Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun
http://blog.ruediger-braun.net

Heinrich-Heine-Universität Düsseldorf

09. Januar 2015

Poissonverteilung

Es sei $\lambda > 0$. Die Poissonverteilung zum Parameter λ ist definiert durch

$$P_\lambda(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

Unter den folgenden Voraussetzungen ist eine Zufallsvariable X poissonverteilt zum Parameter λ:

- X zählt das Auftreten eines Ereignisses pro Zähleinheit
- Im Mittel treten λ Ereignisse pro Zähleinheit auf
- Die Ereignisse beeinflussen sich nicht gegenseitig

Beispiel Tumor

- Ein Tumor aus 160 Zellen wird bestrahlt
- Im Mittel stirbt jede Minute ein Zehntel aller Tumorzellen
- Mit welcher Wahrscheinlichkeit sterben 10 Zellen in der ersten Minute?
- Zwei Modelle sind angemessen
  - Binomialverteilung
  - Poissonverteilung

Beispiel Tumor: Rechnung mit Binomialverteilung

- Modell: 160 unabhängige ja/nein-Experimente

  Erfolg : Tod der Tumorzelle

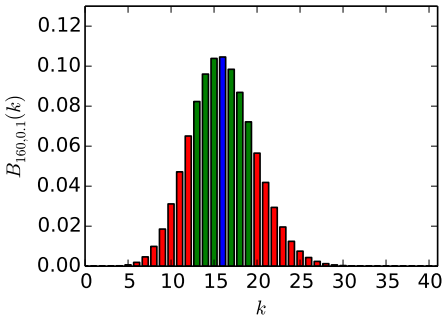
- Erfolgswahrscheinlichkeit im Einzelfall $p = 0.1$
- Anzahl der Erfolge verteilt gemäß $B_{160,\,0.1}$
- Antwort:

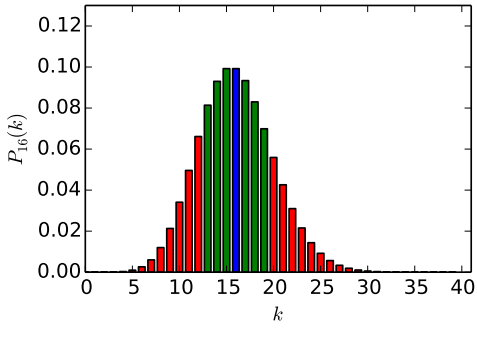
$$B_{160,\,0.1}(10) = \binom{160}{10} \cdot 0.1^{10} \cdot 0.9^{150} = 0.03113$$

Beispiel Tumor: Rechnung mit Poissonverteilung

- Modell: seltenes Ereignis, das im Mittel 16 mal pro Minute auftritt
- Was ist hier selten?
- Für die einzelne Zelle sind Treffer selten
- Parameter der Poissonverteilung ist $\lambda = 16$
- Anzahl der Ereignisse pro Zähleinheit ist verteilt gemäß P_{16}
- Antwort
$$P_{16}(10) = \frac{16^{10}}{10!} e^{-16} = 0.03410$$
- Zum Vergleich $B_{160,\,0.1}(10) = 0.03113$

Vergleich Binomial- und Poissonverteilung

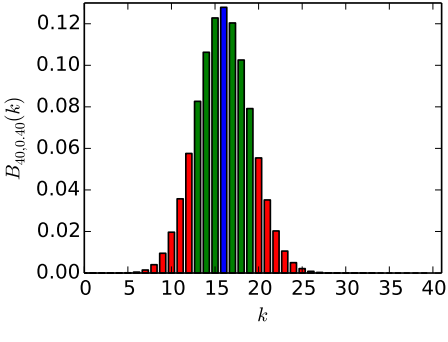


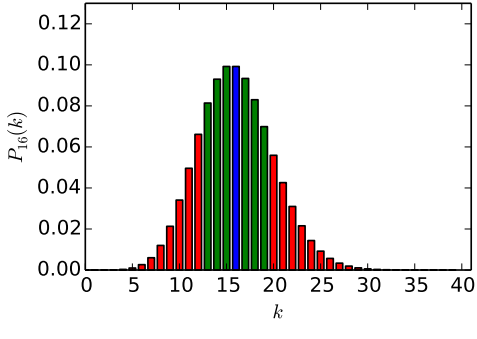


Beide beschreiben einen Prozess mit 16 Erfolgen im Mittel. Der Unterschied ist, dass beim Poissonprozess die Anzahl der Erfolge potenziell unbeschränkt ist.

Vergleich Binomial- und Poissonverteilung, Fortsetzung

$B_{40,0.4}$ und P_{16} besitzen ebenfalls beide den Erwartungswert 16





Gesetz der seltenen Ereignisse

Die Poisson-Verteilung P_λ mit $\lambda = n \cdot p$ ist eine sehr gute Annäherung an die Binomialverteilung $B_{n,p}$, falls $n \geq 100$ und $n \cdot p \leq 10$.

Im Beispiel waren

- $n = 160$
- $p = 0.1$

Die Annäherung ist daher nur gut, nicht sehr gut

Messwiederholungen

- Warum erhöhen mehrere Messungen die Genauigkeit?
- Warum braucht man 100-mal so viele Messungen, um die Genauigkeit zu verzehnfachen?

Rechenregeln für den Erwartungswert

- Für jede Zahl c und jede Zufallsvariable X ist $E(c \cdot X) = c \cdot E(X)$
- Für Zufallsvariablen $X_1, \ldots, X_n$ ist $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$
- X und Y **unabhängige** Zufallsvariable. Dann

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

Rechenregeln für die Varianz

- Für jede Zahl a und jede Zufallsvariable X gilt $Var(a + X) = Var(X)$
- Für Zahl c und jede Zufallsvariable X gilt $Var(c \cdot X) = c^2 \cdot Var(X)$
- X und Y **unabhängige** Zufallsvariable. Dann

$$Var(X + Y) = Var(X) + Var(Y)$$

Zwei unabhängige, identisch verteilte Zufallsvariable

- X_1 und X_2 seien unabhängige Zufallsvariable, die derselben Verteilung gehorchen (also z. B. Messwiederholungen). Sei $Y = \frac{1}{2}(X_1 + X_2)$ der Durchschnittswert
- Der Erwartungswert von X_1 heiße μ, also $E(X_1) = E(X_2) = \mu$
- Die Streuung von X_1 heiße σ, also $Var(X_1) = Var(X_2) = \sigma^2$
- $E(Y) = \frac{1}{2}(E(X_1) + E(X_2)) = \mu$
- $Var(Y) = \left(\frac{1}{2}\right)^2 Var(X_1) + \left(\frac{1}{2}\right)^2 Var(X_2) = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{1}{2}\sigma^2$
- Also ist $\dfrac{\sigma}{\sqrt{2}}$ die Streuung von Y

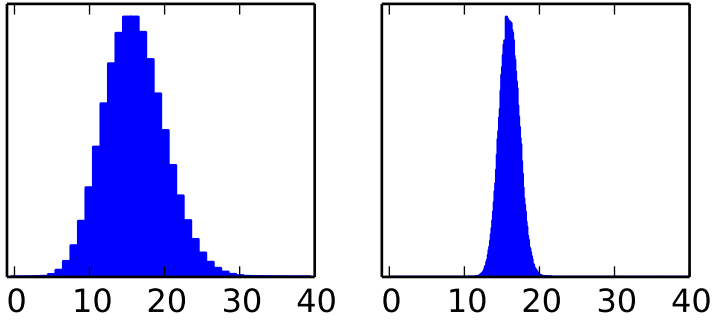
Das schwache Gesetz der großen Zahl

- "Mit ausreichend vielen Messwiederholungen lässt sich jede Genauigkeit erreichen"
- Präziser: $X_1, \ldots, X_n$ unabhängig, alle mit derselben Verteilung
- $\mu = E(X_1) = \cdots = E(X_n)$ und $\sigma^2 = Var(X_1) = \cdots = Var(X_n)$
- $Y = \frac{1}{n}(X_1 + \cdots + X_n)$
- Y ist das arithmetische Mittel der $X_1, X_2, \ldots, X_n$
- Dann $E(Y) = \mu$ und die Streuung von Y beträgt

$$\sigma_Y = \frac{\sigma}{\sqrt{n}}$$

- Das bedeutet: Um die Streuung zu zehnteln, müssen 100 mal so viele Versuche durchgeführt werden

Messwiederholungen: Beispiel



Links: Poissonverteilung P_{16}, Streuung ist 4

Rechts: Durchschnittswerte aus zehn P_{16}-verteilten Zufallsvariablen, Streuung ist $\frac{4}{\sqrt{10}} = 1.26$

Versuchsplanung: α-Strahler

- Der Einschlag von α-Teilchen wird mit der Poisson-Verteilung P_λ modelliert, wobei λ die Zahl der Einschläge pro Sekunde ist
- λ soll bis auf einen Fehler (Streuung) von 0.25 bestimmt werden
- Wie viele Einzelversuche von einer Sekunde Dauer sind erforderlich?
- Dazu müssen wir aus einem Pilotversuch einen Anhaltspunkt für λ kennen. Der Pilotversuch habe $\lambda \cong 25$ ergeben
- $Var(P_\lambda) = \lambda$. Also hat jeder Einzelversuch die Streuung $\sqrt{25} = 5.0$
- Löse Gleichung

$$\frac{5.0}{\sqrt{n}} = 0.25$$

- Also $\sqrt{n} = 20$ und $n = 400$

Teil III

Schließende Statistik

Beispiel *L*-Bakterien

- Ein Bakterium kommt in ungestörtem Boden zu 75% in der *L*-Variante und zu 25% in der *R*-Variante vor
- Es soll getestet wrden, ob ein bestimmtes Pestizid *L*-Bakterien mehr schädigt als *R*-Bakterien
- Dazu wird ein Experiment gemacht, statistisch bewertet und schließlich eine Antwort auf die Frage gegeben:

  Schädigt das Pestizid L-Bakterien mehr als R-Bakterien?

Beispiel L-Bakterien

- Generell sind vier Ausgänge des Experiments möglich
  - Das Pestizid schädigt L-Bakterien nicht mehr als R-Bakterien und das Experiment beantwortet die Frage mit nein
    Korrekte Antwort
  - Das Pestizid schädigt L-Bakterien nicht mehr als R-Bakterien und das Experiment beantwortet die Frage mit ja
    Falsche Antwort
  - Das Pestizid schädigt L-Bakterien mehr als R-Bakterien und das Experiment beantwortet die Frage mit nein
    Falsche Antwort
  - Das Pestizid schädigt L-Bakterien mehr als R-Bakterien und das Experiment beantwortet die Frage mit ja
    Korrekte Antwort
- Durch die Auswahl der Stichprobe kommt Zufall ins Spiel. Falsche Antworten sind unvermeidbar.
- Ziel der Statistik ist es, Schranken für die Wahrscheinlichkeit falscher Antworten zu geben

Nullhypothese und Alternativhypothese

- **Nullhypothese H_0:** Das ist diejenige Hypothese, deren fälschliche Ablehnung man nach Möglichkeit vermeiden will

  Häufig ist die Nullhypothese die Aussage, dass kein Einfluss vorliegt
- **Alternativhypothese H_1:** Das ist die Alternative zur Nullhypothese

Nullhypothese und Alternativhypothese, Fortsetzung

- Wissenschaft ist konservativ. Wer mit einer neuen Idee kommt, muss zeigen, dass sie besser ist als die alte
- Daher ist die Nullhypothese in der Regel die Annahme, dass die bestehende Theorie so gut ist wie die Neuerungen bzw. dass der untersuchte Stoff ohne Einfluss ist
- Neutralitätshypothese in der Genetik: Die Nullhypothese besagt, dass die untersuchte Variation der Gensequenz folgenlos ist.

Fehler erster und zweiter Art

- **Der Fehler 1. Art** ist die fälschliche Ablehnung der Nullhypothese.
- **Der Fehler 2. Art** ist die fälschliche Annahme der Nullhypothese

Die Priorität liegt auf der Vermeidung des Fehlers 1. Art. Diese Asymmetrie ist ein entscheidendes Merkmal der Testtheorie.

Beispiel L-Bakterien

- Ein Bakterium kommt in ungestörtem Boden zu 75% in der L-Variante und zu 25% in der R-Variante vor
- Pilotversuche deuten an, dass ein bestimmtes Pestizid L-Bakterien mehr schädigt als R-Bakterien
- Hypothesen:

  Nullhypthese H_0: "Das Pestizid schädigt L-Bakterien nicht mehr als R-Bakterien"

  Alternative H_1: "Das Pestizid schädigt L-Bakterien mehr als R-Bakterien"
- Experiment: 27 Bakterien zufällig ausgewählt, davon 14 L- und 13 R-Bakterien
- Aus früherer Rechnung wissen wir, dass dieses Ergebnis unter der Nullhypthese sehr unwahrscheinlich ist
- Die Nullhypothese wird abgelehnt

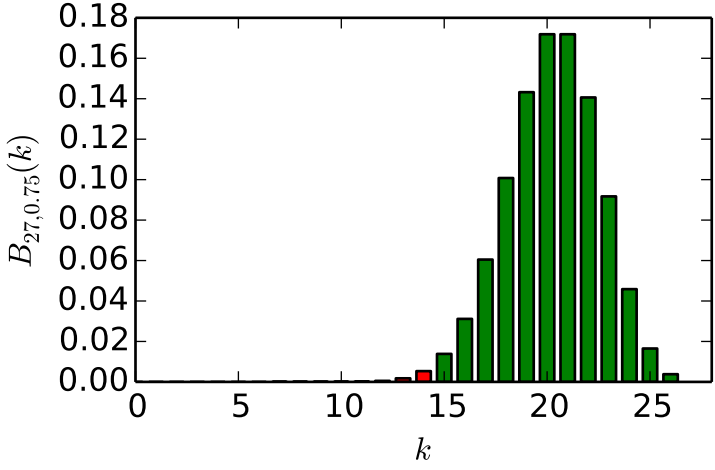
L-Bakterien: Fortsetzung

- Welche Fehlerwahrscheinlichkeit haben wir?
- Wir haben H_0 abgelehnt, es kann sich also höchstens um einen Fehler erster Art handeln
- Seine Wahrscheinlichkeit haben wir bereits ausgerechnet:
  Es ist die Wahrscheinlichkeit, in einem ja/nein-Experiment mit Erfolgswahrscheinlichkeit $p = 0.75$ für den Einzelfall in einer Stichprobe vom Umfang $n = 27$ höchstens 14 Erfolge zu haben
- In Formeln: Die Wahrscheinlichkeit eines Fehlers erster Art beträgt für dieses Ergebnis

$$\sum_{k=0}^{14} B_{27,0.75}(k) = 0.007778$$

- Unsere Antwort hat also eine Fehlerwahrscheinlichkeit von 0.8%

L-Bakterien: Fortsetzung



Unter der Annahme, dass das Pestizid das Zahlenverhältnis zwischen L- und R-Bakterien nicht beeinflusst, markieren die roten Felder die Wahrscheinlichkeit, dass 14 oder weniger L-Bakterien gefunden werden.

Testverfahren: Fehler erster und zweiter Art

- Ein *Test* besteht aus einer Vorschrift, die zu jedem möglichen Versuchsausgang festlegt, ob die Nullhypothese H_0 angenommen oder abgelehnt wird.
- Dabei kann es zu zwei verschiedenen Fehlentscheidungen kommen:

	H_0 wird angenommen	H_0 wird abgelehnt
H_0 trifft zu	richtige Entscheidung	Fehler 1. Art
H_1 trifft zu	Fehler 2. Art	richtige Entscheidung

Signifikanztests

- Für den Fall, dass H_0 zutrifft, bezeichnet man die Wahrscheinlichkeit, dass H_0 trotzdem abgelehnt wird, als *Fehlerwahrscheinlichkeit erster Art*
- Ein Test heißt *Signifikanztest* zum Niveau α, wenn alle Fehlerwahrscheinlichkeiten erster Art $\leq \alpha$ sind
- Das übliche Niveau ist 0.05
- Für den Fall, dass H_0 nicht zutrifft, bezeichnet man die Wahrscheinlichkeit, dass H_0 trotzdem nicht abgelehnt wird, als *Fehlerwahrscheinlichkeit zweiter Art*

Test für die L-Bakterien

- Wir konstruieren einen Test zum Signifikanzniveau $\alpha = 0.05$
- Wären 15 oder gar 16 L-Bakterien immer noch Grund gewesen, H_0 abzulehnen?
- Die Tabelle zeigt, dass 15 L-Bakterien immer noch ein Grund zur Ablehnung sind, 16 aber nicht

Daher lautet die Testvorschrift

- Bei 15 oder weniger L-Bakterien wird H_0 abgelehnt
- bei 16 oder mehr L-Bakterien wird H_0 beibehalten

Tabelle der Werte $\sum_{k=0}^{r} B_{n,p}(k)$ für $n = 27$

r	p	0.75	0.76	0.77	0.78	0.79
9	0.	00001				
10		00003	00002	00001	00001	
11		00016	00010	00006	00003	00002
12		00067	00042	00026	00015	00009
13		00245	00161	00103	00065	00039
14		00778	00538	00364	00240	00153
15		02162	01573	01119	00777	00526
16		05278	04031	03016	02208	01577
17		11325	09067	07126	05488	04136
18		21405	17927	14769	11951	09483
19		35729	31217	26889	22804	19012
20		52917	48050	43120	38195	33350
21		70105	65819	61232	56385	51330
22		84168	81165	77770	73973	69777
23		93340	91729	89806	87530	84864
24		97926	97305	96521	95540	94322
25		99577	99423	99219	98948	98592
26		99958	99939	99914	99878	99828

Fehlerwahrscheinlichkeit zweiter Art im Beispiel

- Wie groß ist die Fehlerwahrscheinlichkeit zweiter Art?
- Das ist keine gute Frage
- Wenn das Pestizid die Wahrscheinlichkeit von L-Bakterien von 75% auf 74.999% senkt, dann ist das sehr schwer nachzuweisen
- Sinnvoll ist folgende Frage

  Angenommen, das Pestizid senkt die Wahrscheinlichkeit von L-Bakterien auf 50%, mit welcher Wahrscheinlichkeit wird unser Test diesen Rückgang entdecken?

- Wenn q die Fehlerwahrscheinlichkeit zweiter Art ist, dann bezeichnet man $1 - q$ als *Power* des Tests
- Die Power hängt also davon ab, welche Annahme man über den Abstand zwischen Nullhypothese und Alternative macht

Fehlerwahrscheinlichkeit zweiter Art: Fortsetzung

- Bei 16 oder mehr L-Bakterien wird H_0 nicht abgelehnt
- Wie wahrscheinlich ist dieses Ergebnis, wenn tatsächlich nur 50% aller Bakterien L-Bakterien sind?
- Gesucht

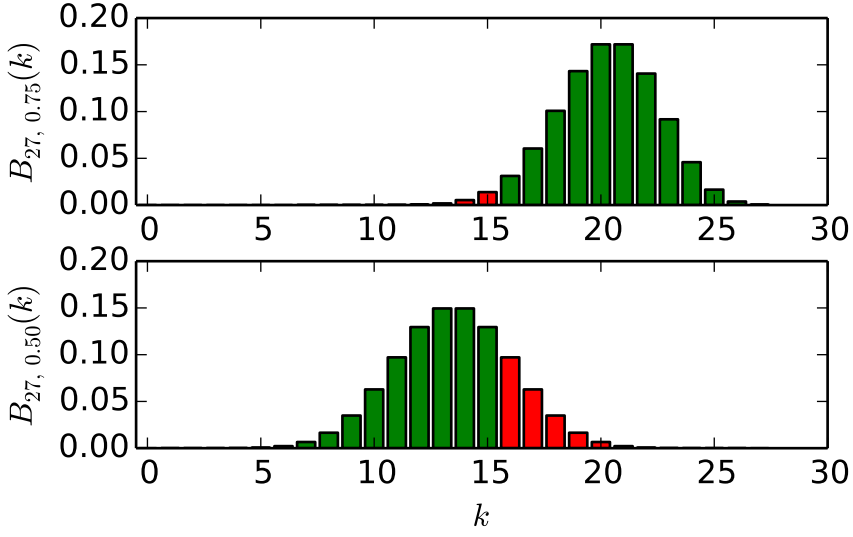
$$\sum_{k=16}^{27} B_{27,0.5}(k) = 1 - \sum_{k=0}^{15} B_{27,0.5}(k) = 1 - 0.77897 = 0.22103$$

- Unter der Annahme beträgt die Fehlerwahrscheinlichkeit zweiter Art 22%

Tabelle der Werte $\sum_{k=0}^{r} B_{n,p}(k)$ für $n = 27$

r	p	0.50	0.51	0.52	0.53	0.54
3	0.	00002	00002	00001	00001	
4		00016	00010	00007	00005	00003
5		00076	00053	00037	00025	00017
6		00296	00216	00155	00111	00078
7		00958	00723	00540	00399	00292
8		02612	02043	01582	01213	00920
9		06104	04944	03966	03150	02476
10		12389	10379	08614	07081	05764
11		22103	19121	16396	13933	11730
12		35055	31253	27637	24235	21068
13		50000	45823	41689	37640	33716
14		64945	60987	56911	52756	48563
15		77897	74666	71203	67528	63669
16		87611	85344	82815	80022	76969
17		93896	92535	90955	89138	87071
18		97388	96693	95854	94850	93660
19		99042	98744	98368	97900	97324
20		99704	99597	99457	99276	99044
21		99924	99893	99851	99794	99717
22		99984	99977	99967	99953	99933
23		99998	99996	99994	99992	99988
24				99999	99999	99998

Fehler 1. und 2. Art



Die grünen Balken gehören zu richtigen Entscheidungen.