8 Ein zweiseitiger Vorzeichentest: Welche Sorte anpflanzen?

Problemstellung

In vielen entwickelten Ländern erzeugt die Landwirtschaft Überschüsse bei gewissen Produkten (Milch, Getreide, Früchte etc.), deren Vermarktung kostspielig sein kann. Deshalb versuchen weitsichtige Bauern auf neue Erzeugnisse umzustellen. Beispielsweise ist die aus China stammende Kiwifrucht um 1960 in Neuseeland eingeführt und bald schon grossmassstäblich kultiviert worden. Inzwischen wird die beliebte Frucht auch in Europa erfolgreich angepflanzt.

Wie Sie sich denken können, gedeiht eine Pflanze nicht ohne weiteres in einem neuen Klima mit einem anderen Boden. In der Regel stehen aber verschiedene Sorten einer bestimmten Pflanze zur Auswahl und man kann in Vorversuchen abklären, welche sich für die neue Umgebung überhaupt eignen.

Wir denken uns nun, dass die zwei vielversprechendsten Sorten A und B bezüglich ihres Ertrages untersucht werden sollen. Zu diesem Zwecke werden insgesamt 20 Parzellen jeweils halb mit Sorte A und halb mit Sorte B bepflanzt. Nach der Ernte soll entscheiden werden, ob sich die Erträge der beiden Sorten signifikant unterscheiden.

Die Lernziele dieses Kapitels

In diesem Kapitel erfahren Sie, wie die statistische Auswertung dieses Versuches *geplant* werden kann. Insbesondere lernen Sie, was man unter einem *Vorzeichentest* - eine Anwendung der Binomialverteilung - versteht.

Erstmals begegnen Sie einem sog. *zweiseitigen* Test, bei welchem der Verwerfungsbereich nicht wie bis anhin aus *einem,* sondern aus *zwei getrennten* Bereichen besteht.

Der Sortenvergleich ist *ein* Beispiel eines häufigen und wichtigen Test-Schemas: Gegeben sind zwei natürlich *gepaarte* Stichproben, hier die Erträge der beiden Sorten. Man möchte abklären, ob sich die beiden Stichproben signifikant unterscheiden.

Sie haben nach jedem Abschnitt in diesem Kapitel die Gelegenheit, anhand einer weiteren Testsituation dieser Art zu prüfen, ob Sie die Erklärungen verstanden haben. Nachstehend die Aufgabenstellung, die Sie parallel zum Sortenvergleich bearbeiten werden.

Aufgabe 8.0: Geburtsgewicht und Intelligenz

Zwei Forscher stellten sich die Frage, ob die Ernährung im Mutterleib einen Einfluss auf die spätere Intelligenz hat. Sie untersuchten zu diesem Zwecke 14 Paare von eineiigen Zwillingen, indem sie Jahre nach der Geburt deren IQ (=Intelligenzquotient) ermittelten. Hier die Untersuchungsergebnisse:

Nr. des Paares	1	2	3	4	5	6	7	8	9	10	11	12	13	14
IQ des schweren Zwilling: X	97	79	100	100	100	124	95	80	91	108	91	90	104	119
IQ des leichten Zwilling: Y	97	70	101	106	85	123	84	70	84	106	97	90	92	104

Vergleichen Sie dieses neue Testproblem mit dem Sortenvergleich und notieren Sie die grundsätzlichen Gemeinsamkeiten. (Aus A. Engel: Stochastik, 1987, Ernst Klett Verlag)

8.1 Die Nullhypothese beim Sortenvergleich

In den folgenden Abschnitten werden wir die statistischen Fachbegriffe H_0, H_1 etc. für den Test "Sortenvergleich" definieren. Parallel dazu werden Sie den gleichen Begriff für den Test "Geburtsgewicht und IQ" bestimmen.

Wir beginnen mit der Nullhypothese: Der Test hat zum Ziel, einen allfälligen Ertrags*unterschied* zwischen den beiden Sorten aufzudecken. Die zukünftigen Testergebnisse sollten nach Möglichkeit die folgende Nullhypothese *widerlegen*:

H_0: "Sorten A und B sind gleich ertragsreich"

Aufgabe 8.1

Im Beispiel "Geburtsgewicht und IQ" interessiert uns, ob es einen signifikanten Unterschied zwischen den IQ's des schweren und des leichten Zwillings gibt.

Formulieren Sie für diesen Test die Nullhypothese.

8.2 Die Alternativhypothese beim Sortenvergleich

Vermutlich erwarten Sie aufgrund der bisher durchgeführten Tests, dass wir allenfalls die Nullhypothese verwerfen zugunsten einer Hypothese folgender Art: "Sorte A ist ertragsreicher als Sorte B" oder umgekehrt. Beachten Sie aber, dass wir *im voraus* keinerlei Hinweis auf eine solchen Tatbestand haben. Uns interessiert nur, ob es *überhaupt einen Unterschied* zwischen den beiden Sorten gibt. Deshalb arbeiten wir mit der folgenden Alternativhypothese:

H_1: "Die Erträge der Sorten A und B sind verschieden"

Sollten wir aufgrund der Testergebnisse schliesslich H_1 akzeptieren, so wird in der Regel auch klar sein, welche der beiden Sorten wir als ertragsreicher zu betrachten haben.

Aufgabe 8.2

Formulieren Sie für den Test "Geburtsgewicht und IQ" eine Alternativhypothese.

8.3 Die Testgrösse beim Sortenvergleich

Wir werden nach der Ernte die Erträge auf den 20 Parzellen wägen und in einer Liste festhalten, etwa so:

Parzellennummer:	1	2	...	...	20
Ertrag der Sorte A: X	5.2	3.9	...	...	7.7
Ertrag der Sorte B: Y	4.3	4.9	...	...	8.8
Vorzeichen von X-Y	**+**	**–**	**...**	**...**	**–**

In der letzten Zeile haben wir die Vorzeichen der Differenzen X–Y notiert.

Wir argumentieren nun so:

Wenn gemäss H_0 die Erträge der beiden Sorten *gleich* sind, so werden diese Vorzeichen in den einzelnen Parzellen zufällig auftreten und es sollten etwa gleich viele Plus- wie Minuszeichen vorkommen.

Weicht aber die Anzahl der Pluszeichen wesentlich von 10 ab, so glauben wir nicht mehr daran, dass H_0 zutrifft. In diesem Fall verwerfen wir H_0 und akzeptieren H_1.

Als Testgrösse verwenden wir hier offenbar:

Z = "Anzahl der positiven Differenzen von X-Y"

oder salopper:

Z = "Anzahl der Pluszeichen"

Deshalb spricht man von einem *Vorzeichentest*.

Aufgabe 8.3

a *Definieren Sie für den Test "Geburtsgewicht und IQ" die Testgrösse Z.*

b *Welchen Wert hat Z? Beachten Sie: Gewisse der Zwillingspaare haben den gleichen IQ, d. h. die Differenz der beiden Zufallsgrössen X und Y ist 0. Es ist üblich und zweckmässig, solche Paare beim Vorzeichentest wegzulassen, weil sie nichts zur Entscheidung beitragen.*

8.4 Der Verwerfungsbereich beim Sortenvergleich

Wir werden, wie gesagt, H_0 verwerfen, wenn die Anzahl Z der Pluszeichen, *auf die eine oder die andere Seite*, zu stark von von der Mitte (10) abweicht. Was wir unter "zu stark" verstehen, ist uns zunächst freigestellt. Um eine *signifikante* Entscheidung treffen zu können, müssen wir die maximale Abweichung so festlegen, dass sie in höchstens 5% der Fälle überschritten wird, wenn H_0 gilt.

Probehalber legen wir fest: Weicht Z um mehr als 4 vom Wert 10 ab, so betrachten wir das als signifikante Abweichung. Dies führt auf den Verwerfungsbereich:

$$\boxed{V = \{0, 1, \ldots, 5, 15, \ldots, 20\}}$$

In Fig. 8.1 ist V grafisch dargestellt: V besteht aus zwei getrennten Intervallen.

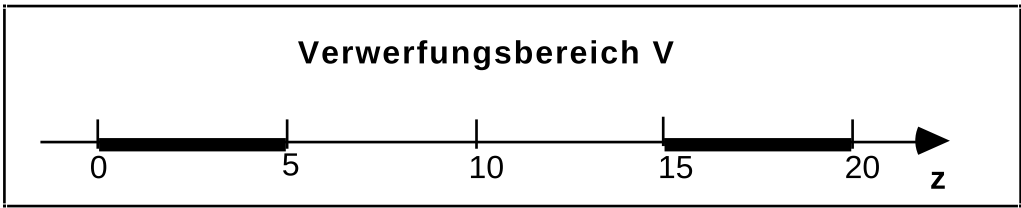

Fig. 8.1: Der Verwerfungsbereich V besteht aus zwei getrennten Intervallen.

Unausgesprochen sind wir bei dieser Festlegung davon ausgegangen, dass auf keiner Parzelle die Erträge genau gleich sind, d. h. dass wir keine Par-zelle weglassen müssen, weil die Differenz 0 nichts zur Entscheidung bei-trägt.

Einseitig - zweiseitig

In den früheren Testbeispielen war der Verwerfungsbereich V meistens ein einziges Intervall in den ganzen Zahlen. Hier ist, wie gesagt, V die Vereinigung von zwei getrennten Bereichen und man spricht deshalb von einem *zweiseitigen* Test. Im Unterschied dazu heissen die früheren Test *einseitig*.

Aufgabe 8.4

Legen Sie einen Verwerfungsbereich für den Test "Geburtsgewicht und IQ" fest. Begründen Sie Ihre Wahl mit den Untersuchungsergebnissen.

Beachten Sie bitte die Ergänzungen zu der gedruckten Lösung.

8.5 Berechnung von α

Wir machen einen Fehler 1. Art, wenn wir die Nullhypothese H_0 verwerfen, obwohl sie zutrifft, d. h. wenn wir zum Schluss kommen, dass die beiden Sorten nicht den gleichen Ertrag abwerfen (H_1), obwohl das Gegenteil (H_0) wahr ist. Wie gross ist die Wahrscheinlichkeit α für diesen Irrtum?

Wenn H_0 zutrifft, wird auf jeder der 20 Parzellen mit der Wahrscheinlichkeit p=0.5 ein "+"-Zeichen erzeugt. Unsere Zufallsgrösse Z ist also binomisch verteilt mit n=20 und p=0.5. Mit der im letzten Kapitel hergeleiteten Formel können wir α somit wie folgt berechnen:

$$
\begin{aligned}
\alpha &= P(Z \in V, \text{ wenn } H_0 \text{ gilt}) \\
&= P(0 \le Z \le 5 \text{ oder } 15 \le Z \le 20, \text{ wenn } p=0.5) \\
&= 1 - P(6 \le Z \le 14, \text{ wenn } p=0.5) \\
&= 1 - \sum_{k=6}^{14} P(Z=k) \\
&= 1 - \sum_{k=6}^{14} \binom{20}{k} p^k \, q^{20-k} \\
&= 1 - 0.5^{20} \sum_{k=6}^{14} \binom{20}{k}
\end{aligned}
$$

Der Taschenrechner liefert:

$$\boxed{\alpha = 0.041 = 4.1\%}$$

Mit dem gewählten Verwerfungsbereich werden wir also im nächsten Oktober entscheiden können, ob sich die beiden Sorten signifikant unterscheiden.

Aufgabe 8.5

a

Beschreiben Sie umgangssprachlich, wie beim Test "Geburtsgewicht und IQ" ein Fehler 1. Art entsteht.

b

Berechnen Sie α für den Verwerfungsbereich V={0,1, 2, 3, 9, 10, 11, 12}.

c

Hängt der IQ signifikant vom Geburtsgewicht ab?

d

Bei welchem zweiseitigen Verwerfungsbereich sind signifikante Entscheidungen möglich?

Zusatzinformation zum Test "Geburtsgewicht und IQ"

Mit dem Vorzeichentest kann also aufgrund der Untersuchungsergebnisse *nicht* geschlossen werden, dass der IQ signifikant vom Geburtsgewicht abhängt. Der Vorzeichentest unterscheidet leider nicht, wie *stark* die IQ der beiden Zwillinge jeweils auseinander liegen.

Es gibt einen verbesserten Vorzeichentest (Rangsummentest von Wilcoxon), der auch die *Beträge* der Differenzen auswertet. Bei diesem Test ergibt sich, dass der IQ signifikant vom Geburtsgewicht abhängt.

Das verwirrt zunächst: Hängt nun der IQ vom Geburtsgewicht ab oder nicht? Beachten Sie aber: Eine Nullhypothese, die beibehalten wurde, ist noch nicht bewiesen, und verschiedene Test führen notwendigerweise in bestimmten Situationen zu anderen Entscheidungen. Der verbesserte Vorzeichentest hat in den meisten Situationen die grössere Macht als der hier besprochene. Man ist daher versucht zu sagen, dass die Nullhypothese nicht zutrifft und der gewöhnliche Vorzeichentest einen Fehler 2. Art begeht. Wenn man diesen Gedanken aber weiterführt, wird man so lange verschiedene Tests durchführen, bis man einmal die gewünschte Entscheidung erhält. Das ist jedoch nicht statthaft. Das Niveau α wird nur eingehalten, wenn man genau einen Test durchführt, und zwar muss man sich im voraus für diesen einen Test entscheiden, ohne die Daten anzuschauen.

8.6 *Berechnung von* β

Der Test soll ja entscheiden, ob die Erträge der beiden Sorten A und B verschieden sind. Es ist nun bekanntlich denkbar, dass die Erträge tatsächlich ungleich sind, dass es aber der Test nicht merkt, weil Z trotzdem nicht in den Verwerfungsbereich V fällt (Fehler 2. Art mit der Wahrscheinlichkeit β).

Berechnen wir β für den Fall, dass die Sorte A nur auf 30% der Parzellen höhere Erträge liefert als Sorte B. Unter dieser Voraussetzung ergibt sich mit der Wahrscheinlichkeit p=0.3 auf jeder der 20 Parzellen ein "+"-Zeichen, die Zufallsgrösse Z ist mit anderen Worten binomialverteilt mit n=20 und p=0.3. Mit diesen Überlegungen erhalten wir für β den Wert:

$$
\begin{aligned}
\beta &= P(\text{"}Z \notin V\text{, wenn } H_1 \text{ gilt"}) \\
&= P(\text{"}6 \le Z \le 14\text{, wenn } p=0.3\text{"}) \\
&= \sum_{k=6}^{14} P(Z=k) \\
&= \sum_{k=6}^{14} \binom{20}{k} p^k \, q^{20-k} \\
&= 58.4\%
\end{aligned}
$$

Die *Macht* des Tests, die beschriebene Ungleichheit aufzudecken, beträgt also nur 41.6%. Man kann diese enttäuschend kleine Macht vergrössern, indem man entweder V oder – noch besser – die Parzellenzahl n vergrössert.

Aufgabe 8.6

a

Beschreiben Sie umgangssprachlich, worin beim Test "Geburtsgewicht und IQ" der Fehler 2. Art besteht.

b

Berechnen Sie für den Verwerfungsbereich V={0, 1, 2, 3, 9, 10, 11, 12} die Wahrscheinlichkeit β für den Fehler 2. Art, wenn Sie annehmen, dass zu 70% der bei der Geburt schwerere Zwilling den höheren IQ hat.

Wenn Sie die bisherigen Aufgaben einigermassen richtig gelöst haben, sind Sie reif für die Lernkontrolle. Wenn Sie diese mit Erfolg bewältigt haben, können Sie sich zum letzten Kapiteltest dieses Leitprogramms anmelden.

Lernkontrolle: Zwei Messmethoden vergleichen

Es stehen zwei Methoden zur Wahl, den Stärkegehalt von Kartoffeln zu bestimmen. Die beiden Methoden wurden verglichen, indem man 16 Kartoffeln halbierte und jeweils beide Methoden anwendete. Es ergaben sich die folgenden Unterschiede des Stärkegehaltes (in %) zwischen den beiden Hälften:

2	0	0	1	2	2	3	-3	1	2	3	0	-1	1	-2	1

Es sind zwei verschiedene Fragen denkbar, die je mit Hilfe eines entsprechenden Vorzeichentests beantwortet werden können.

a

Man hat das Gefühl, die beiden Methoden seien nicht gleichwertig. Prüfen Sie mit einem Vorzeichentest, ob es einen signifikanten Unterschied zwischen den beiden Methoden gibt.

b

Man hat das Gefühl, die erste Methode liefere in der Regel höhere Stärkegehalte als die zweite. Prüfen Sie mit einem (einseitigen!) Vorzeichentest, ob Methode 1 signifikant höhere Werte liefert als Methode 2.

c

Wenn Sie richtig gerechnet haben, sind Sie auf (scheinbar) widersprüchliche Ergebnisse gekommen. Versuchen Sie diesen Widerspruch zu klären.