Dieses Quiz soll Ihnen helfen, Kapitel 4.7 und 4.8 besser zu verstehen.

Frage 1

20 zufällig ausgewählte Patienten mit Fieber werden zufällig in zwei Gruppen aufgeteilt. Die eine Gruppe (12 Personen) wird mit einem neuen Medikament behandelt. Die andere Gruppe (8 Personen) wird mit dem herkömmlichen Medikament behandelt. Es wird bei jeder Person der Rückgang des Fieber nach einer Stunde gemessen. Wir wollen nun prüfen, ob das neue Medikament das Fieber innerhalb einer Stunde signifikant stärker senken kann und verwenden einen Zwei-Stichproben t-Test. Ist ein gepaarter oder ein ungepaarter t-Test angebracht?

◯ Gepaart.

Leider nicht.

√ ◯ Ungepaart

Richtig!

◯ Keine Aussage möglich.

Leider nicht.

Es gibt keine eindeutige Art, auf die man eine Person aus der einen Gruppe einer Person aus der anderen Gruppe zuordnen könnte. Zudem sind die Gruppen nicht gleich gross. Es muss daher ein ungepaarter t-Test gemacht werden.

Frage 2

Es wurden neue Augentropfen entwickelt, die den Augeninnendruck senken sollen. Um die Wirkung des neuen Medikaments zu prüfen, werden 10 Personen zufällig ausgewählt. Jede Person bekommt in das linke Auge die neuen Augentropfen. In das rechte Auge bekommt sie herkömmliche Augentropfen. Gemessen wird nun für jedes Auge und jedem Patienten die Senkung des Augeninnendrucks nach fünf Minuten. Wir wollen nun mit einem Zwei-Stichproben t-Test prüfen, ob der Augeninnendruck mehr sinkt, wenn das neue Medikament verwendet wird. Ist ein gepaarter oder ein ungepaarter t-Test angebracht?

◯ Ungepaart.

Leider nicht.

√ ◯ Gepaart.

Richtig!

◯ Keine Aussage möglich.

Leider nicht.

Wir haben zehn Messungen für die neuen Augentropfen und zehn Messungen für die herkömmlichen Augentropfen. Zu jeder Messung mit den neuen Tropfen (linkes Auge) kann man eindeutig eine Messung mit den herkömmlichen Tropfen zuweisen (rechtes Auge der gleichen Person). Es ist daher ein gepaarter t-Test angebracht.

Frage 3

Ein Verwerfungsbereich ist ein Bereich, der...

◯ ... unplausible Parameterwerte enthält.

Leider nicht.

◯ ... plausible Parameterwerte enthält.

Leider nicht.

◯ ... plausible Werte der Teststatistik enthält, wenn die Nullhypothese stimmt.

Leider nicht.

√ ◯ ... unplausible Werte der Teststatistik enthält, wenn die Nullhypothese stimmt.

Richtig!

Ein Verwerfungsbereich enthält unplausible Werte für die Teststatistik, falls die Nullhypothese stimmt. Er wird im Hypothesentest gebraucht. Ein Vertrauensintervall enthält plausible Parameterwerte. Das Vertrauensintervall kann alternativ zu einem Hypothesentest verwendet werden.

Frage 4

Ein Vertrauensintervall ist ein Bereich, der...

◯ ... unplausible Parameterwerte enthält.

Leider nicht.

√ ◯ ... plausible Parameterwerte enthält.

Richtig!

◯ ... plausible Werte der Teststatistik enthält, wenn die Nullhypothese stimmt.

Leider nicht.

◯ ... unplausible Werte der Teststatistik enthält, wenn die Nullhypothese stimmt.

Leider nicht.

Ein Verwerfungsbereich enthält unplausible Werte für die Teststatistik, falls die Nullhypothese stimmt. Er wird im Hypothesentest gebraucht. Ein Vertrauensintervall enthält plausible Parameterwerte. Das Vertrauensintervall kann alternativ zu einem Hypothesentest verwendet werden.

Frage 5

Ein Forscher versucht ein Gewicht möglichst genau zu bestimmen. Dazu misst er das Gewicht mit seiner Waage 5 mal und berechnet ein 95% Vertrauensintervall für das Gewicht (die Waage hat einen zufälligen Messfehler). Um ganz sicher zu gehen, bittet er noch seinen Kollegen, mit dem gleichen Gewicht und der gleichen Waage nochmals 5 Messungen zu machen und aus seinen 5 Messungen auch ein 95% Vertrauensintervall zu berechnen. Am Schluss vergleichen die beiden Forscher ihre beiden Vertrauensintervalle und stellen fest, dass es zwar einen grossen Überlapp gibt, aber dass sie nicht identisch sind. War es zu erwarten, dass die beiden Vertrauensintervalle nicht genau identisch sind?

√ ◯ Ja, denn die beiden Forscher werden auf Grund der Zufallsschwankung der Waage nicht genau die gleichen Messungen gemacht haben.

Richtig!

◯ Nein, denn der zu Grunde liegende Parameter (wahres Gewicht) ist in beiden Fällen identisch.

Leider nicht.

Ein Vertrauensintervall berechnet sich aus den zufälligen Beobachtungen. Daher ist es ein *zufälliges* Intervall. Wenn ein zweiter Forscher also das selbe Experiment wiederholt, wird er wegen der auftretenden Zufallsfehler leicht andere Messwerte erhalten und deshalb auch ein leicht anderes Vertrauensintervall. Das Beobachtung der beiden Forscher war also zu erwarten. In beiden Fällen enthält das Vertrauensintervall aber den wahren Parameter mit grosser Wahrscheinlichkeit.

Frage 6

Wir machen einen Einstichproben t-Test mit der Nullhypothese $H_0 : \mu = 0$ und der Alternative $H_A : \mu \neq 0$. Die gemessenen Werte sind $x_1 = 0.3$, $x_2 = 0.2$, $x_3 = -0.1$, $x_4 = 0.6$, $x_5 = 0.8$. Der beobachtete Wert der Teststatistik ist $t = 2.30$. Daraus ergibt sich der p-Wert $p = 0.08$. Angenommen, wir haben die Messung x_2 falsch abgetippt und stellen fest, dass in Wirklichkeit gilt $\tilde{x}_2 = 0.7$. Wir rechnen erneut den beobachteten Wert der Teststatistik aus und erhalten nun $\tilde{t} = 2.82$. Ändert sich auch der p-Wert?

◯ Nein, der p-Wert bleibt gleich.

Leider nicht.

◯ Ja, der p-Wert wird grösser.

Leider nicht.

√ ◯ Ja, der p-Wert wird kleiner.

Richtig!

◯ Keine Aussage möglich.

Leider nicht.

Falls die Nullhypothese richtig ist, folgt die Teststatistik der Verteilung $T \sim t_4$. Der p-Wert ist die Wahrscheinlichkeit, dass die Teststatistik einen so extremen Wert wie den beobachteten Wert der Teststatistik oder einen noch extremeren Wert annimmt. Was "extrem" bedeutet, richtet sich dabei nach der Alternativhypothese. Falls $t = 2.30$ ist, kann man den p-Wert folgendermassen berechnen:

$$\begin{aligned} p &= P[T \leq -2.30] + P[T \geq 2.30] = P[T \leq -2.30] + (1 - P[T \leq 2.30]) = \\ &= 0.04 + (1 - 0.96) = 0.08 \end{aligned}$$

Falls wir nun $t = 2.30$ durch $\tilde{t} = 2.82$ ersetzen müssen wir für den p-Wert ausrechnen: $\tilde{p} = P[T \leq -2.82] + P[T \geq 2.82]$. Da jede t_n-Verteilung ungefähr wie eine Glockenkurve aussieht, ist $P[T \leq -2.82]$ kleiner als $P[T \leq -2.30]$. Analog ist $P[T \geq 2.82]$ kleiner als $P[T \geq 2.30]$ (machen Sie sich eine Skizze und schauen Sie sich die Fläche unter der Dichtekurve an). Deshalb ist der neue p-Wert kleiner als der alte.

Frage 7

(Aus *New England Journal of Medicine*, 322(12), 1990, pp. 789-793) Wir betrachten fünfzehn eineiige Zwillingspaare. Die Paare sind so ausgewählt, dass in jedem Paar eine Person unter Schizophrenie leidet und die andere nicht. Per MRI vergleichen wir das Volumen des linken Hippocampus (Teil des Gehirns) innerhalb der Paare (Volumen von Person ohne Schizophrenie: a_i; Volumen von Person mit Schizophrenie: b_i). Ein gepaarter, zweiseitiger t-Test ($H_0 : \mu = 0$, $H_A : \mu \neq 0$) zeigt, dass die mittlere Differenz der Volumen ($x_i = a_i - b_i$, d.h. "Volumen ohne Schizophrenie minus Volumen mit Schizophrenie") auf dem 5%-Niveau signifikant von null verschieden ist (mittlere Differenz $\overline{x}$: 0.199). D.h., das Volumen scheint tatsächlich davon abzuhängen, ob man an Schizophrenie erkrankt ist. Würde ein gepaarter, einseitiger t-Test ($H_0 : \mu = 0$, $H_A : \mu > 0$) die Nullhypothese auch signifikant verwerfen?

◯ Nein.

Leider nicht.

√ ◯ Ja.

Richtig!

◯ Keine Aussage möglich.

Leider nicht.

Bei einem zweiseitigen Test besteht der Verwerfungsbereich aus zwei Bereichen: Ein Bereich, der unplausibel grosse Werte enthält und ein Bereich, der unplausibel kleine Werte enthält. Falls die Nullhypothese stimmt, ist die Wahrscheinlichkeit, dass die Teststatistik in einen der Bereiche fällt jeweils 2.5%, insgesamt also 5%. Bei der einseitigen Alternative $H_A : \mu > 0$ enthält der Verwerfungsbereich nur unplausibel grosse Werte und ist deshalb von der Form $K = [c; \infty)$. Falls die Nullhypothese stimmt, ist die Wahrscheinlichkeit, dass die Teststatistik in K fällt auch 5%. Der Bereich für unplausibel grosse Werte beim einseitigen Test deckt also 5% ab; der Bereich für unplausibel grosse Werte beim zweiseitigen Test deckt aber nur 2.5% ab. Deshalb ist der Bereich für unplausibel grosse Werte beim einseitigen Test grösser als der Wert für unplausibel grosse Werte beim zweiseitigen Test. Wenn die Teststatistik schon in dem relativ kleinen, "oberen" Teil des Verwerfungsbereichs des zweiseitigen Tests liegt, dann liegt sie erst recht im grösseren Verwerfungsbereich des einseitigen Tests. D.h., wenn der zweiseitige Test verwirft, verwirft der einseitige Test, der in die "richtige Richtung" (hier: $H_A : \mu > 0$) sensitiv ist erst recht. Deshalb ist in dieser Frage die Antwort "Ja" richtig. Der einseitige Test, der in die "falsche Richtung" (hier: $H_A : \mu < 0$) sensitiv ist hätte allerdings keine Chance, die signifikante Differenz zu erkennen.

Frage 8

Wir wollen untersuchen, bei welchen Gene sich die Aktivität ändert, wenn wir ein neues Medikament verwenden. Dazu untersuchen wir mit einem Microarray 10.000 Gene bei fünf Patienten ohne Behandlung und sechs Patienten mit Behandlung mit dem neuen Medikament. Für alle 10.000 Gene machen wir nun einen ungepaarten, zweiseitigen t-Test mit Signifikanzniveau 5% und vergleichen die Aktivität des Gens bei den behandelten und den unbehandelten Patienten. Angenommen, das Medikament hat gar keinen Effekt (d.h. die Mittelwerte in den beiden Gruppen sind für jedes Gen in Wahrheit identisch). Wie viele Gene werden wir mit unseren 10.000 t-Tests in etwa fälschlicherweise als signifikant unterschiedlich aktiv bezeichnen?

◯ 0

Leider nicht.

◯ 50

Leider nicht.

√ ◯ 500

Richtig!

◯ 5000

Leider nicht.

Der t-Test mit dem Signifikanzniveau 5% hat einen Fehler 1. Art von 5% (per Definition). D.h., wenn es gar keinen Unterschied in den Mittelwerten gibt, wird der t-Test mit 5% Wahrscheinlichkeit aber doch einen signifikanten Unterschied angeben (und somit eine falsche Aussage machen). Wenn wir 10.000 t-Tests machen und bei jedem mit 5% Wahrscheinlichkeit ein (fälschlicherweise) signifikantes Ergebnis erhalten, müssen wir insgesamt mit ca. $10.000 \cdot 0.05 = 500$ fälschlicherweise signifikanten Ergebnissen rechnen. In der Statistik ist dieses Problem unter dem Begriff “multiples Testen” oder auch “Alphafehler-Kumulierung” bekannt. Es gibt verschiedene mehr oder weniger befriedigende Lösungen zu diesem Problem. Falls es Sie interessiert, schauen Sie auf Wikipedia den Begriff “Alphafehler-Kumulierung” nach.