6. Stochastische Modelle II: Normalverteilungen und Grenzwertsätze

Andreas Eberle
Institut für angewandte Mathematik

November 2008

6.1. Stetige Wahrscheinlichkeitsverteilungen

Wir betrachten eine Zufallsvariable X mit Werten in $\mathbb{R}$ (reelle Zahlen).

Definition

- Eine **stetige Wahrscheinlichkeitsverteilung** auf $\mathbb{R}$ ist festgelegt durch eine **Dichtefunktion** $f(x) \geq 0$ mit

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

- Die Wahrscheinlichkeit für Werte im Intervall $[a, b]$ oder (a, b) ist durch die Fläche $\int_a^b f(x)\, dx$ unter dem Graphen der Dichtefunktion $f(x)$ gegeben.

- Wir sagen, daß X eine **stetige Zufallsvariable mit Dichtefunktion** f_X ist, falls für alle $a \leq b$ gilt:

$$P[a \leq X \leq b] = P[a < X < b] = \int_a^b f_X(x)\ dx.$$

- Die Wahrscheinlichkeit, daß der Wert einer stetigen Zufallsvariable X in einem kleinen Intervall $[x, x + \Delta x]$ liegt, beträgt

$$P[x \leq X \leq x + \Delta x] \approx f_X(x)\,\Delta x$$

- Dies erklärt das Wort *"Dichtefunktion"*:

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P[X \in [x, x + \Delta x]]}{\Delta x} = \lim_{\Delta x \to 0} \frac{\text{W'keit (Intervall)}}{\text{Intervalllänge}}$$

- Durch Aufsummieren über eine Zerlegung $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ eines Intervalls $[a, b]$ in kleine Teilintervalle erhält man die Formel von oben:

$$P[a \leq X \leq b] = \lim_{\Delta x_i \to 0} \sum_i f_X(x_i)\ \Delta x_i = \int_a^b f_X(x)\ dx.$$

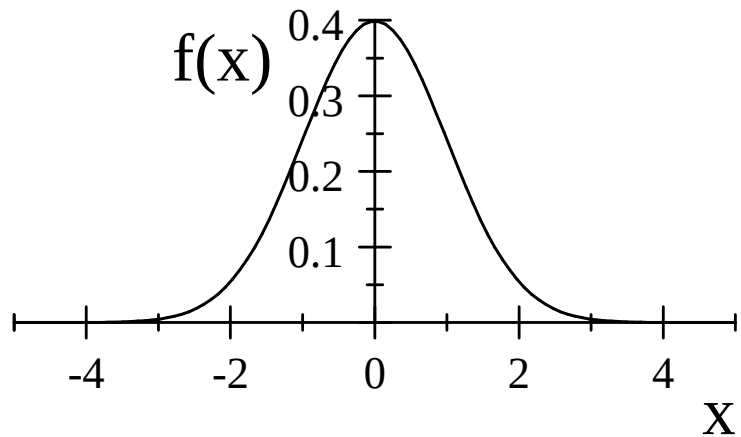
Stetige Wahrscheinlichkeitsverteilungen

Beispiel 1: Standardnormalverteilung

- Die Gaußsche Glockenfunktion

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

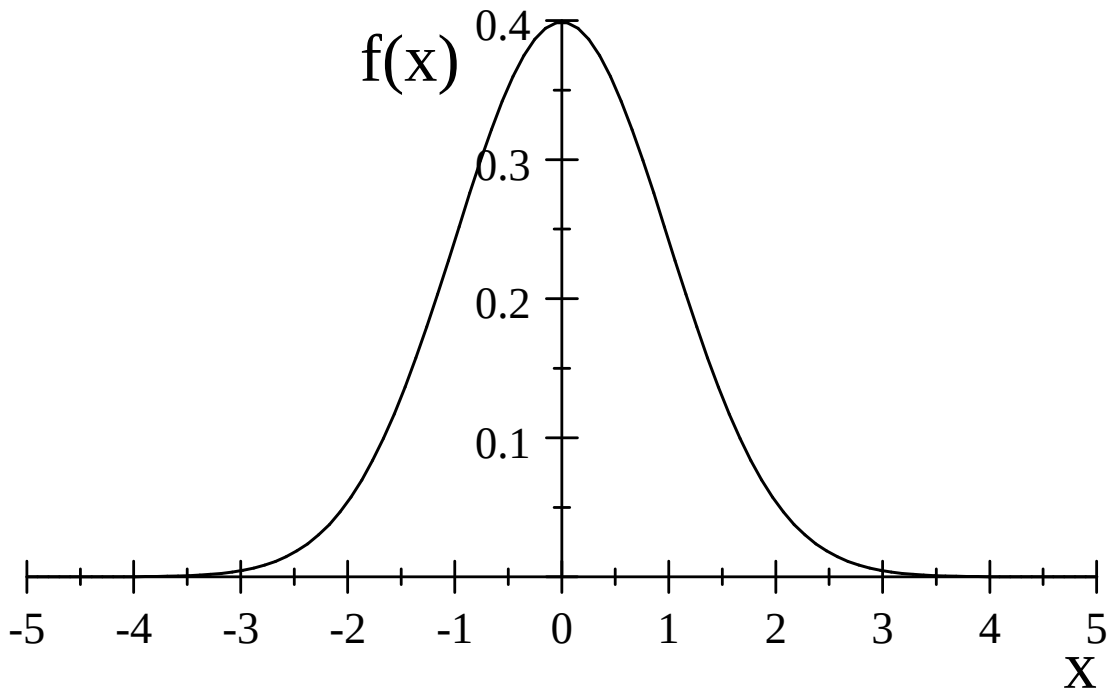
erfüllt $f(x) \geq 0$ und $\int_{-\infty}^{\infty} f(x)\, dx = 1$ (aus diesem Grund wählt man den Vorfaktor $1/\sqrt{2\pi}$).



Definition

Die Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit Dichtefunktion $f(x)$ heißt **Standardnormalverteilung**.

- Ist Z eine *standardnormalverteilte Zufallsvariable*, dann ist die W-keit, dass Z zwischen a und b liegt, gleich dem Integral der Dichte $f(x)$ mit den Grenzen a und b, also gleich der Fläche unter der Kurve über dem Intervall $[a, b]$.



- Ist X eine *standardisierte binomialverteilte Zufallsvariable* mit großem Parameter n, dann gilt dieselbe Aussage näherungsweise:

$$P[a \leq X \leq b] \approx P[a \leq Z \leq b] = \int_a^b f(x)\, dx$$

Verteilungsfunktion

Definition

Die **Verteilungsfunktion** einer reellwertigen Zufallsvariable X ist die durch

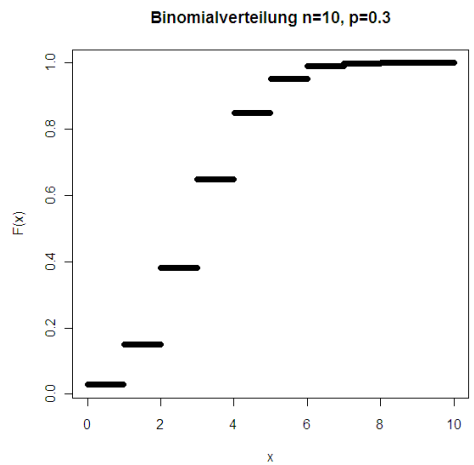
$$F_X(y) = P[X \leq y] \quad , \qquad -\infty < y < \infty,$$

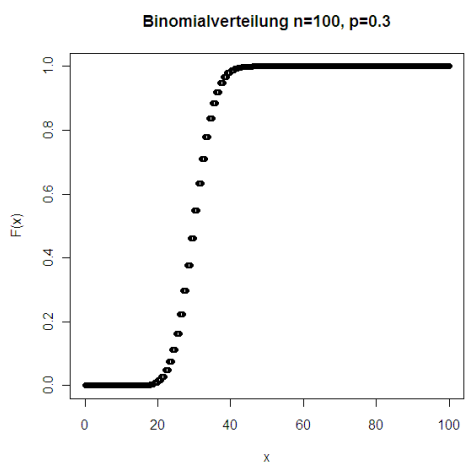
definierte, monoton wachsende Funktion.

- Ist X eine diskrete Zufallsvariable mit Werten $a_1, a_2, \ldots$, dann ergibt sich

$$F_X(y) = \sum_{a_k \leq y} P[X = a_k] = \sum_{a_k \leq y} p_X(a_k) \ .$$

- **Beispiel.**



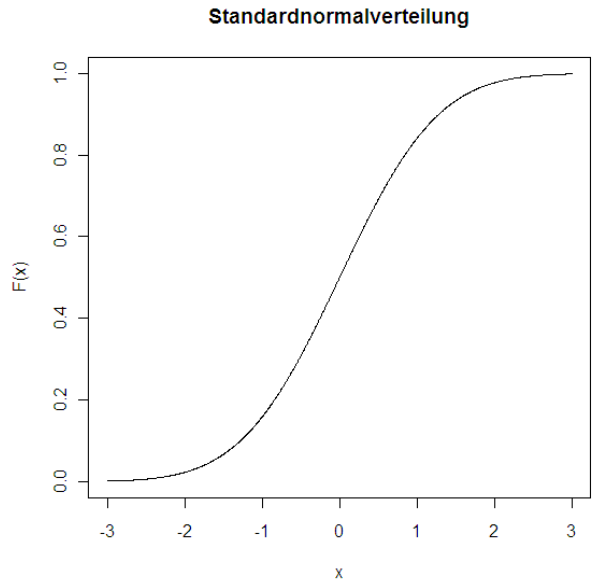


Verteilungsfunktion

- Ist X eine stetige Zufallsvariable, dann ergibt sich

$$F_X(y) = P[X \leq y] = \int_{-\infty}^{y} f_X(x)\, dx\ .$$

- **Die Verteilungsfunktion einer stetigen Zufallsvariable ist also die Stammfunktion der Dichte: $F_X'(x) = f_X(x)$.**
- **Beispiel.**



- Die Verteilungsfunktion $\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$ der Standardnormalverteilung ist in vielen Statistikbüchern tabelliert.

Stetige Wahrscheinlichkeitsverteilungen

Beispiel 2: Exponentialverteilung, Wartezeiten

- Die Exponentialverteilung dient zur Modellierung der *Wartezeit* T *auf das erste Eintreten eines unvorhersehbaren Eregnisses*.
- Wir nehmen an, daß die Häufigkeit N_t des Ereignisses bis zur Zeit t Poissonverteilt ist mit Erwartungswert $\lambda \cdot t$, wobei $\lambda > 0$ eine feste Konstante ist:

$$N_t \sim \text{Poisson}\,(\lambda \cdot t)$$

- Die Wahrscheinlichkeit, daß das Ereignis bis zur Zeit t überhaupt noch nicht eingetreten ist, also daß die Wartezeit auf das erste Eintreten größer als t ist, beträgt dann:

$$P\,[T > t] = P\,[N_t = 0] = \frac{(\lambda \cdot t)}{0!} \cdot e^{-\lambda \cdot t} = e^{-\lambda \cdot t}$$

- Also ergibt sich für die Verteilungsfunktion von T:

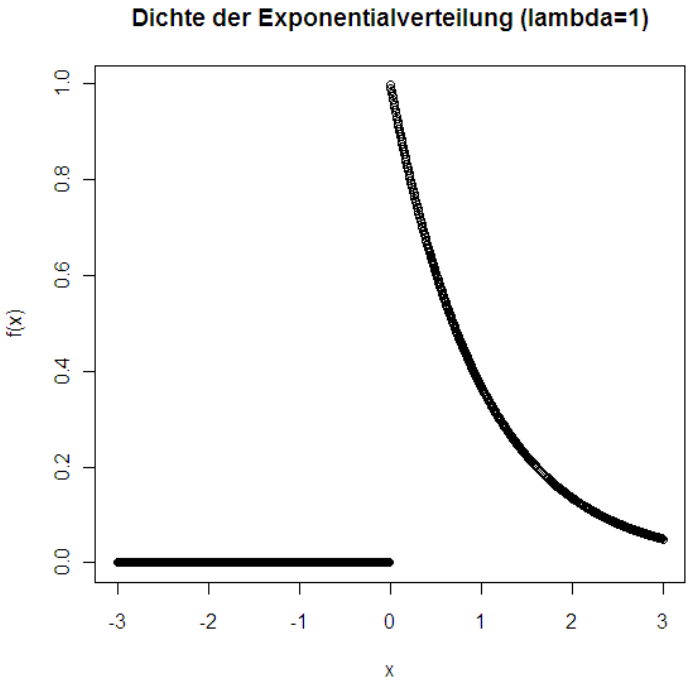
$$F_T(t) = P[T \leq t] = 1 - e^{-\lambda \cdot t} \qquad (t \geq 0)$$

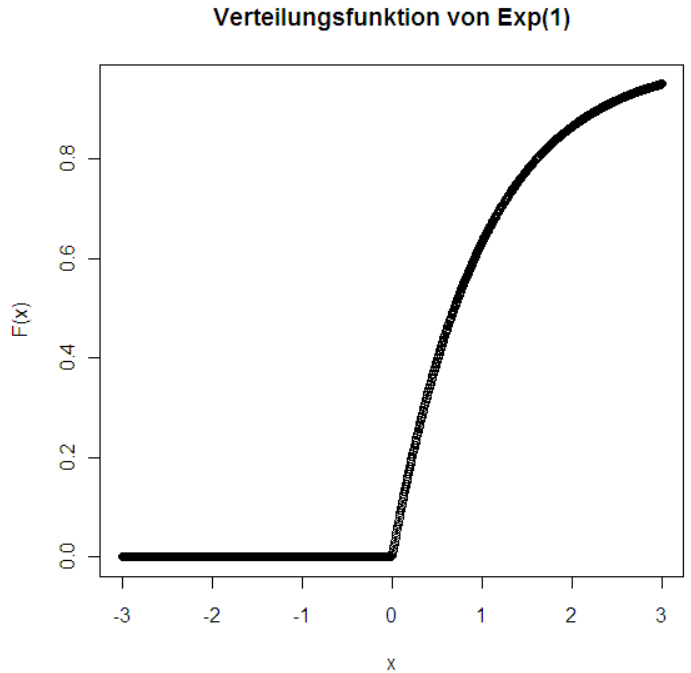
Stetige Wahrscheinlichkeitsverteilungen

Beispiel 2: Exponentialverteilung, Wartezeiten

- $$F_T(t) = P[T \leq t] = 1 - e^{-\lambda \cdot t} \qquad (t \geq 0)$$
- Für die Dichte der Verteilung von T erhalten wir damit:

$$f_T(t) = F_T'(t) = \lambda \cdot e^{-\lambda \cdot t} \text{ für } t > 0, \qquad f_T(t) = 0 \text{ für } t \leq 0.$$





Definition

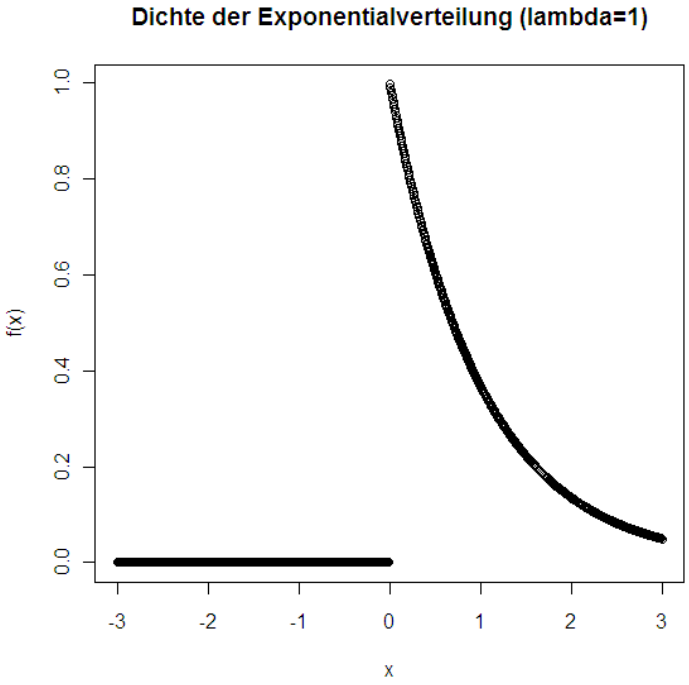
Die Wahrscheinlichkeitsverteilung mit Dichte

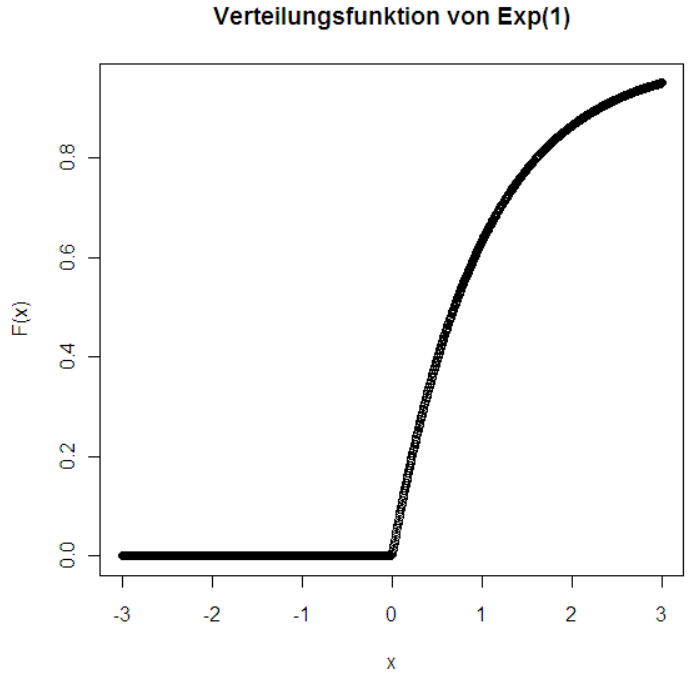
$$f_T(t) = F_T'(t) = \lambda \cdot e^{-\lambda \cdot t}$$

und Verteilungsfunktion

$$F_T(t) = P[T \leq t] = 1 - e^{-\lambda \cdot t}$$

heißt **Exponentialverteilung** zum Parameter $\lambda > 0$ (Exp (λ)).





Stetige Wahrscheinlichkeitsverteilungen

Beispiel 3: Allgemeine Normalverteilung

- Sei Z eine standardnormalverteilte Zufallsvariable. Wir wollen nun die Verteilung der linear transformierten Zufallsvariable

$$X = \sigma Z + m$$

berechnen, wobei $\sigma > 0$ und m Konstanten sind.

- Mithilfe der Substitution $z = (x - m)/\sigma$ erhalten wir

$$\begin{aligned} F_X(y) &= P[X \leq y] = P[\sigma Z + m \leq y] \\ &= P\left[Z \leq \frac{y-m}{\sigma}\right] = \int_{-\infty}^{\frac{y-m}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\ &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/2\sigma^2}\, dx \ . \end{aligned}$$

- Also:

$$f_X(x) = F_X'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/2\sigma^2}$$

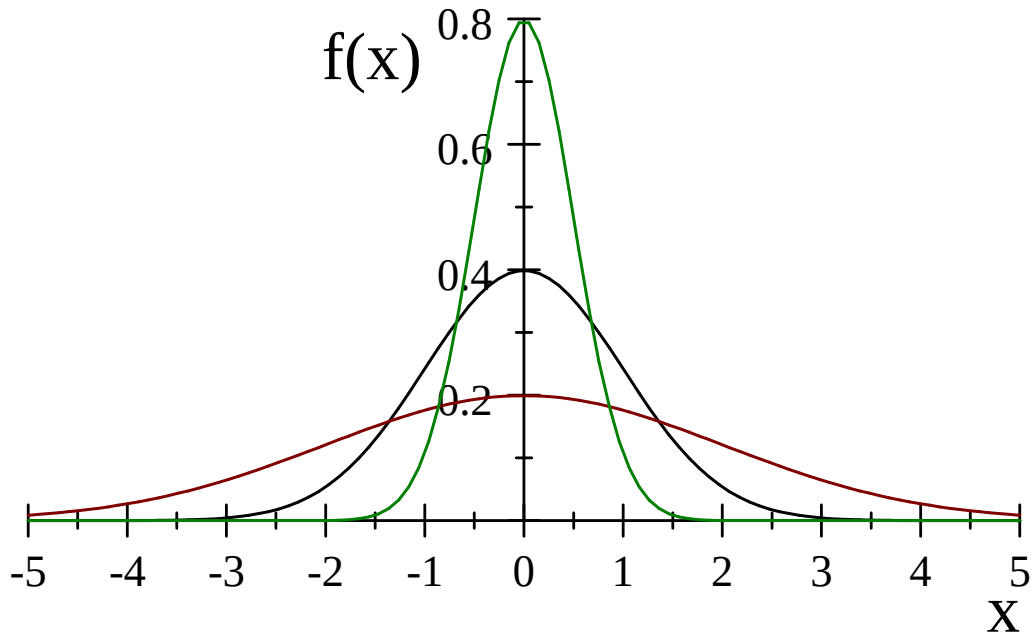
Definition

Die Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit Dichtefunktion

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

heißt **Normalverteilung mit Mittelwert** m **und Varianz** σ^2.

- Normalverteilungen $N(m, \sigma^2)$ mit Mittelwert $m = 0$ und Standardabweichungen $\sigma = 0.5$, 1, bzw. 2:



Normalverteilung $N(m, \sigma^2)$:

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

- Wir werden in kuerze sehen, daß die Parameter m und σ^2 tatsächlich den Erwartungswert und die Varianz einer normalverteilten Zufallsvariable angeben.
- Für $m = 0$ und $\sigma^2 = 1$ ergibt sich die **Standardnormalverteilung**.
- Ist σ^2 klein, dann konzentriert sich die Verteilung $N(m, \sigma^2)$ sehr stark in der Nähe des Mittelwerts.
- Ist σ^2 groß, dann ist die Verteilung $N(m, \sigma^2)$ eher diffus.

Stetige Verteilungen

Transformationseigenschaften der Normalverteilungen

- Für reelle Zahlen a und b gilt:

$$\begin{aligned} X &\sim N(m, \sigma^2) \quad \Longrightarrow \quad X + a \sim N(m + a, \sigma^2) \\ X &\sim N(m, \sigma^2) \quad \Longrightarrow \quad b \cdot X \sim N(b \cdot m, b^2 \cdot \sigma^2) \end{aligned}$$

In Worten:

- Wenn ich zu einer $N(m, \sigma^2)$-verteilten Zufallsvariable a addiere, ist das Ergebnis eine normalverteilte Zufallsvariable zu den Parametern $m + a$ und σ.
- Wenn ich eine normalverteilte Zufallsvariable zu den Parametern m und σ mit b multipliziere, ist das Ergebnis eine normalverteilte Zufallsvariable zu den Parametern $b \cdot m$ und $|b| \cdot \sigma$.

Stetige Wahrscheinlichkeitsverteilungen

Nachtrag zur Normalapproximation der Binomialverteilung

- Ist X eine binomialverteilte Zufallsvariable mit Mittelwert $m = np$ und Varianz $\sigma^2 = np(1-p)$, dann gilt für große n näherungsweise

$$\frac{X - m}{\sigma} \sim N(0, 1),$$

das heißt die standardisierte Zufallsvariable $Y = (X - m)/\sigma$ ist näherungsweise standardnormalverteilt, s.o.

- Wegen

$$X = \sigma Y + m$$

folgt dann aber, daß die Zufallsvariable X selbst für große n näherungsweise normalverteilt ist mit Mittel m und Varianz σ^2. Es gilt also näherungsweise

$$X \sim N(np, np(1-p))$$

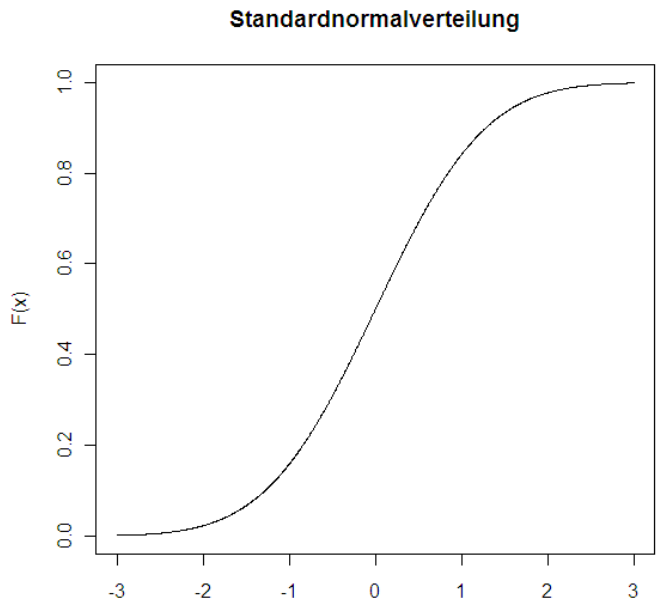
6.2. Quantile

Sei p eine Zahl zwischen 0 und 1.

Definition

Das p-**Quantil** $x_{(p)}$ einer Zufallsvariable X ist der erste Wert x, an dem die Verteilungsfunktion den Level p erreicht oder überschreitet. Bei einer stetigen Zufallsvariable ist $x_{(p)}$ der (kleinste) Wert mit

$$F_X(x_{(p)}) = P\left[X \leq x_{(p)}\right] = p.$$



- Durch Standardisierung kann man zeigen, dass zwischen den Quantilen $x_{(p)}$ einer $N(m, \sigma^2)$ Verteilung und den Quantilen $z_{(p)}$ der Standardnormalverteilung $N(0, 1)$ folgender Zusammenhang besteht:

$$\boxed{x_{(p)} = m + \sigma \cdot z_{(p)}}$$

- Deswegen genügt es, die Quantile $z_{(p)}$ von N(0,1) zu kennen.
- Diese kann man mit der Definition *"$z_{(p)}$ ist die Stelle, an der die Verteilungsfunktion $\Phi(x)$ der Standardnormalverteilung den Wert p annimmt"* aus Tabellen ablesen.
- Die Tabelle der Standardnormalverteilung enthält aber nur Werte $\Phi(x)$ für $x > 0$, deshalb kann man $z_{(p)}$ nur für $p \geq 0.5$ direkt ablesen.
- Für $p < 0.5$ verwende man die Formel

$$z_{(p)} = -z_{(1-p)} \quad ,$$

die aus der Achsensymmetrie der Glockenkurve folgt.

- Bei vielen Fragestellungen zu Normal- oder zu anderen Verteilungen müssen wir Bereiche bestimmen, in denen ein gewisser Anteil der Wahrscheinlichkeitsmasse liegt, d.h. die Werte aus diesem Bereich werden mit einer gewissen W'keit p angenommen.
- Man unterscheidet zwischen einseitigen und zweiseitigen Fragestellungen:
  - Gesucht b so dass $P[X \leq b] = p$ *(einseitig)*
  - Gesucht a so dass $P[X \geq a] = p$ *(einseitig)*
  - Gesucht a, b so dass $P[a \leq X \leq b] = p$ *(zweiseitig)*
  - Gesucht $\varepsilon > 0$ und m so dass $P[|X - m| \leq \varepsilon] = p$ *(zweiseitig symmetrisch)*
- In allen Fällen benötigen wir die Quantile der zugrundeliegenden Verteilung, um die Bereiche zu ermitteln.

Beispiel. (Qualitätskontrolle 1)

- *Wir suchen die minimale Zahl n der Hühner, so dass wir an 99% der Tage 10000 Eier garantieren können, wenn jedes Huhn an $p = 70\%$ der Tage ein Ei legt.*
- *Die mittlere Anzahl gelegter Eier und die Standardabweichung sind*

$$m = np = 0.7 \cdot n \quad \text{und} \quad \sigma = \sqrt{0.7 \cdot 0.3 \cdot n}.$$

- *Bei $n > 10000$ können wir ohne weiteres annehmen, die Zahl X der gelegten Eier sei normalverteilt zu diesen Parametern.*
- *Es soll gelten:*

$$P\left[X < 10000\right] = P\left[\frac{X - m}{\sigma} < \frac{10000 - 0.7 \cdot n}{\sqrt{0.21 \cdot n}}\right] \stackrel{!}{=} 0.01$$

Beispiel. (Qualitätskontrolle 1)

- *Es soll gelten:*

$$P\left[X < 10000\right] = P\left[\frac{X-m}{\sigma} < \frac{10000 - 0.7 \cdot n}{\sqrt{0.21 \cdot n}}\right] \stackrel{!}{=} 0.01$$

- *Da* $\frac{X-m}{\sigma}$ *standardnormalverteilt ist, ist dies gerade dann der Fall, wenn*

$$\frac{10000 - 0.7 \cdot n}{\sqrt{0.21 \cdot n}} = z_{(0.01)} \qquad (*)$$

  das 1% Quantil der Standardnormalverteilung ist.

- *Mithilfe einer Tabelle der Verteilungsfunktion erhalten wir*

$$z_{(0.01)} = -z_{(0.99)} = -2.33$$

- *Wenn wir dies in (*) einsetzen, und nach n auflösen, ergibt sich schließlich*

$$\sqrt{n} = \frac{2.33 \cdot \sqrt{0.21}}{2 \cdot 0.7} + \sqrt{\left(\frac{2.33 \cdot \sqrt{0.21}}{2 \cdot 0.7}\right)^2 + \frac{9999}{0.7}} \Rightarrow n = 14468$$

Beispiel. (Qualitätskontrolle 1)

- *Antwort: Mit 14468 Hühnern, von denen jedes zu 70% täglich ein Ei legt, werden wir an 99% der Tage mindestens 10000 Eier erhalten.*
- *Die mittlere Anzahl gelegter Eier beträgt in diesem Fall*

$$0.7 \cdot 14468 = 10128$$

- *Obwohl die mittlere Anzahl gar nicht so viel größer als 10000 ist, liegt die tatsächlich gelegte Eierzahl fast immer oberhalb von 10000.*
- *Der Grund sind wieder die* **sehr große Anzahl von Hühnern**, *und das* **Gesetz der großen Zahl**.

Beispiel. (Qualitätskontrolle 2)

- *Eine Maschine produziert Nägel der Länge* $m = 1$ *cm. Die Maschine hat aber, wie jede andere, eine kleine Ungenauigkeit, die sich in einer Standardabweichung der Nagellängen von* $\sigma = 0.5$ *mm niederschlägt. Sie wollen einem Abnehmer für die Nägel ein Angebot machen, und wollen dabei angeben, welche Fehlertoleranz bei der Nagellänge zu 99% unterschritten wird. Welche Toleranz geben Sie an?*
- *Zunächst einmal ist es bei solchen Fragestellungen der Qualitätskontrolle üblich, dass man eine Normalverteilung für die Zufallsvariable* X *=Nagellänge voraussetzt. Wir haben also:*

$$X \sim N(m, \sigma^2) \quad \textit{mit} \quad m = 1 \quad \textit{und} \quad \sigma^2 = (0.05)^2 = 0.0025.$$

(Hier ist darauf zu achten, dass μ *und* σ *in derselben Einheit, nämlich cm, gemessen werden)*

- Gesucht ist also $\varepsilon > 0$ so dass

$$P(|X - m| \leq \varepsilon) = P(|X - 1| \leq \varepsilon) = 0.99$$

- Jetzt standardisieren wir und setzen

$$Z = \frac{X-m}{\sigma} = \frac{X-1}{0.05} \Rightarrow Z \sim N(0,1).$$

- Dann haben wir

$$P(|X - 1| \leq \varepsilon) = P(|Z| \leq \tfrac{\varepsilon}{\sigma}).$$

- Als nächstes suchen wir eine Zahl d so dass

$$P(|Z| \leq d) = 0.99.$$

- Aufgund der Achsensymmetrie der Standardnormalverteilung ist $d = z_{0.995}$, dem 99.5%-Quantil der Standardnormalverteilung (Denn wir dürfen rechts und links jeweils nur ein halbes % verlieren).

- Jetzt suchen wir $z_{0.995}$ aus der Tabelle der Verteilungsfunktion der Standardnormalverteilung.
- Wir erhalten:

$$\boxed{z_{0.995} = 2.575}$$

- Also:

$$\boxed{\varepsilon = d \cdot \sigma = z_{0.995} \cdot \sigma = 2.575 \cdot 0.05 = 0.129.} \text{ (Diese Angabe ist in cm)}$$

- Sie können Ihrem Abnehmer garantieren, dass 99% der Nägel nicht mehr als 0.129 cm=1.29 mm von der Solllänge 1 cm abweichen.

6.3. Erwartungswert und Varianz

Erwartungswert von stetigen Zufallsvariablen

- Für die Berechnung des Erwartungswertes einer Zufallsvariable X mit stetiger Verteilung ist die Massenfunktion durch die Dichte, und die Summe durch ein Integral zu ersetzen:

$$\text{Diskret} \quad : \qquad E[X] = \sum_a a \cdot p_X(a)$$

$$\text{Stetig} \quad : \qquad E[X] = \int_{-\infty}^{\infty} a \cdot f_X(a)\ da = \int_{-\infty}^{\infty} x \cdot f_X(x)\ dx$$

- ... bzw. allgemeiner:

$$\text{Diskret:} \qquad E[g(X)] = \sum_a g(a) \cdot p_X(a)$$

$$\text{Stetig:} \qquad E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\ dx$$

Erwartungswert und Varianz

Varianz

- Die Varianz ist wiederum gegeben durch

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

- Dies kann man mit Hilfe der Definition des Erwartungswertes ausschreiben (setze $\overline{x} = E[X]$) :

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{\infty} (x - \overline{x})^2 \cdot f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\, dx - \left(\int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \right)^2 \end{aligned}$$

- Die Standardabweichung ist wieder gegeben als die Wurzel der Varianz:

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

Erwartungswert und Varianz

Beispiel 1: Standardnormalverteilung

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

- Aufgrund der Achsensymmetrie der Glockenkurve ergibt sich sofort:

$$\overline{x} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot e^{-x^2/2}\, dx = 0$$

- Außerdem kann man ausrechnen, daß

$$\text{Varianz } = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-0)^2 \cdot e^{-x^2/2}\, dx = 1$$

gilt (diese Berechnung ist etwas schwieriger).

Erwartungswert und Varianz

Beispiel 2: Exponentialverteilung

- Ist T exponentialverteilt mit Parameter λ, dann gilt

$$f_T(t) = \lambda \cdot e^{-\lambda \cdot t} \text{ für } t > 0, \quad f_T(t) = 0 \text{ für } t \leq 0.$$

- Es folgt:

$$E[T] = \int_0^\infty t \cdot \lambda \cdot e^{-\lambda \cdot t} dt \overset{\text{part. Int.}}{=} \int_0^\infty e^{-\lambda \cdot t} dt = \frac{1}{\lambda}$$

- Ähnlich erhält man:

$$\sigma(T) = \frac{1}{\lambda}$$

Erwartungswert und Varianz

Beispiel 3: Normalverteilung

- Ist X normalverteilt mit Parametern m und σ^2, dann gilt

$$X = \sigma \cdot Z + m \qquad \text{mit} \qquad Z \sim N(0,1).$$

- Daraus ergibt sich, daß X tatsächlich Erwartungswert m und Standardabweichung σ hat:

$$\begin{aligned} E[X] &= \sigma \cdot \underbrace{E[Z]}_{=0} + m = m \\ \sigma(X) &= \sigma \cdot \underbrace{\sigma(Z)}_{=1} = \sigma \end{aligned}$$

Erwartungswert und Varianz

Fluktuationen um den Erwartungswert

Oft drückt man Bereiche um den Erwartungswert in Vielfachen der Standardabweichung aus. Ist X normalverteilt mit Mittelwert m und Standardabweichung σ, dann ist $Z = \frac{X-m}{\sigma}$ standardnormalverteilt, und es ergibt sich:

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

$$P\left[m-\sigma \leq X \leq m+\sigma\right] = P\left[-1 \leq Z \leq 1\right] = 68.2\%$$

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

$$P\left[m-2\sigma \leq X \leq m+2\sigma\right] = P\left[-2 \leq Z \leq 2\right] = 95.4\%$$

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

$$P\left[m-3\sigma \leq X \leq m+3\sigma\right] = P\left[-3 \leq Z \leq 3\right] = 99.7\%$$

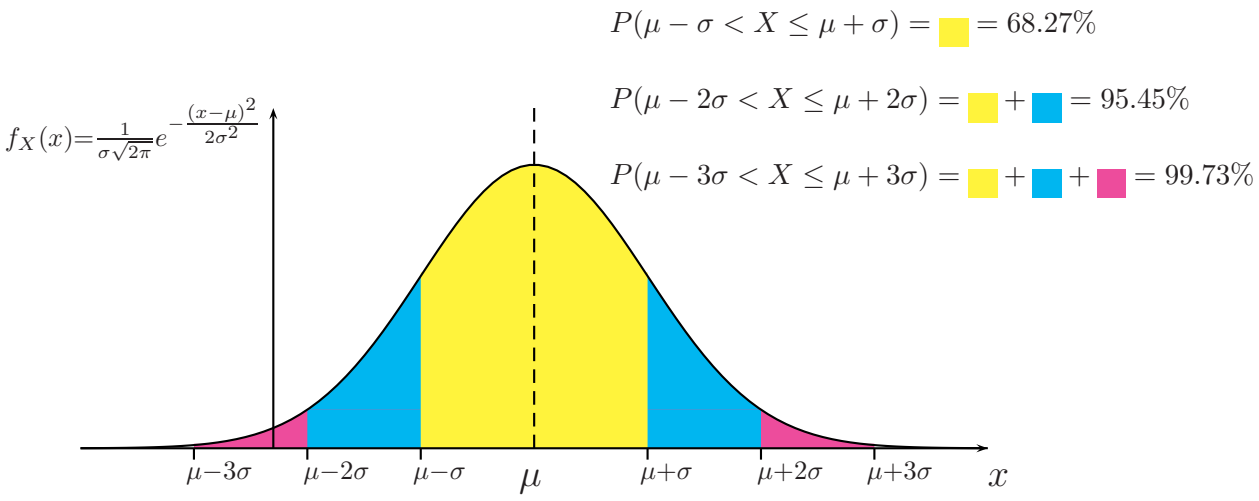
Erwartungswert und Varianz der Gaussverteilung

- Für eine normalverteilte Z.V. $X \sim N(\lambda, \sigma^2)$ erhalten wir

$$\mathbb{E}[X] = \lambda \text{ sowie } V_X = \sigma^2 \text{ also } \sigma_X = \sigma.$$

- Oft drückt man Bereiche um den Erwartungswert in vielfachen von der Standardabweichung aus.

Einfache, Zweifache und dreifache σ-Umgebung um den Erwartungswert



- Der Wert einer *normalverteilten* Zufallsvariable X liegt also fast immer im Bereich $m \pm 3\sigma$, und in der Mehrzahl der Fälle sogar im Bereich $m \pm \sigma$!
- Aufgrund des zentralen Grenzwertsatzes (s.u.) gilt diese Aussage *in vielen Fällen auch näherungsweise für nicht normalverteilte Zufallsvariablen*, da sich deren Verteilung in guter Näherung durch eine Normalverteilung approximieren läßt (siehe z.B. Normalapproximation der Binomialverteilung).
- Für Zufallsvariablen, deren Verteilung sich nicht gut durch eine Normalverteilung approximieren läßt, gilt zumindest noch die folgende allgemeine (aber dafür sehr grobe) **Abschätzung** von Chebyschew:

$$P\left[|X - E[X]| \geq k \cdot \sigma(X)\right] \leq \frac{1}{k^2}$$

Beispiel. (Qualitätskontrolle 1, Fortsetzung von oben)

- *Wir suchen die minimale Zahl n der Hühner, so dass wir 10000 Eier garantieren können, wenn jedes Huhn an 70% der Tage ein Ei legt.*
- *Wie gesehen liegen die Werte einer normalverteilten Zufallsvariable X zu 99.73% in dem Intervall $(m - 3\sigma, m + 3\sigma)$. Wegen der Achsensymmetrie der Verteilung um m haben wir*

$$P[X \leq m - 3\sigma] \;=\; \tfrac{1}{2}(100 - 99.73)\ \% \;=\; 0.135\%$$

- *Wollen wir den Kontrakt sogar an 99.865% aller Tage erfüllen, brauchen wir n Hühner, wobei*

$$10000 = m - 3\sigma = 0.7n - 3\sqrt{n(0.7 \cdot 0.3)}$$

- *Auflösen der quadratischen Gleichung nach $\sqrt{n}$ liefert $n = 14521$. Also kann man sich mit 14521 Hühnern sogar sicher sein, dass der Kontrakt an 99.865% der Tage eingehalten wird.*

6.4. Grenzwertsätze

Mittelwerte von Zufallsvariablen

- Wir betrachten nun die arithmetischen Mittelwerte

$$\overline{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n)$$

  von unabhängigen Zufallsvariablen $X_1, X_{2,} \ldots$.

- **Beipiel 1: Stichprobenmittelwerte**
  - $X_1, X_{2,} \ldots$ sind die beobachteten Ausprägungen $X(\omega_i)$ eines quantitativen Merkmals X bei Entnahme unabhängiger Einzelstichproben $\omega_1, \omega_2, \ldots$ aus der Grundgesamtheit (*Ziehen mit Zurücklegen*).
  - $\overline{X}_n$ ist dann das n-te *Stichprobenmittel*.
  - Wir erwarten, daß für große n der Stichprobenmittelwert ungefähr gleich dem Mittelwert des Merkmals in der Grundgesamtheit ist, da sich Fluktuationen in verschiedene Richtungen "wegmitteln" sollten (-> *Gesetz der großen Zahlen*).
  - Uns interessiert vor allem die Größe und Art der zufälligen Fluktuationen von $\overline{X}_n$ um den Prognosewert (-> *zentraler Grenzwertsatz*).

$$\overline{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n)$$

Beipiel 2: Relative Häufigkeiten

- Wir beobachten die Ausprägungen $Y(\omega_i)$ eines qualitativen oder diskreten Merkmals Y. Uns interessiert die relative Häufigkeit $h_n(a)$ einer bestimmten Merkmalsausprägung a unter den ersten n Beobachtungswerten.
- Setzen wir

$$X_i = \begin{cases} 1 & \text{falls } Y(\omega_i) = a \\ 0 & \text{falls } Y(\omega_i) \neq a \end{cases},$$

dann ergibt sich

$$h_n(a) = \frac{1}{n} \cdot \underbrace{(X_1 + X_2 + \cdots + X_n)}_{\text{Häufigkeit von } a \text{ unter } Y(\omega_1),\ldots,Y(\omega_n)} = \overline{X}_n.$$

- $\overline{X}_n$ ist also gerade die gesuchte relative Häufigkeit.
- Ähnlich wie oben erwarten wir, daß $\overline{X}_n$ für große n ungefähr gleich der W'keit (=relative H'keit in der Grundgesamtheit) der Merkmalsausprägung a ist.

Grenzwertsätze

Das Gesetz der großen Zahlen

- Diese Aussagen kann man im wahrscheinlichkeitstheoretischen Modell beweisen:

Theorem

Sind $X_1, X_2, \ldots$ unabhängige (oder allgemeiner unkorrelierte) Zufallsvariablen mit Erwartungswert m und beschränkten Varianzen, dann konvergieren die Mittelwerte

$$\overline{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n) \longrightarrow m$$

für $n \longrightarrow \infty$ mit Wahrscheinlichkeit 1.

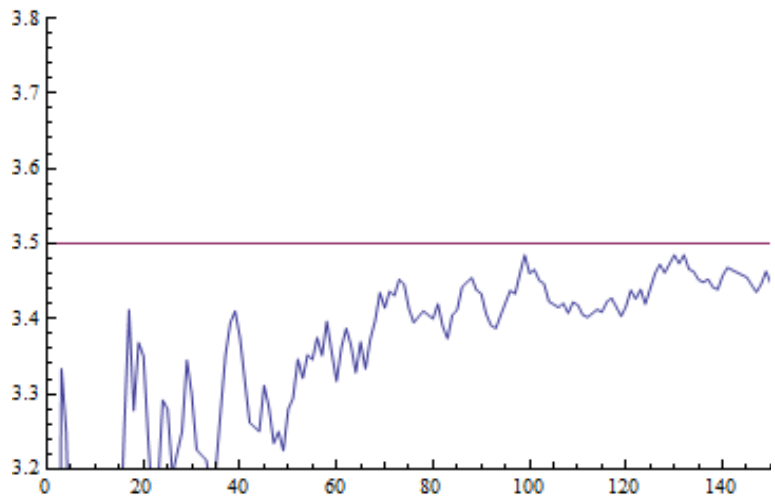
- Für große n gilt also näherungsweise

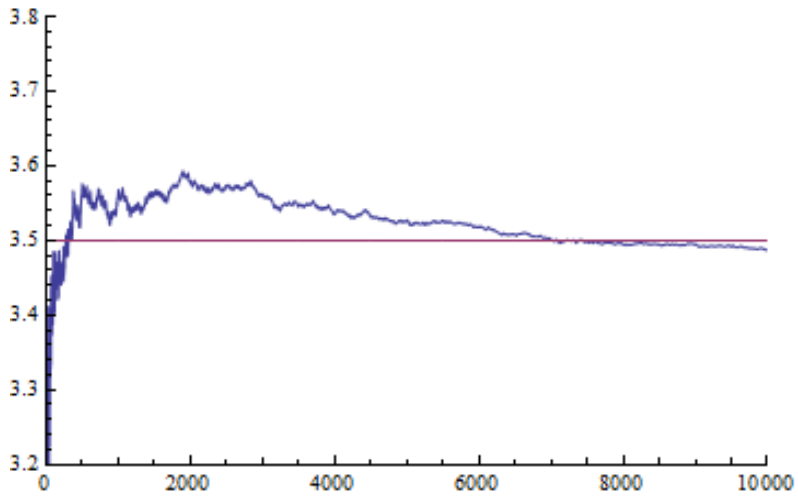
$$\overline{X}_n \approx E\left[\overline{X}_n\right] = m \qquad \text{"}\textbf{Der Zufall mittelt sich weg}\text{"}$$

Grenzwertsätze

Gesetz der großen Zahlen

- Der Stichprobenmittelwert $\overline{X}_n$ von unabhängigen Stichproben einer Zufallsgröße X stimmt für große n in etwa mit dem Erwartungswert der Zufallsgröße (Mittelwert in der Grundgesamtheit) überein.
- Beispiel: Mittlere Augenzahl bei n mal Würfeln:





- Die Konvergenz wird auch in der Mathematica-Demonstration "The Central Limit Theorem" veranschaulicht, siehe *http://demonstrations.wolfram.com/TheCentralLimitTheorem*

Grenzwertsätze

Gesetz der großen Zahlen

Theorem

Sind $X_1, X_2, \ldots$ unabhängige (oder allgemeiner unkorrelierte) Zufallsvariablen mit Erwartungswert m und beschränkten Varianzen, dann konvergieren die Mittelwerte

$$\overline{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n) \longrightarrow m$$

- **Beweisidee**: Für eine vorgegebene Abweichung $\varepsilon > 0$ von Stichprobenmittel und Erwartungswert gilt die Abschätzung

$$\begin{aligned} P\left[\left|\overline{X}_n - m\right| > \varepsilon\right] &= P\left[\left|\overline{X}_n - E\left[\overline{X}_n\right]\right| > \varepsilon\right] \leq \frac{1}{\varepsilon^2}\mathrm{Var}(\overline{X}_n) \\ &= \frac{1}{n^2\varepsilon^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right) \\ &\leq \text{const.} \cdot \frac{1}{n} \end{aligned}$$

Dabei wurde im 2. Schritt die Chebyschev-Ungleichung verwendet.

Grenzwertsätze

Anwendung des Gesetzes der großen Zahlen auf relative Häufigkeiten

- Angewandt auf die relativen Häufigkeiten $h_n(a)$ einer Merkmalsausprägung a in n einzelnen Zufallsstichproben aus einer Grundgesamtheit besagt das Gesetz der großen Zahlen:

- **Für große n gilt näherungsweise:**

$$h_n(a) \approx p$$

- Dabei ist p die Wahrscheinlichkeit der Merkmalsausprägung a, also die relative Häufigkeit von a in der Grundgesamtheit.
- Beipiel: Wenn wir oft hintereinander mit zwei Würfeln würfeln, sollte die relative Häufigkeit eines Paschs ungefähr gleich $\frac{6}{36} = \frac{1}{6}$ sein.

Grenzwertsätze

Fluktuationen von Mittelwerten

- Die Größe der zufälligen Fluktuationen der Mittelwerte $\overline{X}_n$ um den Erwartungswert m kann man grob mithilfe der Ungleichung von Chebyschev abschätzen (vgl. die Beweisidee zum Gesetz der großen Zahlen).
- Die so erhaltene Abschätzung ist aber sehr grob. Wenn wir uns noch einmal die Mathematica-Demonstration "The Central Limit Theorem" ansehen, dann erkennen wir, daß die Fluktuationen für große n näherungsweise normalverteilt sind. Die Standardabweichung der approximierenden Normalverteilung fällt, wenn n zunimmt.

Grenzwertsätze

Der zentrale Grenzwertsatz

- Auch diese Aussagen kann man im wahrscheinlichkeitstheoretischen Modell allgemein beweisen:

Theorem

Sind $X_1, X_2, \ldots$ unabhängige, identisch verteilte Zufallsvariablen mit Erwartungswert m und Varianz σ^2, dann gilt für große n näherungsweise:

$$\overline{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n) \sim N\left(m, \frac{\sigma^2}{n}\right)$$

- Die Verteilung der Mittelwerte $\overline{X}_n$ nähert sich also einer Normalverteilung an, die sich immer stärker in der Nähe des Erwartungswerts m konzentriert.
- Wie stark, hängt von der Varianz σ^2 der gemittelten Zufallsvariablen ab.

Grenzwertsätze

Anwendung des zentralen Grenzwertsatzes auf Stichprobenmittelwerte

- Für große n gilt näherungsweise:

$$\overline{X}_n = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n) \sim N\left(m, \frac{\sigma^2}{n}\right)$$

- Also ist die standardisierte Zufallsvariable

$$Z_n := \sqrt{n} \cdot \frac{\overline{X}_n - m}{\sigma}$$

  näherungsweise **standardnormalverteilt** !

- Dies können wir benutzen, um abzuschätzen, wie stark das Stichprobenmittel vom zu schätzenden Erwartungswert m (=Mittelwert in der Grundgesamtheit) abweicht:

$$\begin{aligned} P\left[\left|\overline{X}_n - m\right| \leq \varepsilon\right] &= P\left[|Z_n| \leq \frac{\varepsilon \cdot \sqrt{n}}{\sigma}\right] \\ &\approx 2 \cdot \left(\Phi\left(\frac{\varepsilon \cdot \sqrt{n}}{\sigma}\right) - \frac{1}{2}\right) \end{aligned}$$

Beispiel. (Schätzung des Mittelwerts bei bekannter Varianz)

- *In einem Experiment machen wir $n = 20$ Messungen einer unbekannten Größe, und erhalten den Stichprobenmittelwert $\overline{x}_n$. Unser Meßverfahren hat eine Ungenauigkeit, die sich in einer Standardabweichung $\sigma = 0.5$ der Meßwerte niederschlägt. Wir nehmen an, daß diese Ungenauigkeit die dominierende Ursache für zufällige Fluktuationen der Meßwerte ist.*

- *Gesucht ist nun zu jedem möglichen Stichprobenmittel ein "**Konfidenzintervall**" $\overline{x}_n \pm \varepsilon(\overline{x}_n)$, so daß die unbekannte Größe m mit 95 % Wahrscheinlichkeit in dem Intervall liegt.*
- *Mithilfe der ZGS-Approximation ergibt sich:*

$$P\left[\left|\overline{X}_n - m\right| \leq \varepsilon\right] \approx 2 \cdot \left(\Phi\left(\frac{\varepsilon \cdot \sqrt{n}}{\sigma}\right) - \frac{1}{2}\right) = 95\% \quad \textit{falls}$$

$$\Phi\left(\frac{\varepsilon \cdot \sqrt{n}}{\sigma}\right) = 0.975, \textit{ also } \varepsilon = \frac{\sigma}{\sqrt{n}} \cdot z_{(0.975)} = \frac{0.5}{\sqrt{20}} \cdot 1.96 = 0.22$$

Grenzwertsätze

Konfidenzintervall für den Mittelwert

- Die gesuchte Größe m liegt also mit W'keit 95% im Intervall $\overline{x}_n \pm 0.22$.
- Wir sagen auch, $[\overline{x}_n - 0.22, \overline{x}_n + 0.22]$ ist ein (approximatives) *95% Konfidenzintervall für den unbekannten Parameter m*.
- Die Breite des Konfidenzintervalls hängt in diesem einfachen Fall nicht vom Schätzwert $\overline{x}_n$ ab, im allgemeinen aber schon.

Grenzwertsätze

Anwendung des zentralen Grenzwertsatzes auf relative Häufigkeiten

- Die relative Häufigkeit $h_n(a)$ einer Ausprägung a eines Merkmals Y bei n unabhängigen Stichproben ist

$$h_n(a) = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n) = \overline{X}_n$$

wobei die Zufallsvariablen $X_i = \begin{cases} 1 & \text{falls } Y(\omega_i)=a \\ 0 & \text{falls } Y(\omega_i)\neq a \end{cases}$ unabhängig und Bernoulli(p) verteilt sind mit p=W'keit von a.

- Aus dem zentralen Grenzwertsatz folgt daher für große n:

$$h_n(a) \sim N\left(p, \frac{p(1-p)}{n}\right) \quad \text{bzw. Abs. H'keit } (a) \sim N\left(np, np(1-p)\right)$$

- Dies ist nichts anderes als die **Normalapproximation der binomialverteilten absoluten Häufigkeit**.

Grenzwertsätze

Verallgemeinerungen des zentralen Grenzwertsatzes

- Der zentrale Grenzwertsatz in der Formulierung von oben läßt sich noch deutlich verallgemeinern.
- Eine mögliche Erweiterungsrichtung ist die Ausdehnung auf verschiedene Klassen *abhängiger* Zufallsvariablen. Eine andere wichtige Erweiterung ist der *Satz von Lindeberg-Feller*, dessen Aussage wir hier nur ganz grob anschaulich wiedergeben wollen:

- **Zentraler Grenzwertsatz von Lindeberg-Feller:**
  "Ist X eine reelle Zufallsgröße, die durch additive Überlagerung vieler kleiner unabhängiger Zufallsgrößen X_i entsteht (d.h. $X = \sum X_i$), dann ist unter geeigneten Voraussetzungen (die wir nicht ausführen wollen) die standardisierte Zufallsvariable

$$\frac{X - E[X]}{\sigma(X)}$$

  näherungsweise standardnormalverteilt."

- **Zentraler Grenzwertsatz von Lindeberg-Feller:**
  "Ist X eine reelle Zufallsgröße, die durch additive Überlagerung vieler kleiner unabhängiger Zufallsgrößen X_i entsteht (d.h. $X = \sum X_i$), dann ist unter geeigneten Voraussetzungen (die wir nicht ausführen wollen) die standardisierte Zufallsvariable

$$\frac{X - E[X]}{\sigma(X)}$$

  näherungsweise standardnormalverteilt."

- Der Satz von Lindeberg-Feller liefert das theoretische Fundament für die häufige mathematische Modellierung von unbekannten Zufallsgrößen durch normalverteilte Zufallsvariablen (Gaußmodelle) !