Wahrscheinlichkeitsrechnung und Statistik
im Biologie-Bachelorstudiengang der

Einige prominente Familien von Wahrscheinlichkeitsverteilungen

Dirk Metzler

9. Juli 2020

Inhaltsverzeichnis

1 Binomialverteilung (und Bernoulli-Verteilung) 1

2 Normalverteilung 2

3 T-Verteilung 4

4 Chi-Quadrat-Verteilung 5

5 F-Verteilung 6

6 Geometrische Verteilung und Exponentialverteilung 7

7 Poisson-Verteilung 9

8 Beta-Verteilung (und uniforme Verteilung) 10

9 Hypergeometrische Verteilung 11

1 Binomialverteilung (und Bernoulli-Verteilung)

Binomialverteilung

Sei K die Anzahl der Erfolge bei n unabhängigen Versuchen mit Erfolgswahrscheinlichkeit von jeweils p. Dann gilt für $k \in \{0, 1, \ldots, n\}$

$$\Pr(K = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

und K heißt *binomialverteilt*, kurz:

$$K \sim \text{bin}(n, p).$$

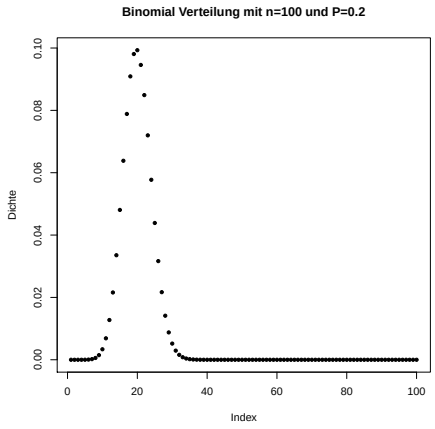
$$\mathbb{E}K = np \qquad \text{Var } K = n \cdot p \cdot (1-p)$$

Eine Zufallsgröße X, die nur zwei verschiedene Werte (Nein/Ja, Misserfolg/Erfolg, ...) annehmen kann, nennt man auch Bernoulli-Zufallsvariable. Kodiert man diese Werte durch 0 und 1, spricht man von der Bernoulli-Verteilung, und nennt $\Pr(X = 1)$ "Erfolgswahrscheinlichkeit".

Sind $X_1, X_2, X_3, \ldots$ alle unabhängig Bernoulli-verteilt mit gleicher Erfolgswahrscheinlichkeit p, so heißt die Folge $X_1, X_2, X_3, \ldots$ auch *Bernoulli-Prozess* oder *Bernoulli-Kette*, und es gilt

$$\sum_{i=1}^{n} X_i \sim \text{bin}(n, p).$$

Wahrscheinlichkeitsgewichte der Binomialverteilung



Benutzung der Binomialverteilung

Ein Problem bei der Binomialverteilung ist: $\binom{n}{k}$ exakt zu berechnen, ist für große n (ohne Computer) sehr aufwändig. Deshalb:

Die Binomialverteilung wird oft durch andere Verteilungen approximiert:

- durch die Normalverteilung, wenn n groß und $n \cdot p \cdot (1-p)$ nicht zu klein,
- durch die Poisson-Verteilung, wenn n groß und p klein.

2 Normalverteilung

Normalverteilung

Eine Zufallsvariable Z mit der Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

heißt *standardnormalverteilt*.

kurz: $Z \sim \mathcal{N}(0,1)$

$$\mathbb{E}Z = 0$$
$$\text{Var } Z = 1$$

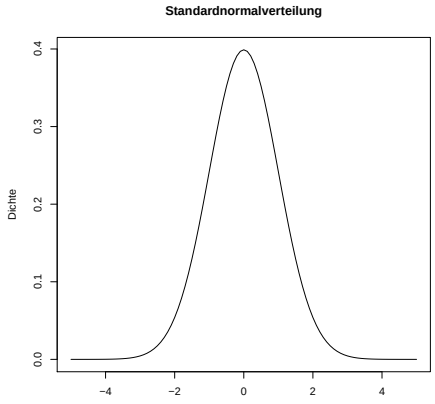
Ist Z $\mathcal{N}(0,1)$-verteilt, so ist $X = \sigma \cdot Z + \mu$ normalverteilt mit Mittelwert μ und Varianz σ^2, kurz:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

X hat dann die Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Dichte der Normalverteilung



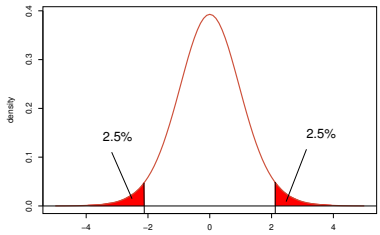
Merkregeln der Normalverteilung

Ist $Z \sim \mathcal{N}(\mu, \sigma^2)$, so gilt:

- $\Pr(|Z - \mu| > \sigma) \approx 33\%$
- $\Pr(|Z - \mu| > 1.96 \cdot \sigma) \approx 5\%$
- $\Pr(|Z - \mu| > 3 \cdot \sigma) \approx 0.3\%$

Berechnung von Quantilen

Sei $Z \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$ standardnormalverteilt. Für welchen Wert z gilt $\Pr(|Z| > z) = 5\%$?



Wegen der Symmetrie bzgl der y-Achse gilt

$$\Pr(|Z| > z) = \Pr(Z < -z) + \Pr(Z > z) = 2 \cdot \Pr(Z < -z)$$

Finde also $z > 0$, so dass $\Pr(Z < -z) = 2.5\%$.

```
> qnorm(0.025,mean=0,sd=1)
[1] -1.959964
```

Antwort: $z \approx 1.96$, also knapp 2 Standardabweichungen

Normalapproximation

Für große n und p, die nicht zu nahe bei 0 oder 1 liegen, kann man die Binomialverteilung durch die Normalverteilung mit dem entsprechenden Erwartungswert und der entsprechenden Varianz approximieren:

Ist $K \sim \text{bin}(n, p)$ und $Z \sim \mathcal{N}(\mu = n \cdot p, \sigma^2 = n \cdot p \cdot (1 - p))$, so gilt

$$\Pr(K \in [a, b]) \approx \Pr(Z \in [a, b])$$

(eine Faustregel: für den Hausgebrauch meist okay, wenn $n \cdot p \cdot (1 - p) \geq 9$)

Zentraler Grenzwertsatz

Eine etwas allgemeinere *Normalapproximation* beschreibt der Zentraler Grenzwertsatz.

Der zentrale Grenzwertsatz besagt, dass die Verteilung von Summen
unabhängiger und identisch verteilter
Zufallsvariablen in etwa die Normalverteilung ist.

Zentraler Grenzwertsatz

Die $\mathbb{R}$-wertigen Zufallsgrößen $X_1, X_2, \ldots$ seien unabhängig und identisch verteilt mit endlicher Varianz $0 < \operatorname{Var} X_i < \infty$. Sei außerdem

$$Z_n := X_1 + X_2 + \cdots + X_n$$

die Summe der ersten n Variablen.

Dann ist die zentrierte und reskalierte Summe im Limes $n \to \infty$ standardnormalverteilt, d.h.

$$\frac{Z_n - \mathbb{E}Z_n}{\sqrt{\operatorname{Var} Z_n}} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

bei $n \to \infty$.

Formal: Es gilt für alle $-\infty \le a < b \le \infty$

$$\lim_{n\to\infty} \Pr\left(a \le \frac{Z_n - \mathbb{E}Z_n}{\sqrt{\operatorname{Var} Z_n}} \le b\right) = \Pr(a \le Z \le b),$$

wobei Z eine standardnormalverteilte Zufallsvariable ist.

3 T-Verteilung

T-Verteilung

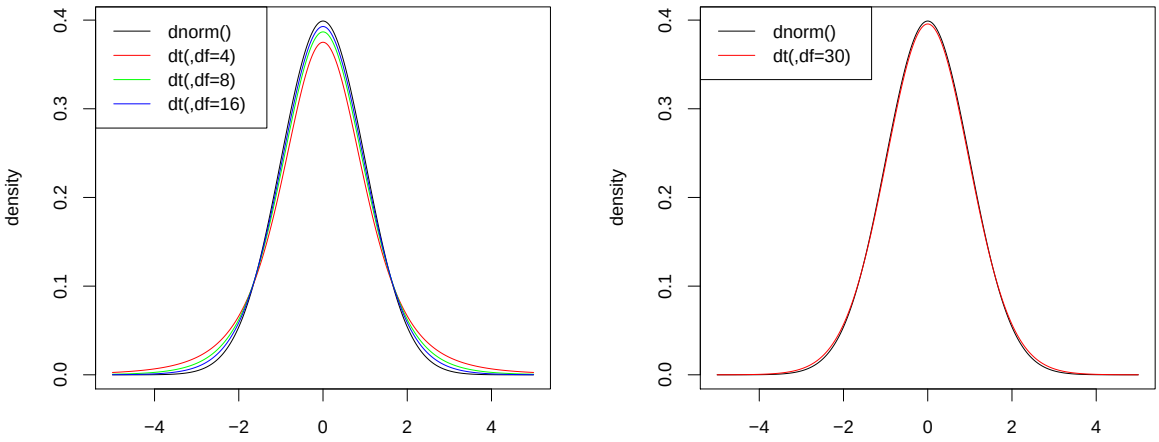
Sind $X_1, \ldots, X_n$ unabhängig aus einer Normalverteilung mit Mittelwert μ gezogen, so ist

$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

t-verteilt mit $n - 1$ Freiheitsgraden (df=*degrees of freedom*). Eine t-verteilte Zufallsvariable bezeichnen wir meist mit T.

Die t-Verteilung heißt auch Student-Verteilung. Die t-Verteilung wurde 1908 von William Gosset veröffentlicht, während Gosset in einer Guinness-Brauerei arbeitete. Da sein Arbeitgeber die Veröffentlichung nicht gestattete, veröffentlichte Gosset sie unter dem Pseudonym *Student*.

Dichte der t-Verteilung



T-Test

Gepaarter t-test

Ein-Stichproben t-test

Zwei-Stichproben t-Test, ungepaart mit gleichen Varianzen

Welch-t-Test, die Varianzen dürfen ungleich sein

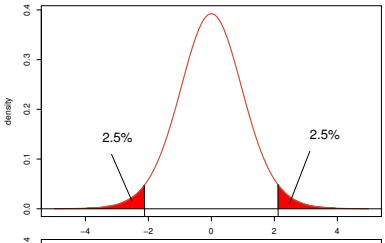
T test : Zweiseitig oder einseitig testen?
In den meisten Fällen will man testen, ob zwei Stichproben sich signifikant unterscheiden. ⇝ zweiseitiger Test

In manchen Fällen

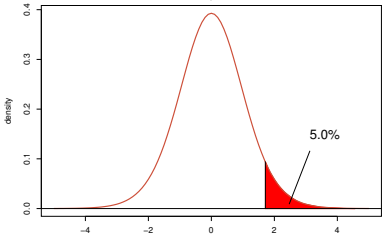
- kann man von vornherein ausschließen, dass die erste Stichprobe kleinere Werte als die zweite Stichprobe hat. Dann will man testen, ob die erste Stichprobe signifikant größer ist.
- will man nur testen, ob der Mittelwert der ersten Stichprobe signifikant größer ist (bzw. kleiner).

⇝ einseitiger Test

T test : Zweiseitig oder einseitig testen?
Wir beobachten einen Wert x, der deutlich größer als der H_0-Erwartungswert μ ist.



p-Wert=$\Pr_{H_0}(|X-\mu| \geq |x-\mu|)$



p-Wert=$\Pr_{H_0}(X-\mu \geq x-\mu)$

4 Chi-Quadrat-Verteilung

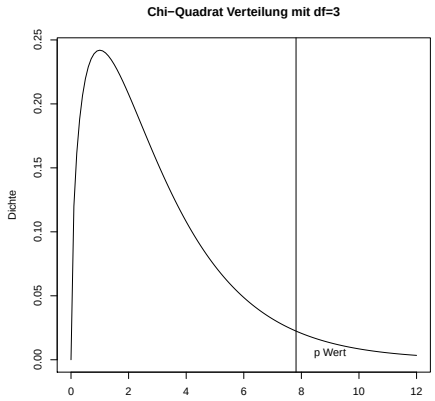
Chi-Quadrat-Verteilung
Seien $X_1, X_2, \ldots, X_n$ n unabhängige standardnormalverteilte Zufallsvariablen, so ist

$$Y = \sum_i X_i^2$$

Chi-Quadrat-verteilt mit n Freiheitsgraden.

$$\mathbb{E}Y = n$$
$$\text{Var } Y = 2n$$

Dichte der Chi-Quadrat-Verteilung



Chi-Quadrat-Test

Gegeben Abweichungen zwischen Daten und eine Verteilung oder zwischen zwei Verteilungen. Wir messen die Abweichungen durch die X^2-Statistik:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

wobei E_i = erwartet Anzahl in Klasse i und O_i = beobachtete (engl. *observed*) Anzahl in Klasse i.

Falls die Nullhypothese gilt und die Erwartungswerte E_i nicht zu klein sind (Faustregel: sie sollten alle ≥ 5 sein), ist X^2 *ungefähr* χ^2*-verteilt*. Die χ^2-Verteilung hängt ab von der Anzahl der Freiheitsgrade **df**.

Nochmal zur t-Verteilung

Angenommen:

- $Z \sim \mathcal{N}(0, 1)$
- X sei χ^2-verteilt mit n Freiheitsgraden.

Dann ist $\frac{Z}{\sqrt{X/n}}$ t-verteilt mit n Freiheitsgraden.

5 F-Verteilung

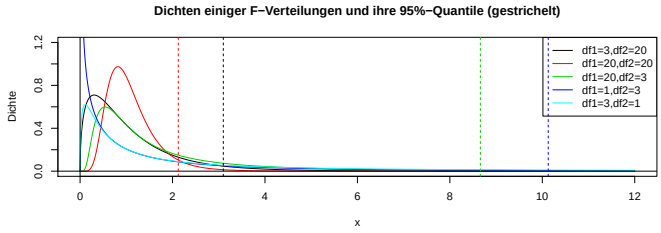
F-Verteilung

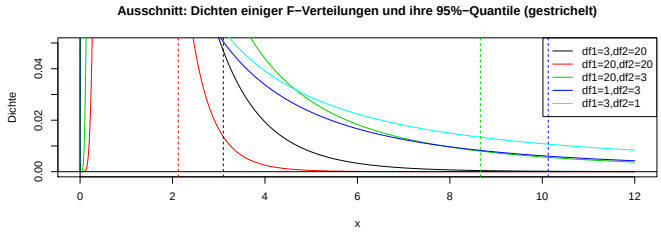
Sind X und Y unabhängige χ^2-verteilte Zufallsvariablen mit Freiheitsgraden m für X und n für Y, so ist

$$F = \frac{X/m}{Y/n}$$

F-verteilt mit m und n Freiheitsgraden.

Dichte der F-Verteilung





Das 95%-Quantil der F-Verteilung mit $df_1 = 3$ und $df_2 = 1$ passte leider nicht in diese Abbildung.
Es beträgt 215.7

F-Test

$X_{ij} = j$-te Beobachtung in der i-ten Gruppe, $j = 1, \ldots, n_i$, Modellannahme: $X_{ij} = \mu_i + \varepsilon_{ij}$.
$\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

$SS_{\text{innerh}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$ Quadratsumme innerhalb d. Gruppen, $n - I$ Freiheitsgrade

$SS_{\text{zw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$ Quadratsumme zwischen d. Gruppen, $I - 1$ Freiheitsgrade

$$F = \frac{SS_{\text{zw}}/(I-1)}{SS_{\text{innerh}}/(n-I)}$$

Unter der Hypothese $H_0 : \mu_1 = \cdots = \mu_I$ („alle μ_i sind gleich“) ist F Fisher-verteilt mit $I-1$ und $n-I$ Freiheitsgraden
(unabhängig vom tatsächlichen gemeinsamen Wert der μ_i).

F-Test

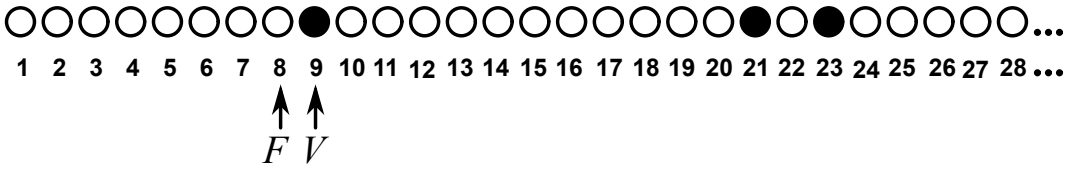
F-Test: Wir lehnen H_0 zum Signifikanzniveau α ab, wenn $F \geq q_\alpha$, wobei q_α das $(1-\alpha)$-Quantil der Fisher-Verteilung mit $I-1$ und $n-I$ Freiheitsgraden ist.

6 Geometrische Verteilung und Exponentialverteilung

Sei $X_1, X_2, \ldots$ eine (unendlich lange) Bernoulli-Kette.

$V = \min\{i : X_i = 1\}$ die Anzahl der Versuche bis zum ersten Erfolg und

$F = V - 1$ die Anzahl der Fehlversuche bis zum ersten Erfolg.



Dann gilt für $i \in \mathbb{N} = \{1, 2, 3, \dots\}$ und $j \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$

$$\begin{aligned} \Pr(V = i) &= (1-p)^{i-1} \cdot p \\ \Pr(F = j) &= (1-p)^{j} \cdot p \end{aligned}$$

V ist geometisch verteilt auf $\mathbb{N}$.
F ist geometisch verteilt auf $\mathbb{N}_0$.

$$\begin{aligned} \Pr(V = i) &= (1-p)^{i-1} \cdot p \quad \text{für } i \in \mathbb{N} \\ \Pr(F = j) &= (1-p)^{j} \cdot p \quad \text{für } j \in \mathbb{N}_0 \end{aligned}$$

Anwendungsbeispiele für die geometrische Verteilung:

- Anzahl Würfelwürfe bis zur ersten 6
- Positionen DNA bis zum nächsten SNP

Eigenschaften der geometrischen Verteilung

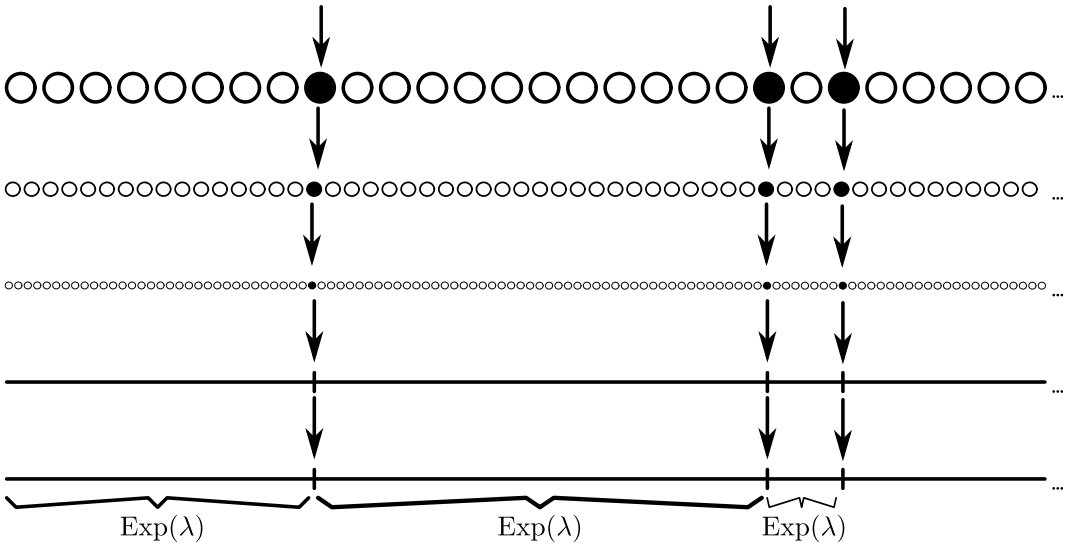
Gedächtnislosigkeit

$$\begin{aligned} \Pr(V = k + i \mid V > k) &= \Pr(V = i) \quad \text{für } i \in \mathbb{N} \\ \Pr(F = k + j \mid F \geq k) &= \Pr(F = j) \quad \text{für } j \in \mathbb{N}_0 \end{aligned}$$

Erwartungswert

$$\begin{aligned} \mathbb{E}V &= p \cdot \mathbb{E}(V|V=1) + (1-p) \cdot \mathbb{E}(V|V>1) = p + (1-p) \cdot (1 + \mathbb{E}V) \\ \Rightarrow \quad \mathbb{E}V &= \frac{1}{p} \quad \text{und} \quad \mathbb{E}F = \mathbb{E}V - 1 = \frac{1-p}{p} \end{aligned}$$

Übergang zu kontinuierlicher Zeitskala bei sehr kleinen p



Beispiele aus der Populationsgenetik:

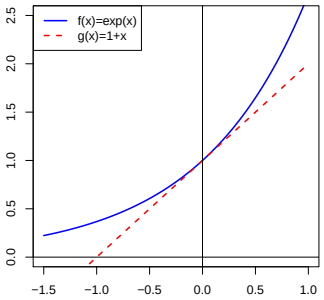
- Positionen auf Chromosom $\longrightarrow$ Centimorgan
- Generationen zum gemeinsamen Vorfahren zweier Allele $\longrightarrow$ Millionen Jahre (oder N_e Generationen als Zeiteinheit)

Geometrische Verteilung $\longrightarrow$ Exponentialverteilung

$$\begin{aligned} \Pr(V = i) &= (1-p)^{i-1} \cdot p \qquad \Pr(V \geq i) = (1-p)^{i-1} \\ \Pr(F = j) &= (1-p)^{j} \cdot p \qquad \Pr(F \geq j) = (1-p)^{j} \end{aligned}$$

Wenn p sehr klein ist und damit $1-p \approx 1$, kommt es auf einen einzelnen Faktor $(1-p)$ mehr oder weniger nicht an und es gilt für große k:

$$(1-p)^{k+1} = \Pr(F > k) \approx \Pr(F \geq k) = \Pr(V > k) = (1-p)^k \approx e^{-p \cdot k}$$



Für $\exp(x) = e^x$ gilt:

$$\begin{array}{rrcll} & \exp(0) & = & 1 & \\ & \exp'(0) & = & \exp(0) & = 1 \\ \Rightarrow & \exp(h) & \approx & 1+h & \text{falls } h \approx 0 \end{array}$$

Es gilt sogar: Wenn $p \to 0$ und $n \to \infty$ so dass $pn \to \lambda$, dann

$$(1-p)^n \to e^{-\lambda}.$$

Exponentialverteilung

Eine Zufallsvariable X mit Werten in $\mathbb{R}_+ = (0, \infty)$ ist exponentialverteilt mit Rate λ (kurz: Exp(λ)-verteilt), wenn für alle $t \in \mathbb{R}_+$ gilt:

$$\Pr(X > t) = e^{-\lambda \cdot t}$$

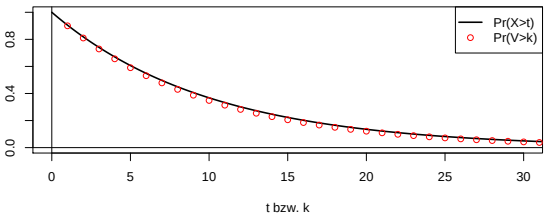
Die Dichte der Exponentialverteilung ist

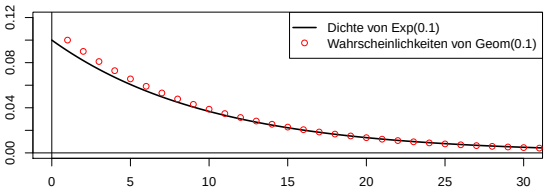
$$f(x) = \lambda \cdot e^{-\lambda x}$$

und es gilt

$$\mathbb{E}X = \sigma_X = 1/\lambda.$$

Vergleich: $X \sim \mathsf{Exp}(0.1)$, $V \sim \mathsf{Geom}(0.1)$ auf $\mathbb{N}$





7 Poisson-Verteilung

Poisson-Verteilung

Binomialverteilt: Wieviele Ereignisse aus einer Bernoulli-Kette der Länge n treten ein?

Poisson-verteilt: Wieviele der mit Rate r kommenden Ereignisse treten bis zur Zeit t ein?

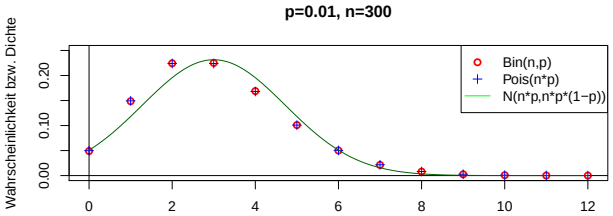
Eine Zufallsvariable Y auf $\mathbb{N}_0$ ist Poisson-verteilt mit Rate λ, wenn gilt

$$\Pr(Y = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

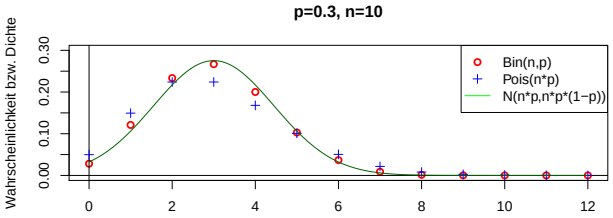
Es gilt dann $\mathbb{E}Y = \mathsf{Var}(Y) = \lambda$.

Poisson-Approximation der Binomialverteilung

Wenn p klein und n groß ist, lässt sich die Binomialverteilung gut durch die Poisson-Verteilung mit $\lambda = n \cdot p$ approximieren.



Für größere p ist die Normalapproximation in der Regel besser geeignet.



Wenn n nicht bekannt ist, ist es bei der Binomialverteilung oft schwer zu schätzen. Vorteil der Poisson-Approximation: nur ein Parameter ($\lambda = np$) muss geschätzt werden.

Anwendungsbeispiele der Poisson-Verteilung (evtl. als Approximation):

- Anzahl an Substitutionen
  - von Nukleotiden einem Chromosom innerhalb einer Generation
  - von Aminosäuren in einem Gen beim Vergleich zweier Arten
- Anzahl der Atome in einer radioaktiven Substanz, die in der nächsten Sekunde zerfallen.
- Wie oft fällt während dieser Vorlesung ein Stift vom Tisch?
- historisches Beispiel: Anzahl der Reitunfälle innerhalb eines Jahres bei der preußischen Armee

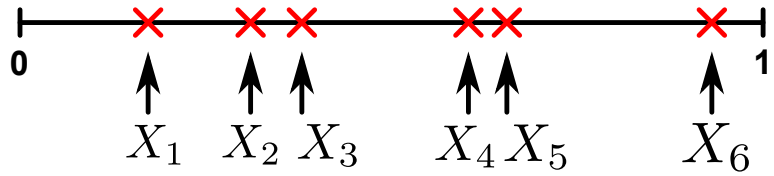
8 Beta-Verteilung (und uniforme Verteilung)

Bedeutung der Beta(a,b)-Verteilung mit $a, b \in \mathbb{N}$

Angenommen, die Anzahl n der Ereignisse im Intervall $[0, 1]$ ist vorgegeben, entweder

- weil wir bei "Ratenprozess" auf diese Anzahl n bedingen oder
- weil wir n unabhängige uniform-verteilte Werte aus $[0, 1]$ gezogen und der Größe nach geordnet haben.

z.B. für $n = 6$:



Dann ist der kt-kleinste Wert X_k Beta$(k, n-k+1)$-verteilt.

Wissenswertes über die Beta-Verteilung

$X \sim \mathsf{Beta}(1,1)$: uniforme Verteilung auf $[0,1]$, d.h. für $0 \le v \le w \le 1$ ist $\Pr(X \in [v,w]) = w - v$.

Für $a, b \in \mathbb{R}_+$ ist die Dichte von Beta(a,b)

$$f(x) = \frac{x^{a-1} \cdot (1-x)^{b-1}}{B(a,b)},$$

mit der Beta-Funktion $B(a,b)$ im Nenner (die dafür sorgt, dass die Fläche unter der Kurve 1 wird).

Ist p die Erfolgswahrscheinlichkeit bei einer Bernoulli-Kette und verwenden wir für die Schätzung von p einen Beta(a, b)-Prior und beobachten dann k Erfolge und m Misserfolge, dann erhalten wir Beta($a+k$, $b+m$) als a-posteriori-Verteilung für p.

9 Hypergeometrische Verteilung

Wenn man aus einer Urne, in der r rote und $m-r$ weiße Kugeln liegen, k Kugel zufällig zieht, ist Anzahl A der roten gezogenen Kugeln,

- falls man **mit** Zurücklegen zieht, binomialverteilt mit $n = k$ und $p = r/m$ und,
- falls man **ohne** Zurücklegen zieht, **hypergeometrisch** verteilt, und es gilt dann

$$\Pr(A = a) = \frac{\binom{k}{a} \cdot \binom{m-k}{r-a}}{\binom{m}{r}}.$$

Anwendungsbeispiele:

- Fishers exakter Test
- Anzahl der Richtigen beim Lotto

Was Sie u.a. erklären können sollten

- Zusammenhänge Normalverteilung, Chi-Quadrat-Verteilung, t-Verteilung, Fisher-Verteilung
- Approximationen der Binomialverteilung durch Normal- oder Poisson-Verteilung
- Was ist bei Bernoulli-Ketten und Urnenexperimenten jeweils binomialverteilt, geometrisch verteilt und hypergeometrisch verteilt?
- Was ist bei "Ratenprozessen" exponentialverteilt, was ist Poisson-verteilt und wie kommt man zu Beta-Verteilungen?
- Rolle der verschiedenen Verteilungen bei statistischen Tests bzw. in der Bayesschen Statistik.