Frequentisten und Bayesianer

Volker Tresp

Frequentisten

Die W-Verteilung eines Datenmusters

- 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 Gaussverteilung 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

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

- Leichter handhabbar ist der Logarithmus dieses Ausdrucks

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

Die Likelihood

- Betrachten wir nun den ganzen Datensatz, so nimmt man häufig an, dass das Rauschen unabhängig ist und man die Wahrscheinlichkeit des Datensatzes gegeben die Parameter als Produkt schreiben läßt

$$P(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{i=1}^{N} P(y_i|\mathbf{x}_i,\mathbf{w})$$

- Diesen Ausdruck bezeichnet man als Likelihood des Modells

$$L(\mathbf{w}) \doteq P(\mathbf{y}|\mathbf{X},\mathbf{w})$$

- Entsprechend ergibt sich die Log-Likelihood zu

$$l(\mathbf{w}) \doteq \log L(\mathbf{w})$$

- Nehmen wir nun unabhängiges Gaussrauschen an, so ergibt sich

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

Der Maximum-Likelihood Schätzer

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

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

Das frequentistische Experiment

- Die Likelihoodfunktion ist bekannt
- Die Natur wählt einen wahren Parametervektor w
- Die Natur generiert unendlich viele Datensätze (Sticproben) $D_1, D_2, \ldots D_L$, $L \rightarrow \infty$, jeder der Größe N
- Für jeden dieser Datensätze D_i berechne ich einen Parameterschätzer $\widehat{w}_i$ (zum Beispiel den ML-Schätzer)
- Man interessiert sich nun zum Beispiel, ob dieser Schätzer, gemittelt über alle Datensätze gleich dem wahren Schätzer ist

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

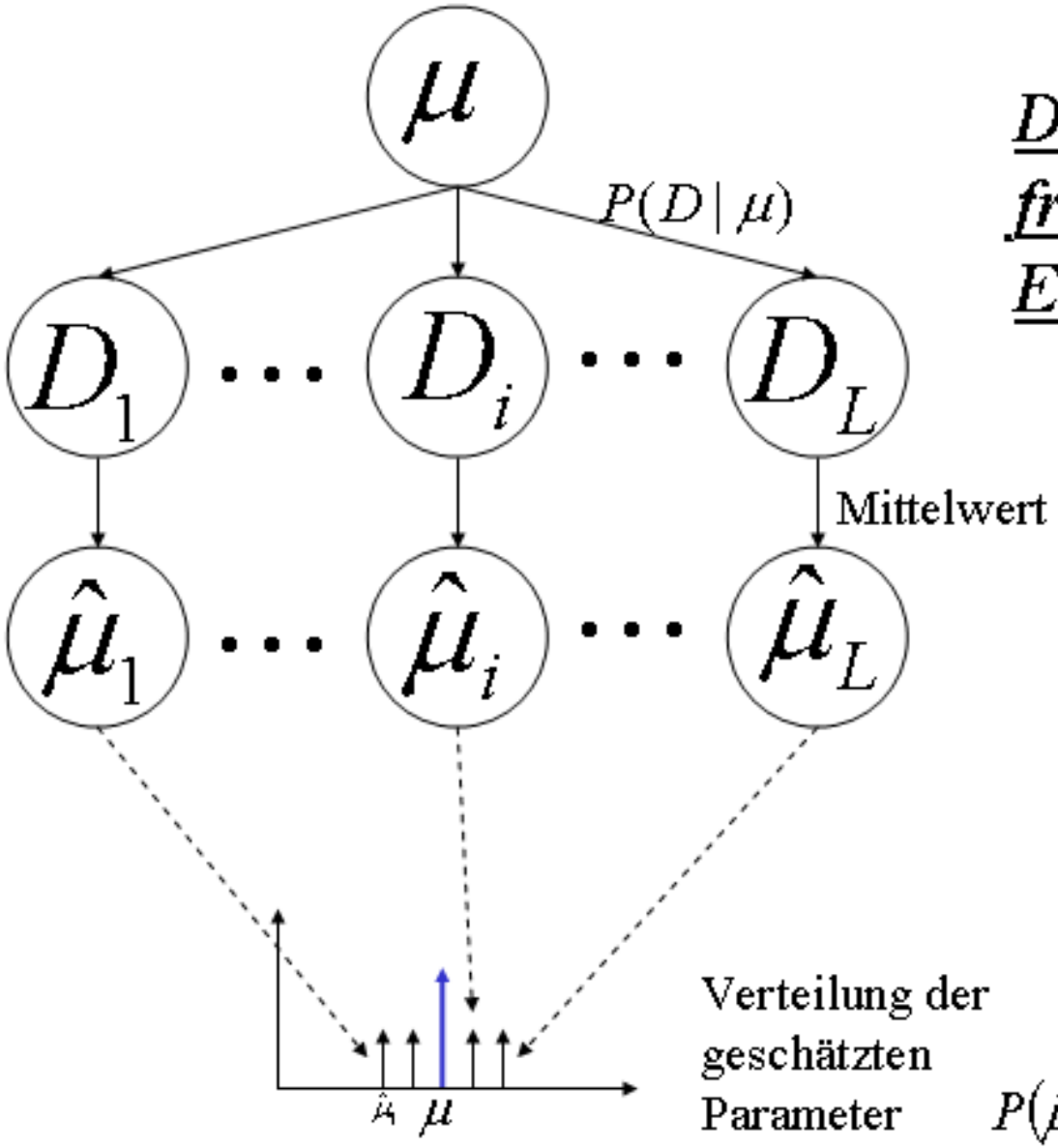
- Wenn $E_D(\widehat{w}) = w$, dann nennt man diesen Schätzer (unverzerrt) unbiased oder erwartungstreu

- Der Bias ist definiert als

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

- Ein Schätzer ist **asymptotisch erwartungstreu**, wenn:

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



Das frequentistische Experiment

Verteilung der geschätzten Parameter

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

Konsistenz

- Varianz eines Schätzers

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

- Erwarteter mittlerer quadratischer Fehler

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

- MSE-Konsistenz: $MSE_{N\to\infty} \to 0$
- $\hat{w}(1)$ ist MSE-wirksamer als $\hat{w}(2)$ falls

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

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

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

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 $N \longrightarrow \infty$
  - unverzerrt (unbiased)
  - MSE-wirksamst (efficient) unter allen asymptotisch unverzerrten Schätzern
  - Gauss-verteilt

Verzerrtheit des ML-Schätzers bei endlichen Daten

- Für endliche Daten können ML-Schätzer verzerrt sein

$$\widehat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\widehat{\mathbf{w}}_{LS})^2)$$

$$\widehat{\sigma}^2_{u} = \frac{1}{N-M-1}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\widehat{\mathbf{w}}_{LS})^2$$

Diskussion: ML

- Auch für komplexere Modelle lässt sich die (Log)-Likelihood in der Regel berechnen (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 endlichen Daten führt der ML-Schätzer zum Überfitten, d.h. komplexere Modelle werden bevorzugt (siehe ERM)
- Die frequentistische Statistik ist stark fokussiert auf die Eigenschaften von Parametern (Signifikanz, ...)
- Zu den Themen Hypothesentests und p-Werte kann man etwas im Appendix finden

Bayesianer

Der Bayes'sche Ansatz

- Der Bayes'sche Ansatz unterscheidet sich zunächst sehr vom frequenstistischen Ansatz
- Der wesentliche Unterschied ist, dass auch Parameter als Zufallsvariable behandelt werden
- Dies bedeutet, dass der Benutzer zunächst eine a prior Annahme über die Verteilung der Parameter machen muss:

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

- Man erhält ein komplettes probabilistisches Modell

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

- In diesem Modell muss man keine Parameter schätzen; Lernen reduziert sich auf probabilistische Inferenz

- Sind Daten D gemessen worden, wird nicht ein bester Parametervektor w_{opt} bestimmt, sondern es wir die Parameterverteilung nach der Bayes'schen Regel adaptiert

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

- **A priori** ist die Prognose

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

dann ist diese nach dem Empfang der Daten (**a posteriori**)

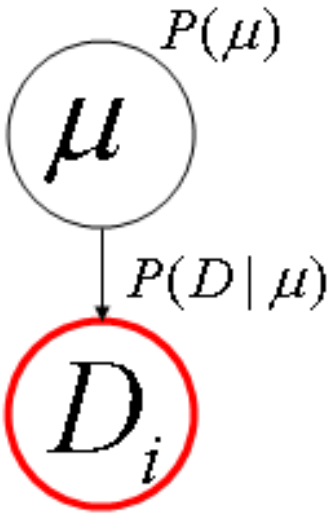
$$P(y|\mathbf{x},D) = \int \frac{P(\mathbf{w})P(D|\mathbf{w})P(y|\mathbf{x},\mathbf{w})d\mathbf{w}}{P(D)} = \int P(\mathbf{w}|D)P(y|\mathbf{w},\mathbf{x})d\mathbf{w}$$

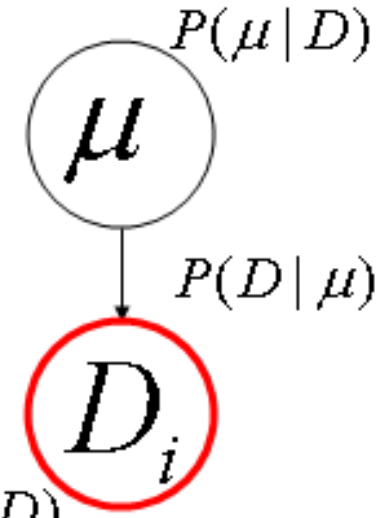
- Anstelle von Optimierungsproblemen muss man Integrale lösen

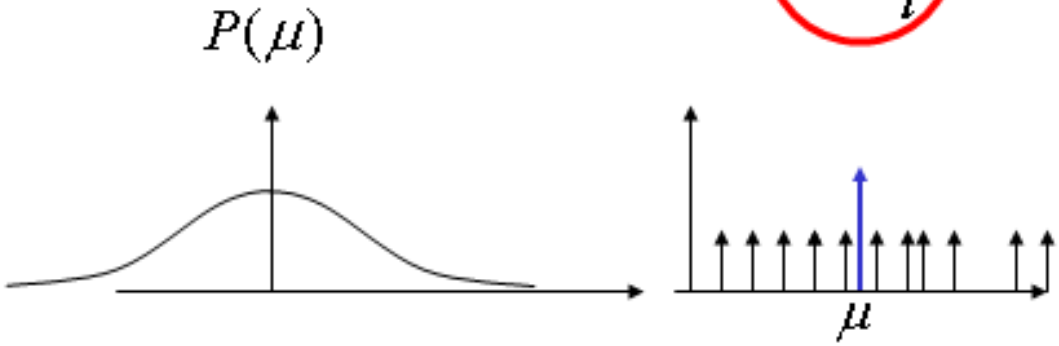
Das Bayes'sche Experiment

- Die Likelihoodfunktion ist bekannt
- Der Prior $P(\mathbf{w})$ ist bekannt
- Die Natur wählt einen wahren Parametervektor $\mathbf{w}$
- Die Natur generiert einen Datensatz der Größe N
- Man berechnet und wertet aus

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







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

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

Das Bayes'sche

Experiment

Lineare Regression: die Bayes'sche Lösung (2)

- Eine typische prior 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 σ^2 und α^2 bekannt sind; sind diese nicht bekannt, so definiert man a priori Verteilungen über diese Größen; der Bayes'sche Programm wird auf dieses Modell angewandt, d.h. es wird entsprechend komplexer
- Occam's Razor: einfache Erklärungen sind komplexeren Erklärungen vorzuziehen

Lineare Regression: die Bayes'sche Lösung (3)

- Likelihood Funktion, wie zuvor

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

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

Lineare Regression: die Bayes'sche Lösung (4)

- Aus der Likelihood-Funktion, der a priori Verteilungsannahme über die Parameter läßt sich mit der Bayes'schen Formel die a posteriori Verteilung über die Parameter berechnen

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

Lineare Regression: die Bayes'sche Lösung (5)

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

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

$$= const - \sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2 - \frac{\sigma^2}{\alpha^2}\sum_{j=0}^{M-1} w_i^2$$

Lineare Regression: die Bayes'sche Lösung (6)

- 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 regularisierten LS-Schätzung (**Penalized Least Squares**)!

Lineare Regression: die Bayes'sche Lösung

- Wir können nun auch die gesamte Verteilung berechnen

  A posterior ist der Parametervektor gaussverteilt

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

mit Mittelwert

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

und Varianz

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

Prädiktive Verteilung

- Man kann ebenfalls die Verteilung einer Vorhersage berechnen
- A posteriori

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

  ist gaussverteilt mit Mittelwert $\mathbf{x}^T \hat{\mathbf{w}}_{MAP}$ und Varianz $\mathbf{x}^T cov(w|D)\mathbf{x} + \sigma^2$
- Beachte, dass in der Vorhersage über alle möglichen Parameterwerte integriert wird
- Dies ist ein wesentlicher Vorteil des Bayes'schen 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 Bayes'schen Lösung: zur Prognose müssen komplex integrale gelöst, b.z.w. approximiert werden!

Lineare Regression: die Bayes'sche Lösung

- Persönlicher Belief wird als Wahrscheinlichkeit formuliert
- Vorwissen kann konsistent integriert werden
- Konsistenter Umgang mit den verschiedenen Formen der Modellierungsunsicherheit
- Bayes'sche 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 genau dann Null, wenn $P = Q$

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

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

- Zeigen Sie, dass die ML-Lösung für $N \to \infty$ die relative Entropie minimiert

- Berechnen Sie die Log-Likelihood für $N \to \infty$ für die wahre Verteilung. Wie nennt man diesen Ausdruck?

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 α bezeichnet. Hier im Beispiel:

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

- $z_{1-\alpha/2}$ 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

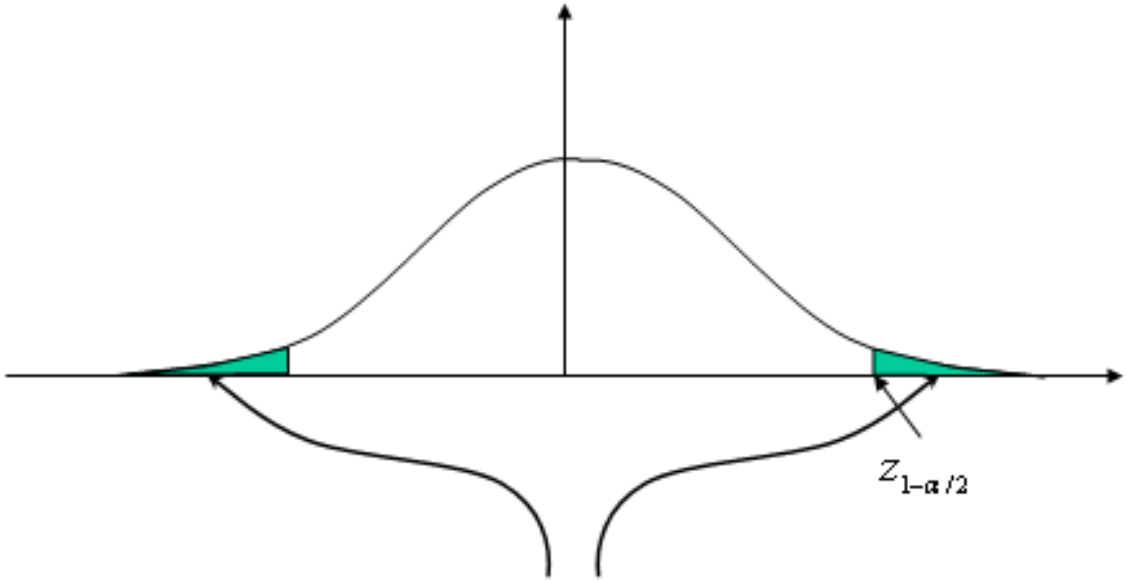
Wahrscheinlichkeit des Fehlers zweiter Art wird als β bezeichnet. Die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, wenn sie tatsächlich falsch $1 - \beta$ ist die Mächtigkeit des Tests

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

$$P(z = \frac{\hat{\mu}}{\sigma^2 / N} \mid \mu = 0)$$

Hypothesentest



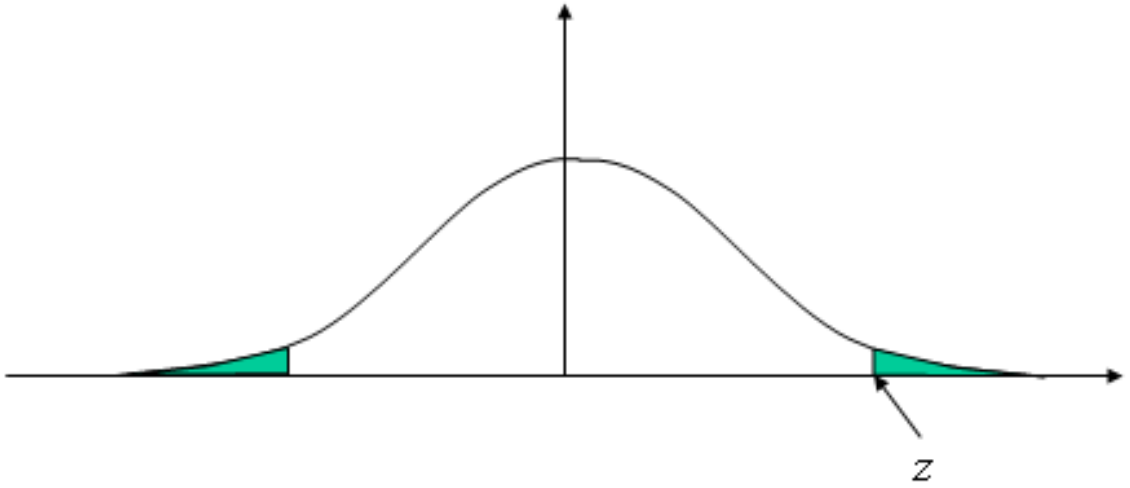
Wenn $\hat{\mu}$ in die grüne Regionen fällt, lehne ich die Null-Hypothese $\mu = 0$

ab. Im Beispiel:

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

$$P(z = \frac{\hat{\mu}}{\sigma^2 / N} \mid \mu = 0)$$

P-Wert



$$\int gruen = p - Wert = 2\int_{z}^{\infty} p(z)dz$$