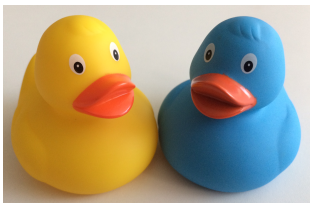
Sind blaue Superenten seltener krank?

Hintergrundinformation

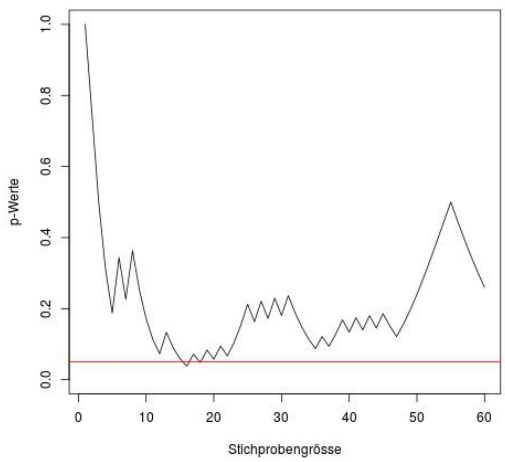


Die blauen Enten in unserem Becken sind in Wirklichkeit auch genau zur Hälfte krank, d.h. trotz unserer „besonderen Züchtung“ bekommen sie die Krankheit auch nicht seltener als herkömmliche Badeenten.

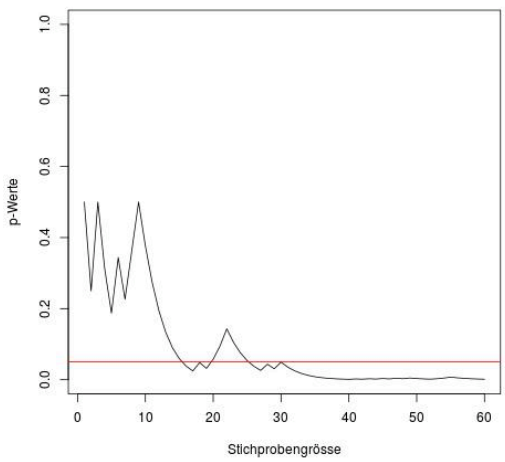
Wie kann es sein, dass der Pfad trotzdem manchmal unter die rote Linie geht? Das bedeutet doch, dass unsere Enten signifikant seltener krank werden?

Der Pfad zeigt den sogenannten p-Wert des statistischen Tests, den wir anwenden. Informell ausgedrückt kann man sagen, dass der p-Wert misst, wie überraschend eine Beobachtung ist, wenn es in der Tat keinen Unterschied gibt[1]. In unserem Beispiel also, wie überraschend ein beobachteter Anteil kranker Enten ist unter der Annahme, dass die blauen Enten genau gleich oft krank werden wie die gelben. Je überaschender der beobachtete Anteil ist, desto weniger spricht dafür, dass es keinen Unterschied gibt.

Mit jeder neuen Ente, d.h. mit jedem neuen Datenpunkt, den wir zufällig (bzw. hier in zufälliger Reihenfolge) ziehen, verändert sich der p-Wert des Tests, den wir aus den bis dahin beobachteten Werten ausrechnen können. Wenn wir alle 60 Enten untersucht haben, sehen wir das Endergebnis für diese Stichprobe: Es gibt keinen signifikanten Unterschied zu den herkömmlichen Bade-Enten. In den Grafiken sind komplette Pfade abgebildet:



So sieht ein möglicher Pfad aus, wenn die blauen Enten in Wirklichkeit genauso häufig krank sind wie die gelben.



So sähe ein möglicher Pfad aus, wenn die blauen Enten in Wirklichkeit seltener krank wären als die gelben.

[1] Formell ist der p-Wert definiert als die Wahrscheinlichkeit, den beobachteten oder einen noch geringeren Anteil von kranken Enten in einer Stichprobe zu finden, wenn die blauen Enten in Wirklichkeit genauso häufig erkranken wie die gelben. Ist diese Wahrscheinlichkeit klein, spricht dies dafür, dass es doch einen Unterschied im Erkrankungsrisiko gibt.

Schaut man jedoch schon vorher immer wieder in die Ergebnisse, sieht man, dass der p-Wert hin- und herschwankt. Er ist so konstruiert, dass für eine feste Stichprobengrösse nur in 5% der Fälle[2] eine falsche Testentscheidung getroffen wird. D.h. wenn wir ganz viele Stichproben mit jeweils 60 Enten untersuchen würden, würden wir nur in 5% der Fälle fälschlicherweise zu dem Ergebnis kommen, unsere neu gezüchteten Enten seien gesünder.

Wenn Forschende hingegen aus Unwissenheit oder aus Neugierde (oder Erfolgsdruck) laufend die Testergebnisse berechnen, bevor die Studie abgeschlossen ist, kann es passieren, dass der Pfad des p-Werts durch die zufällige Reihenfolge von kranken und gesunden Enten manchmal unter die rote Linie rutscht. Würde man an dieser Stelle abbrechen, anstatt die restlichen Enten noch zu untersuchen und das Gesamtergebnis zu betrachten, ist das Risiko viel höher, fälschlicherweise zu dem Schluss zu kommen, unsere neu gezüchteten Enten seien seltener krank.

Deshalb ist es ein wichtiges Qualitätsmerkmal guter wissenschaftlicher Studien, dass die Anzahl untersuchter Personen/Enten/… vorab festgelegt und dann eingehalten wird, und man nicht einfach in dem Moment die Studie abbricht, in dem die Ergebnisse grade zufällig „gut aussehen“.

Es gibt allerdings auch Studien, bei denen es wichtige ethische Gründe dafür gibt, bereits vor Abschluss der Studie einen Teil der Daten zu analysieren. Wird z.B. ein neues Medikament für eine schwere Erkrankung untersucht, wäre es unethisch, den Personen in der Kontrollgruppe das Medikament länger vorzuenthalten, wenn sich z.B. bereits nach der Hälfte der Studie eindeutig zeigt, dass das Medikament wirksam ist. Für solche geplanten Interims-Analysen gibt es deshalb spezielle Versuchspläne. Dabei wird die Anzahl der Interims-Analysen vorab festgelegt und das alpha-Niveau der einzelnen Tests nach einem bestimmten Schema nach unten korrigiert. Dadurch wird sichergestellt, dass es insgesamt nicht zu einer erhöhten Fehlerwahrscheinlichkeit kommt.

Da die Planung um Umsetzung einer korrekten statistischen Auswertung nicht immer einfach ist, gibt es an der Universität Zürich (wie auch an der ETH) statistische Beratungsstellen, die die Forschenden dabei unterstützen (siehe z.B. http://www.crs.uzh.ch/en/expertise.html).

CRS@UZH

www.crs.uzh.ch

Scientifica 2019

[2] 5% ist das Signifikanzniveau des Tests. Das Signifikanzniveau begrenzt die Wahrscheinlichkeit für den Fehler 1. Art, d.h. dass der Test einen Unterschied als signifikant beurteilt, obwohl in Wirklichkeit blaue und gelbe Enten gleich häufig erkranken.