# Frequentistische Statistik und Bayessche Statistik

Volker Tresp

# Frequentistische Statistik

## Herangehensweise

- Die Naturwissenschaft versucht es, der Natur Gesetzmäßigkeiten zu entringen:

$$F = ma$$

- Gesetze gelten unter wiederholbaren idealisierten Situationen: Wiederholter fall eines idealisierten (punktförmigen) Objektes mit vernachlässigtem Luftwiderstand, unter Geschwindigkeiten weit unter der Lichtgeschwindigkeit, ...
- Dies motiviert die frequentistische Statistik: Aussagen probabilistischen Natur unter (im Prinzip) gleichen Bedingungen wiederholbaren Experimenten
- Da die zugrundeliegenden Elementarereignisse typischerweise unbekannt sind, analysiert man direkt die Eigenschaften der Zufallsvariablen

# Grundbegriffe

- Zur statistischen Analyse gehört somit eine genaue Beschreibung eines statistischen Experimentes, zum Beispiel der genauen Art und Weise, wie und an wen ein Medikament verabreicht wird
- Eine **statistische Einheit** ist die Objektklasse, an der Messungen vorgenommen werden, bzw Attribute aufgenommen werde. Im vorangegangenen Beispiel wären dies z.B. Personen
- Als **Population** (Grundgesamtheit) wird die Menge aller statistischen Einheiten bezeichnet, über die eine Aussage gemacht werden soll: Diabetiker, die ein bestimmtes Medikament genommen haben
- Zur Analyse steht nur eine **Stichprobe** (Sample); es wird häufig angenommen, dass die Stichprobe eine zufällig gewählte Untermenge der Population darstellt

## Population

- Eine Population kann endlich, unendlich oder hypothetisch sein
- In einer Bundestagswahl mag die Population aus allen Bürgern bestehen, die zur Wahl gehen
- Im Fall der Wirksamkeit eines Diabetiker Medikamentes mag nicht primär die Wirksamkeit auf alle Diabetiker in Deutschland interessiere, sondern die mittlere Wirksamkeit in Bezug auf einen hypothetischen typischen Diabetiker
- Eine Analyse mit hypothetischen (potentiell unendlichen) Populationen ist oft einfacher

# Typische Annahmen

- Es gibt eine Population aus der eine zufällige Stichprobe von Datenpunkten gezogen wurde
- Zu Datenpunkt $i$ werden Merkmale $\mathbf{x}_i$ bestimmt
- Da Datenpunkte unabhängig gezogen werden, lässt sich schreiben

$$P_{\mathbf{w}}(\mathbf{x}_1, ..., \mathbf{x}_N) = \prod_{i=1}^{N} P_{\mathbf{w}}(\mathbf{x}_i)$$

- Man nimmt an, dass die Daten aus einer Familie von Verteilungsfunktionen $P_{\mathbf{w}}(\mathbf{x}_i)$ kommt, die durch den Parametervektor $\mathbf{w}$ parametrisiert werden; durch die Wahl der Form von $P_{\mathbf{w}}(\mathbf{x}_i)$ werden Annahmen gemacht
- Ziel ist es, die Parameter zu schätzen

# Beispiel: Personengröße

- Man nimmt an, dass die Größe der Personen $\mathbf{x}_i$ in der Population Gauß-verteilt ist, also

$$P_{\mathbf{w}}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x}_i - \mu)^2\right)$$

  mit $\mathbf{w} = (\mu, \sigma)^T$

- Damit ergibt sich

$$P_{\mathbf{w}}(\mathbf{x}_1, ..., \mathbf{x}_N) = \prod_{i=1}^{N} P_{\mathbf{w}}(\mathbf{x}_i) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i; \mu, \sigma^2)$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (\mathbf{x}_i - \mu)^2\right)$$

# Maximum Likelihood

- Eine sinnvolle Schätzung ist der Parametervektor, der die Date am besten erklärt, also die Wahrscheinlichkeit der Daten maximiert. Letztere ist die Likelihood-Funktion (als Funktion der Parameter

$$L(\mathbf{w}) = P_{\mathbf{w}}(\mathbf{x}_1, ..., \mathbf{x}_N)$$

- Bequemer ist es mit der Log-Likelihood-Funktion zu arbeiten,

$$l(\mathbf{w}) = \log L(\mathbf{w}) = \sum_{i=1}^{N} \log P_{\mathbf{w}}(\mathbf{x}_i)$$

- Der Maximum Likelihood (ML) Schätzer ist nun gegeben durch

$$\hat{\mathbf{w}}_{ml} \doteq \arg\max(l(\mathbf{w}))$$

- Dies bedeutet: aus der Familie der betrachteten Wahrscheinlichkeitsverteilungen ist der ML-Schätzer derjenige, der die Daten am besten erklären kann: das beste Modell in der betrachteten Klasse von Modellen!

## Beispiel Personengröße: ML Schätzer

- Die ML Schätzung im Beispiel sind

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

und

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \widehat{\mu})^2$$

# ML-Schätzer für ein lineares Modell

- Nehmen wir an, dass die wahre Abhängigkeit linear ist, wir jedoch nur verrauschte Daten zur Verfügung haben

$$y_i = \mathbf{x}_i^T \mathbf{w} + \epsilon_i$$

- Weiterhin nehmen wir an, dass das Rauschen einer Gauß-Verteilung folgt

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\epsilon_i^2\right)$$

- Daraus folgt, dass

$$P_{\mathbf{w}}(y_i|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2\right)$$

- Leichter handhabbar ist der Logarithmus dieses Ausdrucks

$$\log P_{\mathbf{w}}(y_i|\mathbf{x}_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2$$

## Der Maximum-Likelihood Schätzer

- Die Log-Likelihood Funktion ist somit

$$l = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2$$

- Unter der Annahme von Gauß-Rauschen folgt daher, dass der Maximum Likelihood (ML) Schätzer gleich dem least-squares-Schätzer ist

$$\hat{\mathbf{w}}_{ml} \doteq \arg\max(l(\mathbf{w})) = \hat{\mathbf{w}}_{LS}$$

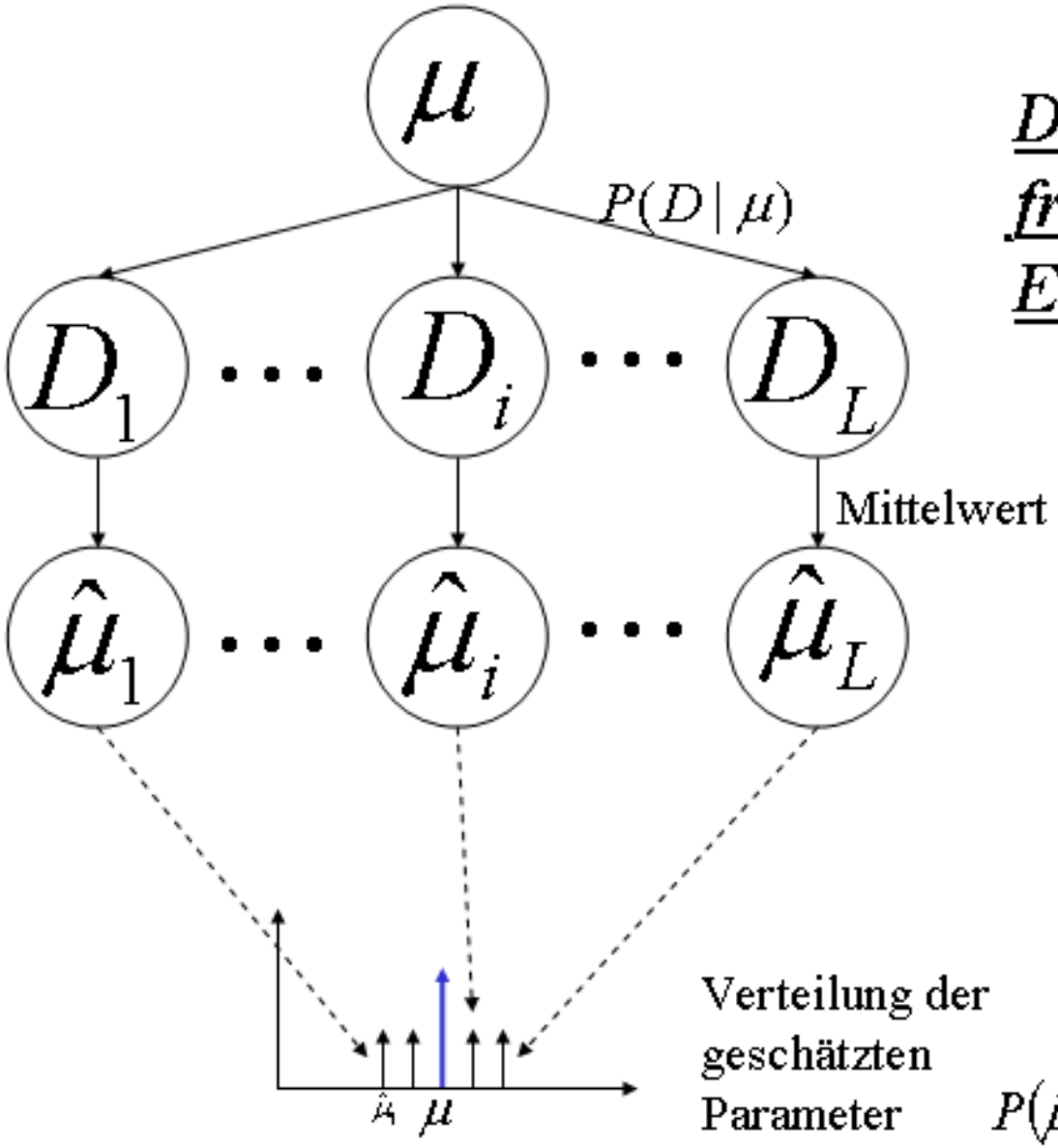
Da, $\hat{\mathbf{w}}_{ml} = \arg\max -\sum_i (y_i - \mathbf{x}_i^T\mathbf{w})^2$ und $\hat{\mathbf{w}}_{ls} = \arg\min \sum_i (y_i - \mathbf{x}_i^T\mathbf{w})^2$

# Analyse von Schätzern

- Der ML-Schätzer ist sicherlich plausibel, aber vielleicht gibt es andere Parameterwerte mit fast der gleichen Likelihood. Wie sicher ist man in Bezug auf die Parameterschätzung?
- Hierzu wird das folgende Gedankenexperiment angestellt (nächste Abbildung)
- Sei $\mu$ der unbekannte aber feste Parameter
- Es werden unendlich viele Datensätze (Stichproben) $D_1, D_2, \ldots D_L$, $L \to \infty$ generiert, jeder der Größe $N$
- Für jeden dieser Datensätze $D_i$ berechnet man einen Parameterschätzer $\widehat{\mu}_i$ (zum Beispiel den ML-Schätzer)
- Ich berechne und analysiere die Verteilung der geschätzten Parameter
- Im Beispiel ergibt sich für die mittlere Personengröße

$$P_\mu(\widehat{\mu}) = \mathcal{N}\left(\widehat{\mu}; \mu, \frac{\sigma^2}{N}\right)$$

- Diese Verteilung lässt sich berechnen, ohne die tatsächlichen Daten zu kennen
- Liegt ein aus Daten geschätztes $\widehat{\mu}$ können Aussagen abgeleitet werden wie: dieses $\widehat{\mu}$ ist sehr unwahrscheinlich, wenn $\mu = 175cm$ beträgt




## *Das frequentistische Experiment*

Verteilung der geschätzten Parameter

$$P(\hat{\mu} \mid \mu) \propto \mathrm{N}\left(\mu, \sigma^2/N\right)$$

## Erwartungstreue

- Man interessiert sich nun zum Beispiel, ob ein Schätzer, gemittelt über alle Datensätze, gleich dem wahren Schätzer ist

$$\lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \hat{w}_i = E_D(\hat{w})$$

- Wenn $E_D(\hat{w}) = w$, dann nennt man diesen Schätzer erwartungstreu (unverzerrt, unbiased)
- In unserem Beispiel ist der Schätzer erwartungstreu, da $E_D(\hat{\mu}) = \mu$

## Asymptotische Erwartungstreue und Bias

- Manchmal ist ein Schätzer für endliche Daten nicht erwartungstreu, wird aber für $N \to \infty$ erwartungstreu. Ein Schätzer ist **asymptotisch erwartungstreu**, wenn:

$$E_{N\to\infty}(\widehat{w}) = w$$

- Die Differenz zwischen wahren Parameter und dem Erwartungswert des geschätzten Parameters ist der Bias (Verzerrung)

$$Bias(\widehat{w}) = E_D(\widehat{w}) - w$$

  Im Beispiel ist der Bias Null.

## Verzerrtheit des ML-Schätzers bei endlichen Daten

- Für endliche Daten können ML-Schätzer verzerrt sein

$$\hat{\sigma}^2_{ml} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\mu})^2$$

$$\hat{\sigma}^2_{unverz} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\mu})^2$$

## Varianz eines Schätzers

- Die Varianz eines Schätzers bewertet, inwieweit die Schätzung um den eigenen Erwartungswert variiert

$$Var(\hat{w}) = E_D\left(\hat{w} - E_D(\hat{w})\right)^2$$

- Im Beispiel ist $Var(\hat{w}) = \sigma^2/N$

# Erwarteter Fehler

- Erwarteter quadratischer Fehler ist die erwartete Abweichung vom Schätzer vom wahren Wert

$$MSE = E_D\left(\hat{w} - w\right)^2$$

- Man kann zeigen, dass sich dieser aus Varianz und dem Quadrat des Bias zusammensetzt

$$MSE = Var_D(\hat{w}) + Bias_D^2(\hat{w})$$

- Wünschenswert ist das der MSE für $N \to \infty$ gegen Null geht. Ein Schätzer ist MSE-Konsistenz, wenn gilt

$$MSE_{N\to\infty} \to 0$$

  In unserem Beispiel ist $MSE = \sigma^2/N$. Der Ausdruck geht gegen Null für $N \to \infty$

## Vergleich von Schätzern

- Wenn man verschiedene Schätzer zur Verfügung hat, will man herausfinden, welcher denn besser ist. Seien $\widehat{w}(1)$ und $\widehat{w}(2)$ zwei unterschiedliche Schätzer. $\widehat{w}(1)$ ist MSE-wirksamer als $\widehat{w}(2)$ falls

$$MSE[\widehat{w}(1)] \leq MSE[\widehat{w}(2)]$$

- Ein Schätzer $\widehat{w}(i)$ ist MSE-wirksamst, falls

$$MSE[\widehat{w}(i)] \leq MSE[\widehat{w}(j)] \quad \forall \widehat{w}(j)$$

# Eigenschaften des ML-Schätzers

Einer der wichtigste Schätzer ist der Maximum-Likelihood (ML)-Schätzer. Der ML-Schätzer hat viele positive Eigenschaften, die seine Beliebtheit begründen:

- Der ML-Schätzer ist asymptotisch $N \to \infty$ unverzerrt (unbiased) (auch wenn er mit endlich vielen Daten verzerrt sein kann)
- Etwas überraschend kann man auch zeigen, dass der ML-Schätzer asymptotisch für $N \to \infty$ immer der beste unverzerrte Schätzer ist: er ist MSE-wirksamst unter allen asymptotisch unverzerrten Schätzern (efficient)
- Der ML-Schätzer ist asymptotisch $N \to \infty$ immer Gauß-verteilt, auch wenn kein Gauß-Rauschen Teil des Modells ist. Im Beispiel ist der Schätzer auch für endlich viele Daten Gauß-verteilt,

$$P_\mu(\hat{\mu}) = \mathcal{N}\left(\hat{\mu}; \mu, \frac{\sigma^2}{N}\right)$$

# Fehlerschranken und Hypothesentests

- Basierend auf der Varianz der Schätzer können nun Fehlerschranken angegeben werden und es können Hypothesentests durchgeführt werden
- Im Beispiel lässt sich die Aussage treffen:

$$P\left(-z(1-\alpha/2) \leq \frac{\widehat{\mu}-\mu}{\sigma/\sqrt{N}} \leq z(1-\alpha/2)\right) = 1-\alpha$$

  wobei $z(1-\alpha/2)$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung ist
- Die Aussage ist, dass in $100\times(1-\alpha)\%$ der Fälle das Intervall den wahren Parameter überdecken.
- Kommentare zu Hypothesentests im Anhang

# Diskussion: ML

- Auch für komplexere Modelle lässt sich die (Log)-Likelihood in der Regel berechnen (z.B. für Modelle mit latenten Variablen)
- Da Daten oft unabhängig generiert werden, ist die Log-Likelihood in der Regel eine Summe über die Anzahl der Datenpunkte

$$l(\mathbf{w}) = \sum_{i=1}^{N} \log P(y_i|\mathbf{w})$$

# Diskussion: ML (2)

- Die Notwendigkeit, das datengenerierende Modell nachzubilden, führt zu interessanten problemangepassten Modellen
- Nachteil: man muss annehmen, das das wahre Modell sich in der Klasse der betrachteten Modelle befindet
- Mit endlich vielen Daten kann der ML-Schätzer zur Überanpassung führen, d.h. komplexere Modelle werden bevorzugt
- Die frequentistische Statistik ist stark fokussiert auf die Eigenschaften von Parametern (Signifikanz, ...)

# Bayessche Statistik

# Der Bayessche Ansatz

- Der wesentliche Unterschied ist, dass auch Parameter als Zufallsvariable behandelt werden
- Dies bedeutet, dass der Benutzer zunächst eine *a priori* Annahme über die Verteilung der Parameter machen muss:

$$P(\mathbf{w})$$

- Man erhält ein komplettes probabilistisches Modell

$$P(\mathbf{w})P(D|\mathbf{w})$$

# A Priori Verteilung

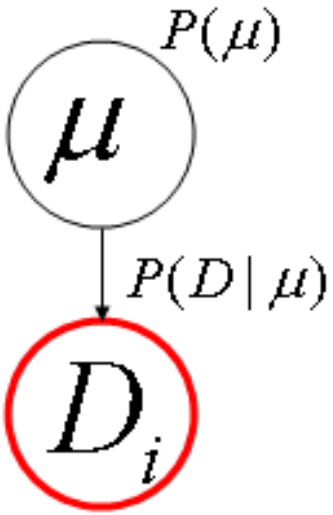
- Im Bayesschen Ansatz wird auch ein Zufallsprozess für die Generierung der Parameter $P(\mathbf{w})$ angenommen; macht es Sinn hier einen Zufallsprozess anzunehmen?
- Hier hilft die Interpretation von Wahrscheinlichkeiten als subjektive Aussage über den Wahrheitsgehalt von Hypothesen: Cox (1946): Wenn man seinen Überzeugungen (Beliefs) Zahlen zuordnen will, kommt man unter wenigen Konsistenzannahmen auf den Bayesschen Formalismus
- Beispiel: Schätzung des Durchmessers von Jupiter. Der Durchmesser mag auch durch Zufallsprozese bestimmt sein, wie den Einschlag von Meteoriten über Jahrmilliarden. D.h. eine a priori Verteilung $P(w)$, wobei hier $w$ der Durchmesser von Jupiter ist, der als Zufallsvariable von Zufallsprozessen bestimmt wurde macht durchaus einen Sinn

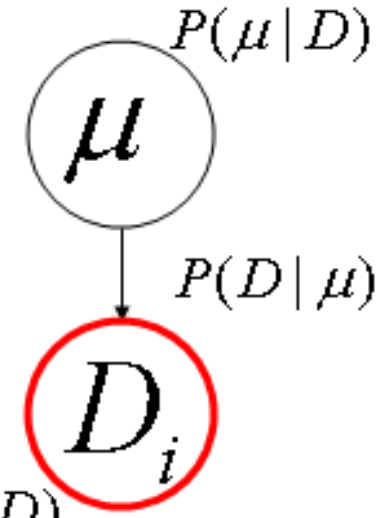
# Das Bayessche Experiment

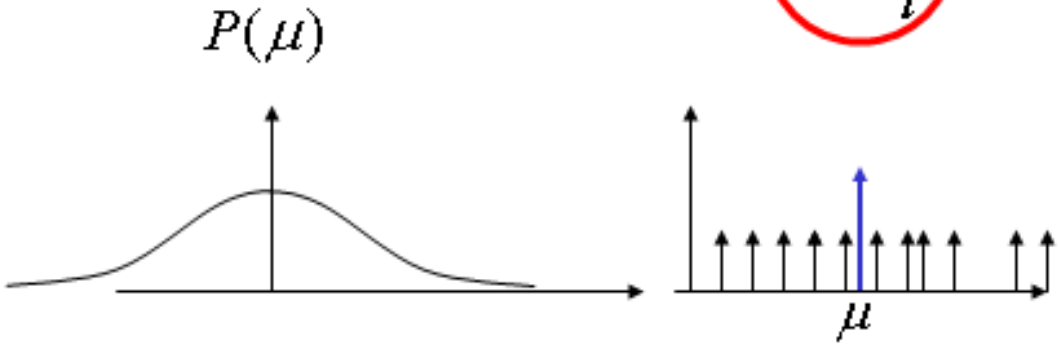
- Im Gegensatz zum frequentistischen Ansatz benötigen wir keine hypothetischen Datensätze (Stichproben) $D_1, D_2, \ldots D_L,\ L \to \infty$ bemühen. Wir arbeiten nur mit den tatsächlichen Daten $D$
- Im Beispiel wird angenommen, dass der wahre Parameter $\mu$ aus einer *a priori* Verteilung $P(\mu)$ generiert wurde. Im Beispiel: $P(\mu) = \mathcal{N}(\mu; 0, \alpha^2)$
- Die Daten werden generiert gemäß $P(D|\mu) = \prod_i \mathcal{N}(x_i; \mu, \sigma^2)$
- Nach dem Satz von Bayes bekomme ich die *a posteriori* Verteilung

$$P(\mu|D) = \frac{P(D|\mu)P(\mu)}{P(D)} = \mathcal{N}\left(\mu; \frac{mean}{1 + \frac{\sigma^2}{N\alpha^2}}, \frac{\sigma^2}{N + \sigma^2/\alpha^2}\right)$$

mit $mean = 1/N \sum_{i=1}^{N} x_i$


$$P(\mu) \propto \mathrm{N}(0, \alpha^2)$$

$$P(\mu \mid D) \propto \mathrm{N}\left( \frac{mean}{1 + \frac{\sigma^2}{N\alpha^2}}, \frac{\sigma^2}{N + \sigma^2 / \alpha^2} \right)$$

***Das Bayes'sche***

***Experiment***

# Analyse

- In der Bayesschen Statistik bekommt man eine klare Aussage über die Verteilung der Parameter, nachdem die Daten gemessen wurden; der frequentistische Ansatz liefert nur einen Schätzer
- Man kann aus dem Bayesschen Ansatz den Maximum *a posteriori* Schätzer ableiten mit

$$\hat{\mathbf{w}}_{map} \doteq \arg\max(P(\mathbf{w}|D))$$

  Im Beispiel ist

$$\hat{\mu}_{MAP} = \frac{mean}{1 + \frac{\sigma^2}{N\alpha^2}}$$

- Beachte, dass der MAP Schätzer für $N \longrightarrow \infty$ zum ML Schätzer konvergiert

## Lineare Regression: Likelihood

- Nehmen wir an, dass die wahre Abhängigkeit linear ist, wir jedoch nur verrauschte Daten zur Verfügung haben

$$y_i = \mathbf{x}_i^T \mathbf{w} + \epsilon_i$$

Wie zuvor erhalten wir

$$P(D|\mathbf{w}) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2\right)$$

# Lineare Regression: a priori Annahme

- Eine typische *a priori* Annahme ist, dass

$$P(\mathbf{w}) = (2\pi\alpha^2)^{-M/2} \exp\left(-\frac{1}{2\alpha^2}\sum_{i=0}^{M-1} w_i^2\right)$$

- Diese Annahme gibt kleineren Gewichten eine höhere *a priori* Wahrscheinlichkeit
- Wir werden im folgenden annehmen, dass sogenannte Hyperparameter wie die Rauschvarianz $\sigma^2$ und $\alpha^2$ bekannt sind; sind diese nicht bekannt, so definiert man *a priori* Verteilungen über diese Größen; der Bayessche Programm wird auf dieses Modell angewandt, d.h. es wird entsprechend komplexer
- Ockhams Rasiermesser: einfache Erklärungen sind komplexeren Erklärungen vorzuziehen

## Lineare Regression: die a posteriori Verteilung

- Aus der Likelihood-Funktion, der *a priori* Verteilungsannahme über die Parameter lässt sich mit dem Satz von Bayes die *a posteriori* Verteilung über die Parameter berechnen

$$P(\mathbf{w}|D) = \frac{P(\mathbf{w})P(D|\mathbf{w})}{P(D)}$$

## Lineare Regression: Berechnung der a posteriori Verteilung

$$P(\mathbf{w}|D) = \frac{P(\mathbf{w})P(D|\mathbf{w})}{P(D)} \propto \exp\left(-\frac{1}{2\alpha^2}\sum_{j=0}^{M-1} w_i^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2\right)$$

$$P(\mathbf{w}|D) = \mathcal{N}\left(\mathbf{w}; \hat{\mathbf{w}}_{map}, cov(\mathbf{w}|D)\right)$$

Mit

$$\hat{\mathbf{w}}_{map} = \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha^2}I\right)^{-1}\mathbf{X}^T\mathbf{y}$$

und Varianz

$$cov(\mathbf{w}|D) = \sigma^2\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha^2}I\right)^{-1}$$

## Lineare Regression: die Bayessche Lösung und die RLS-Lösung

- Somit erhalten wir für den wahrscheinlichsten Parameterwert nach Erhalt der Daten (die maximum *a posteriori* (MAP) Schätzung)

$$\hat{\mathbf{w}}_{map} \doteq \arg\max(P(\mathbf{w}|D)) = \hat{\mathbf{w}}_{Pen}$$

  mit $\lambda = \frac{\sigma^2}{\alpha^2}$.

- Dies bedeutet, dass trotz unterschiedlicher Herangehensweise der Frequentisten und der Bayesianer die Ergebnisse sehr ähnlich sind; die MAP Schätzung entspricht der RLS-Schätzung

## Lineare Regression: Vorhersage mit der Bayesschen Lösung

- Soweit scheint der Unterschied zwischen der Bayesschen Lösung und der frequentistischen Lösung nicht allzu groß zu sein: beide liefern Schätzer mit Unsicherheit
- Ein wesentlicher Unterschied besteht in der Prognose für einen neuen Eingang. Während in der frequentistischen Lösung der Schätzer eingesetzt wird, $\widehat{y}_i = \mathbf{x}_i^T \mathbf{w}_{ml}$, wird in der Bayesschen Lösung mit den Regeln der Wahrscheinlichkeitslehre gearbeitet
- Mit

$$P(y, \mathbf{w}|x, D) = P(\mathbf{w}|D)P(y|\mathbf{w}, \mathbf{x})$$

folgt

$$P(y|\mathbf{x}, D) = \int P(\mathbf{w}|D)P(y|\mathbf{w}, \mathbf{x})d\mathbf{w}$$

# Prädiktive Verteilung für das lineare Modell

- *A posteriori* wird die prädiktive Verteilung

$$P(y|\mathbf{x}, D) = \int P(y|\mathbf{w}, \mathbf{x}) P(\mathbf{w}|D) d\mathbf{w} = \mathcal{N}(y|\mathbf{x}^T \widehat{\mathbf{w}}_{map}, \mathbf{x}^T cov(w|D)\mathbf{x} + \sigma^2)$$

  ist Gauß-verteilt mit Mittelwert $\mathbf{x}^T \widehat{\mathbf{w}}_{map}$ und Varianz $\mathbf{x}^T cov(w|D)\mathbf{x} + \sigma^2$

- Die Vorhersage berücksichtigt sowohl die das Rauschen auf den Daten als auch die Unsicherheit im Parametervektor!

- Beachte, dass in der Vorhersage über alle möglichen Parameterwerte integriert wird

- Dies ist ein wesentlicher Vorteil des Bayesschen Ansatzes: er berücksichtigt nicht nur den wahrscheinlichsten Parameterwert sondern wertet auch die Parameterverteilung aus; dadurch können zum Beispiel auch Nebenoptima in der Lösung berücksichtigt werden!

- Dies ist jedoch auch ein wesentliches technisches Problem der Bayesschen Lösung: zur Prognose müssen komplex integrale gelöst, b.z.w. approximiert werden!

# Lineare Regression: die Bayessche Lösung

- Persönlicher Belief wird als Wahrscheinlichkeit formuliert
- Vorwissen kann konsistent integriert werden
- Konsistenter Umgang mit den verschiedenen Formen der Modellierungsunsicherheit
- Bayessche Lösungen führen zu Integralen, die in der Regel nicht analytisch lösbar sind
- Im Folgenden werden wir spezielle Näherungen kennen lernen (Monte-Carlo Integration, Evidence Framework)
- Die vielleicht einfachste Näherung ist

$$P(y|\mathbf{x}, D) = \int P(y|\mathbf{w}, \mathbf{x}) P(\mathbf{w}|D) d\mathbf{w} \approx P(y|\mathbf{x}, \mathbf{w}_{map})$$

d.h. man macht eine Punktschätzung des unbekannten Parameters und setzt dies in das Modell ein (analog zum frequentistischen Ansatz)

# APPENDIX: Likelihood und Entropie

- Der Abstand zweier Verteilungen wird durch die relative Entropie oder die Kullack-Leibler Divergenz bestimmt

$$D(P\|Q) = \int P(x) \log \frac{P(x)}{Q(x)}\, dx$$

  Dieses ist nur Null, wenn $P = Q$

- Die relative Entropie zwischen der wahren unbekannten Verteilung $P(y)$ und der approximativen Verteilung $P(y|\mathbf{w})$ ist

$$\int P(y) \log \frac{P(y)}{P(y|\mathbf{w})}\, dy = const - \int P(y) \log P(y|\mathbf{w})$$

$$\approx const - \frac{1}{N} \sum_{i=1}^{N} \log P(y_i|\mathbf{w}) = const - \frac{1}{N} l(\mathbf{w})$$

- Dies bedeutet, dass der ML-Ansatz asymptotisch $N \to \infty$ diejenige Verteilung findet, die der wahren Verteilung in Bezug auf die relative Entropie am ähnlichsten ist

- Der beste Fit ist die wahre Verteilung selber, für die gilt für $N \longrightarrow \infty$

$$\frac{1}{N} l(\mathbf{w}) \rightarrow Entropy(Y)$$

# APPENDIX: Hypothesentests

- Beispiel: ist ein Parameter von Null verschieden?
- Nullhypothese: $H_0 : \mu = 0$, Alternativhypothese: $H_a : \mu \neq 0$
- Teststatistik: normierter Mittelwert $z = \frac{\hat{\mu}}{\sigma^2/N}$;
- Die Nullhypothese soll verworfen werden, wenn $|z| > 2.58$; dann ist die Wahrscheinlichkeit, dass die Nullhypothese verworfen wird, obwohl sie wahr ist 0.01% (Fehler erster Art). Die Wahrscheinlichkeit des Fehlers erster Art wird als $\alpha$ bezeichnet. Hier im Beispiel:

$$\alpha = 0.01 = 2 \int_{x=z_{1-\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx$$

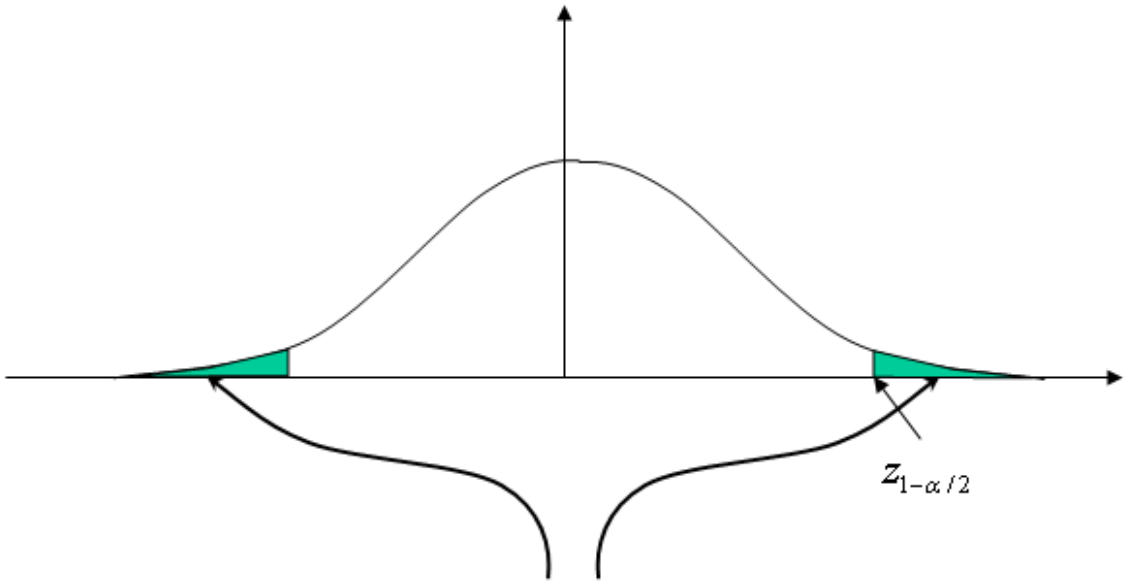
- $z_{1-\alpha/2}$ findet man in Standardtabellen
- Der Fehler zweiter Art ist die Annahme der Nullhypothese, obwohl die alternative Hypothese wahr ist; dieser Fehler ist häufig schwer oder unmöglich zu berechnen; Die

Wahrscheinlichkeit des Fehlers zweiter Art wird als $\beta$ bezeichnet. Die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, wenn sie tatsächlich falsch $1 - \beta$ ist die Mächtigkeit des Tests

- Um einen Eindruck zu gewinnen über den Fehler 2ter Art kann man die Gütefunktion berechnen $g(\mu) = 1 - P(H_0 \text{ wird angenommen}|\mu)$. Dies gibt einen Eindruck vom Fehler 2ter Art für die möglichen Werte der Parameter der alternativen Hypothese; für einen guten Test ist die Gütefunktion nahe 1.

- Der p-Wert ist der beobachtete Signifikanzlevel, der definiert ist als die Wahrscheinlichkeit, unter $H_0$ den beobachteten Prüfwert $z$ zu erhalten. Im Beispiel ist der p-Wert das $\alpha$, für welches $z = z_{1-\alpha/2}$

# ***Hypothesentest***



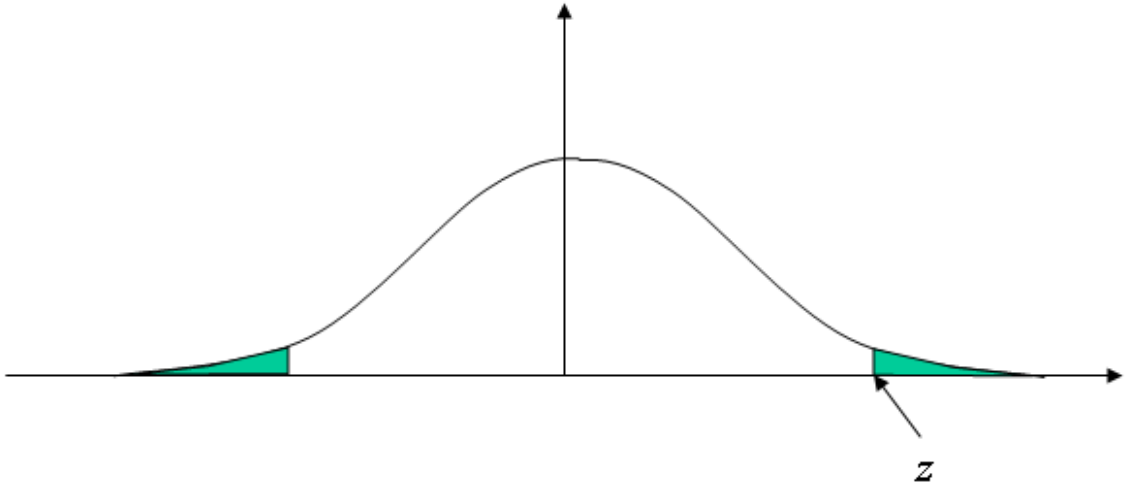

Wenn $\hat{\mu}$ in die grüne Regionen fällt, lehne ich die Null-Hypothese $\mu = 0$

ab. Im Beispiel:

$$\alpha = \int gruen = 0.01 = 2 \int_{z_{1-\alpha/2}}^{\infty} p(z)dz$$

$$z = \frac{\hat{\mu}}{\sigma^2 / N}$$

## ***P-Wert***




$$\int gruen = \alpha_{obs} = \text{pWert} = 2\int_{z}^{\infty} p(z)dz$$