Artificial Intelligence - Data Science

Grundlagen des Maschinellen Lernens und Fehlerabschätzung

Alexandra Posekany

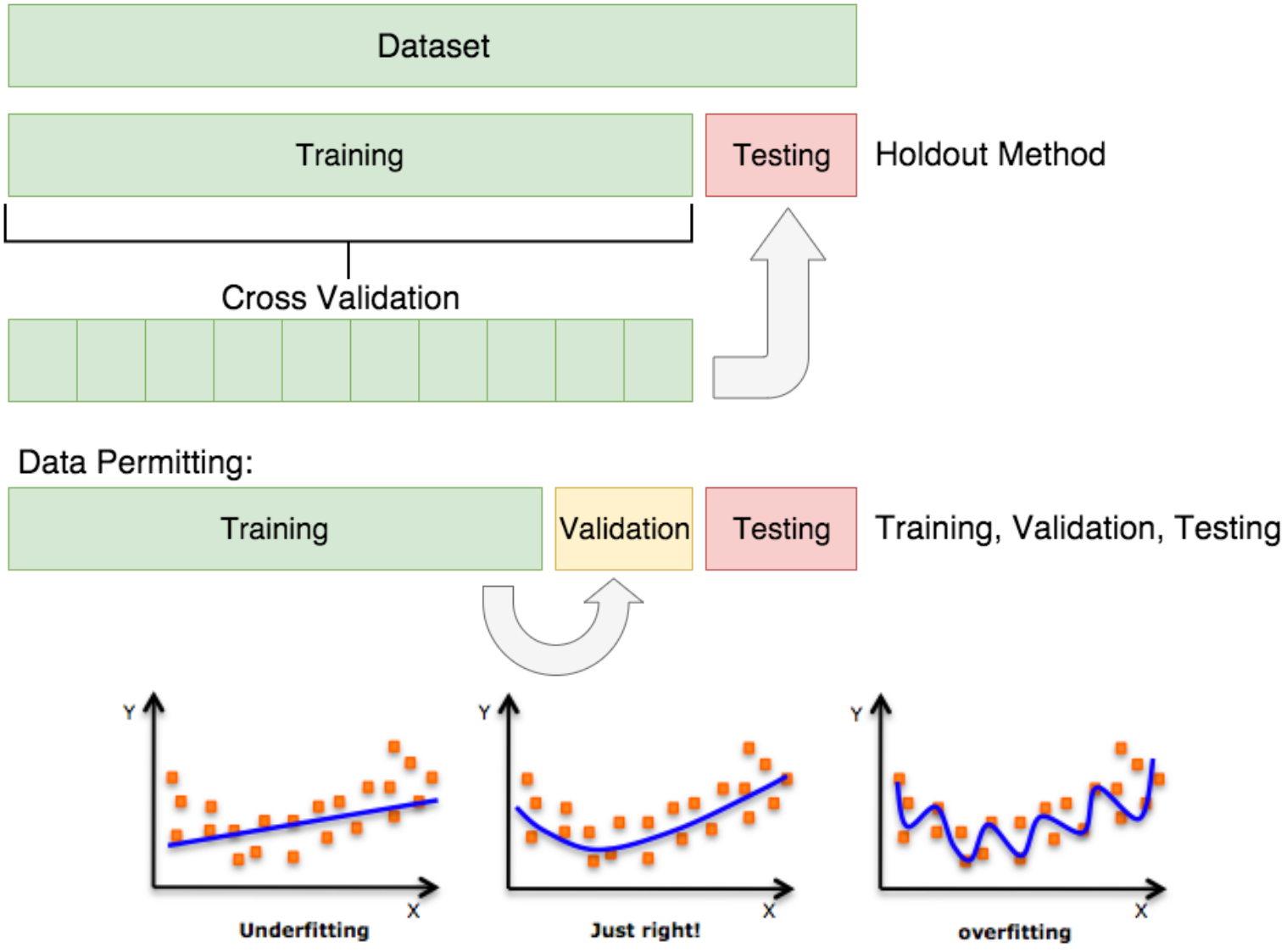
WS 2020

Was erwartet euch in diesem Semester?

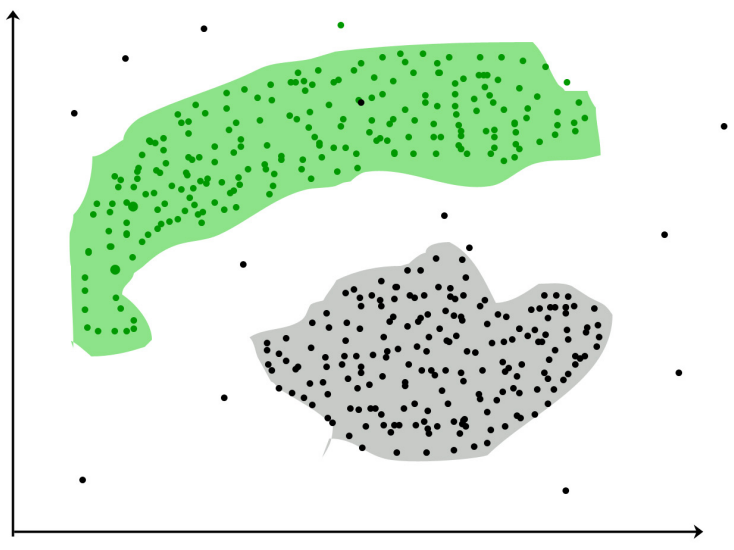
SYT - AI - "Data Science"

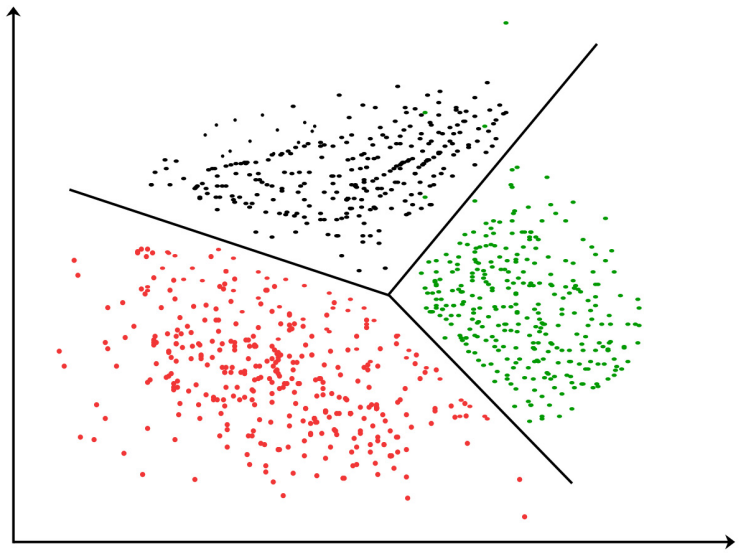
- **Ablauf von Datenanalyse oder maschinellen Lernprozessen**
  durch Exploration Daten in Trainings- und Testdatensätze aufteilen; Kenntnis von In-sample Schätzungen und Prädiktionen und Out-of-sample Schätzungen und Prädiktionen; Qualitätsprüfung von Algorithmen; Probleme der Modellanpassung
- **Mustererkennung**
  Kenntnis von Konzepten von Distanzmaßen (z.B. lineare Diskriminanzanalyse, Cluster Analyse (k-nearest neighbors, model based Clustering), Support Vector Machines)
- **Methoden im Rahmen aktueller Anwendungsgebiete umsetzen**
  Fallstudie: Anwendung von Methoden der Exploration, Modellierung und Qualitätsprüfung anhand eines realen Datensatzes oder einer realen Fallstudie

Was erwartet euch im Fachbereich? Ablauf von Datenanalyse oder maschinellen Lernprozessen



Was erwartet euch im Fachbereich? Mustererkennung





Was erwartet euch im Fachbereich? Fallstudien

Entweder im Rahmen von Projekten oder mit einem vorgegebenen Datensatz bzw. Datensatz eigener Wahl

1. Durchführung von Exploration (tabellarisch/numerisch und graphisch)
2.
   - Anpassung von geeigneten Modellen unter Verwendung von maschinellen Lernmethoden (Training, Validation, Test)
   - Anwendung von geeigneten Hypothesentests unter Verwendung von vorgefertigten oder Sampling-basierten Methoden
3. optional: als interaktive Application mit shiny

WH: Was ist AI, ML, DL, DA, DM?

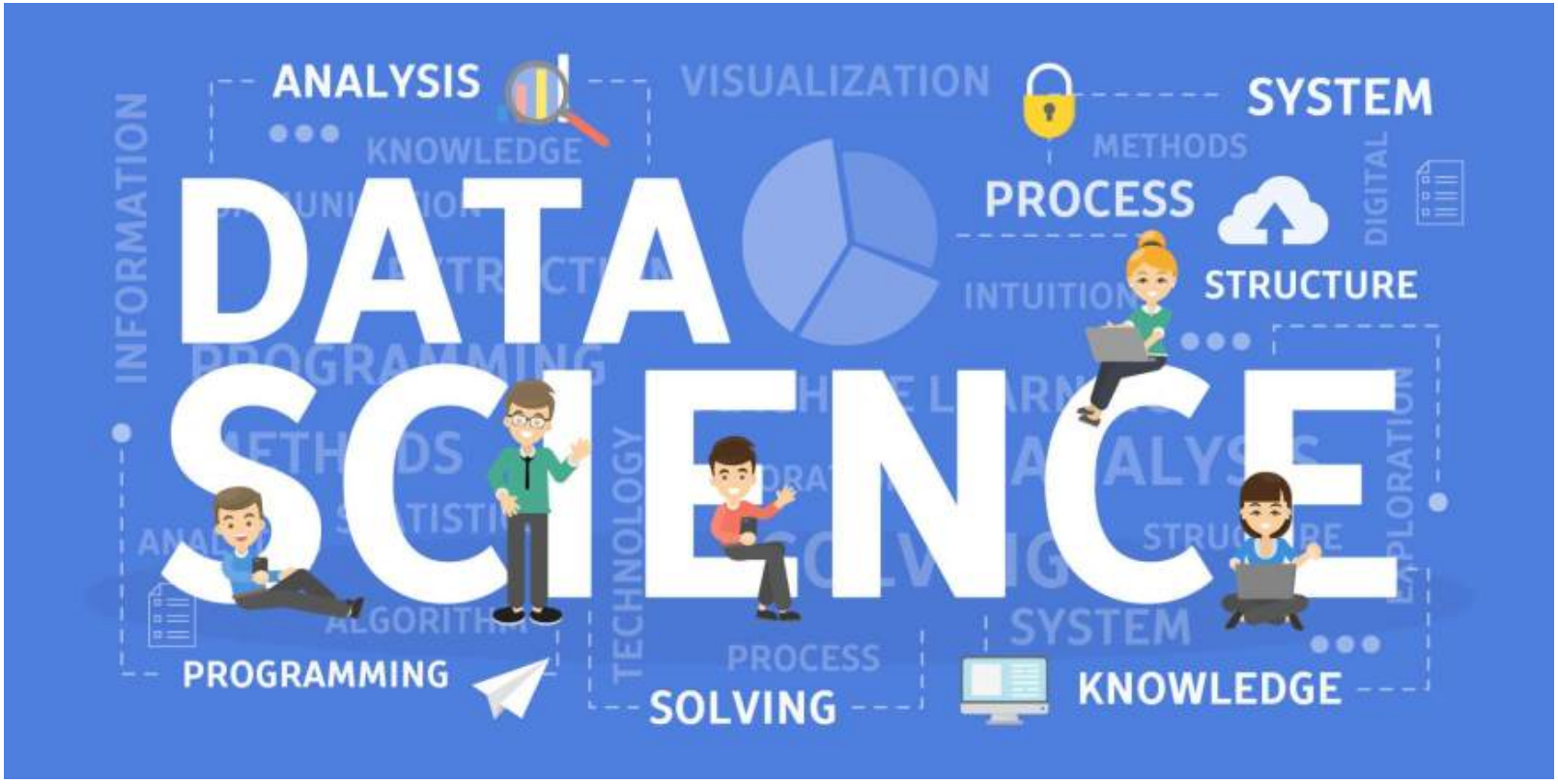
Artificial Intelligence (AI)
Teilbereich der Informatik mit Automatisierung von Prozessen als Ziel, wobei kognitive Prozesse als Vorbild dienen. Nachahmung menschlicher Intelligenzleistungen wie "Lernen" oder "Probleme zu lösen".

Machine Learning (ML)
Teilbereich der AI zum Erkennen von Mustern und Gesetzmäßigkeiten basierend auf Datenbanken und Algorithmen

Deep Learning (DL)
Optimierung künstlicher neuronaler Netze mit mehreren Zwischenschichten zwischen Eingabe- und Ausgabeschicht.

Data Science (DS)
Verknüpfung von Statistik mit Softwareentwicklung, Pipelining von Datenbanksystemen und Maschinellem Lernen zur Erkennung von Mustern und Gesetzmäßigkeiten in großen Datenmengen

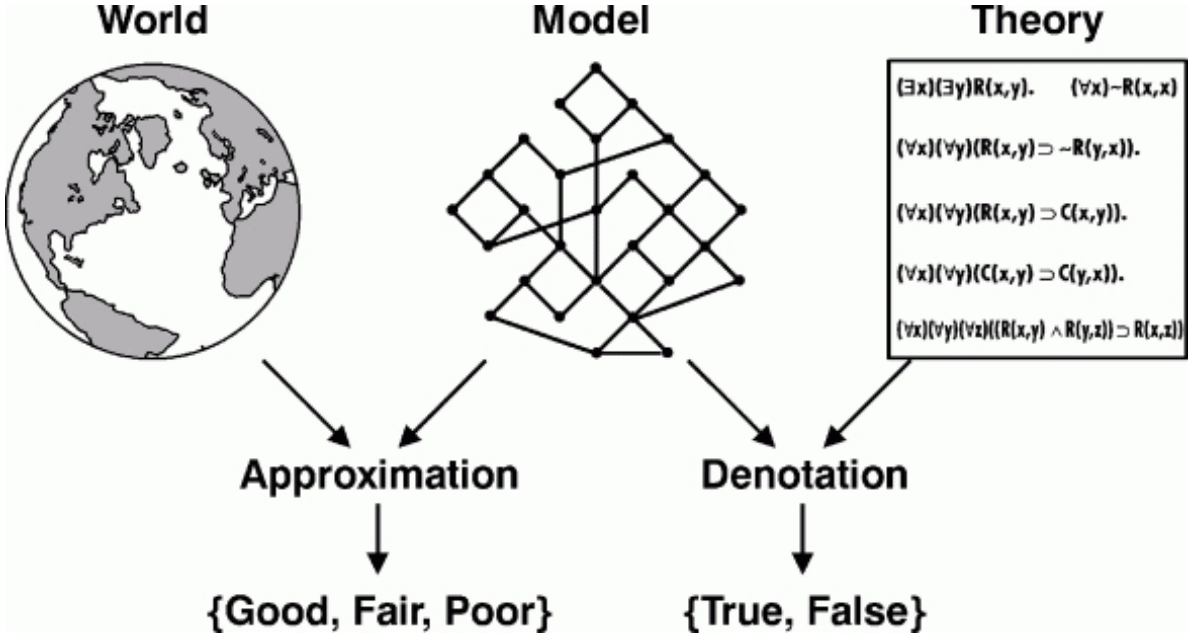
WH: Was ist also Data Science?



WH: Moral of the story

Essentially, all models are wrong,
but some are useful.

[George Box]



Eindimensionale Modelle im Überblick

Wir betrachten die uns bekannten Modelle mit einer Einflussvariable

$y \sim 1$ triviales Modell (lm)
(1 Mittelwert für die gesamten Daten)

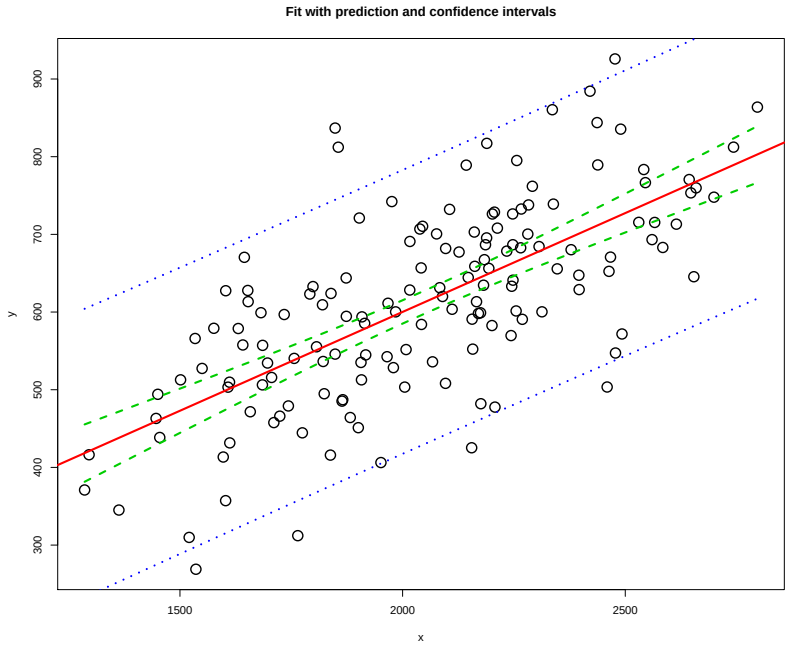
$y \sim a$ Einweg-Varianzanalyse (lm)
(2 Mittelwerte für 2 Kategorien der Variable a)

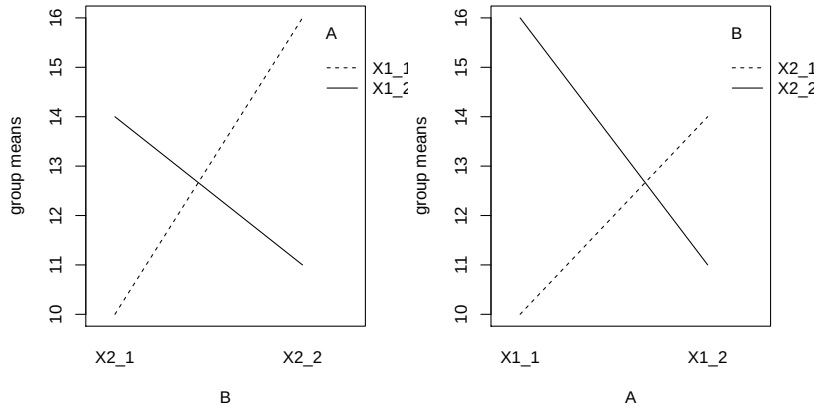
$y \sim x$ Lineare Regression (lm)
(1 Regressionsgerade $y = \alpha + \beta \cdot x$)

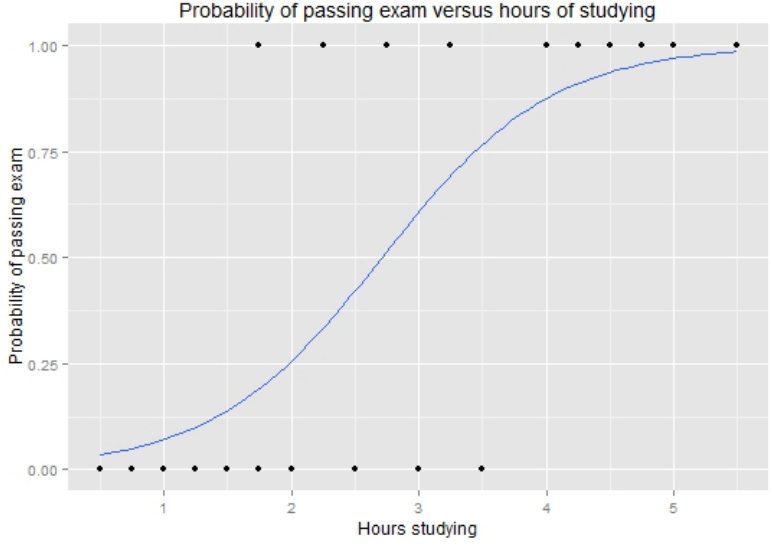
$y \sim x$ Regression Tree
Aufplittung in Teilintervalle von x und
Zuordnung von mittleren geschätzten Werten

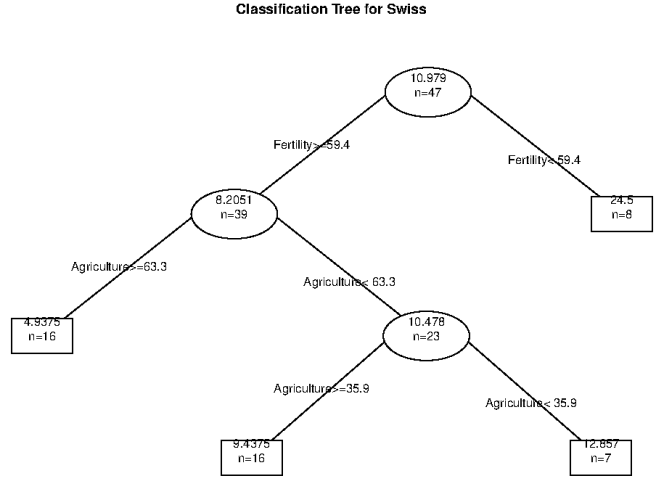
$y \sim x$ Logistische Regression (glm(,family=binomial("logit")))
(Klassifikation der binären Variable y mithilfe von Daten x)

Modelle und Szenarien









Zweidimensionale Modelle im Überblick

Wir betrachten die uns bekannten Modelle mit zwei Einflussvariablen

$y \sim a1 + a2$ Zweiweg-Varianzanalyse (lm)
(4 Mittelwerte für je 2 Kategorien der Variable a1
und je 2 Kategorien der Variable a2)

$y \sim a1 * a2$ Zweiweg-Varianzanalyse (lm)
(4 Mittelwerte für je 2 Kategorien der Variable a1
und je 2 Kategorien der Variable a2
plus Interaktionen zwischen a1 und a2)

$y \sim x1 + x2$ Lineare Regression (lm)
(1 Regressionsebene $y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$)

$y \sim x1 + x2$ Regression Tree
Aufplittung in Teilintervalle von x1 und x2
und Zuordnung von mittleren geschätzten Werten

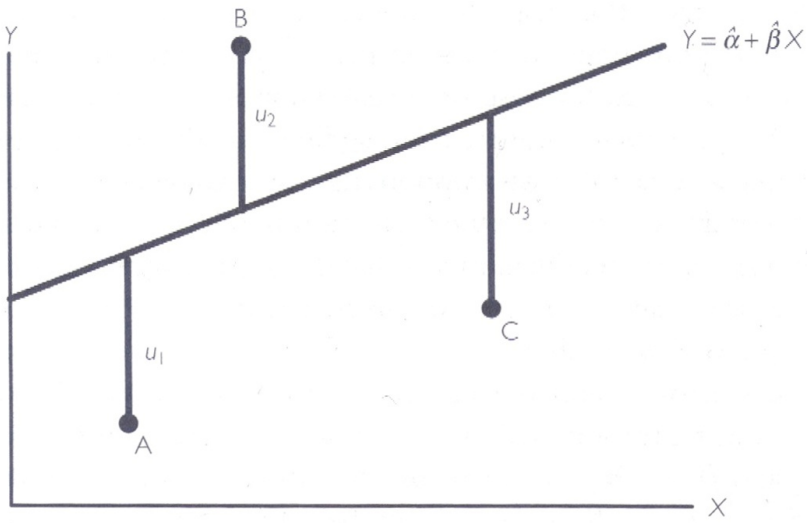
$y \sim x1 + x2$ Logistische Regression (glm(,family=binomial("logit")))
(Klassifikation der binären Variable y
mithilfe von Daten x1 und x2)

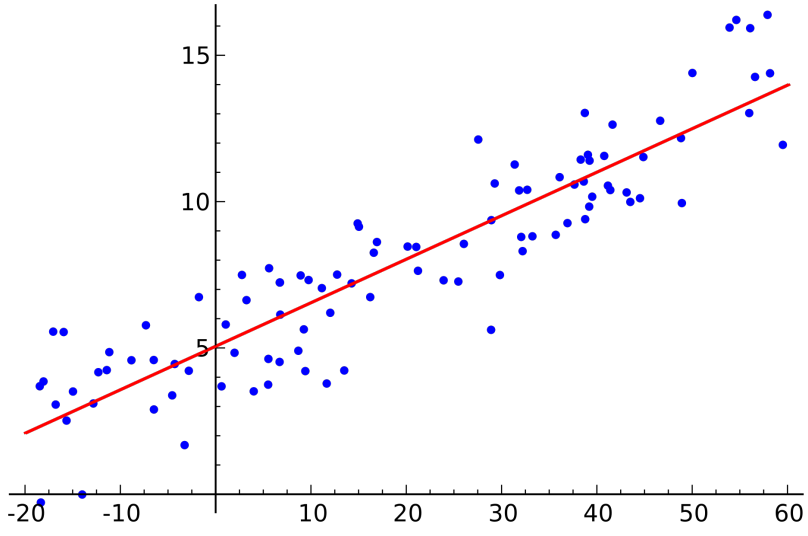
Woher kommen die Schätzungen: Kleinste Quadrate Schätzung

Die Kleinste Quadrate Schätzung (Least Squares Estimation) der linearen, nichtlinearen und logistischen Regression, Regression Trees und ANOVA bedeutet, dass die Summe der quadratischen Residuen (residual sum of squares RSS)

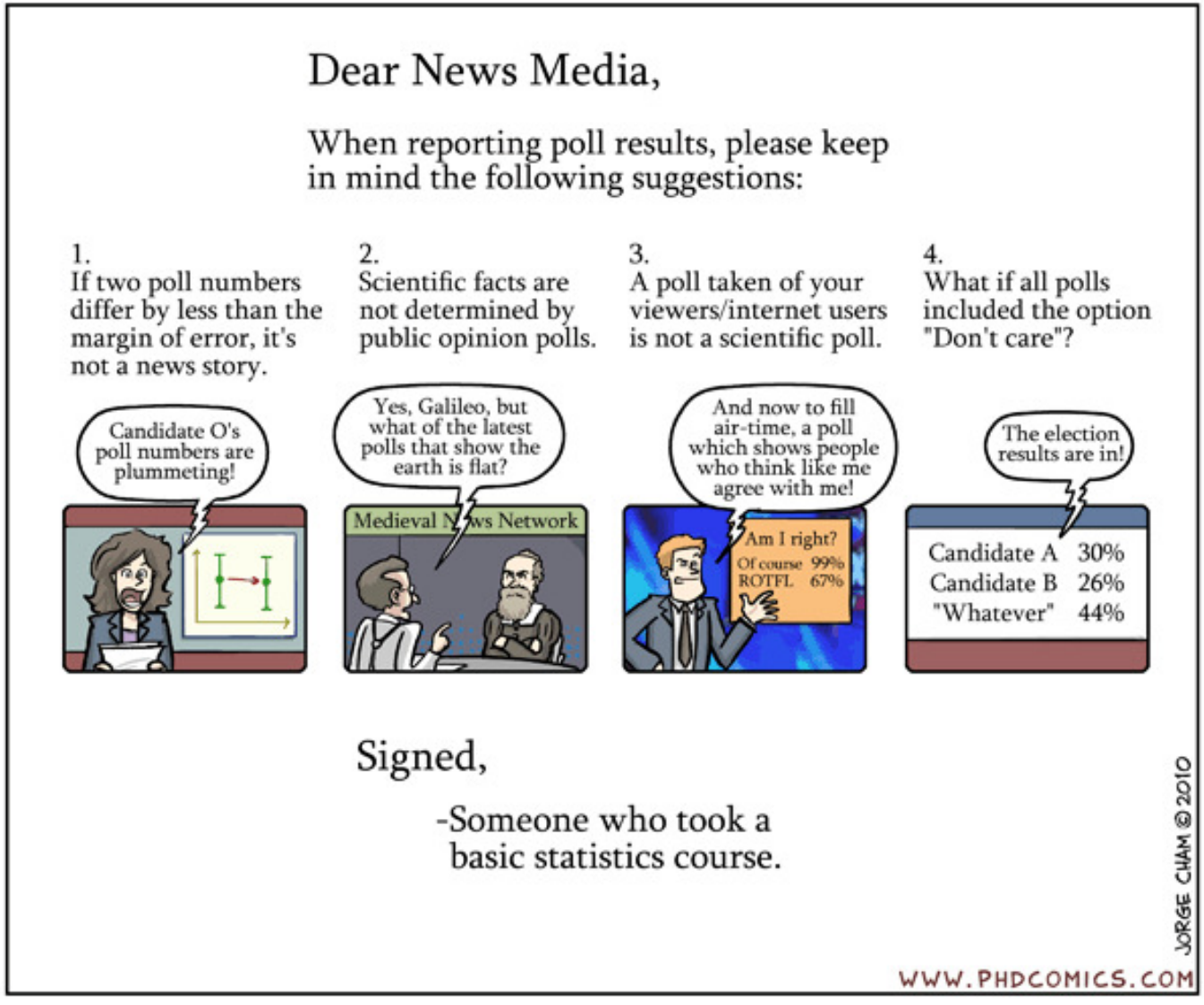
$$e_i = (y_i - \hat{\alpha} - \hat{\beta} y_i)^2, i = 1, 2, \ldots, N.$$

minimiert wird, um $\hat{\alpha}$ und $\hat{\beta}$ zu schätzen.





Was wollen wir mit Fehlerdarstellungen und Intervallen?



Arten von Fehlerberechnungen

- Daten können nicht exakt gemessen werden
- Modelle und Formeln sind nur eine Näherung der Realität
- Simulationen sind grundsätzlich fehlerbehaftet
- **reduzierbaren Modellierungsfehler**

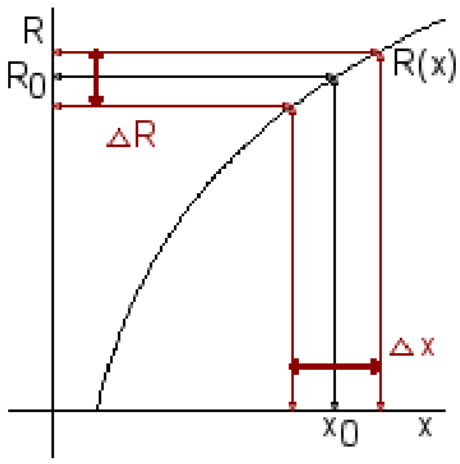
  Methoden, Modelle oder Algorithmen verbessern
- **nichtreduzierbaren Beobachtungsfehler**

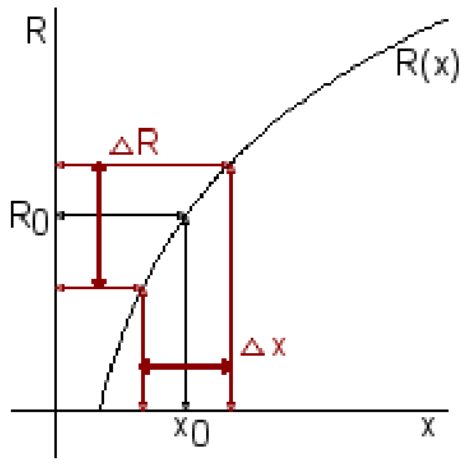
  Heisenbergsche Unschärferelation; Börsenkurse, Körpergrößen inherent zufallsbehaftet

Arten von Fehlerberechnungen

Gauß'sche Fehlerrechnung = lineare Fehlerfortpflanzung Die lineare Fehlerrechnung nähert an, wie sich die Schwankung der Eingangswerte auf die Schwankung eines aus einer (deterministischen) Berechnung resultierenden Wertes (nichtreduzierbare Fehler) auswirkt.

$$\Delta Res(x_1, x_2, \ldots) = \frac{\partial R}{\partial x_1}\Delta x_1 + \frac{\partial R}{\partial x_2}\Delta x_2 + \ldots$$

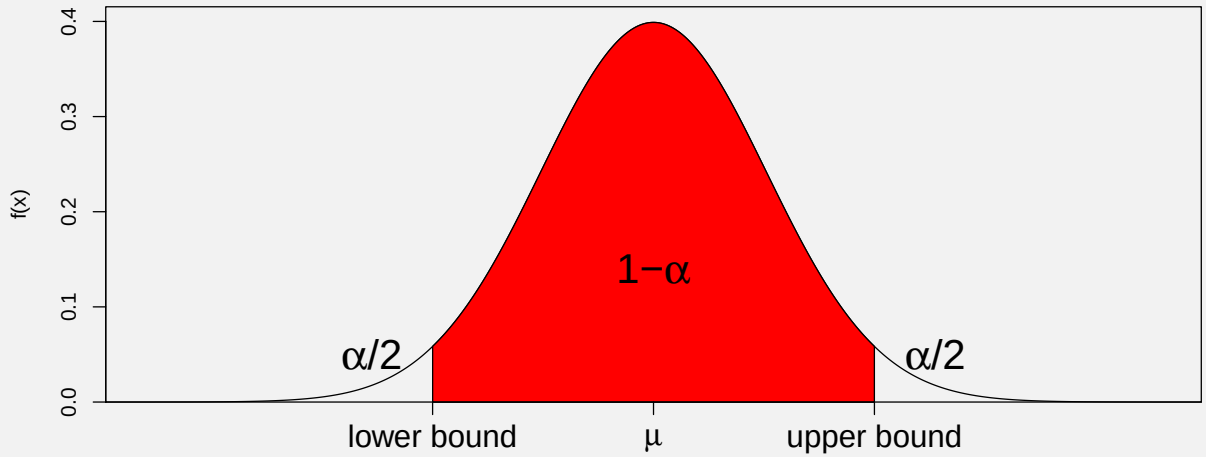




Arten von Fehlerberechnungen

Konfidenzintervalle (= Vertrauensbereiche)

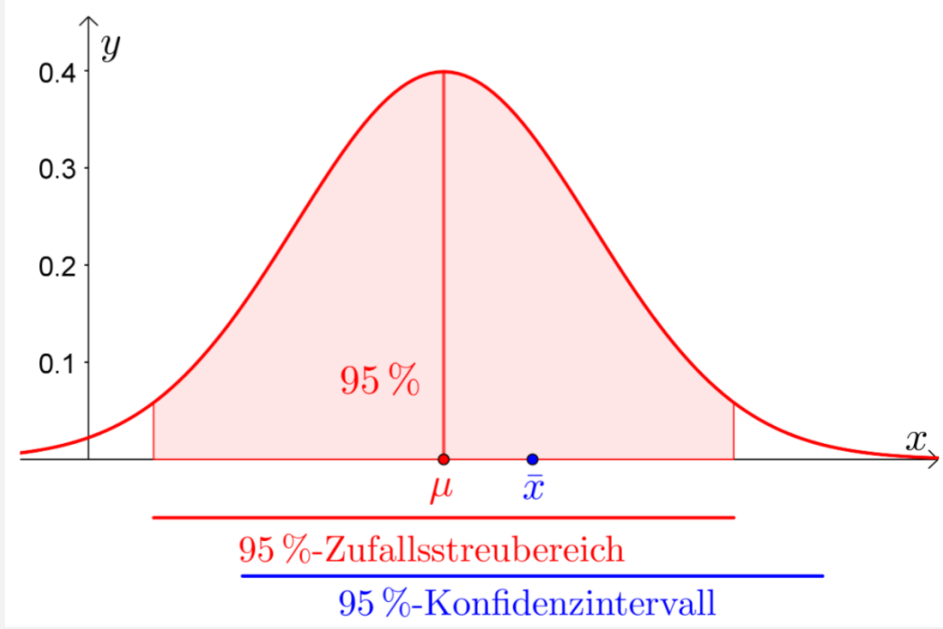
Wenn ein Wert (arithmetischer Mittelwert, Proportion, ...) aus einer Stichprobe geschätzt wird, um auf den zugrunde liegenden "wahren Wert" des Zentrums, Anteils etc. Rückschlüsse zu ziehen, dann gibt der Konfidenzbereich den Bereich an, der diesen Wert mit einer Wahrscheinlichkeit von 95 % oder 99 %, allgemein $1 - \alpha$, überdeckt. Dann ist die Wahrscheinlichkeit, nicht zu überdecken, also den wahren Wert nicht zu enthalten, gleich α, etwa $\alpha = 5\%$.



Arten von Fehlerberechnungen

Prädiktionsintervalle (= Vorhersagebereiche, Zufallsstreubereiche)

Wenn ein Wert (arithmetischer Mittelwert, Proportion, ...) aus einem bekannten Modell oder auf Basis bekannter gemessener Werte vorhergesagt werden soll, dann gibt der Vorhersagebereich den Bereich an, in dem sich der zu messende Werte mit einer Wahrscheinlichkeit von 95 % oder 99 % befindet.



Was steckt hinter Vertrauensbereichen?

Wahrscheinlichkeitsrechnung und Statistik - konkret die **Normalverteilung**

Die Normalverteilung hat eine Wahrscheinlichkeitsdichte von

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

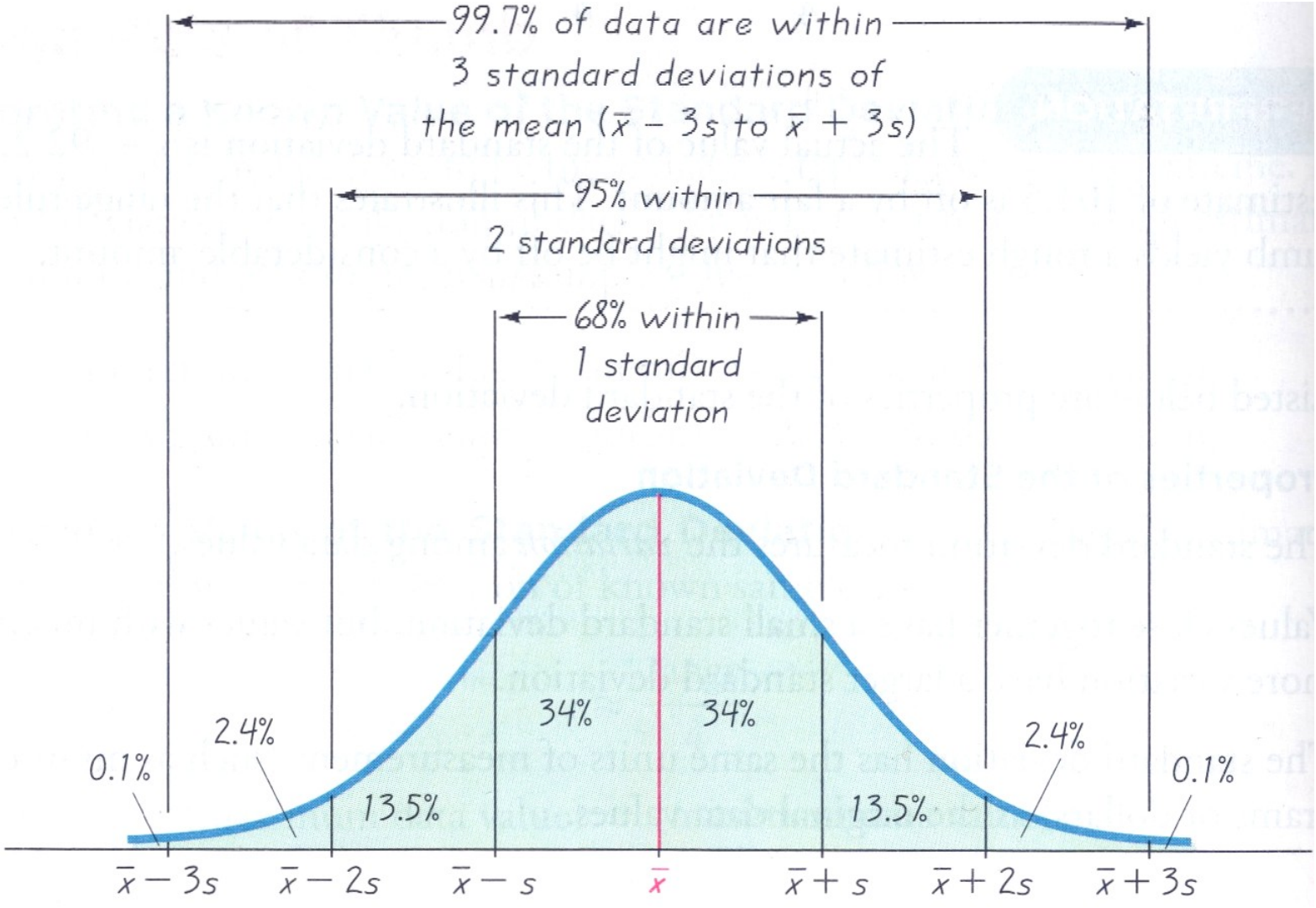
mit den Parametern, die dem Erwartungswert μ und der Varianz σ^2 bzw. Standardabweichung σ der Daten entsprechen:

$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \mathbb{V}[X] &= \sigma^2 \end{aligned}$$

Die Kurve zu dieser Dichtefunktion heißt “Gauß’sche Glockenkurve’ ’

Gauß'sche Glockenkurve

Faustregel (68-95-99.7 Regel) der Normalverteilung

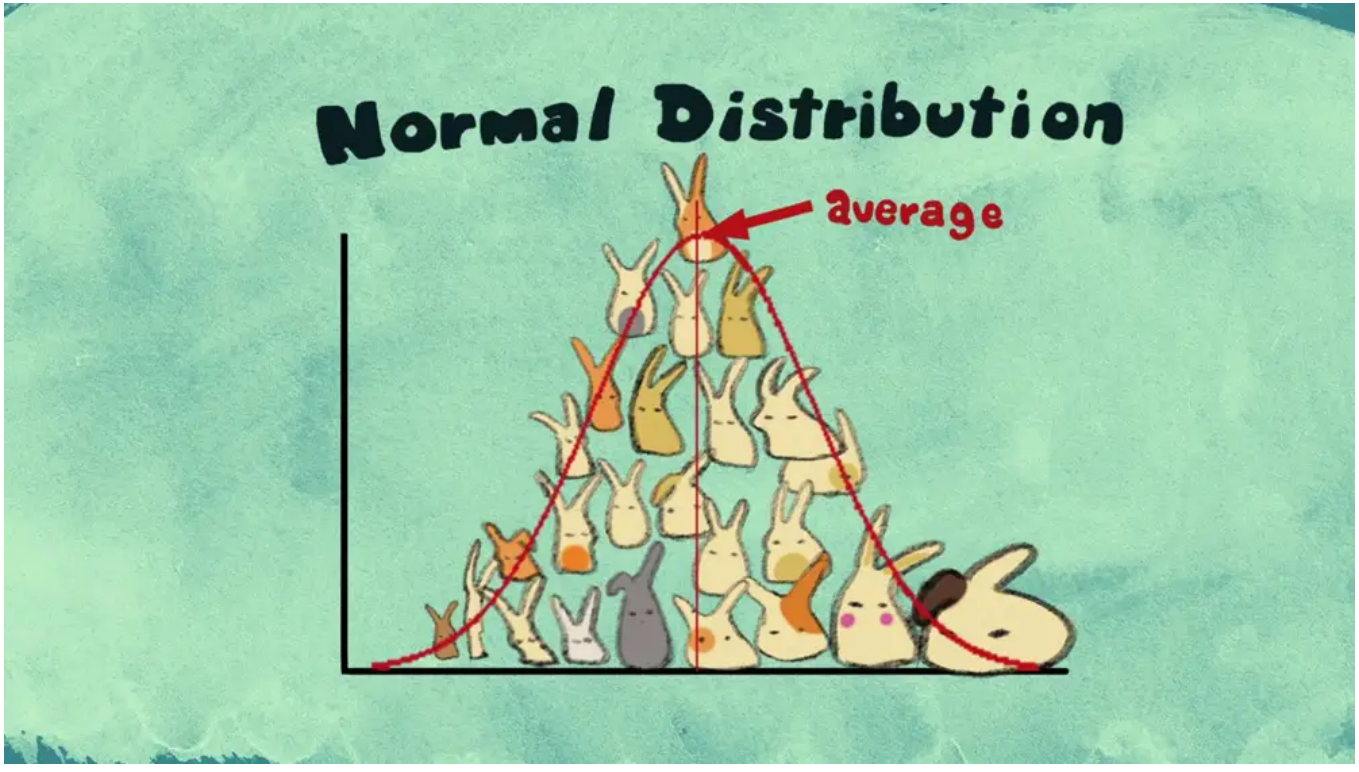


Zentraler Grenzverteilungssatz

Unabhängig davon, wie die Daten in Wahrheit verteilt sind, ist der arithmetische Mittelwert der Daten immer um den wahren zugrundeliegenden erwarteten Zentrumswert (Erwartungswert) μ mit einer Normalverteilung verteilt, deren Standardabweichung $\frac{\sigma}{\sqrt{n}}$ umso kleiner wird je mehr Daten man beobachtet.

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Video: Bunnies and Dragons



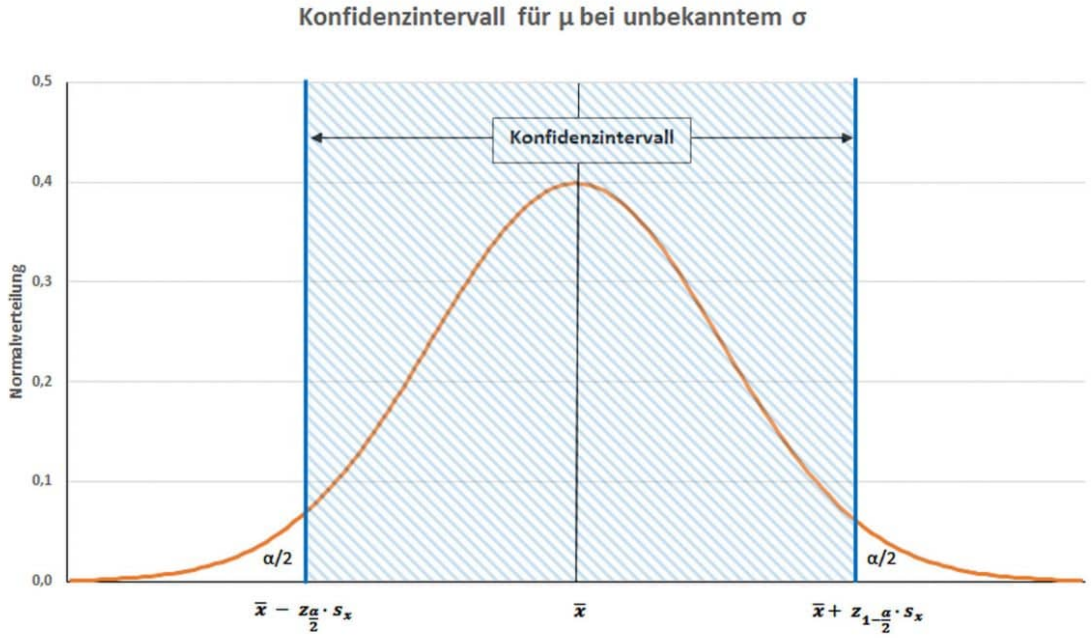
https://youtu.be/L67S87kAXPI

Konfidenzintervall für den Mittelwert

Für den arithmetischen Mittelwert $\bar{x}$ gilt als Schätzung für den erwarteten Wert der Lage μ, dass dieser mit Wahrscheinlichkeit 95% vom Intervall

$$\left(\bar{x} - 1,96 \cdot \frac{standardabweichung}{\sqrt{n}}; \bar{x} + 1,96 \cdot \frac{standardabweichung}{\sqrt{n}}\right)$$

überdeckt wird, die Quantile der Normalverteilung $q_{0.025}$=-1,96 und $q_{0.975}$=1,96 sind.



Berechnung in R

```r
# Konfidenzintervall
mittelwert <- 12.34
standardabweichung <- 1.67
anzahl <- 50
quantil <- qnorm(0.005,mean=0,sd=1)

# Standardisieren
# subtrahieren Mittelwert
# dividieren durch Standardabweichung

mittelwert-quantil*standardabweichung/sqrt(anzahl)
```

```
## [1] 12.94834
```

```r
mittelwert+quantil*standardabweichung/sqrt(anzahl)
```

```
## [1] 11.73166
```

```r
qnorm(0.005,mean=mittelwert,sd=standardabweichung/sqrt(anzahl))
```

```
## [1] 11.73166
```

```r
qnorm(0.995,mean=mittelwert,sd=standardabweichung/sqrt(anzahl))
```

```
## [1] 12.94834
```

Vertrauensbereiche für Kategorienhäufigkeiten

Die Binomialverteilung zählt die Anzahl der "interessanten Ereignisse'' bei n Experimenten, wenn

- nur zwei mögliche Ereignisse - `Erfolg''` und Misserfolg'' - existieren
- die Erfolgswahrscheinlichkeit immer gleichbleibend p ist
- n Experimente, die nur in Erfolg oder Misserfolg ende können, aneinandergereiht werden

  Die Binomialverteilung hat eine Wahrscheinlichkeitsfunktion für die gezählten Erfolge k von

$$\mathbb{P}[X = k] = \left(\begin{array}{c} n \\ k \end{array} \right) \cdot p^k \cdot (1-p)^{n-k}.$$

mit den Parameter, die der Stichprobengröße n und Wahrscheinlichkeit für "interessanten Ereignisse'' p der Daten entsprechen, gilt:

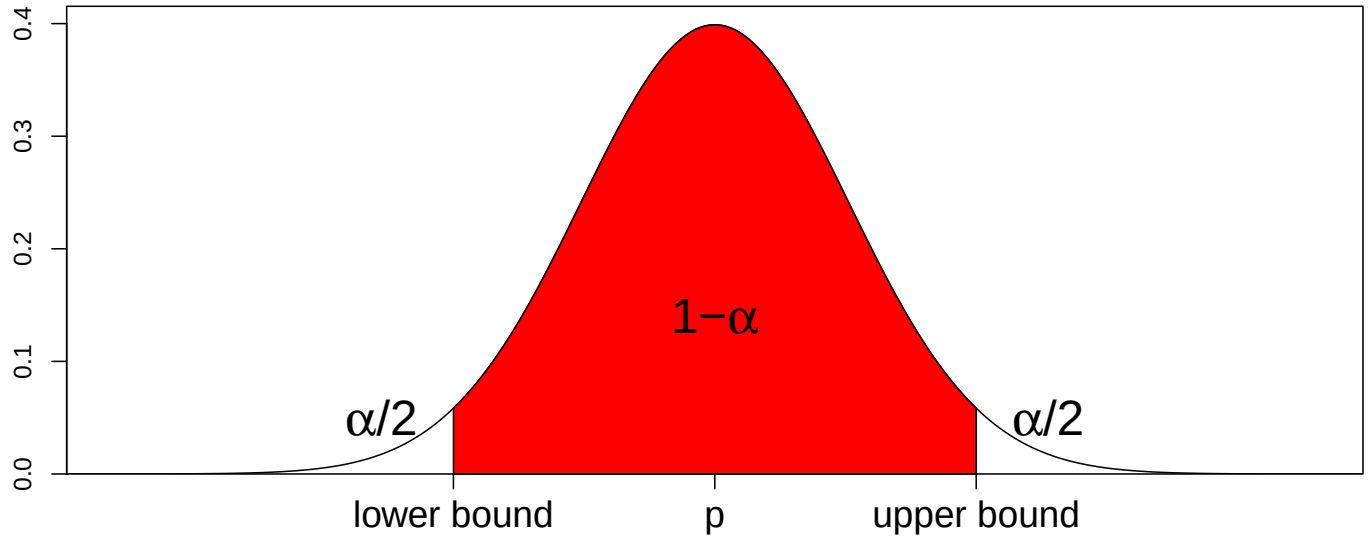
$$\begin{aligned} \mathbb{E}[X] &= n \cdot p \\ \mathbb{V}[X] &= n \cdot p \cdot (1-p) \end{aligned}$$

Konfidenzintervall für Proportionen (relative Häufigkeiten)

Für die Proportion $\hat{p}$ gilt als Schätzung für die Wahrscheinlichkeit einer Kategorie p, dass diese mit Wahrscheinlichkeit 95% vom Intervall

$$\left(\hat{p} - 1{,}96 \cdot \frac{\hat{p} \cdot (1-\hat{p})}{\sqrt{n}}; \hat{p} + 1{,}96 \cdot \frac{\hat{p} \cdot (1-\hat{p})}{\sqrt{n}}\right)$$

überdeckt wird, die Quantile der Normalverteilung $q_{0.025}$=-1,96 und $q_{0.975}$=1,96 sind.



Berechnung in R

```r
# Medikamente

gesamt = 30*10^6
positivewirkung = 25*10^6

quantil = qnorm(0.025,mean = 0,sd = 1)

wirkungswahrscheinlichkeit = positivewirkung/gesamt

wirkungswahrscheinlichkeit + quantil * wirkungswahrscheinlichkeit*(1-wirkungswahrscheinlichkeit)/sqrt(gesa
```

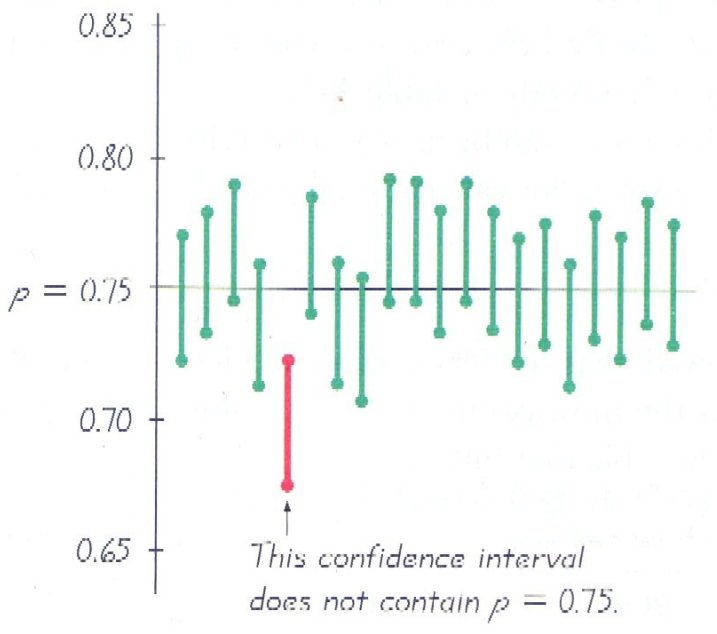
```
## [1] 0.8332836
```

```r
wirkungswahrscheinlichkeit - quantil * wirkungswahrscheinlichkeit*(1-wirkungswahrscheinlichkeit)/sqrt(gesa
```

```
## [1] 0.833383
```

Interpretationen von (95%) Konfidenzintervallen

- **(wiederholte) Stichproben**
  "Würde mehrfach auf dieselbe Weise Stichproben gezogen werden, dann würden in 95% der Fälle die berechneten Konfidenzintervalle den wahren zugrunde liegenden (aber uns unbekannten) Wert überdecken."



Interpretationen von (95%) Konfidenzintervallen

- **Einzelstichprobe**
  "Mit 95%-iger Wahrscheinlichkeit enthält das ermittelte Konifdenzintervall den wahren zugrunde liegenden (aber uns unbekannten) Wert."

- **Akzeptanzbereich eines Hypothesentests**
  "Das Konfidenzintervall enthält alle Werte, die zum zugrunde liegenden Testszenario passen. Wenn ein beobachteter Wert außerhalb dieses Konfidenzintervalls liegt, kann man dieses Testszenario mit einer Irrtumswahrscheinlichkeit (p-Wert) von höchstens 5% verwerfen."

R Code Regression

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8885 -0.6617  0.1172  0.6318  2.1516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.297247   0.350697   248.9   <2e-16 ***
## x            3.497239   0.005171   676.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9237 on 148 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 4.573e+05 on 1 and 148 DF,  p-value: < 2.2e-16
```

Standardfehler der Regressionskoeffizienten

Mithilfe des Konfidenzintervalls der einzelnen Regressionskoeffizienten erfolgt die Modellselektion, die sich auf p-Werte stützt.

Die Standardfehler = Breite des Konfidenzintervalls hängen von unterschiedlichen Faktoren ab:

- **Stichprobengröße**: Je mehr Beobachtungen gemacht wurden, desto schmaler ist das Konfidenzintervall.
- **Messfehler**: Je geringer der Messfehler der Daten, desto schmaler ist das Konfidenzintervall.

  Der Messfehler entspricht der Varianz der Residuen ε, die möglichst klein sein sollte, also dass die Punkte eng um die Gerade liegen.
- **x-Range**: Je breiter die Spannweite der unabhängigen Variablen, desto schmaler ist das Konfidenzintervall.

  Daher sind "gute Hebelpunkte", die oft weit weg von den anderen x-Werten liegen und dadurch die Spannweite vergrößern, zusätzlich wichtig für das Modell.

Prädiktionsintervalle

μ Bei einem bekannten Wert der Lage μ lässt sich vor der Durchführung der Beobachtung für den arithmetischen Mittelwert $\bar{x}$ vorhersagen, dass dieser mit Wahrscheinlichkeit 95% im Intervall

$$\left(\mu - 1,96 \cdot \frac{\textit{standardabweichung}}{\sqrt{n}}; \mu + 1,96 \cdot \frac{\textit{standardabweichung}}{\sqrt{n}}\right)$$

liegen wird, wenn man diese Standardabweichung und Stichprobengröße n vorher kennt.

p Bei einem bekannten Wert der Erfolgswahrscheinlichkeit p lässt sich vor der Durchführung der Beobachtung für die relative Häufigkeit $\hat{p}$ vorhersagen, dass diese mit Wahrscheinlichkeit 95% im Intervall

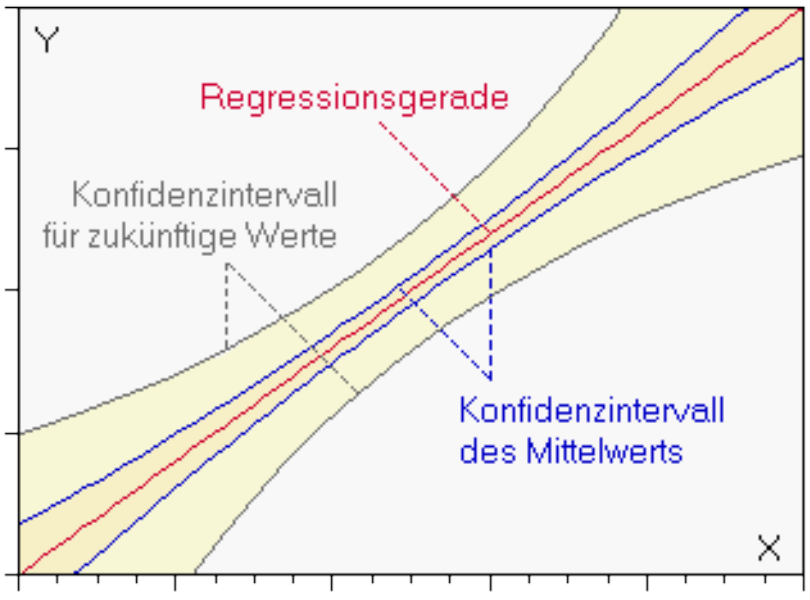
$$\left(p - 1,96 \cdot \frac{p \cdot (1-p)}{\sqrt{n}}; p + 1,96 \cdot \frac{p \cdot (1-p)}{\sqrt{n}}\right)$$

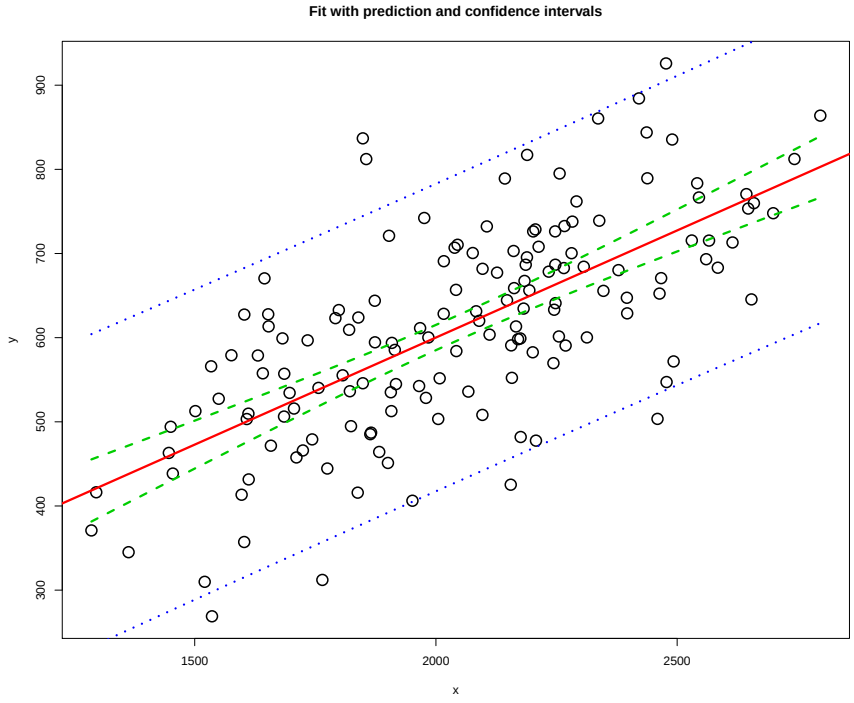
liegen wird, wenn man die Stichprobengröße n vorher kennt.

Vertrauensbereiche und Vorhersagebereiche der Regression

- Der **Vorhersagebereich einer Prädiktion aus dem Modell** wird an einer Stelle $\boldsymbol{x}_i$ als Prädiktionsbereich mit dem Mittelwert gleich dem Wert der Regressionsgerade und mit Konfidenzniveau α berechnet.
  An jeder Stelle wird dieser Vorhersagebereich "punkteweise" ermittelt und bezieht sich auf zukünftige y-Werte, die vorhergesagt werden.
- Der **Vertrauensbereich des Modells** wird so ermittelt, dass nicht nur an jeder Stelle, sondern über die gesamte Länge der Geraden hinweg ein Konfidenzniveau α gültig bleibt.
  Der Vertrauensbereich des Modells bezieht sich auf den Verlauf der Regressionsgeraden bzw. -(hyper)ebene und NICHT auf die y-Werte.

Vertrauensbereiche und Vorhersagebereiche der Regression






```
predict(linearesmodell,neuedaten,level = 1-alpha,interval =
```

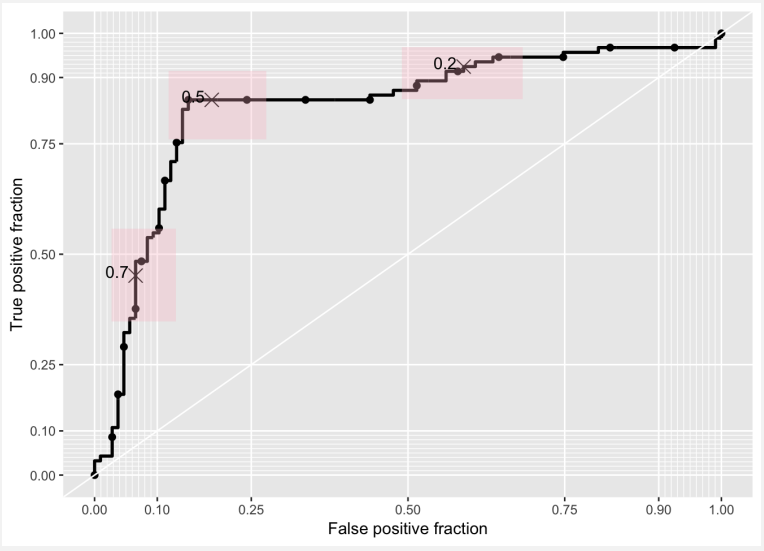
Arten von Fehlerberechnungen

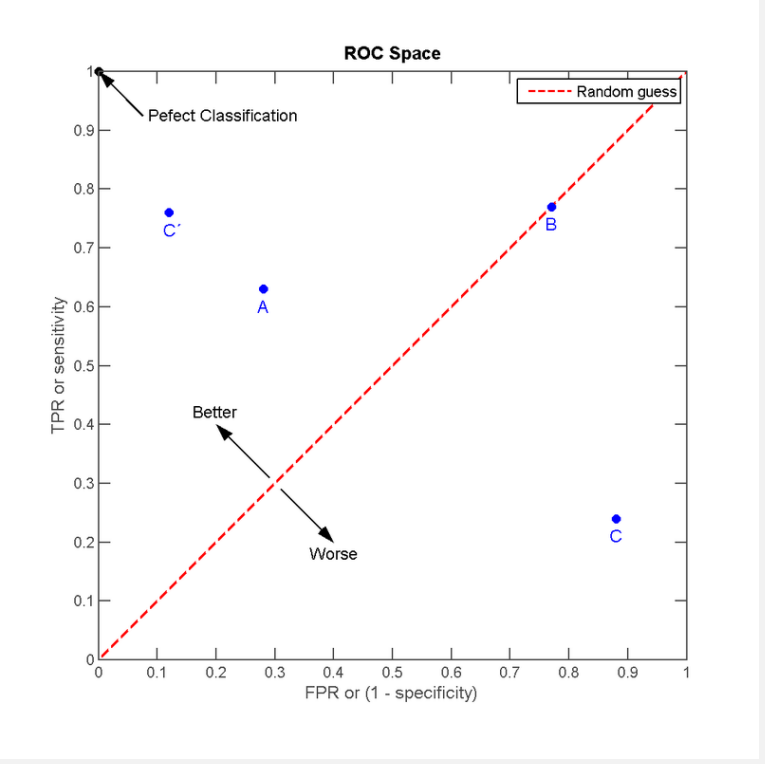
Klassifikationsfehler

	True condition				
Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
	True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Receiver Operating Characteristic (ROC) Curve

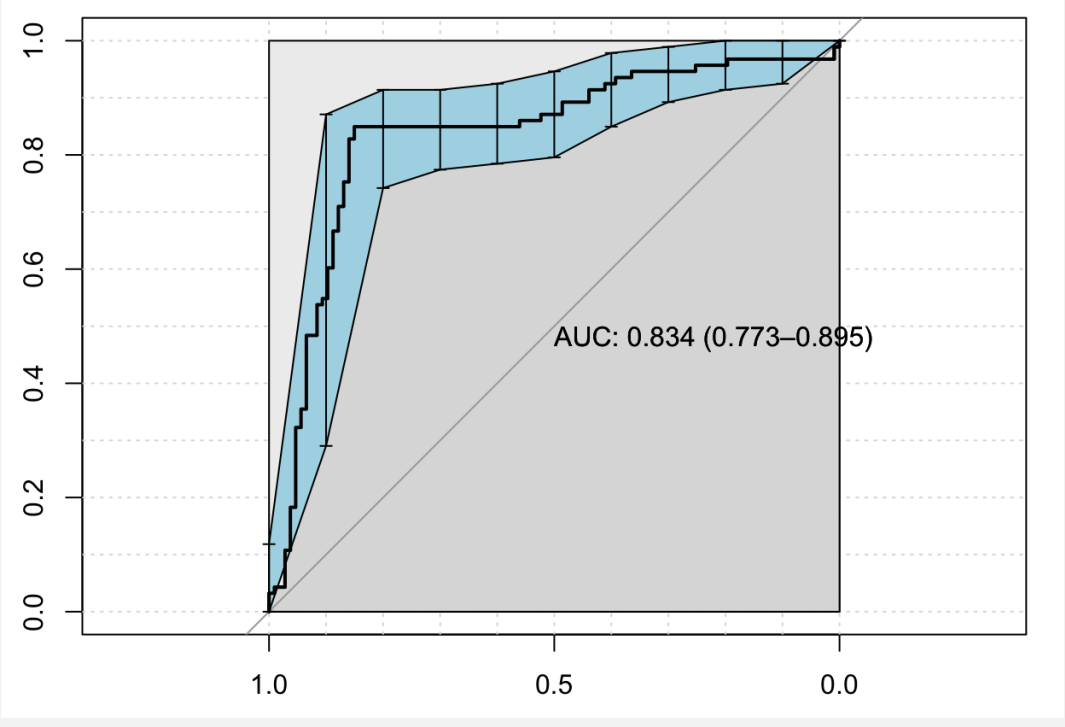
Receiver Operating Characteristic (ROC) Curve Die Receiver Operating Characteristic (ROC) Curve ist ein graphisches Diagnostikwerkzeug für binäre Klassifikationsalgorithmen. Sie stellt die Sensitivität = True Positive Rate gegen 1-Spezifizität = False Positive Rate dar.





Area under the Curve (AUC)

Area under the Curve (AUC) Die Fläche unterhalb der Receiver Operating Characteristic (ROC) Curve, Area under the Curve (AUC), ist ein Maß für die Klassifikationsqualität von binären Klassifikationsalgorithmen.



Confusion Matrix und ROC im Beispiel

	Wert = 0	Wert = 1	
Präd. =0	TN = 67	FN = 33	N = 100
Präd. =1	FP = 28	TP = 72	P = 100

Sensitivität = True Positive Rate = $\frac{TP}{P} = 0.72$

Spezifizität = True Negative Rate = $\frac{TN}{P} = 0.67$

Diese Werte werden dann für jeden möglichen Klassifikationsthreshold ausgerechnet, wodurch erst die ROC entstehen kann.

Confusion Matrix

```
Pima.tr$diabetes<-as.numeric(Pima.tr$type)-1
DiabetesModell <- glm(diabetes~.-skin-npreg-type,family = binomial(link="logit"),data=Pima.tr)

InSamplePrediction<-as.numeric(predict.glm(object = DiabetesModell,
                    newdata = Pima.tr,type = "response")>0.5)
OutOfSamplePrediction<-as.numeric(predict.glm(object = DiabetesModell,
                    newdata = Pima.te,type = "response")>0.5)

ConfusionMatrixInSample<-table(Pima.tr$diabetes,InSamplePrediction)
rownames(ConfusionMatrixInSample)<-c("kein Diabetes","Diabetes")
```

```
print.xtable(xtable(ConfusionMatrixInSample),comment = FALSE)
```

	0	1
kein Diabetes	116	16
Diabetes	30	38

ROC in R

```
library(pROC)
wahreWerte=labels; Vorhersagen=predictions;
pROCobj = roc(wahreWerte,Vorhersagen, smoothed = TRUE,
        ci=TRUE, ci.alpha=0.9, stratified=FALSE, plot=TRUE,
        auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
        print.auc=TRUE, show.thres=TRUE)
```

calculate confidence region

```
sens.ci <- ci.se(pROCobj)
```

colours confidence region lightblue and adds error bars to plot

```
plot(sens.ci, type="shape", col="lightblue")
plot(sens.ci, type="bars")
```

Was passiert bei mehrfacher Anpassung der Modelle?

- Bei Methoden, die auf **numerischen Optimierungsalgorithmen** basieren (Kleinste Quadrate [OLS], Maximum Likelihood [ML], etc.), ändert sich der Outcome nur minimal aufgrund der Ungenauigkeit der numerischen Schätzung, wenn dieselben Daten verwendet werden.
- Bei **(Re-)Sampling-basierten** Methoden (Bootstrapping [BS], Jackknifing [JK], Monte Carlo [MC] Simulation, MCMC, etc.) ändert sich die Schätzung bei jeder neuen Berechnung, selbst wenn dieselben Inputdaten verwendet werden.

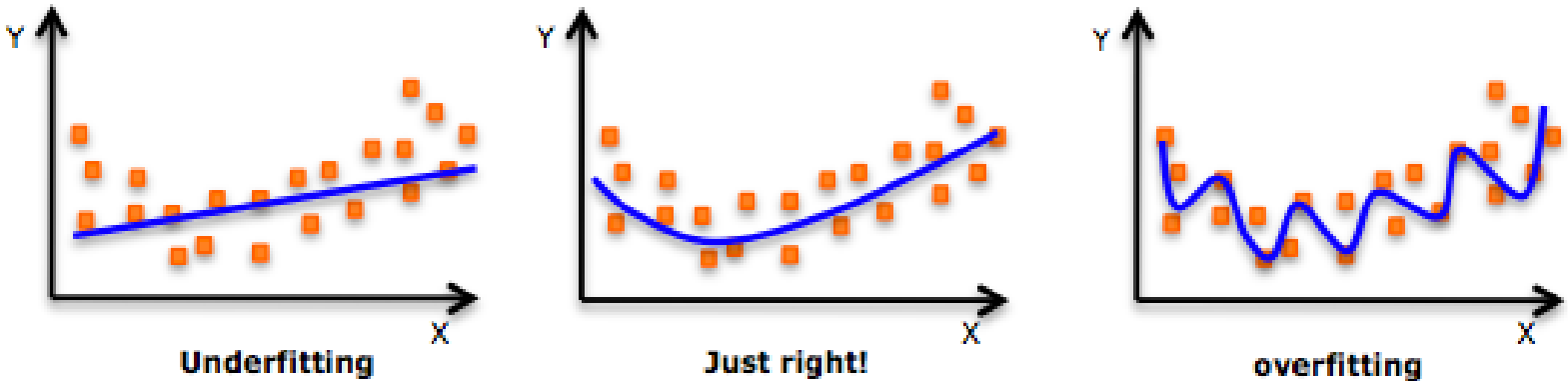
Wenn sich die Inputdaten ändern, ändert sich das geschätzte Modell und die daraus resultierenden Prädiktionen.

Was ist daher eine "gute Modellschätzung''?

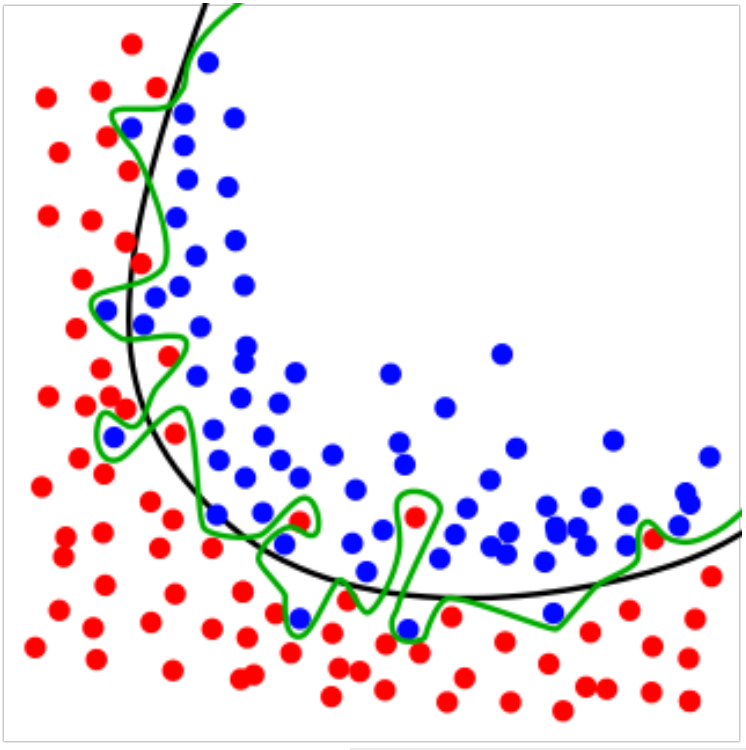
Wie evaluieren wir Schätzungen?

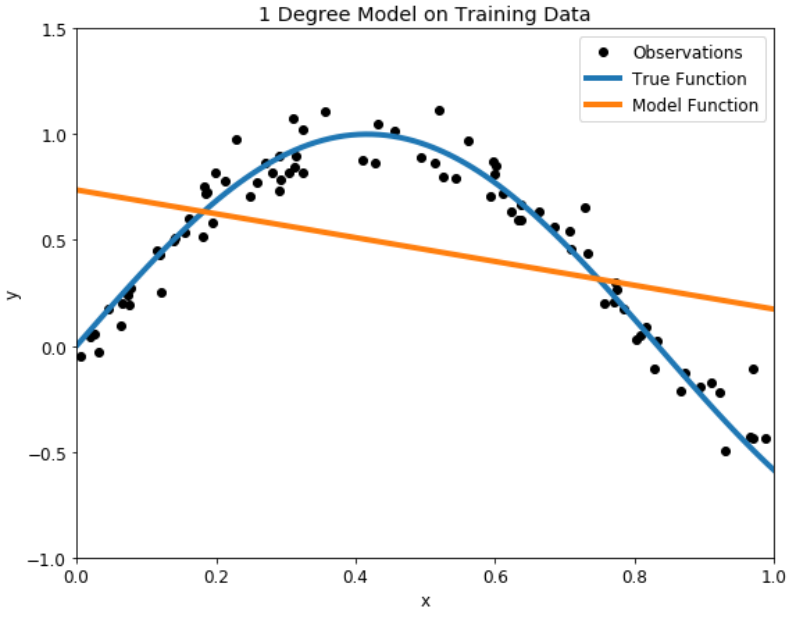
Probleme der Modellanpassung

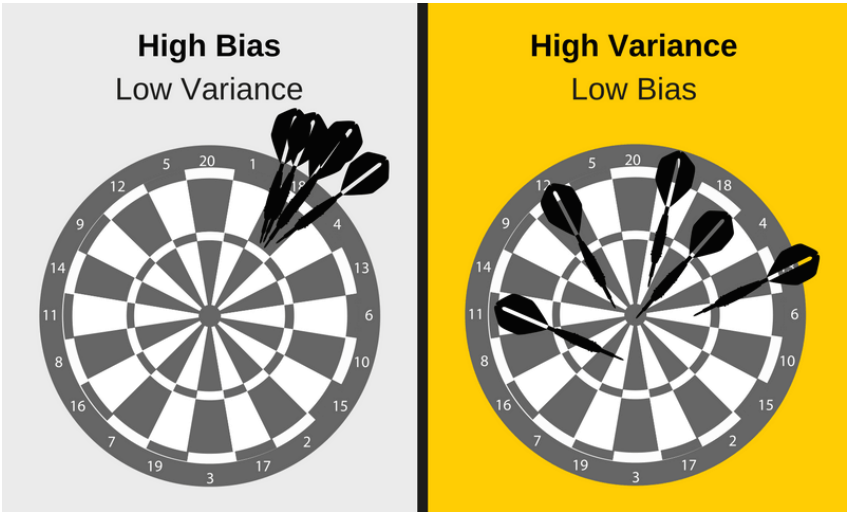
- **Underfitting**: zu wenig komplexes Modell wird angepasst (z.B. Gerade, obwohl die Daten eher polynomial oder exponentiell sind)
- **Overfitting**: zu komplexes Modell wird angepasst (z.B. Polynom hoher Ordnung, obwohl die Daten eher ein einfaches Polynom oder exponentiell sind)



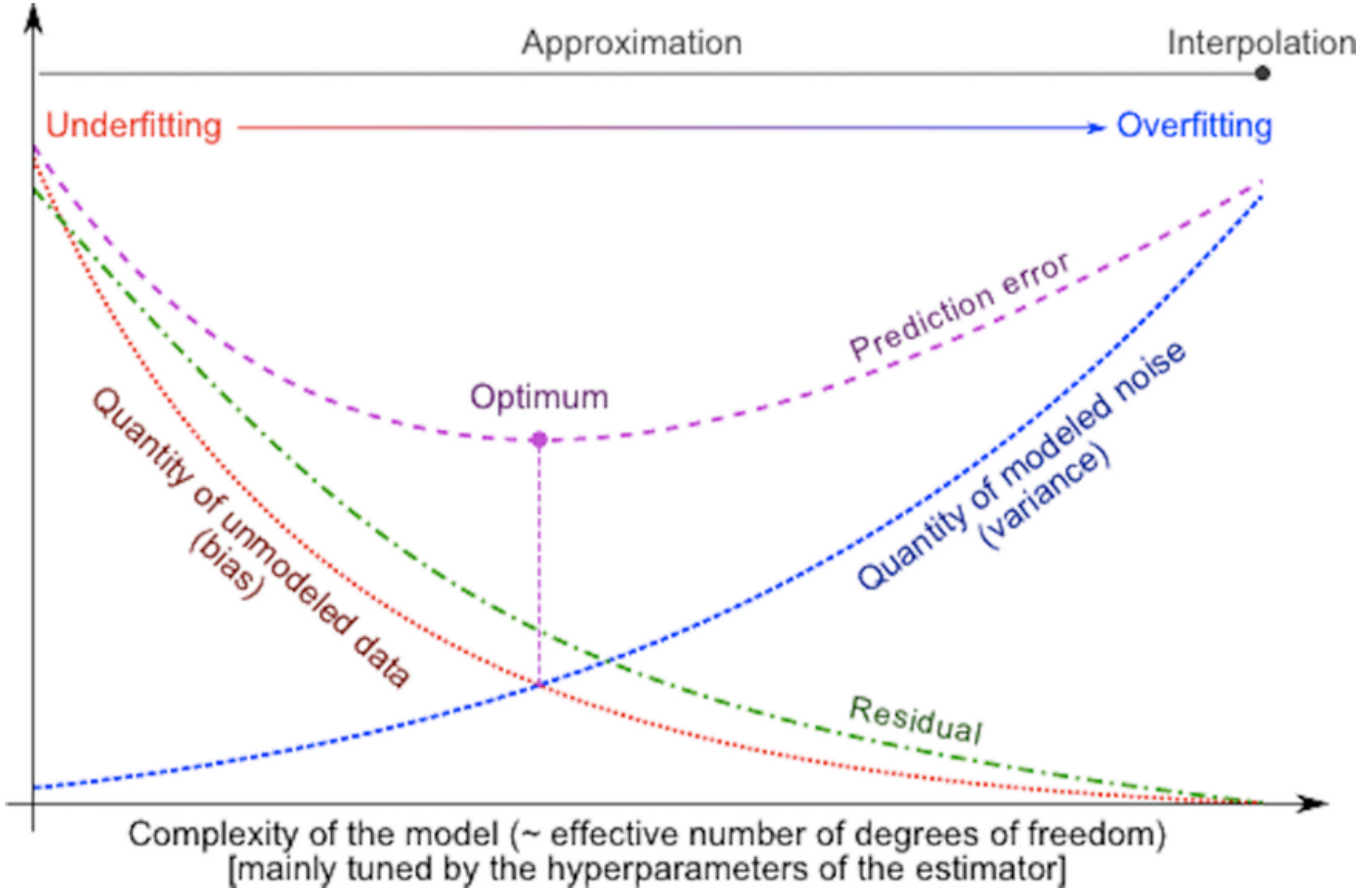
Probleme der Modellanpassung







Over- vs. Underfitting

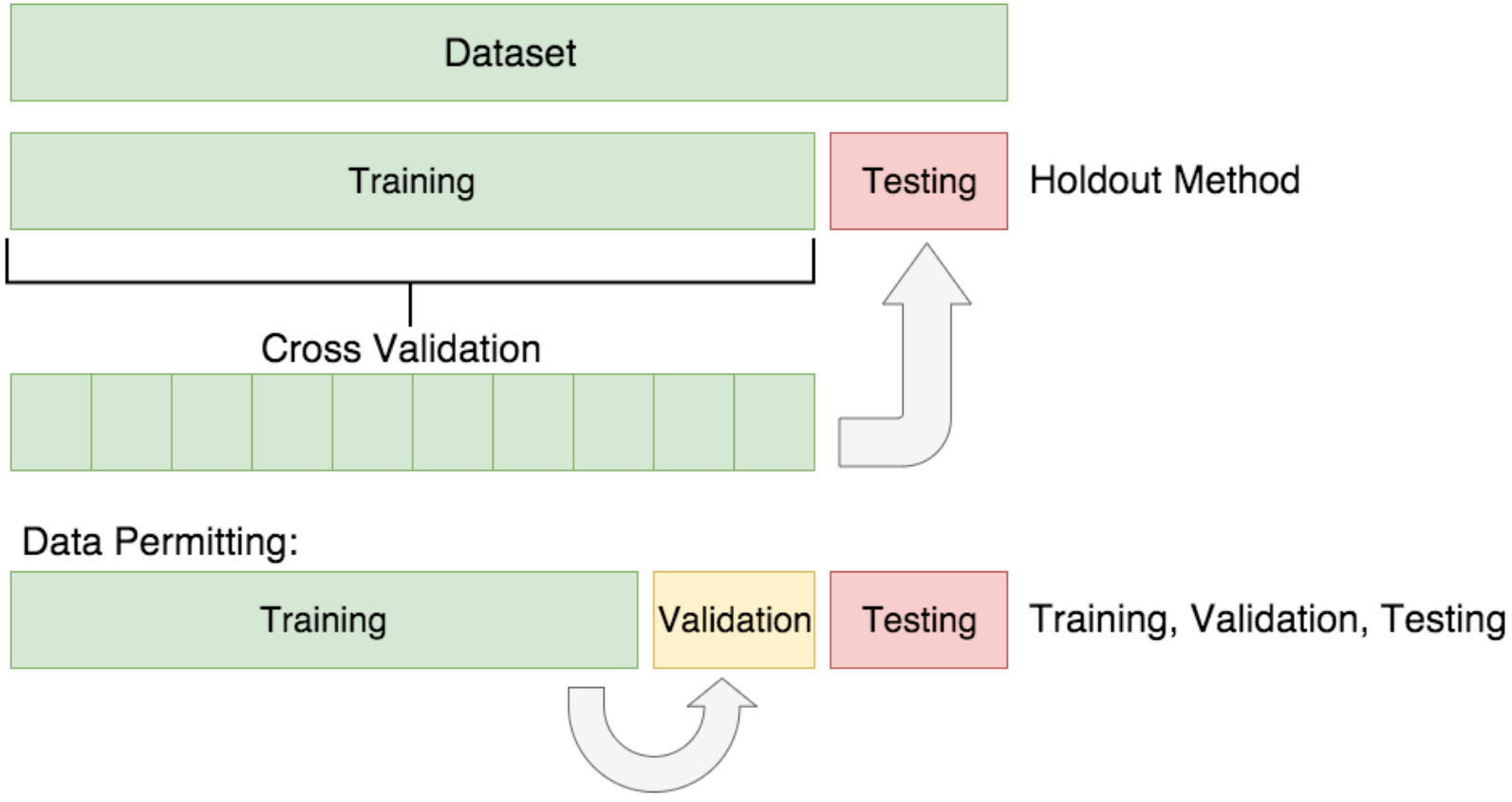


Validierung von Modellen

- **Modellselektion**: die Qualität von Modellen einschätzen, um das "Beste" auszuwählen
- **Modellevaluation**: nach Selektion eines bestimmten Modells den Prädiktionsfehler abschätzen

- **In-sample Validation**: Einsetzen der Daten, die zur Modellschätzung verwendet wurden in das Modell. Vergleich der geschätzten Werte mit den wahren Werten.
- **Out-of-sample Validation**: Einsetzen anderer Daten, die nicht zur Modellschätzung verwendet wurden in das Modell. Vergleich der geschätzten Werte mit den wahren Werten.

Datenaufteilung zur Validierung



Methoden zur Qualitätsprüfung und Validierung

- **Jackknifing**

  Unterteilt die Daten in Teildatensätze und schätzt auf jedem Teildatensatz separat

- **Crossvalidation (Kreuzvalidierung)**

  unterteilt die Daten in Teildatensätze zur Validierung (out-of-sample-prediction), während auf dem Rest der Daten geschätzt wird

- **Bootstrapping**

  zieht mit zurücklegen aus den Daten und kreiert viele virtuelle Stichproben über die geschätzt wird

- **Permutationstests**

  vertauscht die Gruppenzugehörigkeiten und wiederholt die Modellanpassung

```
library(boot)
cv.glm(model) # crossvalidation mit (generalised) linear mo
boot #  bootstrapping
```

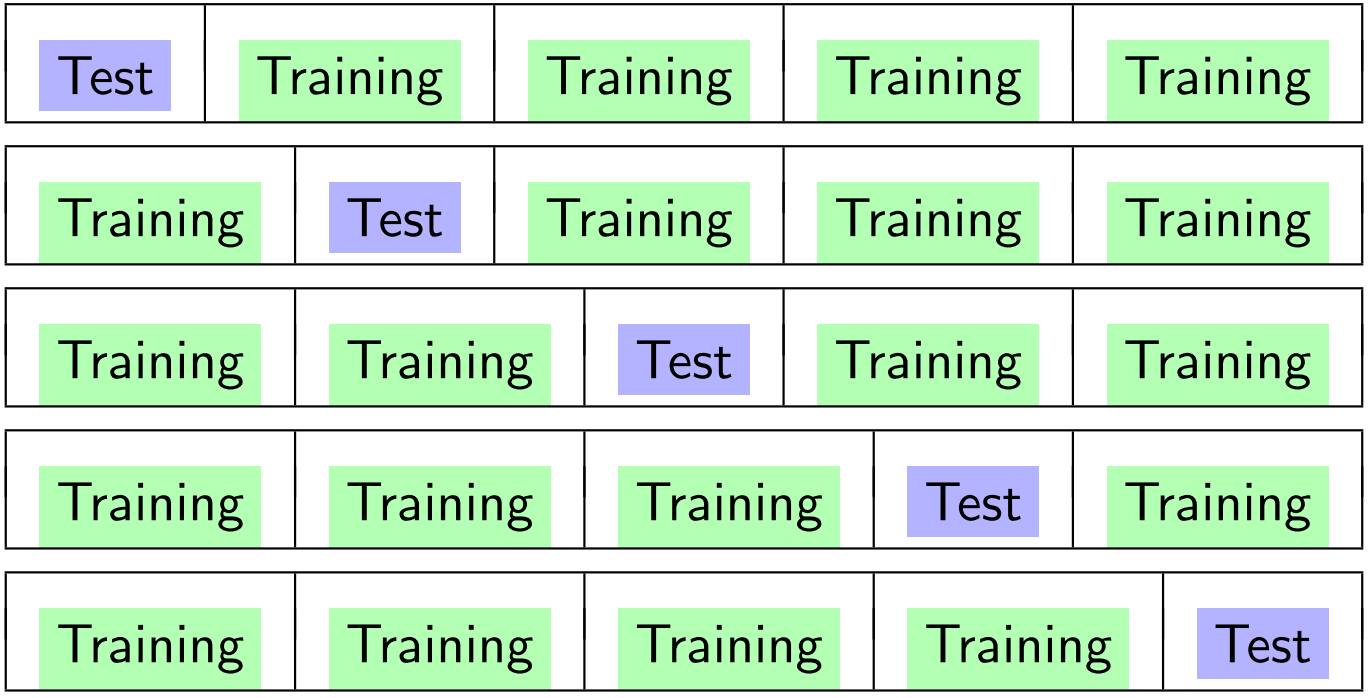
Kreuzvalidierung (Crossvalidation)

- **k-fache Kreuzvalidierung (k-fold Crossvalidation)**

  Bsp: Für k=5 wird der Datensatz in 5 gleich große Teile aufgeteilt.

  Wichtig ist, dass das zufällig geschieht und hier nicht systematisch die "besten" und "schlechtesten" Beobachtungen zusammengenommen werden.

  Je ein Teil wird als Testdatensatz herangezogen, während die anderen 4 Teile zum Training = Modellschätzung verwendet werden.



Wie wählt man k?

k ist stark abhängig von den Daten und der Datenmenge

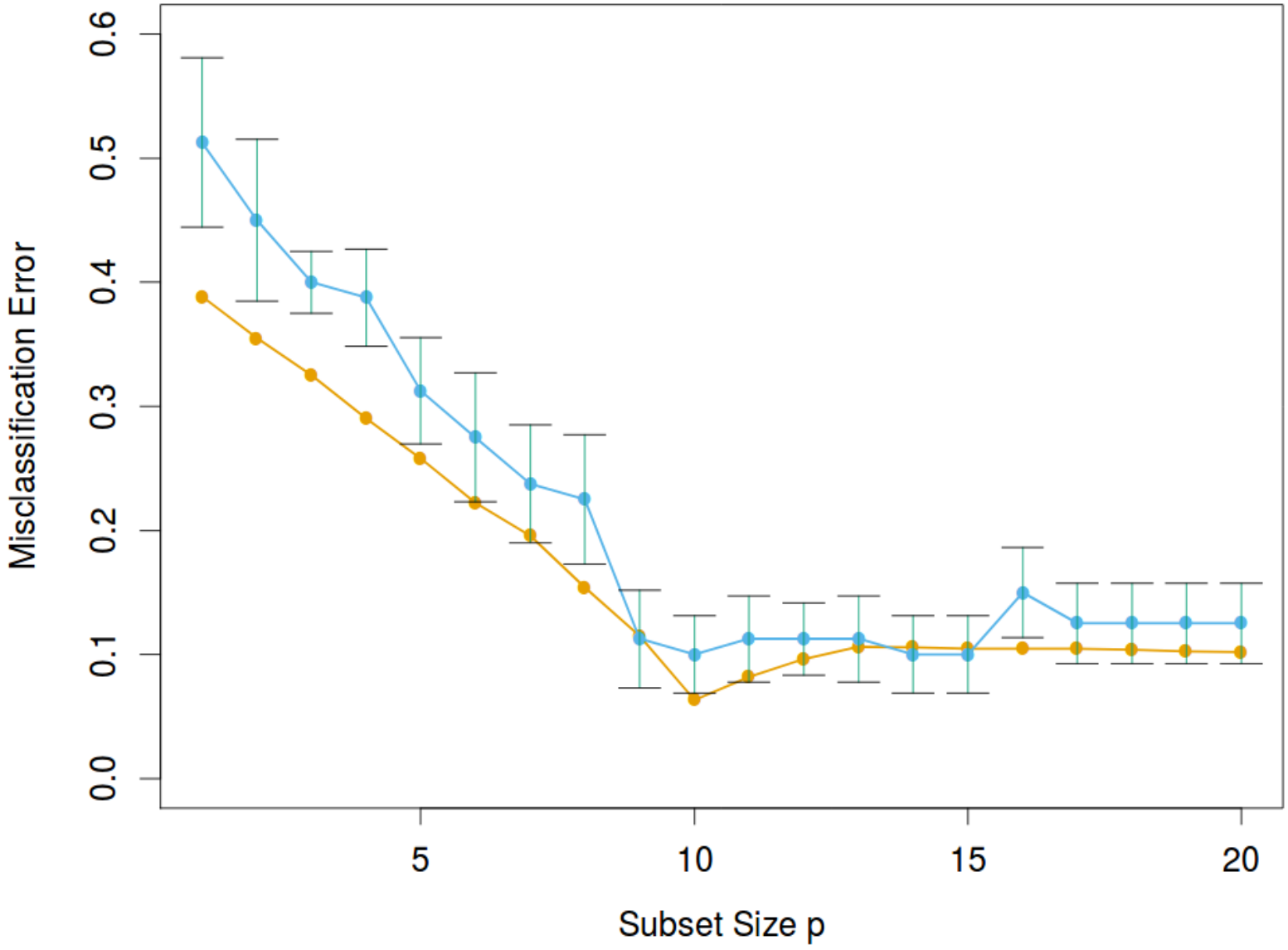
- k=n: **“Leave one out” Crossvalidation**

  sinnvoll für nicht allzu große Datensätze (n $\leq$ 1000)
  die Datensätze sind einander sehr ähnlich
  wird für größere Datenmengen und Modelle extrem rechenaufwändig hier werden Hebelpunkte und Ausreißer besonders schlagend und sollten auffallen

- k=5 oder k=10: einfache Kreuzvalidierung in 5-10 gleich großen Teildatensätzen
  schnell berechenbar und auch für große Datenmengen noch computational durchführbar

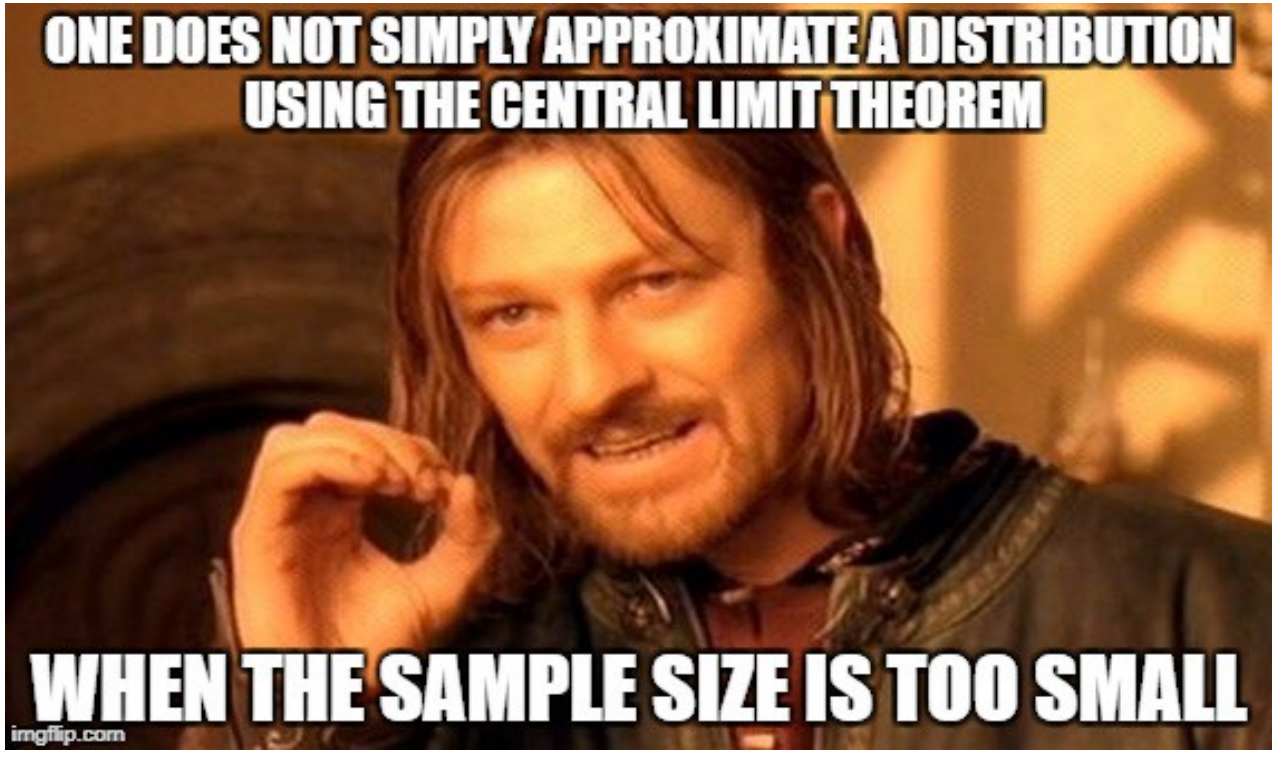
- andere Werte von k stellen eine Balance zwischen Genauigkeit der Erfassung von Unterschieden innerhalb des Datensatzes und der verfügbaren Rechenleistung und -zeit dar

Was bringt Crossvalidation?



Bootstrapping

- keine Ahnung von zugrundeliegender Datenverteilung
- wenige Beobachtungen oder Wiederholungen von Experimenten sind zu teuer, unethisch, ...



Bootstrapping

Ablauf des Algorithmus

- aus den Daten selbst simulieren (**ziehe** Beobachtungen **mit Zurücklegen**)

  kreiere sehr viele virtuelle Stichproben

- auf **jeder virtuellen Stichprobe** das **Modell** anpassen, Werte aus**rechnen** und **prädiktieren**, etc.

- **Verteilung** der vielen **Einzelergebnisse** erlaubt Aussagen über Signifikanz, Konfidenz, Genauigkeit (Akkuratheit und Präzision), Sensitivität, Spezifizität

  beim Simulieren von Mittelwerten tritt der Zentraler Grenzverteilungssatz in Kraft

Probleme bei Evaluation mittels Algorithmen

- Crossvalidation

  Die Aufteilung darf keine Systematik haben (Werte aufsteigend/absteigend anordnen und dann aufteilen, etc.)

  die ideale Anzahl von Validationsdatensätzen ist nicht allgemein definiert

- Bootstrapping

  Ausreißer können massiv verstärkt werden und Modalität etc. kreieren

- für große Datenmengen ist beides extrem ressourcenaufwändig

  leave-on-out crossvalidation ist fast unmöglich ab einer bestimmten Datenmenge

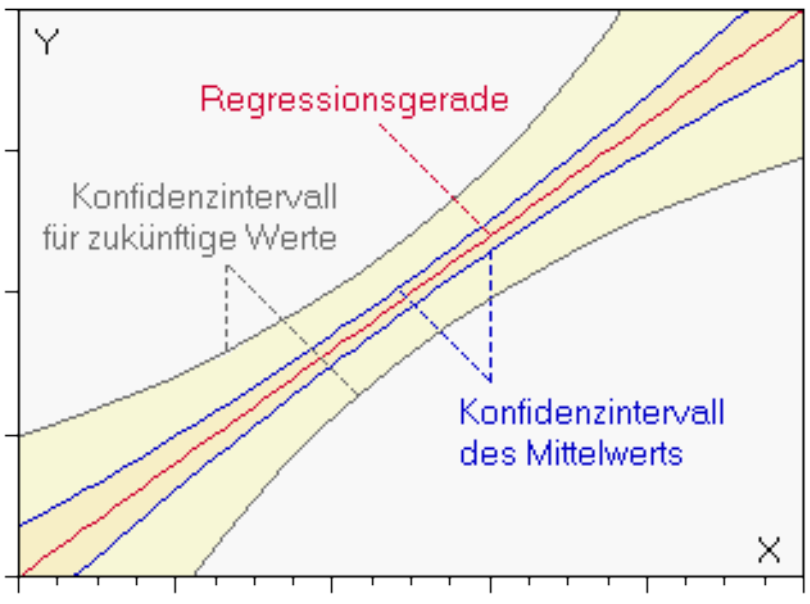
Maße für Modellanpassung

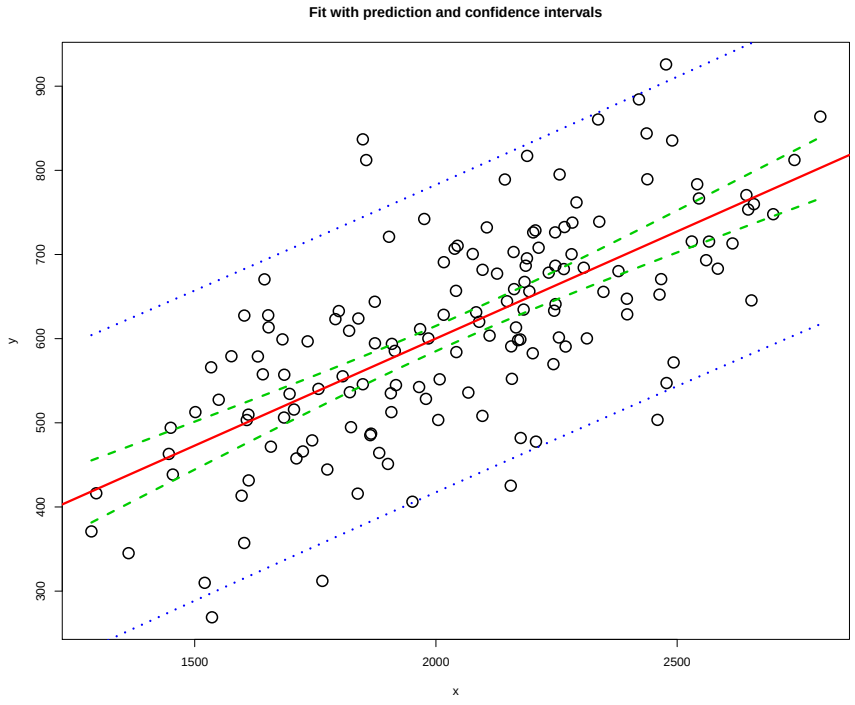
Metrischer Outcome	Mean Squared Error	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
	Root Mean Squared Error	$\text{RMSD} = \sqrt{\text{MSE}(\hat{\theta})}$ $\text{RMSD} = \sqrt{\text{E}((\hat{\theta} - \theta)^2)}$ $\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$
	Median absolute Deviation	$\text{MAD} = \text{median}(\lvert X_i - \mathit{median}(X)\rvert)$

Klassification

Categorial Dichotomous Outcome	Sensitivity True Positive Rate	$\frac{\sum \text{True Positive}}{\sum \text{Condition Positive}}$
	Specificity True Negative Rate	$\frac{\sum \text{True Negative}}{\sum \text{Condition Negative}}$
Classification	Accuracy	$\frac{\#\text{correctly classifed units}}{\#\text{all units}}$
	Missclassification Rate	$\frac{\#\text{incorrectly classifed units}}{\#\text{all units}}$

Vertrauensbereiche und Vorhersagebereiche der Regression





Arten von Fehlerberechnungen

Klassifikationsfehler

	True condition				
Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
	True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Area under the Curve (AUC)

Die Fläche unterhalb der Receiver Operating Characteristic (ROC) Curve, Area under the Curve (AUC), ist ein Maß für die Klassifikationsqualität von binären Klassifikationsalgorithmen.