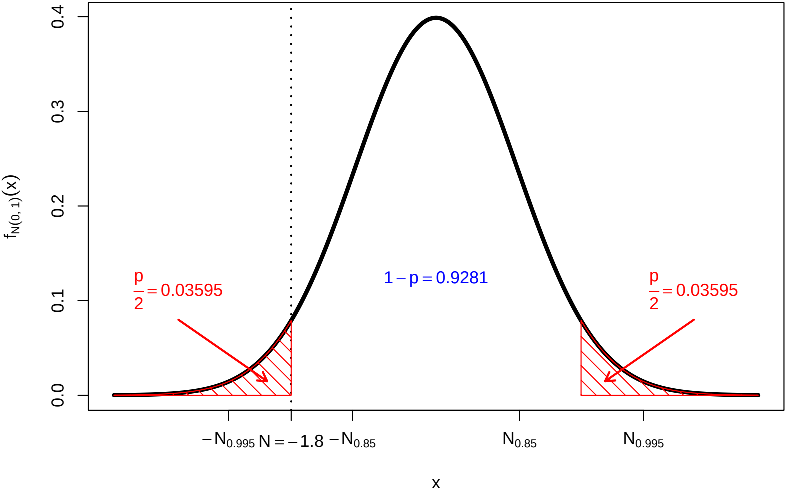
Beispiel: p-Werte bei rechtsseitigem Gauß-Test (Grafik)

Realisierte Teststatistik $N = 1.6$, p-Wert: 0.0548



Beispiel: p-Werte bei zweiseitigem Gauß-Test (Grafik)

Realisierte Teststatistik $N = -1.8$, p-Wert: 0.0719



Entscheidung mit p-Wert

- Offensichtlich erhält man auf der Grundlage des p-Werts p zur beobachteten Stichprobenrealisation die einfache Entscheidungsregel

$$H_0 \text{ ablehnen} \qquad \Leftrightarrow \qquad p < \alpha$$

  für Hypothesentests zum Signifikanzniveau α.
- Sehr niedrige p-Werte bedeuten also, dass man beim zugehörigen Hypothesentest H_0 auch dann ablehnen würde, wenn man die maximale Fehlerwahrscheinlichkeit 1. Art sehr klein wählen würde.
- Kleinere p-Werte liefern also stärkere Indizien für die Gültigkeit von H_1 als größere, **aber** (wieder) Vorsicht vor Überinterpretation: Aussagen der Art *„Der p-Wert gibt die Wahrscheinlichkeit für die Gültigkeit von H_0 an"* sind unsinnig!

Warnung!

Bei der Entscheidung von statistischen Tests mit Hilfe des p-Werts ist es **unbedingt** erforderlich, das Signifikanzniveau α **vor** Berechnung des p-Werts festzulegen, um nicht der Versuchung zu erliegen, α im Nachhinein so zu wählen, dass man die „bevorzugte" Testentscheidung erhält!

Tests und Konfidenzintervalle

- Enger Zusammenhang zwischen zweiseitigem Gauß-Test und (symmetrischen) Konfidenzintervallen für den Erwartungswert bei bekannter Varianz.
- Für Konfidenzintervalle zur Vertrauenswahrscheinlichkeit $1-\alpha$ gilt:

$$
\begin{aligned}
& \widetilde{\mu} \in \left[\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot N_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot N_{1-\frac{\alpha}{2}}\right] \\
\Leftrightarrow \quad & \widetilde{\mu} - \overline{X} \in \left[-\frac{\sigma}{\sqrt{n}} \cdot N_{1-\frac{\alpha}{2}}, \frac{\sigma}{\sqrt{n}} \cdot N_{1-\frac{\alpha}{2}}\right] \\
\Leftrightarrow \quad & \frac{\widetilde{\mu} - \overline{X}}{\sigma}\sqrt{n} \in \left[-N_{1-\frac{\alpha}{2}}, N_{1-\frac{\alpha}{2}}\right] \\
\Leftrightarrow \quad & \frac{\overline{X} - \widetilde{\mu}}{\sigma}\sqrt{n} \in \left[-N_{1-\frac{\alpha}{2}}, N_{1-\frac{\alpha}{2}}\right]
\end{aligned}
$$

- Damit ist $\widetilde{\mu}$ also **genau dann** im Konfidenzintervall zur Sicherheitswahrscheinlichkeit $1-\alpha$ enthalten, **wenn** ein zweiseitiger Gauß-Test zum Signifikanzniveau α die Nullhypothese $H_0 : \mu = \widetilde{\mu}$ **nicht** verwerfen würde.
- Vergleichbarer Zusammenhang auch in anderen Situationen.

Zusammenfassung: Gauß-Test für den Mittelwert

bei bekannter Varianz

Anwendungs-voraussetzungen	exakt: $Y \sim N(\mu, \sigma^2)$ mit $\mu \in \mathbb{R}$ unbekannt, σ^2 bekannt approximativ: $E(Y) = \mu \in \mathbb{R}$ unbekannt, $\text{Var}(Y) = \sigma^2$ bekannt $X_1, \ldots, X_n$ einfache Stichprobe zu Y		
Nullhypothese Gegenhypothese	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$
Teststatistik	$N = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$		
Verteilung (H_0)	N für $\mu = \mu_0$ (näherungsweise) $N(0,1)$-verteilt		
Benötigte Größen	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$		
Kritischer Bereich zum Niveau α	$(-\infty, -N_{1-\frac{\alpha}{2}}) \cup (N_{1-\frac{\alpha}{2}}, \infty)$	$(N_{1-\alpha}, \infty)$	$(-\infty, -N_{1-\alpha})$
p-Wert	$2 \cdot (1 - \Phi(\|N\|))$	$1 - \Phi(N)$	$\Phi(N)$

Approximativer Gauß-Test für Anteilswert p

- Wichtiger Spezialfall des (approximativen) Gauß-Tests für den Mittelwert einer Zufallsvariablen mit bekannter Varianz:

  Approximativer Gauß-Test für den Anteilswert p einer alternativverteilten Zufallsvariablen

- *Erinnerung:* Für alternativverteilte Zufallsvariablen $Y \sim B(1, p)$ war Konfidenzintervall für Anteilswert p ein Spezialfall für Konfidenzintervalle für Mittelwerte von Zufallsvariablen mit **unbekannter** Varianz.
- **Aber:** Bei der Konstruktion von Tests für $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$ für ein vorgegebenes p_0 (sowie den einseitigen Varianten) spielt Verteilung der Teststatistik unter H_0, insbesondere für $p = p_0$, entscheidende Rolle.
- Da Varianz für $p = p_0$ bekannt $\rightsquigarrow$ approximativer Gauß-Test geeignet. Für $p = p_0$ gilt genauer $\text{Var}(Y) = \text{Var}(X_i) = p_0 \cdot (1 - p_0)$ und damit

$$\text{Var}(\widehat{p}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot n \cdot \text{Var}(Y) = \frac{p_0 \cdot (1 - p_0)}{n} \ .$$

Als Testgröße erhält man also: $N = \dfrac{\widehat{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)}}\sqrt{n}$

Zusammenfassung: (Approx.) Gauß-Test für Anteilswert p

Anwendungs-voraussetzungen	approximativ: $Y \sim B(1, p)$ mit $p \in [0, 1]$ unbekannt $X_1, \dots, X_n$ einfache Stichprobe zu Y		
Nullhypothese Gegenhypothese	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p \leq p_0$ $H_1 : p > p_0$	$H_0 : p \geq p_0$ $H_1 : p < p_0$
Teststatistik	$N = \frac{\widehat{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)}} \sqrt{n}$		
Verteilung (H_0)	N für $p = p_0$ näherungsweise $N(0, 1)$-verteilt		
Benötigte Größen	$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$		
Kritischer Bereich zum Niveau α	$(-\infty, -N_{1-\frac{\alpha}{2}}) \cup (N_{1-\frac{\alpha}{2}}, \infty)$	$(N_{1-\alpha}, \infty)$	$(-\infty, -N_{1-\alpha})$
p-Wert	$2 \cdot (1 - \Phi(\|N\|))$	$1 - \Phi(N)$	$\Phi(N)$

Beispiel: Bekanntheitsgrad eines Produkts

- Untersuchungsgegenstand: Hat sich der Bekanntheitsgrad eines Produkts gegenüber bisherigem Bekanntheitsgrad von 80% reduziert, nachdem die Ausgaben für Werbemaßnahmen vor einiger Zeit drastisch gekürzt wurden?
- Annahmen: Kenntnis des Produkts wird durch $Y \sim B(1, p)$ beschrieben, wobei p als Bekanntheitsgrad des Produkts aufgefasst werden kann.
- Stichprobeninformation aus Realisation einfacher Stichprobe **(!)** zu Y: Unter $n = 500$ befragten Personen kannten 381 das Produkt $\rightsquigarrow \widehat{p} = 0.762$.
- Gewünschtes Signifikanzniveau (max. Fehlerwahrscheinlichkeit 1. Art): $\alpha = 0.05$

Geeigneter Test: **(Approx.) linksseitiger Gauß-Test für den Anteilswert** p

1. Hypothesen: $H_0 : p \geq p_0 = 0.8$ gegen $H_1 : p < p_0 = 0.8$
2. Teststatistik: $N = \frac{\widehat{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)}} \sqrt{n} \overset{\bullet}{\sim} N(0, 1)$, falls H_0 gilt ($p = p_0$)
3. Kritischer Bereich zum Niveau $\alpha = 0.05$: $K = (-\infty, -N_{0.95}) = (-\infty, -1.645)$
4. Realisierter Wert der Teststatistik: $N = \frac{0.762 - 0.8}{\sqrt{0.8 \cdot (1 - 0.8)}} \sqrt{500} = -2.124$
5. Entscheidung: $N \in K \rightsquigarrow H_0$ wird abgelehnt, der Bekanntheitsgrad des Produkts hat sich signifikant reduziert.

t-Test für den Mittelwert

bei unbekannter Varianz

- Konstruktion des (exakten) Gauß-Tests für den Mittelwert bei bekannter Varianz durch Verteilungsaussage

$$N := \frac{\overline{X} - \mu}{\sigma} \sqrt{n} \sim N(0,1) ,$$

falls $X_1, \ldots, X_n$ einfache Stichprobe zu normalverteilter ZV Y.

- Analog zur Konstruktion von Konfidenzintervallen für den Mittelwert bei unbekannter Varianz: Verwendung der Verteilungsaussage

$$t := \frac{\overline{X} - \mu}{S} \sqrt{n} \sim t(n-1) \qquad \text{mit} \qquad S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2} ,$$

falls $X_1, \ldots, X_n$ einfache Stichprobe zu normalverteilter ZV Y, um geeigneten Hypothesentest für den Mittelwert μ zu entwickeln.

- Test lässt sich genauso wie Gauß-Test herleiten, lediglich
  - Verwendung von S statt σ,
  - Verwendung von $t(n-1)$ statt $N(0,1)$.

- Beziehung zwischen symmetrischen Konfidenzintervallen und zweiseitigen Tests bleibt wie beim Gauß-Test erhalten.
- Wegen Symmetrie der $t(n-1)$-Verteilung bleiben auch alle entsprechenden „Vereinfachungen" bei der Bestimmung von kritischen Bereichen und p-Werten gültig.
- p-Werte können mit Hilfe der Verteilungsfunktion der $t(n-1)$-Verteilung bestimmt werden (unproblematisch mit Statistik-Software).
- Zur Berechnung der Gütefunktion: Verteilungsfunktion der „nichtzentralen" $t(n-1)$-Verteilung benötigt (unproblematisch mit Statistik-Software).
- Zur Berechnung von p-Werten und Gütefunktionswerten für große n: Näherung der $t(n-1)$-Verteilung durch Standardnormalverteilung bzw. der nichtzentralen $t(n-1)$-Verteilung durch Normalverteilung mit Varianz 1 (vgl. Gauß-Test) möglich.
- Analog zu Konfidenzintervallen:
  Ist Y nicht normalverteilt, kann der t-Test auf den Mittelwert bei unbekannter Varianz immer noch als approximativer (näherungsweiser) Test verwendet werden.

Zusammenfassung: t-Test für den Mittelwert

bei unbekannter Varianz

Anwendungs-voraussetzungen	exakt: $Y \sim N(\mu, \sigma^2)$ mit $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{++}$ unbekannt approximativ: $E(Y) = \mu \in \mathbb{R}, \text{Var}(Y) = \sigma^2 \in \mathbb{R}_{++}$ unbekannt $X_1, \ldots, X_n$ einfache Stichprobe zu Y		
Nullhypothese Gegenhypothese	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$
Teststatistik	$t = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$		
Verteilung (H_0)	t für $\mu = \mu_0$ (näherungsweise) $t(n-1)$-verteilt		
Benötigte Größen	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)}$		
Kritischer Bereich zum Niveau α	$(-\infty, -t_{n-1;1-\frac{\alpha}{2}}) \cup (t_{n-1;1-\frac{\alpha}{2}}, \infty)$	$(t_{n-1;1-\alpha}, \infty)$	$(-\infty, -t_{n-1;1-\alpha})$
p-Wert	$2 \cdot (1 - F_{t(n-1)}(\lvert t \rvert))$	$1 - F_{t(n-1)}(t)$	$F_{t(n-1)}(t)$

Beispiel: Durchschnittliche Wohnfläche

- Untersuchungsgegenstand: Hat sich die durchschnittliche Wohnfläche pro Haushalt in einer bestimmten Stadt gegenüber dem aus dem Jahr 1998 stammenden Wert von 71.2 (in $[m^2]$) **erhöht**?
- Annahmen: Verteilung der Wohnfläche Y im Jahr 2009 unbekannt.
- Stichprobeninformation: Realisation einer einfachen Stichprobe vom Umfang $n = 400$ zu Y liefert Stichprobenmittel $\overline{x} = 73.452$ und Stichprobenstandardabweichung $s = 24.239$.
- Gewünschtes Signifikanzniveau (max. Fehlerwahrscheinlichkeit 1. Art): $\alpha = 0.05$

Geeigneter Test:
Rechtsseitiger approx. t-Test für den Mittelwert bei unbekannter Varianz

1. Hypothesen: $H_0 : \mu \leq \mu_0 = 71.2 \quad$ gegen $\quad H_1 : \mu > \mu_0 = 71.2$
2. Teststatistik: $t = \frac{\overline{X}-\mu_0}{S}\sqrt{n} \overset{\bullet}{\sim} t(399)$, falls H_0 gilt $(\mu = \mu_0)$
3. Kritischer Bereich zum Niveau $\alpha = 0.05$: $K = (t_{399;0.95}, \infty) = (1.649, \infty)$
4. Realisierter Wert der Teststatistik: $t = \frac{73.452-71.2}{24.239}\sqrt{400} = 1.858$
5. Entscheidung: $t \in K \rightsquigarrow H_0$ wird abgelehnt; Test kommt zur Entscheidung, dass sich durchschnittliche Wohnfläche gegenüber 1998 erhöht hat.

Beispiel: p-Wert bei rechtsseitigem t-Test (Grafik)

Wohnflächenbeispiel, realisierte Teststatistik $t = 1.858$, p-Wert: 0.032