Modul Statistische Aspekte der Analyse molekularbiologischer und genetischer Daten

Übungsblatt 2: Statistische Grundlagen

Janne Pott

WS 2021/22

Sie können Ihre Lösung zu Aufgabe 1 als PDF in Moodle hochladen (Frist: 08.11.2021).

Aufgabe 1: Diagnostischer Test

Die Bayesianische Interpretation der Wahrscheinlichkeit kommt bei diagnostischen Tests zur Anwendung. Als Beispiel wird hier der qRT-PCR-Assay auf SARS-Cov-2 angeführt, der virale RNA nachweisen kann. Seien T+/- die Ergebnisse des PCR-Tests und K+/- das Vorliegen bzw. das Nichtvorliegen einer Corona-Infektion.

a) Definieren Sie **Sensitivität**, **Spezifität** und **Prävalenz**, und stellen sie eine schematische 4-Feldertafel auf.
b) Leider gibt es (bislang) keine konkreten Angaben zu Sensitivität und Spezifität der verschiedenen Testverfahren. Ein Review dazu schätzt allerdings die Sensitivität auf 70% und die Spezifität auf 95% (PMID: 32398230). Der **positiv prädiktiver Wert** (PPW: WSK krank zu sein, wenn der Test positiv ausfällt) und **negativ prädiktiven Wert** (NPW: WSK gesund zu sein, wenn der Test negativ ausfällt) können in Abhängigkeit von Sensitivität, Spezifität und Prävalenz ausgerechnet werden. Bestimmen sie PPW und NPW für folgende Prävalenzen und stellen Sie diese graphisch da:
    - 3% (z.B. Hausarztpraxis)
    - 20% (z.B. Altenheim)
    - 80% (z.B. Isolierstation)
c) Erstellen Sie für eine der Prävalenzen eine 4-Feldertafel (Gesamtfallzahl 1000).

Hinweise: Die Funktion für PPW und NPW ist Ihnen auch schon in der ersten R-Übung begegnet!

Aufgabe 2: LogLikelihood

Bei der Genotypisierung eines biallelischen Markers in 10 diploiden Individuen haben sie viermal das Allel A beobachtet.

a) Warum muss man von n=20 Allelen ausgehen? Welche Verteilung kann man annehmen?
b) Bestimmen Sie die $-$**Log-Likelihood** Ihrer Daten unter der Annahme, dass die Allelhäufigkeit 50% beträgt.
c) Zeigen Sie allgemein, dass der **Maximum-Likelihood-Schätzer** der Allelhäufigkeit bei dieser Art Experiment gleich der Anzahl der Allele A geteilt durch die Anzahl der untersuchten Allele ist.

d) Bestimmen Sie den **Maximum-Likelihood-Schätzer** für Ihre Daten, die zugehörige –Log-Likelihood und vergleichen Sie diese Werte mit denen unter a.
e) Ab wie vielen Treffern könnten Sie die Annahme, dass die wahre Allelfrequenz 50% beträgt, nicht mehr ablehnen? (Signifikanzniveau von 5%)

Aufgabe 3: Konfidenzintervall

Beim ARD-Deutschlandtrend werden monatlich etwa 1000 Wahlberechtigte befragt. In der Anlage wird die Schwankungsbreite zwischen 1.4 und 3.1 Prozentpunkten angegeben (s. Info-Box unten). Überprüfen Sie diese Aussage zur Fehlertoleranz, indem Sie die Konfidenzintervalle für die Anteilswerte $p_1 = 0.05$ und $p_2 = 0.5$ berechnen.

Hinweise: Bei der Befragung von n Wählern gibt es einen Anteil p von Wählern und $q = 1 - p$ Nichtwählern einer Partei x. Überlegen Sie sich zuerst, von welcher Verteilung damit ausgegangen wird. Anschließend können sie die Varianz dieser Verteilung nutzen, um die Standardfehler zu bestimmen.

Untersuchungsanlage	
Grundgesamtheit	Wahlberechtigte Bevölkerung ab 18 Jahren
Stichprobe	Repräsentative Zufallsauswahl / Dual Frame (Relation Festnetz-/Mobilfunknummern 60:40), Disproportionaler Ansatz (West/Ost 70:30)
Erhebungsverfahren	Telefoninterviews (CATI)
Fallzahl	1027 Befragte
Erhebungszeitraum	31.08. bis 02.09.2020
Schwankungsbreite	1.4 bis 3.1 Prozentpunkte (bei einem Anteilswert von 5 bzw. 50 Prozent)

Aufgabe 4: FDR & FWER

Es wurden 9 SNPs auf Assoziation mit Krankheit X getestet. Die daraus resultierenden p-Werte sind in der untenstehenden Tabelle festgehalten. Die Ergebnisse wurden zunächst zum Signifikanzniveau von 5% ausgewertet, was das multiple Testen nicht berücksichtigte.

a) Definieren Sie die Begriffe false discovery rate (FDR) und family-wise error rate (FWER).
b) Geben Sie die jeweiligen Schranken zu den drei genannten Verfahren an und notieren Sie, ob der SNP nach Korrektur signifikant mit Krankheit X assoziiert ist.
c) Was kontrolliert welches Verfahren (FDR oder FWER)?

SNP-ID	p-Wert	Bonferroni	Bonferroni-Holm	Benjamini-Hochberg
rs1001	0.023			
rs1002	0.006			
rs1003	0.025			
rs1004	0.350			
rs1005	0.300			
rs1006	0.040			
rs1007	0.200			
rs1008	0.002			
rs1009	0.015			