Statistische Inferenz | 02 | Zentraltendenz, Streuung, Standardfehler

Musterlösung

Prof. Dr. Roland Schäfer | Germanistische Linguistik FSU Jena

14. November 2024

Hinweis: Wo nicht anders angegeben, runden Sie die Ergebnisse auf zwei Nachkommastellen.

1 Skalenniveaus

Bestimmen Sie das Skalenniveau von folgenden Messgrößen:

1. Prozentwerte Verhältnisskala
2. Wortfrequenz-Rang (häufigstes Wort, …, seltenstes Wort) Ordinalskala
3. Kasus Nominalskala
4. Geschwindigkeit Intervallskala
5. Akzentsitz (Erstsilbe, Mittelsilbe, Endsilbe) Nominalskala
6. Satzlänge, gemessen in Wörtern Intervallskala
7. Frequenz eines Wortes im Korpus (absolute Zahl) Intervallskala
8. Höhe über NN Verhältnisskala
9. DSH-Prüfungsniveau (I – III) Ordinalskala
10. Verhältnis Satzlänge in Wörtern zu Wortlänge in Silben in einem Text Verhältnisskala
11. Wortklasse (= Wortart) Nominalskala
12. Beschleunigung Verhältnisskala
13. Textniveau (leicht, mittel, schwer) Ordinalskala
14. Frequenz eines Wortes im Korpus pro eine Millionen Wörter Intervallskala
15. Textsorte Nominalskala

2 Modus und Median

Ermitteln Sie den Modus und wo möglich den Median für folgende Messreihen von Hand (ohne Software):

1. x_1 = [Nom, Akk, Akk, Akk, Nom, Dat, Gen, Nom, Nom, Akk, Dat, Dat, Akk, Akk]
   Modus: Akk, Median: n. d.

2. x_2 = [4, 5, 3, 3, 3, 2, 1, 2, 2, 1, 5, 4, 2, 2, 1, 3, 2]
   Modus: 2, Median: 2

3. x_3 = [4.3, 5.0, 3.0, 3.3, 3.7, 2.3, 1.3, 2.7, 2.0, 1.0, 5.0, 4.3, 2.0, 2.0, 1.3, 3.0, 2.7]
   Modus: 2.0, Median: 2.7

3 Mittel und Streuung

Ermitteln Sie von Hand für die untenstehenden Messreihen das arithmetische Mittel, die Varianz und die Standardabweichung:

1. $x_4 = [2.73, 1.85, 21.24, 17.97, 5.49, 18.90, 12.46, 0.97, 6.45, 7.43]$
   $\bar{x_4} = 9.55$, $var(x_4) = 57.05$, $s(x_4) = 7.55$
2. $x_5 = [1.00, 1.91, 3.12, 4.38, 4.72, 5.29, 3.82, 3.25, 2.04, 0.93]$
   $\bar{x_5} = 3.05$, $var(x_5) = 2.36$, $s(x_5) = 1.54$
3. $x_6 = [1.07, 1.06, 0.94, 1.84, 3.04, 3.22, 4.18, 5.27, 6.27, 6.75]$
   $\bar{x_6} = 3.36$, $var(x_6) = 4.79$, $s(x_6) = 2.19$

4 z-Werte und Standardfehler

Ermitteln Sie für die Messreihen aus Aufgabe 3 die z-Werte für die Messpunkte und die Standardfehler von Hand. Formulieren Sie in eigenen Worten (jeweils ein Satz), was z-Werte und Standardfehler angeben.

1. z-Werte x_4: $[-0.9, -1.02, 1.55, 1.11, -0.54, 1.24, 0.39, -1.14, -0.41, -0.28]$, $SF = 2.39$
2. z-Werte x_5: $[-1.33, -0.74, 0.05, 0.87, 1.09, 1.46, 0.5, 0.13, -0.66, -1.38]$, $SF = 0.49$
3. z-Werte x_6: $[-1.05, -1.05, -1.11, -0.7, -0.15, -0.07, 0.37, 0.87, 1.33, 1.55]$, $SF = 0.69$

Die z-Werte sind die in Standardabweichungen normierten Abweichungen der Messwerte vom Mittelwert der Stichprobe. Der Standardfehler ist die mittlere Abweichung des beobachteten Stichprobenmittels vom Populationsmittel in (unendlich) vielen Stichproben der gegebenen Größe (falls die Varianz in der Population der Varianz in der Stichprobe entspricht). (Mit „gegebener Größe“ ist hier die Größe der Stichprobe gemeint.) Vereinfacht gesagt ist er der Erwartungswert, um den Stichproben der gegebenen Größe vom wahren Mittelwert abweichen.

5 Konfidenzintervalle (Anteilswerte)

5.1 Berechnung des Konfidenzintervalls für Anteilswerte

Berechnen Sie für folgende Anteilswerte (q) die Konfidenzintervalle bei den Stichprobengrößen $n = 10$ und $n = 100$ auf den Konfidenzniveaus $\alpha = 0.9$ und und $\alpha = 0.99$ (also je vier Mal den unteren und den oberen Wert des Konfidenzintervalls). Die kritischen Werte der Normalverteilung entnehmen Sie bitte der zur Verfügung gestellten Tabelle. Runden Sie auf drei Nachkommastellen.

1. $q_1 = 0.21$
2. $q_2 = 0.49$
3. $q_3 = 0.89$

q	n=10, α=0.9	n=100, α=0.9	n=10, α=0.99	n=100, α=0.99
0.21	$[-0.002, 0.422]$	$[0.143, 0.277]$	$[-0.122, 0.542]$	$[0.105, 0.315]$
0.49	$[0.230, 0.750]$	$[0.408, 0.572]$	$[0.083, 0.897]$	$[0.361, 0.619]$
0.89	$[0.727, 1.053]$	$[0.839, 0.941]$	$[0.635, 1.145]$	$[0.809, 0.971]$

Wie interpretieren Sie ein negatives Vorzeichen beim unteren Wert des KIs?

5.2 Konfidenzintervalle für Mittelwerte

5.2.1 Fehler finden

Warum hätte folgende Tabelle ganz nicht gedruckt werden dürfen? Der Fehler ist ohne nachzurechnen erkennbar?

Measure	*M*	*SD*	95% CI	
			Lower	Upper
Age at testing (years)	20.23	2.94	19.59	20.88
Age of onset of L2 learning (years)	5.13	1.78	5.74	5.53

Ingrid Mora-Plaza, Joan C. Mora, Mireia Ortega and Cristina Aliaga-Garcia. Is L2 pronunciation affected by increased task complexity in pronunciation-unfocused speaking tasks? *Studies in Second Language Acquisition*. First View. https://doi.org/10.1017/S0272263124000470

Die Werte für das CI bei *Age of onset* können nicht korrekt sein. Erstens ist die untere Grenze höher als die obere Grenze. Zweitens liegt der Wert für M (das Mittel; bei uns $\bar{x}$) liegt nicht zwischen den beiden Grenzen des KIs. Das gilt selbst dann, wenn *Lower* und *Upper* vertauscht wurden. Drittens sind die Differenzen zwischen den Werten $5.13 - 5 - 53 = 0.4$ und $5.74 - 5.53 = 0.21$. Da das gewöhnliche Konfidenzintervall immer symmetrisch ist, können also auch nicht einfach die Zahlen vertauscht worden sein.

5.2.2 Transfer: Stichprobengröße

Aus den Zahlen für *Age at testing* können wir den Standardfehler und die Stichprobengröße rekonstruieren **auch ohne in den Originalartikel zu schauen**. Finden Sie zuerst den Standardfehler und dann die Stichprobengröße. Mit der Stichprobengröße können Sie dann das korrekte KI für *Age of onset* berechnen.

(1) Da es sich um das 95%-Konfidenzintervall handelt, ist $z = 1.96$, und die Grenzen des KIs sollten $1.96 \cdot SF$ vom Stichprobenmittel entfernt sein. Wir betrachten der Einfachheit halber nur die untere Grenze.

$$1.96 \cdot SF = 20.23 - 19.59 \tag{1}$$

$$1.96 \cdot SF = 0.64 \tag{2}$$

$$SF = 0.33 \tag{3}$$

(2) Die Stichprobengröße kann man nun auf verschiedene Weise herleiten. Der aus meiner Sicht einfachste, aber etwas umständliche Weg: Wir kennen die Gleichung für den Standardfehler SF bei einer bekannten Standardabweichung σ (geschätzt aus einer Stichprobe als s) und einer gegebenen Stichprobengröße n. Es ergibt sich.

$$SF = \frac{s}{\sqrt{n}} \tag{4}$$

$$\frac{SF}{s} = \frac{1}{\sqrt{n}} \tag{5}$$

$$\frac{s}{SF} = \sqrt{n} \tag{6}$$

$$\frac{s}{SF}^2 = n \tag{7}$$

Wenn wir hier die bekannten Werte einsetzen, erhalten wir $n = 79.37$. Es scheinen also (bis auf Rundungsfehler) 79 Probanden gewesen zu sein. Die angegebenen Werte sind auch mit leicht anderen Stichprobengrößen kompatibel. Es handelt sich um eine Schätzung, die aber die Größenordnung der Stichprobe hinreichend genau rekonstruiert. (Wir rechnen mit $n = 79$ weiter, aber der Artikel informiert uns, dass es $n = 82$ waren. Sie sehen daran, wie schnell sich durch Rundungen Fehler in der Rechnung fortsetzen und zu zumindest nicht trivialen Abweichungen führen können.)

$$n = \frac{2.94}{0.33}^2 \tag{8}$$

$$n = 79.37 \tag{9}$$

(3) Für das KI des zweiten Werts gilt weiterhin $z = 1.96$. Wir erhalten:

$$KI = \bar{x} \pm 1.96 \cdot \frac{1.78}{\sqrt{79}} \tag{10}$$

Da wir noch nicht wissen, was von dem präsentierten Datensalat das Mittel ist, betrachten wir $1.96 \cdot \frac{1.78}{\sqrt{79}} = 0,39$. Die Differenz $5.53 - 5.13 = 0.4$ ist dem gleich bis auf Rundungsfehler, während 5.74 nicht in eine solche Beziehung zu den anderen Werten gesetzt werden kann. Vermutlich ist 5.74 verschrieben zu $4.74 = 5.13 - 0.39$, und das Mittel ist tatsächlich 5.13.

6 Transfer: Anteilswerte und Mittelwerte

In dieser Transferaufgabe geht es darum, zu zeigen, wie sich der Standardfehler für Anteilswerte aus dem Standardfehler für Mittelwerte ableiten lässt. Der allgemeine Standardfehler für Normalverteilungen wird berechnet mit:

$$SE = \frac{s}{\sqrt{n}} = \frac{\sqrt{s^2}}{\sqrt{n}} = \sqrt{\frac{s^2}{n}} \tag{11}$$

In der Wurzel der dritten Variante steht also die Varianz (hier der Einfachheit halber die nicht korrigierte Version mit n statt $n - 1$ im Nenner). Der Standardfehler für Anteilswerte — also für Stichproben von n Einsen und Nullen (also letztlich Zähldaten aus einem Bernoulli-Experiment) wie $x_7 = [0, 1, 1, 0, 0, 1, 1, 1, 1, 0]$ — wird berechnet mit:

$$SE = \sqrt{\frac{q(1-q)}{n}} \tag{12}$$

Was bei dem SE für Mittelwerte die Varianz ist, erscheint hier als $p(1 - p)$. Um zu zeigen, dass dies tatsächlich die Varianz für Bernoulli-Stichproben ist, können wir die zunächst genauso berechnen wie für Mittelwerte, wenn wir den Anteilswert von Einsen q als Mittelwert ansetzen.

$$q = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{13}$$

Die Varianz für eine solche Stichprobe (oder Grundgesamtheit) ist dann wie zu erwarten gegeben:

$$s^2(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} \tag{14}$$

Zeigen Sie für solche Fälle, dass:

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{q(1-q)}{n}} \tag{15}$$

Es ist also hinreichend, zu zeigen, dass:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = q(1-q) \tag{16}$$

Berechnen Sie anschließend zur Kontrolle auf beiden Wegen das Ergebnis für $x_7 = [0, 1, 1, 0, 0, 1, 1, 1, 1, 0]$.

Wir kürzen $\sum_{i=1}^{n}$ als $\sum$ ab.

$$\frac{\sum(x_i - \bar{x})}{n} \tag{17}$$

$$\frac{\sum x_i}{n} \cdot \frac{\sum(1 - x_i)}{n} \tag{18}$$

$$q \cdot (1-q) \tag{19}$$

Und das war es eigentlich schon. Die Definition der Anteilswerte für Einsen und Nullen des entspricht den Multiplikanden, wenn man die Summenformel auflöst. Die Beispielrechnung:

$$x_7 = [0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0] \tag{20}$$

$$n_7 = 10 \tag{21}$$

$$\bar{x_7} = q_7 = 0.6 \tag{22}$$

$$s^2(x_7) = \frac{\sum(x_{7_i} - \bar{x_7})^2}{n_7} = 0.24 = q_7(1-q_7) = 0.6 \cdot 0.4 \tag{23}$$

$$\tag{24}$$