Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Hypothesenbewertung

Christoph Sawade/Niels Landwehr
Tobias Scheffer

Überblick

- Wiederholung: Hypothesenbewertung
  - Verfahren
  - Anwendungen
  - Konfidenzintervalle
- ROC-Analyse
- Statistische Tests
  - p-Wert
  - Vorzeichen-, Wald-, t- und Pearsons –Test

Überblick

- Wiederholung: Hypothesenbewertung
  - Verfahren
  - Anwendungen
  - Konfidenzintervalle
- ROC-Analyse
- Statistische Tests
  - p-Wert
  - Vorzeichen-, Wald-, t- und Pearsons-Test

Hypothesenbewertung

- Klassifikation, Regression: Lernproblem
  - Eingabe: Trainingsdaten $L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m) \rangle$
  - Ausgabe: Hypothese (Modell) $f : X \to Y$

$f(\mathbf{x}) = ? \in Y \qquad \mathbf{x} \in X$ Testbeispiel

- Ziel des Lernens: genaue Vorhersagen treffen

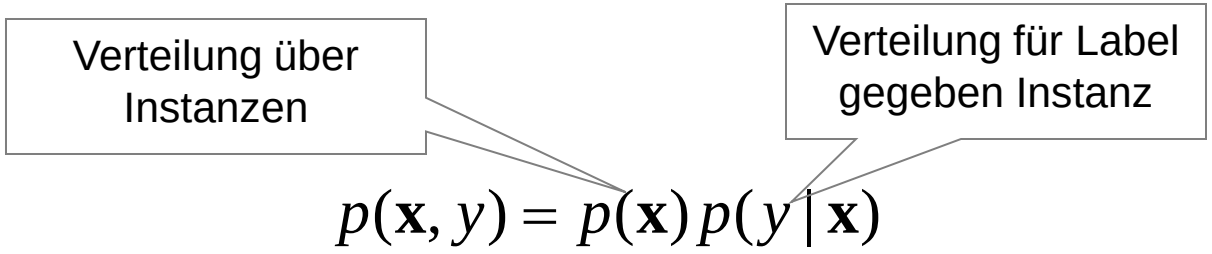
- „Hypothesenbewertung“: Abschätzung der Genauigkeit von Vorhersagen
  - Schätzproblem: was ist eine gute Schätzung des erwarteten Fehlers?

Verlustfunktionen

- Instanz $(\mathbf{x}, y)$, Hypothese sagt $f(\mathbf{x})$.
- Verlustfunktion definiert, wie schlecht das ist.
  - $\ell(y, f(\mathbf{x}))$ Verlust der Vorhersage $f(\mathbf{x})$ auf Instanz $(\mathbf{x}, y)$
  - Nicht-negativ: $\forall y, y' : \ell(y, y') \geq 0$
  - Problem-spezifisch, gegeben.
- Verlustfunktionen für Klassifikation
  - Zero-one loss: $\ell(y, y') = 0, \quad wenn \quad y = y'; \quad 1, \quad sonst$
  - Klassenabhängige Kostenmatrix: $\ell(y, y') = c_{yy'}$
- Verlustfunktionen für Regression
  - Squared error: $\ell(y, y') = (y - y')^2$

Hypothesenbewertung

- Zentrale Annahme: dem Lernproblem liegt eine (unbekannte) Verteilung $p(\mathbf{x},y)$ zugrunde



- Empirisches Risiko:

$$\hat{R}(f) = \frac{1}{m} \sum_{j=1}^{m} l(y_j, f(\mathbf{x}_j))$$

Fehler eines Schätzers

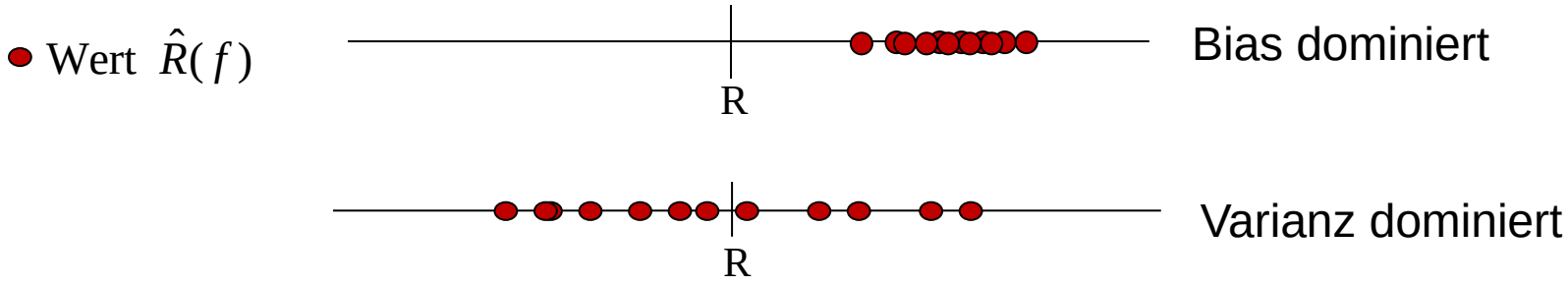
- Empirischer Fehler ist Schätzer

$$\hat{R}(f) = \frac{1}{m} \sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j))$$

- Schätzer ist Zufallsvariable.
  - Wert hängt von Zufallsexperiment „Messung des empirischen Risikos“ ab

$$(\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y)$$ Welche $(\mathbf{x}_i, y_i)$ werden gezogen?

- $$\mathbb{E}\left[\left(\hat{R} - R\right)^2\right] = Bias\left(\hat{R}\right)^2 + Var\left(\hat{R}\right)$$



Bias eines Schätzers

- Schätzer $\hat{R}(f)$ ist erwartungstreu, genau dann wenn:
  - $E[\hat{R}(f)] = R(f)$
- Ansonsten hat $\hat{R}(f)$ einen Bias:
  - $Bias = E[\hat{R}(f)] - R(f)$
- Schätzer ist optimistisch, wenn
  - *Bias* < 0.
- Schätzer ist pessimistisch, wenn
  - *Bias* > 0.
- Schätzer ist erwartungstreu, wenn
  - *Bias* = 0.

Varianz eines Schätzers

- Schätzer $\hat{R}(f)$ hat eine Varianz

$$Var[\hat{R}(f)] = E[(\hat{R}(f) - E[\hat{R}(f)])^2] = E[\hat{R}(f)^2] - E[\hat{R}(f)]^2$$

- Je größer die Stichprobe ist, die zum Schätzen verwendet wird, desto geringer ist die Varianz.
- Genaue Form der Varianz hängt von der Verlustfunktion ab.

- Hohe Varianz: großer „Zufallsanteil“ bei der Bestimmung des empirischen Risikos.
- Großer Bias: systematischer Fehler bei der Bestimmung des empirischen Risikos.

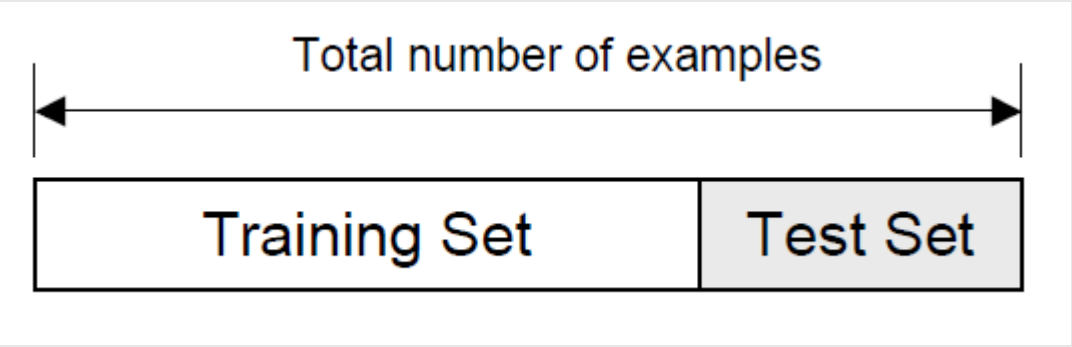
Hypothesenbewertung: Risikoschätzung

- Empirisches Risiko auf Daten $T = \langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \rangle$:

- Wichtig: Wo kommt T her?
  - Trainingsdaten (T=L)?
  - Hold-out: Verfügbare Daten in disjunkte L und T aufteilen.
  - Cross-Validation (Spezialfall: Leave-one-out)

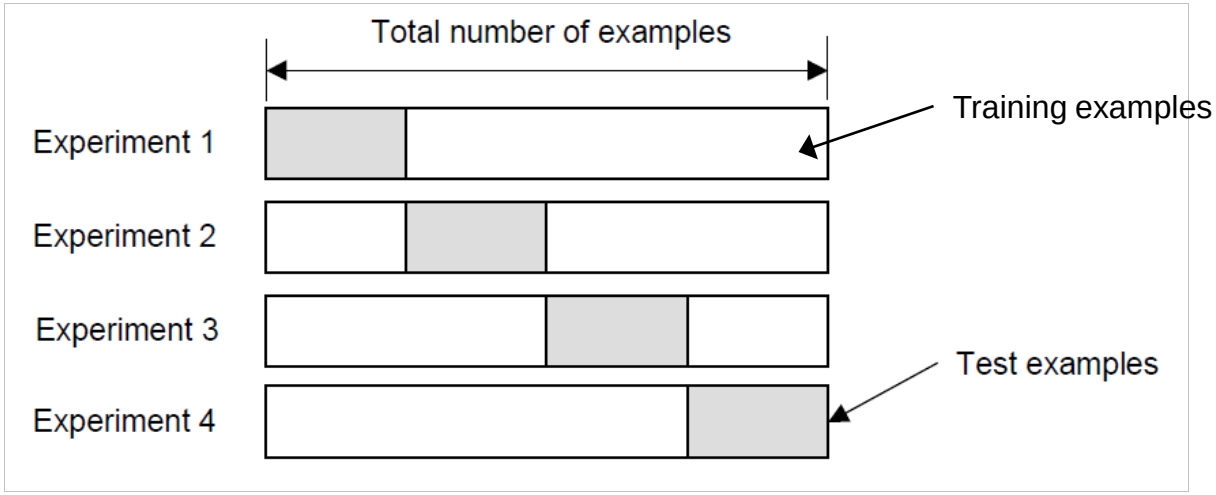
Holdout-Testing

- Starte Lernalgorithmus mit Daten L, gewinne so Hypothese f_L .
- Ermittle empirisches Risiko $\hat{R}_T(f_L)$ auf Daten T.
- Starte Lernalgorithmus auf Daten *D*, gewinne so Hypothese f_D .
- Ausgabe: Hypothese f_D , benutze $\hat{R}_T(f_L)$ als Schätzer für das Risiko von f_D



Cross Validation

- Gegeben: Daten $D = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_d, y_d) \rangle$
- Teile D in n Abschnitte $D_i = \langle (\mathbf{x}_{i_1}, y_{i_1}), ..., (\mathbf{x}_{i_k}, y_{i_k}) \rangle, \quad k = d / n$ mit $D = \bigcup_{i=1}^{n} D_i$ und $D_i \cap D_j = 0$
- Wiederhole für i=1..n
  - Trainiere f_i mit $L_i = D \setminus D_i$.
  - Bestimme empirisches Risiko $\hat{R}_{D_i}(f_i)$ auf D_i.

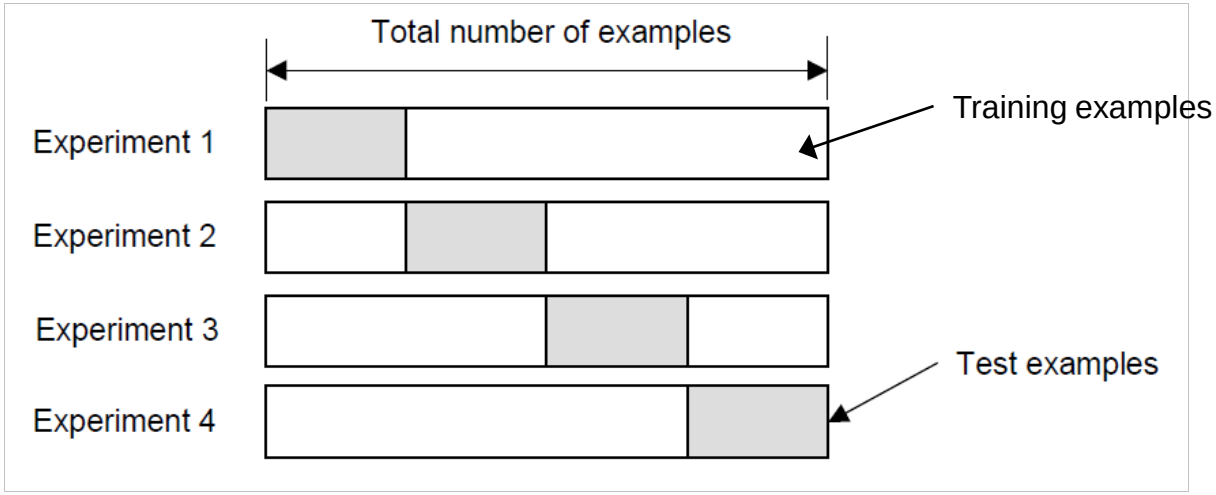


Cross Validation

- Mittle empirische Risikoschätzungen auf den jeweiligen Testmengen D_i:

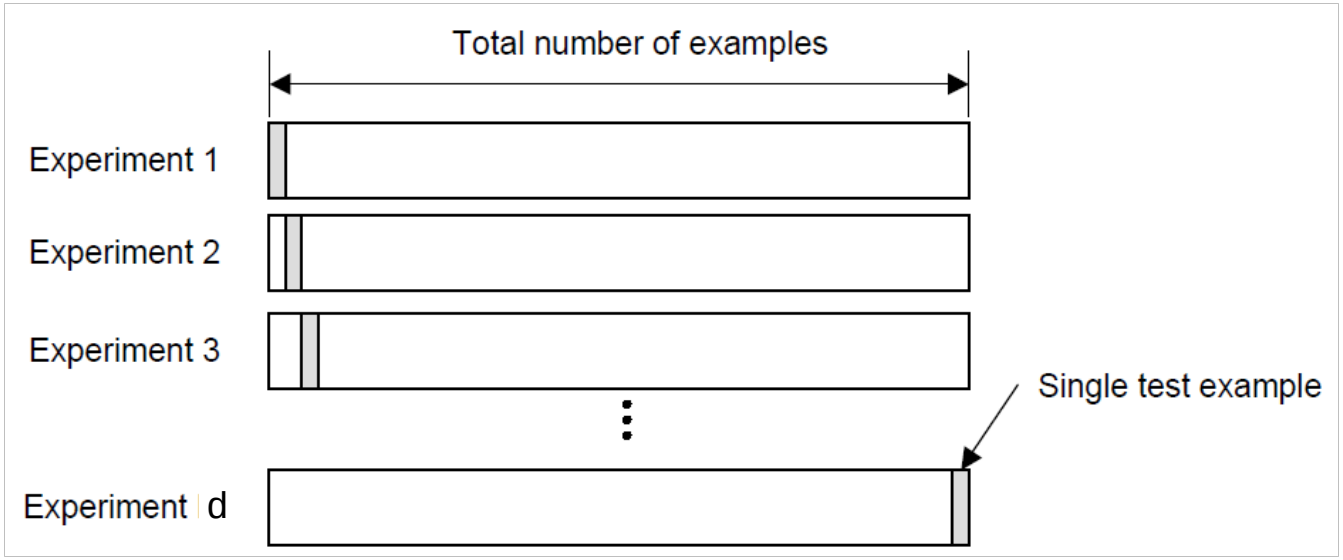
$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} \hat{R}_{D_i}(f_i)$$

- Trainiere f_D auf allen Daten D.
- Liefere Hypothese f_D und Schätzer $\overline{R}$.



Leave-One-Out Cross-Validation

- Spezialfall $n=d$ heisst auch *leave-one-out* Fehlerschätzung



Überblick

- Wiederholung: Hypothesenbewertung
  - Verfahren
  - Anwendungen
  - Konfidenzintervalle
- ROC-Analyse
- Statistische Tests
  - p-Wert
  - Vorzeichen-, Wald-, t- und Pearsons –Test

Anwendungen Hypothesenevaluierung

- Verfahren hat einen Parameter, den wir einstellen müssen
  - Regularisierungsparameter λ

$$f_{\mathbf{w}*} = \arg\min_{f_{\mathbf{w}}} \sum_i \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \lambda \| \mathbf{w} \|^2 \qquad \lambda=?$$

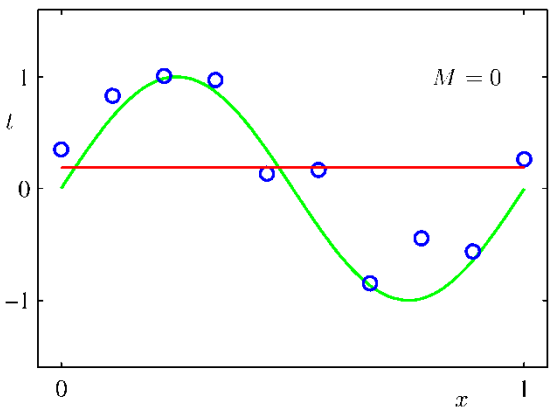
  - (Hyper)Parameter, der Modellklasse bestimmt, z.B. Polynomgrad bei polynomieller Regression

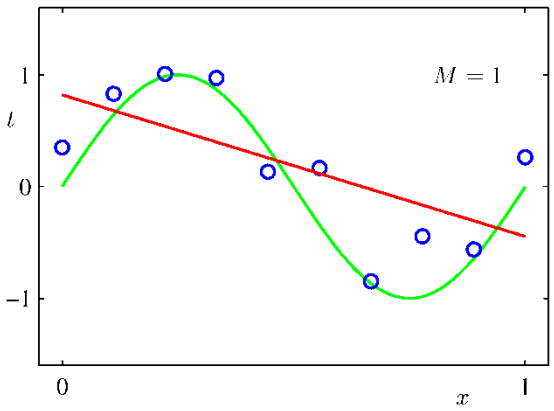
$$f_{\mathbf{w}}(x) = \sum_{i=0}^{M} w_i x^i \qquad \text{M=?}$$

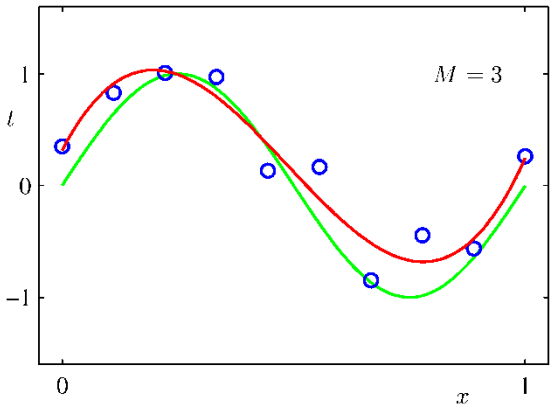
- In allen diesen Fällen ist der Trainingsfehler kein geeignetes Entscheidungskriterium!
  - Besser Fehlerschätzung mit Holdout-Menge oder Cross-Validierung

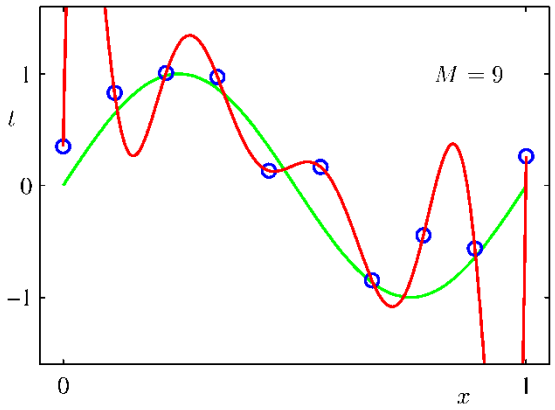
Beispiel polynomielle Regression: Training vs. Testfehler

- Erfolg des Lernens hängt vom gewählten Polynomgrad M ab, der Komplexität des Modells kontrolliert (hier besonders stark, weil Modell nicht regularisiert)



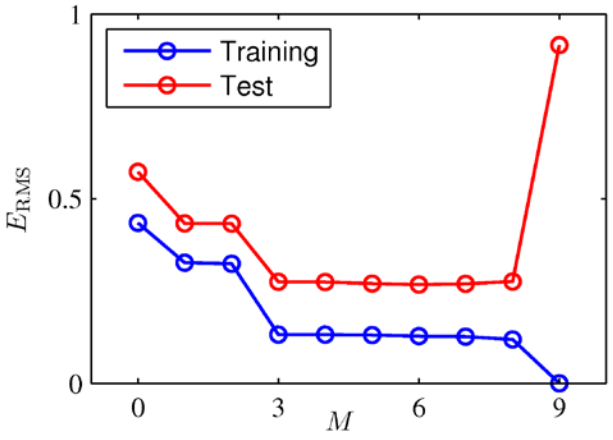


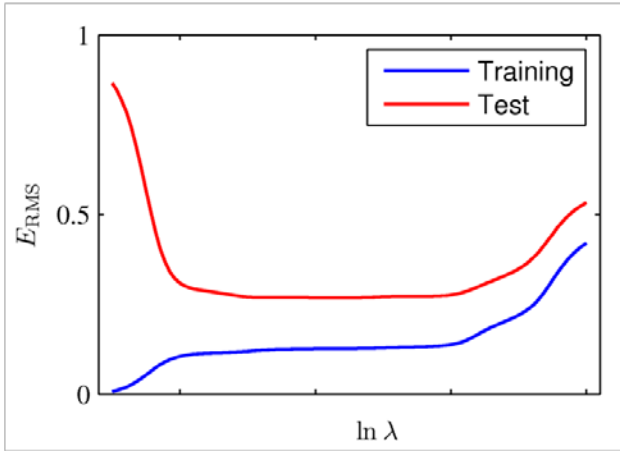




Regularisierte Polynomielle Regression

- Regularisierer wirkt wie eine Begrenzung der Modellkomplexität und verhindert Überanpassung
- In der Praxis am besten, Modellkomplexität durch Regularisierung zu kontrollieren (direkter Parameter wie bei Polynomen oft nicht verfügbar)
- Regularisierer kann durch Fehlerschätzung (Holdout-Testing oder Cross-Validation) eingestellt werden.





Triple-Cross-Validation

- Ziel: Abschätzung der Genauigkeit von Vorhersagen unter optimalen Parametern

Triple-Cross-Validation

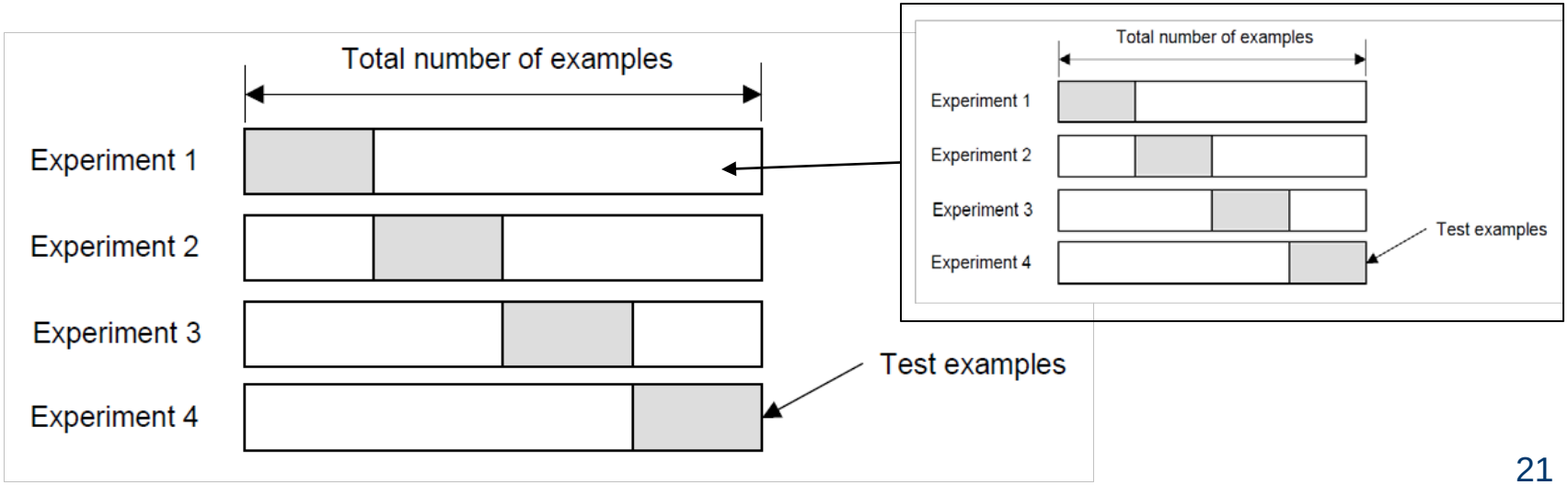
- Gegeben: Daten $D = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_d, y_d) \rangle$
- Teile D in n Abschnitte $D_i = \langle (\mathbf{x}_{i_1}, y_{i_1}), ..., (\mathbf{x}_{i_k}, y_{i_k}) \rangle, \quad k = d / n$ mit $D = \bigcup_{i=1}^{n} D_i$ und $D_i \cap D_j = 0$
- Wiederhole für i=1..n
  - Teile $D \setminus D_i$ in m Abschnitte mit $D \setminus D_i = \bigcup_{j=1}^{m} D_{i,j}$ und $D_{i,j} \cap D_{i,k} = 0$
  - Wiederhole für j=1..m
    - Trainiere $f_{i,j,C}$ mit $D \setminus D_i \setminus D_{i,j}$ f.a. möglichen Parameter C
    - Bestimme empirisches Risiko $\hat{R}_C(f_{i,j,C})$ auf $D_{i,j}$.
    - Bestimme C^* mit minimalen Risiko $\hat{R}_C$
  - Trainiere f_i mit $D \setminus D_i$ und C^*
  - Bestimme empirisches Risiko $\hat{R}_{D_i}(f_i)$ auf D_i.

Triple-Cross-Validation

- Mittle empirische Risikoschätzungen auf den jeweiligen Testmengen D_i:

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} \hat{R}_{D_i}(f_i)$$

- Trainiere f_D auf allen Daten D.
- Liefere Hypothese f_D und Schätzer $\overline{R}$.



Überblick

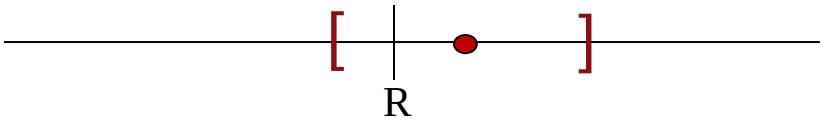
- Wiederholung: Hypothesenbewertung
  - Verfahren
  - Anwendungen
  - Konfidenzintervalle
- ROC-Analyse
- Statistische Tests
  - p-Wert
  - Vorzeichen-, Wald-, t- und Pearsons –Test

Konfidenzintervalle

- Idee Konfidenzintervall:
  - Intervall um den geschätzten Fehler $\hat{R}(f)$ angeben
  - so dass der echte Fehler „meistens“ im Intervall liegt
  - Quantifiziert Unsicherheit der Schätzung

- Weg zum Konfidenzintervall: Analyse der Verteilung der Zufallsvariable $\hat{R}(f)$

$\hat{R}(f)$



Zero-One Loss und Fehlerwahrscheinlichkeit

- Für Konfidenzintervalle betrachten wir Risikoschätzung im Spezialfall Klassifikation mit Zero-One Loss
- Verlustfunktion Zero-One Loss:
  - $l(y, y') = 0, \quad wenn \quad y = y'; \quad 1, \quad sonst$
- $\rightarrow$ Risiko = Fehlerwahrscheinlichkeit.
  - $$R(f) = \int \ell(y, f(\mathbf{x}))\, p(\mathbf{x}, y)\, d\mathbf{x}dy$$
    $$= \int [[y \neq f(\mathbf{x})]]\, p(\mathbf{x}, y)\, d\mathbf{x}dy$$
    $$= p(y \neq f(\mathbf{x}))$$

$[[Ereignis]]$: binäre Indikatorvariable für "Ereignis"

Verteilung für Fehlerschätzer

- Hypothese f wird auf separater Testmenge mit m unabhängigen Beispielen evaluiert:

$$\hat{R}_T(f) = \frac{1}{m}\sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j)) \qquad T = \langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \rangle$$

- Fehlerschätzer ist erwartungstreu, $E[\hat{R}_T(f)] = R(f)$

- Fehlerschätzer ist Summe über Beispielverluste: Bei jedem Beispiel kann ein korrektes oder falsches Ergebnis beobachtet werden

$$\ell_j = \ell(y_j, f(\mathbf{x}_j)) \in \{0,1\}$$ unabhängig, Bernouilli-verteilt mit Parameter $R(f)$

$$\ell_j \sim \text{Bern}(\ell \mid R(f))$$

- Entspricht m Münzwürfen

Schranken für echtes Risiko

- Was sagt das empirische Risiko $\hat{r} = \hat{R}_T(f)$ jetzt also über das echte Risiko?
- Empirisches Risiko $\hat{r} \rightarrow$ empirische Varianz $s_{\hat{r}}^2 = \frac{\hat{r}(1-\hat{r})}{m-1}$
- Einseitige Schranke für echtes Risiko:

$$P\left(R(f) \le \hat{R}_T(f) + \varepsilon\right) = P\left(R(f) - \hat{R}_T(f) \le \varepsilon\right)$$

$$= P\left(\frac{e}{s_{\hat{r}}} \le \frac{\varepsilon}{s_{\hat{r}}}\right)$$

$$\approx \Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right)$$

$$e = R(f) - \hat{R}(f)$$

$$P\left(\frac{e}{s_{\hat{r}}} \mid r\right) \approx N\left(\frac{e}{s_{\hat{r}}} \mid 0,1\right)$$

$\Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right)$ kumulative Verteilungsfunktion der Normalverteilung

Schranken für echtes Risiko

- Was sagt das empirische Risiko $\hat{r} = \hat{R}_T(f)$ jetzt also über das echte Risiko?
- Empirisches Risiko $\hat{r} \rightarrow$ empirische Varianz $s_{\hat{r}}^2 = \frac{\hat{r}(1-\hat{r})}{m-1}$

- Zweiseitige Schranke:

$$P\left(|R(f) - \hat{R}_T(f)| \leq \varepsilon\right) = 1 - P\left(R(f) - \hat{R}_T(f) > \varepsilon\right) + 1 - P\left(\hat{R}_T(f) - R(f) > \varepsilon\right)$$

$$= \ldots$$

$$\approx 2\left(1 - \left(1 - \Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right)\right)\right)$$

$$= 2\Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right)$$

Konfidenzintervalle

- Idee: ε so wählen, dass Schranke mit vorgegebener Wahrscheinlichkeit von $1-\delta$ (z.B. $\delta = 0.05$) gilt.

- Einseitiges $1-\delta$-Konfidenzintervall: Schranke ε, so dass
$$P\left(R(f) \leq \hat{R}_T(f) + \varepsilon\right) \geq 1-\delta$$

- Zweiseitiges $1-\delta$-Konfidenzintervall: Schranke ε, so dass
$$P\left(|R(f) - \hat{R}_T(f)| \leq \varepsilon\right) \geq 1-\delta$$

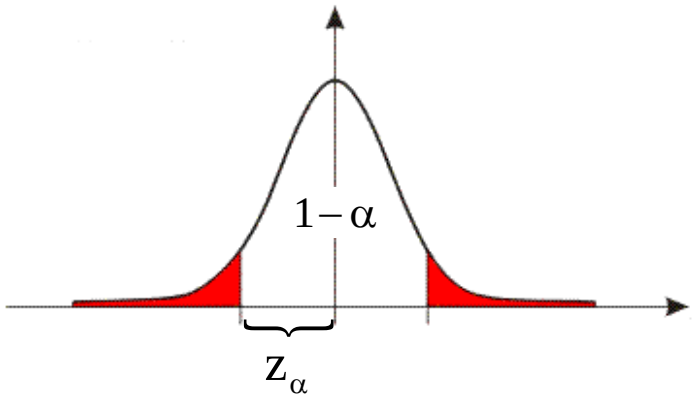
- Bei symmetrischer Verteilung gilt immer:
  - ε zu einseitigem $1-\delta$-Konfidenzintervall
    = ε zu zweiseitigem $1-\delta/2$-Konfidenzintervall.

Konfidenzintervalle

- $\hat{R}_T(f)$ ist annähernd normal-verteilt

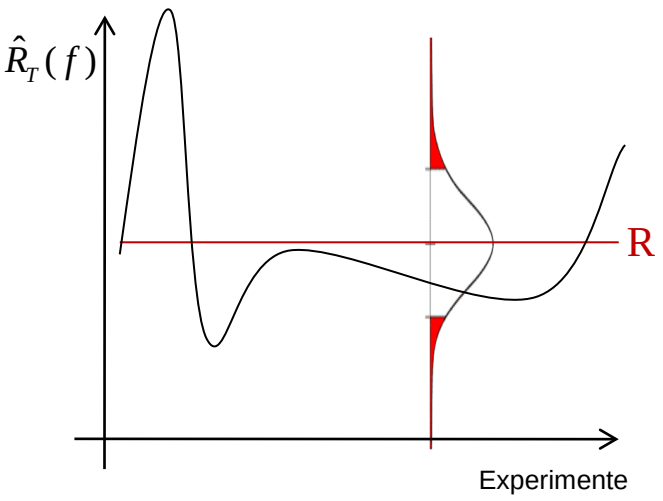
$$P\left(|R(f)-\hat{R}_n(f)|>z_\alpha\right)=1-\alpha$$

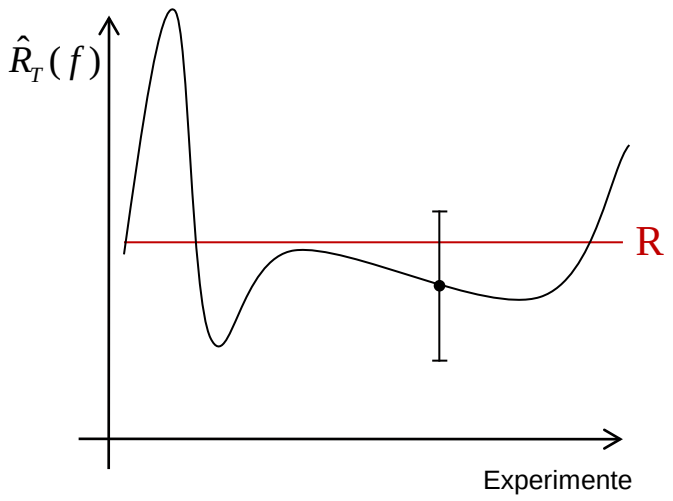
$$z_\alpha=\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$$



Konfidenzintervalle

- $\hat{R}_T(f)$ ist annähernd normal-verteilt





Students t-Verteilung

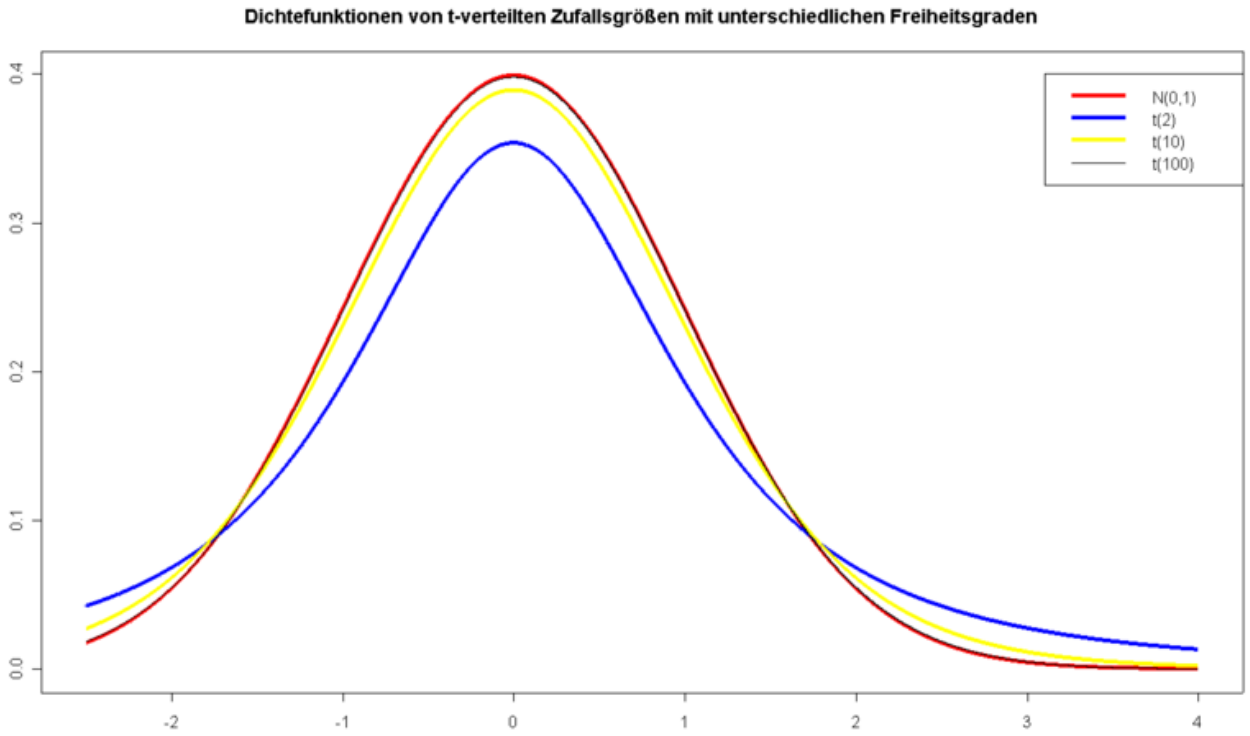
- Empirisches Risiko annähernd normalverteilt:
  - $$p\left(\hat{R}_T(f)=\hat{r}\mid r\right)=B\left(m\hat{r}\mid r,m\right)$$
    $$\approx N\left(\hat{r}\mid r,\frac{r(1-r)}{m}\right)$$
    $$=N\left(\frac{\hat{r}-r}{\sigma_{\hat{r}}}\mid 0,1\right)$$

    Einfache Charakterisierung der Verteilung des empirischen Fehlers
- Problem: Risiko muss bekannt sein, damit wir Varianz bzw. Standardfehler bestimmen können.
  - $$\sigma_{\hat{r}}^2=\frac{r(1-r)}{m};\quad \sigma_{\hat{r}}=\sqrt{\frac{r(1-r)}{m}}$$
- Nur das empirische Risiko ist gegeben.

Students t-Verteilung

- Schätzen der Varianz durch empirische Varianz:
  - $s_{\hat{r}}^2 = \frac{\hat{r}(1-\hat{r})}{m-1}, \qquad s_{\hat{r}} = \sqrt{\frac{\hat{r}(1-\hat{r})}{m-1}}$
- Empirisches Risiko folgt bei geschätzter Varianz Students t-Verteilung (ähnlich Gauß-Verteilung, aber mehr Wahrscheinlichkeitsmasse in den Außenbereichen).
- Für große m konvergiert Students t-Verteilung gegen die Normalverteilung.

Students t-Verteilung



$$\lim_{m\to\infty} t\left(\frac{\hat{r}-r}{s_{\hat{r}}}|m\right) = N\left(\frac{\hat{r}-r}{s_{\hat{r}}}|0,1\right)$$

$$f_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t$$

Überblick

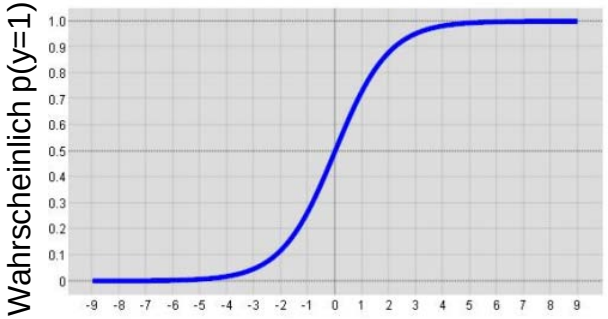
- Wiederholung: Hypothesenbewertung
  - Verfahren
  - Anwendungen
  - Konfidenzintervalle
- ROC-Analyse
- Statistische Tests
  - Vorzeichen-, Wald-, t- und Pearsons –Test
  - p-Wert

Klassifikator / Entscheidungsfunktion

- Für eine binäre Klassifikation (y= +1 oder -1) wird oft eine kontinuierliche Entscheidungsfunktion $f(x)$ gelernt.
  - Z.B. lineares Modell

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{m} w_i x_i$$

- Je größer $f(x)$, desto wahrscheinlicher ist, dass x zur Klasse +1 gehört
  - Z.B. logistische Regression



$$\sigma(f(x)) = \frac{1}{1+\exp(-f(x))}$$

Klassifikator / Entscheidungsfunktion

- Wie bestimmen wir Klassenentscheidung +1/-1 aus f(x)?
- Allgemeine Lösung:

$$\text{Vorhersage} = \begin{cases} +1 : f(\mathbf{x}) \geq \theta \\ -1 : \text{sonst} \end{cases}$$

- Der Wert für θ verschiebt „false positives“ zu „false negatives“.
- Optimaler Wert hängt von Kosten einer positiven oder negativen Fehlklassifikation ab.

Evaluation von Klassifikatoren und Entscheidungsfunktionen

- Fehlklassifikationswahrscheinlichkeit
  - Häufig nicht aussagekräftig, weil P(+1) sehr klein.
  - Wie gut sind 5% Fehler, wenn P(+1)=3%?
  - Idee: Nicht Klassifikator bewerten, sondern Entscheidungsfunktion.

- Receiver Operating Characteristic (ROC-Kurve)
  - Bewertet Entscheidungsfunktion,
  - Jeder Punkt auf der ROC Kurve entspricht einem Schwellwert θ
  - Fläche unter ROC-Kurve = P(positives Beispiel hat höheren f-Wert als negatives Beispiel)

ROC-Analyse

- Entscheidungsfunktion + Schwellwert = Klassifikator

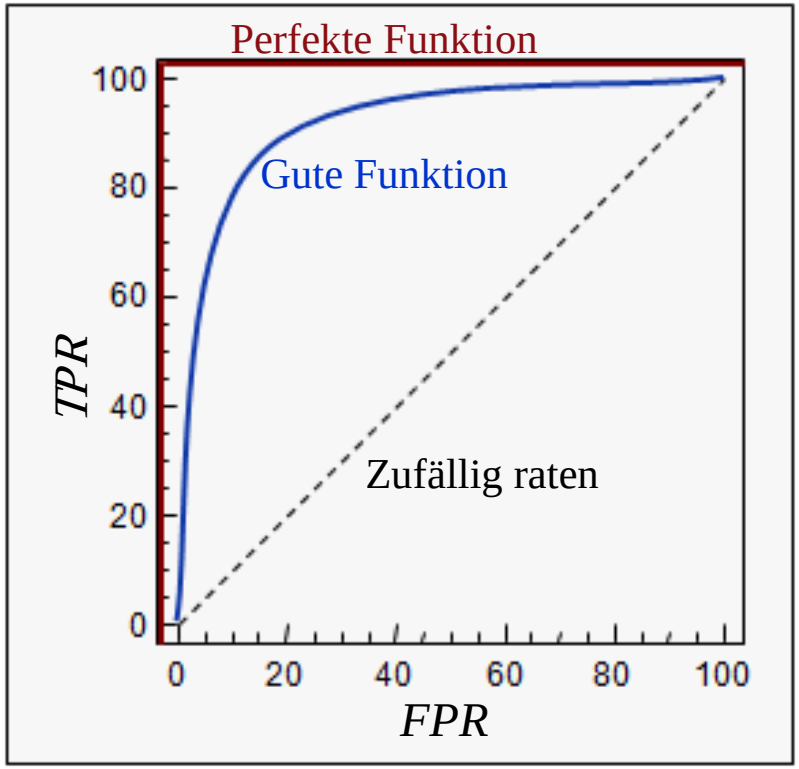
$$\text{Vorhersage} = \begin{cases} +1 : f(\mathbf{x}) \geq \theta \\ -1 : \text{sonst} \end{cases}$$

  - Fehler hängen vom Schwellwert ab
  - Großer Schwellwert: Mehr positive Bsp falsch.
  - Kleiner Schwellwert: Mehr negative Bsp falsch.

- ROC-Analyse: Bewertung der Entscheidungsfunktion unabhängig vom konkreten Schwellwert.
- Charakterisieren das Verhalten des Klassifikators für alle möglichen Schwellwerte.

ROC-Kurven

- Rate der „False Positives“ und „True Positives“ in Abhängigkeit des Schwellwertes
  - X-Achse: „False Positive Rate“
  - Y-Achse: „True Positive Rate“



	Vorhersage „+“	Vorhersage „-“
Echtes Label „+“	TP	FN
Echtes Label „-“	FP	TN

$$FPR = \frac{FP}{N} \qquad TPR = \frac{TP}{P}$$

$$N = FP + TN \qquad P = TP + FN$$

Bestimmen der ROC-Kurve von f

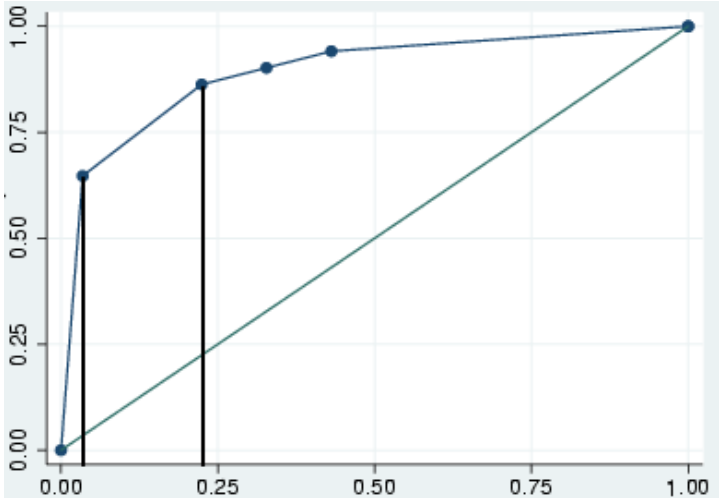
- Annahme: kein $f(\boldsymbol{x}) = f(\boldsymbol{x}')$ für $\boldsymbol{x} \neq \boldsymbol{x}'$.

- Generiere Liste L aller Instanzen $\boldsymbol{x}$, absteigend sortiert nach $f(\mathbf{x})$
- P = Anzahl positiver Instanzen, N = Anzahl negativer Instanzen
- TP = FP = 0

- Für i = 1 bis Länge(L)
  - x = i-tes Element von L
  - Wenn x positive Instanz: increment(TP)
  - Wenn x negative Instanz: increment(FP)
  - Zeichne neuen Punkt mit Koordination (FP/N,TP/P)

Flächeninhalt der ROC-Kurve

- Flächeninhalt AUC kann durch Integrieren (Summieren der Trapez-Flächeninhalte) bestimmt werden.



- x_+ = zufällig gezogenes Positivbeispiel
- x_- = zufällig gezogenes Negativbeispiel
- Theorem: AUC = $P(f(x_+) > f(x_-))$.

Precision / Recall

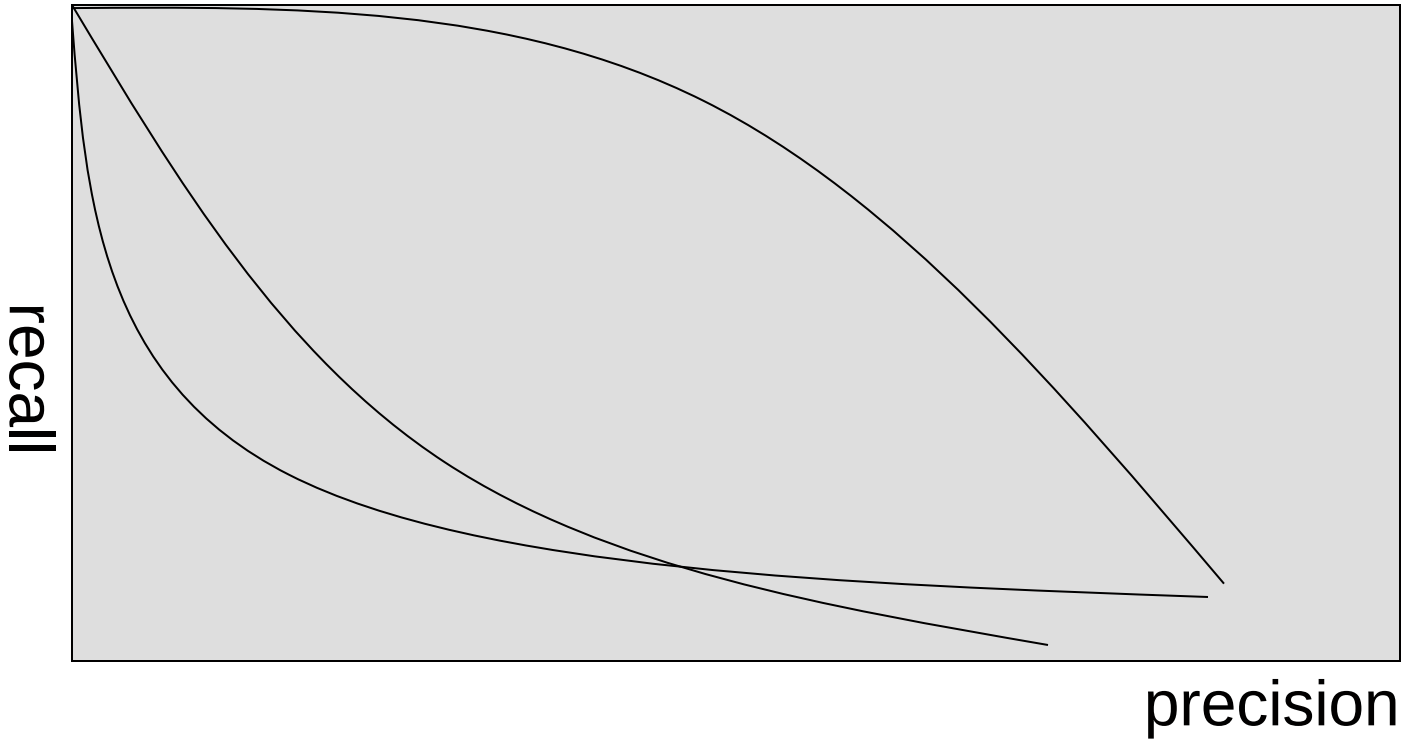
- Alternative zur ROC-Analyse.
- Stammt aus dem Information Retrieval.

- $\text{Precision} = \frac{TP}{TP + FP}$ ← Alle Instanzen mit Vorhersage „+“

- $\text{Recall} = \frac{TP}{TP + FN}$ ← Alle Instanzen mit echtem Label „+“

- Precision: P(positiv | positiv vorhergesagt)
- Recall: P(positiv vorhergesagt | ist positiv)

Precision / Recall Trade-Off



- Precision-/Recall-Kurven
- Welcher Klassifikator ist der Beste / Schlechteste

F-Measure, Breakeven Point

- Zusammenfassungen der Kurve in einer Zahl:
  - F-Measure: Harmonisches Mittel über Precision und Recall, maximiert über Schwellwert θ

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

  - Precision-Recall-Breakeven-Point: Es gibt einen Punkt θ auf der Kurve für den gilt Precision(θ) = Recall(θ) =: PRBEP

Evaluation von Hypothesen: Zusammenfassung

- Verlustfunktion, Risiko
- Empirisches Risiko $\rightarrow$ Aussagen über echtes Risiko.
  - Holdout-Testing, Cross Validation.
  - Ein-/zweiseitige Konfidenzschranken.
- Qualitäts-/Risikomaße
  - Fehlerrate,
  - ROC-Analyse, AUC,
  - Precision-Recall-Kurven.

Überblick

- Wiederholung: Hypothesenbewertung
  - Verfahren
  - Anwendungen
  - Konfidenzintervalle
- ROC-Analyse
- Statistische Tests
  - p-Wert
  - Vorzeichen-, Wald-, t- und Pearsons –Test

Statistische Tests

- Welche Schlussfolgerungen über die Realität erlauben uns Beobachtungen *wirklich?*
- Ein Test ist eine Prozedur mit den Eingaben
  - Nullhypothese,
  - Beobachtungen
  - Parameter α .
- Ein Test hat die möglichen Ausgaben
  - „Nullhypothese abgelehnt“ – das Gegenteil der Nullhypothese gilt.
  - „nicht abgelehnt“ – keine Schlussfolgerung möglich, kein neues Wissen gewonnen.

Statistische Tests

- Nullhypothese:
  - Aussage von der wir bis auf weiteres ausgehen,
  - die wir aber überprüfen möchten und zu widerlegen bereit sind.

- Bedingung für einen statistischen Test:
  - Wenn die Nullhypothese gilt, dann darf sie nur mit einer Wahrscheinlichkeit von höchstens α abgelehnt werden.

Statistische Tests

- Ausgabe „Nullhypothese abgelehnt“:
  - Wir ziehen die Schlussfolgerung, dass die Nullhypothese nicht die Realität beschreibt.
  - Neues Wissen gewonnen, Publikation!
- Ausgabe „nicht abgelehnt“:
  - Wir können keine Schlussfolgerung ziehen.
  - Vielleicht gilt die Nullhypothese, vielleicht nicht.

Statistische Tests

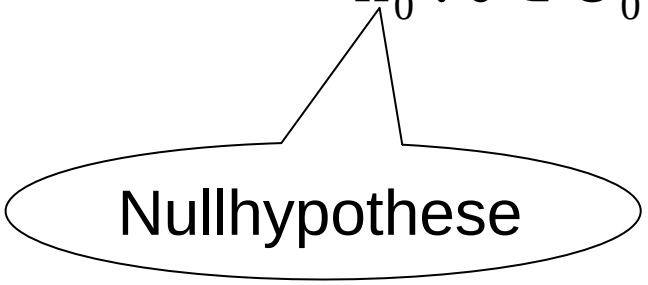
- Beispiel: Wirksamkeit von Medikamenten
  - Nullhypothese: „Medikament ist nicht wirksam“.
- Beobachtungen: Symptome bei einer Test- und einer Kontrollgruppe.
- Wenn sich Symptome bei Testgruppe so stark von Kontrollgruppe unterscheiden, dass
  - P(Beobachteter Unterschied | Nullhypothese) $< \alpha$, dann sagen wir dass die Unterschiede zwischen den Gruppen *signifikant* sind und lehnen die Nullhypothese ab. Medikament ist wirksam.
- Ansonsten kein Ergebnis.

Statistische Tests

- Ziel: anhand vorliegender Beobachtungen $x \in \mathcal{X}$ einer Zufallsvariable X eine *begründete* Entscheidung über die Gültigkeit oder Ungültigkeit einer Hypothese zu treffen

- Formal:

$$h_0 : \theta \in \Theta_0 \text{ vs. } h_1 : \theta \in \Theta_1$$

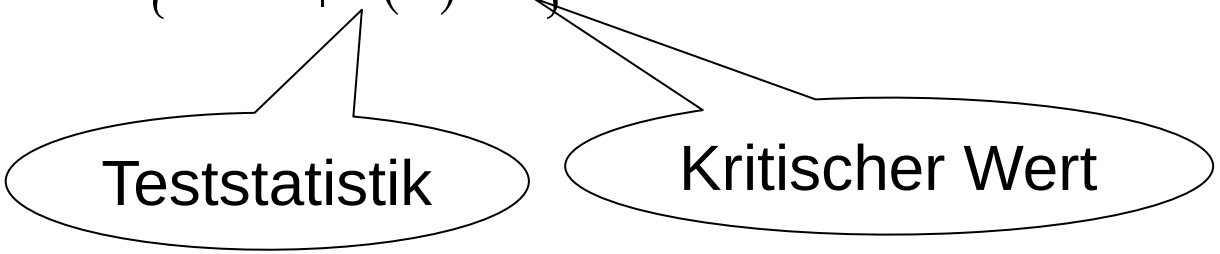


Statistische Tests

- Im Allgemeinen ist ein statistischer Tests durch seinen kritischen Bereich

$$R = \{x \in \mathcal{X} \mid T(x) > c\}$$

definiert.



- Wenn $X \in R$, lehnen wir die Nullhypothe ab, sonst nicht
- Woher kommen $T(x)$ und c?
  - problemabhängig
  - bestimmen die Aussagekraft (Verteilungsannahmen, Vorwissen)

Statistische Tests

- Viele Tests haben die folgende Form
  - einseitiger Test: $h_0 : \theta \leq \theta_0 \text{ vs. } h_1 : \theta > \theta_0$
  - zweiseitiger Test: $h_0 : \theta = \theta_0 \text{ vs. } h_1 : \theta \neq \theta_0$

- Weitere Unterscheidungen
  - 1 vs. 2 Stichproben-Tests
  - nach zu schätzenden Parametern (Mittelwert, Varianz)
  - Varianz bekannt / unbekannt
  - paired / unpaired

- Signifikanz-Niveau eines Tests: $\alpha = \sup_{\theta \in \Theta_0} P\left(X \in R \mid \theta\right)$

p-Wert

- Die Aussage „Nullhypothese abgelehnt“ ist nicht sehr informativ

- p-Wert: kleinste Signifikanz-Niveau α, für das die Nullhypothese abgelehnt wird
  - Wahrscheinlichkeit unter Annahme der Nullhypothese, dass die *wirkliche* Teststatistik größer ist, als die beobachtete
  - Achtung: keine Wahrscheinlichkeit, dass Nullhypothese richtig ist!
- Ursache für großen p-Wert
  - Nullhypothese richtig ODER
  - Nullhypothese falsch, aber Test zu schwach

p-Wert

- p-Wert: kleinste Signifikanz-Niveau α, für das die Nullhypothese abgelehnt wird
  - Wahrscheinlichkeit unter Annahme der Nullhypothese, dass die *wirkliche* Teststatistik größer ist, als die beobachtete

- Ein p-Wert von
  - <5% gilt als signifikant
  - <1% gilt als sehr signifikant
  - <0,1% gilt als hoch signifikant

Beispiel

- 12 Patienten wurden zwei unterschiedliche Schmerzmittel A und B verabreicht und die Wirkung in Stunden gemessen

Patient	1	2	3	4	5	6	7	8	9	10	11	12
A	2	3,6	2,6	2,6	7,3	3,4	14,9	6,6	2,3	2	6,8	8,5
B	3,5	5,7	2,9	2,4	9,9	3,3	16,7	6	3,8	4	9,1	20,9

- Gibt es Unterschiede zwischen den Medikament in der Wirkung?

Vorzeichen-Test

- Seien $x_1, \ldots, x_n$ unabhängig und identisch verteilt mit Median m

- $h_0 : m = \mu_0$ vs. $h_1 : m \neq \mu_0$

unter h_0 binomial-verteilt

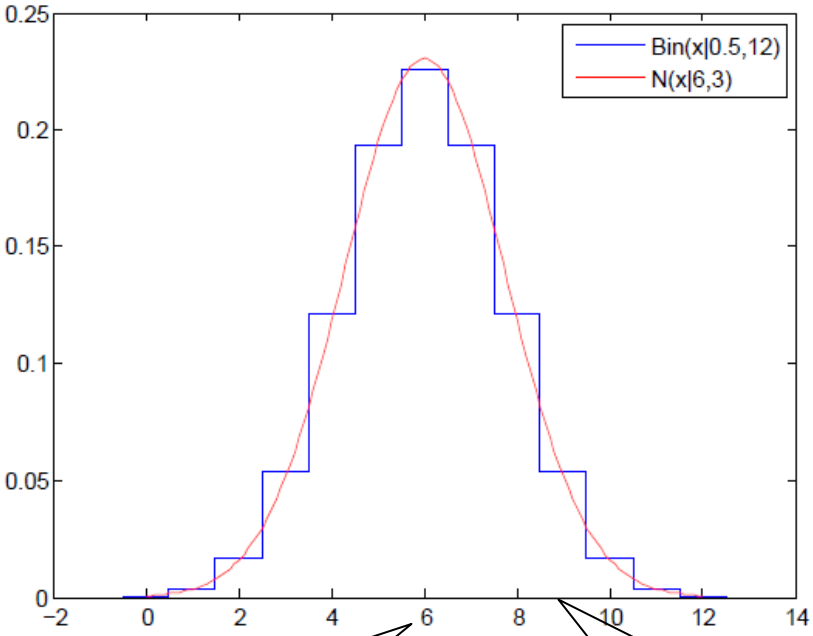
- Lehne Nullhypothese ab, gdw. $T(x) > c$

  - $$T(x) = \max\left(\sum_{i=1}^{n} [\![x_i - \mu_0 > 0]\!], \sum_{i=1}^{n} [\![x_i - \mu_0 < 0]\!] \right)$$

  - $$c = \text{BinCDF}_{n,0.5}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Vorzeichen-Test

- Lehne Nullhypothese ab, gdw. $T(x) > c$



unter h_0 binomial-verteilt

$h_0 : m = \mu_0$

Wie wahrscheinlich ist $T(x) = 9$?

Beispiel

- 12 Patienten wurden zwei unterschiedliche Schmerzmittel A und B verabreicht und die Wirkung in Stunden gemessen

Patient	1	2	3	4	5	6	7	8	9	10	11	12
A	2	3,6	2,6	2,6	7,3	3,4	14,9	6,6	2,3	2	6,8	8,5
B	3,5	5,7	2,9	2,4	9,9	3,3	16,7	6	3,8	4	9,1	20,9
x=B-A	1,5	2,1	0,3	-0,2	2,6	-0,1	1,8	-0,6	1,5	2	2,3	12,4

- Gibt es Unterschiede zwischen den Medikament in der Wirkung?
  - Nullhypothese: beide gleich $h_0 : m = 6$

- $T(x) = 9$

Beispiel
p-Wert

Patient	1	2	3	4	5	6	7	8	9	10	11	12
A	2	3,6	2,6	2,6	7,3	3,4	14,9	6,6	2,3	2	6,8	8,5
B	3,5	5,7	2,9	2,4	9,9	3,3	16,7	6	3,8	4	9,1	20,9
x=B-A	1,5	2,1	0,3	-0,2	2,6	-0,1	1,8	-0,6	1,5	2	2,3	12,4

- Nullhypothese $\mathrm{h}_0 : \mathrm{m} = 6$
- $\mathrm{T}(\mathrm{x}) = 9$

- $$p - Wert = p\left(T\left(X\right) > T\left(X_n\right) \mid h_0\right)$$
$$= p\left(Z > 9\right) + p\left(Z < 3\right), \quad Z \sim Bin\left(k \mid 12; 0,5\right)$$
$$= 2BinCDF\left(3 \mid 12; 0,5\right)$$
$$\approx 14,6\%$$

Wald-Test

- Gegeben eine normalverteilte Schätzung $\hat{\theta} \in \mathbb{R}$ für einen Parameter θ aus $x_1, \ldots, x_n$

$$\frac{\hat{\theta} - \theta_0}{\widehat{se}} \sim \mathcal{N}(0,1)$$

- $h_0 : \theta = \theta_0$ vs. $h_1 : \theta \neq \theta_0$
- Lehne Nullhypothese ab, gdw. $T(x) > c$
  - $T(x) = \dfrac{\left|\hat{\theta} - \theta_0\right|}{\widehat{se}}$
  - $c = \Phi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$

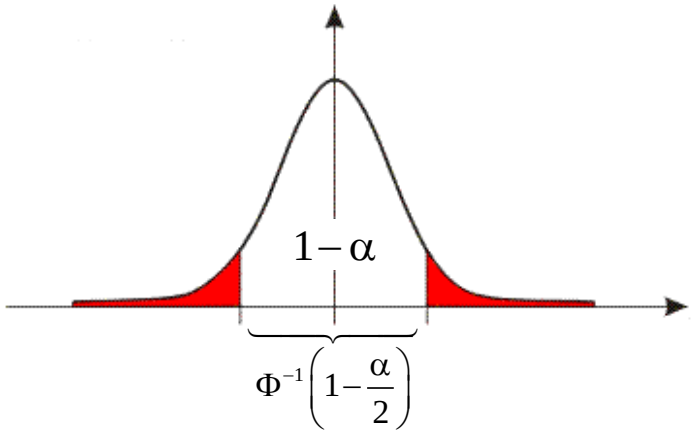
unter h_0 normalverteilt

Wald-Test

- Wald-Test: Lehne Nullhypothese ab, gdw.

$$\frac{\left|\hat{\theta} - \theta_0\right|}{\widehat{se}} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

- Wald-Test hat Signifikanz-Niveau α
  - Beweis: $\sup_{\theta \in \Theta_0} P_\theta\left(X \in R\right) = P_{\theta_0}\left(\frac{\left|\hat{\theta} - \theta_0\right|}{\widehat{se}} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$



$$\rightarrow P\left(\left|Z\right| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

$$= \alpha$$

Beispiel

- 12 Patienten wurden zwei unterschiedliche Schmerzmittel A und B verabreicht und die Wirkung in Stunden gemessen

Patient	1	2	3	4	5	6	7	8	9	10	11	12
A	2	3,6	2,6	2,6	7,3	3,4	14,9	6,6	2,3	2	6,8	8,5
B	3,5	5,7	2,9	2,4	9,9	3,3	16,7	6	3,8	4	9,1	20,9
x=B-A	1,5	2,1	0,3	-0,2	2,6	-0,1	1,8	-0,6	1,5	2	2,3	12,4

- Gibt es Unterschiede zwischen den Medikament in der Wirkung?
  - Nullhypothese: beide gleich $h_0 : \theta = 0$

- $$T(x) \approx \frac{2,133}{0,984} \approx 2,168$$

Beispiel
p-Wert

Patient	1	2	3	4	5	6	7	8	9	10	11	12
A	2	3,6	2,6	2,6	7,3	3,4	14,9	6,6	2,3	2	6,8	8,5
B	3,5	5,7	2,9	2,4	9,9	3,3	16,7	6	3,8	4	9,1	20,9
x=B-A	1,5	2,1	0,3	-0,2	2,6	-0,1	1,8	-0,6	1,5	2	2,3	12,4

- Nullhypothese $h_0 : \theta = 0$
- $T(x) \approx 2{,}168$

- $$p - Wert = p\left(T(X) > T(X_n) \mid h_0\right)$$
$$= p(Z > 2{,}168) + p(Z < -2{,}168), \quad Z \sim N(0;1)$$
$$= 2\,NormCDF(-2{,}168 \mid 0;1)$$
$$\approx 3\%$$

t-Test

- Seien $x_1, \ldots, x_n$ unabhängig normalverteilt mit Erwartungswert μ und unbekannter Varianz
- $h_0 : \mu = \mu_0$ vs. $h_1 : \mu \neq \mu_0$
- Lehne Nullhypothese ab, gdw. $T(x) > c$
  - $T(x) = \frac{\left|\overline{X}_n - \mu_0\right|}{\widehat{se}}, \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - $c = F_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right)$

unter h_0 t-verteilt
(n-1 Freiheitsgrade)

- Für kleine n besser geeignet als Wald-Test

Beispiel

- 12 Patienten wurden zwei unterschiedliche Schmerzmittel A und B verabreicht und die Wirkung in Stunden gemessen

Patient	1	2	3	4	5	6	7	8	9	10	11	12
A	2	3,6	2,6	2,6	7,3	3,4	14,9	6,6	2,3	2	6,8	8,5
B	3,5	5,7	2,9	2,4	9,9	3,3	16,7	6	3,8	4	9,1	20,9
x=B-A	1,5	2,1	0,3	-0,2	2,6	-0,1	1,8	-0,6	1,5	2	2,3	12,4

- Gibt es Unterschiede zwischen den Medikament in der Wirkung?
  - Nullhypothese: beide gleich $h_0 : \theta = 0$

- $$T(x) \approx \frac{2,133}{0,984} \approx 2,168$$

Beispiel
p-Wert

Patient	1	2	3	4	5	6	7	8	9	10	11	12
A	2	3,6	2,6	2,6	7,3	3,4	14,9	6,6	2,3	2	6,8	8,5
B	3,5	5,7	2,9	2,4	9,9	3,3	16,7	6	3,8	4	9,1	20,9
x=B-A	1,5	2,1	0,3	-0,2	2,6	-0,1	1,8	-0,6	1,5	2	2,3	12,4

- Nullhypothese $h_0 : \theta = 0$
- $T(x) \approx 2{,}168$

- $$p - Wert = p\left(T(X) > T(X_n) \mid h_0\right)$$
$$= p(Z > 2{,}168) + p(Z < -2{,}168), \quad Z \sim F_{n-1}(0)$$
$$= 2\,tCDF_{n-1}(-2{,}168)$$
$$\approx 5{,}3\%$$

Pearsons χ^2-Test

- Seien $x_1, \ldots, x_n$ unabhängig multinomial-verteilt mit Erwartungswert $\mu = (\mu_1, \ldots, \mu_k)$

$x_i = (x_i^1, \ldots, x_i^k), \quad x_i^j \in \{0,1\}$

- $h_0 : \mu = \mu_0$ vs. $h_1 : \mu \neq \mu_0$

- Lehne Nullhypothese ab, gdw. $T(x) > c$
  - $T(x) = \sum_{j=1}^{k} \frac{(x_1^j - \mu_j)^2}{\mu_j}$
  - $c = \left[\chi^2_{k-1}\right]^{-1}\left(1 - \frac{\alpha}{2}\right)$

unter h_0 χ^2-verteilt
(k-1 Freiheitsgrade)

Dualität

- Ein Test mit Signifikanzniveau α verwirft die Nullhypothese $h_0 : \mu = \mu_0$, genau dann nicht, wenn μ_0 innerhalb des $1 - \alpha$ -Vertrauensintervalls liegt.

Zusammenfassung

- Ein statistischer Test ist spezifiziert durch eine Statistik und einen kritischen Wert
- Wir lehnen die Nullhypothese ab, wenn $X \in R$

$$R = \{x \in \mathcal{X} \mid T(x) > c\}$$

- Da X unbekannt, hängt Test von Beobachtungen ab: Die Nullhypothese soll nur mit Wahrscheinlichkeit α fälschlicher Weise abgelehnt werden
- Verschiedene Tests: Vorzeichen-Test, Wald-Test, t-Test, Pearson χ^2