KARLSRUHER INSTITUT FÜR TECHNOLOGIE

Institut für Volkswirtschaftslehre (ECON)

Lehrstuhl für Statistik, Ökonometrie und Mathematische Finanzwirtschaft

D-76128 Karlsruhe, Postfach 6980

Prof. Dr. Wolf-Dieter Heller

Florian Jacob

PC-Praktikum für die Statistikausbildung im Wintersemester 2013/2014

Lektion 1-3

(Statistik II)

Institut für Statistik und Mathematische Wirtschaftstheorie Prof. Heller

9. Exponentialverteilung, Faltung und Erlangverteilung

Bitte laden Sie für diese Übung zunächst **EXCEL** (per Doppelklick auf das EXCEL-Icon auf dem Desktop bzw. durch Wahl des entsprechenden Eintrags aus dem Startmenü). Öffnen Sie anschließend die Datei **Aufgabe 9.xls**, die Sie auf Laufwerk **k:** finden. Zum Öffnen einer Datei wählen Sie bitte in EXCEL nach Anklicken des **Microsoft-Office-Buttons** (oben links) den Eintrag **Öffnen**. Im Dialog **Öffnen** müssen Sie nun unter **Suchen in** über den Arbeitsplatz das Laufwerk **k:** auswählen. Anschließend sollten Sie im Dateifenster die oben genannte Datei sehen. Wählen Sie diese per Mausklick aus und bestätigen Sie abschließend per Mausklick auf den Schalter **Öffnen** Ihre Wahl. Die Datei sollte jetzt geöffnet werden. EXCEL wird Ihnen dabei mitteilen, dass die Datei Makros enthält und fragen, ob diese aktiviert werden sollen. Bestätigen Sie dies durch Anklicken von **"Diesen Inhalt aktivieren"**. Nach dem Laden wird EXCEL zunächst einige Einstellungen initialisieren.

In dieser Aufgabe werden Sie sich zunächst mit der Exponentialverteilung beschäftigen. Anschließend werden Sie über die Faltung zur Erlangverteilung gelangen.

Zur Aufgabe: Im Call-Center einer Fluggesellschaft gibt es drei Abteilungen, die für unterschiedliche Aufgaben zuständig sind:

1. Die Abteilung **Service** benötigt im Mittel 30 Minuten, um einen Kunden zu beraten.
2. Die zweite Abteilung führt Buchungen durch. Im Mittel dauert dies 12 Minuten.
3. Der Bereich **Fluginformationen** teilt Fluggästen An- und Abflugszeiten mit. Erwartungsgemäß dauert ein solches Gespräch 6 Minuten.

Auswertungen vergangener Gesprächszeiten führten zu dem Ergebnis, dass die Anrufzeiten für jede Abteilung mit den oben angegebenen Mittelwerten exponentialverteilt sind. Für die Abteilung **Service** gilt also z.B.:

$$E_{Service}(X) = 30 \text{ min} = 0{,}5 \text{ h}.$$

Erwartungswert und Varianz einer Exponentialverteilung lassen sich wie folgt berechnen:

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

Somit kann aus den oben genannten Mittelwerten der Parameter λ für jede der drei Exponentialverteilungen näherungsweise bestimmt werden.

Für die Abteilung Service ergibt sich z.B. $\lambda = 2$ ($\lambda = \frac{1}{E_{Service}(X)} = \frac{1}{0{,}5} = 2$).

(a) Vervollständigen Sie die beiden Tabellen **Dichtefunktionen** und **Verteilungsfunktionen** auf dem Blatt **Daten a**. Benutzen Sie dabei für die Berechnung der λ die Erwartungswerte in Stunden und nicht in Minuten.

Tipps für die Bearbeitung:

In die linke Spalte der beiden Tabellen müssen jeweils die x-Werte eingetragen werden. Für die Tabelle **Dichtefunktionen** *verwenden Sie bitte Werte im Intervall* $[0; 2{,}5]$*, für die Tabelle* **Verteilungsfunktionen** *Werte im Intervall* $[0; 5]$*, jeweils im Abstand von* 0,1 *(d.h.* 0*;* 0,1*;* 0,2*; usw.). In die Zeile rechts von der Zelle* $x \setminus \lambda$ *tragen Sie bitte die* λ*-Werte für die unterschiedlichen Verteilungen ein. Für die Werte der Dichte- bzw. der Verteilungsfunktion können Sie die EXCEL-Funktion* **EXPONVERT** *benutzen. Die Syntax dieser Funktion lautet folgendermaßen:*
EXPONVERT(X; Lambda; Kumuliert)

Kumuliert *ist eine Boolean-Variable, die angibt, ob man einen Wert der Dichte- oder der Verteilungsfunktion berechnen möchte. Um die Werte der Dichtefunktionen zu berechnen, muss* **Kumuliert** *also gleich null (=falsch) gesetzt werden, für die Berechnung der Werte der Verteilungsfunktion verwenden Sie bitte den Wert eins (=wahr). Bitte benutzen Sie für* **Lambda** *einen Verweis auf die entsprechende Zelle (um die Lösung der folgenden Aufgaben zu erleichtern). Achten Sie dabei auf die absolute Adressierung der Zellen.*

Weitere Hilfen:
↪ **Adressierung von Zellen**
↪ **Excel-Funktionen**
↪ **Zellen automatisch ausfüllen**
↪ **EXPONVERT**

(b) Um die Ergebnisse aus (a) leichter interpretieren zu können, sollen Dichte- und Verteilungsfunktion in dieser Teilaufgabe graphisch dargestellt werden. Auf dem Tabellenblatt **Graphen1** sind zwei graue Flächen vorbereitet, auf denen die Diagramme platziert werden sollen.

Tipps, um die Graphen zu erstellen:

Bitte erstellen Sie je ein Diagramm für die Dichte- und die Verteilungsfunktionen. Dazu markieren Sie am besten den Wertebereich der entsprechenden Tabelle und rufen dann den Diagramm-Manager auf. Sie finden den Diagrammmanager unter der Registerkarte ***Einfügen****. Wählen Sie als Diagrammtypen das zweidimensionale Liniendiagramm. Wenn Sie doppelt auf das Diagramm klicken, erscheinen die Diagrammtools in der Multifunktionsleiste. Klicken Sie nun auf "Daten auswählen". In Feld "Legendeneinträge" kann man die Zellen angeben, in dene die y-Werte stehen. Wenn sie die Zellen vorher markiert haben, sollten die richtigen Zellen dort schon angegeben sein. Im Feld "Horizontale Achsenbeschriftung" müssen Sie den Datenbereich der x-Werte angeben. Dazu klicken Sie auf "Bearbeiten" und können dann mit der Maus die entsprechenden Zellen im Tabellenblatt markieren. Wenn Sie in der Multifunktionsleiste ein anderes "Diagrammlayout" wählen, wird der Titel des Diagramms hinzugefügt. Als Diagrammtitel tragen Sie "Dichtefunktionen" oder Verteilungsfunktionen" ein. Die Rubrikenachse kann einfach "x-Werte" und die Größenachse genauso wie der Diagrammtitel genannt werden. Jetzt müssen Sie die Grafik nur noch an die richtige Stelle ziehen.*

(c) Um den Einfluss des Parameters λ auf die Dichte- und Verteilungsfunktion **Fluginfo** besser erkennen zu können, soll der λ-Wert veränderbar gestaltet werden. Auf dem Tabellenblatt **Graphen1** ist dafür schon ein Feld vorgesehen (rechts neben $\lambda =$). Bitte ändern Sie sowohl für die Verteilungs- als auch für die Dichtefunktion die Zelle, in der der λ-Wert steht, in einen Verweis auf die Zelle **I2** auf dem Tabellenblatt **Graphen1** um. Jetzt können Sie auf dem Tabellenblatt **Graphen1** diesen λ-Wert ändern und gleichzeitig beobachten, wie sich die roten Kurven verändern.

Hilfen:
↪ **Zellen in anderen Tabellenblättern markieren**

(d) Im weiteren sollen Sie die Wahrscheinlichkeit für eine Beobachtung in einem bestimmten Intervall $I = (a,b)$ berechnen. Dies kann allgemein mit der Formel

$$P(a < X < b) = F(b) - F(a)$$

geschehen. In EXCEL gibt es mehrere Möglichkeiten, diese Wahrscheinlichkeit zu berechnen.

(1) Sie können die Werte F(a) und F(b) wie schon in der ersten Übung durch die Funktion **EXPONVERT** bestimmen. Der Parameter **kumuliert** muss dabei auf **1** gesetzt werden.

(2) Die zweite Möglichkeit besteht darin, die Werte aus der Tabelle der Verteilungsfunktionen zu übernehmen.

(3) Speziell für die Exponentialverteilung gilt:

$$F(a < X < b) = (1 - e^{-\lambda b}) - (1 - e^{-\lambda a}) = e^{-\lambda a} - e^{-\lambda b}$$

Diese Formel kann natürlich auch benutzt werden. Die e-Funktion **EXP** finden Sie unter der Registerkarte "Formeln" durch klicken auf die Schaltfläche "Funktion einfügen" (Kategorie: **Math.&Trigonom.**).

Überprüfen Sie für $P(2 < X < 3)$ mit $\lambda = 2$, ob die drei verschiedenen Wege zum gleichen Ergebnis führen.

(e) Die Fluggesellschaft interessiert sich dafür, nach welcher Zeit ein Telefonat mit einer Wahrscheinlichkeit von α% beendet ist. Diese Information gibt das α-Quantil q_α :

$$P(X \leq q_\alpha) = \alpha$$

Speziell für die Exponentialverteilung gilt:

$$1 - e^{-\lambda q_\alpha} = \alpha$$

Bestimmen Sie zunächst die Formel, mit der die Quantile der Exponentialverteilung berechnet werden können und berechnen Sie dann die Quantile in der auf Blatt **Aufgabe d+e** angegebenen Tabelle. Vergleichen Sie Ihre Ergebnisse mit den Werten der Verteilungsfunktion aus Aufgabe (a). Was sagen die unterschiedlichen 0,75-Quantile aus?

Formel für q_α:

Hilfen:
↪ **Berechnungshilfe LN**

Für die Organisation des Call-Centers ist es von Interesse zu wissen, wie die Summe der Zeiten mehrerer Telefonate verteilt ist.

Grundsätzlich gilt, dass die Summe mehrerer unabhängiger, mit gleichem Parameter λ exponentialverteilter Zufallszahlen Erlang-verteilt ist mit den Parametern λ und n:

$$\sum_{i=1}^{n} X_i \sim Erl(\lambda,n) \qquad \text{mit } X_i \sim Exp(\lambda)$$

Die Anzahl der Freiheitsgrade n der Erlangverteilung entspricht dabei der Anzahl der exponentialverteilten Zufallszahlen. Eine Erlangverteilung mit n Freiheitsgraden ist also die Faltung von n Exponentialverteilungen mit gleichem Parameter λ.

Um das Prinzip der Faltung graphisch veranschaulichen zu können, sollen Sie zunächst eine Erlangverteilung mit zwei Freiheitsgraden betrachten, also das Ergebnis der Faltung von zwei gleichen Exponentialverteilungen. Ihre Dichte kann folgendermaßen berechnet werden:

$$\begin{aligned}
f(z) &= \int\limits_{x+y=z} f(x)f(y)dx \\
&= \int\limits_{\substack{x\geq 0 \\ z-x\geq 0}} f(x)f(z-x)dx \\
&= \int\limits_{\substack{x\geq 0 \\ z-x\geq 0}} \lambda e^{-\lambda x}\lambda e^{-\lambda(z-x)}dx \\
&= \lambda^2 e^{-\lambda z} \int\limits_{\substack{x\geq 0 \\ z-x\geq 0}} 1dx \\
&= \lambda^2 e^{-\lambda z}[x]_0^z \\
&= \lambda^2 z e^{-\lambda z}
\end{aligned}$$

EXCEL unterstützt leider keine numerische Berechnung von Integralen, deshalb müssen die beiden Exponentialverteilungen zunächst diskretisiert werden, um anschließend das Integral durch eine Summe approximieren zu können.

Die Aufgaben (f) und (g) dienen der Vorbereitung für die eigentliche Faltung in Aufgabe (h). In Aufgabe (f) wird die gemeinsame Verteilung zweier diskretisierter Exponentialverteilungen berechnet, indem die unabhängigen Verteilungsfunktionen intervallweise miteinander multipliziert werden.

(f) Tragen Sie in die Tabelle auf Blatt **Aufgabe f+g** die Wahrscheinlichkeiten dafür ein, dass exponentialverteilte Zufallszahlen (mit Parameter $\lambda = 10$) in den angegebenen Intervallen der gemeinsamen Verteilung liegen. In den grauen Feldern stehen als Hilfe die Werte der Verteilungsfunktion für die darüber bzw. daneben stehenden x-Werte. Im weißen Bereich sind links und oberhalb der "bunten Tabelle" Intervallgrenzen $I = (a,b]$ angegeben. In der weißen Zeile bzw. Spalte unterhalb bzw. rechts von der "bunten Matrix" tragen Sie bitte mit Hilfe der Summenfunktion die Randwahrscheinlichkeiten für x_1 bzw. x_2 ein. Um die Farben der Matrix nicht zu verändern, benutzen Sie beim Kopieren der Werte bitte die Funktion "Inhalte einfügen –> Formeln" bzw. schauen Sie sich die Hilfe "Formeln ausfüllen ohne das Format zu ändern" an.

Hilfen:
↪ **Diskretisierung**
↪ **Formeln ausfüllen ohne das Format zu ändern**

(g) Stellen Sie die Daten des "bunten Bereichs" bitte graphisch dar. Ändern Sie den λ-Wert von 10 in 2 um, damit die Graphik besser zu erkennen ist. Schauen Sie sich das Diagramm von verschiedenen Seiten an.

Hilfen:
↪ **Tipps zur Diagrammerstellung**
↪ **Diagramm drehen**, auf Blatt **Graphen2**

(h) In dieser Aufgabe soll nun die eigentliche Faltung durchgeführt werden. Dabei müssen jeweils die Wahrscheinlichkeiten derjenigen $(x_1; x_2)$-Kombinationen addiert werden, die das gleiche Ergebnis für $(x_1 + x_2)$ liefern. In der Tabelle aus Aufgabenteil (f) sind diese Kombinationen in den gleichen Farben markiert. Um eine in EXCEL komplizierte Diagonal-Addition zu vermeiden, wurden für Sie zwei Makros programmiert, die die Zellen vertikal versetzen bzw. diesen Vorgang rückgängig machen:

- **Tabelle verschieben,**
- **Verschieben rückgängig**

Um die Darstellung der Funktion in der Tabelle zu erleichtern, wird im weiteren auf eine Angabe der Intervallgrenzen verzichtet und stattdessen die Intervallmitte als Repräsentant gewählt. Verwenden Sie $\lambda = 10$ und tragen Sie mit Hilfe der Summenfunktion die Wahrscheinlichkeiten für $P(x_1 + x_2)$ ein.

(i) Können Sie anhand der eben berechneten Wahrscheinlichkeit für den Bereich $I = (0; 2]$ eine Aussage über die Genauigkeit der Häufigkeitsverteilung von $(x_1 + x_2)$ machen?

-
-

Vergleichen Sie die Ergebnisse, die Sie durch die Diskretisierung der Exponentialverteilungen erhalten haben, mit den theoretischen Werten für die Erlangverteilung mit zwei Freiheitsgraden. Die Dichte- und die Verteilungsfunktion der Erlangverteilung finden Sie im Funktionenmanager unter **Benutzerdefiniert**. Benutzen Sie für λ einen Verweis auf Zelle **N17** und ändern Sie λ in der Tabelle auf Blatt **Aufgabe h** ebenfalls in einen Verweis auf Zelle **N17** um.

Sind die Werte $P_{diskret}(x_1 + x_2)$ gute Näherungen für $P_{Erlang}(I)$? Vergleichen Sie $P_{diskret}(x_1 + x_2)$ und $P_{Erlang}(I)$ für unterschiedliche λ-Werte, indem Sie Zelle **N17** ändern.

-
-

Hilfen:
↪ **Berechnungshilfe P(Intervall)**

(j)* In der letzten Teilaufgabe wurde die Verteilung der Gesamtlänge zweier Telefonate betrachtet. Für die Fluggesellschaft ist allerdings interessanter, wie die Summe einer größeren Anzahl Telefonate verteilt ist.

Untersuchen Sie bitte, welchen Einfluss die Anzahl der Telefonate auf die Verteilung ihrer Summe hat. Verwenden Sie die benutzerdefinierten Funktionen **ErlangVerteilung** und **ErlangDichte**. Nehmen Sie einen λ - Wert von 30 an. Benutzen Sie bitte Verweise auf λ (Achtung auf Tabellenblatt **Graphen4**!) und die Freiheitsgrade und achten Sie auf die richtige Adressierung.

Erstellen Sie nun wie in Aufgabe (b) je eine Graphik für die Dichte- bzw. die Verteilungsfunktionen und fügen Sie sie in **Graphen4** ein. Wie ändert sich die Dichte und die Verteilungsfunktion bei steigender Anzahl von Freiheitsgraden? Wie ändert sich die Verteilung mit zunehmendem λ?

-
-

Hilfestellungen zu Übung 9

EXCEL-Funktionen (Aufgabe (a))

In EXCEL kann man mit Hilfe von Funktionen Berechnungen in den Zellen durchführen. Die Eingabe dazu muss mit einem Gleichheitszeichen begonnen werden. Möchte man zum Beispiel die Berechnung **2+3** durchführen, muss man **=2+3** in die Zelle schreiben. Nachdem die Eingabe mit **Return** abgeschlossen ist, erscheint in der Zelle das Ergebnis "5". Klickt man erneut auf die Zelle kann man in der Eingabezeile die Formel sehen.

Zusätzlich verfügt EXCEL über einen großen Katalog vordefinierter Funktionen. Um diese Funktionen nutzen zu können, muss man die Zelle, in der das Ergebnis der Funktion stehen soll, markieren. Dann klickt man unter der Registerkarte "Formeln" auf das Funktionen-Zeichen (f_x). Jetzt erscheint ein neues Fenster (der Funktionenmanager). Im Bereich **Kategorie** kann die Art der Kategorie gewählt werden. Im Bereich rechts daneben werden die Funktionen dieser Kategorie angezeigt. Klickt man die Kategorie **alle** an, werden alle EXCEL zur Verfügung stehenden Funktionen aufgelistet. Das kann relativ unübersichtlich sein, deshalb sind die Kategorien zusätzlich in kleinere Untergruppen unterteilt.

Unter dem Kategorie-Fenster finden Sie jeweils eine Kurzbeschreibung zu der rechts markierten Funktion. Mit einem Doppelklick kann man die gewünschte Funktion aktivieren.

EXPONVERT (Aufgabe (a))

Die Funktion EXPONVERT finden Sie im Funktionenmanager unter der Kategorie **Statistik**. Mit ihr kann entweder der Wert der Dichte- bzw. Verteilungsfunktion einer Exponentialverteilung berechnet werden. Zu gegebenem λ berechnet **EXPONVERT** also $f(x)$ oder $F(x)$. Dazu müssen drei Werte als Parameter angegeben werden:

- **X**: Der x-Wert, für den die Dichte- oder die Verteilungsfunktion berechnet werden soll. Für diese Aufgabe bietet es sich an, keine Zahlenwerte für x zu verwenden, sondern die relativen Adressen anzugeben. Wenn der Cursor in dem Textfeld **X** blinkt, kann man die Zelle mit dem x-Wert ganz einfach anklicken, dann übernimmt EXCEL die entsprechende Adresse.
- **Lambda**: ist der λ-Wert der Exponentialverteilung. Hier empfiehlt es sich, absolute Adressen anzugeben. Die Zelle kann am besten wie bei **X** per Mausklick markiert werden. EXCEL übernimmt Adressen allerdings immer mit relativen Bezug. D.h. die Adresse muss noch nachgebessert werden, indem die **$**-Zeichen für den absoluten Bezug eingefügt werden.
- **Kumuliert**:**Kumuliert** ist eine Boolean-Variable, die angibt, ob man einen Wert der Dichte- oder die Verteilungsfunktion berechnen möchte.

  Um die Werte der Dichtefunktionen zu berechnen, muss **Kumuliert** also gleich null (=falsch) gesetzt werden, für der Berechnung der Werte der Verteilungsfunktion verwenden Sie bitte den Wert eins (=wahr). Bitte benutzen Sie für **Lambda** einen Verweis auf die entsprechende Zelle (um die Lösung der folgenden Aufgaben zu erleichtern). Achten Sie dabei auf die absolute Adressierung der Zellen.

Beispiel: Der Wert der Dichtefunktion an der Stelle $x = 0$,$f_{Service}(0)$, soll in Zelle **C60** stehen (d.h. diese Zelle muss markiert sein, bevor man den Funktionenmanager aufruft). Nachdem die korrekte Eingabe im Funktionenmanager mit **OK** abgeschlossen ist, schreibt EXCEL =**EXPONVERT(B60,C59,0)** in die Zelle (oben in der Befehlszeile zu sehen) und auf dem Bildschirm erscheint das Ergebnis **10** ($= f_{Service}(0)$).

Zellen in anderen Tabellenblättern markieren (Aufgabe (c))

Für einen Verweis geben Sie in der entsprechenden Zelle der Tabelle ein = ein. Dann müssen Sie auf das Tabellenblatt **Graphen1** wechseln und dort die Zelle **I2** markieren. Wenn Sie nun die **Return**-Taste drücken, übernimmt EXCEL den Verweis auf das andere Tabellenblatt.

Alternativ können Sie auch **Graphen1!I2** eingeben. Das Ausrufezeichen symbolisiert, dass sich die Zelle **I2** auf dem davor genannten Tabellenblatt befindet.

Berechnungshilfe ln ((Aufgabe (e))

Der natürliche Logarithmus wird in EXCEL mit **LN(x)** berechnet. Sie finden sie im Funktionenmanager in der Kategorie **Math.&Trigonom**.

Diskretisierung (Aufgabe (f))

Die Diskretisierung der Exponentialverteilung ist relativ einfach zu erreichen: Betrachtet werden die Wahrscheinlichkeiten in unterschiedlichen Bereichen (in der Tabelle angegeben). Da die Verteilungsfunktion bekannt ist, muss man lediglich deren Wert an der Untergrenze des Bereichs von deren Wert an der Obergrenze abziehen.

Beispiel:

$$P(0{,}3 < x < 0{,}4) = F(0{,}4) - F(0{,}3)$$

Für die Faltung ist entscheidend, dass die Wahrscheinlichkeiten der gemeinsamen Verteilung der beiden Exponentialverteilungen betrachtet werden. Diese erhält man durch Multiplikation der beiden Einzel-Wahrscheinlichkeiten. Beachten Sie, dass dies nur möglich ist, da die Exponentialverteilungen unabhängig sind.

Beispiel:

$$P(0{,}3 < x_1 < 0{,}4; 0{,}6 < x_2 < 0{,}7) = [F1(0{,}4) - F(0{,}3)] * [F(0{,}7) - F(0{,}6)]$$

Formeln ausfüllen ohne die Formatierung zu ändern (Aufgabe (f))

Die Formeln für die Wahrscheinlichkeit können Sie wieder automatisch ausfüllen lassen. Standardmäßig übernimmt EXCEL dabei nicht nur die Formeln, sondern auch die Formate. Für Aufgabe (h) sollen die Farben aber so bestehen bleiben, damit der Effekt der Faltung besser zum Ausdruck kommt. Um einen Bereich automatisch auszufüllen, können Sie prinzipiell wie in Aufgabe (a) verfahren. Markieren Sie aber die Zellen mit Hilfe der linken statt der rechten Maustaste. Es erscheint ein Fenster, in dem Sie die Funktion **Inhalte einfügen** –> **Formeln** wählen müssen. Jetzt werden nur die Formeln kopiert, nicht aber die Formatierung.

Tipps zur Diagrammerstellung der diskretisierten Dichte zweier Exponentialverteilungen (Aufgabe (g))

- Markieren Sie die y-Werte der bivariaten Verteilung und die x_2-Werte (Zellen **F60:O70**).
- Rufen Sie den Diagrammassistenten auf und wählen Sie den benutzerdefinierten Diagrammtyp **"3-D-Säule"**.
- Geben Sie als x-Werte die x_1-Werte an.
- Wählen Sie einen Diagrammtitel (z.B. "Massenverteilung zweidimensionale diskretisierte Exponentialverteilung")
- Fügen Sie die Graphik als Objekt auf dem Tabellenblatt **Graphen2** ein, und schieben Sie es auf den dafür vorgesehenen Platz.

Diagramm drehen (Aufgabe (h))

Sie können ein Diagramm drehen, indem Sie mit dem Mauszeiger auf eine der Diagrammecken zeigen und kurz die linke Maustaste drücken. Jetzt erscheint an jeder der Diagrammecken ein schwarzes Quadrat. Wenn Sie wieder mit dem Mauszeiger auf eine der Ecken zeigen, erscheint ein kleines Kreuz. Halten Sie die linke Maustaste gedrückt und drehen Sie das Diagramm in die gewünschte Richtung.

Berechnungshilfe $P_{Erlang}(I)$ (Aufgabe (i))

Für die Berechnung der Wahrscheinlichkeiten der Erlang-Verteilung müssen wieder die vollen Intervalle um die Intervallmitten benutzt werden. Für den ersten und letzten Wert wird dieses Intervall etwas vergrößert, um den gesamten Definitionsbereich abzudecken. D.h. um die Wahrscheinlichkeit für einen Bereich anzugeben, muss wieder die Verteilungsfunktion benutzt werden. Für **X** wird ein Verweis auf die Bereichsober- bzw. Untergrenze verwendet. Die Zahl der Freiheitsgrade n kann in dieser Aufgabe als absoluter Zahlenwert eingetragen werden und bei **Lambda** muss ein Verweis auf die Zelle **N11** benutzt werden (Achtung: absolute Adressierung).

Institut für Statistik und Mathematische Wirtschaftstheorie Prof. Heller

10. Normalverteilung, χ^2-Verteilung, zentraler Grenzwertsatz

Diese Lektion beschäftigt sich zunächst mit der Normalverteilung. Sie werden dabei sowohl die ein- als auch die zweidimensionale Normalverteilung untersuchen. Anschließend werden Sie mit der χ^2-Verteilung arbeiten und sich an diesem Beispiel die Bedeutung des zentralen Grenzwertsatzes verdeutlichen. Abschließend werden Sie anhand der Poisson-Verteilung untersuchen, ob der zentrale Grenzwertsatz auch für diskrete Verteilungen anwendbar ist.

Laden Sie bitte die Datei **Aufgabe 10.xls**.

(a) In dieser Aufgabe sollen Sie sich mit der Normalverteilung vertraut machen. Zunächst sollen Sie dazu normalverteilte Zufallszahlen generieren, diese dann in Klassen einteilen und sie schließlich als Histogramm darstellen. Die Dichte der Normalverteilung hat folgende Gestalt:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$$

wobei μ der Erwartungswert und σ die Standardabweichung bezeichnet.

Auf dem Tabellenblatt **Daten a** sollen die Zufallszahlen generiert und bearbeitet, anschließend ein Histogramm auf dem Blatt **Graphen 1** erzeugt werden.

Füllen Sie zunächst die zwei fehlenden Spalten in **Tabelle 1** auf dem Tabellenblatt **Daten a** aus. In EXCEL gibt es sogenannte **Add-Ins**. Dies sind Zusatzfunktionen, die man nur bei Bedarf installiert, da sie in der Regel relativ viel Speicherplatz benötigen und nicht von jedem Benutzer gebraucht werden. Für diese und die folgenden Aufgaben benötigen Sie das Add-In **Analysefunktionen**. In ihm sind mehrere Datenanalyse-Funktionen enthalten. Unter anderem finden Sie hier eine Funktion **Zufallszahlengenerierung**, mit der Sie die Daten für **Tabelle 1** erzeugen können (falls Sie dieses Add-In nicht unter der Registerkarte "Daten" in der Gruppe "Datenanalyse" **Extras** finden, müssen Sie es mit Hilfe des **Add-In-Manager**s einfügen).

Klassieren Sie nun die Daten aus den ersten beiden Spalten aus **Tabelle 1** mit Hilfe der Funktion **Histogramm** aus dem Add-In **Analysefunktionen**. Berechnen Sie anschließend die relativen Klassenhäufigkeiten und erzeugen Sie daraus ein Histogramm. Nähere Angaben zur Vorgehensweise finden Sie in den Hilfefunktionen auf Tabellenblatt **Daten a** bzw. im Anhang zum Skript.

Wie wirken sich Änderungen der Werte von μ und σ auf die Gestalt der Histogramme aus? Stellen die Histogramme eine gute Näherung für die entsprechenden Normalverteilungen dar? Wie könnte man die Histogramm-Form der Gestalt der theoretischen Verteilung noch näher bringen? Wie kann man eine N(μ,σ^2)-Verteilung in eine Standardnormalverteilung überführen?

-
-
-

Hilfen:
↪ **Analysefunktion Zufallszahlengenerieren**, Blatt **Daten a**

↪ **Analysefunktion Histogramm**, Blatt **Daten a**
↪ **Zeile / Spalte löschen**, Blatt **Daten a**
↪ **Relative Häufigkeiten**, Blatt **Daten a**
↪ **Histogramm erzeugen (1)**, Blatt **Daten a**

(b) Als nächstes sollen Sie $N(\mu,\sigma^2)$-verteilte Zufallzahlen in standard-normalverteilte Zufallszahlen überführen. Mit Hilfe der Transformationsfunktion

$$f(x) = \sigma x + \mu$$

können standard-normalverteilte Zufallszahlen in $N(\mu,\sigma^2)$-verteilte übergeführt werden (wie lautet die Transformationsformel, mit der Sie normalverteilte in standard-normalverteilte Zufallszahlen transformieren können?).

Auf dem Tabellenblatt **Daten b** finden Sie die Tabelle **Transformation**. Erzeugen Sie zunächst in der linken Spalte 2000 $N(2,3^2)$-verteilte Zufallszahlen. Standardisieren Sie diese danach in der Spalte **standardisiert**.

Zum Vergleich erzeugen Sie bitte 2000 standard-normalverteilte Zufallszahlen. Klassieren Sie die Daten und berechnen Sie die relativen Häufigkeiten der Klassen wie in Aufgabe (a). Benutzen Sie dabei die angegebenen Intervalluntergrenzen. Erzeugen Sie nun für die klassierten Daten Diagramme. Hierfür wurde das Tabellenblatt **Graphen 2** für Sie vorbereitet. Hilfestellungen finden Sie auf dem Tabellenblatt **Daten b** bzw. im Anhang des Skriptes.

Was fällt Ihnen auf? Warum decken sich die Histogramme **standardisiert** und **mü=0**, **sigma=1** nicht genau? Wie kann man eine bessere Deckung erreichen?

-
-

Hilfen:
↪ **Histogramm erzeugen (2)**, Blatt **Daten b**

(c) In dieser Aufgabe geht es darum, die Einflussfaktoren einer bivariaten Normalverteilung zu untersuchen. Dabei sollen sowohl Zufallszahlen als auch die theoretische Verteilung betrachtet werden.

1. Zum Vergleich sollen Sie sich jetzt noch die theoretische Verteilung ansehen. Die Dichte der bivariaten Normalverteilung sieht folgendermaßen aus:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)}{\sigma_x}\frac{(y-\mu_y)}{\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]}, \qquad x,y \in \mathbb{R}$$

Zunächst sollen Sie die Tabelle im Tabellenblatt **Daten c1** mit den entsprechenden Dichtewerten ausfüllen. Hierfür steht Ihnen die benutzerdefinierte Funktion **ZweiDimNormVert** zur Verfügung. Das Ergebnis können Sie auf dem Tabellenblatt **Graphen3** anschauen.

Betrachten Sie die Graphik wieder aus unterschiedlichen Perspektiven. Achten Sie besonders auf die Ansicht von oben. Welche Bedeutung haben die Ellipsen?

Wie verändert sich die Dichte, wenn man die verschiedenen Parameter verändert? Entspricht dies Ihren Beobachtungen aus Aufgabe a?

-
-
-

2. Um bivariat normalverteilte Zufallszahlen zu generieren, erzeugt man am besten standard-normalverteilte ZZ und transformiert diese anschließend. Erzeugen Sie bitte zunächst zweimal 5000 standardnormalverteilte Zufallszahlen auf dem Tabellenblatt **Daten c2** (Spalten U und V). In den Spalten X und Y sollen die eigentlichen Zufallszahlen stehen, die $N(\mu_x,\sigma_x^2)$- bzw. $N(\mu_y,\sigma_y^2)$-verteilt sein sollen. Der Korrelationskoeffizient sei r.

   Dazu müssen die $N(0,1)$-verteilten Zufallszahlen in den Spalten U und V wie folgt transformiert werden[1]:

$$X = \mu_x + \sigma_x * U$$

$$Y = \mu_y + \sigma_y * \sqrt{1-r^2} * V + r * \sigma_y * U$$

   Füllen Sie die Spalten X und Y auf dem Tabellenblatt **Daten c2** entsprechend aus. Benutzen Sie für U und V relative Adressen, für die übrigen Werte absolute Verweise.

   Je zwei nebeneinanderstehende Zufallszahlen in den Spalten X und Y stellen nun ein bivariat normalverteiltes Zufallszahlenpaar (X,Y) dar. Wenn Sie auf das Feld **Klassierung durchführen** klicken, werden die Zahlenpaare klassiert und die Ergebnisse in die rechts daneben stehende Tabelle eingetragen.

   Die Tabelle **Relative Häufigkeiten der klassierten ZZ** soll jetzt in zwei Diagrammen graphisch dargestellt werden. Markieren Sie zunächst die relativen Häufigkeiten inklusive der Kopfzeile mit den Klassenbereichen der y-Werte und rufen Sie den Diagramm-Manager auf. Wählen Sie nacheinander die Formate **"3-D-Säulen"** und **3-D-Oberfläche**. Geben Sie als **Beschriftung der Rubrikenachse** die Klassen der x-Werte auf dem Register **Reihen** an (Zellen **M6:N55**).

   Fügen Sie beide Diagramme im Tabellenblatt **Graphen4** ein, schieben Sie sie an die dafür vorgesehenen Stelle und bringen Sie sie auf die richtige Größe. Wenn Sie auf die Diagrammecken klicken, können Sie das Diagramm beliebig drehen. Schauen Sie sich die Verteilung aus unterschiedlichen Perspektiven an.

   Wo liegen die Vor- und Nachteile der beiden Diagrammarten? Welche ist theoretisch "korrekter"?

   -
   -
   -

[1] Die Formel können Sie sich recht einfach selbst herleiten: Gehen Sie dazu von der Dichte der bivariaten Normalverteilung aus (siehe unten). Spalten Sie nun zunächst die Dichtefunktion der Zufallsvariablen $X \sim N(\mu_X;\sigma_X^2)$ ab und formen Sie anschließend den zweiten Faktor, der die Dichte von Y darstellen muß, um. Nutzen Sie dabei Ihre Kenntnis des Erwartungswertes und der Varianz von Y.

Verändern Sie nacheinander μ_x,σ_x^2,μ_y,σ_y^2 und r. Damit die Veränderung auch in den Graphen umgesetzt wird, müssen Sie die Daten neu klassieren (also einfach noch mal auf **Klassierung durchführen** klicken). Wie wirkt sich eine Änderung der Werte der einzelnen Parameter auf die Gestalt der Graphen aus? Warum brauchen Sie die Zufallszahlen nicht jedes mal neu generieren?

-
-
-

(d) In dieser Aufgabe sollen Sie χ^2-verteilte Zufallszahlen generieren und graphisch darstellen. χ^2-verteilte Zufallszahlen ergeben sich als Summe mehrerer quadrierter standard-normalverteilter Zufallszahlen. Die Anzahl der Freiheitsgrade, n, der χ^2-Verteilung entspricht dabei der Anzahl der in die Summe eingehenden standard-normalverteilten Zufallszahlen:

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2 \,,$$

wobei n die Anzahl der Freiheitsgrade bezeichnet und $X_i \sim N(0;1)$.

Erzeugen Sie bitte zunächst 25-mal 1000 standard-normalverteilte Zufallszahlen in Tabellen **Normalverteile ZZ** (je nach Rechner kann die Berechnung ein bisschen dauern!). Quadrieren Sie die Zufallszahlen in der Tabelle **Quadrierte Normalverteilte ZZ**. Bilden Sie in der Tabelle **Summe der quadrierten ZZ** die Summe der ersten 5, der ersten 10 und dann aller 25 Vektoren (benutzen Sie dabei die Summenfunktion). In den Spalten dieser Tabelle stehen jetzt χ^2-verteilte Zufallszahlen mit 5, 10 und 25 Freiheitsgraden.

Um die Zufallszahlen in einem Histogramm darzustellen, müssen sie zuvor klassiert werden. Benutzen Sie den Vektor **Intervalluntergrenzen** als Intervallgrenzen in der Analysefunktion **Histogramm**. Wählen Sie die Zellen **BH4**, **BJ4** und **BL4** als Ausgabezelle. Erstellen Sie zu den drei Verteilungen ein Diagramm, und fügen Sie es auf dem Tabellenblatt **Graphen5** ein. Wählen Sie den Diagrammtyp **"Gestapelte Säulen"** (vergessen Sie nicht, vorher die beiden überflüssigen Klassenspalten zu löschen - vgl. Aufgabenteil (a)!).

Wie ändert sich die Verteilung mit wachsender Anzahl von Freiheitsgraden? Wie lautet der theoretische Wert für den Mittelwert und die Varianz der Verteilungen[2] ?

-
-

[2]Den Stichprobenmittelwert und die Stichprobenvarianz können Sie mit den EXCEL-Funktionen **MITTELWERT** und **VARIANZ** berechnen (welcher Unterrschied besteht zwischen den Funktionen **VARIANZ** und **VARIANZEN**?).

Als nächstes sollen die drei Vektoren mit den χ^2-verteilten ZZ standardisiert werden. Welche Formel können Sie dazu verwenden? Füllen Sie die Tabellen auf dem Tabellenblatt **Daten d** entsprechend aus. Erstellen Sie aus der letzten Tabelle eine Graphik, die neben den drei standardisierten χ^2-verteilten ZZ-Vektoren auch die Dichte der Standardnormalverteilung darstellt. Wählen Sie den benutzerdefinierten Diagrammtyp **Häufigkeitsverteilung** (dabei müssen Sie die x-Werte mit markieren!).

Um was für eine Diagrammart handelt es sich dabei? Was fällt Ihnen auf? Was kann man über χ^2-Verteilungen mit vielen Freiheitsgraden sagen? Verdeutlichen Sie sich anhand der Graphik die Bedeutung des folgenden, Ihnen aus der Vorlesung bekannten Satzes:

-
-
-

Zentraler Grenzwertsatz: (Lindeberg und Lévy)

Sei $X_1, \ldots, X_n$ eine Folge unabhängiger Zufallsvariablen, die alle derselben Verteilung genügen und eine endliche, von Null verschiedene Varianz besitzen.

Dann konvergiert die Folge $U_1, \ldots, U_n$ der standardisierten Variablensummen

$$U_n = \frac{\sum_{i=1}^{n} X_i - E(\sum_{i=1}^{n} X_i)}{\sqrt{Var(\sum_{i=1}^{n} X_i)}} = \frac{\sum_{i=1}^{n} X_i - n\,E(X_i)}{\sqrt{n}\sqrt{Var(X_i)}}$$

der Verteilung gegen eine **normalverteilte Zufallsvariable** *mit Erwartungswert 0 und Varianz 1. Durch Erweitern von Zähler und Nenner in U_n mit $\frac{1}{n}$ lässt sich die Aussage des Zentralen Grenzwertsatzes leicht auf Mittelwerte anwenden.*

Hilfen:
↪ **Chi-Quadrat standardisieren**, Blatt **Daten d**
↪ **Relative Häufigkeitsdichte**, Blatt **Daten d**

(e) Zum Abschluss dieses Abschnitts soll die Summe von Zufallzahlen einer diskreten Verteilung betrachtet werden. Von unabhängigen poissonverteilten Zufallszahlen wissen Sie, dass die Summe wieder poissonverteilt ist:

$$\sum_{i=1}^{n} X_i \sim Pois(\sum_{i=1}^{n} \lambda_i) \quad (*)$$

mit

$$X_i \sim Pois(\lambda_i)$$

Erzeugen Sie auf dem Tabellenblatt **Daten e** je 1000 poissonverteilte Zufallszahlen mit dem Parameter λ wie in der Tabelle **Poissonverteilte ZZ** angegeben. In der Tabelle **Summe der ZZ** müssen Sie einzelne Spalten der Tabelle **Poissonverteilte ZZ** so addieren, dass sich der in der Kopfzeile angegebene λ-Wert ergibt. Klassieren Sie die Werte mit der Analysefunktion **Histogramm** unter Verwendung der angegebenen Klassenuntergrenzen und fügen Sie ein entsprechendes Diagramm auf dem Tabellenblatt **Graphen6** ein (Diagrammtyp **"Gestapelte Säulen"**).

Wie verändert sich die Verteilung mit wachsendem λ? Welche Auswirkungen hat demnach die zunehmende Anzahl von Zufallszahlen auf die Verteilung ihrer Summe?

-
-
-

In der Tabelle **ZZ $\sim$ Pois(5)** sollen Sie nun die Gültigkeit von $(*)$ überprüfen. Hierfür addieren Sie bitte in der ersten Spalte von **ZZ $\sim$ Pois(5)** die ersten fünf Spalten $\lambda = 1$ der Tabelle **Poissonverteilte ZZ**, in der zweiten Spalte die Spalten $\lambda = 2$ und $\lambda = 3$ und übernehmen Sie in der dritten Spalte die Spalte $\lambda = 5$. Klassieren Sie anschließend die Daten und stellen Sie sie als Histogramm in **Graphen6** dar.

Handelt es sich um die gleichen Verteilungen? Falls es Abweichungen gibt, was könnte der Grund dafür sein?

Vergleichen Sie die Histogramme mit der Ihnen nun vertrauten Dichte der Standardnormalverteilung. Was stellen Sie mit zunehmender Anzahl von Summanden (bzw. zunehmendem λ) fest? Versuchen Sie, analog zu Aufgabe (d) den Graphen der Normalverteilung mit in das Diagramm zu zeichnen (welchen Erwartungswert und welche Varianz muss die Normalverteilung hier haben?).

-
-
-

Hilfestellungen zu Übung 10

Analysefunktion Zufallszahlengenerieren (Aufgabe (a), Blatt **Daten a**)

Rufen Sie die Funktion **Zufallszahlen** aus dem Add-In **Analysefunktionen** (Registerkarte "Daten", Gruppe "Datenanalyse") auf.

Anzahl der Variablen beschreibt wie viel Spalten mit Zufallszahlen belegt werden sollen. Da in **Tabelle 1** in jeder Spalte Zufallszahlen mit unterschiedlichem Mittelwert und verschiedenen Standardabweichungen eingetragen werden sollen, muss hier **1** angegeben werden.

Unter **Anzahl der Zufallszahlen** können Sie die Menge der zu generierenden Zufallszahlen bestimmen. Bitte geben Sie hier **1000** an.

Als Verteilung müssen Sie **Standard** für die Normalverteilung auswählen. Jetzt erscheint ein Feld, in dem Sie den Mittelwert und die Standardabweichung angeben können.

In das Feld **Ausgangswert** muss nichts eingetragen werden.

Als **Ausgabebereich** muss jeweils die oberste Zelle der aktuellen Spalte eingetragen werden (d. h. also **B10...F10**). Damit ist die Eingabe abgeschlossen.

Diese Schrittfolge muss für jede der Spalten mit den entsprechenden Mittelwerten und Standardabweichungen durchgeführt werden, insgesamt also fünfmal.

Analysefunktion Histogramm (Aufgabe (a), Blatt **Daten a**)

Rufen Sie die Analyse-Funktion **Histogramm** auf. Diese Funktion klassiert zunächst eine Menge von Zahlen und bestimmt dann die Klassenhäufigkeiten. Außerdem kann man sich ein Diagramm ausgeben lassen. Von dieser Funktion wird hier allerdings kein Gebrauch gemacht, da EXCEL hier ein Stabdiagramm an Stelle eines Histogramms erstellt.

Im **Eingabebereich** müssen Sie jeweils die erste Zelle einer der Spalten aus **Tabelle 1** angeben (d.h. **B10:B1009.... F10:F1009**).

Der **Klassenbereich** ist immer gleich. Hier geben Sie die Werte der Spalte **Intervalluntergrenzen** an (**H10:H130**). EXCEL benutzt diese Werte als Untergrenzen bei der Intervallklasseneinteilung.

Als Ausgabebereich geben Sie immer die nächste freie Zelle in der ersten Zeile an. Also für die erste Berechnung **J9**, für die zweite Berechnung **L9** usw. (es werden von der Funktion immer zwei Spalten ausgefüllt).

Die restlichen Felder können Sie ignorieren.

Nachdem Sie die Tabelle fertig ausgefüllt haben, löschen Sie bitte alle Spalten, in denen die Klassen aufgeführt sind, mit Ausnahme der ersten (d.h. die Spalten **L, N, P** und **R**).

Zeile / Spalte löschen (Aufgabe (a), Blatt **Daten a**)

Um mehrere Spalten bzw. Zeilen auf einmal zu löschen, müssen Sie mit dem Mauszeiger auf eines der entsprechenden Felder des grauen Spalten- bzw. Zeilenkopfes klicken (Wenn Sie auf den Kopf klicken, wird automatisch die ganze Zeile/Spalte markiert). Dann drücken Sie die Steuerungstaste. Jetzt können noch beliebig viele weitere Zeilen/Spalten zum Löschen markiert werden. Wenn alle Zellen markiert sind, rechte Maustaste drücken und **Zellen löschen** wählen.

Relative Häufigkeiten (Aufgabe (a), Blatt **Daten a**) In dieser Tabelle sollen die relativen Häufigkeiten der Klassen eingetragen werden. Insgesamt wurden in jeder Spalte 1000 Werte erzeugt, d.h. jeder Wert aus der Tabelle **Absolute Häufigkeiten** muss einfach durch 1000 geteilt werden.

Am leichtesten ist dies zu realisieren, wenn in der erste Zelle der Tabelle die relative Adresse der entsprechenden Zelle der Tabelle **Absolute Häufigkeiten** (**K10**) "durch 1000 geteilt" eingetragen wird. Dann können die Werte mit der rechten Maustaste automatisch nach rechts und unten ausgefüllt werden.

Histogramm erzeugen (1) (Aufgabe (a), Blatt **Daten a**)
Bevor Sie den Diagramm-Manager aufrufen, markieren Sie bitte den Bereich der relativen Häufigkeiten inklusive der Kopfzeile (**R8:V130**).

Wählen Sie als Diagrammtyp den Typen "Gestapelte Säulen".

Tragen Sie als x-Werte die Klassengrenzen ein (**Q10:Q130**), und geben Sie der Graphik einen Namen (z.B. "Histogramm Normalverteilte ZZ").

Fügen Sie das Histogramm als Objekt auf dem Tabellenblatt **Graphen 1** ein.

Histogramm erzeugen (2) (Aufgabe (b), Blatt **Daten b**)
Erzeugen Sie bitte zwei Diagramme:

Das erste Histogramm soll die $N(2,3^2)$-verteilten Zufallzahlen und deren Standardisierung enthalten. Markieren Sie die entsprechenden Zeilen inklusive der Überschriften (**N8:O120**), und wählen Sie wieder das Diagrammformat "Gestapelte Säulen". Als x-Werte müssen Sie die Spalte der Klassengrenzen angeben (**F10:F120**).Wählen Sie einen Titel (z.B. "Histogramm $N(2,9)$-verteilte und standardisierte ZZ") und fügen Sie die Graphik auf **Graphen2** ein.

Das zweite Histogramm soll die standardisierten, ursprünglich $N(2,3^2)$- verteilten und die standardnormalverteilten Zufallszahlen enthalten. Beachten Sie hier nur die Klassen im Bereich $(-4;4)$. Markieren Sie diese Werte (**O40:P80**) und die beiden Überschriften (mit gedrückter Steuerungstaste zusätzlich **O8**und **P8** anklicken). Wählen Sie den Diagrammtyp "Gestapelte Säulen" und geben sie als x-Werte die Klassengrenzen von -4 bis 4 an (**F40:F80**). Als Titel können sie z.B. "Histogramm standardisierte und standardnormalverteilte ZZ" wählen. Fügen Sie auch diese Graphik auf Blatt **Graphen2** ein.

Institut für Statistik und Mathematische Wirtschaftstheorie Prof. Heller

11. Schätzfunktionen, Erwartungstreue und Varianzminimalität

Auf der Suche nach der "richtigen" Schätzfunktion

In dieser Aufgabe sollen Sie sich mit den Eigenschaften von Schätzfunktionen vertraut machen. Nach einer kurzen (hoffentlich nicht zu theoretischen) Einführung sollen Sie sich anhand von Beispielen die Begriffe Erwartungstreue, Varianzminimalität und Konsistenz verdeutlichen. Zunächst stellt sich die Frage, wofür man Schätzfunktionen überhaupt braucht. Meistens will man mit Hilfe einer Schätzfunktion einen unbekannten Parameter schätzen. Am leichtesten kann man sich die Situation an einem Beispiel klarmachen:

In der Schraubenfabrik von Herrn Fischer werden nicht alle Schrauben fehlerfrei produziert. Herr Fischer möchte gerne wissen, wie groß der Anteil kaputter Schrauben an der Gesamtproduktion ist. Dazu läßt er eine Stichprobe vom Umfang 10 entnehmen.

Intuitiv wird man jetzt die kaputten Schrauben zählen und durch die Gesamtanzahl der Stichprobe teilen, um eine Schätzung für den Anteil der kaputten Schrauben zu erhalten. Nur warum ist das richtig?

Machen wir uns zunächst die Situation nochmal klar: Wir betrachten eine Bernoulli-Verteilung mit Parameter p (eine entnommene Schraube ist entweder heil=0 oder kaputt=1. Dabei ist p, der Anteil kaputter Schrauben, die Wahrscheinlichkeit, dass eine kaputte Schraube gezogen wird).

Formal können wir die Entscheidungssituation jetzt so darstellen:

Zustandsraum Z:

Bezeichnet die Menge aller Zustände. Mit Zuständen sind alle möglichen Werte für den gesuchten Parameter g gemeint (das ist in unserem Fall die Wahrscheinlichkeit p, dass eine Schraube kaputt ist, also alle Werte im Intervall [0,1]). Der wahre Parameter kommt aus diesem Zustandsraum und ist uns unbekannt.

Aktionenraum Γ:

Bezeichnet die Menge der möglichen Entscheidungen. Eine Entscheidung besteht aus der Angabe eines Parameterwertes. Der Aktionenraum muss also auch alle möglichen Parameterwerte enthalten. Mit einer Schätzung des Parameters entscheidet man sich für den Wert, der einem am sinnvollsten erscheint.

Informationenraum X:

Ist gleich dem Stichprobenraum. Er enthält alle möglichen Stichproben. In diesem Falle alle Vektoren aus dem Raum $\{0,1\}^n$. $x = (x_1,\ldots,x_n)$, mit $x_i = 0$ oder 1, ist dabei die realisierte Stichprobe.

Schätzfunktion δ:

Ist eine Entscheidungsfunktion. Mit ihrer Hilfe wollen wir alle Informationen, die uns durch die Stichprobe vorliegen, verarbeiten und unseren Schätzwert für g berechnen. Wir bilden also den Informationenraum auf den Aktionenraum ab. Damit gilt: $\delta : X -> \Gamma$

$$\delta(x) = \hat{\gamma}, \quad \text{wobei} \quad \hat{\gamma} \quad \text{unsere Entscheidung ist.}$$

Herr Fischer möchte, falls p sehr groß ist, seine Schraubenmaschine durch ein neues besseres Exemplar ersetzen. Wenn er den Parameter nicht richtig einschätzt und deshalb die falsche Entscheidung trifft, entsteht ihm ein Schaden (entweder er schätzt p zu hoch ein, dann kauft er sich eventuell eine neue Maschine, obwohl die alte gut funktioniert; oder er schätzt p zu niedrig, dann kauft er keine neue Maschine, sein wahrer Ausschuß ist aber höher als er denkt).

Eine genaue Berechnung dieses Schadens ist leider nicht möglich, deshalb greift Herr Fischer auf die Quadratische Schadensfunktion zurück.

a) Geben Sie bitte die quadratische Schadensfunktion an. Benutzen Sie dazu den Formeleditor in EXCEL (wählen Sie bitte aus dem Pull-down-Menü "Einfügen" die Option "Objekt", dann auf dem Register "neu erstellen" den "Microsoft Formel-Editor 3.0").

$$S(\gamma,\hat{\gamma}) = (\gamma - \hat{\gamma})^2$$

Nennen Sie Vor- und Nachteile der quadratischen Schadensfunktion:

–

–

Ausgehend von der Schadensfunktion kann jetzt die sogenannte Risikofunktion bestimmt werden. Dabei handelt es sich um nichts anderes als den Erwartungswert der Schadensfunktion, d.h. den mittleren quadratischen Fehler.

$$E_\gamma(S(\gamma,\hat{\gamma})) = E_\gamma(S(\gamma,\delta(x))) = E_\gamma[(\gamma - \delta(x))^2] = R(\delta,\gamma)$$

Der Parameter $\hat{\gamma}$ ist abhängig von der gewählten Schätzfunktion $\delta(x)$. Herr Fischer will sein Risiko natürlich minimieren und möchte deshalb wissen, was für eine Schadensfunktion er wählen soll. Er weiss, dass man das Risiko in zwei Schritten minimieren kann. Dazu muss man wissen, dass

$$R(\delta,\gamma) = Var_\gamma(\delta(x)) + (\gamma - E_\gamma[\delta(x)])^2$$

mit:

$Var_\gamma(\delta(x))$ = **Streuung der Schätzfunktion:** zeigt an, wie stark der Schätzwert durch die Zufälligkeit der Stichprobe streut, dies wird durch die Schätzfunktion $\delta(x)$ beeinflusst, und natürlich durch den wahren Parameter γ.

$(\gamma - E_\gamma[\delta(x)])^2$ = **Quadrat der Abweichung des mittleren Schätzwertes $\hat{\gamma}$ vom wahren Parameter γ.**

Da beide Terme positiv sind, müssen wir beide minimieren. Die Streuung der Schätzfunktion kann leicht minimiert werden, indem man die Schätzfunktion von der Stichprobe löst, d.h. sie unabhängig von der Stichprobe wählt, z.B. könnte man als Schätzfunktion

$$\delta(x) = 0{,}3$$

wählen.

b) Warum ist dieses Vorgehen nicht sinnvoll? Überlegen Sie sich, was für Auswirkungen diese Wahl der Stichprobe auf den ersten und auf den zweiten Term der Risikofunktion hat.

–

–

Es ist anscheinend nicht sinnvoll mit dem ersten Term der Gleichung zu beginnen, versuchen wir es daher mit dem zweiten. Will man das Quadrat der Abweichung des mittleren Schätzwertes $\hat{\gamma}$ vom wahren Parameter γ minimieren, muss gelten:

$$\gamma = E_\gamma[\delta(x)] \quad \text{für} \quad \gamma \quad \epsilon \quad \Gamma$$

Wenn dies für eine Schätzfunktion gilt, ist sie **erwartungsteu**, d.h. um die beste Schätzfunktion zu finden sollte man nur erwartungstreue Funktionen in Betracht ziehen.

$$E_\gamma[\delta(x)] - \gamma$$

wird auch Bias oder Verzerrung von δ genannt. **Eine erwartungstreue Schätzfunktion hat also eine Verzerrung von 0.**

c) Herr Fischer betrachtet die folgenden fünf Schätzfunktionen für den Parameter p. Welche der fünf Schätzfunktionen kann er ausschliessen, weil sie nicht erwartungstreu sind?

1.) $\hat{p} = \frac{1}{2}(x_1 + x_2)$

2.) $\hat{p} = 0{,}5 + \frac{1}{10}\sum_{i=1}^{10} x_i$

3.) $\hat{p} = \frac{2}{3}(\frac{1}{9}\sum_{i=2}^{10} x_i) + \frac{1}{3}x_1$

4.) $\hat{p} = \frac{1}{10}\sum_{i=1}^{10} x_i$

5.) $\hat{p} = \sum_{i=1}^{10} \frac{1}{i}x_i$

–

–

d) Herr Fischer konnte zwei der oben genannten Funktionen ausschließen. Er weiss, dass eine der drei übrigen Schätzfunktionen die gleichmäßig beste erwartungstreu Schätzfunktion für seinen Parameter ist. Der obigen Überlegung folgend will er nun den ersten Term der Risikofunktion $Var_\gamma(\delta(x))$ minimieren. D.h. für die gleichmäßig beste Schätzfunktion δ^* soll gelten:

$$Var_\gamma(\delta^*(x)) \leq Var_\gamma(\delta(x)) \quad \text{für alle} \quad \gamma \quad \text{aus} \quad \Gamma$$

Wie kann Herr Fischer ermitteln welche der drei Funktionen die gleichmäßig beste erwartungstreue Schätzfunktion ist?

–

–

e) Die Stichprobe hat folgenden Vektor x ergeben:

$$x = (x_1,\ldots,x_n)^T = (1,0,0,0,1,0,0,1,0,0)^T$$

Dabei gilt:

$x = (x_1,\ldots,x_{10})^T$, mit

$x_i = 0$ falls die entnommene Schraube heil ist, oder
$x_i = 1$ falls die entnommene Schraube kaputt ist.

Nutzen Sie die unter d) identifizierte gleichmäßig beste erwartungstreue Schätzfunktion, um p, den Anteil kaputter Schrauben, zu schätzen. Verwenden Sie die Summenfunktion von EXCEL bei Ihrer Berechnung.

$\hat{\gamma} = \hat{p} =$

Institut für Statistik und Mathematische Wirtschaftstheorie Prof. Heller

12. Maximum-Likelihood-Schätzung

Die Likelihood-Schätzung ist ein anderes Verfahren der Parameterschätzung als das in der letzten Aufgabe betrachtete. Auch hierbei wird die Information einer Stichprobe verwendet, aus der zusammen mit der Verteilungsannahme die Likelihood-Funktion erstellt wird.

Idee der Likelihood-Schätzung ist, dass man für jeden Stichprobenwert den dazugehörigen "wahrscheinlichsten" Parameter angeben möchte. Hat man eine Stichprobe mit mehreren Werten, werden diese natürlich aggregiert, um das Ergebnis genauer zu machen.

Die Likelihood-Funktion erstellt man, indem der gezogene x-Wert als fest in der Dichtefunktion bzw. Wahrscheinlichkeitsverteilung angenommen wird, und der gesuchte Parameter als variabel. Mit Hilfe der Likelihood-Funktion kann jetzt der "wahrscheinlichste" Parameter berechnet werden. Man muss einfach das Maximum der Funktion bestimmen.

a) Ferdinand hat sich einen neuen Computer gekauft. Nach kurzer Zeit stellt er fest, dass bei bestimmten Progammen immer wieder Softwarefehler auftreten. Er nimmt an, dass die Zeiten, die zwischen den einzelnen Fehlern liegen exponential verteilt sind.

Um eine genauere Aussage über die Verteilung machen zu können, misst er dreimal die Zeit zwischen zwei Fehlern. Seine Stichprobe enthält folgende Werte (in Stunden):

$$x_1 = 4{,}x_2 = 2{,}x_3 = 0{,}9$$

Ferdinand möchte zunächst für jeden x-Wert eine eigene Likelihoodfunktion angeben und das zugehörige optimale λ berechnen. Geben Sie bitte die Likelihoodfunktionen für die drei oben genannten x-Werte an (benutzen Sie dazu den Formeleditor).

b) Um die drei Funktionen graphisch darzustellen, füllen Sie bitte die untenstehende Tabelle aus (bitte benutzen sie Verweise auf die x- und λ-Werte). Die dazugehörige Graphik können Sie auf dem Tabellenblatt "Graphen 1" betrachten.

c) Bestimmen Sie die Schätzwerte λ_i^* anhand der Graphik. Berechnen Sie danach die exakten Werte. Warum können Sie anstatt der Funktion $L_i(\lambda)$ die Funktion $\ln L_i(\lambda)$ maximieren?

–

–

Ferdinand stellt fest, dass die drei λ-Werte unterschiedlich groß sind, und er mit dem Ergebnis eigentlich wenig anfangen kann. Wie kann er das Ergebnis verbessern?

–

–

d) Die Likelihoodfunktion der aggregierten Daten ist das Produkt der einzelnen Likelihoodfunktionen. Es gilt also:

$$L(\lambda) = \Pi_{i=1}^{n} L_i(\lambda)$$

Geben sie die aggregierte Likelihood-Funktion an und bestimmen sie λ^*.

- $L(\lambda) =$
- $\lambda^* =$

Warum ist es zulässig, dass die Daten aggregiert werden? Betrachten Sie die Graphik auf dem Tabellenblatt "Graphen 2". Warum hat die Funktion $L(\lambda)$ im Vergleich zu den anderen drei Funktionen so niedrige Funktionswerte? Wie verändert sich $L(\lambda)$ wenn man die unterschiedlichen x_i verändert?

-
-
-

e)* Ferdinand ist nicht nur an den Zeiten zwischen den Ausfällen interessiert, sondern möchte auch wissen, wie häufig seine Progamme während eines Tages abstürzen. Er zählt die Ausfälle an drei verschiedenen Tagen jeweils in einer Zeit von fünf Stunden. Sein Ergebnis sieht wie folgt aus (Anzahl der Ausfälle in fünf Stunden).

$$x_1 = 6, x_2 = 3, x_3 = 1$$

Da Ferdinand in Aufgabe a) angenommen hatte, dass seine Zwischenausfallzeiten exponential verteilt sind, nimmt er nun als logischen Schluss an, dass die Anzahl seiner Ausfälle poissonverteilt ist (siehe "Poissonprozess").

Geben Sie bitte die drei einzelnen Likelihoodfunktionen für die x-Werte und die gemeinsame Likelihoodfunktion an und berechnen Sie die Werte für λ_1^*, λ_2^*, λ_3^* und λ^*:

- $L_1(\lambda) =$ bzw. $\lambda_1^* =$
- $L_2(\lambda) =$ bzw. $\lambda_2^* =$
- $L_3(\lambda) =$ bzw. $\lambda_3^* =$
- $L(\lambda) =$ bzw. $\lambda^* =$

Auf dem Tabellenblatt "Graphen 3" können Sie den Kurvenverlauf der Funktionen betrachten. Wie ändern sich die Funktionen wenn die x-Werte verändert werden? Wie häufig glaubt Ferdinand sein Programm neu aufrufen zu müssen, wenn er zehn Stunden lang damit arbeitet?

-
-
-

f)* Überlegen Sie sich, ob die Likelihoodfunktion der Poissonverteilung stetig oder diskret ist. Betrachten Sie zum besseren Verständnis die Graphik auf Tabellen Blatt "Graphen 4".

–

–

g)* Wenn man einen Schätzer für λ mit Hilfe der Likelihoodfunktion berechnet, macht man eigentlich nichts anderes, als den Parameter mit dem besten "Fit" zu den gegebenen Daten zu suchen.

Diesen besten "Fit" kann man auch bestimmen (wenn auch nicht so genau), indem man die theoretische Verteilungsfunktion einer Verteilung der empirischen anpasst, d.h. die theoretische Verteilung mit unterschiedlichen Parametern über die empirische legt und diejenige aussucht, die der empirischen am nächsten kommt.

Auf dem Tabellenblatt "Daten g" ist eine Tabelle mit der empirischen und der theoretischen Verteilungsfunktion exponentialverteilter Zufallszahlen vorbereitet. Füllen Sie die Spalte der theoretischen Verteilung aus. Benutzen Sie dabei einen Verweis auf λ.

Auf dem Tabellenblatt "Graphen 5" können Sie jetzt die beiden Funktionen betrachten. Ändern Sie λ auf dem Tabellenblatt "Daten g" in einen Verweis auf λ auf dem Tabellenblatt "Graphen 5".

Jetzt können Sie λ verändern und dabei beobachten, wie sich die Kurve ändert. Für welchen λ-Wert gibt es den besten "Fit"?

–

–

Institut für Statistik und Mathematische Wirtschaftstheorie Prof. Heller

13. Intervallschätzung

Im Gegensatz zur Punktschätzung wird bei der Intervallschätzung nicht nur ein Schätzer angegeben, sondern ein Intervall, in dem der gesuchte Wert liegen kann.

Dieses Intervall überdeckt den wahren (aber unbekannten) Parameter mit einer Wahrscheinlichkeit (1-a). Es heißt deshalb auch Konfidenzintervall zum Niveau (1-a) oder (1-a)*100

Je höher das Niveau, d.h. je kleiner die Irrtumswahrscheinlichkeit a ist, desto größer ist die Wahrscheinlichkeit, dass der wahre Parameter vom Konfidenzintervall überdeckt wird. Das Konfidenzintervall wird dann aber auch immer breiter, so dass ab einem bestimmten Niveau keine sinnvolle Aussage mehr möglich ist.

a) Auf dem Tabellenblatt "Daten a" ist ein Programm für Sie vorbereitet, das 50 normalverteilte Stichproben des Umfangs n zieht.

Anschließend berechnet das Programm das Konfidenzintervall zu jeder der Stichproben für den Parameter μ. Dann werden die Stichproben, ihre Konfidenzintervalle und die Stichprobenmittelwerte ausgegeben. Zusätzlich wird die Verteilung der kumulierten Stichproben und der Mittelwerte als Histogramm angezeigt.

Zur Erzeugung der Stichproben muss der Parameter μ angegeben werden, im weiteren Verlauf der Aufgabe sollen Sie aber davon ausgehen, dass Sie μ nicht kennen.

Zunächst müssen Sie folgende Parameter festlegen:

1. die Irrtumswahrscheinlichkeit α ($0 < \alpha <= 1$)
2. den Stichprobenumfang n ($0 <$ n $<= 50$)
3. ob Sie σ als bekannt oder nicht bekannt annehmen wollen (1 = bekannt; 0 = unbekannt)
4. den wahren Parameter μ, den Sie im weiteren als unbekannt annehmen
5. den wahren Parameter σ, den Sie je nach 3. als bekannt oder unbekannt annehmen

(die Zellen für $\bar{x}$, Ober- und Untergrenze enthalten Verweise. Löschen Sie diese Felder bitte nicht)

Mit den beiden Buttons "Histogramm erzeugen" bzw. "Histogramm löschen" können Sie das Programm starten, bzw. die Graphik wieder löschen (**bevor sie das Programm starten, müssen Sie die Graphik löschen!!**).

Wählen Sie folgende Werte für die Parameter:

α	0,1
n	9
$\sigma_{bekannt}$	1
μ	0
σ	1

Welche Werte werden durch das rote, welche durch das blaue Histogramm angezeigt? Beschreiben Sie die Gestalt der Histogramme (Lage, Symmetrie, Spannweite, etc.), wie unterscheiden sie sich?

–

–

Was stellen die roten Punkte, was die blauen vertikalen Striche dar? Was zeigen die horizontalen Linien an? Warum sind einige von ihnen hellblau, andere dunkelblau? Was symbolisiert die vertikale Linie in der Mitte der Graphik?

–

–

–

Wie groß ist das Konfidenzintervall der Gesamtverteilung (Ober- und Untergrenze sind neben der Graphik angegeben) im Vergleich zu der von Ihnen geschätzten durchschnittlichen Länge der Vertrauensbereiche. Wie viele hellblaue Linien gibt es? Wie steht die Anzahl der hellblauen Linien mit α in Bezug?

–

–

–

b) Wiederholen Sie die Parameterschätzung für nachfolgende Werte. Füllen Sie die unten stehende Tabelle dazu aus.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**
α	0,1	0,05	0,05	0,4	0,05	0,05	0,1
n	9	9	50	9	9	9	50
$\sigma_{\textbf{bekannt}}$	1	1	1	1	0	1	0
μ	1	1	1	5	5	5	5
σ	1	1	1	1	1	3	3
Anzahl hellblauer Linien							
Durchschn. Länge der KI							
Mittelwert der Gesamtziehung							
Breite des Gesamt-KI							

Schätzen Sie grob die durchschnittliche Länge der Vertrauensbereiche ab. Gibt es Punktschätzer für μ, die ungefähr gleichgroß sind, aber deutlich unterschiedlich lange Konfidenzintervalle haben? Was könnte der Grund dafür sein?

–

–

Wie ändern sich die Konfidenzintervalle, wenn man den Umfang der Stichprobe erhöht? Was verändert sich, wenn die Standardabweichung als unbekannt angenommen wird?

–

–

Welchen Einfluss hat die Irrtumswahrscheinlichkeit auf die Breite der Konfidenzintervalle? Wie verändern sich die Konfidenzintervalle, wenn man μ und σ verändert? Was für Auswirkungen haben diese Änderungen noch?

–

–

Was würde sich an der Konfidenzintervallbreite ändern, wenn Sie statt 50 hundert Stichproben ziehen könnten? Wie breit wäre ein Konfidenzintervall des Niveaus 1 bzw. 0?

–

–

c) Als nächstes sollen Sie Konfidenzintervalle der Varianz einer Normalverteilung betrachten. Auf dem Tabellenblatt "Daten c" ist ein Programm vorbereitet, das analog zu dem der Aufgaben a und b arbeitet. Der Unterschied besteht darin, dass Konfidenzintervalle der Varianz berechnet und 100 statt nur 50 Stichproben gezogen werden.

Beginnen Sie mit folgenden Variablen:

α	0,05
n	9
$\mu_{bekannt}$	1
μ	0
σ	1

Warum werden die Stichproben nicht als Einzelwerte ausgegeben? Warum wird nur ein Histogramm angezeigt? Beschreiben Sie die Gestalt des Histogramms (Lage, Symmetrie, Spannweite, etc.).

Was ist an den Vertrauensbereichen der Varianz anders als bei denen der Mittelwerte?

Was zeigen die horizontalen Linien an? Warum sind einige von ihnen hellblau, andere dunkelblau? Was symbolisiert die vertikale Linie?

–

–

–

Schätzen Sie grob die durchschnittliche Länge der Vertrauensbereiche ab. Gibt es - wie im Aufgabenteil b) Schätzung von μ bei unbekannter Varianz - auch hier Punktschätzer für σ^2, die ungefähr gleich groß sind, aber deutlich unterschiedlich lange Konfidenzintervalle haben?

–

–

–

Wie viele hellblaue Linien gibt es? Wie steht die Anzahl der hellblauen Linien mit α in Bezug?

–

–

d) Wiederholen Sie die Parameterschätzung für nachfolgende Werte. Füllen Sie die unten stehende Tabelle dazu aus.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**
α	0,1	0,05	0,05	0,4	0,05	0,05	0,1
n	9	9	50	9	9	9	50
$\mu_{\textbf{bekannt}}$	1	1	1	1	0	1	0
μ	0	0	0	5	5	5	5
σ	1	1	1	1	1	3	3
Anzahl hellblauer Linien							
Durchschn. Länge der KI							
Mittelwert der Gesamtziehung							
Breite des Gesamt-KI							

Wie ändern sich die Konfidenzintervalle, wenn man den Umfang der Stichprobe erhöht? Was verändert sich, wenn der Mittelwert als unbekannt angenommen wird? Welchen Einfluss hat die Irrtumswahrscheinlichkeit auf die Breite der Konfidenzintervalle?

–

–

–

Wie verändern sich die Konfidenzintervalle, wenn man μ und σ verändert? Was für Auswirkungen haben diese Änderungen noch? Wie breit wäre ein Konfidenzintervalle des Niveaus 1 bzw. 0?

–

–

–

Institut für Statistik und Mathematische Wirtschaftstheorie Prof. Heller

14. Parametertests

Bitte laden Sie für diese Übung zunächst **EXCEL** (per Doppelklick auf das EXCEL-Icon auf dem Desktop bzw. durch Wahl des entsprechenden Eintrags aus dem Startmenü). Öffnen Sie anschließend die Datei **Aufgabe 14.xls**, die Sie auf Laufwerk **k:** finden. Zum Öffnen einer Datei wählen Sie bitte in EXCEL nach Anklicken des Microsoft-Office-Buttons (oben links) den Eintrag **Öffnen**. Im Dialog **Öffnen** müssen Sie nun das Laufwerk **k:** auswählen. Anschließend sollten Sie im Dateifenster die oben genannte Datei sehen. Wählen Sie diese per Mausklick aus und bestätigen Sie abschließend per Mausklick auf den Schalter **Öffnen** Ihre Wahl. Die Datei sollte jetzt geöffnet werden. EXCEL wird Ihnen dabei mitteilen, dass die Datei Makros enthält und fragen, ob diese aktiviert werden sollen. Bestätigen Sie dies durch Anklicken von "Diesen Inhalt aktivieren".

In dieser Aufgabe sollen Sie sich mit Parametertests vertraut machen.

Mit Tests versucht man allgemein, Annahmen bezüglich der Eigenschaften einer Zufallsvariable zu verifizieren. Hierzu wird untersucht, ob die Ergebnisse einer Stichprobenziehung mit einer zuvor formulierten Annahme verträglich sind.

Analog der Parameterschätzung wird die Information einer Stichprobe durch eine Teststatistik geeignet zusammengefasst. Diese stellt ebenfalls eine Zufallsvariable dar.

Man kann dann eine Stichprobe ziehen, den Wert der Teststatistik berechnen und diesen mit den Eigenschaften der theoretischen Verteilung vergleichen, die bei Gültigkeit der Annahme gegeben ist.

Im folgenden sollen Sie daher zunächst die Verteilungseigenschaften der Teststatistiken, die bei Tests basierend auf normalverteilten Beobachtungen von Interesse sind, untersuchen.

Teststatistik für μ:

Die Teststatistik für einen Test auf den Mittelwert μ ist der Stichprobenmittelwert:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Die folgenden beiden Situationen müssen unterschieden werden:

1. σ^2 ist bekannt:

   Wenn σ^2 bekannt ist, gilt:

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

Da man für den Vergleich der theoretischen Verteilung mit der Stichprobe die Quantile benötigt, muss $\bar{x}$ standardisiert werden, damit man diese aus der Tabelle der Standard-Normalverteilung ablesen kann. Diese standardisierte Teststatistik wird im folgenden mit z bezeichnet. Für z gilt:

$$\begin{aligned} z &= \sqrt{n}\,\frac{\bar{x}-\mu}{\sigma} \\ z &\sim N(0,1) \end{aligned}$$

2. σ^2 ist unbekannt:

   Ist σ^2 unbekannt, muss für diesen Parameter anhand der Stichprobe ein Wert geschätzt werden. Ein erwartungstreuer Schätzer für σ^2 ist die korrigierte Stichprobenvarianz:

$$\bar{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

   Für z gilt dann:

$$\begin{aligned} z &= \sqrt{n}\,\frac{\bar{x}-\mu}{\bar{s}_n} \\ z &\sim t_{n-1} \end{aligned}$$

Teststatistik für σ^2:

Auch hier muss danach unterschieden werden, ob der Wert des zweiten Parameters (μ) bekannt ist oder ebenfalls geschätzt werden muss.

1. μ ist bekannt:

   Ist μ bekannt, kann die Stichprobenvarianz als Teststatistik für σ^2 benutzt werden:

$$\hat{\sigma}^2 = s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

   Für z gilt dann:

$$\begin{aligned} z &= \frac{n}{\sigma^2}\,s_n^2 \\ z &\sim \chi_n^2 \end{aligned}$$

2. μ ist unbekannt:

   Ist μ unbekannt, muss die korrigierte Stichprobenvarianz als Teststatistik für σ^2 benutzt werden, die auch bei unbekanntem μ erwartungstreu ist:

$$\hat{\sigma}^2 = \bar{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

   Für z gilt jetzt:

$$\begin{aligned} z &= \frac{n-1}{\sigma^2}\,\bar{s}_n^2 \\ z &\sim \chi_{n-1}^2 \end{aligned}$$

a) Nach dieser theoretischen Vorüberlegung sollen Sie nun anhand von Beispieldaten die obigen Aussagen über Teststatistiken verifizieren. Hierzu wurden auf den Tabellenblättern "Daten a1" - "Daten a4" Stichproben des Umfangs 5 aus einer $N(\mu,\sigma^2)$-verteilten Grundgesamtheit vorbereitet.

Auf den Tabellenblättern "Daten a1" und "Daten a2" sollen Sie die Schätzer für μ untersuchen, auf den Tabellenblättern "Daten a3" und "Daten a4" die für σ^2. Die Tabellenblätter "Graphen a*" enthalten jeweils graphische Darstellungen der Verteilungen der Teststatistiken.

1. μ bei bekannter Varianz:
   Berechnen Sie zunächst die Teststatistik für μ und den dazugehörigen standardisierten Wert z.
   Benutzen Sie bei der Berechnung von z für μ und σ Verweise auf die Zellen "AA10" und "AA12" (absolute Adressierung!). Wie groß ist n?

   - $n =$

   Sie können die Intervallgrenzen berechnen lassen, indem Sie auf die Schaltfläche "Intervallgrenzen berechnen" klicken. Klassieren Sie nun $\bar{x}$ und z mit Hilfe der Analysefunktion "Histogramm" (Registerkarte "Daten", Gruppe "Datenanalyse") und diesen Intervallgrenzen (bitte keine Spalten löschen!). In der Tabelle "Wahrscheinlichkeitsdichten" werden Ihnen jetzt die Werte für $\bar{x}$,z,$N(\mu,\sigma^2/5)$,$N(0,1)$,t_4,χ^2_4 und χ^2_5 angezeigt. In der ersten Zeile dieser Tabelle ("Kurve anzeigen") können Sie mit 1 (= anzeigen) oder 0 (= nicht anzeigen) angeben, welche der Funktionen im Diagramm auf Tabellenblatt "Graphen a1" angezeigt werden sollen. Betrachten Sie die graphische Darstellung der Daten.
   Welcher der vier theoretischen Verteilungen scheinen $\bar{x}$ und z zu entsprechen?

   - $\bar{x}$:
   - z :

   Wiederholen Sie das Experiment für Stichproben, deren Beobachtungen einer Normalverteilung mit einem anderen Erwahrungswert μ und / oder einer anderer Varianz σ^2 gehorchen. Ändern Sie dazu μ und σ in den Zellen "AA10" und "AA12". Sie müssen anschließend die Intervallgrenzen neu berechnen lassen und die Klassierungen wiederholen, damit die Änderung in der Graphik sichtbar wird.
   Wie ändern sich die Verteilungen von $\bar{x}$ und z?

   - $\bar{x}$:
   - z :

2. μ bei unbekannter Varianz:

   Berechnen Sie zunächst die korrigierten Stichprobenvarianzen. Dazu können Sie die EXCEL-Funktion "VARIAN**Z**(bereich)" verwenden.

   Achtung! Die Funktion "VARIAN**ZEN**(bereich)" berechnet die Varianz einer bekannten Grundgesamtheit und nicht die korrigierte Stichprobenvarianz.

   Das heißt:

   - VARIANZ(Zelle$_1$:Zelle$_n$) berechnet $\bar{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
   - VARIANZEN(Zelle$_1$:Zelle$_n$) berechnet $s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

   Als nächstes berechnen Sie bitte z. Verwenden Sie für μ und $\bar{s}_n^2$ Verweise auf die entsprechenden Zellen ("AC10" bzw. die richtige Zelle in Spalte "O"). Achten Sie darauf, dass $\bar{s}_n$ und nicht $\bar{s}_n^2$ in der Formel verwendet wird. Verfahren Sie im weiteren wie in Aufgabe 1. (d.h. Intervallgrenzen berechnen lassen und Daten klassieren). In der Tabelle "Wahrscheinlichkeitsdichten" können Sie wieder angeben, welche Graphen Sie betrachten wollen.

   Welche Verteilung entspricht der von z?

   •

   Wiederholen Sie das Experiment wie in Teil 1. für Beobachtungen mit anderem Mittelwert bzw. Varianz. Ändern Sie dazu die Werte für μ und σ in den Zellen "AC10" und "AC12". Sie müssen die Intervallgrenzen neu berechnen lassen und die Klassierungen wiederholen, damit die Änderung in der Graphik sichtbar wird.

   Wie ändern sich die Verteilungen von $\bar{x}$ und z?

   • $\bar{x}$:

   • z :

3. σ^2 bei bekanntem Mittelwert:

   Berechnen Sie die Teststatistik für σ^2 bei bekanntem μ und das dazugehörige z. Warum können Sie dazu nicht die oben genannte Funktion "VARIANZEN(bereich)" verwenden? Vergessen Sie nicht Verweise für μ und σ zu verwenden!

   •

Klassieren Sie die Daten und betrachten Sie die Graphen auf Tabellenblatt "Graphen a3". Welcher Verteilung scheint z hier zu gehorchen?

•

Ändern Sie wie in den vorangegangenen Abschnitten μ und σ in den Zellen "AA10" und "AA12". Sie müssen die Intervallgrenzen neu berechnen lassen und die Klassierungen wiederholen, damit die Änderung in der Graphik sichtbar wird.
Wie ändern sich die Verteilungen von s_n^2 und z? Müssen Sie Ihre eben gegebene Antwort revidieren?

- s_n^2 :
- z :
-

4. σ^2 bei unbekanntem Mittelwert:
   Verfahren Sie wie in Teil 3. Können Sie für die korrigierte Stichprobenvarianzen die Funktion "VARIANZ(bereich)" verwenden?

   •

   Verwenden Sie bei der Berechnung wieder Verweise für μ und σ. Klassieren Sie die Daten und betrachten Sie die Graphen auf Tabellenblatt "Graphen a4".
   Ändern Sie μ und σ in den Zellen "AA10" und "AA12". Sie müssen die Intervallgrenzen wieder neu berechnen lassen und die Klassierungen wiederholen, damit die Änderung in der Graphik sichtbar wird.
   Wie ändern sich die Verteilungen von $\bar{s}_n^2$ und z in diesem Fall?

   - $\bar{s}_n^2$:
   - z :

In den nächsten Aufgaben sollen Sie verschiedene Hypothesen anhand von Stichproben testen. Tests können im Sinne der statistischen Entscheidungstheorie als Entscheidungsfunktionen bezüglich einer zuvor getroffenen Annahme über die Eigenschaft einer Zufallsvariablen aufgefasst werden. Wir behandeln in den folgenden Aufgaben den Fall, dass eine Annahme mit Hilfe einer solchen Entscheidungsfunktion nicht abgelehnt oder abgelehnt werden soll. Diese Annahme wird als Nullhypothese, ihre Negation als Gegenhypothese bezeichnet. Die Nullhypothese ist hier die Behauptung, dass ein Parameter γ aus einem Parameterraum Γ in einen Teilbereich Γ_0 liegt, die Gegenhypothese besagt, dass der Parameter γ in Γ_1 liegt, wobei

$$\Gamma = \Gamma_0 \cup \Gamma_1 \text{ und } \Gamma_0 \cap \Gamma_1 = \emptyset$$

gelten.

In Aufgabe b) werden Sie einen Test für μ bei bekannter Varianz durchführen. Im Anschluss daran sollen Sie sich die Begriffe Fehler 1. und 2. Art anhand dieses Beispiels verdeutlichen. Danach soll in Aufgabe c) μ bei unbekannter Varianz getestet werden. Zusätzlich werden Sie eine leicht abgewandelte Form der Hypothesentests kennenlernen, die häufig in Statistik-Software-Programmen benutzt wird. Abschließend werden in Aufgabe d) noch zwei Test für σ^2 bei bekanntem bzw. unbekannten Erwartungswert μ durchgeführt.

b) In einer Molkerei wird Milch in 1 Liter Plastikbeutel abgefüllt. Für den Besitzer Herrn Müller ist wichtig, dass seine Abfüllanlage genau funktioniert:

Auf der einen Seite hat der Verbraucherverband "Molkereiprodukte" mit einer Klage gedroht, falls sich herausstellen sollte, dass in den Plastikbeuteln im Schnitt weniger als ein Liter Milch ist, auf der anderen Seite dürfen die Beutel auch nicht zu viel Milch enthalten, weil sie sonst platzen.

Herr Müller entschließt sich, die Justierung seiner Anlage mit einem Hypothesentest zu überprüfen. Er entscheidet sich dafür, eine Stichprobe vom Umfang 10 zu entnehmen und den Inhalt dieser Beutel genau zu messen. Seine Ergebnisse finden Sie auf dem Tabellenblatt "Daten b1". Zusätzlich weiß Herr Müller, dass seine Anlage mit einer Varianz von 400 ml^2 bei der Abfüllung arbeitet.

1. Welche Art Test muss Herr Müller durchführen, um beide Seiten zufrieden zu stellen? Geben Sie zunächst die Art des Tests sowie die Null- und Gegenhypothese auf dem Tabellenblatt "Daten b1" an.

   - H_0 :
   - H_1 :

   Herr Müller möchte, dass die Nullhypothese in höchstens 10% der Fälle fälschlicherweise abgelehnt wird. Das heißt, die Irrtumswahrscheinlichkeit α soll 0,1 sein.

   Daraus lässt sich nun der Annahmebereich für den Test ableiten: Die Verteilung der Teststatistik ist bekannt:

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

Da $\alpha = 10\%$ ist, soll der Annahmebereich so gewählt werden, dass $\bar{x}$ unter H_0 mit 90% Wahrscheinlichkeit ($= 1 - \alpha$) in ihm liegt. Es gilt also:

$$\begin{aligned} 1-\alpha &= P(Untergrenze \leq \bar{x} \leq Obergrenze) \\ &= P(\bar{x} \leq Obergrenze) - P(\bar{x} \leq Untergrenze) \\ &= P(\bar{x} \leq \mu_0 + \frac{\sigma}{\sqrt{n}}\, Q_{1-\frac{\alpha}{2}}) - P(\bar{x} \leq \mu_0 + \frac{\sigma}{\sqrt{n}}\, Q_{\frac{\alpha}{2}}) \end{aligned}$$

Daraus kann der Annahmebereich für einen zweiseitigen Test direkt abgelesen werden:

$$I = \left[\mu_0 + \frac{\sigma}{\sqrt{n}} Q_{\frac{\alpha}{2}} \;;\; \mu_0 + \frac{\sigma}{\sqrt{n}} Q_{1-\frac{\alpha}{2}}\right]$$

Geben Sie die Werte für die Parameter α, σ und μ an. Basierend auf diesen Werten können Sie nun die Quantile der Standard-Normalverteilung mit Hilfe der EXCEL-Funktion "STANDNORMINV" (siehe Hilfe "Quantile der Normalvert") berechnen.

- α :
- σ :
- μ :
- $Q_{\frac{\alpha}{2}}$:
- $Q_{1-\frac{\alpha}{2}}$:

Bestimmen Sie anhand dieser Informationen den Annahmebereich und schließen Sie auf den / die kritischen Bereiche. Für $-\infty$ bzw. ∞ können Sie eine große negative bzw. positive Zahl einsetzen. Überlegen Sie, wie groß diese Zahl mindestens sein sollte und warum. (Tipp: Überprüfen Sie mit Hilfe des entsprechenden Tabellenblattes und Übung a), in welchem Bereich Beobachtungen zu erwarten sind.)

- Annahmebereich :
- Kritischer Bereich :

Berechnen Sie anschließend den Wert der Teststatistik für μ. Benutzen Sie - wo möglich - Verweise.

- Wert der Teststatistik :

Tragen Sie abschließend ein, ob die Nullhypothese abgelehnt werden muss oder nicht (in der Hilfe "Logische Verknüpfungen" ist erklärt, wie Sie diese Werte dynamisch eintragen lassen können).

- Entscheidung :

Anhand dieses Beispiels kann man sich den Unterschied zwischen Fehler 1. und 2. Art sehr gut deutlich machen:

- Ein **Fehler 1. Art** tritt dann auf, wenn man die Nullhypothese ablehnt, obwohl sie richtig ist. In unserem Beispiel hieße das, der Wert der Teststatistik, also $\bar{x}$, liegt außerhalb des Annahmebereichs, obwohl der vermutete Parameter μ_0 richtig ist.
- Ein **Fehler 2. Art** hingegen wird gemacht, wenn die Nullhypothese nicht abgelehnt wird, obwohl sie falsch ist. Mit anderen Worten, $\bar{x}$ liegt im Annahmebereich obwohl $\mu \neq \mu_0$.

Den Annahmebereich haben Sie oben nach Vorgabe einer maximal zulässigen Irrtumswahrscheinlichkeit berechnet. Diese entspricht dem Fehler 1. Art. Da bei Ablehnung von H_0 mit Wahrscheinlichkeit $(1 - \alpha)$ davon ausgegangen werden kann, dass die Gegenhypothese korrekt ist, wird α klein gewählt.

Die Wahl einer Fehlerwahrscheinlichkeit 1. Art hängt also von den Interessen des Testenden und nicht von den Daten an sich ab. Mit abnehmendem α wächst der Annahmebereich des Tests.

Auf dem Tabellenblatt "Graphen b1" ist hierzu eine Graphik vorbereitet, die den Annahme- und Ablehnungsbereich sowie die Teststatistik zeigt. Überprüfen Sie die Veränderung im Annahmebereich durch nachträgliches Ändern von α.

2. Der Fehler 2. Art ist nicht so einfach darzustellen oder zu berechnen, weil man in diesem Fall keine Aussage über den wahren Parameter machen kann. Man weiß nur, dass es nicht μ_0 ist. Da man den wahren Parameter μ nicht kennt, kann keine Aussage darüber gemacht werden wie groß die Wahrscheinlichkeit des Fehlers 2. Art ist, sondern nur wie groß sie unter der Annahme eines bestimmten μ's und der Nullhypothese ist.

   In dieser Aufgabe soll deshalb die Wahrscheinlichkeit des Fehlers 2. Art in Abhängigkeit von μ dargestellt werden. Der Annahmebereich für die Nullhypothese ist Ihnen aus Aufgabe 1. bekannt. Ein Fehler 2. Art tritt dann auf, wenn der wahre Parameter μ ungleich μ_0 ist und der Wert der Teststatistik trotzdem im Annahmebereich liegt.

   D.h. es ist P(Fehler 2. Art) = P(Annahmebereich der Nullhypothese) für unterschiedliche μ_1 gesucht:

   "= NORMVERT(Obergrenze; μ_1: σ / WURZEL(n); 1)
   - NORMVERT(Untergrenze; μ_1; σ / WURZEL(n); 1) "

   Benutzen Sie für alle Werte Verweise und außer für die unterschiedlichen μ_1-Werte absolute Adressierungen.

Schauen Sie sich die Graphik der Wahrscheinlichkeiten für den Fehler 1. und 2. Art auf dem Tabellenblatt "Graphen b2" an. Wie heißen die beiden Kurven?

-
-

Ändern Sie den Wert α der Irrtumswahrscheinlichkeit. Wie verändern sich die beiden Kurven? Speziell, was für Auswirkungen hat ein sehr kleiner Fehler 1. Art auf den Fehler 2. Art, wenn der wahre Parameter μ nur unwesentlich von μ_0 abweicht?

-
-

Ändern Sie nun die Größe der Stichprobe (n). Welchen Einfluss hat n auf die Wahrscheinlichkeiten der Fehlers 1. und 2. Art?

-
-

Welche Vorgehensweise für die Auslegung eines Tests, d.h. die Wahl von α und n, würden Sie basierend auf den Ergebnissen dieser Aufgabe wählen? Welche sonstigen hier nicht betrachteten Restriktionen müssen eventuell noch beachtet werden?

-
-

c) Der Verbraucherverband verstärkt seinen Druck auf Herrn Müller und fordert, dass die Molkerei eindeutig nachweist, dass in den Milchbeuteln nicht weniger als 1 Liter Milch ist. Er beschuldigt Herrn Müller dabei, sein Testergebnis durch die Wahl einer falschen Varianz σ^2 verfälscht zu haben.

Herr Müller überzeugt den Verbraucherverband davon, dass ein Hypothesentest ein ausreichender Nachweis dafür sei, dass in seinen Milchbeuteln mindestens 1000 ml Milch sind. Der Verbraucherverband akzeptiert ein Signifikanzniveau von 10%.

1. Verfahren Sie wie in Aufgabe b1:
   Füllen Sie die Felder auf Tabellenblatt "Daten c1" aus: Nennen Sie zunächst die Art des Tests und die Null- bzw. Gegenhypothese, die Herr Müller wählen sollte, um den Verbraucherverband zu beruhigen.

   - H_0 :
   - H_1 :

   Geben Sie als nächstes die Werte für α, $\bar{s}_n$ und μ_0, sowie das Quantil der t-Verteilung an. Berechnen Sie das Quantil mit Hilfe der EXCEL-Funktion "TINV".

   - α :
   - μ :
   - $Q_{1-\alpha}$:

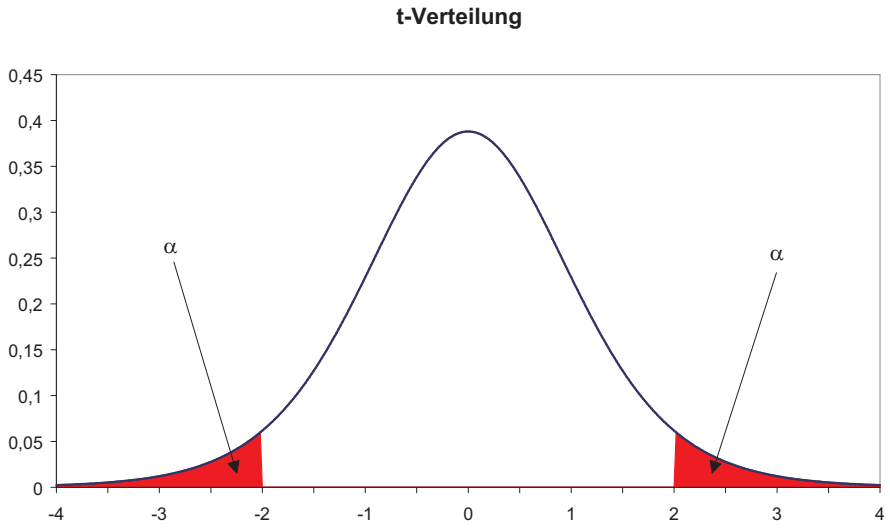
   "TINV" benötigt ähnlich wie "NORMINV" zwei Argumente: Die Wahrscheinlichkeit "Wahrsch" und die Anzahl der Freiheitsgrade. Im Gegensatz zu "NORMINV" kann man mit "TINV" aber nicht direkt negative Quantile berechen:

   – "TINV" berechnet den positiven x-Wert für zweischwänzige Quantile einer t-Verteilung mit Mittelwert = 0. D.h.

   $$\text{TINV}(\text{P}(\text{T}<\text{-Q}_\alpha) + \text{P}(\text{T} > \text{Q}_\alpha); \text{Freiheitsgrade}) = Q_\alpha$$

   Für das Beispiel unten (siehe Abbildung) heißt das:

   $$\text{TINV}(2\alpha; \text{Freiheitsgrade}) = 2$$



– Die Berechnung der Quantile, wie Sie es aus der Vorlesung kennen - also einschwänzig - gestaltet sich leider etwas komplizierter. Da "TINV" nur positive Quantilswerte ausgibt, muss eine Fallunterscheidung gemacht werden:
$\alpha < 0{,}5$ oder $\alpha > 0{,}5$.

Die folgende Tabelle enthält eine Übersicht.

	bekannte Notation	"TINV"-Notation	Beispiel (Graphik)
zweischwänzig	P(T < -Q_α) + P(T > Q_α)	TINV(2 α ; Freiheitsgrade)	TINV(2 α ; Freiheitsgrade) = 2
einschwänzig ($\alpha < 0{,}5$)	P(T < Q_α)	-TINV(2 α; Freiheitsgrade)	-TINV (2 α; Freiheitsgrade) = -2
einschwänzig ($\alpha > 0{,}5$)	P(T < Q_α)	TINV((1-α) 2; Freiheitsgrade)	TINV(1 - α) 2; Freiheitsgrade) = 2

- α :
- σ :
- μ_0 :
- Quantil:

Berechnen Sie mit den vorhandenen Informationen den Annahmebereich und schließen Sie daraus den / die kritischen Bereich(e). Bestimmen Sie anschließend den Wert der Teststatistik für μ. Benutzen Sie - wo möglich - Verweise.

- Annahmebereich :
- Kritischer Bereich :
- Wert der Teststatistik :

Tragen Sie ein, ob die Hypothese angenommen werden kann oder abgelehnt werden muss. Benutzen Sie dazu wieder die logischen Funktionen (vgl. Aufgabe b1).

- Entscheidung :

Sie können sich die Verteilung der Teststatistik auf Tabellenblatt "Graphen c1" anschauen. Verändern Sie α und betrachten Sie, wie sich die Kurve ändert.

-

2. Bis jetzt haben Sie die Tests "zu Fuß" durchgeführt. Das hat den Vorteil, dass man die einzelnen Rechenoperationen relativ gut verfolgen kann. EXCEL bietet jedoch auch einige vordefinierte Testfunktionen an, von denen die meisten allerdings schwer nachvollziehbar sind.

Für die Hypothesentests auf μ bei bekannter und unbekannter Varianz gibt es je eine Funktion: "GTEST" und "TTEST". Da die beiden Funktionen relativ ähnlich sind, soll im folgenden nur der "TTEST" genauer untersucht werden, also der Test auf μ bei unbekannter Varianz.

"TTEST" hat vier Argumente: "Matrix1", "Matrix2", "Seiten" und "Typ".

- "Seiten" bestimmt, ob ein ein- oder zweiseitiger Test durchgeführt werden soll.
- In "Matrix1" müssen die Daten der Stichprobe stehen, in "Matrix2" der Wert μ_0 ("Matrix1" und "Matrix2" müssen gleich groß sein, d.h. man muss μ_0 n mal anführen).
- Über "Typ" können unterschiedliche Arten von t-Tests gewählt werden. Für diese Aufgabe ist nur der gepaarte t-Test entscheidend. Dieser wird durch eine "1" symbolisiert.

Interpretation des Ergebnisses:

"TTEST" berechnet zunächst die Prüfgröße $\bar{x}$ und standardisiert diese:

$$t = \frac{\bar{x} - \mu_0}{\bar{s}_n} \sqrt{n}$$

- **"einseitig":** Dann wird eine sogenannte Übertretungswahrscheinlichkeit berechnet und ausgegeben. Für $t > 0$ ist das die Wahrscheinlichkeit dafür, dass der wahre t-Wert zwischen t und ∞ liegt, für $t < 0$ die Wahrscheinlichkeit, dass er zwischen t und $-\infty$ liegt. Ist dieser Wert größer als die Irrtumswahrscheinlichkeit α, kann die Nullhypothese nicht abgelehnt werden. Bei diesem Vorgehen wird also kein Annahmebereich berrechnet, sondern anhand der Überschreitungswahrscheinlichkeit entschieden, ob die Prüfgröße innerhalb des Annahmebereiches liegt.
- **"zweiseitig":** Auch beim zweiseitigen Test wird eine Überschreitungswahrscheinlichkeit ausgegeben. Das ist $p(-\infty; -t) + P(t; \infty)$, also die Wahrscheinlichkeit dafür, dass der wahre t-Wert größer als t bzw. kleiner als $-t$ ist. Der Wert den diese Funktion für den zweiseitigen Test berechnet, ist immer doppelt so groß wie der für den einseitigen Test (Symmetrie der t-Verteilung). Auch hier kann die Nullhypothese nicht abgelehnt werden, wenn die Übertretungswahrscheinlichkeit$> \alpha$ ist.

Berechnen Sie die standardisierte Teststatistik für μ (entspricht t) und überprüfen Sie das Ergebnis von "TTEST", indem Sie mit Hilfe der EXCEL-Funktion "TVERT" die Wahrscheinlichkeiten $P(T > t)$ und $P(T < -t) + P(T > t)$ berechnen (beachten Sie die Hilfe zu "TVERT").

Außer den Testfunktionen gibt es noch einige Analysefunktionen, die auch zu Tests verwendet werden können: Im folgenden sollen Sie die Analysefunktion "Zweistichproben t- Test bei abhängigen Stichproben" genauer betrachten.

Führen Sie für das vorliegende Beispiel einen "Zweistichproben t-Test bei abhängigen Stichproben" durch (diesen Test finden Sie unter den Analysefunktionen).

Als Variable A bzw. B geben Sie bitte die Werte der Stichprobe bzw. den gleichgroßen Vektor mit μ_0's ein (wie bei "TTEST"). Die hypothetische Differenz der Mittelwerte ist "0". Wählen Sie "Alpha" gleich "0,1" und geben Sie als Ausgabebereich die Zelle "J3" an.

Als Ausgabe erhalten Sie eine Reihe von unterschiedlichen Kenngrößen: Zusätzlich zu Mittelwert und Stichprobenvarianz wird Ihnen der Wert der t-Statistik ausgegeben sowie die Überschreitungswahrscheinlichkeiten, die Sie schon von "TTEST" kennen. Darüber hinaus werden Ihnen noch zwei weitere Werte gegeben: "Kritischer t- Wert bei einseitigem t-Test" und "Kritischer t-Wert bei zweiseitigem t-Test". Diese beiden Werte geben das Quantil für $1-\alpha$ bzw. $1-\alpha/2$ an. Berechnen Sie zum Vergleich die vier Quantile Q_α, $Q_{1-\alpha}$,$Q_{\alpha/2}$ und $Q_{1-\alpha/2}$ der t-Verteilung (beachten Sie die Erklärung zu "TINV" unter c1).

Transformieren Sie das Quantil $Q_{1-\alpha}$ mit μ_0 und der korrigierten Stichprobenvarianz und vergleichen Sie das Ergebnis mit den Grenzen des Annahmebereichs aus Aufgabe c1. Was fällt Ihnen dabei auf? Warum kommt es zwangsläufig zu diesem Ergebnis?

-
-

Egal wie man das Testergebnis berechnet - ob mit Überschreitungswahrscheinlichkeit oder mit Hilfe eines kritischen Bereiches - man kommt zu der gleichen Entscheidung. Wo liegen Ihrer Meinung nach die Vorzüge oder Nachteile der jeweiligen Methode?

-
-

d) Herr Müller hat Probleme mit seiner Jogurt-Abfüllanlage, deshalb möchte er sich eine neue kaufen. Der Hersteller verspricht, dass er eine Abfüllanlage mit einer Varianz von höchstens $25g^2$ liefern kann. Herr Müller glaubt ihm nicht.

Der Hersteller will ihn mit einem Hypothesentest davon überzeugen, dass seine Bedenken unbegründet sind. Man einigt sich auf eine Fehlerwahrscheinlichkeit von $\alpha = 0{,}1$.

(Tipp: Überlegen Sie sich anhand Ihrer Ergebnisse aus Aufgabe a), welche Teststatistik Sie nehmen müsen und welcher Verteilung diese gehorcht)

1. Nehmen Sie an, ein Mittelwert von 200 g sei bekannt. Welchen Test muss der Hersteller durchführen? Formulieren Sie die Null- und die Gegenhypothese. Geben Sie anschließend die Freiheitsgrade und das $(1-\alpha)$-Quantil an (Beachten Sie die Hilfe dazu).

   - H_0 :
   - H_1 :
   - Freiheitsgrade :
   - Q_α :

   Berechnen Sie als nächstes den Annahmebereich und den kritischen Bereich sowie die Teststatistik (die Formel muss leider "zu Fuß" eingegeben werden). Tragen Sie in den Zellen "G21" und "G23" wie in den vorangegangenen Aufgaben die richtigen logischen Verknüpfungen ein.

- Annahmebereich :
- Kritischer Bereich :
- Wert der Teststatistik :
- Entscheidung :

Wie ändert sich die Entscheidung wenn man α verändert?

-

2. Führen Sie den gleichen Test bei unbekanntem μ durch. Für die Teststatistik können Sie jetzt die Funktion "VARIANZ" benutzen, die die korrigierte Stichprobenvarianz berechnet.

   - H_0 :
   - H_1 :
   - Freiheitsgrade :
   - Q_α :
   - Wert der Teststatistik :
   - Annahmebereich :
   - Kritischer Bereich :
   - Entscheidung :

   Wie unterscheiden sich die Annahmebereiche aus 1. und 2.? Wie ändert sich die Teststatistik und warum? Ändert sich die Entscheidung dadurch?

   -
   -
   -