Signifikanztest

Grundlagen zum Verständnis

Heinz Tüchler

Vorbemerkungen:

wichtige Bereiche der klinischen Statistik:

- Studienplanung
- Studienauswertung:
  - Beschreibung der Daten
  - Exploration von Datenstrukturen
  - Schätzen von Kenngrößen aus einer Stichprobe

Stichprobe = Begrenzte Zahl von Objekten, die stellvertretend für eine (oft als unendlich gedachte) Grundgesamtheit untersucht werden.

  - Prüfen von Hypothesen

Hypothese = Annahme über eine Kenngröße in der Grundgesamtheit

Die klinische Statistik kommt bei der Studienplanung sowie bei der Sudienauswertung zur Anwendung. Die Hypothesenprüfung ist ein Teil der Auswertung. Das sei betont, da andere Bereiche, vor allem das Explorieren und das Schätzen, verglichen mit dem Hypothesenprüfen, zu wenig Aufmerksamkeit erhalten.

Signifikanztest ist eine Form der Prüfung von Hypothesen

Hypothesen werden vor allem mit Signifikanztests geprüft. Daneben gibt es aber auch sequentielle Testverfahren, und Bayessche Statistik.

Im Folgenden soll die allgemeine Grundvorstellung eines Signifikanztests dargestellt werden.

Welches Merkmal mit Wahrscheinlichkeit 50:50 eignet sich als Demonstrationsbeispiel?

Dazu suchen wir ein einfaches Beispiel. Wir wollen von einem Merkmal ausgehen, das nur zwei Ausprägungen hat. Diese sollen "im Normalfall" gleich wahrscheinlich sein. Beispiele dafür wären die Tumorlokalisation - eingeteilt in "rechts" und "links", das Geschlecht eines Neugeborenen, das Auftreten einer Krankheit im Sommer oder im Winter usw.
Natürlich kann in all diesen Fällen bereits Vorwissen bestehen, das der Annahme gleicher Wahrscheinlichkeit entgegensteht.

Wie drückt man Wahrscheinlichkeiten aus?

Wahrscheinlichkeit kann verschieden ausgedrückt werden. Umgangssprachlich verwendet man gerne ein Verhältnis (z.B.: "die Chancen stehen 10 zu 1") oder Prozent (z.B.: "Fremdverschulden kann mit 99% ausgeschlossen werden").
In der Statistik wird Wahrscheinlichkeit vorwiegend in Anteilen von 1 angegeben; statt 100% sagt man p=1, statt 37% p=0.37. Oft wird auch noch die Null vor dem Komma weggelassen (statt p=0.05, p=.05).
Außerdem werden für die Notation von Parametern einer Grundgesamtheit gerne, aber nicht immer, griechische Buchstaben verwandt (statt p=0.5, π=0.5).

Wie kann man einen Zufallsprozeß im Urnenmodell darstellen?

Die oben genannten Beispiele kann (muß?) man als Zufallsprozesse betrachten. Aus der Sicht des einzelnen Patienten ist es dem Zufall überlassen, ob er rechts oder links einen Tumor hat. Dabei handelt es sich nicht zwingend um einen Zufall im philosophischen Sinn, sondern um einen Zufall "mangels besseren Wissens".

Manche Menschen (z.B. der Verfasser) halten es für anschaulich Zufallsprozesse, wie die oben genannten durch physikalisch nachvollziehbare Vorgänge darzustellen. Eine Möglichkeit dafür ist das sogenannte Urnenmodell.

Man nimmt einen Behälter (die Urne) und gibt z.B. verschiedenfarbige Kugeln hinein. Jeder Farbe symbolisiert ein Ereignis und das unbesehene Herausnehmen einer Kugel gleicht dann dem Zufallsprozeß. Merkmale mit zwei Ausprägungen wird man durch zwei Farben darstellen.

Wie sieht das Modell für den Fall - π(rot)=0.5 aus?

Soll das Modell einen Vorgang mit zwei gleichwahrscheinlichen Ausprägungen darstellen, dann gibt man gleich viele Kugeln jeder Farbe in die Urne. Die Gesamtzahl spielt dabei keine Rolle. Jedoch darf immer nur eine Kugel entnommen werden. Sie muß vor der nächsten Entnahme wieder zurückgelegt werden.

Was erwartet man, wenn man 10 Kugeln zieht?

Man erwartet sich etwa 50% rote und 50% blaue Kugeln zu ziehen.

Beispiel - Mehr Milchkühe:

- angenommener Normalfall: etwa gleich viele weibliche wie männliche Rinder
- Interesse an mehr Milchkühen

Für die nächsten Überlegungen wollen wir als Beispiel das Geschlecht von Rindern verwenden. Wir nehmen an, daß im Normalfall etwa gleich viele weibliche wie männliche Rinder geboren werden. Und wir nehmen weiters an, daß jemand Interesse an mehr Milchkühen hätte und deshalb nach Mitteln sucht, das Geschlechtsverhältnis dahingehend zu beeinflussen.

Wie kann das Beispiel im Urnenmodell dargestellt werden?

Der tierische Normalfall könnte im Urnenmodell durch eine Urne mit gleich vielen blauen (Stiere) wie roten (Kühe) Kugeln dargestellt werden.

- Angeblich wirksame Substanzen werden gefunden

Nun nehmen wir an, es existierten 6 angeblich wirksame Substanzen.

Dazu muß man klären, was "wirksam" bedeuten soll. Als wirksam würde man eine Substanz dann betrachten, wenn sie die Wahrscheinlichkeit weiblicher Geburten erhöhte. Sie müßte deshalb nicht ausschließlich zu weiblichen Geburten führen.

Wie stellt man Wirksamkeit im Urnenmodell dar?

Im Modell stellt man die Wirksamkeit durch mehr rote als blaue Kugeln in der Urne dar.

- Milchkuh-Versuch wird mit 6 Substanzen durchgeführt

Wir nehmen für jede Substanz eine Urne, ohne die Zahl der darin enthaltenen roten und blauen Kugeln zu kennen und führen einen Versuch durch.

Versuchsplan: 10 mal 1 Kugel ziehen

Der Versuch besteht im zehnmaligen Ziehen einer Kugel.

Welche Daten brauchen wir?

Spielt die Reihenfolge der gezogenen roten und blauen Kugeln eine Rolle?

Zur Auswertung des Versuches brauchen wir nur für jede Urne die Zahl roter Kugeln. Die Reihenfolge ist für die Fragestellung bedeutungslos (; sie könnte bestenfalls das Modell an sich in Frage stellen).

Angenommen wir hätten folgende Ergebnisse erhalten:

Ergebnisse:	1	2	3	4	5	6	7	8	9	10	Anzahl rot	Anzahl blau
Substanz 1	r	b	b	b	r	b	b	r	b	r	4	6
Substanz 2	r	b	r	r	r	b	b	r	b	r	6	4
Substanz 3	r	b	r	b	r	r	b	r	b	b	5	5
Substanz 4	b	b	b	b	r	r	r	b	b	b	3	7
Substanz 5	r	r	r	r	r	r	r	r	r	r	10	0
Substanz 6	r	r	r	r	r	r	r	b	b	b	7	3

- Welche Fragen stellen sich zu den Ergebnissen?

Wir würden uns fragen, welche Substanzen wirksam sind, d.h. in welchen Urnen sich mehr rote als blaue Kugeln befinden.
Weiters würden wir uns nach der Stärke der Wirkung, also nach dem Verhältnis von roten und blauen Kugeln in jeder Urne fragen.
Wir suchen eine Regel, nach der wir für oder gegen "Wirkung" entscheiden.
Außerdem sind wir uns nicht im Klaren, wie man das Ergebnis von Substanz 4 bewerten sollen. Hier haben wir entgegen unserer Erwartung weniger als die Hälfte rote Kugeln gezogen.

Was war das Ziel des Versuches?

Das Ziel des Versuches war die Entscheidung zwischen den in Frage kommenden Hypothesen, nämlich "es besteht keine Wirkung" (d.h. es sind gleichviele rote wie blaue Kugeln in der Urne) oder "die Substanz fördert weibliche Geburten" (d.h. es sind mehr rote als blaue Kugeln in der Urne).

Wie beschreibt man die in Frage kommenden Hypothesen in der üblichen formalen Notation?

- Nullhypothese: H_0: $\pi = 0.5$
- Alternativhypothese: H_A:$\pi > 0.5$

Nullhypothese (H_0): Eine Hypothese, die man ohne besonderen Beleg in allgemeiner Übereinstimmung für zutreffend hält.

Alternativhypothese (H_A): Eine "fordernde" Hypothese; sie wird nicht von vornherein und ohne Beweis für wahr gehalten.
Anmerkung: Statt "H_A" wird oft die Notation "H_1" verwandt.

Kann man eine beliebige Hypothese als Nullhypothese verwenden?

Im Allgemeinen gibt es einen Konsens, was eine zulässige Nullhypothese und was eine entsprechende Alternativhypothese ist. In manchen Fällen ist das aber eher eine Übereinkunft als eine inhaltliche Selbstverständlichkeit.

Welche Ergebnisse sind bei Wirkungslosigkeit zu erwarten?

Welche Ergebnisse sind möglich?

Sind alle Ergebnisse gleich wahrscheinlich?

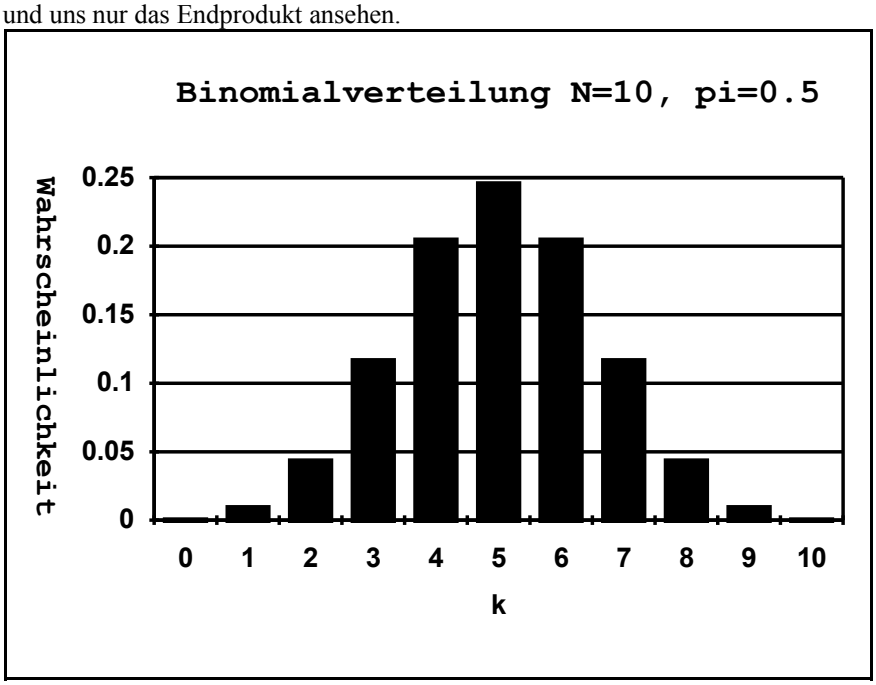
Grundsätzlich ist jedes Ergebnis zwischen keiner und zehn roten Kugeln möglich, aber es ist intuitiv klar, daß unter der Voraussetzung gleich vieler roter wie blauer Kugeln in der Urne (=Wirkungslosigkeit) nicht alle gleich wahrscheinlich sind. Wir wollen nun die Intuition durch Überlegung stützen.

Wie bestimmt man die Wahrscheinlichkeit der möglichen Ergebnisse?

- Grenzwert der relativen Häufigkeit d.h. Ergebnisse sehr vieler Versuche
- Berechnung

Man kann sich auf zwei Arten ein Bild über die Wahrscheinlichkeit jedes möglichen Ergebnisses machen.

Erstens könnte man eine Urne mit gleich vielen roten und blauen Kugeln nehmen und "sehr oft" den Versuch zehn Kugeln zu ziehen wiederholen. Dann erhielte man die Wahrscheinlichkeit jedes Ergebnisses von 0 roten bis 10 roten Kugeln als Grenzwert der relativen Häufigkeit dieser Resultate.
Zweitens kann man die Wahrscheinlichkeit jedes möglichen Ergebnisses berechnen. Hier werden wir allerdings nicht darauf eingehen, wie man das macht, und uns nur das Endprodukt ansehen.



Unserem Beispiel entspricht eine mathematische Funktion, die sogenannte Binomialverteilung - unter der Nullhypothese jene mit N=10 (=10mal eine Kugel ziehen) und π=0.5 (=50% rote Kugeln in der Urne). Man schreibt das BN (N=10, π=0.5). Im obigen Histogramm ist die Wahrscheinlichkeitsfunktion der Binomialverteilung mit diesen Parametern dargestellt.

Erklärung der Wahrscheinlichkeitsfunktion:

Wie wahrscheinlich ist k = 5 ?
Wie wahrscheinlich ist k ≠ 5 ?
Wie wahrscheinlich sind k=10, k=9, k=8?

Die Höhe einer Säule entspricht der Wahrscheinlichkeit des Ergebnisses, k bezeichnet die Zahl roter Kugeln. Das Ergebnis k=5 hat eine Wahrscheinlichkeit von p=.246 d.h. in rund einem Viertel der Versuche wird man fünf rote von zehn Kugeln erhalten. Das heißt aber auch, daß in rund drei Viertel der Versuche nicht 5 rote und 5 blaue Kugeln gezogen werden, obwohl in der Urne ein Verhältnis von 1 zu 1 besteht.
Zehn rote Kugeln (k=10) zieht man allerdings durchschnittlich nur in einem von tausend Versuchen denn p(k=10)=.001 (sprich: p für k ist 10 ist .001).
Auch 9 von 10 Kugeln sind nur in einem von hundert Versuchen rot.
Acht rote Kugeln sollte man in rund vier von hundert Versuchen erhalten.
Je näher das mögliche Ergebnis dem wahren Verhältnis in der Urne kommt, desto wahrscheinlicher tritt es auch auf.

Was setzt man bei der Berechnung der Wahrscheinlichkeitsfunktion voraus?

- Gültigkeit des Modells
- Gültigkeit von H_0

Zur Erinnerung: die Berechnung setzt sowohl die Gültigkeit des Modells, als auch die Nullhypothese voraus.

Wie bewertet man nun die Ergebnisse des Milchkuh-Experiments?

Die Ergebnisse der Urnen (Substanzen) 1,3 und 4 weisen nicht auf mehr rote als blaue Kugeln in der jeweiligen Urne hin, bei den anderen drei wurden aber mehr rote als blaue Kugeln gezogen.
Allerdings ersehen wir aus der Grafik, daß das Ergebnis 6 rote Kugeln eine Wahrscheinlichkeit von .20 besitzt d.h. in etwa einem von 5 Versuchen zu er-

warten ist. Auch das Ergebnis 7 rote Kugeln ist nicht allzu selten. Es tritt durschschnittlich in einem von zehn Versuchen auf.

Will man bei k=10 noch an H_0 glauben?

Bedenkt man, daß das Ergebnis von Urne 5 (10 rote 0 blaue Kugeln) unter der Nullhypothese nur in einem von tausend Versuchen auftreten sollte so ist man geneigt nicht an die Nullhypothese zu glauben, sondern zu vermuten, daß in dieser Urne mehr rote Kugeln sind.

Wollte man auch hier die Nullhypothese beibehalten, hätte man sich den Versuch sparen können.

Wenn man die extremsten 5% an möglichen Ergebnissen als kritisch betrachtet, welche k-Werte fallen dann in diese kritische Region?

Betrachtete man die 5% extremsten möglichen Ergebnisse als kritisch für die Nullhypothese, dann würde das (großzügig gerechnet) jene mit k=8, k=9 und k=10 umfassen, da die Summe der Wahrscheinlichkeiten 0010+0.0098+0.0439=0.0547 d.h. rund 5% ergibt.

Wenn man die möglichen Ergebnisse acht, neun oder zehn rote Kugeln als Anlaß zur Ablehnung von H_0 nähme, dann würde das einem Signifikanztest mit einem Signifikanzniveau von (rund) 5% entsprechen.

Lehnte man hingegen H_0 nur bei k=10 ab, würde das Signifikanzniveau α=.001 (1 Promille) betragen.

Kritische Region = die Menge aller möglichen Ergebnisse, die zur Ablehnung der Nullhypothese führen.

Signifikanzniveau (Alpha) = die Summe der Wahrscheinlichkeiten aller möglichen Ergebnisse, die in die kritische Region fallen.

p-Wert eines beobachteten Ergebnisses = die Summe der Wahrscheinlichkeiten aller möglichen Ergebnisse, die so extrem wie oder extremer als das beobachtete Ergebnis sind.

Welchen p-Wert hat das Ergebnis k=9?

Das Ergebnis k=9 (neun rote Kugeln) hat einen p-Wert von .001 + .0098 = .0108.

Was ist ein signifikantes, was ist ein hochsignifikantes Ereignis?

Ein Ergebnis bezeichnet man als signifikant, wenn sein p-Wert nicht größer als das Signifikanzniveau ist (d.h. wenn es in die kritische Region fällt).

Es ist eine verbreitete Angewohnheit, Ergebnisse, die auf einem sehr kleinen Niveau (1% oder 1 Promille) signifikant sind als "hochsignifikant" zu bezeichnen. Viele Statistiker lehnen das ab, insbesondere, wenn ein Signifikanzniveau für alle Ergebnisse einer Studie angewandt wird (z.B.: alle Resultate mit einem p-Wert kleiner gleich .05 werden als signifikant betrachtet) und jene mit einem wesentlich kleinerem p-Wert (z.B.: .01) als hochsignifikant bezeichnet werden.

Welche Milchkuh-Ergebnisse sind auf dem 5%-Niveau signifikant?

Nur das Ergebnis der Urne 5 ist signifikant.

Kann ein signifikantes Ergebnis ein Irrtum sein?

Natürlich kann ein Ergebnis signifikant sein und damit zur Ablehnung der Nullhypothese führen, obwohl diese in Wahrheit zutrifft.

Wie wahrscheinlich ist dieser Irrtum?

Unter der Voraussetzung, daß die Nullhypothese zutrifft, tritt ein signifikantes Ergebnis, und damit eine falsche Entscheidung mit der Wahrscheinlichkeit Alpha (=Signifikanzniveau) ein. Die kritische Region wurde gerade so festgelegt, daß unter H_0 ein Anteil Alpha aller möglichen Ergebnisse in sie fällt.

Bei einem Signifikanzniveau von 5% erhält man also im Durchschnitt 5% irrtümlich signifikante Ergebnisse, wenn bei den entsprechenden Versuchen in Wahrheit die Nullhypothese zutrifft. Deshalb stellt Alpha zugleich die Irrtumswahrscheinlichkeit (unter der Nullhypothese) dar.

Irrtumswahrscheinlichkeit = Wahrscheinlichkeit für H_A zu entscheiden, wenn in Wahrheit H_0 richtig ist (=Signifikanzniveau).

Wie wählt man das Signifikanzniveau?

Prinzipiell ist man frei in der Wahl des Signifikanzniveaus. Praktisch haben sich einige Alpha-Werte eingebürgert, die man kaum ignorieren kann. Üblicherweise setzt man Alpha mit .05 oder .01 - sehr selten auch mit .001 fest.

Zusammenfassung:

Wie wird ein Signifikanztest ausgeführt?

- Formulieren von H_0 und H_A
- Planung eines Experimentes
- Festlegen des Signifikanzniveaus (der kritischen Region)
- Durchführung des Experiments
- Errechnen der entsprechenden Prüfgröße und Bestimmen des p-Wertes
- Entscheidung für HA, wenn p <= Alpha, d.h. wenn das Ergebnis in die kritische Region fällt

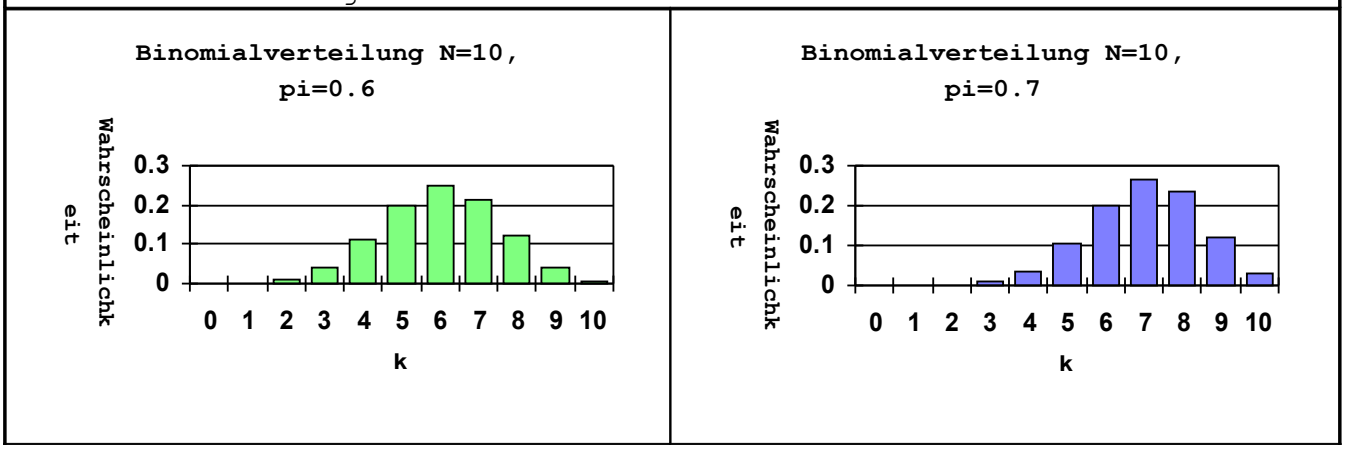
Im Beispiel besteht das "Berechnen" der Prüfgröße einfach im Zählen der roten Kugeln. Bei vielen Tests sind Berechnungen anzustellen, auf die wir hier jedoch nicht eingehen.

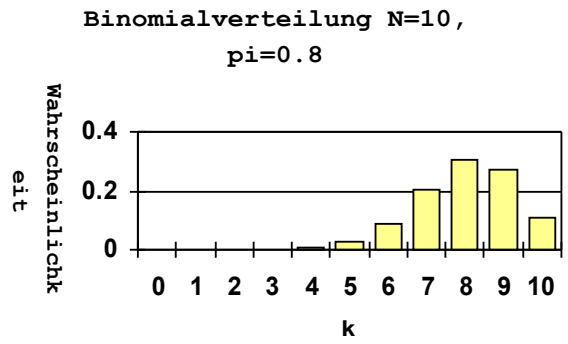
Sind die Ergebnisse k=10, k=9, k=8 unter H_A wahrscheinlicher?
Wie wahrscheinlich sind sie unter H_A?

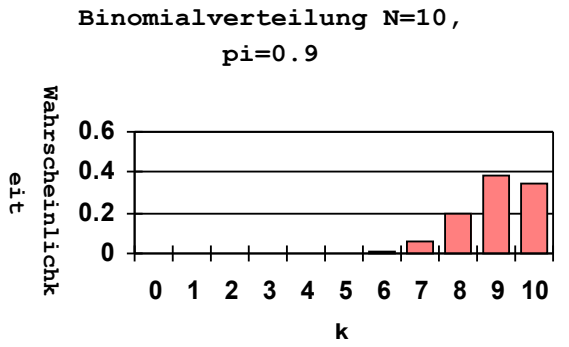
Wie man erwarten würde, ist es unter der Alternativhypothese wahrscheinlicher, acht, neun oder zehn rote Kugeln zu ziehen als unter der Nullhypothese. Wie wahrscheinlich es ist, läßt sich nur berechnen, wenn man sogenannte "einfache" Alternativhypothesen formuliert, das sind solche, bei denen π ein fester Wert gegeben wird (z.B.: H_A: π=0.6 oder H_A: π=0.75).

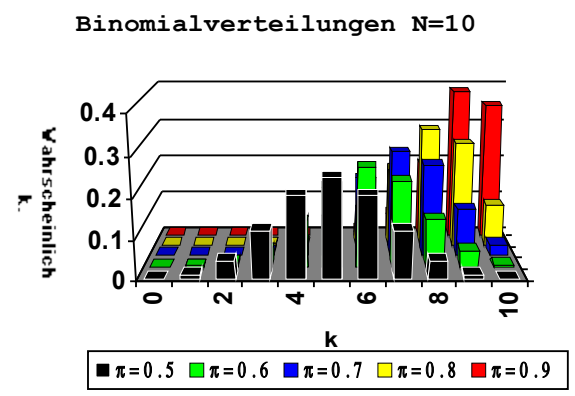
Wie sehen die Wahrscheinlichkeitsverteilungen unter genauer festgelegten Alternativhypothesen aus?

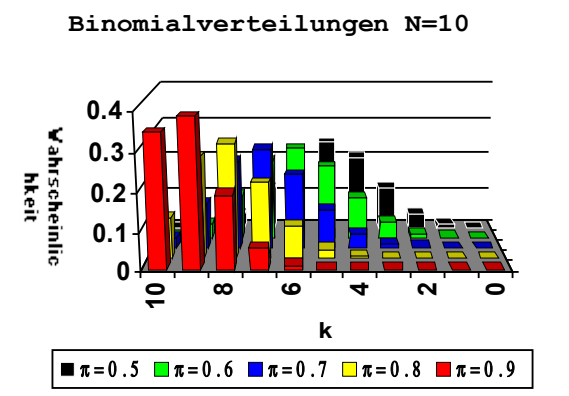
Binomialverteilungen N=10











Binomialverteilungen N=10, π=0.5, π=0.6, π=0.7, π=0.8, π=0.9;
Tabelle der Wahrscheinlichkeitsfunktionen

	k										
	0	1	2	3	4	5	6	7	8	9	10
π=0.5	0,0010	0,0098	0,0439	0,1172	0,2051	0,2461	0,2051	0,1172	0,0439	0,0098	0,0010
π=0.6	0,0001	0,0016	0,0106	0,0425	0,1115	0,2007	0,2508	0,2150	0,1209	0,0403	0,0060
π=0.7	0,0000	0,0001	0,0014	0,009	0,0368	0,1029	0,2001	0,2668	0,2335	0,1211	0,0282
π=0.8	0,0000	0,0000	0,0001	0,0008	0,0055	0,0264	0,0881	0,2013	0,3020	0,2684	0,1074
π=0.9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0015	0,0112	0,0574	0,1937	0,3874	0,3487

Die Grafiken und die Tabelle zeigen die Wahrscheinlichkeitsfunktionen von Binomialverteilungen mit unterschiedlichem π. Das entspricht Urnen mit einem Anteil von 60%, 70%, 80% und 90% roten Kugeln.

Die letzten beiden Grafiken zeigen alle vier Wahrscheinlichkeitsfunktionen samt jener für π=0.5 hintereinander gestellt (von zwei Seiten!). Man sieht, daß sich die Verteilung mit zunehmendem π nach rechts verschiebt, daß also die Ergebnisse mit hohem k (großer Zahl roter Kugeln) bei zunehmendem π wahrscheinlicher werden.

Wie wahrscheinlich sind die Ergebnisse k=10, k=9, k=8 unter HA: π=0.6 ?

Während k=10 unter H0 nur eine Wahrscheinlichkeit von 1 Promille besitzt, ist es unter HA: π=0.6 schon sechsmal so wahrscheinlich. Auch k=9 steigert seine Wahrscheinlichkeit von rund 1% auf rund 4% usw.

Das bedeutet, daß die kritische Region insofern vernünftig gewählt ist, als die darin enthaltenen Ergebnisse unter der Alternativhypothese eine höhere Wahrscheinlichkeit haben als unter der Nullhypothese.

Schema zur Bewertung eines statistischen Tests (Fehler I. und II. Art):

	Ergebnis ist "nicht signifikant"	Ergebnis ist "signifikant"
H0 ist wahr	richtig negativ	Fehler I.Art (α)
HA ist wahr	Fehler II.Art (β)	richtig positiv

Fehler I.Art = Entscheidung für H_A, wenn in Wahrheit H_0 gilt.

Alpha = Wahrscheinlichkeit, einen Fehler I.Art zu machen, vorausgesetzt H_0 gilt.

Fehler II.Art = Entscheidung für H_0, wenn in Wahrheit H_A gilt.

Beta = Wahrscheinlichkeit, einen Fehler II.Art zu machen vorausgesetzt H_A gilt.

Teststärke: Wahrscheinlichkeit eines signifikanten Ergebnisses, wenn H_A wahr (d.h. H_0 falsch) ist.

(Teststärke = 1 - Beta)

Die grundsätzliche Problemstellung bei einem Test läßt sich mit obiger Tabelle darstellen. Es gibt zwei mögliche Hypothesen und es gibt zwei mögliche Entscheidungen. Daraus ergeben sich vier Kombinationen, zwei wünschenswerte und zwei unerfreuliche.

In zwei Fällen entscheidet man richtig, in zwei Fällen irrt man. Auf den Fehler I.Art sind wir bereits eingegangen. Seine Größe ist uns bekannt. Sie entspricht dem Signifikanzniveau Alpha.

Ebenso ist es möglich, einen sogenannten Fehler II.Art zu begehen, nämlich irrtümlich für H_0 zu entscheiden.

Wie groß ist Alpha im Beispiel?

Alpha beträgt 5 Prozent.

Wie groß ist Beta im Beispiel bei gegebenem Alpha und gegebener Stichprobengröße, wenn H_A: $\pi = 0.7$?

Beta bezeichnet die Wahrscheinlichkeit irrtümlich für H_0 zu entscheiden. Das tun wir genau dann, wenn das Ergebnis nicht in die kritische Region fällt, wenn wir also weniger als acht rote Kugeln ziehen, obwohl in der Urne 70% rote Kugeln (π =0.7) sind.

Aus der Tabelle der Binomialverteilung entnehmen wir, daß bei π=0.7 die Wahrscheinlichkeiten für acht rote Kugeln p(k=8)=0.2335, für neun rote Kugeln p(k=9)=0.1211 und für zehn p(k=10)=0.0282 betragen. Das sind in Summe p(k>=8)=0.3838.

In rund 38% der Versuche mit einer 7:3-Urne ist also mit einem signifikanten Ergebnis und damit mit einer richtigen Entscheidung zu rechnen. Daraus ergibt sich, daß man in 62% irrtümlich ein nicht signifikantes Ergebnis erhält und damit eine falsche Entscheidung trifft. Beta ist demnach 0.62 .

Wie groß ist Beta für H_A: $\pi = 0.9$?

Beta ist 0.07.

Wie groß ist Beta wenn H_A: π=1.0?

Diese Frage läßt sich auch ohne Tabelle lösen. Wie wahrscheinlich ist es weniger als acht mal Rot zu ziehen, wenn die Urne nur rote Kugeln enthält? Beta ist 0.0 .

Was kann man über Beta für H_A: $\pi>0.7$ sagen?

Vergleicht man Beta von verschiedenen Alternativhypothesen, so stellt man fest, daß Beta mit zunehmendem Abstand der Alternativhypothese von der Nullhypothese kleiner wird.

Beträgt Beta (für H_A: $\pi=0.7$) 62% so ist es für jede Alternativhypothese $\pi>0.7$ kleiner als 62%.

Wie groß ist Beta höchstens?

Unter der Alternativhypothese sind jedenfalls mehr signifikante Ergebnisse als unter der Nullhypothese zu erwarten. Daher muß Beta kleiner als 1 - Alpha sein. Wenn sich H_A kaum von H_0 unterscheidet, ist Beta nahezu 1 - Alpha. Wenn man eine Urne mit 500 roten und 500 blauen Kugeln (H_0: $\pi=0.5$) von einer Urne mit 501 roten und 499 blauen Kugeln (H_A: $\pi=0.501$) unterscheiden will, dann muß man auch wenn man aus der "Alternativ"-Urne zieht mit nahezu 95% nicht signifikanten Ergebnissen rechnen.

Was bewirkt eine Verminderung von Alpha auf 1%?

Verringert man Alpha auf 1%, dann erhält man unter H_0 nur 1% fälschlich signifikante Ergebnisse. Man ist sozusagen "vorsichtiger".

Welche Ergebnisse fallen unter Alpha = 0.01 in die kritische Region?

Im Beispiel wären dann nur die Ergebnisse k=9 und k=10 signifikant (genaugenommen beträgt Alpha dann 1.08%).

Beeinflußt eine Änderung von Alpha auch Beta?

Eine Verkleinerung der kritischen Region verringert auch unter der Alternativhypothese die Chance auf ein signifikantes Ergebnis und erhöht damit automatisch Beta.

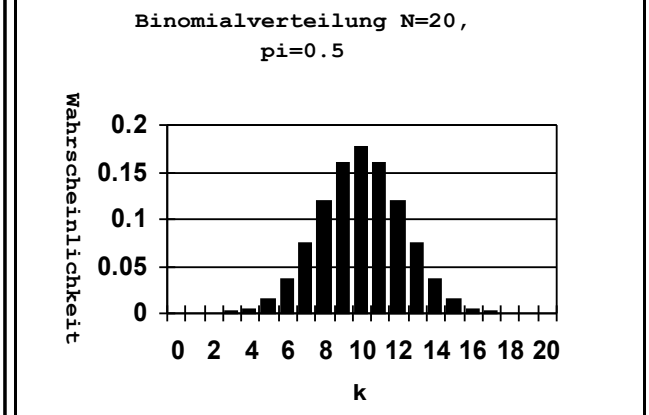
Was ändert sich, wenn man die Stichprobengröße auf 20 erhöht?

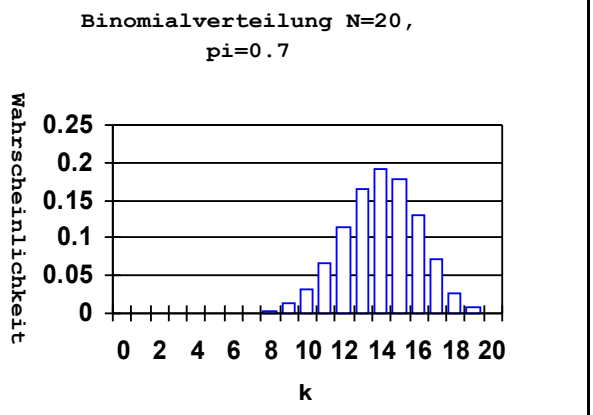
(vorausgesetzt H0, HA und Alpha bleiben unverändert)

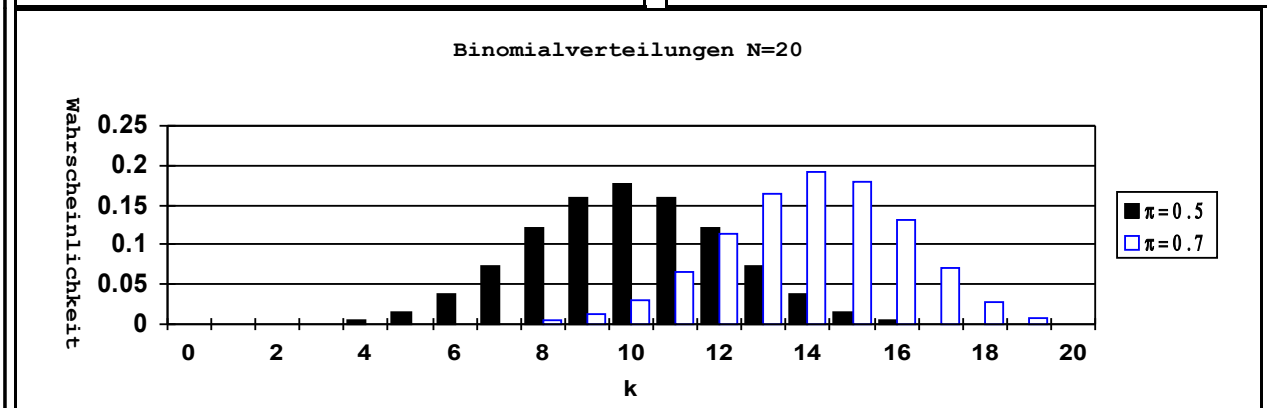
Alle bisherigen Überlegungen sind davon ausgegangen, daß man bei einem Versuch nur zehn mal eine Kugel zieht. Man könnte aber auch mehr Kugeln ziehen und man würde sich erwarten, damit mehr Information über den Inhalt zu erhalten. Das sollte auch die Gefahr eines Irrtums verringern.

Wir wollen überlegen, wie sich eine Vergrößerung der Stichprobe von 10 auf 20 Kugeln auswirkt und uns dazu die Wahrscheinlichkeitsfunktionen der Binomialverteilungen BN(N=20, $\pi=0.5$) und BN(N=20, $\pi=0.7$) ansehen.

Binomialverteilungen N=20







Binomialverteilungen N=20 (teilw.)

		k										
Anz R	...	10	11	12	13	14	15	16	17	18	19	20
π=0.5	...	.176	.160	.120	.074	.037	.015	.005	.001	.000	.000	.000
π=0.7	...	.031	.065	.114	.164	.192	.179	.130	.072	.028	.007	.001

Welche Ergebnisse fallen bei N=20 in die kritische Region?

Bei einem Stichprobenumfang von 20 Kugeln fallen - großzügig gerechnet - die Ergebnisse mit 14 und mehr roten Kugeln (k>=14) in die kritische Region. Alpha beträgt dann 5.7%.

Wie groß ist Beta für HA: π=0.7?

Unter der Alternativhypothese π=0.7 beträgt Beta 0.39 .

Die Chance, eine 7:3-Urne zu entdecken ist damit von 38% auf 61% gestiegen. Je größer die Stichprobe ist, desto größer ist auch die Chance eine richtige Entscheidung zu fällen.

Was ist wirklich in den Urnen?

Ergebnisse:	1	2	3	4	5	6	7	8	9	10	Anzahl rot	Anzahl blau	**wahre Anzahl rot**	**wahre Anzahl blau**
Substanz 1	r	b	b	b	r	b	b	r	b	r	4	6	**5**	**5**
Substanz 2	r	b	r	r	r	b	b	r	b	r	6	4	**5**	**5**
Substanz 3	r	b	r	b	r	r	b	r	b	b	5	5	**5**	**5**
Substanz 4	b	b	b	b	r	r	r	b	b	b	3	7	**4**	**6**
Substanz 5	r	r	r	r	r	r	r	r	r	r	10	0	**10**	**0**
Substanz 6	r	r	r	r	r	r	r	b	b	b	7	3	**7**	**3**

Die beiden letzten Spalten zeigen den wirklichen Inhalt der sechs Urnen. Die Daten stammen aus einem tatsächlich durchgeführten Experiment.

Man sieht, daß die Nullhypothese in 3 Fällen zu recht und in 1 Fall zu unrecht beibehalten wurde. Substanz 4 zeigt das Problem einseitig formulierter Alternativhypothesen.

Nur die Substanz 5 wurde zurecht als wirksam erkannt.

Zusammenfassung:

Wie hängen Alpha, Beta, H_A und der Stichprobenumfang zusammen?

Unter sonst gleichen Bedingungen gilt:

- je kleiner Alpha, desto größer Beta
- je kleiner der Unterschied zwischen H_0 und H_A, desto größer Beta
- je größer die Stichprobe, desto kleiner Beta

Beta bzw. der notwendige Stichprobenumfang können für verschiedene - nicht jedoch für alle - Signifikanztests bestimmt werden.

Für viele Situationen kann bei gegebenen Alpha, Beta, Nullhypothese und Alternativhypothese der notwendige Stichprobenumfang errechnet werden. Das ist jedoch nicht für alle Tests möglich.

Was ist eine einseitige, was eine zweiseitige Alternativhypothese?

Bisher wurde der Einfachheit halber immer von einer einseitigen Alternativhypothese ausgegangen. Wir haben im Beispiel unterstellt, daß in einer Urne entweder gleichviele rote wie blaue Kugeln sind, oder daß die roten überwiegen.

Hätten wir auch den Fall in Betracht ziehen wollen, daß weniger rote als blaue Kugeln in der Urne sind, wäre die Alternativhypothese H_A: $\pi \neq 0.5$ zu formulieren und zu prüfen gewesen.

Wie wirkt sich eine zweiseitige Formulierung der Alternativhypothese auf Alpha und die kritische Region aus?

Grundsätzlich kann man das Signifikanznivau beibehalten, aber das hat eine Änderung der kritischen Region zur Folge. Man muß dann extreme Ergebnisse auf beiden Seiten beachten. Das Ergebnis "keine rote Kugel" spricht ebenso gegen die Nullhypothese wie "alle Kugeln rot".

Im Beispiel mit 10 gezogenen Kugeln läßt sich allerdings keine kritische Region konstruieren, die halbwegs die Bedingung des vorher verwandten Alpha von 5% erfüllt. Dieses Problem ist aber prinzipiell lösbar.

Einem Alpha von 2% würde die kritische Region k=0, k=1, k=9 und k=10 entsprechen (genau Alpha = .0216).

Ändert sich Beta?

Da die kritische Region bei zweiseitiger Alternativhypothese an jedem Ende nur mehr halb so groß ist, vergrößert sich Beta ebenso, wie wenn man bei einer einseitigen Alternativhypothese Alpha halbierte.

Allgemeine Voraussetzungen des Hypothesenprüfens mit einem Signifikanztest:

Vor der Durchführung des Experimentes (der Studie) muß Folgendes festgelegt werden:

- Nullhypothese
- Alternativhypothese (einseitig oder zweiseitig)
- Signifikanzniveau (Alpha)
- Umfang oder Dauer des Experimentes
- der spezifische Signifikanztest

Alle genannten Bedingungen sind notwendig. Besonders die Festlegung des Umfanges wird oft übergangen und es werden so lange Daten gesammelt, bis man ein signifikantes Ergebnis erreicht. Das ist grob falsch, da sich dadurch das reale Alpha unkontrolliert erhöht.

Gültig ist der Test nur, wenn das ihm zugrunde liegende Modell und das Experiment im Einklang stehen!

Wenn man z.B. das obige Beispiel so veränderte, daß man jene Kugeln, die man entnommen hat, nicht wieder zurücklegte, würden sich die Wahrscheinlichkeiten der möglichen Ergebnisse ändern und der Test somit falsch sein.

Wie dokumentiert man Testergebnisse?

Man gibt an:

- den angewandten Test
- die errechnete Prüfgröße
- (die dazugehörigen Freiheitsgrade)
- p-Wert

Beispiel: Östrogenrezeptorbestimmung
biochemisch und immunhistochemisch

Anzahl %		immunhistochem. positiv	 negativ	
bio-chem-isch	positiv	170 70.8	13 5.4	183 76.3
	negativ	25 10.4	32 13.3	57 240

Chi-Square	Value	df	p-value
Pearson	68.60	1	.00000
Continuity Correction	65.42	1	.00000
Phi	.53		.00000

im Text: Chi^2 (kontinuitätskorrigiert) = 65.42, df=1, p<0.001

Stand: 10.4.1993