**Tag 7: Beschreibende Statistik**

**A) Mittelwert und Standardabweichung**

**B) Standardfehler des Mittelwerts**

**C) Median und Quantile**

**D) Hausaufgabe**

**Z) Zusatzthema: Signifikanztest**

**A) Mittelwert und Standardabweichung**

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 Experimente 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.) Selbst wenn man davon ausgeht, dass es eine eindeutige Abhängigkeit zwischen dem variierten Parameter und der untersuchten Größe gibt, werden die Messwerte unterschiedlich ausfallen, sie streuen um den erwarteten Wert.

Schaut man sich sehr viele Messwerte für ein in immer gleicher Weise wiederholtes Experiment an, ergibt sich eine Häufigkeitsverteilung. Diese gibt an, wie häufig ein bestimmter Messwert beobachtet wurde und dient dazu, die Wahrscheinlichkeit dieses Messwertes abzuschätzen. 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. Eine andere wichtige Verteilung ist die Gleichverteilung, bei der alle Werte in einem bestimmten Bereich mit gleicher Wahrscheinlichkeit auftreten.

Für eine normal verteilte Zufallsvariable x entspricht die Wahrscheinlichkeitsdichte folgender Formel:



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

Die wichtigsten Formeln der beschreibenden Statistik sind für normal verteilte Messwerte der Mittelwert und die Standardabweichung.

Für n normal verteilte Messwerte x1 bis xn ermittelt man diese empirischen folgendermaßen:

Empirischer Mittelwert:



empirische Standardabweichung:



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

T7A1) Plotten Sie wie gestern noch einmal Verteilungen vieler mit rand bzw randn erzeugter Zufallszahlen und versuchen Sie aus der Grafik abzuschätzen:

Was ist der Mittelwert? Was ist die Standardabweichung?

T7A2) Berechnen Sie Mittelwerte, Standardabweichungen, Varianzen, Minima und Maxima Ihrer beiden Vektoren mit mean, std, var, min und max.

Wie wirkt es sich auf diese Werte aus, wenn Sie eine feste Zahl zu den jeweiligen Vektoren addieren?

Wie wirkt sich eine Multiplikation mit einem Faktor aus?

T7A3) 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...)

\*T7A4) Setzen Sie die oben angegebenen Formel für die Wahrscheinlichkeitsdichte einer Normalverteilung in Matlab um. Schreiben Sie eine Funktion, die für die Eingabeargumente x (ein Vektor, der den Definitionsbereich angibt, z.B. x=-4:0.01:4), Mittelwert und Standardabweichung die entsprechende Wahrscheinlichkeitsdichteverteilung als Vektor zurück gibt und als Kurve grafisch darstellt (bitte mit Titel und beschrifteten Achsen).

Spielen Sie etwas mit den Parametern Mittelwert und Standardabweichung. Wie verändern diese die Kurve?

**B) Standardfehler des Mittelwerts**

Die Berechnung von Mittelwert und Standardabweichung ist die wahrscheinlich am häufigsten benutzte Methode der Datenanalyse. Allerdings ist bei ihrer empirischen Berechnung 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

****

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 wie oben beschrieben empirisch geschätzt werden).

T7B1) Sie haben die Aufgabe, das Fressverhalten von Mäusen zu charakterisieren. Dazu füttern Sie die Mäuse ausschließlich mit genormten Futterpellets und zählen täglich, wie viele Pellets aus dem Futterbehälter verschwinden.

Schreiben Sie eine Funktion, die zufällig die Anzahl der von einer Maus gefressenen Pellets ermittelt und zurückgibt. Dabei soll der Mittelwert der gesamten Mäusepopulation 30 Pellets und die Standardabweichung 5 Pellets betragen.

T7B2) Benutzen Sie die Funktion aus T7A1 in einer weiteren Funktion, die als Eingabeargument die Stichprobengröße bekommt und als Ausgabe die ermittelten Werte für Mittelwert und Standardabweichung zurückgibt. Außerdem soll diese Funktion die Verteilung der Werte grafisch als Histogramm darstellen.

Lassen Sie diese Funktion für verschiedene Stichprobengrößen laufen, z.B. N=1; N=3; N=5; N=10; N=20; N=50; N=100; N=1000. Wie wirkt sich die Stichprobengröße auf Mittelwert, Standardabweichung und Histogramm aus?

\*T7B3) Programmieren Sie eine Funktion, die für Sie eine ganze Messreihe des Mäusefressverhaltens steuert. Die Funktion bekommt als Eingabeargumente die jeweilige Stichprobengröße (also wie viele Mäuse pro Tag beobachtet werden sollen) und die Anzahl der Stichproben (also an wie vielen Tagen gezählt werden soll) und liefert als Rückgabewert den 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 Stichproben mit 3 Tieren, 10 Stichproben mit 3 Tieren, 3 Stichproben mit 10 Tieren, 10 Stichproben mit 10 Tieren, 10 Stichproben mit 100 Tieren, 100 Stichproben mit 10 Tieren, 100 Stichproben mit 100 Tieren, 1000 Stichproben mit 10 Tieren, 10 Stichproben mit 1000 Tieren. Wie wirken sich die beiden Parameter auf den Standardfehler des Mittelwerts aus? Wie auf die Verteilung der Mittelwerte?

T7B4) 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 30 Tagen dargestellt, wobei die Werte eines Tieres jeweils in der gleichen Zeile stehen: [pellets.mat]

Schreiben Sie ein Skript, das die Mittelwerte und Standardabweichungen einerseits zwischen den Tagen, andererseits zwischen den Tieren berechnet und diese in zwei Abbildungen grafisch darstellt. Der Befehl, um eine Kurve mit Fehlerbalken zu zeichnen heißt errorbar(x,mw,standabw), wobei mw der Vektor der Mittelwerte ist, der gegen den Vektor x aufgetragen wird und standabw ist der Vektor der Standardabweichungen, die als symmetrische Balken zu beiden Seiten des Mittelwerts aufgetragen werden.

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.

\*)Führen Sie die gleichen Berechnungen noch einmal für folgende Messreihe durch [pellets2.mat], in der ebenfalls die Daten eines Tieres jeweils in einer Zeile stehen. Was sind die Unterschiede?

**C) 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). 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 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 “ganz normal“ oder “auffällig“ ist.

In Matlab gibt es für den Median den Befehl m=median(x). Für allgemeine Perzentile ist die Syntax Z=prctile(x,p) mit Datenvektor x und gewünschter Prozentzahl p. Allerdings ist diese Funktion nicht im Standardumfang von Matlab enthalten, sondern in der Statistics Toolbox. (Sie lässt sich aber einfach selber schreiben, s.u.)

T7C1) 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?

T7C2) Berechnen Sie für den gleichen Datensatz den Mittelwert und den Median der Reaktionszeiten. Warum unterscheiden sich diese so stark?

T7C3) Genauere Betrachtung der Daten 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.

\*T7C4) Wiederholen Sie die Betrachtung von Verteilung, 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.

\*T7C5) Machen Sie eine Statistik darüber, wie stark sich Mittelwerte und Mediane für die Versuchspersonen unterscheiden. Sollte man hier mitteln?

\*T7C6) Wie stark unterscheiden sich Mittelwerte und Mediane bei dem Beispiel der Sonnenblumenkerne [sbkerne.mat] ?

\*T7C7) Schreiben Sie eine Funktion perzentil, die als Eingabeparameter einen Datenvektor und eine Zahl N bekommt und den Wert des N%-Perzentils des Datenvektors zurückgibt.

**D) Hausaufgabe**

T7H1) Benutzens Sie noch einmal die Daten von gestern [stimulus.mat] und [antwort.mat] und die von Ihnen in T6H3 bestimmte Schwelle für den Beginn der Antwort der Modellzelle: Bestimmen Sie Mittelwert und Standardabweichung der Reaktion der Modellzelle vor, während und nach der Injektion des Strompulses. (Hier handelt es sich um eine Mittelung über die Zeit.)

T7H2) Wie wir gesehen haben, ist die Antwort einer Apparatur selbst ohne biologisches Präparat nicht perfekt rauschfrei. Um das Geräterauschen abzuschätzen, wurden für die Apparatur mit der Modellzelle 100 Messungen (sogenannte "trials") mit dem gleichen Reiz [stimulus.mat] durchgeführt und die Antworten als Matrix unter [antworten1khz.mat] abgespeichert.

Schauen Sie sich ein beliebiges "trial" zusammen mit dem Reiz an. Berechnen und plotten Sie den Zeitverlauf der über die trials gemittelten Antwort.

Berechnen und plotten Sie den Mittelwert und die Standardabweichung der jeweils letzten 300ms für jedes trial. Gibt es eine Tendenz? Gibt es Ausreißer?

Sind Mittelwert und / oder Standardabweichung vor, während und nach der Reizung unterschiedlich?

T7H3) Folgende Daten wurden im Rahmen eines Fortgeschrittenen-Praktikums gewonnen: [spikedaten\_kurz.mat]. Es handelt sich dabei um intrazelluläre Messungen des Membranpotentials eines Blutegelneurons. Es wurden 10 Antworten dieser Zelle auf einen jeweils gleichen Reiz aufgezeichnet, einen Strompuls, dessen Zeitverlauf (in nA) im gleichzeitig gespeicherten Vektor stimulus abgelegt ist. Das Neuron antwortet auf diesen Strompuls mit einer Depolarisation des Membranpotentials, die mit sogenannten "spikes" (Aktionspotentialen, immer ähnlich aussehenden Ausschlägen der Spannung) überlagert ist. Die Daten wurden mit einer Frequenz von 10000 Datenpunkten pro Sekunde aufgenommen.

a) Schreiben Sie ein Skript, um sich alle Antwortspuren gemeinsam mit dem Zeitverlauf der Reizung anzusehen (das darf auch nacheinander sein).

b) Legen Sie "per Augenmaß" eine Schwelle fest, mit der sich die Aktionspotentiale finden lassen. Achtung: jeweils am Ende des Strompulses gibt es in der Antwortspur ein Artefakt. Dieses wollen wir möglichst nicht als Aktionspotential erkennen.

\*\*T7H4) Schreiben Sie eine Funktion, die als Eingabeargumente die Matrix mit den Messdaten der letzten Aufgabe und einen Schwellwert bekommt und als Ausgabe einen Vektor liefert, wie viele Spikes in den einzelnen Durchläufen jeweils ausgelöst wurden. (Die Schwierigkeit an dieser Aufgabe ist, dass Blutegel-spikes eine zeitliche Länge von mehreren Millisekunden besitzen aber jeweils nur einmal gezählt werden sollen).

Bestimmen Sie damit Mittelwert und Standardabweichung der Spikeanzahlen.

\*T7H5) Etwas 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 oben aufgeführten Formel 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?

\*\*) Schreiben Sie eine Funktion, die für die Angabe von zwei Mittelwerten und zwei Standardabweichungen ausgibt, bei welchem Wert man die Grenze ziehen sollte.

**Z) Zusatzthema: 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. Interessierte können aber mit zwei einfachen Beispielen die Anwendung von Signifikanztests in Matlab ausprobieren. Signifikanztests sind nicht im Standard-Programmunfang von Matlab enthalten, sondern finden sich in der Toolbox "statistics" (die hoffentlich auf allen Rechnern im Raum installiert sein sollte.)

Das erste Beispiel ist der t-Test für den Erwartungswert einer normal verteilten Stichprobe. Bei diesem Test ist die Nullhypothese, dass eine Menge von Messwerten (unabhängige, normal verteilte Zufallsvariablen) einer Verteilung mit einem gegebenen Mittelwert μ0 und unbekannter Varianz entstammt, also dass μ0 = μ. Dafür wird mit dem empirischen Stichprobenmittelwert Bild (s.o.) und der empirischen Stichprobenstandardabweichung s (s.o., dort als sx bezeichnet) die Testprüfgröße t berechnet:



Die Nullhypothese μ0 = μ wird zum Signifikanzniveau α abgelehnt wenn



 also der Betrag von t größer dem -Quantil der t-Verteilung mit n − 1 Freiheitsgraden (diese sind normalerweise in Tabellen abgelegt und Matlab natürlich bekannt). In Matlab wird dieser Test durch den Befehl ttest umgesetzt (die Interpretation nimmt Matlab einem allerdings nicht ab.)

Die Syntax lautet h=ttest(vektor,mittelwert) bzw h=ttest(vektor,mittelwert,alpha), der Rückgabewert ist 1 wenn die Nullhypothese abgelehnt wird (also wenn der erwartete und der empirische Mittelwert verschieden sind), sonst ist er 0.

Das zweite Beispiel ist ein t-Test für zwei unabhängige Stichproben, ttest2. Bei diesem lautet die Nullhypothese, dass zwei Stichproben x und y zwei Normalverteilungen mit gleichem Mittelwert (aber eventuell verschiedener Varianz) entstammen, also H0: μx = μy. Hierzu wird mit den empirischen Stichprobenvarianzen und Stichprobenmittelwerten die sogenannte gewichtete Varianz



bestimmt, um damit die Prüfgröße



zu berechnen. Mittels der Ungleichung



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

Die Syntax lautet h=ttest2(vektor1,vektor2) bzw. h=ttest2(vektor1,vektor2,alpha)

T7Z1) Ein superschlauer Futtermittelhersteller behauptet, dass eine Maus im Durchschnitt 32 Futterpellets am Tag frisst.

a) Überprüfen Sie diese Aussage für ein Signifikanzniveau von 5% für Ihre gesamte Mäusepopulation anhand der Messdaten [pellets] und [pellets2].

b) Wie sieht es bei einem Signifikanzniveau von 10% aus?

\*c) Trifft die Behauptung auf irgendeine der Mäuse zu? Für welches Niveau?

T7Z2) Untersuchen Sie für einen Ihrer beiden Datensätze:

a) Haben die ersten beiden Mäuse signifikant unterschiedliche Mengen gefressen?

b) Sind am ersten und am fünften Tag von der gesamten Mäusegruppe signifikant unterschiedlich viele Pellets gefressen worden?

\*c) Gibt es Unterschiede zwischen irgendeinem Mäusepaar?

\*d) Gibt es Unterschiede zwischen irgendeinem Paar von Tagen?