

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  $L = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \rangle$
  - ◆ Ausgabe: Hypothese (Modell)  $f : X \rightarrow Y$

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

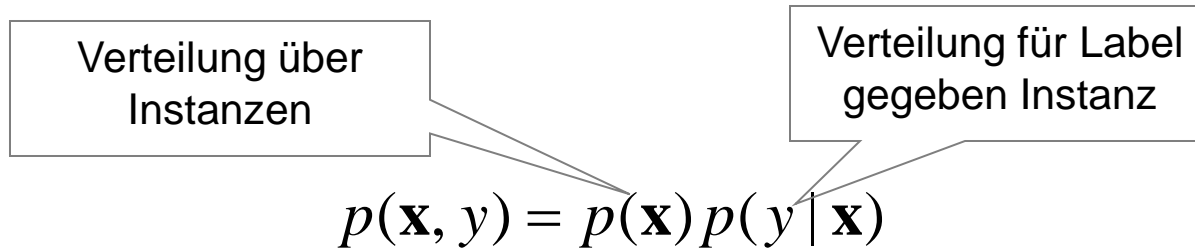
- Ziel des Lernens: genaue Vorhersagen treffen
- „Hypothesenbewertung“: Abschätzung der Genauigkeit von Vorhersagen
  - ◆ Schätzproblem: was ist eine gute Schätzung des erwarteten Fehlers?

# 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:  $\ell(y, y') = 0$ , wenn  $y = y'$ ; 1, 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

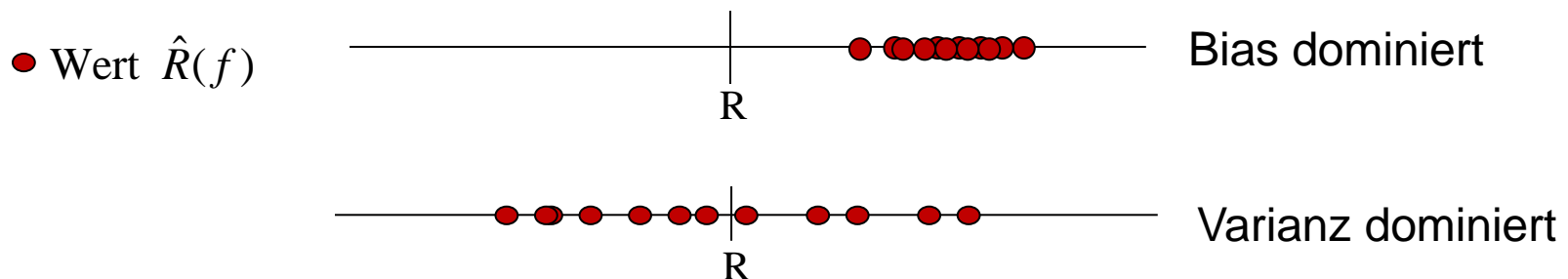
$$\hat{R}(f) = \frac{1}{m} \sum_{j=1}^m \ell(y_j, f(\mathbf{x}_j))$$

- 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?

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



# 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

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

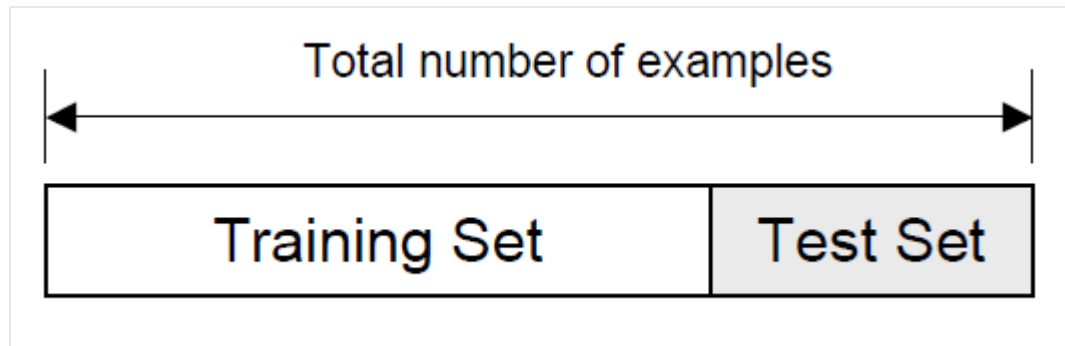
- Je größer die Stichprobe ist, die zum Schätzen verwendet wird, desto geringer ist die Varianz.
- Genaue Form der Varianz hängt von der Verlustfunktion ab.
- Hohe Varianz: großer „Zufallsanteil“ bei der Bestimmung des empirischen Risikos.
- Großer Bias: systematischer Fehler bei der Bestimmung des empirischen Risikos.

# Hypothesenbewertung: Risikoschätzung

- Empirisches Risiko auf Daten  $T = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \rangle$ :
- Wichtig: Wo kommt  $T$  her?
  - ◆ Trainingsdaten ( $T=L$ )?
  - ◆ Hold-out: Verfügbare Daten in disjunkte  $L$  und  $T$  aufteilen.
  - ◆ Cross-Validation (Spezialfall: Leave-one-out)

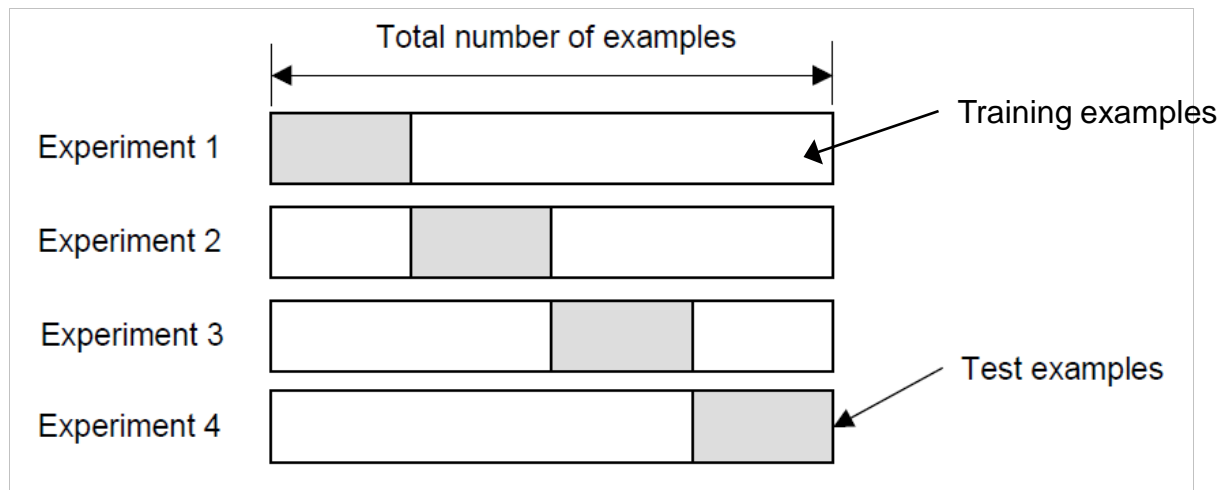
# Holdout-Testing

- 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  $D = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_d, y_d) \rangle$
- Teile  $D$  in  $n$  Abschnitte  $D_i = \langle (\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_k}, y_{i_k}) \rangle$ ,  $k = d / n$   
mit  $D = \bigcup_{i=1}^n D_i$  und  $D_i \cap D_j = \emptyset$
- 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  $D_i$

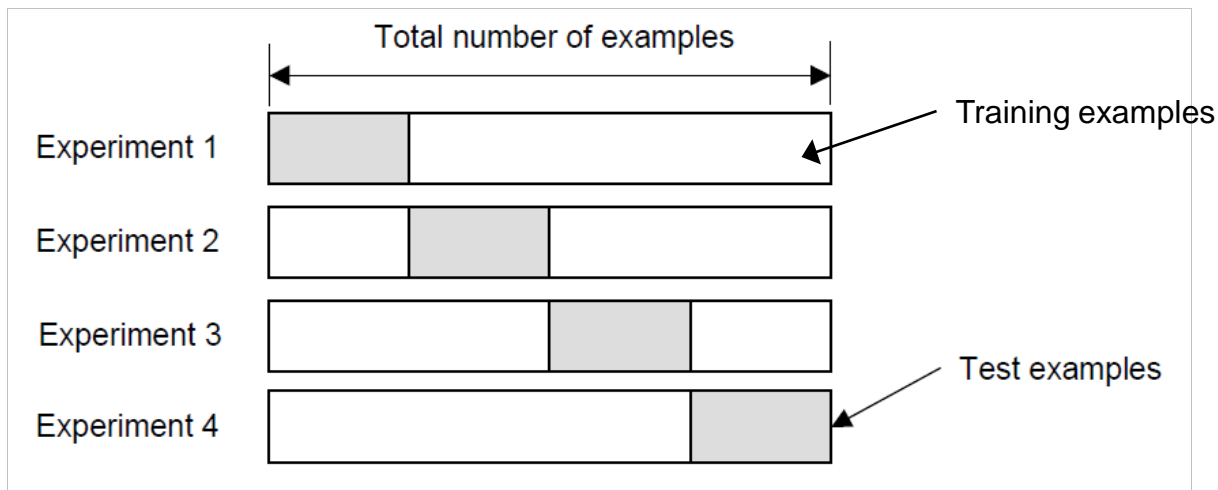


# Cross Validation

- Middle empirische Risikoschätzungen auf den jeweiligen Testmengen  $D_i$ :

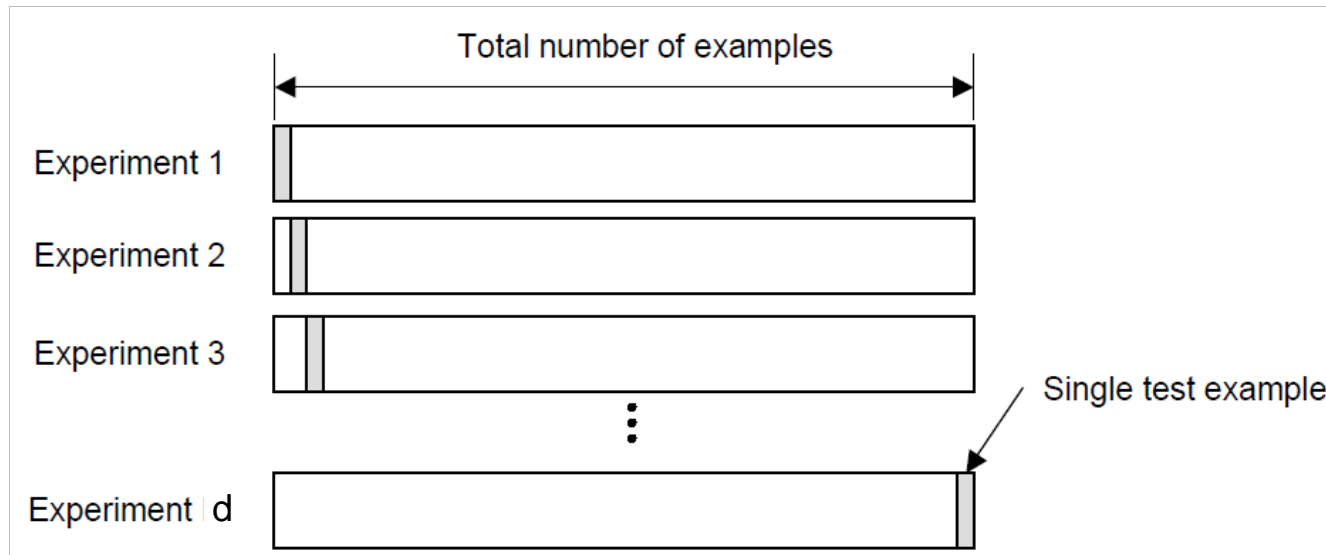
$$\bar{R} = \frac{1}{n} \sum_{i=1}^n \hat{R}_{D_i}(f_i)$$

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



# 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  $\lambda$

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

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

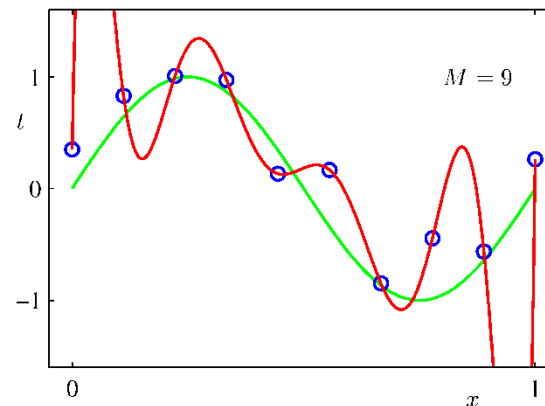
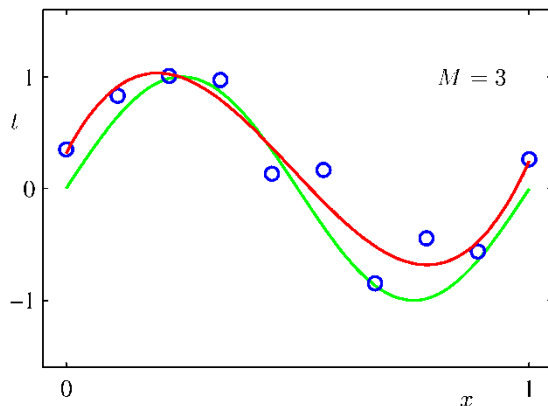
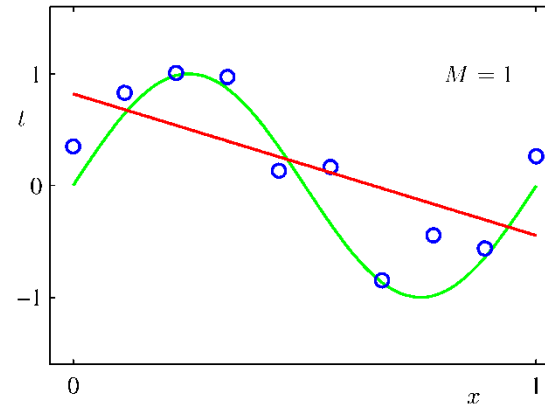
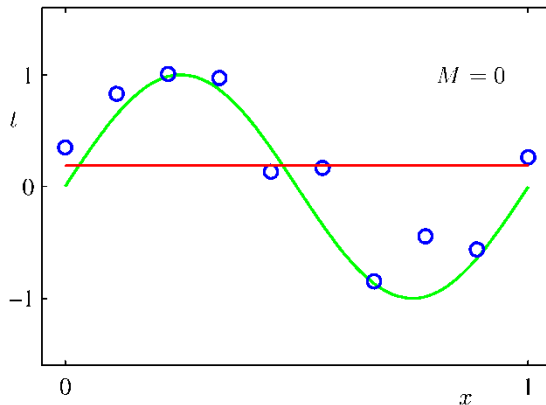
$$f_{\mathbf{w}}(x) = \sum_{i=0}^M w_i x^i \quad M=?$$

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



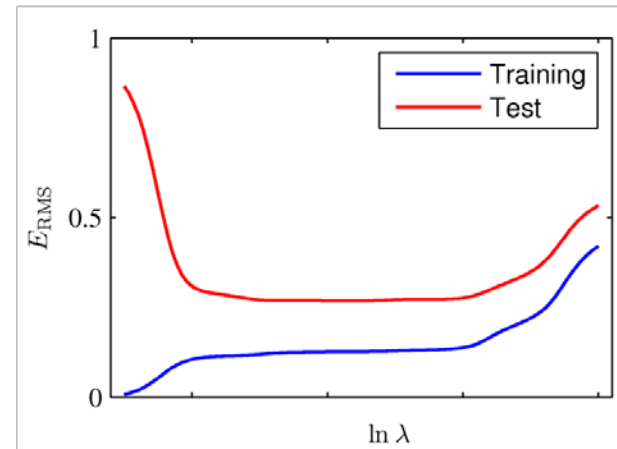
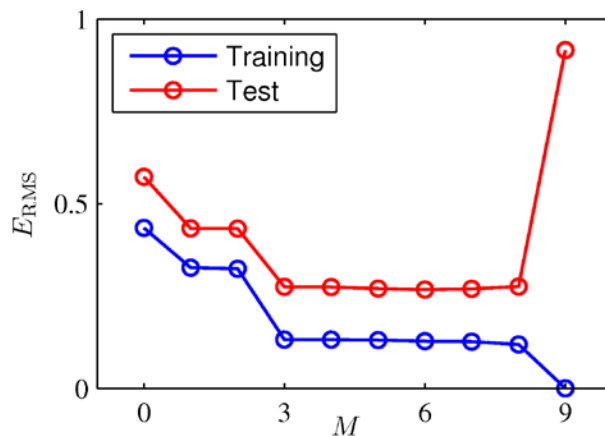
# Beispiel polynomielle Regression: Training vs. Testfehler

- 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

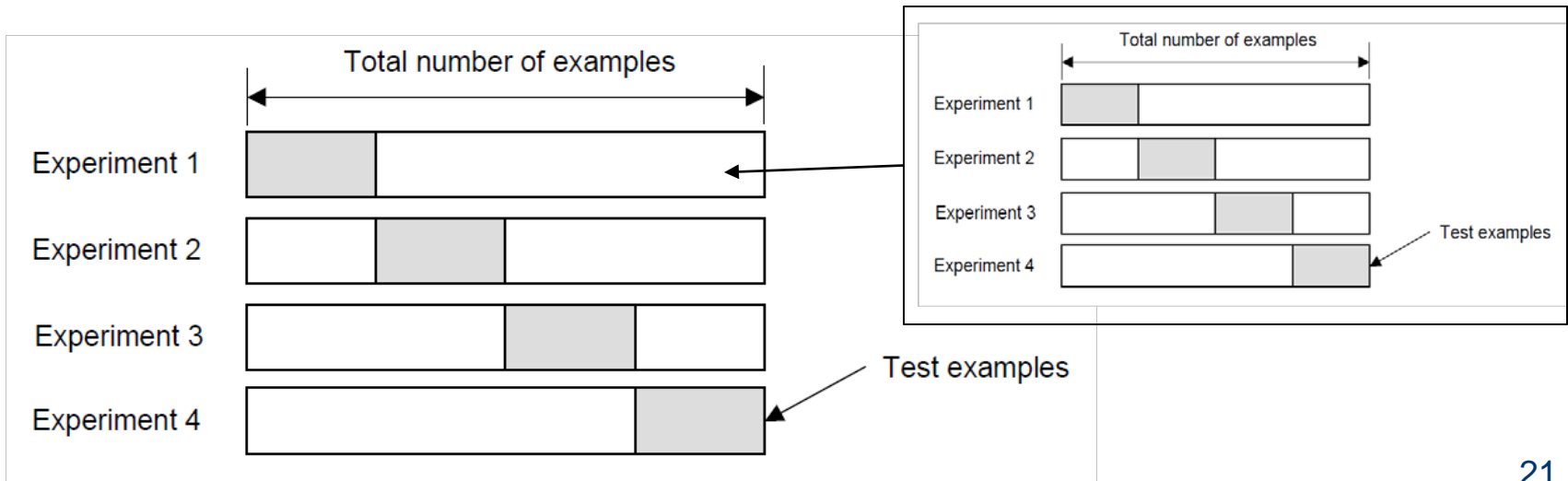
- Gegeben: Daten  $D = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_d, y_d) \rangle$
- Teile  $D$  in  $n$  Abschnitte  $D_i = \langle (\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_k}, y_{i_k}) \rangle$ ,  $k = d / n$   
mit  $D = \bigcup_{i=1}^n D_i$  und  $D_i \cap D_j = \emptyset$
- Wiederhole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

# Triple-Cross-Validation

- Middle empirical risk estimates on the respective test sets  $D_i$ :

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

- Train  $f_D$  on all data  $D$ .
- Deliver hypothesis  $f_D$  and estimator  $\bar{R}$ .



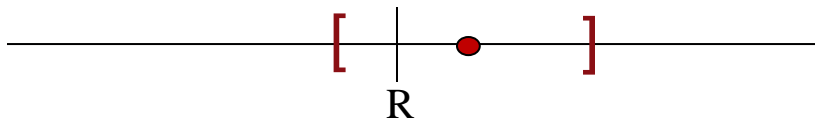
# Überblick

- 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:

- ◆  $l(y, y') = 0$ , wenn  $y = y'$ ; 1, sonst

- → Risiko = Fehlerwahrscheinlichkeit.

- ◆ 
$$\begin{aligned} R(f) &= \int \ell(y, f(\mathbf{x})) p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \int [[y \neq f(\mathbf{x})]] p(\mathbf{x}, y) d\mathbf{x} dy \\ &= p(y \neq f(\mathbf{x})) \end{aligned}$$

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



# Verteilung für Fehlerschätzer

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

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

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

$$\begin{aligned} P\left(R(f) \leq \hat{R}_T(f) + \varepsilon\right) &= P\left(R(f) - \hat{R}_T(f) \leq \varepsilon\right) & e &= R(f) - \hat{R}(f) \\ &= P\left(\frac{e}{s_{\hat{r}}} \leq \frac{\varepsilon}{s_{\hat{r}}}\right) & P\left(\frac{e}{s_{\hat{r}}} \mid r\right) &\approx N\left(\frac{e}{s_{\hat{r}}} \mid 0,1\right) \\ &\approx \Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right) \end{aligned}$$

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

$$\begin{aligned} P\left(|R(f) - \hat{R}_T(f)| \leq \varepsilon\right) &= 1 - P\left(R(f) - \hat{R}_T(f) > \varepsilon\right) + 1 - P\left(\hat{R}_T(f) - R(f) > \varepsilon\right) \\ &= \dots \\ &\approx 2 \left( 1 - \left( 1 - \Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right) \right) \right) \\ &= 2\Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right) \end{aligned}$$

# Konfidenzintervalle

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

- Einseitiges  $1-\delta$ -Konfidenzintervall: Schranke  $\varepsilon$ , so dass

$$P\left(R(f) \leq \hat{R}_T(f) + \varepsilon\right) \geq 1 - \delta$$

- Zweiseitiges  $1-\delta$ -Konfidenzintervall: Schranke  $\varepsilon$ , so dass

$$P\left(|R(f) - \hat{R}_T(f)| \leq \varepsilon\right) \geq 1 - \delta$$

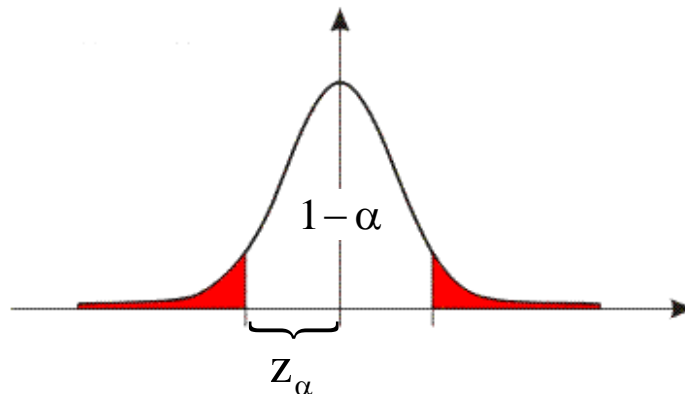
- Bei symmetrischer Verteilung gilt immer:
  - ◆  $\varepsilon$  zu einseitigem  $1-\delta$ -Konfidenzintervall  
=  $\varepsilon$  zu zweiseitigem  $1-\delta/2$ -Konfidenzintervall.

# Konfidenzintervalle

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

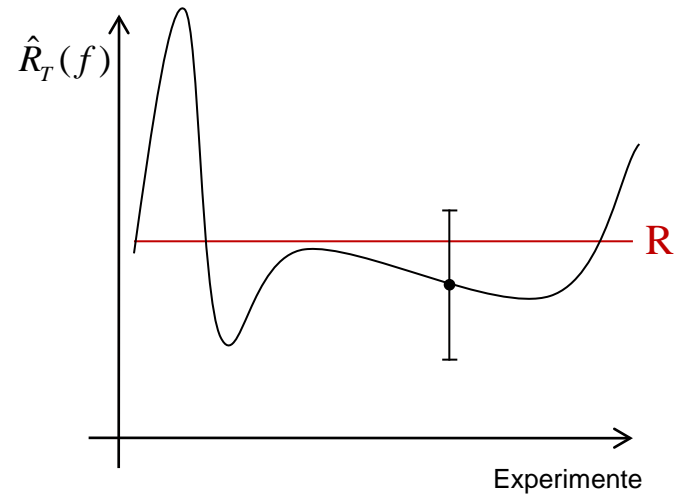
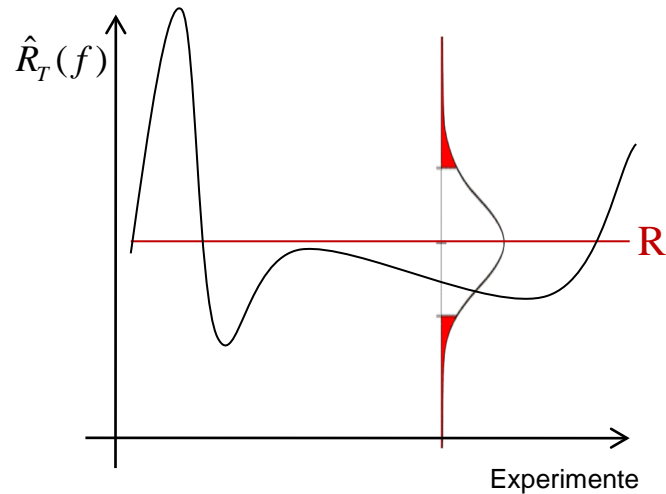
$$P\left(|R(f) - \hat{R}_n(f)| > z_\alpha\right) = 1 - \alpha$$

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



# Konfidenzintervalle

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



# Students t-Verteilung

- Empirisches Risiko annähernd normalverteilt:

- ◆ 
$$p(\hat{R}_T(f) = \hat{r} | r) = B(m\hat{r} | r, m)$$
$$\approx N\left(\hat{r} | r, \frac{r(1-r)}{m}\right)$$
$$= N\left(\frac{\hat{r} - r}{\sigma_{\hat{r}}} | 0, 1\right)$$
 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:

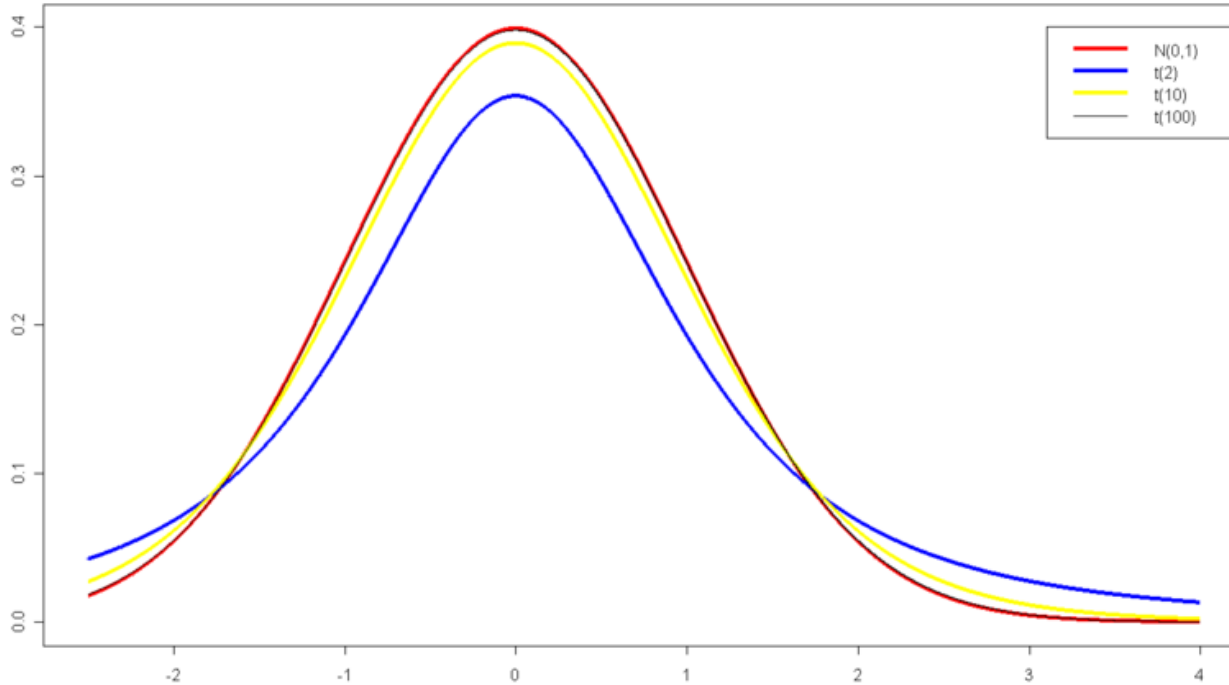
- ◆  $s_{\hat{r}}^2 = \frac{\hat{r}(1-\hat{r})}{m-1}, \quad s_{\hat{r}} = \sqrt{\frac{\hat{r}(1-\hat{r})}{m-1}}$

- 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

Dichtefunktionen von t-verteilten Zufallsgrößen mit unterschiedlichen Freiheitsgraden



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

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

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

# Überblick

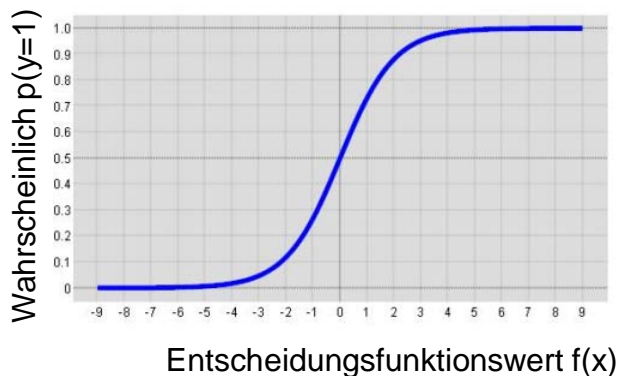
- Wiederholung: Hypothesenbewertung
  - ◆ Verfahren
  - ◆ Anwendungen
  - ◆ Konfidenzintervalle
- ROC-Analyse
- Statistische Tests
  - ◆ Vorzeichen-, Wald-, t- und Pearsons –Test
  - ◆ p-Wert

# Klassifikator / Entscheidungsfunktion

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

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

- 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  $f(\mathbf{x})$ ?
- Allgemeine Lösung:

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

- Der Wert für  $\theta$  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  $\theta$
  - ◆ Fläche unter ROC-Kurve =  $P(\text{positives Beispiel hat höheren f-Wert als negatives Beispiel})$

# ROC-Analyse

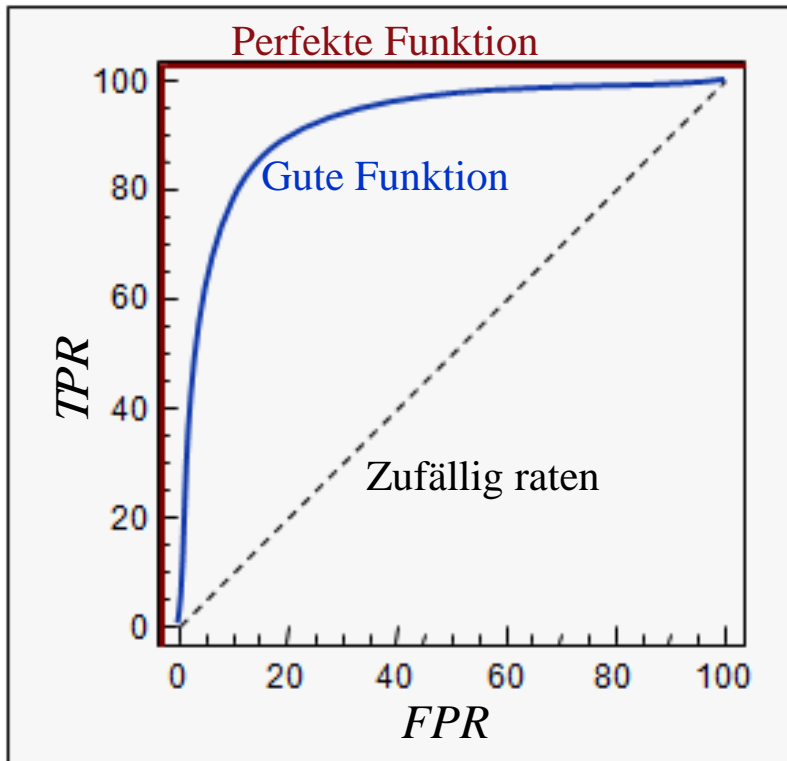
- Entscheidungsfunktion + Schwellwert = Klassifikator

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

- ◆ 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

$$FPR = \frac{FP}{N}$$

$$TPR = \frac{TP}{P}$$

$$N = FP + TN$$

$$P = TP + FN$$

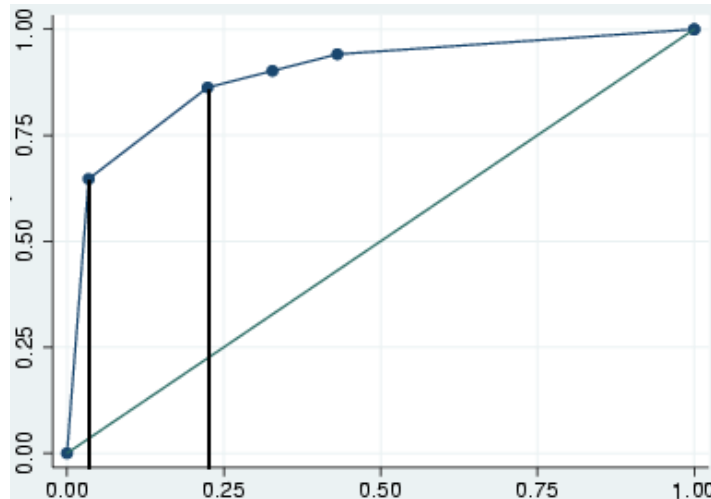
# Bestimmen der ROC-Kurve von $f$

- Annahme: kein  $f(\mathbf{x}) = f(\mathbf{x}')$  für  $\mathbf{x} \neq \mathbf{x}'$ .
- Generiere Liste  $L$  aller Instanzen  $\mathbf{x}$ , absteigend sortiert nach  $f(\mathbf{x})$
- $P$  = Anzahl positiver Instanzen,  $N$  = Anzahl negativer Instanzen
- $TP = FP = 0$
- Für  $i = 1$  bis  $\text{Länge}(L)$ 
  - ◆  $\mathbf{x} = i$ -tes Element von  $L$
  - ◆ Wenn  $\mathbf{x}$  positive Instanz:  $\text{increment}(TP)$
  - ◆ Wenn  $\mathbf{x}$  negative Instanz:  $\text{increment}(FP)$
  - ◆ Zeichne neuen Punkt mit Koordination  $(FP/N, TP/P)$



# Flächeninhalt der ROC-Kurve

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

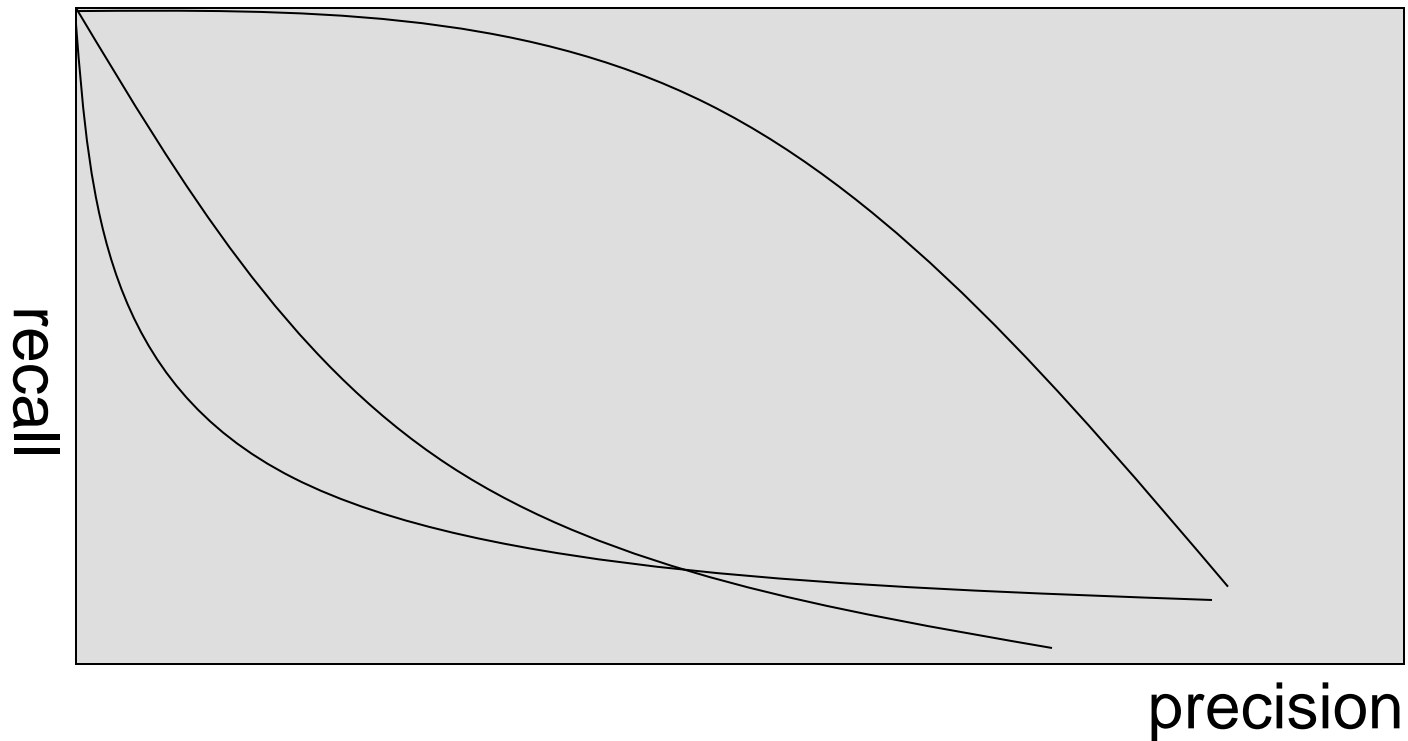


- $x_+$  = zufällig gezogenes Positivbeispiel
- $x_-$  = zufällig gezogenes Negativbeispiel
- Theorem:  $AUC = P(f(x_+) > f(x_-))$ .

# 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  $\theta$

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

- ◆ Precision-Recall-Breakeven-Point: Es gibt einen Punkt  $\theta$  auf der Kurve für den gilt  $\text{Precision}(\theta) = \text{Recall}(\theta) =: \text{PRBEP}$

# Evaluation von Hypothesen: Zusammenfassung

- Verlustfunktion, Risiko
- Empirisches Risiko → 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  $\alpha$ .
- 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  $\alpha$  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
  - ◆  $P(\text{Beobachteter Unterschied} \mid \text{Nullhypothese}) < \alpha$ , 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  $\mathbf{x} \in \mathcal{X}$  einer Zufallsvariable  $\mathbf{X}$  eine *begründete* Entscheidung über die Gültigkeit oder Ungültigkeit einer Hypothese zu treffen
- Formal:

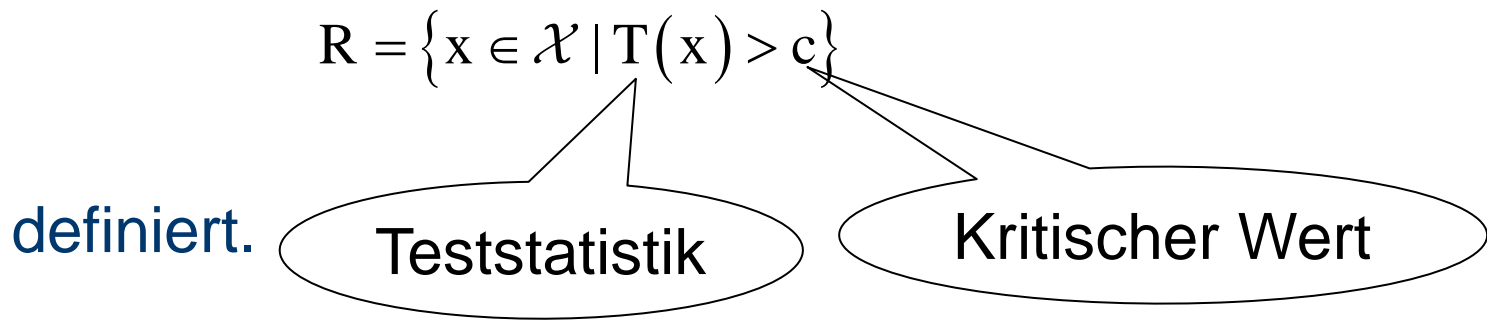
$$h_0 : \theta \in \Theta_0 \text{ vs. } h_1 : \theta \in \Theta_1$$



Nullhypothese

# Statistische Tests

- Im Allgemeinen ist ein statistischer Tests durch seinen kritischen Bereich



- Wenn  $X \in R$ , lehnen wir die Nullhypothese 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:  $h_0 : \theta \leq \theta_0$  vs.  $h_1 : \theta > \theta_0$
  - ◆ zweiseitiger Test:  $h_0 : \theta = \theta_0$  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:  $\alpha = \sup_{\theta \in \Theta_0} P(X \in R \mid \theta)$

# p-Wert

- Die Aussage „Nullhypothese abgelehnt“ ist nicht sehr informativ
- p-Wert: kleinste Signifikanz-Niveau  $\alpha$ , 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  $\alpha$ , 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	8	3,8	4	9,1	20,9

- Gibt es Unterschiede zwischen den Medikamenten in der Wirkung?



# Vorzeichen-Test

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

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

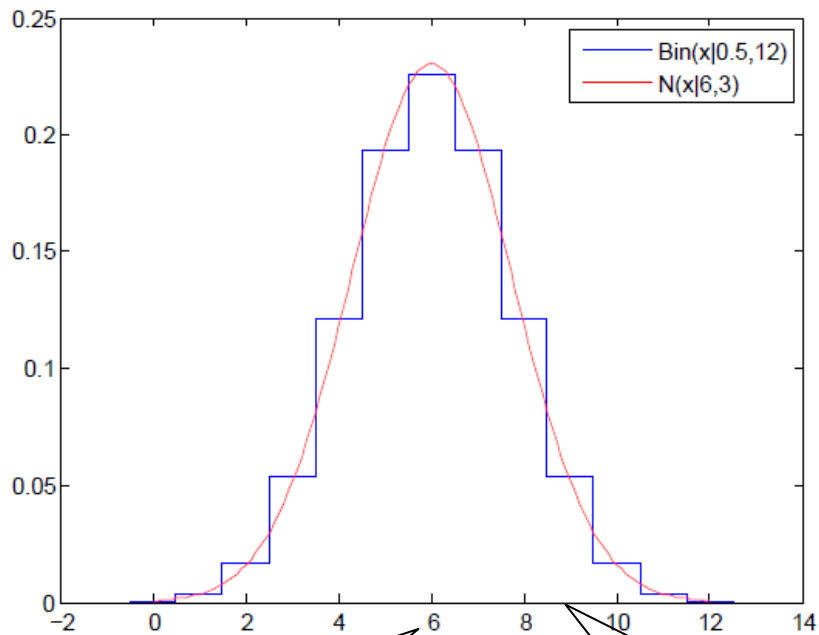
- Lehne Nullhypothese ab, gdw.  $T(\mathbf{x}) > c$

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

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

# Vorzeichen-Test

- Lehne Nullhypothese ab, gdw.  $T(\mathbf{x}) > c$



unter  $h_0$  binomial-verteilt

$$h_0 : m = \mu_0$$

Wie wahrscheinlich ist  
 $T(\mathbf{x}) = 9$ ?

# 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 Medikamenten in der Wirkung?
  - ◆ Nullhypothese: beide gleich  $h_0 : \mu = 6$
- $T(\mathbf{x}) = 9$

# 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

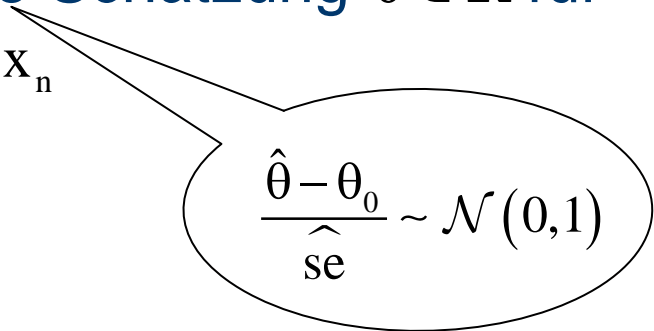
- Nullhypothese  $h_0 : m = 6$

- $T(\mathbf{x}) = 9$

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

# Wald-Test

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


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

- $h_0 : \theta = \theta_0$  vs.  $h_1 : \theta \neq \theta_0$

- Lehne Nullhypothese ab, gdw.  $T(\mathbf{x}) > c$

- ◆  $T(\mathbf{x}) = \frac{|\hat{\theta} - \theta_0|}{\widehat{se}}$
- ◆  $c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$



unter  $h_0$  normalverteilt

# Wald-Test

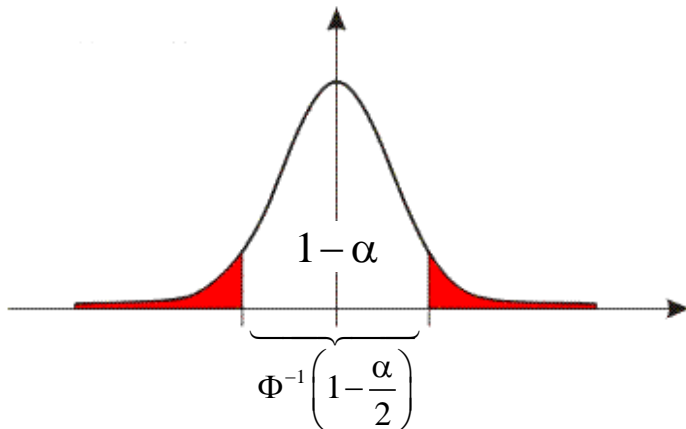
- Wald-Test: Lehne Nullhypothese ab, gdw.

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

- Wald-Test hat Signifikanz-Niveau  $\alpha$

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

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



# 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 Medikamenten in der Wirkung?
  - ◆ Nullhypothese: beide gleich  $h_0 : \theta = 0$

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

# 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  $h_0 : \theta = 0$
- $T(\mathbf{x}) \approx 2,168$
- $p\text{-Wert} = p(T(X) > T(X_n) | h_0)$   
 $= p(Z > 2,168) + p(Z < -2,168), Z \sim N(0;1)$   
 $= 2\text{NormCDF}(-2,168 | 0;1)$   
 $\approx 3\%$



# t-Test

- Seien  $x_1, \dots, x_n$  unabhängig normalverteilt mit Erwartungswert  $\mu$  und unbekannter Varianz

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

- Lehne Nullhypothese ab, gdw.  $T(\mathbf{x}) > c$

- ◆  $T(\mathbf{x}) = \frac{|\bar{X}_n - \mu_0|}{\widehat{se}}$ ,  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n x_i$

- ◆  $c = F_{n-1}^{-1} \left( 1 - \frac{\alpha}{2} \right)$

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 Medikamenten in der Wirkung?
  - ◆ Nullhypothese: beide gleich  $h_0 : \theta = 0$

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

# 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  $h_0 : \theta = 0$
- $T(\mathbf{x}) \approx 2,168$
- $p\text{-Wert} = p(T(\mathbf{X}) > T(\mathbf{X}_n) | h_0)$   
 $= p(Z > 2,168) + p(Z < -2,168), Z \sim F_{n-1}(0)$   
 $= 2tCDF_{n-1}(-2,168)$   
 $\approx 5,3\%$

# Pearsons $\chi^2$ -Test

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

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

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

- Lehne Nullhypothese ab, gdw.  $T(\mathbf{x}) > c$

$$\diamond T(\mathbf{x}) = \sum_{j=1}^k \frac{(x_1^j - \mu_j)^2}{\mu_j}$$

$$\diamond c = \left[ \chi_{k-1}^2 \right]^{-1} \left( 1 - \frac{\alpha}{2} \right)$$

unter  $h_0$   $\chi^2$ -verteilt  
( $k-1$  Freiheitsgrade)

# Dualität

- Ein Test mit Signifikanzniveau  $\alpha$  verwirft die Nullhypothese  $h_0 : \mu = \mu_0$ , genau dann nicht, wenn  $\mu_0$  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  $\alpha$  fälschlicher Weise abgelehnt werden

- Verschiedene Tests: Vorzeichen-Test, Wald-Test, t-Test, Pearson  $\chi^2$