5. Stochastische Modelle I: Diskrete Zufallsvariablen

Andreas Eberle
Institut für angewandte Mathematik

Oktober 2008

Zufallsgrößen

Eine **Zufallsgröße** X ist eine Größe, deren Wert wir nicht exakt kennen bzw. vorhersagen können (aufgrund unbekannter Einflüsse, mangelnder Information, oder echten Zufalls). Wir können den möglichen Werten nur **Wahrscheinlichkeiten** zuordnen. Genau wie Merkmale können Zufallsgrößen auch sowohl Werte mit quantitativer als auch mit qualitativer Ausprägung annehmen (z.B. könnte der Wert einer Zufallsgröße X, die das Geschlecht einer Versuchsperson beschreibt mit Wahrscheinlichkeit $1/2$ "weiblich" und mit Wahrscheinlichkeit $1/2$ "männlich" sein).

- Für die mathematische Modellierung ist es zweckmäßig zwischen **diskreten** und **kontinuierlichen** Zufallsgrößen zu unterscheiden.
- Eine **diskrete Zufallsgröße** X nimmt nur bestimmte Werte

$$a_1, a_2, a_{3,...}$$

an. Die Anzahl der möglichen Werte ("Merkmalsausprägungen") kann endlich oder abzählbar unendlich sein (z.B. könnte jede ganze Zahl als Wert auftreten).

- Eine **kontinuierliche Zufallsgröße** X kann dagegen jede reelle Zahl als Wert annehmen. Die Wahrscheinlichkeit, daß **genau** eine bestimmte Zahl a als Wert auftritt (mit allen Nachkommastellen) ist gleich 0.

- Im mathematischen Modell beschreibt man eine Zufallsgröße X durch eine Funktion $X(\omega)$, die von einem "Zufallsparameter" ω abhängt. Eine solche Funktion heißt eine **Zufallsvariable**. Wir gehen hier (zunächst) nicht auf die genaue Definition ein, und verwenden den Begriff eher intuitiv.

5.1. Diskrete Wahrscheinlichkeitsverteilungen

Wir betrachten nun eine diskrete Zufallsvariable X mit Wertebereich

$$W = \{a_1, a_2, a_{3,\ldots}\}.$$

Definition

Eine **Wahrscheinlichkeitsverteilung** auf W ist festgelegt durch Gewichte (Elementarwahrscheinlichkeiten) $p(a_i) \geq 0$ mit

$$\sum_i p(a_i) = 1$$

Wir sagen, daß die Zufallsvariable X die **Verteilung** bzw. **Massenfunktion** p_X hat, falls gilt:

$$P[X = a_i] = p_X(a_i)$$

Diskrete Wahrscheinlichkeitsverteilungen

Beispiel 1: Gleichverteilung

- Gibt es nur eine endliche Anzahl möglicher Werte $a_1, a_2, \ldots, a_m$, dann wird durch

$$p(a_i) = \frac{1}{m} = \frac{1}{\text{Anzahl der möglichen Werte}}$$

  eine Wahrscheinlichkeitsverteilung festgelegt, unter der jeder mögliche Wert *dieselbe Wahrscheinlichkeit* hat. Eine Zufallsvariable mit einer solchen Verteilung heißt **gleichverteilt**.

- **Beispiele:** Augenzahl beim Würfeln ($p(i) = 1/6$), Entnehmen einer *Zufallsstichprobe* aus einer Grundgesamtheit ($p(\omega_i) = 1/n$).

- **Warnung:** In der Umgangssprache wird der Begriff *"zufällig"* oft gleichbedeutend mit *"gleichverteilt"* verwendet. In der Wahrscheinlichkeitstheorie bezeichnen wir aber jeden Vorgang, der vom Zufall beeinflusst ist, als zufällig - die Wahrscheinlichkeiten der Ausgängen müssen nicht unbedingt alle gleich groß sein !

- Die Gleichverteilung wird häufig dann zur stochastischen Modellierung verwendet, wenn man keine Information über die Wahrscheinlichkeiten der einzelnen Ausgänge hat, und daher davon ausgeht, daß alle Ausgänge gleich wahrscheinlich sind.
- Dabei ist allerdings **Vorsicht** geboten: Fassen wir z.B. mehrere mögliche gleich wahrscheinliche Ausgänge/Merkmalsausprägungen zu einer neuen Merkmalsausprägung zusammen (*Klasseneinteilung*), dann sind die neuen Merkmalsausprägungen (Klassen) nicht mehr unbedingt gleich wahrscheinlich !
- In vielen praktischen Anwendungen ist eine natürliche Einteilung in Merkmalsausprägungen nicht offensichtlich, d.h. es ist nicht klar, was genau gleich wahrscheinlich sein sollte.

Diskrete Wahrscheinlichkeitsverteilungen

Beispiel 2: Empirische Verteilung

Sind $x_1, x_2, \ldots, x_n$ die Ausprägungen eines Merkmals in einer Grundgesamtheit oder einer Stichprobe, dann wird durch die relativen Häufigkeiten

$$p(a_k) = h_k = \frac{n_k}{n}$$

eine Wahrscheinlichkeitsverteilung auf der Menge $\{a_1, a_2, \ldots, a_m\}$ aller möglichen Merkmalsausprägungen festgelegt.

Definition

Die Verteilung mit Massenfunktion p heißt **empirische Verteilung der Grundgesamtheit bzw. der Stichprobe**.

$$p(a_k) = h_k = \frac{n_k}{n}$$

- Die Gewichte $p(a_k)$ der empirischen Verteilung der *Grundgesamtheit* geben an, mit welcher Wahrscheinlichkeit wir den Wert a_k erhalten, wenn wir eine einzelne (einelementige) Zufallsstichprobe aus der Grundgesamtheit entnehmen. Eine Zufallsvariable X mit Verteilung p ist also das mathematische Modell für das Zufallsexperiment *"Entnehmen einer Zufallsstichprobe aus der Grundgesamtheit"*.
- Wenn unsere ganze Information über ein Zufallsexperiment aus einer statistischen Untersuchung stammt, dann kann es naheliegend sein, in zukünftigen mathematischen Modellen anzunehmen, daß die Wahrscheinlichkeiten der möglichen Merkmalsausprägungen gleich den beobachteten relativen Häufigkeiten in der Stichprobe sind. Wir könnten also die (unbekannte) empirische Verteilung der Grundgesamtheit in unserem Modell näherungsweise durch die *empirische Verteilung der Stichprobe* ersetzen.

Diskrete Wahrscheinlichkeitsverteilungen

Beispiel 3: Bernoulliverteilung

- Wir betrachten eine Zufallsgröße X, die angibt, ob eine bestimmte Merkmalsausprägung vorliegt/ ein bestimmtes Ereignis eintritt ($X = 1$) oder nicht ($X = 0$). Ist p die Erfolgswahrscheinlichkeit, dann gilt

$$p_X(1) = P[X = 1] = p, \qquad p_X(0) = P[X = 0] = 1 - p.$$

- Die entsprechende Wahrscheinlichkeitsverteilung auf dem Wertebereich $\{0, 1\}$ heißt **Bernoulliverteilung mit Erfolgswahrscheinlichkeit p**.
- **Beispiele:**
  - Wählerbefragung: $X = 1$ falls sich ein zufällig ausgewählter Wähler für den Kandidaten entscheidet; p beschreibt den Stimmenanteil des Kandidaten unter allen Wählern.
  - Tieruntersuchung: $X = 1$ falls ein zufällig ausgewähltes Tier krank ist; $p =$ Anteil der kranken Tiere.

- Angenommen, wir untersuchen nun eine Zufallsstichprobe von n Tieren. Sei $X_i = 1$, falls das i-te untersuchte Tier krank ist, und $X_i = 0$ sonst. Wir nehmen an, daß die n untersuchten Tiere unabhängig voneinander aus der Gesamtheit ausgewählt wurden. Dies ist normalerweise nicht exakt der Fall (da z.B. meistens ein Tier, was schon einmal in der Stichprobe vorkommt, nicht noch ein zweites Mal untersucht wird), gilt aber oft in guter Näherung. Ist p der Anteil kranker Tiere in der Grundgesamtheit, dann sollte also beispielsweise gelten:

$$\begin{aligned} P[\text{"alle krank"}] &= P[X_1 = 1 \text{ und } X_2 = 1 \text{ und } \ldots\ldots \text{ und } X_n = 1] \\ &= P[X_1 = 1] \cdot P[X_2 = 1] \cdot \ldots\ldots \cdot P[X_n = 1] \\ &= p^n \end{aligned}$$

- ... und entsprechend

$$\begin{aligned} P[\text{"alle gesund"}] &= P[X_1 = 0 \text{ und } X_2 = 0 \text{ und } \ldots\ldots \text{ und } X_n = 0] \\ &= P[X_1 = 0] \cdot P[X_2 = 0] \cdot \ldots\ldots \cdot P[X_n = 0] \\ &= (1-p)^n \end{aligned}$$

Diskrete Wahrscheinlichkeitsverteilungen

Mehrdimensionale Bernoulliverteilung

- Allgemeiner können wir nach der Wahrscheinlichkeit fragen, daß wir für das erste Tier einen gewissen Gesundheitszustand x_1 ($x_1 = 1$ für krank, $x_1 = 0$ für gesund) beobachten, für das zweite Tier den Gesundheitszustand x_2, usw.
- Wegen der Unabhängigkeit erhalten wir:

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= P[X_1 = x_1 \text{ und } X_2 = x_2 \text{ und } \ldots\ldots \text{ und } X_n = x_n] \\ &= \underbrace{P[X_1 = x_1]}_{p \text{ oder } 1-p} \cdot \underbrace{P[X_2 = x_2]}_{p \text{ oder } 1-p} \cdot \cdots\cdots \cdot \underbrace{P[X_n = x_n]}_{p \text{ oder } 1-p} \\ &= p^k \cdot (1-p)^{n-k} \end{aligned}$$

wobei k die Anzahl der Einsen (für kranke Tiere) unter den Beobachtungswerten ist.

- Also z.B.

$$p(0,1,0,0,0,0) = p \cdot (1-p)^5, \quad p(1,1,0,0,1) = p^3 \cdot (1-p)^2.$$

5.2. Verteilungen für Häufigkeiten

- Wir betrachten nun allgemein n unabhängige Ereignisse, von denen jedes mit Wahrscheinlichkeit p eintritt.
- Die Anzahl N der Ereignisse, die davon eintreten, ist dann eine Zufallsvariable mit Werten in

$$W = \{0, 1, 2, \ldots, n\} .$$

Beispiel. (Zufallsstichproben, WICHTIG !)

Wir betrachten (wie im Beispiel oben) die Merkmalsausprägungen $X_1, X_2, \ldots, X_n$ von n einzelnen unabhängigen Zufallsstichproben aus einer Grundgesamtheit. Ist p die relative Häufigkeit einer bestimmten Merkmalsausprägung a in der **Grundgesamtheit**, *dann sind die Ereignisse "$X_1 = a$", "$X_2 = a$",...,"$X_n = a$" unabhängig mit Wahrscheinlichkeit p. Die Anzahl N der Ereignisse, die eintreten, ist gerade die Häufigkeit der Merkmalsausprägung a in der gesamten* **Zufallsstichprobe** $X_1, X_2, \ldots, X_n$.

- Beispielsweise können wir nach der Anzahl N der Wähler fragen, die in einer Zufallsstichprobe von 1000 Wahlberechtigten für Obama gestimmt haben, wenn insgesamt 53% für Obama gestimmt haben.
- In diesem Fall gilt $p = 0.53$ und $n = 1000$.
- Wir würden erwarten, daß von den befragten Wählern ca. $0.53 \times 1000 = 530$ für Obama votiert haben (**Gesetz der großen Zahl**: *rel. Häufigkeit* $\approx$ *Wahrsch'keit*).
- Andererseits wissen wir aber, daß N meistens nicht **genau** den Wert 530 annimmt (**Zufällige Fluktuationen**).
- Um Rückschlüsse von der Stichprobe auf die Grundgesamtheit ziehen zu können, müssen wir wissen, wie weit N "in der Regel" von 530 abweicht, also mit welcher W'keit N einen bestimmten Wert k annimmt.

Verteilungen für Häufigkeiten

Binomialverteilung

Theorem

Wir betrachten n unabhängige Ereignisse mit W'keit p. Dann gilt für die Anzahl N der Ereignisse, die eintreten:

$$P[N = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n).$$

Hierbei ist der Binomialkoeffizient $\binom{n}{k}$ definiert durch

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k \cdot (k-1) \cdot \cdots \cdot 1} = \frac{n!}{k!(n-k)!},$$

wobei $n! = n \cdot (n-1) \cdot \cdots \cdot 1$ das Produkt der Zahlen von 1 bis n bezeichnet.

$$P[N = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n).$$

Herleitung der Formel:

- Wir greifen zunächst aus den Ereignissen k heraus, die eintreten. Dafür gibt es $\binom{n}{k}$ Möglichkeiten, nämlich:
  - n für das erste Ereignis, das wir herausgreifen; $n-1$ für das zweite; .... ; $n-k+1$ für das k-te; also insgesamt $n \cdot (n-1) \cdot \cdots \cdot (n-k+1)$ Möglichkeiten.
  - Da es aber auf die Reihenfolge, in der wir die k Ereignisse herausgreifen, nicht ankommt, müssen wir noch durch die Anzahl möglicher Anordnungen der k Ereignisse teilen, also durch $k \cdot (k-1) \cdot \cdots \cdot 1$.
  - Insgesamt gibt es also

$$\frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k \cdot (k-1) \cdot \cdots \cdot 1} = \binom{n}{k} \quad \text{Möglichkeiten.}$$

- Nun berechnen wir die Wahrscheinlichkeit, daß die k herausgegriffenen Ereignisse eintreten, und die anderen nicht. Wegen der Unabhängigkeit der Ereignisse beträgt diese

$$p^k \cdot (1-p)^{n-k} \qquad \text{(siehe oben)}.$$

- Insgesamt erhalten wir also:

$$\begin{aligned} P[N &= k] \\ &= \sum P[A_{i_1}, A_{i_2}, \ldots, A_{i_k} \text{ treten ein, und die anderen nicht}] \\ &= \sum p^k \cdot (1-p)^{n-k}, \end{aligned}$$

wobei über alle k-elementigen Teilmengen $\{i_1, \ldots, i_k\}$ von $\{1, \ldots, n\}$ summiert wird, also

$$P[N = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Binomialverteilung

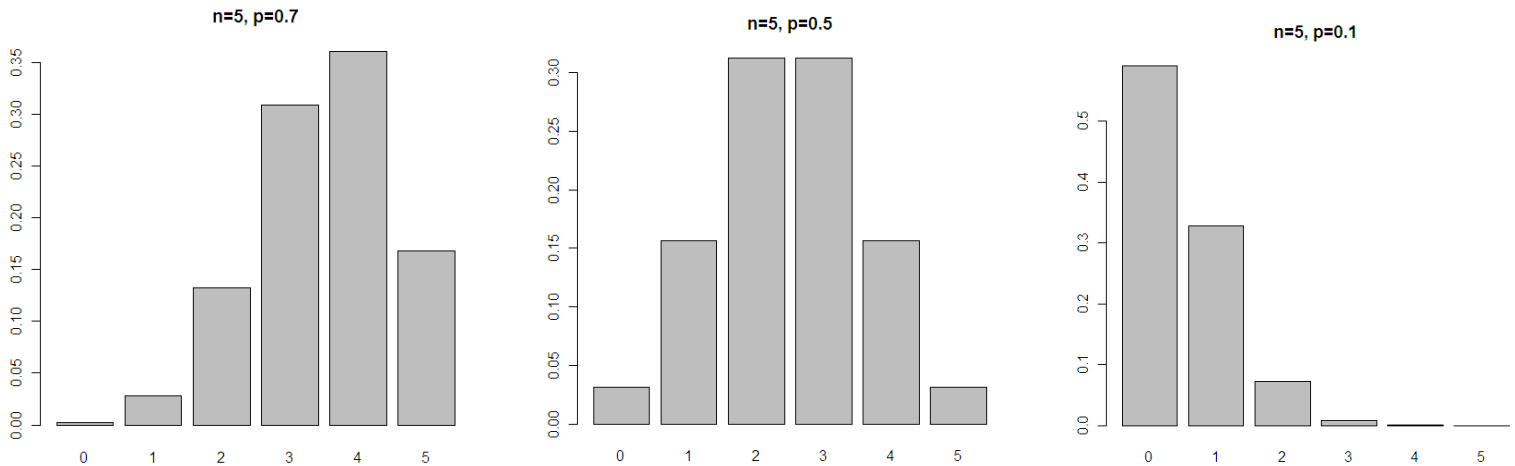
Definition

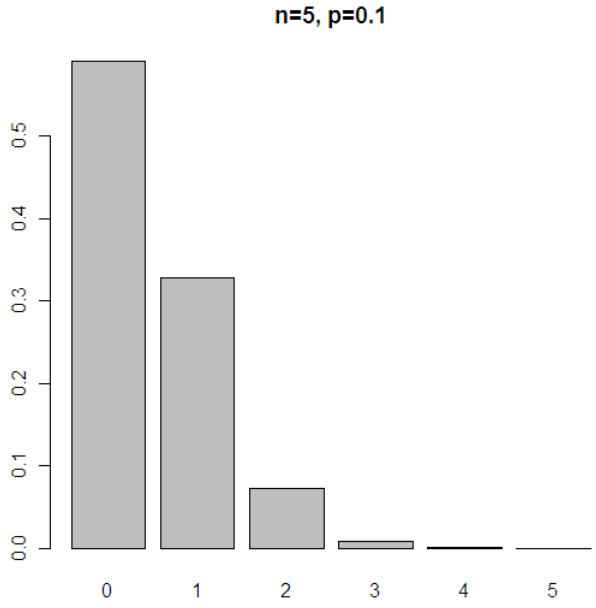
Die Wahrscheinlichkeitsverteilung mit Gewichten

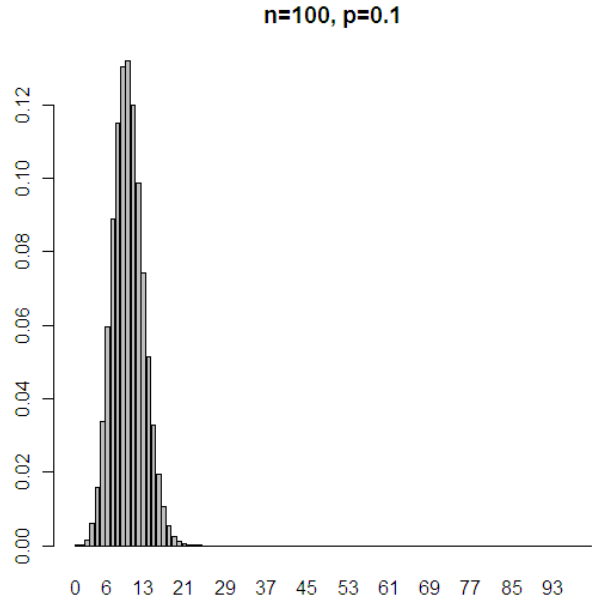
$$p(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n)$$

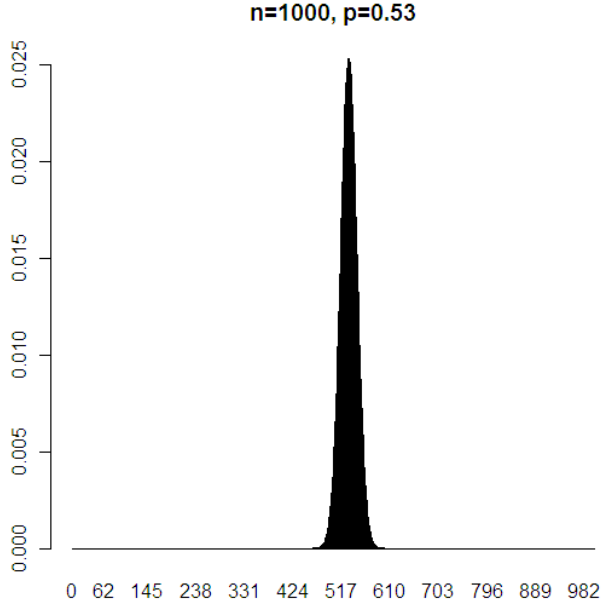
heißt **Binomialverteilung mit Parametern n und p ("Bin(n,p)")**.

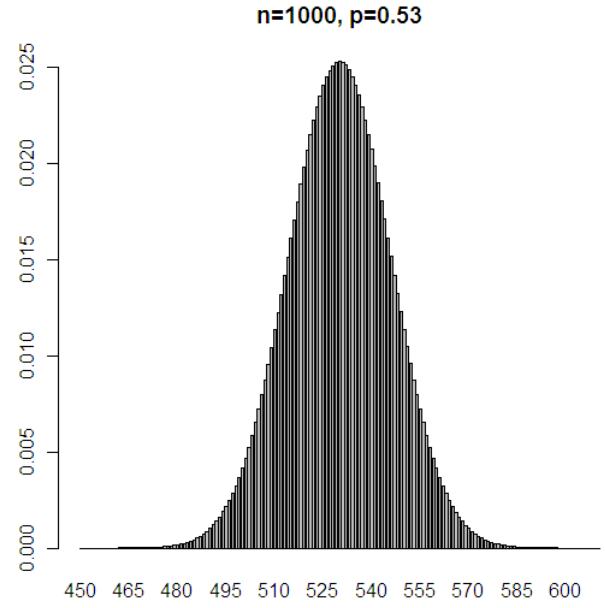
Stabdiagramme von Massenfuntionen von Binomialverteilungen:




n=5, p=0.1
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5


n=100, p=0.1
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0 6 13 21 29 37 45 53 61 69 77 85 93


n=1000, p=0.53
0.000
0.005
0.010
0.015
0.020
0.025
0 62 145 238 331 424 517 610 703 796 889 982


n=1000, p=0.53
0.000
0.005
0.010
0.015
0.020
0.025
450 465 480 495 510 525 540 555 570 585 600

Beispiel.

- *Ein Huhn im Hühnerstall legt an einem bestimmten Tag mit Wahrscheinlichkeit* $p = 70\%$ *genau ein Ei, und mit Wahrscheinlichkeit* $1 - p$ *kein Ei. Insgesamt sind 10 Hühner im Stall. Wie groß ist die Wahrscheinlichkeit, daß an einem Tag weniger als 4 Eier gelegt werden ?*
- *Wir nehmen an, daß die Ereignisse "Huhn* i *legt Ei",* $1 \leq i \leq 10$*, unabhängig sind. Dann ist die Anzahl* N *der gelegten Eier binomialverteilt mit Parametern* $n = 10$ *und* $p = 0.7$*.*
- *Also erhalten wir:*

$$\begin{aligned} P[N < 4] &= P[N = 0] + P[N = 1] + P[N = 2] + P[N = 3] \\ &= \binom{10}{0}(0.7)^0(0.3)^{10} + \binom{10}{1}(0.7)^1(0.3)^9 \\ &\quad + \binom{10}{2}(0.7)^2(0.3)^8 + \binom{10}{3}(0.7)^3(0.3)^7 \\ &= 0.0106 = 1.06\% \end{aligned}$$

Folgerungen, die man aus diesem Ergebnis ableiten könnte, wären etwa:

- An ca. einem von 100 Tagen werden nicht genügend Eier da sein.
- Die Wahrscheinlichkeit, daß an zwei Tagen hintereinander nicht genügend Eier da sind, beträgt etwa $\frac{1}{10000}$, ist also vernachlässigbar gering.
- Wenn wir trotzdem beobachten, daß an zwei Tagen hintereinander nicht genügend Eier gelegt werden, dann stimmt vermutlich etwas an unseren Modellannahmen nicht. Zum Beispiel könnte die zugrundeliegende Hypothese, daß jedes Huhn mit Wahrscheinlichkeit 70 % ein Ei legt, falsch sein.

- *Dies ist ein erstes Beispiel für einen* **Hypothesentest** *(schließende Statistik): Wir untersuchen, ob eine dem Modell zugrundeliegende Hypothese ($p = 0.7$) aufgrund der Beobachtungsdaten plausibel ist.*

Verteilungen für Häufigkeiten

Stichproben mit und ohne Zurücklegen

- Wenn wir aus einer Grundgesamtheit n einzelne Zufallsstichproben $\omega_1, \omega_2, \ldots, \omega_n$ entnehmen, und jede einzelne Stichprobe jeweils vor Ziehen der nächsten Stichprobe wieder in die Grundgesamtheit **zurücklegen** (eine statistische Einheit kann dann also mehrmals ausgewählt werden), dann ist die Häufigkeit N einer bestimmten Merkmalsausprägung a unter den beobachteten Merkmalswerten $X_1, X_2, \ldots, X_n$ eine binomialverteilte Zufallsgröße mit Parametern n und p. Hierbei ist p die relative Häufigkeit des Merkmals a in der Grundgesamtheit.

Verteilungen für Häufigkeiten

Stichproben mit und ohne Zurücklegen

- In vielen Fällen wählt man die einzelnen Zufallsstichproben $\omega_1, \omega_2, \ldots, \omega_n$ aber so, daß keine statistische Einheit mehrmals vorkommt (**"Ziehen ohne Zurücklegen"**).
- In diesem Fall sind die einzelnen Stichproben (und damit die beobachteten Merkmalsausprägungen) **nicht mehr unabhängig**: Wenn ω_1 eine bestimmte Einheit ist, dann kann ω_2 nicht dieselbe Einheit sein - es gibt also einen Zusammenhang zwischen ω_1 und ω_2.
- Dementsprechend ist die Binomialverteilung eigentlich NICHT mehr das korrekte Modell für die Häufigkeit einer bestimmten Merkmalsausprägung unter den Beobachtungswerten. Stattdessen muß man die **hypergeometrische Verteilung** verwenden.
- Glücklicherweise unterscheiden sich beide Verteilungen nur sehr wenig, wenn die Anzahl $\widehat{n}$ der Einheiten in der Grundgesamtheit viel größer ist als die Anzahl n der Einzelstichproben, da dann dieselbe Einheit auch bei Zurücklegen nur sehr selten zweimal herausgegriffen wird.

Verteilungen für Häufigkeiten

Hypergeometrische Verteilung

Theorem

Sei $\widehat{n}$ die Anzahl der Einheiten, und $\widehat{k}$ die Häufigkeit des Merkmals a in der Grundgesamtheit. Dann gilt für die Häufigkeit N des Merkmals a in einer Zufallsstichprobe ohne Zurücklegen der Größe n:

$$p_N(k) = P[N = k] = \frac{\binom{\widehat{k}}{k} \cdot \binom{\widehat{n}-\widehat{k}}{n-k}}{\binom{\widehat{n}}{n}}, \qquad k = 0, 1, \ldots, n.$$

Die Wahrscheinlichkeitsverteilung mit den Gewichten $p_N(k)$ heißt **hypergeometrische Verteilung mit Parametern** $n, \widehat{k}, \widehat{n}$.

- Für $\widehat{n} \gg n$ gilt

$$p_N(k) \approx \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \ , \qquad \text{wobei} \quad p = \frac{\widehat{k}}{\widehat{n}}$$

  die relative Häufigkeit des Merkmals a in der Grundgesamtheit ist.

- Daher kann man in der Praxis oft die Binomialverteilung statt der hypergeometrischen Verteilung verwenden, obwohl eigentlich nicht zurückgelegt wird.

Verteilungen für Häufigkeiten

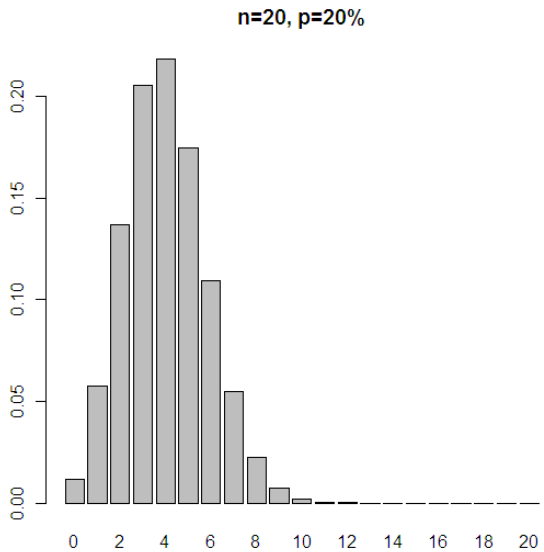
Seltene Ereignisse

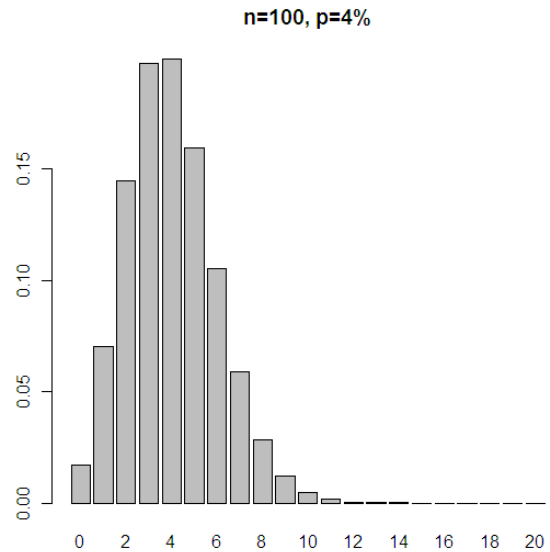
- Wir betrachten nun ein Ereignis, das nur sehr selten (mit einer sehr kleinen Wahrscheinlichkeit p) eintritt,z.B.
  - Schadensfall bei Versicherung
  - radioaktiver Zerfall
  - Tippfehler im Generalanzeiger
  - positiver Test auf Vogelgrippe
  - Börsencrash, Erdbeben
- Wir wollen wissen, wie oft das Ereignis bei einer sehr großen Anzahl n unabhängiger Wiederholungen eintritt.
- Dies können wir im Prinzip mit der Binomial(n,p)-Verteilung ausrechnen. Wenn n sehr groß ist, ist es aber unpraktisch, die Gewichte der Binomialverteilung zu berechnen und aufzusummieren. Stattdessen werden wir die Binomialverteilung durch eine einfachere Verteilung, die *Poissonverteilung*, approximieren.

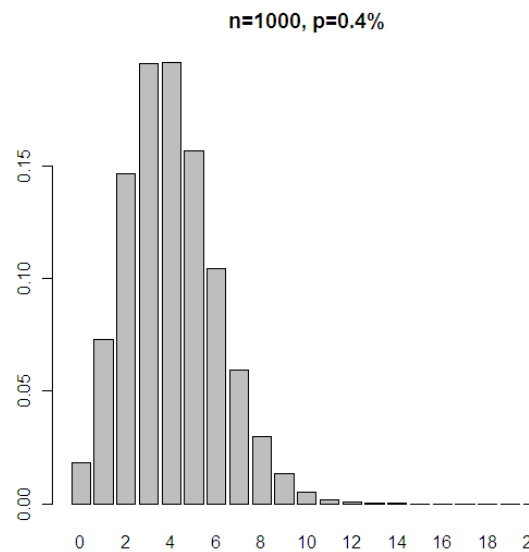
Verteilungen für Häufigkeiten

Poissonapproximation der Binomialverteilung

Um die Approximation zu verstehen, betrachten wir verschiedene Werte für n und p, für die jeweils im Mittel gerade 4 Ereignisse eintreten, d.h. es gilt jeweils $n \cdot p = 4$. Die Stabdiagramme der entsprechenden Binomialverteilungen sehen wie folgt aus:







- Ist die Anzahl n unabhängiger Wiederholungen groß, dann hängt die Massenfunktion offensichtlich kaum noch von n ab, falls der Mittelwert $n \cdot p$ gleich bleibt ! Dies kann man auch beweisen:

-> Interaktive Demonstration dazu siehe

http://www-wt.iam.uni-bonn.de/~eberle/BinomialPoisson.nbp

(benötigt Mathematica Player, kostenloser Download auf Mathematica-Webseite)

Verteilungen für Häufigkeiten

Poissonverteilung

Theorem

Für große n und für p mit $n \cdot p = \lambda$ (also $p = \lambda / n$) gilt

$$\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \approx \frac{1}{k!} \lambda^k \cdot e^{-\lambda}$$

Definition

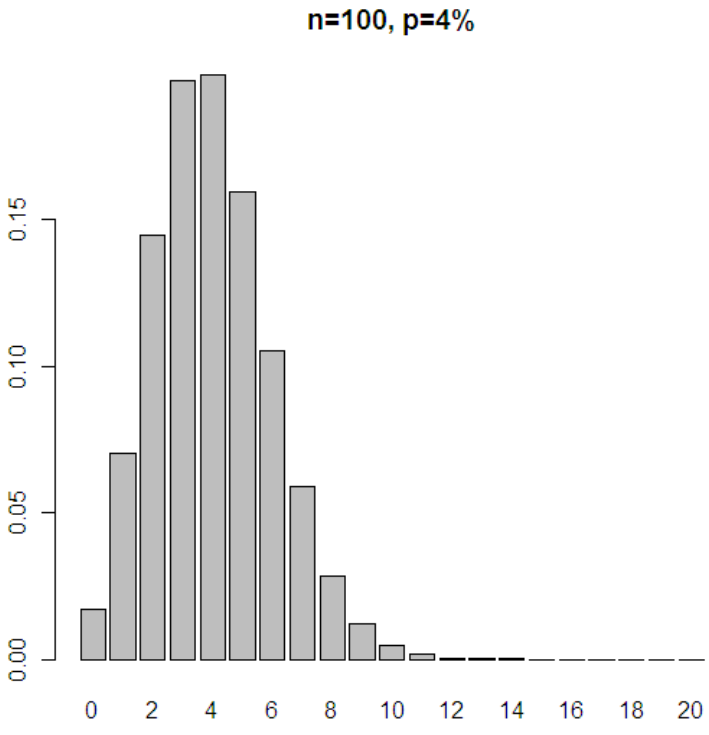
Sei $\lambda > 0$. Die Wahrscheinlichkeitsverteilung mit den Gewichten

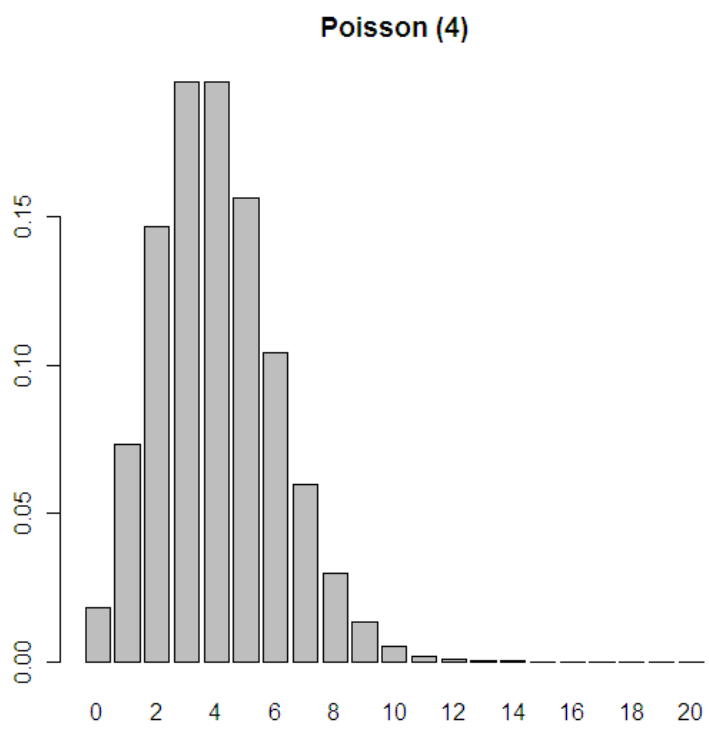
$$p(k) = \frac{1}{k!} \lambda^k \cdot e^{-\lambda}, \qquad k = 0, 1, 2, ...,$$

heißt **Poissonverteilung zum Parameter** λ.

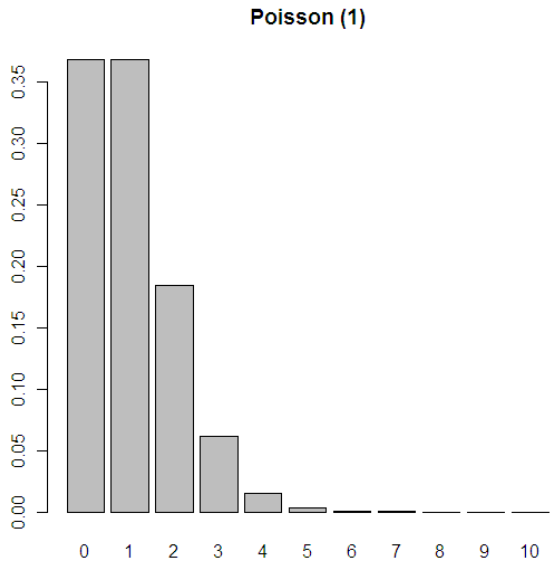
Hierbei gibt λ die mittlere Anzahl von Ereignissen, die eintreten, an.

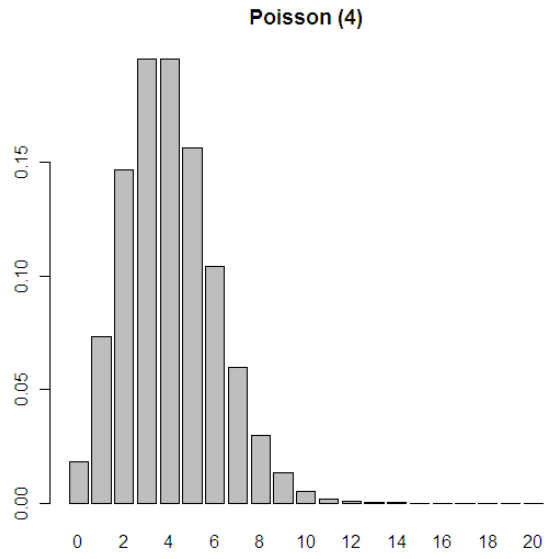
Poissonapproximation der Binomialverteilung:

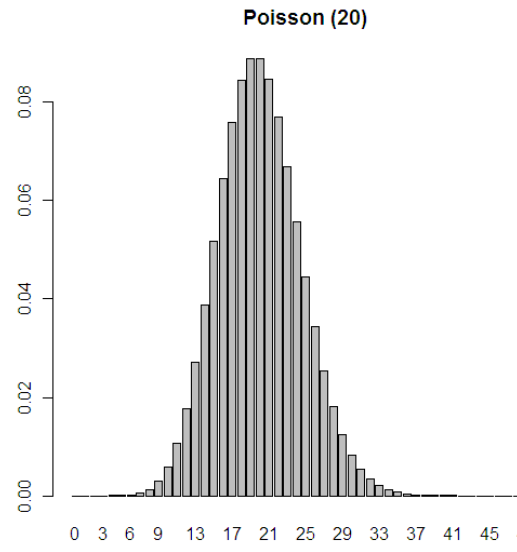




Verschiedene Poissonverteilungen:



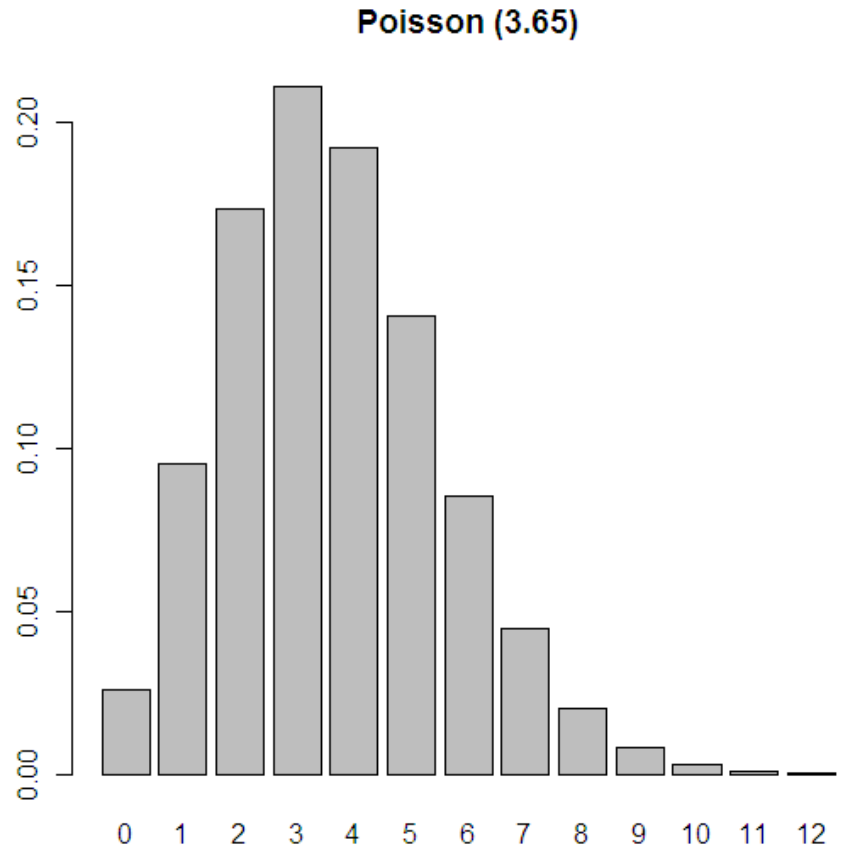




Beispiel. (Hühner im Stall)

- *Die Wahrscheinlichkeit p, daß die Hühner im Hühnerstall an einem Tag nicht genug Eier legen, betrug etwa $\frac{1}{100}$. Wie groß ist die Wahrscheinlichkeit, daß die Hühner an mehr als 5 Tagen im Jahr zuwenig Eier legen ?*
- *Es gibt $n = 365$ Tage im Jahr, also legen die Hühner im Schnitt an $n \cdot p = 3,65$ Tagen im Jahr nicht genug Eier.*
- *Wenn wir annehmen, daß das Eierlegen an verschiedenen Tagen unabhängig voneinander ist, dann ist die Anzahl N der Tage mit zuwenig Eiern in guter Näherung Poissonverteilt mit Parameter $\lambda = 3,65$.*
- *Wir erhalten also:*

$$\begin{aligned} P[N > 5] &= P[N = 6] + P[N = 7] + P[N = 8] + \cdots \\ &\approx \left[\frac{(3.65)^6}{6!} + \frac{(3.65)^7}{7!} + \frac{(3.65)^8}{8!} + \cdots \right] \cdot e^{-3.65} \\ &= [3.28 + 1.71 + 0.78 + \cdots] \cdot e^{-3.65} = 0.16 \end{aligned}$$



- In Wirklichkeit sind die Anzahlen der an verschiedenen Tagen gelegten Eier natürlich nicht unabhängig voneinander. Bei direkt aufeinanderfolgenden Tagen sollte zumindest ein Zusammenhang bestehen. Trotzdem können wir das erhaltene Ergebnis als eine erste Näherung für die gesuchte Wahrscheinlichkeit ansehen.

5.3. Erwartungswert und Varianz

Wir betrachten nun eine diskrete Zufallsvariable X, deren Werte $a_1, a_2, a_3, \ldots$ reelle Zahlen sind. Die Anzahl der möglichen Werte kann endlich oder abzählbar unendlich sein. Sei

$$p_X(a_i) = P[X = a_i]$$

die Massenfunktion/Verteilung von X.

Erwartungswert und Varianz

Definition

Der **Erwartungswert** (Mittelwert, Prognosewert) der Zufallsvariablen X bzw. der Verteilung p_X ist definiert als

$$E[X] = \sum_k a_k \cdot P[X = a_k] = \sum_k a_k \cdot p_X(a_k).$$

Die **Varianz** von X bzw. p_X ist

$$Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2.$$

Die **Standardabweichung** ist

$$\sigma(X) = \sqrt{Var(X)}.$$

$$E[X] = \sum_k a_k \cdot P[X = a_k] = \sum_k a_k \cdot p_X(a_k).$$

- Der Erwartungswert ist also das mit den Wahrscheinlichkeiten $p_X(a_k)$ **gewichtete Mittel** der Werte a_k.
- Allgemeiner können wir den Erwartungswert einer beliebigen Funktion $f(X)$ der Zufallsvariablen X folgendermaßen berechnen:

$$E[f(X)] = \sum_k b_k \cdot P[f(X) = b_k] = \sum_k f(a_k) \cdot P[X = a_k],$$

also

$$E[f(X)] = \sum_k f(a_k) \cdot p_X(a_k)$$

- **Zum Beispiel:**

$$E[X^2] = \sum_k a_k^2 \cdot p_X(a_k), \qquad Var(X) = \sum_k (a_k - E[X])^2 \cdot p_X(a_k)$$

Erwartungswert und Varianz

Beispiel 1: Empirische Verteilung

- Der Erwartungswert einer **empirischen Verteilung** mit relativen Häufigkeiten $h(a_k)$ ist das arithmetische Mittel der Beobachtungswerte:

$$\text{Erw.wert} = \sum_k a_k \cdot h(a_k) = \frac{1}{n} \sum_k a_k \cdot n(a_k) = \frac{1}{n} \sum_i x_i$$

- Entsprechend ist die Varianz einer empirischen Verteilung die empirische Varianz der Beobachtungswerte (*definiert mit Vorfaktor 1/n, nicht mit 1/(n-1)*).

Erwartungswert und Varianz

Beispiel 2: Gleichverteilung

- Der Erwartungswert einer auf $\{1, 2, ..., n\}$ **gleichverteilten** Zufallsvariable X ist

$$\begin{aligned} E[X] &= \sum_{k=1}^{n} k \cdot P[X = k] = \sum_{k=1}^{n} \frac{k}{n} \\ &= \frac{1}{n} \sum_{k=1}^{n} k = \frac{n \cdot (n+1)}{2n} = \frac{n+1}{2} \end{aligned}$$

- Entsprechend ergibt sich

$$E[X^2] = \sum_{k=1}^{n} k^2 \cdot P[X = k] = \frac{1}{n} \sum_{k=1}^{n} k^2 = \frac{(2n+1)(n+1)}{6},$$

also

$$Var(X) = E[X^2] - E[X]^2 = ... = \frac{n^2 - 1}{12}$$

Erwartungswert und Varianz

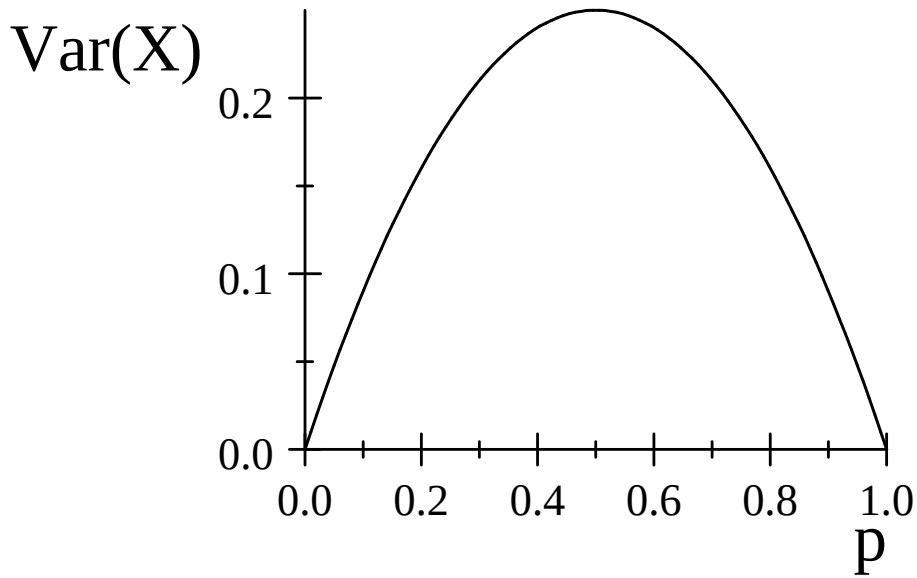
Beispiel 3: Bernoulliverteilung

- Der Erwartungswert einer Bernoulli(p)-verteilten Zufallsvariable X ist

$$E[X] = 1 \cdot P[X = 1] + 0 \cdot P[X = 0] = p.$$

- Für die Varianz erhalten wir:

$$\begin{aligned} E[X^2] &= 1^2 \cdot P[X = 1] + 0^2 \cdot P[X = 0] = p\ , \qquad \text{also} \\ Var(X) &= E[X^2] - E[X]^2 = p - p^2 = p \cdot (1 - p) \end{aligned}$$



Erwartungswert und Varianz

Lineare Transformationen

- Um Erwartungswert und Varianz in anderen Fällen effizient berechnen zu können, bemerken wir zunächst, wie sich diese unter linearen Tranformationen verhalten:

Theorem

Für beliebige Zufallsvariablen X, Y und reelle Konstanten a, b gilt

$$\begin{aligned} E[a \cdot X + b \cdot Y] &= a \cdot E[X] + b \cdot E[Y] \qquad \textit{und} \\ Var(a \cdot X) &= a^2 \cdot Var(X). \end{aligned}$$

Sind X und Y **unabhängig** *oder, allgemeiner,* **unkorreliert** *(s.u.), dann gilt zudem*

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

- Dabei verwenden wir den folgenden Unabhängigkeitsbegriff für diskrete Zufallsvariablen:

Definition

Zwei diskrete Zufallsvariablen X und Y heißen **unabhängig**, wenn für alle möglichen Werte a_k von X und b_l von Y die Ereignisse "$X = a_k$" und "$Y = b_l$" unabhängig sind, d.h. wenn

$$\begin{aligned} p_{X,Y}(a_k, b_l) &= P[X = a_k \text{ und } Y = b_l] \\ &= P[X = a_k] \cdot P[Y = b_l] = p_X(a_k) \cdot p_Y(b_l) \end{aligned}$$

für alle a_k und b_l gilt.

- Erwartungswerte von Zufallsvariablen, die sich durch lineare Transformationen aus anderen Zufallsvariablen ergeben, können wir also aus den Erwartungswerten dieser Zufallsvariablen berechnen.
- Entsprechendes gilt für die Varianzen, falls die zugrundeliegenden Zufallsvariablen unabhängig (bzw. unkorreliert) sind.

Erwartungswert und Varianz

Beispiel 4: Binomialverteilung

- Eine mit Parametern n und p binomialverteilte Zufallsvariable N können wir als Anzahl der Ereignisse interpretieren, die von n unabhängigen Ereignissen $A_1, A_2, \ldots, A_n$ mit W'keit p eintreten.
- Also erhalten wir folgende Darstellung:

$$N = X_1 + X_2 + \cdots + X_n,$$

  wobei die Zufallsvariable X_i den Wert 1 hat, falls das Ereignis A_i eintritt, und den Wert 0 falls nicht.
- Die Zufallsvariablen $X_1, X_2, \ldots, X_n$ sind unabhängig und Bernoulli(p)-verteilt - es gilt also

$$E[X_i] = p \qquad \text{und} \qquad Var(X_i) = p \cdot (1-p)$$

- ... und damit:

$$\begin{aligned} E[N] &= E[X_1] + E[X_2] + \cdots + E[X_n] = n \cdot p \quad , \qquad \text{und} \\ Var(N) &= Var(X_1) + \cdots + Var(X_n) = n \cdot p \cdot (1-p) \end{aligned}$$

Erwartungswert und Varianz

Beispiel 5: Poissonverteilung

- Die Poissonverteilung mit Parameter λ ergibt sich als Grenzwert für $n \longrightarrow \infty$ der Binomialverteilung mit Parametern n und $p = \lambda / n$.
- Die Binomialverteilung mit diesen Parametern hat Erwartungswert $n \cdot p = \lambda$ und Varianz

$$\sigma^2 = n \cdot p \cdot (1 - p) = \lambda \cdot (1 - \frac{\lambda}{n}).$$

- Für $n \longrightarrow \infty$ erhalten wir also als Erwartungswert und Varianz einer Poissonverteilten Zufallsvariable N :

$$E[N] = \lambda \qquad \text{und} \qquad Var(N) = \lim_{n \longrightarrow \infty} \lambda \cdot (1 - \frac{\lambda}{n}) = \lambda.$$

- Dasselbe Ergebnis kann man auch direkt durch nachrechnen aus der Definition von Erwartungswert und Varianz erhalten.

5.4. Normalapproximation der Binomialverteilung

- Ist n groß, dann ist es sehr aufwändig, Wahrscheinlichkeiten für binomialverteilte Zufallsvariablen exakt auszurechnen.
- Ist außerdem noch p klein, also das zugrundeliegende Ereignis selten, dann können wir die Poissonapproximation verwenden.

- Wir wollen uns nun überlegen, wie wir die Binomialverteilung approximieren können, wenn das zugrundeliegende Ereignis **nicht selten** und n **hinreichend groß** ist. Dazu "*standardisieren*" wir eine binomialverteilte Zufallsvariable zunächst auf Erwartungswert 0 und Varianz bzw. Standardabweichung 1:

Normalapproximation der Binomialverteilung

Standardisierte Zufallvariablen

- Eine beliebige reellwertige Zufallsvariable X können wir zentrieren, d.h. in eine Zufallsvariable mit Erwartungswert 0 verwandeln, indem wir den Erwartungswert subtrahieren:

$$\widetilde{X} = X - E[X] \qquad \text{ist Zufallsvariable mit} \quad E\left[\widetilde{X}\right] = 0.$$

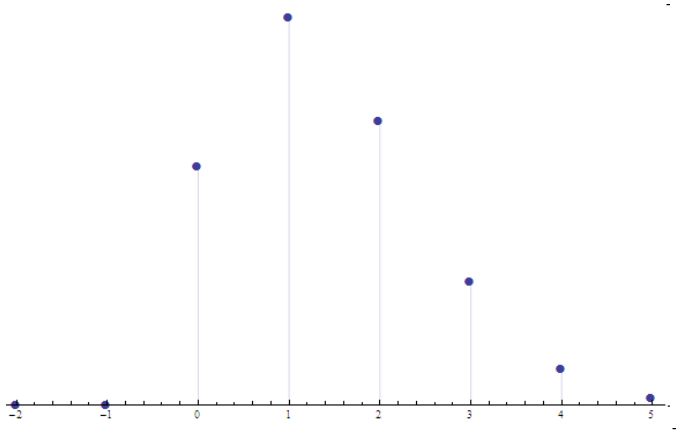
- Wenn wir die zentrierte Zufallsvariable $\widetilde{X}$ anschließend noch durch die Standardabweichung $\sigma(X)$ teilen, erhalten wir eine Zufallsvariable mit Erwartungswert 0 und Standardabweichung 1:

$$Y = \frac{X - E[X]}{\sigma(X)} \quad \text{ist ZV mit} \quad E[Y] = 0 \quad \text{und} \quad \sigma(Y) = 1.$$

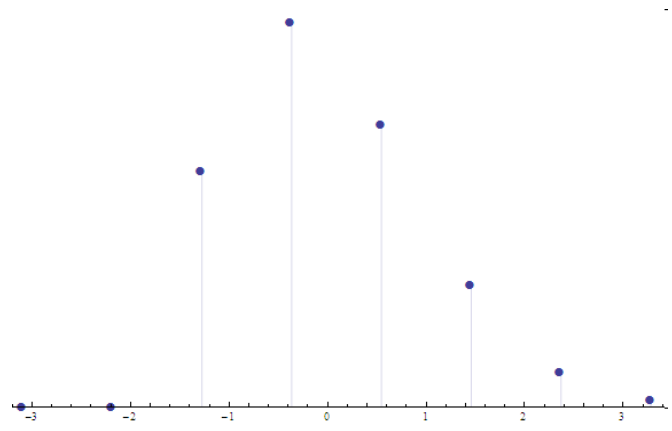
- Ist X binomialverteilt mit Parametern n und p, dann sieht die standardisierte Zufallsvariable so aus:

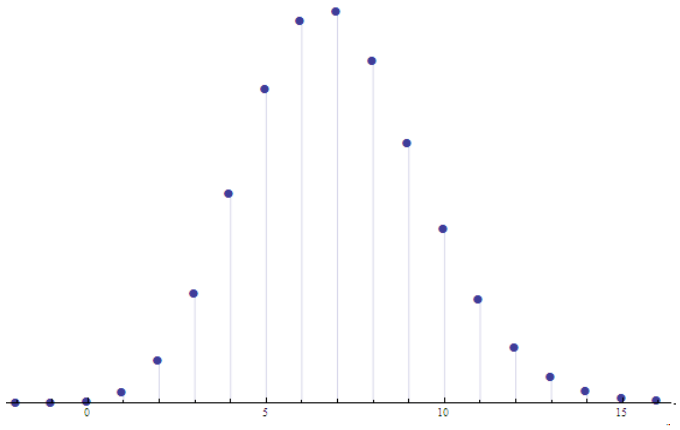
$$Y = \frac{X - np}{\sqrt{np \cdot (1-p)}}.$$

Binomialverteilung und Standardisierung, p=0.14

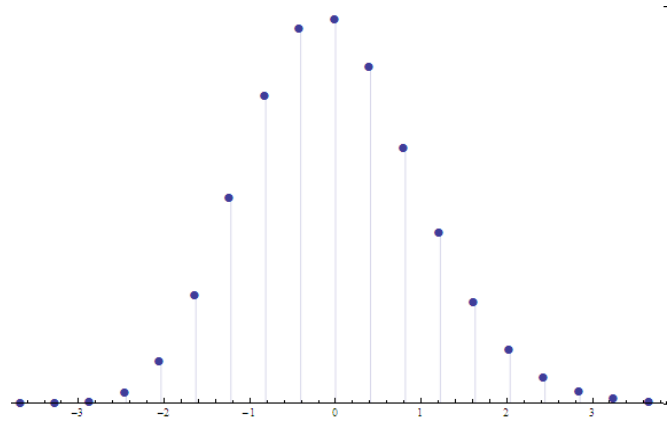


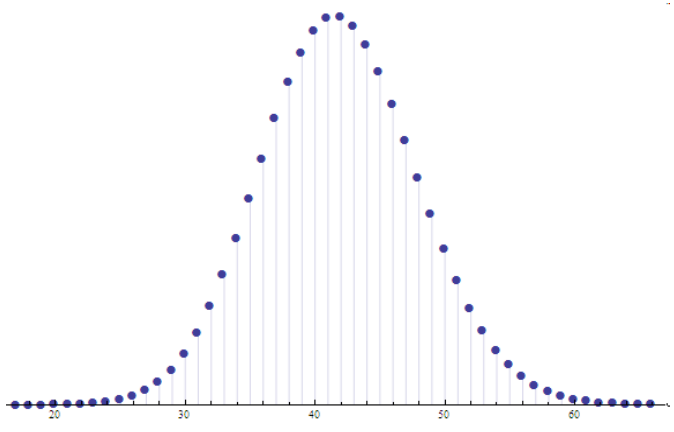
n=10



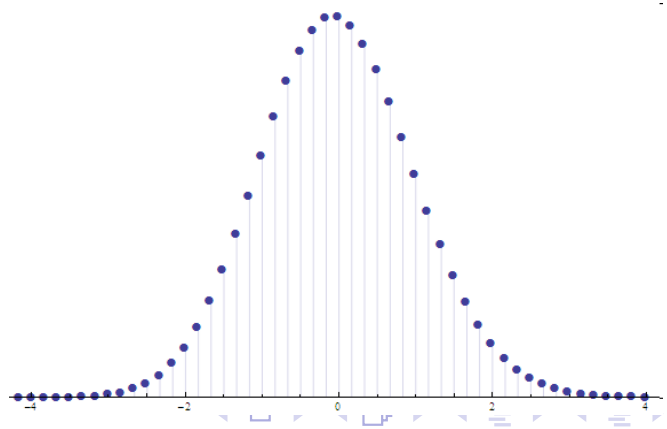


n=50





n=300



-> Interaktive Demonstrationen siehe

http://www-wt.iam.uni-bonn.de/~eberle/BinomialNormal.nbp

http://www-wt.iam.uni-bonn.de/~eberle/BinomialNormalB.nbp
(benötigt Mathematica Player, kostenloser Download auf Mathematica-Webseite)

Normalapproximation der Binomialverteilung

- Man erkennt, daß sich die Massenfunktion einer standardisierten binomialverteilten Zufallsvariable für große n sehr rasch einer "Gaußschen Glockenkurve" annähert.
- Mithilfe von Grenzwertaussagen aus der Analysis (Stirlingsche Formel) kann man dies auch beweisen:

Theorem

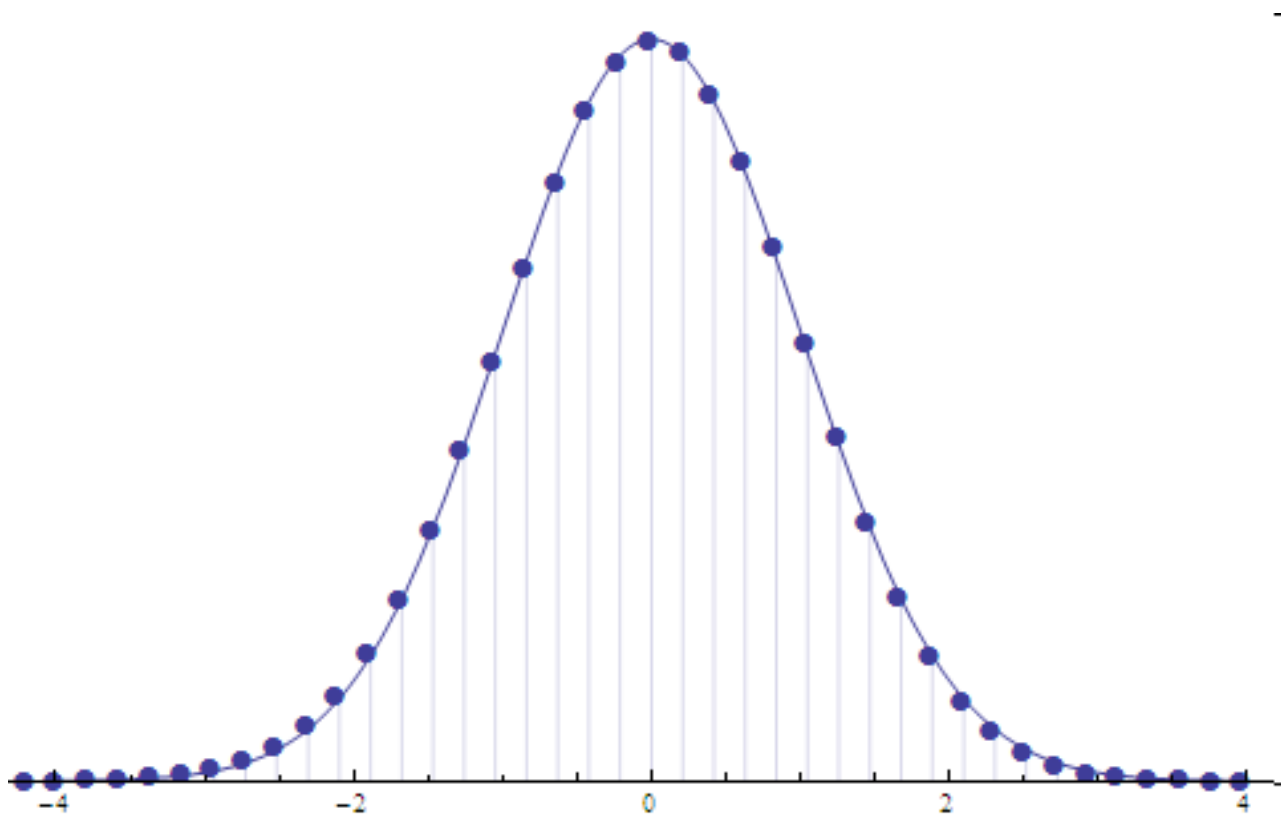
Ist X eine Binomial (n, p) verteilte Zufallsvariable, und n hinreichend groß, dann gilt für die standardisierte Zufallsvariable

$$Y = \frac{X - np}{\sqrt{np(1-p)}}$$

näherungsweise

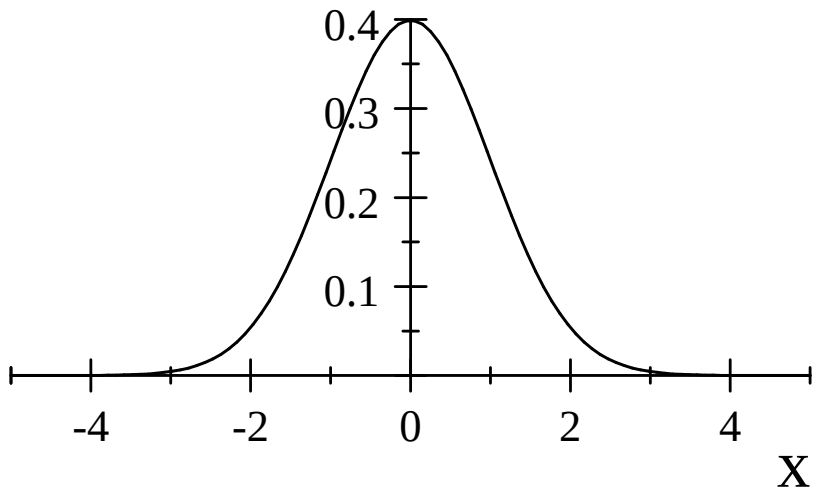
$$P\left[a \leq Y \leq b\right] \approx \int_a^b \underbrace{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}}_{\text{Gaußsche Glockenkurve}} \, dx$$

Standardisierte Binomialverteilung ($n = 100$ und $p = 0.65$) und Gaußsche Glockenkurve (Standardnormalverteilung)



- Ist X eine Binomial (n, p) verteilte Zufallsvariable, und n hinreichend groß, dann gilt für die standardisierte Zufallsvariable $Y = \frac{X - np}{\sqrt{np(1-p)}}$ näherungsweise:

$$P\left[a \leq Y \leq b\right] \approx \int_a^b \underbrace{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}}_{\text{Gaußsche Glockenkurve}} \, dx$$

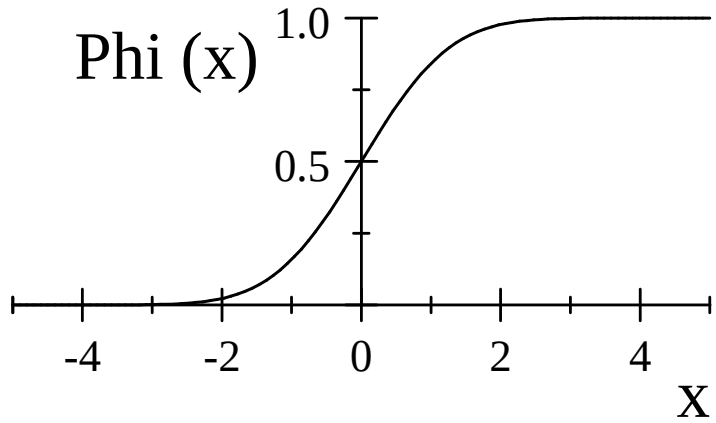


- Die Wahrscheinlichkeit, daß Y zwischen a und b liegt, ist also in etwa durch den **Flächeninhalt** unter der Gaußschen Glockenkurve über dem Intervall $[a, b]$ gegeben.

- Insbesondere erhalten wir

$$P\left[Y \leq x\right] \approx \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy,$$

wobei $\Phi(x)$ die *Stammfunktion* der Gaußschen Glockenkurve ist.



- $\Phi(x)$ ist im allgemeinen nicht explizit berechenbar. Die Werte können approximativ (mit dem Computer) berechnet, oder aus einer Tabelle abgelesen werden.
- Zum Beispiel gilt: $\Phi(-\infty) = 0$, $\Phi(0) = 1/2$,

$$\Phi(1) = 0.841,\ \Phi(2) = 0.977,\ \Phi(3) = 0.9987,\quad \Phi(\infty) = 1.$$

Normalapproximation der Binomialverteilung

Anwendbarkeit der Normalapproximation

- **Faustformel**: Die Näherung ist relativ genau, wenn

$$n \cdot p \geq 10 \qquad \text{und} \qquad n \cdot (1-p) \geq 10 \qquad \text{gilt.}$$

- Ist p sehr klein, dann ist das Ereignis selten, und man sollte besser die **Poissonapproximation** verwenden.
- Entsprechend ist für p nahe bei 1 das Gegenereignis selten, und man sollte eine Poissonapproximation für die Häufigkeit des Gegenereignisses verwenden.

Beispiel.

- *Wie im Beispiel von oben legt ein Huhn im Hühnerstall an einem bestimmten Tag mit Wahrscheinlichkeit* $p = 70\%$ *genau ein Ei, und mit Wahrscheinlichkeit* $1 - p$ *kein Ei. Insgesamt sind jetzt aber 100 Hühner im Stall. Wie groß ist die Wahrscheinlichkeit, daß an einem Tag weniger als 60 Eier gelegt werden ?*
- *Wir nehmen wieder an, daß die Ereignisse "Huhn* i *legt Ei",* $1 \leq i \leq 100$*, unabhängig sind. Dann ist die Anzahl* N *der gelegten Eier binomialverteilt mit Parametern* $n = 100$ *und* $p = 0.7$*.*
- *Die mittlere Anzahl der gelegten Eier beträgt* $E[N] = np = 70$ *und die Standardabweichung* $\sigma(N) = \sqrt{np(1-p)} = \sqrt{21}$*.*
- *Also erhalten wir:*

$$\begin{aligned} P[N < 60] &= P[N - E[N] < 60 - 70] \\ &= P\left[\frac{N - E[N]}{\sigma(N)} < \frac{60 - 70}{\sqrt{21}}\right] \approx \Phi\left(\frac{60 - 70}{\sqrt{21}}\right) = 1.45\% \end{aligned}$$

- Sind dagegen 1000 Hühner im Stall, und wir fragen nach der Wahrscheinlichkeit, daß diese mindestens 600 Eier legen, dann ergibt sich analog:

$$E[N] = np = 700 \quad \text{und} \quad \sigma(N) = \sqrt{np(1-p)} = \sqrt{210}.$$

- Also erhalten wir:

$$\begin{aligned} P[N &< 600] = P\left[N - E[N] < 600 - 700\right] \\ &= P\left[\frac{N - E[N]}{\sigma(N)} < \frac{600 - 700}{\sqrt{210}}\right] \approx \Phi\left(\frac{600 - 700}{\sqrt{210}}\right) \\ &= 0.0000000026\% \end{aligned}$$

- **Die Wahrscheinlichkeit einer großen Abweichung vom Erwartungswert nimmt also bei wachsender Gesamtgröße sehr rasch ab !**
- Dies ist die Grundlage statistischer Erhebungen - aber auch von Versicherungen und Spielcasinos: Wenn das Casino im Schnitt immer etwas gewinnt, und eine große Zahl von Spielern beteiligt ist, dann ist es sehr unwahrscheinlich, daß das Casino insgesamt Verlust macht.

Größenordnung der Fluktuationen um den Erwartungswert

Aus der Normalapproximation ergeben sich folgende Abschätzungen für die Abweichung einer binomialverteilten Zufallsvariable X von ihrem Erwartungswert $E[X]$:

1. **Typische Fluktuation:** *Eine Standardabweichung*

$$\begin{aligned} & P\left[|X - E[X]| \leq \sigma(X)\right] \\ = \quad & P\left[\left|\frac{X - E[X]}{\sigma(X)}\right| \leq 1\right] \approx \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy = 2 \cdot \int_{0}^{1} \ldots\, dy \\ = \quad & 2 \cdot (\Phi(1) - \Phi(0)) = 2 \cdot (0.841 - 0.5) = 68.2\% \end{aligned}$$

2. **Gelegentliche Fluktuation:** *Zwei Standardabweichungen*

$$P\left[|X - E[X]| \leq 2\sigma(X)\right] = P\left[\left|\frac{X - E[X]}{\sigma(X)}\right| \leq 2\right]$$

$$= \quad 2 \cdot (\Phi(2) - \Phi(0)) = 2 \cdot (0.977 - 0.5) = 95.4\%$$

3. **Seltene Fluktuation:** *Drei Standardabweichungen*

$$P\left[|X - E[X]| \leq 3\sigma(X)\right] = 2 \cdot (\Phi(3) - \Phi(0)) = 99.7\%$$

- Ist n groß genug, dann liegt der Wert einer binomialvertelten Zufallsvariable X also fast immer im Bereich $E[X] \pm 3\sigma(X)$, und in der Mehrzahl der Fälle sogar im Bereich $E[X] \pm \sigma(X)$!
- Die Normalapproximation und diese Aussage gilt aber nicht nur für binomialverteilte Zufallsvariablen, sondern zum Beispiel allgemein für jede Zufallsvariable, die sich als Summe von vielen kleinen unabhängigen Zufallsvariablen auffassen läßt !!! Dies ist die Aussage des **zentralen Grenzwertsatzes**, siehe nächster Abschnitt.