Konstruktion eines näherungsweisen Konfidenzintervalls für $p_2 - p_1$

Für die Konstruktion eines Konfidenzintervalls für die Differenz $p_2 - p_1$ benötigt man die folgende Eigenschaft:
Für zwei unabhängige Zufallsgrößen gilt: $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
Mit $\sigma^2_{X-Y} = \mathrm{Var}(X - Y)$ kann man folgern: $\sigma_{X-Y} = \sqrt{\mathrm{Var}(X - Y)} = \sqrt{\mathrm{Var}(X) + \mathrm{Var}(Y)}$

Mit der Abkürzung $k := z_{1-\alpha/2} \cdot \sqrt{\frac{h_1(1-h_1)}{n_1} + \frac{h_2(1-h_2)}{n_2}}$ erhält man analog zum einfachen Wald- Konfidenzintervall für die Differenz der beiden unbekannten Parameter $p_2 - p_1$ das $(1-\alpha)-$Konfidenzintervall $\left[(h_2 - h_1) - k\,;(h_2 - h_1) + k\right]$.
Das Bilden der Differenz (heuristisches Mittel auch in der Analysis) erlaubt die folgende Interpretation:
Ist die linke Grenze des $(1-\alpha)-$Konfidenzintervall o.B.d. A. größer Null, so gilt:
Da nach Konstruktion die (unbekannte) Differenz $p_2 - p_1$ in ca. $(1-\alpha)\cdot 100\%$ der Fälle von den Konfidenzintervallen überdeckt wird, kann man mit der Sicherheitswahrscheinlichkeit $(1-\alpha)$ folgern, dass $p_2 - p_1 > 0$, also $p_2 > p_1$ gilt.
Der Fall „kleiner Null“ liefert natürlich $p_2 < p_1$.

Nun haben wir (in der Schule) zwei Möglichkeiten, die plausibel klingen:

Zwei Größen sind **statistisch signifikant verschieden**, wenn

- die zugehörigen Konfidenzintervalle sich nicht überdecken (Verfahren 1),
- das zugehörige Konfidenzintervall zu $p_2 - p_1$ die Null nicht enthält (Verfahren 2).

Vorsicht: Die beiden Verfahren liefern nicht dasselbe Ergebnis. Falls (1) nicht erfüllt ist, kann (2) noch erfüllt sein, aber nicht umgekehrt!
Auch kann bei Verfahren (1) nicht mit der Sicherheitswahrscheinlichkeit argumentiert werden. Falls Sie es trotzdem machen, müssen Sie hier mit der Bonferronie-Korrektur rechnen. Bei einer geforderten 5% Unsicherheit müssen Sie für beide Konfidenzintervalle mit der Unsicherheit von 2,5% rechnen. Das übersteigt aber nun wirklich das Schulniveau – aber es ist vielleicht für den einen oder anderen Lehrer gut zu wissen.
Warum das Verfahren (2) besser ist, kann mit einfachen algebraischen Mitteln gezeigt werden. In der Literatur habe ich dazu leider nichts gefunden, also muss man „selbst ran“:
Das hat jetzt nicht unbedingt etwas mit Stochastik zu tun, aber interessant ist es schon.

Vorher einige Abkürzen – um Strukturen zu erkennen:

$$x := \frac{h_1(1-h_1)}{n_1};\ y := \frac{h_2(1-h_2)}{n_2};\ z := z_{1-\alpha/2}$$

Verfahren (1)

$\left[h_1 - z \cdot \sqrt{x}\,;\, h_1 + z \cdot \sqrt{x}\right];\ \left[h_2 - z \cdot \sqrt{y}\,;\, h_2 + z \cdot \sqrt{y}\right]$

O.B.d.A. sei $h_2 > h_1$:

Die Intervalle überdecken sich nicht, falls $(h_2 - h_1) - z \cdot (\sqrt{x} + \sqrt{y}) > 0$

Verfahren (2)

Die Null ist nicht im Intervall enthalten, falls $(h_2 - h_1) - z \cdot \sqrt{x + y} > 0$

Ja, nun könnte man gleich unter den Schülern eine Umfrage starten und fragen, ob $\sqrt{x + y} = \sqrt{x} + \sqrt{y}$ gilt. Das Ergebnis kennen wir eigentlich schon!

Da aber für $x, y \neq 0$ auch $\sqrt{x} + \sqrt{y} > \sqrt{x + y}$ gilt, wird bei Verfahren (1) von der Differenz der Punktschätzer immer eine größere Zahl subtrahiert. Also gibt es Fälle, in denen (1) schon nicht mehr funktioniert, während Verfahren (2) noch eine Signifikanz liefert.

Zusatz:

Der Satz

Für zwei unabhängige Zufallsgrößen gilt: $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

soll im Unterricht nicht bewiesen werden, er kann aber plausibel gemacht werden, wenn man die mittleren quadratischen Abweichungen vom Erwartungswert betrachtet.
Geht man (hoffentlich) nicht den algebraischen Weg, so könnte man die Eigenschaft mithilfe einer Simulation als Vermutung entdecken lassen. So eine Simulation hat oftmals mehr Überzeugungscharakter als ein formaler Beweis.

Beispiel:

$n_1 = 350\,;\ p_2 = 0,6\,;\ n_2 = 200\,;\ p_2 = 0,25$

Jeweils 5000 Wiederholungen und Berechnung der Stichprobenvarianzen:

S2: $\mathrm{Var}(X - Y)\,;$ S3 $= \mathrm{Var}(X) + \mathrm{Var}(Y)$

Die Werte unterscheiden sich erst in der 5. Dezimalen voneinander. Mehrere Simulationen und Variationen der Eingangsparameter stützen die Eigenschaft.

Kollektion 1

h_differenz	0,00039785714
	0,0019324375
	0,0019114956

S1 = aMittel ( )
S2 = Varianz ( )
S3 = Varianz (h_1) + Varianz (h_2)