Wie kommt die statistische Hürde zustande, die die Kandidaten im Erfolgsfall überspringen müssen?

Die „statistische Hürde“ für das Bestehen des Tests bestand in einer gewissen Mindestanzahl von erfolgreichen Test-Durchgängen. Warum das notwendig war, soll im Folgenden erläutert werden.

Prinzipiell können Testvariante und Anzahl der Durchgänge sehr flexibel gehandhabt werden. Um jedoch die Tests einfacher miteinander vergleichen und auswerten zu können, wurde versucht, sie möglichst gleichartig anzulegen. Voraussetzung ist dabei immer, dass der Kandidat mit diesem Versuchsdesign zurechtkommt und ihm zustimmt.

Bei der Berechnung der „statistischen Hürde“ müssen unter anderem zwei Werte beachtet werden, der Alpha- und der Powerwert. Mit Alpha bezeichnet man die Wahrscheinlichkeit, dass der Kandidat den gesamten Test (also alle einzelnen Durchgänge zusammengefasst) besteht, obwohl er die behauptete Fähigkeit gar nicht besitzt. Der Alpha-Wert muss sehr gering sein, denn nur dann bedeutet ein positives Ergebnis, dass der Kandidat aufgrund seiner Fähigkeit und nicht durch Raten bestanden hat. Der 1/Alpha-Wert macht denselben Sachverhalt besser verständlich: Der Kehrwert von Alpha (1 durch Alpha) gibt an, wie oft ein Kandidat, der die behauptete Fähigkeit nicht besitzt, diesen Test durchführen müsste, damit er ihn durchschnittlich ein Mal rein zufällig besteht. Bei einem sehr kleinen Alpha-Wert ergibt sich also ein sehr hoher 1/Alpha-Wert, d.h.: Wenn zum zufälligen Bestehen sehr viele Tests durchgeführt werden müssten, bedeutet ein Bestehen bei unserem Test, dass wahrscheinlich mehr als nur Glück dafür verantwortlich ist.

Für die GWUP-Tests wurden Alpha-Werte von etwa 10^{-10} angesetzt; die 1/Alpha-Werte liegen dementsprechend etwa bei etwa 10 Milliarden. Die Chance, dass ein Kandidat, der über keine besondere PSI-Fähigkeit verfügt, allein durch bloßes Raten (und ohne Schummeln) den Test erfolgreich meistert und die GWUP den Preis irrtümlich zahlen muss, ist damit etwa gleich der Chance, zwei Mal hintereinander einen „Fünfer“ im Lotto „6 aus 49“ zu gewinnen (ca. 10^{-10}). In der normalen Wissenschaft sind solche kleinen Werte für Alpha nicht üblich. Man beachte aber, dass es hier nicht nur um wissenschaftliche Aussagen geht, sondern auch um das Zahlen einer sehr hohen Prämie. Außerdem ist zu bedenken, dass eventuell im Laufe der Zeit recht viele Kandidaten getestet werden möchten. Die Wahrscheinlichkeit, dass die GWUP irgendwann irrtümlich den Preis zahlen muss, vervielfacht sich um die Anzahl dieser Kandidaten.

Die Wahrscheinlichkeit, dass ein Kandidat, der die behauptete Fähigkeit tatsächlich besitzt, den Test besteht, wird als Power bezeichnet. Wir achteten darauf, dass der Power-Wert bei etwa 0,99 liegt. Wenn also ein Kandidat tatsächlich die von ihm behauptete Fähigkeit hat, dann wird diese beim vorliegenden Testdesign auch mit einer Wahrscheinlichkeit von 99% erkannt. Die Power hängt aber auch von der Stärke der Fähigkeit ab, also von der Wahrscheinlichkeit, Treffer zu erzielen. Diese Trefferwahrscheinlichkeit ist von dem Kandidaten selbst einzuschätzen. Der eine vermutet beispielsweise, dass seine Trefferwahrscheinlichkeit 90% beträgt, bei einem anderen sind es vielleicht nur 60%. Der Begriff „Trefferwahrscheinlichkeit“ darf hier nicht mit „Trefferquote“ verwechselt werden. Beim Würfeln ist bekanntlich die Wahrscheinlichkeit, eine Eins zu erhalten, 1/6. Dies ist die Trefferwahrscheinlichkeit. Wenn man jedoch zehnmal würfelt, ist der Anteil der Einsen nicht unbedingt ein Sechstel, er kann auch davon abweichen. Dieser tatsächlich erzielte Anteil wird hier als Trefferquote bezeichnet. Während also die Trefferwahrscheinlichkeit eine feste theoretische Größe darstellt, die eine bestimmte Fähigkeit charakterisiert, ist die Trefferquote das konkrete Ergebnis, das zufallsbedingt von der Trefferwahrscheinlichkeit abweichen kann.

In einem Test mit n Durchgängen muss nun eine Trefferquote k/n erzielt werden (also k Treffer), die zwischen der vom Kandidaten angegebenen und der bei fehlender Fähigkeit zu erwartenden Trefferwahrscheinlichkeit (z.B. 10% bei der 1:10-Variante) liegt. Ist die erzielte Trefferquote mindestens k/n, so wird die Fähigkeit des Kandidaten akzeptiert, andernfalls nicht. Mit diesem Entscheidungsverfahren müssen wir also zwischen der vom Kandidaten behaupteten Trefferwahrscheinlichkeit von z.B. 90% und der ohne Fähigkeiten zu erwartenden Trefferwahrscheinlichkeit von 10% unterscheiden. Wie oben bereits erwähnt, ist die erzielte Trefferquote zufällig, sie entspricht nicht genau der Stärke der Fähigkeit. Daher kann es bei diesem Entscheidungsverfahren Verwechslungen geben: Fehler erster Art („falsch positiv“, d.h. der Kandidat ist nur zufällig erfolgreich) oder Fehler zweiter Art („falsch negativ“, ein fähiger Kandidat hatte zufällig Pech). Natürlich kann mit größerer Durchgangsanzahl n die Trefferwahrscheinlichkeit genauer geschätzt werden (wenn man zehnmal würfelt, wird der Anteil der Einsen mitunter recht stark von 1/6 abweichen. Würfelt man jedoch 10 000 Mal, so liegt dieser Anteil sehr dicht bei 1/6.) Mit höherer Zahl an Durchgängen wird also die Verwechslungsgefahr gesenkt.

Wir haben nun die erforderliche Durchgangsanzahl n und die zu fordernde Mindest-Trefferanzahl k so gewählt, dass solche Verwechslungen möglichst selten auftreten. Das heißt, bei fehlender Fähigkeit (Trefferwahrscheinlichkeit 10% bei der 1:10-Variante) soll eine irrtümliche Akzeptanz der behaupteten Fähigkeit nur mit der sehr kleinen Wahrscheinlichkeit Alpha vorkommen. Andererseits soll mit einer Wahrscheinlichkeit von 99% (Power) die Fähigkeit des Kandidaten erkannt und damit akzeptiert werden, sofern er tatsächlich über die behauptete Trefferwahrscheinlichkeit von 90% verfügt. Sollte der Kandidat aber seine eigene Trefferwahrscheinlichkeit nicht mit 90%, sondern mit 60% angeben, muss die Durchgangsanzahl n vergrößert werden, um Verwechslungen möglichst zu vermeiden. Bei sehr schwach ausgeprägten Fähigkeiten wird daher der Testaufwand unzumutbar hoch. Im Extremfall kann es sogar dazu kommen, dass wir auf den Test verzichten müssen – wir können also nicht beliebig kleine Fähigkeiten testen.

Andererseits bedeutet das, dass ein Kandidat, der seine eigene Trefferwahrscheinlichkeit mit 90% angab und den Test nicht bestand, eventuell trotzdem über die behauptete Fähigkeit verfügt, wenn auch mit bedeutend kleinerer Trefferwahrscheinlichkeit. Der GWUP-Test prüft also die Fähigkeit bezogen auf die angegebene Trefferwahrscheinlichkeit und nicht die Fähigkeit „an sich“.. Wie aus dieser sehr vereinfachten Darstellung ersichtlich wird, ist die genaue Festlegung der geforderten Treffer ein Abwägen zwischen Strenge und Fairness: Ein kleiner Alpha-Wert bewahrt die GWUP davor, einem Kandidaten besondere Fähigkeiten zu attestieren, ob-

wohl er einfach nur Glück hatte, und ein hoher Power-Wert bewahrt den Kandidaten davor, ungerechterweise mit dem Etikett „erfolglos“ belegt zu werden.

Wie wird aber nun der erforderliche Versuchsumfang ermittelt? Dazu erinnern wir zunächst an die aus der Theorie der Binomialverteilung bekannte Tatsache, dass die Wahrscheinlichkeit, bei n Versuchen gerade k Treffer zu erzielen, durch

$$b(p,n,k) = \frac{n!}{k!*(n-K)!}\, p^k (1-p)^{n-k}$$

gegeben ist. Dabei bezeichnet p die Trefferwahrscheinlichkeit und n! (sprich: n Fakultät) das Produkt der Zahlen 1 bis n. Die Wahrscheinlichkeit, mindestens k Treffer zu erzielen, ist dann gerade

$B(p, n, k)$ = Summe der $b(p, n, i)$ von $i = k$ *bis* $i = n$.

Zur Versuchsplanung nahmen wir zunächst, wie in wissenschaftlichen Studien üblich, die Null-Hypothese an, dass also keine besondere Fähigkeit vorliegt. Die Trefferwahrscheinlichkeit hieße dann p_0 ($p_0 = 0{,}1$ bei der Variante 1:10 bzw. $p_0 = 0{,}5$ bei der Variante 1:2). Wir beginnen dann mit einem nicht zu großen n-Wert und suchen den kleinsten k-Wert, für den $B(p_0, n, k)$ nicht größer als unser vorgegebenes Alpha ist. Wenn wir also bei einem Versuch mit *n* Durchgängen mindestens *k* Treffer verlangen, so wäre bei Gültigkeit der Nullhypothese die Wahrscheinlichkeit, diese Forderung per Zufall zu erfüllen, nicht größer als Alpha. Natürlich muss auch der Kandidat eine faire Chance haben. Nehmen wir also an, dass er die behauptete Fähigkeit tatsächlich besitzt. Diese Annahme bezeichnet man als Alternativhypothese. Die vom Kandidaten behauptete Trefferwahrscheinlichkeit sei p_a. Für die oben genannte Durchgangsanzahl *n* und die Mindesttrefferanzahl *k* erhält man dann die Power aus $B(p_a, n, k)$. Vermutlich wird diese Power noch nicht dem geforderten Wert von 0,99 entsprechen. Daher müssen wir diese Berechnung wiederholen, und zwar mit einer größeren Durchgangsanzahl n. Dieses n wird solange vergrößert, bis $B(p_a, n, k)$ etwa 0,99 beträgt. Die so erhaltenen Werte für *n* und *k* legen den Versuchsplan fest.

Die zwei Teststufen

Die GWUP führt Tests in zwei weniger „strengen“ Stufen durch, von denen beide bestanden werden müssen. Das Gesamtverfahren weist dann die gewünschte Strenge auf. Dies hat den praktischen Vorteil, dass nach Nichtbestehen der ersten Stufe abgebrochen und somit der Aufwand reduziert werden kann. Diese erste Stufe bezeichnen wir als *Vortest*, und ihr Bestehen ist Voraussetzung für die zweite Stufe, den *Haupttest*, dessen Bestehen Vorbedingung für den Randi-Test ist. Zur Planung dieser beiden Stufen wählen wir zwei Alpha-Werte Alpha1 und Alpha2 so, dass das Produkt Alpha1·Alpha2 unseren vorgegebenem Alpha-Wert ergibt und beide Werte ähnlich groß sind. Entsprechend wählen wir zwei Power-Werte Power1 und Power2, so dass Power1·Power2 = 0,99 gilt.

Die Versuchspläne

Bei der 1:2-Variante ($p_0 = 0{,}5$) haben wir angenommen, dass die Kandidaten ihre eigene Trefferwahrscheinlichkeit mit mindestens 90% ($p_a = 0{,}9$) einschätzen, was auch der Fall war (andernfalls hätten wir einen neuen Versuchsplan durchrechnen müssen). Dabei müssen wir berücksichtigen, dass hier n der Anzahl der zu testenden Gläser entspricht, und bei nicht vorhandener PSI-Fähigkeit die Trefferrate 50% beträgt. Für das Bestehen des Tests wird gefordert, dass mindestens *k* Gläser richtig beurteilt wurden. Der ermittelte Wert n wurde hier so gerundet, dass er durch 10 teilbar ist, damit man diesen Test in Durchgängen mit je zehn Gläsern durchführen kann. Die Werte Alpha, Power und k wurden entsprechend korrigiert. Damit erhielten wir nach obiger Methode für die zwei Stufen folgende Pläne:

Versuchspläne für die Variante 1:2 ($p_0 = 0{,}5$, $p_a = 0{,}9$)

	n	k	Alpha	1/Alpha	Power
Erster Teiltest:	50	40	0.000011931	83815	0.99065
Zweiter Teiltest:	60	46	0.000021119	47350	0.99933
Gesamt:	110	86	$2.5196*10^{-10}$	3968883950	0.98999

Für die Variante 1:2a, bei welcher in jedem der fünf Durchgänge genau fünf Gläser behandelt waren, ist zur Berechnung der Wahrscheinlichkeiten die Theorie der hypergeometrischen Verteilungen anzuwenden. Für den ersten Teiltest erhält man dann Alpha = 0,000054 und Power = 0,9996. Diese Variante war also für die Kandidaten günstiger.

Aus früheren Versuchen ist bekannt, dass die 1:10-Variante ($p_0 = 0{,}1$) als schwerer empfunden wird. Wir haben deswegen zur Planung nur eine Trefferwahrscheinlichkeit von $0{,}9^2 = 0{,}81$ statt 0,9 angenommen, auch wenn der Kandidat die eigene Trefferwahrscheinlichkeit (zu seinen Ungunsten) höher eingeschätzt hat. Damit erhielten wir nach obiger Methode für die zwei Stufen folgende Pläne:

Versuchspläne für die Variante 1:10 (p0 = 0,1 , pa = 0,81)

	n	k	Alpha	1/Alpha	Power
Erster Teiltest:	13	7	0.000099285	10072	0.993
Zweiter Teiltest:	18	10	0.000002046	488998	0.996575
Gesamt:	31	17	2.0316*10-10	4922228785	0.98877

Dr. Volker Guiard