Tag 6: Statistik

Version vom 26.8.2016

A) ZUFALLSZAHLEN UND WAHRSCHEINLICHKEITSVERTEILUNGEN

Hintergrund:

Mit Hilfe von experimentellen Messungen versucht man, allgemeingültige Aussagen zu treffen und Regeln für untersuchte Zusammenhänge aufzustellen. Man variiert einen Parameter (z.B. Menge an Dünger) und beobachtet den dadurch hervorgerufenen Effekt auf eine Messgröße. Dies wäre sehr einfach, wenn grundsätzlich jede Beobachtung immer gleich ausfallen würde, wenn man sie mehrfach wiederholt. In der Realität ist dies allerdings nicht der Fall: Messdaten hängen grundsätzlich zumindest in einem bestimmten Rahmen vom Zufall ab, denn in einem Experiment können niemals alle Zufallsfaktoren ausgeschlossen werden. (z.B. könnte es bei einer Studie über die Wirksamkeit eines Medikaments einen Einfluss haben, wie viel die Patienten geraucht haben oder ob sie gerade Stress hatten.) Auch wenn es eine eindeutige Abhängigkeit zwischen dem variierten Parameter und der gemessenen Größe gibt, werden die Messwerte unterschiedlich ausfallen, sie streuen um den erwarteten Wert.

Wahrscheinlichkeitsverteilung:

Wiederholt man ein Experiment in genau gleicher Weise sehr häufig, ergibt sich eine Häufigkeitsverteilung. Diese gibt an, wie häufig ein bestimmter Messwert beobachtet wurde und dient dazu, die Wahrscheinlichkeit dieses Messwertes abzuschätzen.

Für Messdaten (auch künstlich vom Computer erzeugte Zufallszahlen) werden Häufigkeitsverteilungen empirisch ermittelt, um dadurch auf die zugrundeliegende Wahrscheinlichkeitsfunktion zu schließen. Dafür benutzt man Histogramme. Diese teilen den gesamten Wertebereich der Variablen in mehrere Bereiche auf. Das Histogramm gibt für jeden der Bereiche an, wie häufig der Wert der Variable in einer Messung innerhalb des jeweiligen Bereichs lag. Dazu wird für jeden Teilbereich (sogenannte Klassen) ein Rechteck dargestellt, dessen Fläche die gemessene Häufigkeit repräsentiert. In Matlab werden Histogramme mit dem Befehl *hist* erzeugt. Dieser kann vielfältig eingesetzt werden:

hist(v)	Teilt den Wertebereich des Vektors *v* in 10 gleich große Klassen ein. Wenn *hist* ohne Ausgabeargument aufgerufen wird, stellt es die Häufigkeit des Auftretens der Klassen als Balkengrafik dar.
h=hist(v)	Wenn *hist* mit einem Ausgabeargument aufgerufen wird, produziert es keine grafische Ausgabe, sondern gibt den Vektor der Häufigkeiten zurück. (Kann mit mehreren Eingabeargumenten kombiniert werden.)
hist(v,nbins)	Teilt den Wertebereich des Vektors *v* in *nbins* gleich große Klassen ein. (Mit oder ohne Ausgabeargument verwendbar)
hist(v,centers)	Benutzt den Vektor *centers* als Mittelpunkte der Klassen, in die die Elemente von *v* aufgeteilt werden. Wenn *hist* ohne Ausgabeargument aufgerufen wird, stellt es die Häufigkeit des Auftretens der Klassen als Balkengrafik dar. (Mit oder

ohne Ausgabeargument verwendbar)

[h,xout]=hist(v)

Wenn *hist* mit zwei Ausgabeargumenten aufgerufen wird, produziert es keine grafische Ausgabe, sondern gibt zwei Vektoren zurück: den Vektor *h* der Häufigkeiten und den Vektor *xout* der Klassenmittelpunkte. (Kann mit mehreren Eingabeargumenten kombiniert werden)

Normalverteilung:

Die meisten biologischen Daten lassen sich durch eine Normalverteilung(auch Gauß-Verteilung genannt) beschreiben, bei der Messwerte umso häufiger auftreten, je näher sie am Erwartungswert, dem Mittelwert der Verteilung liegen.

Für eine normalverteilte Zufallsvariable x entspricht die Wahrscheinlichkeitsdichte folgender Formel:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

wobei μ den Mittelwert und σ die Standardabweichung der Wahrscheinlichkeitsverteilung angibt.

Für normalverteilte Messwerte (oder auch mit einem Zufallsgenerator erzeugte Zufallszahlen) kennt man diese Kennwerte der den Daten zugrunde liegenden Normalverteilung nicht. Man kann sie jedoch aus den Messwerten x_1 bis x_n schätzen:

Empirischer Mittelwert:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Empirische Standardabweichung:

$$s_X := \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Auch wenn es keine schlechte Übung ist, diese Formeln einmal in Matlab umzusetzen, kann man stattdessen auch einfach die Befehle *mean* und *std* benutzen.

Achtung: Die Berechnung von empirischem Mittelwert und empirischer Standardabweichung macht ausschließlich für symmetrisch verteilte (am besten normal verteilte) Werte Sinn!

Eine andere wichtige Verteilung, die in diesem Kurs auch betrachtet wird, ist die Gleichverteilung, bei der alle Werte in einem bestimmten Bereich mit gleicher Wahrscheinlichkeit auftreten. (Für gleich verteilte Werte ist es absolut sinnlos, Mittelwert und Standardabweichung zu berechnen.)

Stichprobengröße:

Bei der Berechnung und Interpretation des empirischen Mittelwerts und der empirischen Standardabweichung ist Vorsicht geboten:

Man kennt nur eine begrenzte Stichprobe, die nicht unbedingt die gesamte Population repräsentieren muss. Je größer diese Stichprobe ist, desto sicherer kann man sich sein, den tatsächlichen Werten der ganzen Population nahe zu kommen.

Um abzuschätzen, wie gut verwendete Stichprobengrößen eine Population charakterisieren, verwendet man den Standardfehler des Mittelwerts (standard error of the mean, SEM). Dieses Maß gibt die Streuung der Mittelwerte von verschiedenen, zufällig aus der Population gezogenen gleich großen Stichproben um den Erwartungswert (den wahren Populationsmittelwert) an. Der Standardfehler der Mittelwerte ist definiert als

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Wobei

- n die Größe der Stichproben angibt (nicht die Anzahl der Stichproben!) und
- σ die Standardabweichung der Verteilung (diese ist normalerweise nicht bekannt und muss aus den Daten empirisch geschätzt werden).

Grafische Darstellung:

Sowohl für die Standardabweichung als auch für den Standardfehler des Mittelwertes ist es üblich, Kurven mit Fehlerbalken zur graphischen Darstellung zu nutzen.

- In Matlab lautet der Befehl, um eine Kurve mit Fehlerbalken zu zeichnen, *errorbar(x,mw,fehler)*.
- Dabei ist

x der Vektor der x-Werte, gegen die Mittelwert und Fehler aufgetragen werden

mw der Vektor der Mittelwerte,

fehler der Vektor der Standardabweichungen / Standardfehler der Mittelwerte

- Die Fehler werden als symmetrische Balken zu beiden Seiten des Mittelwerts aufgetragen. (Falls Sie asymmetrische Fehlerbalken brauchen, schauen Sie in der Hilfe nach).
- Da Fehlerbalken für die Darstellung unterschiedlicher Größen (insbesondere Standardabweichung und Standardfehler, teilweise auch Quartile...) verwendet werden, ist es unbedingt notwendig, in der Abbildungsunterschrift zu schreiben, was die Fehlerbalken bedeuten - und beim Lesen von wissenschaftlichen Veröffentlichungen auf diese Angabe zu achten.

Zufallszahlen:

Bevor wir uns mit der statistischen Auswertung von echten Messdaten beschäftigen, erzeugen wir zunächst einmal selber "Messdaten" mit Matlab, nämlich Zufallszahlen. Diese werden beispielsweise gebraucht, wenn man Experimente plant, in denen bestimmte Reize in zufälliger Reihenfolge präsentiert werden sollen. Eine weitere wichtige Anwendung von Zufallszahlen sind Simulationen biologischer Prozesse. Wenn man Zufallszahlen künstlich erzeugt, ist (im Gegensatz zur Auswertung von Messdaten) die Wahrscheinlichkeitsverteilung bekannt (also die Wahrscheinlichkeit dafür, dass eine Zufallsvariable einen bestimmten Wert annimmt).

Im Rahmen des Kurses erzeugen wir folgende Zufallszahlen:

M1=randn(Z,S)	erzeugt eine ZxS-Matrix mit normalverteilten Zufallszahlen mit Mittelwert 0 und Standardabweichung 1
M2=rand(Z,S)	erzeugt eine ZxS-Matrix mit gleichverteilten Zufallszahlen zwischen 0 und 1
v=randperm(n)	liefert einen Vektor der ganzen Zahlen von *1* bis *n* in zufälliger Reihenfolge.

Aufgaben:

T6A1) Probieren Sie die Funktionen *randn* und rand aus:

- Erzeugen Sie einige Beispiele normalverteilter und gleichverteilter Zufallszahlen: Was passiert, wenn man die gleiche Funktionen mehrfach hintereinander in gleicher Weise aufruft?
- In welchen Bereich liegen die Werte für die beiden Funktionen?

T6A2) Erzeugen Sie jeweils einen sehr langen Vektor (z.B. 10000 Elemente) mit jeder der beiden Funktionen *rand* und *randn*.

- Schauen Sie sich die jeweilige Verteilung der Zufallszahlen mit dem Befehl *hist* an.
- Was sind die Unterschiede zwischen den beiden Verteilungen?
- Mit *hist(v,n)* teilt *hist* den Vektor *v* in *n* gleich große Bereiche ein. Sehen Sie sich die Verteilungen für verschiedene Werte von *n* an.
- Schätzen Sie aus der Abbildung ab: Was ist der Mittelwert? Was die Standardabweichung?
- Berechnen Sie Mittelwerte, Standardabweichungen, Varianzen, Minima und Maxima Ihrer beiden Vektoren mit *mean, std, var, min* und *max*.

T6A3) Modifizieren Sie Ihre Zufallsvektoren, indem Sie diese

- mit verschiedenen Faktoren multiplizieren
- verschiedene Zahlen hinzuaddieren
- Wie wirken sich diese Änderungen auf die Verteilungen aus?
- Wie wirken sie sich auf Mittelwert, Standardabweichung, Minimum und Maximum aus?

T6A4) Laden Sie den Vektor mit den Anzahlen an Sonnenblumenkernen von 100 Blumen [sbkerne.mat]. Berechnen Sie Mittelwert, Varianz und Standardabweichung. Wie habe ich diesen Vektor erzeugt? (Nein, ich habe mich nicht in den Garten gesetzt und gezählt...)

*T6A5) Setzen Sie die oben angegebenen Formel für die Wahrscheinlichkeitsdichte einer Normalverteilung in Matlab um.

- Schreiben Sie eine Funktion, die drei Eingabeargumente bekommt
  - einen Vektor, der den Definitionsbereich angibt, z.B. *x=-4:0.01:4*,
  - den gewünschten Mittelwert und
  - die gewünschte Standardabweichung
- Als Ausgabeargument soll die Funktion die berechnete Wahrscheinlichkeitsdichteverteilung als Vektor zurück geben
- Die Funktion soll die Wahrscheinlichkeitsdichte außerdem als Kurve grafisch darstellen (bitte mit Titel und beschrifteten Achsen).
- Variieren Sie die Parameter Mittelwert und Standardabweichung. Wie verändern diese die Kurve?

T6A6) Schreiben Sie eine Funktion *wuerfel*, die jeweils eine ganze Zahl zwischen 1 und 6 zurückgibt.

T6A7) Benutzen Sie diese Funktion in einer weiteren Funktion *wuerfel_verteilung*, die als Eingabewert bekommt, wie oft gewürfelt wird, und als Ausgabe die Verteilung (als in einem Vektor gespeichertes Histogramm) der erzielten Würfelergebnisse zurückliefert.

T6A8) Sie haben die Aufgabe, das Fressverhalten von Mäusen zu charakterisieren, wobei die Mäuse ausschließlich mit genormten Futterpellets gefüttert werden, deren gefressene Anzahl jeden Tag notiert wird. Schreiben Sie eine Funktion, die diese Datenerhebung für eine Maus simuliert:

- Die Funktion bekommt als Eingabeargument die Anzahl der zu simulierenden Tage übergeben.
- Die Funktion erzeugt für jeden Tag eine Zufallszahl, die die Anzahl der von der Maus gefressenen Pellets darstellen soll.
- Dabei soll der Mittelwert der pro Tag gefressenen Pellets 30 betragen und die Standardabweichung 5.
- Die Funktion gibt den Vektor der gefressenen Pellets zurück.

T6A9) Benutzen Sie die Funktion aus T6A8 in einer weiteren Funktion:

- Als Eingabeargument bekommt die Funktion die Anzahl betrachteter Tage (Stichprobenöße).
- Ausgaben sind die errechneten Werte für Mittelwert, Standardabweichung, und Standardfehler des Mittelwerts.
- Außerdem soll diese Funktion die Verteilung der Werte grafisch als Histogramm darstellen.
- Lassen Sie diese Funktion für verschiedene Stichprobengrößen (also Anzahlen der Tage) laufen, z.B. *N=1; N=3; N=5; N=10; N=20; N=50; N=100; N=1000*. (Dafür könnten Sie ein Skript schreiben, das die Funktion mit den jeweiligen Stichprobengrößen aufruft.)

- Wie wirkt sich die Stichprobengröße auf Mittelwert, Standardabweichung, Standardfehler des Mittelwertes und Histogramm aus?

*T6A10) Programmieren Sie eine Funktion, die für Sie eine ganze Messreihe des Mäusefressverhaltens steuert.

- Die Funktion bekommt als Eingabeargumente
  - die Anzahl der Mäuse, die pro Tag beobachtet werden sollen (Anzahl Stichproben) und
  - die Anzahl der Tagen, an denen die gefressenen Pellets gezählt werden sollen (Stichprobengröße).
- Rückgabewert ist der Standardfehler des Mittelwerts.
- Außerdem zeigt sie die Verteilung der erzielten Mittelwerte als Histogramm grafisch an.
- Probieren Sie diese Funktion für verschiedene Kombinationen aus Stichprobengröße und Anzahl der Stichproben aus, z.B. 3 Tage mit 3 Tieren, 10 Tage mit 3 Tieren, 3 Tage mit 10 Tieren, 10 Tage mit 10 Tieren, 10 Tage mit 100 Tieren, 100 Tage mit 10 Tieren, 100 Tage mit 100 Tieren, 1000 Tage mit 10 Tieren, 10 Tage mit 1000 Tieren.
- Wie wirken sich die beiden Parameter auf den Standardfehler des Mittelwerts aus?
- Wie wirken sie sich auf die Verteilung der Mittelwerte aus?

T6A11) Die vorige Aufgabe war insofern unrealistisch, als alle Tiere statistisch gleich viel Hunger hatten. Natürlich gibt es aber bei echten Tieren individuelle Unterschiede. In folgender Matrix sind die Messungen von 30 Tieren an 100 Tagen dargestellt, wobei die Werte eines Tieres jeweils in der gleichen Zeile stehen: [Maeusepellets.mat]

- Schreiben Sie ein Skript, das die Mittelwerte und Standardabweichungen einerseits zwischen den Tagen, andererseits zwischen den Tieren berechnet.
- Stellen Sie die beiden Verläufe von Mittelwert und Standardabweichung in zwei Abbildungen mit Fehlerbalken grafisch dar. (Vergessen Sie nicht die Beschriftungen, damit Sie sich später beim Vergleich der Abbildungen zurecht finden.)
- Inwiefern unterscheiden sich die Ergebnisse für die beiden Arten, Mittelwerte und Standardabweichungen zu berechnen (zwischen Tagen vs. zwischen Tieren)?
- Berechnen Sie für beide Wege den resultierenden Standardfehler des Mittelwerts und stellen Sie auch diesen in gesonderten Abbildungen mit Fehlerbalken dar.
- Wie unterscheiden sich die Abbildungen? Welche Aussagen kann man jeweils daraus ableiten?

B) MEDIAN UND QUANTILE

Zwar gibt es viele Datensätze, die sich gut durch Normalverteilungen erklären lassen. Aber bei manchen Datensätzen ist diese Bedingung eben doch nicht erfüllt, sondern man misst "schiefe" Verteilungen. Das kommt insbesondere dann zustande, wenn es im Datensatz Ausreißer gibt (also besonders große oder besonders kleine Werte, s.u.). Diese verfälschen den Mittelwert. Deshalb ist es in diesen Fällen häufig ratsamer, statt des Mittelwertes den Median zu berechnen, um

den "typischen" Messwert zu betrachten. Der Median gibt denjenigen Wert an, bei dem die Hälfte der Messwerte kleiner und die andere Hälfte größer ist, unabhängig davon, wie groß oder klein die Werte sind.

Diese Sortierung der Daten nach Größe und anschließende Unterteilung in Klassen mit gleich vielen Datenpunkten nennt man Quantile. Neben dem Median (Aufteilung in 50%-Stücke) spielen insbesondere die Quartile (Aufteilung in 25%-Stücke) und Perzentile (Aufteilung in 1%-Stücke) eine Rolle. Beispielsweise sind das 3% und das 97%-Perzentil übliche Größen für die Auswertung, um zu entscheiden, ob ein Messwert "normal" oder "auffällig" ist.

Eine übliche graphische Darstellung der Datenauswertung basierend auf Medianen und Quantilen ist der Boxplot. Dieser enthält folgende Angaben:

- für jeden gegebenen x-Wert wird der Bereich vom 25% bis zum 75% Perzentil der y-Werte als ein Kasten dargestellt
- innerhalb dieses Kastens wird der Median mit einer weiteren Linie markiert.
- Balken nach oben und unten (im Englischen bezeichnet als "whiskers", also Fühler) geben den Bereich an, in dem die restlichen Datenpunkte liegen, die nicht als Ausreißer zu betrachten sind.
- Im Boxplot werden Ausreißer als einzelne Datenpunkte ober- bzw unterhalb der "whiskers" eingezeichnet. Daten werden als Ausreißer betrachtet, wenn sie größer als $q_{75}+1.5*(q_{75}-q_{25})$ oder kleiner als $q_{25}-1.5*(q_{75}-q_{25})$ sind, wobei q_{25} das 25% und q_{75} das 75% Perzentil bezeichnen. Für normalverteilte Daten entspricht das einem Wert von etwa +/-2.7*Standardabweichung, was etwa 99.3% aller Daten entspricht.

Matlab:

mx=median(x)	berechnet den Median des Vektors x.
mM=median(M)	berechnet für jede Spalte der Matrix M den Median.
mM2=median(M,2)	berechnet für jede Zeile der Matrix M den Median.
Z=prctile(x,p)	berechnet für den Datenvektor *x* (bzw für jede Spalte der Matrix *x*) das *p*-te Perzentil. Allerdings ist diese Funktion nicht im Standardumfang von Matlab enthalten, sondern in der Statistics Toolbox. (Sie lässt sich aber einfach selber schreiben, s.u.)
boxplot(X)	erzeugt einen Boxplot. Wenn *X* eine Matrix ist, wird für jede Spalte Median, Perzentile und Ausreißer berechnet und als eigene "box" aufgetragen.

Aufgaben:

T6B1) Schiefe Verteilungen sieht man oft bei der Messung von Reaktionszeiten. Sehen Sie sich die Verteilung der Reaktionszeiten (in ms)dieser Versuchsperson an: [rt_VP5.mat] (Hinweis: wenn Sie ein Histogramm mit vielen Klassen nehmen, sehen Sie mehr!) Warum ist das keine Normalverteilung?

T6B2) Berechnen Sie für den gleichen Datensatz den Mittelwert und den Median der Reaktionszeiten. Warum unterscheiden sich diese so stark?

T6B3) Erstellen Sie für die Daten einen Boxplot. Dieser wird Ihnen zeigen, dass es einen einzelnen extrem großen Wert gibt. Löschen Sie diesen aus dem

Datensatz und vergleichen Sie noch einmal Mittelwert und Median.

T6B4) Wiederholen Sie die Betrachtung von Verteilung, Boxplot, Mittelwert und Median noch einmal für den gesamten Datensatz [rt_all.mat], bei dem jeweils 180 Reaktionszeiten von 24 Versuchspersonen gemessen wurden. Betrachten Sie dabei zunächst den gesamten Datensatz gemeinsam, ohne auf individuelle Unterschiede zwischen den Versuchspersonen zu achten.

T6B5) Machen Sie eine Statistik darüber, wie stark sich Mittelwerte und Mediane für die Versuchspersonen unterscheiden. Sollte man hier mitteln?

T6B6) Wie stark unterscheiden sich Mittelwerte und Mediane bei dem Beispiel der Sonnenblumenkerne [sbkerne.mat] ?

T6B7) Schreiben Sie eine Funktion *perzentil*, die als Eingabeparameter einen Datenvektor und eine Zahl N bekommt und den Wert des N%-Perzentils des Datenvektors zurückgibt.

C) SIGNIFIKANZTEST

Sehr häufig ist bei der Auswertung biologischer Daten nach Signifikanz gefragt. Wir haben im Kurs leider keine Zeit dafür umfangreich auf Signifikanztests und ihren mathematischen Hintergrund einzugehen. Wir werden aber mit zwei einfachen Beispielen die Anwendung von Signifikanztests in Matlab ausprobieren. Signifikanztests sind nicht im Standard-Programmumfang von Matlab enthalten, sondern finden sich in der Toolbox "statistics" (die hoffentlich auf allen Rechnern im Raum installiert sein sollte).

Achtung: Die Anwendung von Signifikanztests macht nur Sinn, wenn die Stichprobe groß genug ist! (Wikipedia gibt n>30 als Faustregel an, wenn die Daten nicht unbedingt einer Normalverteilung entstammen - aber es kommt auch auf die Standardabweichung der Verteilung an, wie groß die Stichprobe sein muss, um sinnvolle Ergebnisse zu liefern.)

Das erste Beispiel ist der t-Test für den Erwartungswert einer normalverteilten Stichprobe.

- Bei diesem Test ist die Nullhypothese, dass eine Menge von n Messwerten (unabhängige, normal verteilte Zufallsvariablen) einer Verteilung mit einem gegebenen Mittelwert μ_0 und unbekannter Varianz entstammt, also dass $\mu_0 = \mu$.
- Dafür wird mit dem empirischen Stichprobenmittelwert und der empirischen Stichprobenstandardabweichung s (siehe oben, dort als s_x bezeichnet) die Testprüfgröße t berechnet:

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

- Wie man sieht, geht hier (ebenso wie beim Standardfehler des Mittelwerts) die Stichprobengröße n zusätzlich zur empirischen Standardabweichung s und dem Abstand zwischen gemessenem und zu testendem Mittelwert ein. Je größer die Stichprobe, je größer der Abstand zwischen den Mittelwerten und je kleiner die Standardabweichung ist, desto größer ist der Betrag der Testprüfgröße t.
- Die Nullhypothese $\mu_0 = \mu$ wird zum Signifikanzniveau α abgelehnt wenn

$$|t| > t(1 - \frac{\alpha}{2}, n - 1)$$

- also der Betrag von *t* größer als das *(1-a/2)*-Quantil der *t*-Verteilung mit *n−1* Freiheitsgraden ist (diese Werte sind normalerweise in Tabellen abgelegt und Matlab natürlich bekannt).
- Wenn die Nullhypothese zum Beispiel zum Signifikanznivieau 5% abgelehnt wird, bedeutet das, dass die Messwerte mit 95% Wahrscheinlichkeit nicht einer Normalverteilung mit dem Mittelwert μ_0 entstammen, also die Werte wirklich verschieden sind.
- In 5% der Fälle kann der signifikante Unterschied aber durch Zufall innerhalb der Verteilung der Nullhypothese zustande gekommen sein.
- Wenn die Nullhypothese nicht abgelehnt wird, ist es nicht zulässig daraus zu schliessen, dass die Messwerte der zu testenden Verteilung entstammen!

Das zweite Beispiel ist ein t-Test für zwei unabhängige Stichproben. Bei diesem lautet die Nullhypothese, dass zwei Stichproben *x* und *y* zwei Normalverteilungen mit gleichem Mittelwert entstammen, also H_0: $\mu x = \mu y$. Hierzu wird mit den empirischen Stichprobenvarianzen und Stichprobenmittelwerten die sogenannte gewichtete Varianz

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

bestimmt, um damit die Prüfgröße

$$t = \sqrt{\frac{nm}{n+m}} \frac{\bar{x} - \bar{y}}{s}$$

zu berechnen. Mittels der Ungleichung

$$|t| > t(1 - \frac{\alpha}{2},\ n+m-2)$$

wird überprüft, ob die Nullhypothese zum Signifikanzniveau *a* abgelehnt werden kann und somit von einem signifikanten Unterschied der beiden Stichproben ausgegangen werden kann.

Matlab:

h=ttest(vektor,mittelwert)	testet, ob die Nullhypothese abgelehnt werden kann, dass die im Vektor *vektor* gespeicherten Messdaten einer Normalverteilung mit dem Mittelwert *mittelwert* entstammen. Das Standard-Signifikanznivieau ist 5%.
h=ttest(vektor,mittelwert, alpha)	wie oben, aber mit Angabe des Signifikanznivieaus *alpha*
h=ttest2(vektor1,vektor2)	testet, ob für zwei Stichproben *vektor1* und *vektor2* zum Standard-Signifikanzniveau 5% die Nullhypothese abgelehnt werden kann, dass beide Stichproben der gleichen Verteilung

	entstammen.
h=ttest2(vektor1,vektor2, alpha)	wie oben, aber mit Angabe des Signifikanznivieaus *alpha*

Für alle *ttest*-Funktionen gilt: Der Rückgabewert ist

- *1* wenn die Nullhypothese abgelehnt wird (also wenn der erwartete und der empirische Mittelwert mit 100-α% Wahrscheinlichkeit verschieden sind).
- *0* wenn die Nullhypothese nicht abgelehnt werden kann.

Eine gute Hilfe für die Einschätzung von Signifikanz bietet wiederum der Boxplot. Diesen kann man in Matlab mit der Option 'notch' aufrufen, so dass eine Einkerbung der Box das 5% Konfidenzintervall darstellt (unter der Annahme, dass es sich um normalverteilte Daten handelt). Überlappen sich die Einkerbungen zweier Boxen, sind die Daten wahrscheinlich nicht auf einem 5% Niveau signifikant verschieden.

- *boxplot(matrix,'notch','on')*

Aufgaben:

T6C1) Ein superschlauer Futtermittelhersteller behauptet, dass eine Maus im Durchschnitt 31 Futterpellets am Tag frisst.

- Überprüfen Sie diese Aussage für ein Signifikanzniveau von 5% für Ihre gesamte Mäusepopulation anhand der Messdaten [Mauesepellets.mat]
- Wie sieht es bei einem Signifikanzniveau von 10% aus?
- *) Für wie viele Mäuse kann die Aussage des Futtermittelherstellers nicht abgelehnt werden?
- Sehen Sie sich die Daten im Boxplot an. Bestätigen sich dort Ihre Testergebnisse?

T6C2) Untersuchen Sie für den gleichen Datensatz:

- Haben die ersten beiden Mäuse signifikant unterschiedliche Mengen gefressen?
- Sind am ersten und am fünften Tag von der gesamten Mäusegruppe signifikant unterschiedlich viele Pellets gefressen worden?
- *) Bei wie vielen Mäusepaaren gibt es einen signifikanten Unterschied der durchschnittlich gefressenen Pellets?
- *) Bei wie vielen Paaren von Tagen gibt es einen signifikanten Unterschied der von allen Mäusen durchschnittlich gefressenen Pellets?
- *) Gibt es Paare von Mäusen oder von Tagen, die hoch signifikant (alpha=0.01) verschiedene Futtermengen aufweisen?

HAUSAUFGABEN:

T6H1) 4 Punkte Im Programm [vogelfang.m] werden drei verschiedene Arten von Zufallszahlen benutzt. Vollziehen Sie dieses Programm nach. Nehmen Sie schrittweise folgende Änderungen vor: a) Bei Amseln gibt es 60% Weibchen. b) Bei Spatzen streut das Gewicht von Weibchen 3 Mal mehr als das Gewicht von

Männchen. c) Es kommen 25% Meisen und 25% Spatzen in der Gegend vor.

T6H2) Generieren Sie mit Ihrer auf [vogelfang.m] aufbauenden Lösung von T4H7 (oder wahlweise der Musterlösung [vogelmatrix.m]) 10 Vogelfang-Matrizen und berechnen Sie für jede Kombination von Art und Geschlecht jeweils Mittelwerte und Standardabweichungen des Gewichts, sowie den Standardfehler Ihrer Gewichtsmessungen.

- Ist das Gewicht der Arten signifikant verschieden?
- Ist das Gewicht der Geschlechter einer der Arten signifikant verschieden?

T6H3) 6 Punkte In einem psychophysikalischen Experiment sollen einem Versuchstier drei verschiedene Töne in zufälliger Reihenfolge vorgespielt werden, aber jeder Ton soll genau 5 Mal vorkommen.

- Wir kümmern uns erstmal nicht um die Generierung der Töne, sondern nennen sie einfach Bedingung 1, 2 und 3.
- Überlegen Sie sich einen Algorithmus, der die Reizbedingungen in die richtige Reihenfolge bringt und setzen Sie diesen in ein Programm um.
- Testen Sie das Programm, indem Sie es mehrfach laufen lassen. Macht es immer, was es soll? Sind die Ergebnisse jedes Mal gleich?
- Erweitern Sie Ihr Programm so, dass es *N* (eine beliebige Anzahl) Reize, die *M* mal (also beliebig oft) vorgespielt werden sollen, in eine Reihenfolge bringt.
- Tipp: Benutzen Sie für diese Aufgabe die Funktion *repmat*. Diese erzeugt eine große Matrix durch mehrfache Wiederholung einer kleineren. Z.B. *B = repmat(A,2,5)* erzeugt eine Matrix *B*, in der insgesamt 10 Kopien der Matrix *A* enthalten sind, wobei *A* zweimal untereinander und fünfmal nebeneinander angeordnet wird. (*B* hat also die doppelte Zeilen- und fünffache Spaltenzahl von *A*.)

*T6H4) Die Messwerte einer Apparatur ist selbst ohne biologisches Präparat nicht perfekt rauschfrei. Um das Geräterauschen abzuschätzen, wurden im Elektrophysiologie-Praktikum für die Apparatur mit einer Modellzelle (einem elektronischen Schaltkreis, der die Membraneigenschaften einer Nervenzelle nachbaut) 100 Messungen mit dem gleichen
Reiz [stimulus1khz.mat] durchgeführt und die Antworten als Matrix unter[antworten1khz.mat] abgespeichert.

- Schauen Sie sich eine beliebige einzelne Messung zusammen mit dem Reiz an (entsprechend Aufgabe T5C3).
- Berechnen und plotten Sie in ein neues Grafikfenster den Zeitverlauf der über die 100 Messungen gemittelten Antwort.
- Berechnen und plotten Sie in ein neues Grafikfenster den Mittelwert und die Standardabweichung der jeweils letzten 300ms für jede Messung (Mittelung über die Zeit). Gibt es eine Tendenz? Gibt es Ausreißer?
- Berechnen und geben Sie als Textausgabe im Kommandofenster aus: Sind Mittelwert und / oder Standardabweichung vor, während und nach der Reizung unterschiedlich?

*T6H5) (Für mathematisch Interessierte) Häufig sehen Messdaten zunächst recht kompliziert verteilt aus. Bei genauerer Untersuchung stellt sich dann manchmal heraus, dass sie aus zwei überlappenden Verteilungen stammen. Beispielsweise überlappen sich die Verteilungen der Körpergrößen von Männern und Frauen

(denn es gibt Frauen, die größer sind als viele Männer).

Stellen Sie sich vor, Sie bekommen die Aufgabe, aus der Körpergröße auf das Geschlecht zurückzuschließen und kennen die Verteilungen der Körpergrößen. Für solche Aufgaben wird oft das Prinzip "Maximum Likelihood" verwendet: Tippe auf die Verteilung mit der höheren Wahrscheinlichkeit für den gegebenen Wert. Mit dieser Idee lässt sich ein Schwellwert bestimmen, unterhalb dessen man auf die Verteilung mit dem kleineren Mittelwert tippen sollte. Dieser Schwellwert ist der Schnittwert der Verteilungen.

- Erzeugen Sie sich zwei Zufallszahlen, die verschiedenen Normalverteilungen entstammen, die eine mit Mittelwert 5 und Standardabweichung 2, die andere mit Mittelwert 3 und Standardabweichung 1.
- Berechnen Sie mit der gestern eingeführten Formel der Wahrscheinlichkeitsdichte für jede der beiden Zufallszahlen die Wahrscheinlichkeiten, dass sie der einen oder der anderen Verteilung entstammten.
- Erweitern Sie dieses Programm für zwei Vektoren aus Zufallszahlen aus den oben genannten Verteilungen.
- Berechnen Sie den Anteil der Zufallszahlen, die nach dem Maximum Likelihood Prinzip der falschen Verteilung zugeordnet würden.
- Schauen Sie sich die beiden Verteilungen grafisch an. Wo sollte man die Grenze ziehen?
- Variieren Sie Mittelwerte und Standardabweichungen der beiden Verteilungen. Wann gibt es mehr und wann weniger Fehler?

**T6H6) Bonuspunkte möglich (Für mathematisch Interessierte) Erweitern Sie die letzte Aufgabe zu einer Funktion, die für die Angabe von zwei Mittelwerten und zwei Standardabweichungen ausgibt, bei welchem Wert man die Grenze ziehen sollte, um die Verteilungen optimal zu trennen.