Lösungen zu ausgewählten Aufgaben

(6.3)

Verblüffenderweise gibt es schon bei $n = 13$ mehr Permutationen als 32-bittige Zahlen (da die Inverse der Fakultätsfunktion ohne weitere Tricks etwas kompliziert ist, empfiehlt sich hier schlichtes Probieren und eine Binärsuche nach der Grenze, bei der $n!$ größer wird als 2^{32}). Bei $n = 20$ deckt ihr nur noch ein gutes Milliardstel der möglichen Permutationen ab, bei $n = 100$ könnt erwischt ihr nur noch jede 10^{366}ste Permutation, also praktisch keine mehr (übrigens hilft da auch nicht der Übergang zu 64-bittigen Zahlen, mit denen erwischt ihr jede 10^{356}ste Permutation, was genauso „praktisch nichts“ mehr ist).

Abhilfe könnte sein, mehr Zufallszahlen zu ziehen – z.B. könntet ihr eine Menge von Positionen haben und aus dieser „ohne Zurücklegen“ ziehen (es lohnt sich, sowas mal in Python zu schreiben). Effektiv zieht ihr dann 2^{32n} bits „Zufall“. Allerdings muss man solche Verfahren sehr sorgfältig durchdenken, weil es leicht sein kann, dass diese auch keine Gleichverteilung auf den Permutationen mehr erzeugen. Ein Übriges tut dann der Umstand, dass bei manchen der Zufallszahlengeneratoren, die z.B. in der C-Bibliothek stehen, aufeinanderfolgende (oder in gewissem Abstand aufeinanderfolgende) Zufallszahlen gar nicht mehr so schrecklich zufällig sind...

(7.4)

Mit den Ereignissen S (Spam) und H (Ham, gute Mail), C („click“ kommt in Mail vor) ergibt sich

$$P(C) = P(C \mid S)P(S) + P(C \mid H)P(H) = 0.361$$
$$P(S \mid C) = \frac{P(C \mid S)P(S)}{P(C)} = 0.997$$
$$P(H \mid C) = \frac{P(C \mid H)P(H)}{P(C)} = 0.00277.$$

Ist es Zufall, dass die beiden letzten Zahlen zusammengenommen ziemlich genau 1 ergeben?

(12.2)

Sie fragt einfach die beiden Schätzer, welches p sie für ein Bernoulli-Experiment, das bei `N` Versuchen n Erfolge liefert, schätzen würden. Eure Beobachtung sollte irgendetwas zu systematischem Über- und Unterschätzen umfassen.

(13.1)

In den Zeilen soll immer etwas wie $P_\vartheta(\omega)$ stehen.

In der Definition der Konfidenzintervalle wird ja dafür gesorgt, dass bei einer Summe über die Wahrscheinlichkeiten von Stichproben ein ausreichend großer Wert herauskommt, während wir nachher wissen wollen, „wie viele“ ϑ wir in Betracht ziehen müssen, um auf das gewünschte Niveau zu kommen. Darum sammeln wir erst (in dieser Darstellung) $A(\vartheta)$, deren summierte Wahrscheinlichkeit ganz rechts steht und können dann senkrecht das Konfidenzintervall ablesen. Beachtet, dass die Spaltensummen durchaus beliebig verschieden von Eins sein können – dies sagt letztlich wieder nichts anderes, als dass wir eben keine Verteilung haben, aus der etwas wie „Wahrscheinlichkeit, dass ϑ in C ist“ zu bekommen wäre – unsere Konstruktion gibt sowas einfach nicht her.

(14.1)

Es ist $\log n = \log Cm^{-s} = \log C - s \log m$ – also ist $x = \log n$, $y = \log m$ und $z = \log C$.

(14.2)

Abgesehen von einigem Gewackel, das bei so statistischem Kram immer da ist, dürftet ihr einerseits sehen, dass bei hohen Rängen Rr kleiner als erwartet ist. Höhe Ränge sind selten auftretende Wörter (einmal, zweimal, dreimal) – von denen gibt es also offenbar weniger, als Zipf's Law vorhersagt. Vielleicht wäre aber eine bessere Interpretation, dass einfach ihr Rang zu niedrig ist, weil etwas häufigere Wörter fehlen, also z.B. die Häufigkeit 20 nicht besetzt ist? Ein erster Schritt zur Klärung dieser Frage wäre, nachzusehen, ob die Probleme besser oder schlechter werden, wenn man längere Texte verwendet.

Je nach Dokument seht ihr vielleicht auch bei ganz niedrigen Rängen Abweichungen. Hier fehlen bestimmt nicht Wörter bei niedrigeren Rängen (bei welchen auch?) – hier sind die häufigen Wörter also gewiss unterhäufig.

(14.3)

Hier sehen die Zahlen wohl deutlich wüster aus, dürften aber speziell bei den seltenen Wörtern (also denen mit hohem Rang) weniger Schwierigkeiten machen als beim Rechnen mit den Rängen. Dafür siehts bei den häufigen Wörtern (also denen mit niedrigem Rang) vermutlich ganz schlecht aus – hier scheint r weit zu groß sein, wobei allerdings durch das Quadrieren von r das (tatsächlich vorhandene) Problem größer aussieht als es real ist. Genau mit diesem Bereich werden wir uns in unserer Ableitung noch beschäftigen müssen. In der grafischen Darstellung der folgenden Aufgaben wird klarer werden, was hier vorgeht.

(20.2)

$t = 9$. Bei $t = 8$ ist $\alpha = \beta(0.5) \approx 0.056$, also zu viel. Bei $t = 9$ ist $\beta(0.5) \approx 0.01$ – damit ist der Test deutlich schärfer als gewollt, aber das ist bei $n = 10$ eben nicht zu vermeiden. Natürlich hat man auch bei $t = 10$ noch einen Test zum Niveau 0.05 (und in der Tat einen zum Niveau 0.01), aber wenn Leute schon so fragen, ist natürlich immer der Test mit dem größten β unter allen mit verträglichem α gemeint.