

2. Teilprüfung Statistik und Wahrscheinlichkeitsrechnung

FS 2010

Prof. Dr. Michael Havbro Faber

ETH Zürich

27.5. 2010
08:00 – 09:30

Vorname:

Name:

Stud. Nr.:

Studienrichtung:

2. Teilprüfung: Statistik und Wahrscheinlichkeitsrechnung Bau-, Umwelt- und Geomatikingenieurwissenschaften

Datum und Dauer:

Donnerstag, 27. Mai 2010

Beginn: 8:00 Uhr

Zeitdauer: 90 Minuten

Hilfsmittel:

- Alle Unterlagen (Skripte, Bücher, andere Ausdrücke, etc.) sind erlaubt.
- Taschenrechner (ohne Kommunikationsmittel) sind erlaubt.
- Kommunikationsmittel (z.B. Telefon) sind nicht erlaubt.

Hinweise:

- Bitte kontrollieren Sie zuerst, ob Sie das Material vollständig erhalten haben:
 - Aufgabenstellung inkl. genereller Information (15 Seiten).
 - Papierbogen kariert und gestempelt (1 mal).
- Bitte legen Sie Ihre Legi vor sich auf den Tisch.
- Alle Lösungsblätter müssen mit Namen und Vornamen versehen werden.
- Nur die zur Verfügung gestellten Blätter dürfen verwendet werden.
- Legen Sie am Ende der Prüfung alle Aufgaben- und Lösungsblätter in das Couvert zurück und lassen Sie dieses am Platz liegen.
- Wenn Sie vor 9:00 Uhr fertig sind, dann benachrichtigen Sie einen Assistierenden; er/sie wird dann Ihre Prüfung einsammeln. Sie dürfen bis 9:00 Uhr den Saal verlassen; danach warten Sie bitte, bis die Prüfung zu Ende ist (9:30 Uhr).

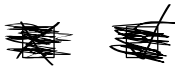
Teil 1: Multiple Choice (maximal 54 Punkte)

In den folgenden Multiple Choice Fragen können für die gleiche Frage (mindestens) eine oder mehrere Antworten zutreffend sein.

Bitte markieren Sie alle richtigen Antworten mit einem Häkchen oder Kreuz:



Wenn Sie ein bereits markiertes Kästchen rückgängig machen wollen, dann tun Sie das bitte deutlich:



Auf jede Aufgabe bzw. Teilaufgabe werden maximal 2 Punkte vergeben.

1.1 Die Zufallsvariable Y ist als eine Summe von Zufallsvariablen X_i definiert:

$$Y = \sum_{i=1}^{10000} X_i .$$

Die Zufallsvariablen X_i sind voneinander unabhängig und gleichverteilt. Welcher Verteilung folgt Y ?

- Normalverteilung
- Gleichverteilung
- Lognormalverteilung
- Keiner Verteilung

1.2 Welche der folgenden Aussagen ist/sind richtig?

Die Fläche unter der Dichtefunktion einer Normalverteilung im Intervall von $[m-s \leq X \leq m+s]$ entspricht 13.5 % der Gesamtfläche.

Die Fläche unter der Dichtefunktion einer Normalverteilung im Intervall von $[m-s \leq X \leq m+s]$ entspricht 63.8 % der Gesamtfläche.

Die Parameter der Standardnormalverteilung sind $m = 1, s = 1$

Die Parameter der Standardnormalverteilung sind $m = 0, s = 1$

1.3 Welche der folgenden Aussagen ist/sind richtig?

Bei Bernoulli-Versuchen gibt es nur zwei sich gegenseitig ausschliessenden mögliche Ergebnisse.

Die Wahrscheinlichkeit $p_Y(y)$ der Anzahl an Erfolgen y in n unabhängigen Bernoulli-Versuchen kann mit der Binomialverteilung beschrieben werden, wenn die Erfolgswahrscheinlichkeit der Versuche konstant ist.

Die Wahrscheinlichkeit $p_Y(y)$ der Anzahl an Erfolgen y in n unabhängigen Bernoulli-Versuchen kann mit der Binomialverteilung beschrieben werden, wenn die Erfolgswahrscheinlichkeit der Versuche mit steigender Anzahl an Versuchen ansteigt.

Bei der Binomialverteilung sinkt mit zunehmender Anzahl an Versuchen die Wahrscheinlichkeit, dass die Anzahl der Erfolge Null ist.

1.4 Wie gross ist die Wahrscheinlichkeit $p_Y(y)$, dass man mit einem Würfel nach 5 Versuchen genau zwei mal eine Augenzahl grösser als 4 gewürfelt hat?

$p_Y(y) = 0.5$

$p_Y(y) = \binom{5}{2} (2/6)^2 (1 - (2/6))^{5-2} = 0.329$

$p_Y(y) = \binom{10}{2} (2/6)^2 (1 - (2/6))^{10-2} = 0.195$

$p_Y(y) = \binom{5}{3} (2/6)^3 (1 - (2/6))^{5-3} = 0.165$

1.5 Welche der folgenden Aussagen trifft / treffen zu?

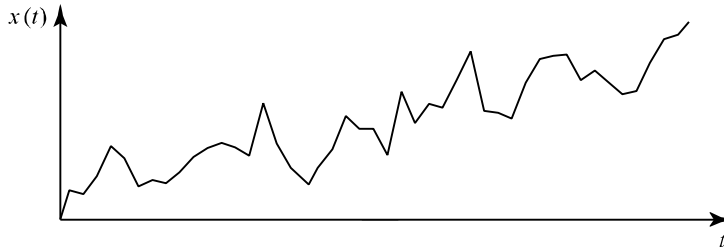
Ein Poissonprozess heisst homogen, falls seine Intensität $\nu(t)$ konstant ist.

Bei Poissonprozessen ist die Zeit bis zum ersten Ereignis exponentialverteilt.

Bei Poissonprozessen ist die Zeit zwischen zwei Ereignissen exponentialverteilt.

Der Poissonprozess ist ein diskreter Zufallsprozess.

1.6 Die folgende Abbildung zeigt eine Realisation eines stochastischen Prozesses:



Welche der folgenden Aussagen trifft / treffen auf die oben dargestellte Abbildung zu?

- Streng stationärer stochastischer Prozess.
- Stationärer stochastischer Prozess.
- Nicht stationärer stochastischer Prozess.
- Man kann keine Aussage über die Stationarität treffen.

1.7 Eine Wetterstation zeichnet die Windgeschwindigkeiten an einer bestimmten Stelle auf. Nach Auswertung der Daten liegt der Stichprobenmittelwert der maximalen Windgeschwindigkeit pro Monat bei 14 m/s. Aus denselben Daten soll nun der Stichprobenmittelwert der maximalen Windgeschwindigkeit pro Jahr bestimmt werden. Welche der folgenden Stichprobenmittelwerte kann / können ausgeschlossen werden?

- Der Stichprobenmittelwert der maximalen Windgeschwindigkeit pro Jahr ist 16 m/s.
- Der Stichprobenmittelwert der maximalen Windgeschwindigkeit pro Jahr ist 20 m/s.
- Der Stichprobenmittelwert der maximalen Windgeschwindigkeit pro Jahr ist 12 m/s.
- Es kann keine Aussage über die maximalen Windgeschwindigkeit pro Jahr getroffen werden.

1.8 Welche Anforderungen muss die Wahrscheinlichkeitsdichtefunktion einer Zufallsvariablen X besitzen, damit die Verteilungsfunktion der Extrema während einer Referenzperiode T einer Gumbelverteilung für extreme Maxima (Gumbel max) folgen kann.

- Nach oben unbeschränkt
- Nach unten unbeschränkt
- Exponentielle Abnahme im oberen Bereich
- Exponentielle Abnahme im unteren Bereich

1.9 Die jährliche Wahrscheinlichkeit für die Überschreitung eines bestimmten Abflusses $Q = 25 [m^3 / s]$ beträgt 0.001. Wie gross ist die Wiederkehrperiode dieses Ereignisses?

- 10 Jahre
- 1000 Jahre
- 2000 Jahre
- 500 Jahre

1.10 Welche der folgenden Aussagen ist/sind richtig?

Wahrscheinlichkeitspapier wird verwendet, um die Korrelation zwischen zwei Datensätze abzuschätzen.

Mit Wahrscheinlichkeitspapier lässt sich abschätzen, ob es plausibel ist, dass die Daten einer gewählten Verteilungsfamilie folgen.

Um Wahrscheinlichkeitspapier zu erstellen, muss immer die Wurzel aus der Verteilungsfunktion gezogen werden.

Um Wahrscheinlichkeitspapier zu erstellen, muss die Umkehrfunktion der Verteilungsfunktion gebildet werden.

1.11 Es soll ein Modell für die Körpergrössen aller Studenten erstellt werden. Aus unserer Stichprobe sind der Mittelwert $\bar{x} = 178$ cm und die Standardabweichung $s_x = 18$ cm der Körpergrösse bekannt. Sie bestimmen den Parameter I einer Exponential-Verteilung unter Verwendung der Methode der Momente. Welche der folgenden Aussagen ist/sind richtig?

Der Parameter I berechnet sich zu $m_1 = \bar{x} = I = 178$ cm.

Der Parameter I berechnet sich zu $m_1 = \bar{x} = \frac{1}{I}$, $I = \frac{1}{178 \text{ cm}}$.

Der Parameter I berechnet sich zu $\sqrt{m_2 - m_1^2} = s_x = I = 18$ cm.

Der Parameter I berechnet sich zu $m_2 - m_1^2 = s_x^2 = \frac{1}{I} = (18 \text{ cm})^2$, $I = 0.0031$ cm

1.12 Welche der folgenden Aussagen ist/sind richtig?

Mit der Methode der Momente lässt sich die Unsicherheit der Parameterschätzung
ermitteln.

Mit der Maximum-Likelihood-Methode werden die Parameter einer gewählten
Verteilung so geschätzt, dass die Stichproben-Likelihood maximal wird.

Bei der Parameterschätzung unter Verwendung der Maximum-Likelihood-Methode
werden die Stichprobenwerte maximiert.

Bei der Parameterschätzung unter Verwendung der Methode der Momente werden
die Momente der Stichprobe mit den Momenten der Verteilungsfunktion gleich
gesetzt.

1.13 Welcher der folgenden Terme muss berechnet werden, um die Parameter einer
Lognormalverteilung unter Verwendung der Maximum-Likelihood-Methode zu
ermitteln?

$$\sum_{i=1}^n \frac{1}{\hat{x}_i z \sqrt{2p}} \exp\left(-\frac{1}{2} \left(\frac{\ln(\hat{x}_i) - l}{z}\right)^2\right) \quad \input type="checkbox"/>$$

$$\min_{l,z} \left(\prod_{i=1}^n \frac{1}{\hat{x}_i z \sqrt{2p}} \exp\left(-\frac{1}{2} \left(\frac{\ln(\hat{x}_i) - l}{z}\right)^2\right) \right) \quad \input type="checkbox"/>$$

$$\max_{l,z} \left(\prod_{i=1}^n \frac{1}{\hat{x}_i z \sqrt{2p}} \exp\left(-\frac{1}{2} \left(\frac{\ln(\hat{x}_i) - l}{z}\right)^2\right) \right) \quad \input type="checkbox"/>$$

$$\prod_{i=1}^n \frac{1}{s \sqrt{2p}} \exp\left(-\frac{1}{2} \left(\frac{\hat{x}_i - m}{s}\right)^2\right) \quad \input type="checkbox"/>$$

1.14 Mit der Maximum-Likelihood-Methode kann unter Verwendung der Fisher-
Informationsmatrix neben den Mittelwerten der geschätzten Parameter

die Unsicherheit der geschätzten Parameter bestimmt werden.

die Wölbung der Stichproben ermittelt werden.

die Kovarianz zwischen den geschätzten Parameter ermittelt werden.

die aleatorische Modellunsicherheit bestimmt werden.

1.15 Die Körpergrösse der Studenten wird als normalverteilt angenommen. Der Mittelwert m und die Standardabweichung s wurden anhand $n' = 243$ Messungen aus dem Jahr 2009 geschätzt. Durch die Berücksichtigung der statistischen Unsicherheit wird der Mittelwert als normalverteilt angenommen. Die Verteilung des Mittelwert m soll mit $n = 265$ neuen Messungen aus dem Jahr 2010 aktualisiert werden. Welche der folgenden Aussagen ist/sind richtig?

Durch das Aktualisieren mit den neuen Daten bleibt der Mittelwert des Mittelwert m gleich und die Standardabweichung des Mittelwertes wird kleiner.

Durch das Aktualisieren können sich Standardabweichung und Mittelwert des Mittelwertes m verändern.

Ein einzelner Messwert aus dem Jahr 2009 hat einen grösseren Einfluss auf das Modell als einer aus dem Jahr 2010, da die Anzahl der Messungen aus dem Jahr 2009 kleiner ist als aus dem Jahr 2010 ($n' < n$).

Ein einzelner Messwert aus dem Jahr 2009 hat genau den gleichen Einfluss auf das Modell wie ein Messwert aus dem Jahr 2010.

1.16 Die Parameter b_0 und b_1 für das Regressionsmodell $y = b_0 + b_1x + e$ werden so gewählt, dass

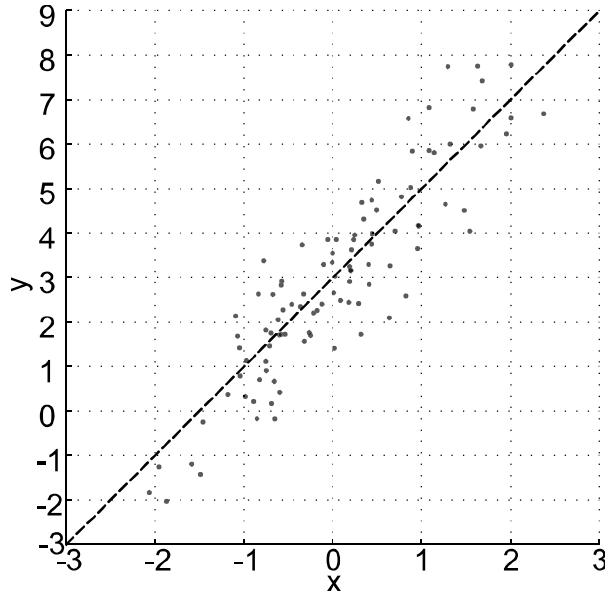
die Quadrate der Differenzen zwischen den gemessenen Werten und dem Regressionsmodell möglichst gross sind.

die Quadrate der Differenzen zwischen den gemessenen Werten und dem Regressionsmodell möglichst klein sind.

das Quadrat der gemessenen Werte möglichst klein wird.

die Stichproben-Likelihood für das Regressionsmodell maximal wird.

1.17 Die folgende Abbildung enthält die grafische Darstellung einer linearen Regression:



Schätzen Sie anhand der Abbildung die Werte der Regressionsparameter b_0 und b_1 für das Regressionsmodell $y = b_0 + b_1x + e$.

- $y = 3 + x + e$
- $y = -3 + x + e$
- $y = 3 + 2x + e$
- $y = -3 + 2x + e$

1.18 Der Residualwert eines Regressionsmodells, welcher durch die normalverteilte Zufallsvariable e repräsentiert wird, ...

- ist der Mittelwert der Abweichungen der gemessenen Werte zum Regressionsmodell.
- ist die Summe aller Abweichungen zwischen den gemessenen Werten und dem Regressionsmodell.
- hat einen Mittelwert von b_1 .
- hat einen Mittelwert von 0.

1.19 Die drei Zufallsvariablen X_1 , X_2 und X_3 seien voneinander unabhängig und standardnormalverteilt. Die zusammengesetzte Zufallsvariable Y ist folgendermassen definiert: $Y = (X_1)^2 + (X_2)^2 + (X_3)^2$

a) Welche der folgenden Aussagen ist / sind richtig?

Y ist Chi-verteilt mit 2 Freiheitsgraden.

Y ist Chi-Quadrat-verteilt mit 3 Freiheitsgraden.

Für den Wertebereich von Y gilt $y \geq 0$.

Der Mittelwert m_Y der Zufallsvariablen Y ist Null.

b) Welche der folgenden Aussagen ist / sind richtig?

Die Summe von zwei Chi-Quadrat-verteilten Zufallsvariablen ist ebenfalls Chi-Quadrat-verteilt.

Durch Quadrieren einer Chi-Quadrat-verteilten Zufallsvariable ergibt sich eine Chi-verteilt Zufallsvariable.

Für eine grosse Anzahl Freiheitsgrade nähert sich die Chi-Quadrat-Verteilung einer Normalverteilung an.

Mittelwert und Standardabweichung einer Chi-Verteilung berechnen sich aus der Anzahl Freiheitsgrade der Verteilung.

1.20 X sei eine lognormalverteilte Zufallsvariable. Sie planen Versuche, mit denen Sie einen Datensatz mit $n=1000$ Beobachtungen der Zufallsvariablen X erhalten.

Welche der folgenden Aussagen ist / sind richtig?

Der Stichprobenmittelwert der Versuchsergebnisse, \bar{X} , ist näherungsweise normalverteilt.

Nach den Versuchen kann für den dann gegebenen Datensatz $\hat{\mathbf{x}} = (\hat{x}_1, \mathbf{K}, \hat{x}_{1000})^T$ der Stichproben-Mittelwert exakt berechnet werden.

Die Varianz des Stichprobenmittelwertes \bar{X} ist grösser als die Varianz der Zufallsvariablen X .

Die Varianz des Stichprobenmittelwertes, $Var[\bar{X}]$, verdoppelt sich, wenn statt $n=1000$ nur $n=500$ Versuche durchgeführt werden können.

1.21 An 10 Stahlträgern aus derselben Produktion wurden Zugversuche gemacht.

- c) Mit den Versuchsdaten wurde der Stichproben-Mittelwert \bar{x} und die Stichproben-Standardabweichung s_x der Zugfestigkeit von Stahl berechnet:

$$\bar{x} = \sum_{i=1}^{10} \hat{x}_i \qquad s_x = \sqrt{\frac{1}{10} \sum_{i=1}^{10} (\hat{x}_i - \bar{x})^2}$$

Welche der folgenden Aussagen ist / sind richtig?

Aufgrund der begrenzten Stichprobengrösse sind \bar{x} und s_x als Schätzer für den wahren Mittelwert m_x und die wahre Standardabweichung s_x mit Unsicherheit behaftet.

\bar{x} ist ein erwartungstreuer Schätzer für den wahren Mittelwert m_x .

s_x^2 dient als Schätzer für die Varianz des Stichproben-Mittelwertes, $Var[\bar{x}]$.

Der erwartungstreue Schätzer für die Stichprobenvarianz berechnet sich wie folgt: $s_x^2 = \frac{n}{n-1} s_x^2$

- d) Aus den Versuchsdaten wurde das folgende symmetrische 95%-Konfidenzintervall für den wahren Mittelwert m_x der Fliessspannung bestimmt (s_x ist aus Erfahrung bekannt):

$$I = [\bar{x} - \Delta \leq m_x \leq \bar{x} + \Delta] = [358MPa \leq m_x \leq 389MPa]$$

Welche der folgenden Aussagen ist / sind richtig?

Mit einer Konfidenz von 95% liegt der wahre Mittelwert m_x im Intervall I .

Die Breite des Konfidenzintervalls lässt sich für $\alpha = 0.05$ folgendermassen

berechnen: $\Delta = \frac{0.05 \cdot \sqrt{n}}{s_x}$.

Der Stichprobenmittelwert \bar{x} liegt mit einer Wahrscheinlichkeit von 0.05 ausserhalb des Intervalls I .

Das 95%-Konfidenzintervall wird kleiner, wenn man für seine Berechnung auf eine grössere Anzahl Versuche zurückgreifen kann.

- e) Mit den Versuchsdaten wurde ein Hypothesentest für den wahren Mittelwert m_x bei bekannter Varianz s_x durchgeführt. Die Nullhypothese $H_0 : m_x = 360MPa$ wurde auf einem Signifikanzniveau von $\alpha = 0.05$ abgelehnt.

Welche der folgenden Aussagen ist / sind richtig?

Die Wahrscheinlichkeit, dass die Hypothese fälschlicherweise abgelehnt wurde, beträgt 0.05.

$\alpha = 0.05$ ist die Wahrscheinlichkeit eines Fehlers 1. Art.

Die Stichprobengröße n hat keinen Einfluss auf die Wahrscheinlichkeit eines Fehlers 2. Art.

Die Alternativhypothese ist $H_1 : m_x < 360MPa$.

1.22 Welche der folgenden Aussagen trifft / treffen zu (in Bezug auf Hypothesentests)?

Die Wahl des Signifikanzniveaus beeinflusst das Ergebnis des Hypothesentests.

Die Formulierung der Nullhypothese beeinflusst die Wahrscheinlichkeit eines Fehlers 2. Art nicht.

Die operative Regel besagt, wann die Nullhypothese verworfen bzw. akzeptiert werden kann.

Bei einem Test für die Güte der Anpassung besagt die Nullhypothese, dass der Datensatz durch eine bestimmte Verteilungsannahme gut repräsentiert werden kann.

1.23 Welche der folgenden Aussagen trifft / treffen zu (in Bezug auf Tests für die Güte der Anpassung)?

Der Kolmogorov-Smirnov-Test ist nur für diskrete Verteilungsfunktionen anwendbar.

Möchte man den Chi-Quadrat-Test auf ein kontinuierliches Verteilungsmodell anwenden, so muss man zunächst den Stichprobenraum durch Bildung von Intervallen diskretisieren.

Die Stichprobenstatistik e_m^2 beim Chi-Quadrat-Test ist Chi-Quadrat-verteilt mit $k - 1$ Freiheitsgraden, wobei k die Anzahl der Parameter ist, die aus den Daten geschätzt wurden.

Beim Kolmogorov-Smirnov-Test wird die empirische (beobachtete) Verteilung der Daten mit der postulierten Verteilungsfunktion verglichen.

1.24 Sie haben eine Normalverteilung an einen Datensatz $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)^T$ mit Werten für eine Materialfestigkeit angepasst. Alternativ betrachten Sie eine Lognormalverteilung, deren Parameter Sie aus der Literatur entnehmen konnten. Angaben zur Modellevaluation sind in der folgenden Tabelle enthalten (Alle Zahlenwerte wurden mit dem Datensatz $\hat{\mathbf{x}}$ berechnet):

Modell	Parameter	Chi-Quadrat-Statistik e_m^2	Log-Likelihood $l(\boldsymbol{\theta} \hat{\mathbf{x}})$
Normalverteilung	m, s (geschätzt)	0.407	- 50.1
Lognormalverteilung	I, V (Literatur)	0.531	-27.7

Welche der folgenden Aussagen ist / sind richtig?

Ein direkter Vergleich der beiden Chi-Quadrat-Stichprobenstatistiken ist möglich, da beide Modelle zwei Parameter haben.

Die Lognormalverteilung eignet sich besser zur Beschreibung der Daten, da die Log-Likelihood für dieses Modell grösser ist.

Ein Vergleich der Log-Likelihoods ist sinnlos, da bei der Lognormalverteilung die Parameter nicht aus den Daten geschätzt wurden.

Unabhängig von den Werten in der Tabelle eignet sich die Lognormalverteilung besser zur Modellierung einer Materialfestigkeit, da eine lognormalverteilte Zufallsvariable keine negativen Werte annehmen kann.

Teil 2: Rechenaufgabe (maximal 30 Punkte)

Während einer Projektarbeit an der ETH Zürich wurden die Rohdichten von $n = 11$ Holzbrettern der Holzart Fichte ermittelt. Die Ergebnisse sind in Tabelle 2.1 zusammengefasst.

Tabelle 2.1: Experimentell ermittelte Werte der Rohdichte von Fichtenbrettern

Brettnummer i	Rohdichte \hat{r}_i [kg/m ³]
1	439.1
2	441.5
3	380.6
4	395.3
5	443.4
6	321.7
7	448.8
8	442.7
9	480.3
10	377.3
11	456.2

a) (5 Punkte)

Passen Sie an die Beobachtungen aus Tabelle 2.1 eine Normalverteilung an, indem Sie die entsprechenden Maximum-Likelihood-Schätzer der Verteilungsparameter m_r und s_r bestimmen.

b) (5 Punkte)

Bestimmen Sie mithilfe der gegebenen Verteilungsparameter $m_r = 420$ und $s_r = 40$ die Wahrscheinlichkeit, mit der Bretter der Holzart Fichte einen Wert der Rohdichte von mindestens 500 kg/m³ annehmen.

Im Rahmen von früheren experimentellen Untersuchungen wurde für die Rohdichte von Fichtenholz eine Normalverteilung mit einer Standardabweichung von $s_r = 45.0$ [kg/m³] (deterministisch) ermittelt.

Der Mittelwert dieses Modells ist unbekannt. Es wird jedoch angenommen, dass die *a priori* Verteilung des Mittelwertes normalverteilt ist mit $m_{m_r}' = 432$ [kg/m³] und $s_{m_r}' = 3.49$ [kg/m³].

c) (4 Punkte)

Berechnen Sie anhand der Standardabweichungen s_r und s_{m_r}' die Gewichtung n' des a priori Modells, die zur analytischen Erstellung eines a posteriori Modells für den Mittelwert verwendet wird.

d) (8 Punkte)

Berechnen Sie unter Verwendung der Daten aus Tab. 2.1 die Parameter der *a posteriori* Verteilung (m_{m_r}'' und s_{m_r}'') für den Mittelwert der Rohdichte.

e) (8 Punkte)

Berechnen Sie die *prädiktive* Dichtefunktion für die Rohdichte $f_p(r|\hat{\rho})$ von Fichtenholz unter Berücksichtigung der *a posteriori* Verteilungsparameter des Mittelwertes. Formulieren Sie hierfür zuerst die allgemeine Form der prädiktiven Dichtefunktion und berechnen Sie anschliessend deren Parameter $m_{r'''}$ und $s_{r'''}$.