

2. Testen von Hypothesen

Signifikanztest (Normalverteilung)

Einführendes Beispiel:

Einer Kandidatin werden n Fragen mit zwei Auswahlantworten vorgelegt, von denen genau eine richtig ist (einfacher Multiple choice-Test). Es soll geprüft werden, ob die Kandidatin eine Könnlerin ist ($p > 0.5$) oder nur rät ($p = 0.5$).

Hypothese H_1 :	Die Kandidatin ist eine Könnlerin	$p > 0.5$ (Test einseitig)
Nullhypothese H_0 :	Die Kandidatin rät nur	$p = 0.5$

Planung eines Experiments:

Es werden $n = 100$ Fragen gestellt

Testgrösse X : Anzahl der richtigen Antworten.

Die Zufallsvariable X ist binomialverteilt mit dem Erwartungswert $\mu = 50$ und der Standardabweichung $\sigma = 5$. Da die Faustregel erfüllt ist, kann die Binomialverteilung durch die Normalverteilung approximiert werden.

Beim gebräuchlichen Signifikanzniveau $\alpha = 5\%$, ist u so zu wählen, dass $\Phi(u) = 0.95$ gilt. Für die kritische Grenze erhält man damit: $1.645 \cdot \sigma = 1.645 \cdot 5 = 8.225$

Entscheidungsregel (einseitiger Test):

Beantwortet die Kandidatin mindestens $\mu + 1.645 \cdot \sigma \approx 50 + 8.2 \approx 59$ Fragen richtig, so stuft man sie als Könnlerin ein.

Fehlermöglichkeiten:

Fehler 1. Art:

Die Nullhypothese wird abgelehnt, obwohl sie richtig ist.

Im Beispiel bedeutet dies: Eine ratende Kandidatin beantwortet, obwohl sie nur rät, mindestens 59 Fragen richtig und wird damit fälschlicherweise als Könnerin eingestuft.

Fehler 2. Art:

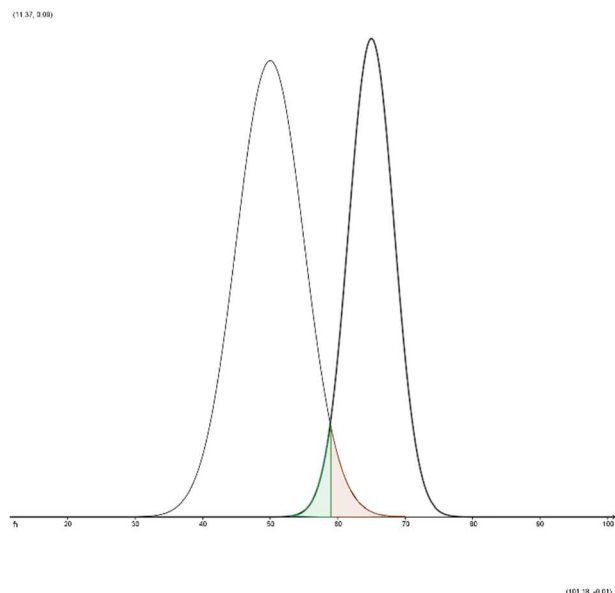
Die Nullhypothese ist falsch, wird aber fälschlicherweise beibehalten.

Im Beispiel bedeutet dies: Eine Könnerin (z.B. $p = 0.65$) hat Pech, beantwortet zu wenig Fragen richtig und wird deshalb als Raterin eingestuft.

In der Abbildung sind die beiden Wahrscheinlichkeiten für den Fehler 1. und 2. Art (unter der Annahme $p = 0.65$) zu erkennen.

Die Wahrscheinlichkeit, dass eine ratende Kandidatin die Prüfung besteht (d.h. mindestens 59 Fragen richtig beantwortet ist) höchstens 5%.

Die Wahrscheinlichkeit, dass eine Kandidatin mit $p = 0.65$ die Prüfung nicht besteht (d.h. weniger als 59 Fragen richtig beantwortet) ist 8.8%.



Die beiden Fehler in zwei anderen Kontexten:

Fehler erster und zweiter Art		
	Schuldig	unschuldig
Schuld-spruch	Richtig	Justizirrtum zu Ungunsten des Angeklagten
Freispruch	Justizirrtum zu Gunsten des Angeklagten	Richtig

(Bohley, 1992, S. 568 zitiert nach Conzelmann, 1999, S. 251)

Fehler erster und zweiter Art		
	H0 wahr	H0 falsch
H0 abgelehnt	Fehler 1. Art (α -Fehler, Irrtumswahrscheinlichkeit) $\Rightarrow$ Trainingsverfahren überschätzt	richtige Entscheidung ($1 - \beta$, Power) $\Rightarrow$ Beleg für besseres Trainingsverfahren
H0 beibehalten	richtige Entscheidung $1 - \alpha$ $\Rightarrow$ Beleg, dass Trainingsverfahren nicht besser	Fehler 2. Art (β -Fehler) $\Rightarrow$ Trainingsverfahren unterschätzt

Je nach den möglichen Auswirkungen des Fehlers wird man versuchen den Fehler 1. oder 2. Art oder beide zu minimieren (Stichwort: Pilzsammeln).

Signifikanztest (Binomialverteilung)

Das folgende Beispiel stammt von Sir Ronald Aylmer Fisher (1890-1962), der als Begründer der modernen mathematisch orientierten Statistik gilt:

Die englische Lady trinkt ihren Tee stets mit einem Zusatz Milch. Eines Tages verblüfft sie ihre Teerunde mit der Behauptung, sie könne allein am Geschmack entscheiden, ob zuerst die Milch oder zuerst der Tee eingegossen worden sei. Sie behauptet, dass sie im Vergleich zum blinden Raten öfter die richtige Eingiess-Reihenfolge treffen würde. Wie könnte die Behauptung der Lady überprüft werden

Planen eines Zufallsexperiments

Die Lady erhält $n = 20$ -mal je 2 Tassen Tee überreicht, jeweils eine vom Typ Milch vor Tee, die andere vom Typ Tee vor Milch. Die Reihenfolge dieser beiden Tassen wird durch den Wurf einer L-Münze festgelegt

Hypothese H_1 : Die Lady hat die behauptete Fähigkeit, d.h. ihre Trefferwahrscheinlichkeit p ist grösser als $\frac{1}{2}$

Nullhypothese H_0 : Die Lady rät nur: $p = \frac{1}{2}$

Entscheidungsregel: sei X die Anzahl der Treffer

$X \geq 14$ Entscheid für H_1

$X < 14$ Entscheid für H_0

Fehler 1. Art:

H_0 wird verworfen, obwohl sie richtig ist (d.h. man attestiert der Dame eine Fähigkeit, die nicht vorhanden ist (die Dame erzielt also mindestens 14 Treffer durch Raten))

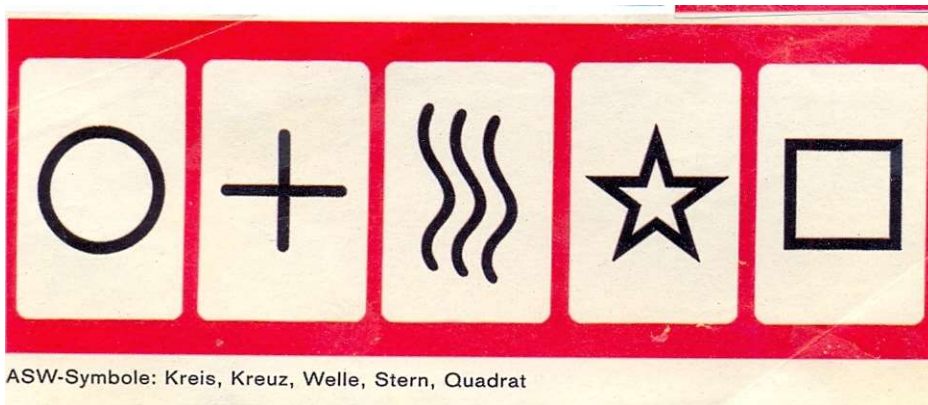
Fehler 2. Art.

H_0 wird beibehalten, obwohl H_1 richtig ist. Die Dame hat tatsächlich die behauptete Begabung, erzielt aber weniger als 14 Treffer.

Die Entscheidungsregel wird oft so gewählt, dass die Wahrscheinlichkeit α des Fehlers erster Art kleiner als 0.05 ist. Ein solcher Test heisst Signifikanztest zum Signifikanzniveau $\alpha = 5\%$. Bei dieser Vorgehensweise wird der Fehler 1. Art im Vergleich zum Fehler 2. Art als schwerwiegender erachtet. Die meisten Tests werden in der Hoffnung auf eine signifikante Ablehnung der Nullhypothese angelegt. Die Nullhypothese behauptet die Abwesenheit eines Effekts, den man eigentlich nachweisen möchte!

Ein Beispiel aus der Parapsychologie (ESP: Extra Sensory Perception)

Gibt es Menschen, die über eine Begabung verfügen, die sich naturwissenschaftlich nicht erklären lässt?



Der amerikanische Parapsychologe B. Rhine (1895 bis 1980) hat folgendes Experiment vorgeschlagen:

Die Testperson muss versuchen, eine aus 5 verschiedenen Karten ausgewählte Karte zu identifizieren, ohne sie gesehen zu haben.

Hypothese H_1 : Die Person verfügt über eine besondere Begabung: $p > 0.2$

Nullhypothese H_0 : Die Person hat keine besondere Begabung, rät nur: $p = 0.2$

2. Planen eines Experiment:

Es werden 10 Versuche durchgeführt. Die Testgröße Z : Anzahl der richtig geratenen Karten ist binomialverteilt.

3. Signifikanzniveau:

Eine Person, die nur rät, wird im Mittel bei 10 Versuchen 2 Treffer erzielen. Wir werden einer Person erst dann besondere Begabung zuerkennen, wenn Z deutlich (man sagt auch signifikant) höher liegt.

Entscheidungsregel:

$Z \leq 3 \rightarrow H_0$ beibehalten

$Z > 3 \rightarrow H_0$ ablehnen, Entscheid für H_1

4. Fehler 1. Art:

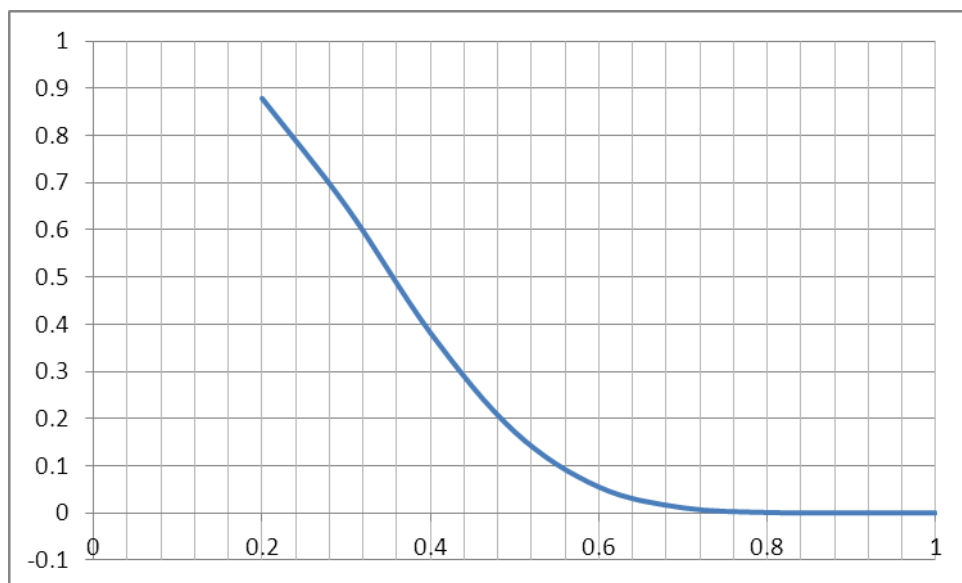
Wahrscheinlichkeit, dass einer "normalen" Person irrtümlich besondere Begabung zugeschrieben wird ($p = 0.2$, aber $Z > 3$) die zugehörige Wahrscheinlichkeit α ist 3.3%

Fehler 2. Art:

Wahrscheinlichkeit bei einer Person die besondere Begabung nicht anzuerkennen.

Das Risiko 2. Art hängt von Trefferwahrscheinlichkeit der Testperson ab:

Je grösser der Grad ausser sinnlicher Begabung ist (das heisst je grösser p ist) umso kleiner ist die Wahrscheinlichkeit, diese besondere Begabung nicht festzustellen.



Verschärft man die Entscheidungsregel ($Z > 5$) so verkleinert sich zwar der Fehler 1. Art, aber der Fehler 2. Art vergrössert sich.

Will man beide Fehler verkleinern, dann bedingt dies eine Erhöhung der Anzahl Versuche.

Zusammenfassung Signifikanztest:

1. Hypothese H_1 :

Aufgrund der Daten ergibt sich eine bestimmte Vermutung über den unbekannten Parameter einer Wahrscheinlichkeitsverteilung (z.B. über die Erfolgswahrscheinlichkeit bei einem Bernoulli-Versuch).

2. Nullhypothese H_0 . Gegenhypothese zu H_1 :

Die Nullhypothese behauptet Abwesenheit eines Effekts, den man eigentlich nachweisen möchte.

3. Planen eines Zufallsexperiments, Festlegen einer Entscheidungsregel

4. Mögliche Fehlentscheide:

Fehler 1. Art: H_0 wird verworfen, obwohl H_0 richtig ist. Die zugehörige Wahrscheinlichkeit wird mit α bezeichnet, α heisst Signifikanzniveau. Übliches Signifikanzniveau ist $\alpha = 5\%$.

Fehler 2. Art: H_0 wird nicht verworfen, obwohl H_0 falsch ist.

Weitere Beispiele für Signifikanztests:

a)

Ist das neue Medikament besser als das alte?
Ist ein konkreter Würfel gefälscht?

Nullhypothese: Sie sind beide gleich gut
Nullhypothese: Es handelt sich um einen Laplace-Würfel.

Statistische Qualitätskontrolle:

Der Ausschussanteil ist zu gross?

Die Lieferung ist in Ordnung

3. Beispiel eines Anpassungstests : Chiquadrattest

Die Güte der Übereinstimmung von Daten mit einer bestimmten Wahrscheinlichkeitsverteilung kann mit dem sogenannten Chi-Quadrattest geprüft werden.

Beispiele:

Sind die Augenzahlen eines bestimmten Würfels gleichverteilt?

Ist die Anzahl der Mädchengeburten in Vierkinderfamilien binomialverteilt?

Sind die Widerstände normalverteilt?

Nach Pearson wird folgende Testgrösse verwendet:

$$\chi^2 = \sum_{i=1}^k \frac{(b_i - e_i)^2}{e_i}$$

wobei b_i bzw. e_i die beobachteten bzw. erwarteten Häufigkeiten bedeuten. Zu grosse bzw. zu kleine Werte der Testgrösse sind verdächtig vgl. dazu die Tabellen der kritischen Werte in der Formelsammlung.

Beispiel:

Ist ein bestimmter Würfel ein Laplacewürfel?

Der Würfel wird $n = 542$ -mal geworfen:

Augenzahl X	1	2	3	4	5	6
beob. abs. Häufigkeit b_i	78	91	86	92	90	105
erwartete Häufigkeit e_i	$\frac{542}{6}$	$\frac{542}{6}$	$\frac{542}{6}$	$\frac{542}{6}$	$\frac{542}{6}$	$\frac{542}{6} \approx 90.3$

Der gerechnete Wert von $\chi^2 = 4.3$ ist mit dem kritischen Wert zu vergleichen. Dieser hängt von der Anzahl k der Klassen ($k = 6$) bzw. von $v = k - 1 = 5$ ab, v heisst Freiheitsgrad. In der Tabelle liest man als kritischen Wert 11.1 bei einem Signifikanzniveau von 5% ab. Da der gerechnete Wert $\chi^2 = 4.3 < 11.1$ ist, kann der Laplacewürfel als unverdächtig eingestuft werden.

Zur Beurteilung können auch Erwartungswert und Standardabweichung der Chiquadratverteilung zugezogen werden. Es gilt nämlich

für den Erwartungswert: $\mu = k - 1 = v$ Freiheitsgrad

für die Varianz $\sigma \approx 2 \cdot (k - 1) = 2v$

im Beispiel also $\mu = 5$ bzw. $\sigma = 10$

Entscheidungsregel:

Weicht der Wert von Chiquadrat um mehr als 2 Standardabweichungen vom Erwartungswert ab, dann gilt die Abweichung als auffällig und man würde in diesem Fall annehmen, dass der Würfel kein Laplacewürfel ist.

Die kritischen Werte findet man in der Formelsammlung tabelliert.

Wie schon erwähnt sind auch zu kleine Werte von Chiquadrat auffällig

Beispiel:

Hat Mendel Daten manipuliert?

Mendels Theorie sagt aus: Beim Kreuzen von gelben runden mit grünen eckigen Erbsen gibt es vier Nachkommen im Verhältnis

	9	:	3	:	3	:	1
	gelb/rund		gelb/eckig		grün/rund		grün eckig
beobachtete Häufigkeit b_i :	315		101		108		32
erwartete Häufigkeit e_i :	$\frac{9}{16} \cdot 556$		$\frac{3}{16} \cdot 556$		$\frac{3}{16} \cdot 556$		$\frac{1}{16} \cdot 556$
	312.75		104.25		104.25		34.75

$$\chi^2 = \frac{(315 - 312.75)^2}{312.75} = \frac{(108 - 104.25)^2}{104.25} = \frac{(101 - 104.25)^2}{104.25} = \frac{(32 - 34.75)^2}{34.75} = 0.470$$

Freiheitsgrad $v = k - 1 = 3$

Der kritische Wert 7.81 bei einem Signifikanzniveau von 5% ist weit unterschritten. Es handelt sich um einen verdächtig kleinen Wert, denn aus der Tabelle schätzt man ab, dass in etwa 93% aller Fälle grössere Werte von Chiquadrat auftreten.

Die neueste Deutung der zu guten Werte: Mendel hat jedes Mal so lange Material gesammelt, bis seine Werte nahe bei den Erwartungswerten lagen!

Mit dem Chiquadrattest können die Daten bei der Untersuchung der Widerstände im ersten Abschnitt überprüft werden:

Messdaten		Werte der NV mit Erw.wert 15.11 Standabw. 0.052		Chiquadrat
<i>Klasse</i>	<i>Häufigkeit</i>			
15	3	2		0.95
15.025	1	3		1.68
15.05	6	7		0.24
15.075	16	13		0.91
15.1	20	17		0.41
15.125	17	19		0.21
15.15	18	17		0.12
15.175	11	12		0.02
15.2	5	6		0.30
15.225	1	3		1.18
und größer	2	1		0.31
Summe	100	100		6.34

15.11 Erwartungswert

(Schätzung durch emp.Mittelwert der Daten)

0.052 Standardabweichung

(Schätzung durch emp. Standardabweichung)

Anpassungstest

Chiquadrat

Chiquadrat aus den Daten 6.34

kritisches Chiquadrat Signifikanzniveau 5%

Freiheitsgrad 10 (11 Klassen)
aus Tabelle 18.3

Da der Chiquadratwert 6.34 kleiner als der kritische Wert 18.3 ist, kann die Hypothese, es liege eine Normalverteilung vor beibehalten werden.

Die Übereinstimmung einer Verteilung mit der Normalverteilung kann auch graphisch überprüft werden.