Lebensdauerverteilungen I

Zur Modellierung von Lebensdauern (im weiteren Sinne) ist die Normalverteilung selten geeignet, da

- Nur Werte > 0 möglich
- Symmetrie selten sinnvoll

Typische Verteilungen sind die Exponentialverteilung, die Gammaverteilung, und die Weibullverteilung

Lebensdauern und Ereignisanalyse

Moderner Zweig vieler empirischer Untersuchungen: Lebensdaueranalyse bzw. allgemeiner Ereignisanalyse. Im Folgenden nur eine kurze Einführung, weiterführende Texte sind

- Rohwer und Pötter (2001): *Grundzüge der sozialwissenschaftlichen Statistik*, Teil III. Juventa, Soziologische Grundlagentexte.
- Blossfeld, Hamerle, Mayer (1986): *Ereignisanalyse: Statistische Theorie und Anwendungen in den Wirtschafts- und Sozialwissenschaften*. Campus.
- Diekmann und Mitter (1984): *Methoden zur Analyse von Zeitverläufen*. Teubner.
- Blossfeld und Rohwer (1995): *Techniques of Event History Modelling*. Erlbaur.

Lebensdauerverteilungen II

Betrachtet wird die Zufallsgröße „Zeit bis zu einem Ereignis“, z.B. Tod, Rückkehr aus Arbeitslosigkeit, Konkurs. Um den zeitlichen Aspekt (time) zu betonen, wird die interessierende Zufallsvariable häufig mit T statt mit X bezeichnet.

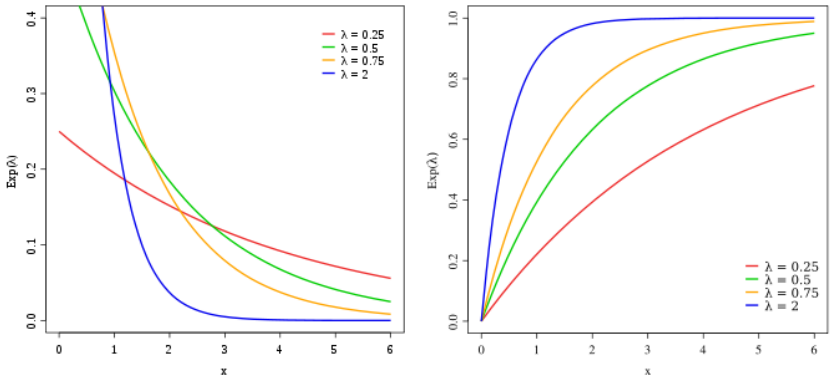
Die Exponentialverteilung

In Zeichen $\mathbf{X} \sim \mathbf{Ex}(\lambda)$ Der Parameter λ charakterisiert die Verteilung. Der Erwartungswert (Lebenserwartung) ist $\frac{1}{\lambda}$.

1. **Modell:** X ist die Lebensdauer eines Objekts, das nicht altert.
2. **Dichte, Verteilungsfunktion und Momente**

$$\begin{aligned} f_X(x) &= \lambda e^{-\lambda x} \\ F_X(x) &= 1 - e^{-\lambda x} \\ E(X) &= \frac{1}{\lambda} \\ V(X) &= \frac{1}{\lambda^2} \end{aligned}$$

Dichte und Verteilungsfunktion



Die Survivorfunktion und Hazardrate

Die Survivorfunktion einer Verteilung ist definiert durch:

$$S(x) := P(X \geq x) = 1 - F(x)$$

Die Hazardrate ist definiert durch:

$$\lambda(x) := \lim_{h \to 0} \frac{P(x \leq X \leq x + h | X \geq x)}{h}$$

Zur Interpretation der Hazardrate

- Teil 1: bedingte Wahrscheinlichkeit mit Argument $\{x \leq X \leq x + h\}$ (Tod zwischen den Zeitpunkten x und $x + h$)
- Teil 2: bedingendes Ereignis $\{X \geq x\}$: Überleben bis mindestens zum Zeitpunkt x
- Teil 3: Intensität relativ zur Größe des betrachteten Intervalls $[x, x + h]$ mit Breite h.
- Teil 4: Grenzwert h gegen 0 betrachten, d.h. h sehr klein machen.
- Insgesamt: grobe, anschauliche Deutung:

  Risiko, im nächsten Moment zu „sterben“, wenn man bis zum Zeitpunkt x „überlebt“ hat.
- Beachte: $\lambda(\cdot)$ ist keine Wahrscheinlichkeit, kann Werte zwischen 0 und unendlich annehmen.
- Sehr anschauliches Instrument zur Beschreibung von Lebensdauerverteilungen.

Zusammenhänge zwischen S, F und Hazard

Es gelten folgende Zusammenhänge:

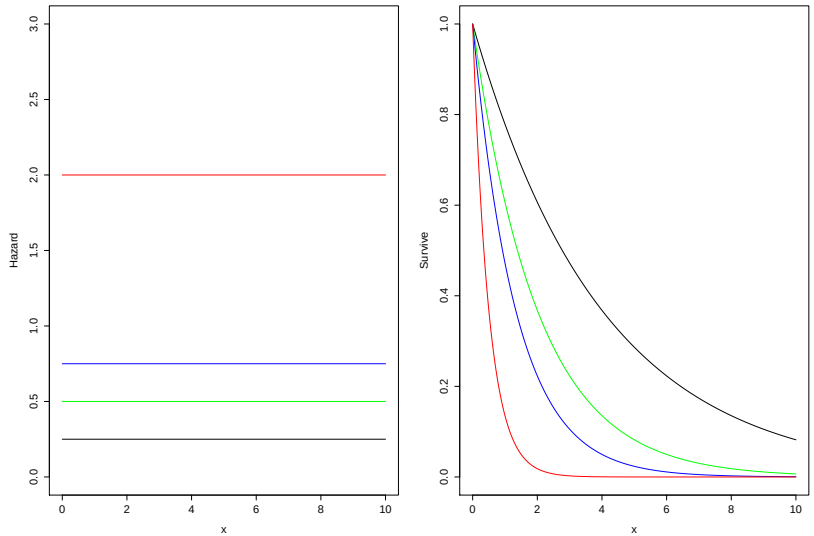
$$\begin{aligned}
\lambda(x) &= \frac{f(x)}{S(x)} \\
S(x) &= \exp\left(-\int_0^x \lambda(u)du\right) \\
F(x) &= 1 - \exp\left(-\int_0^x \lambda(u)du\right) \\
f(x) &= \lambda(x) \cdot S(x)
\end{aligned}$$

Beispiel: Warten auf den Bus

X ist die Zeit bis der Bus kommt (10 Minuten Takt).
Für $x \in (0; 10)$ gilt:

$$\begin{aligned} f(x) &= 0.1 \\ F(x) &= 0.1x \\ S(x) &= 1 - 0.1x = 0.1 * (10 - x) \\ \lambda(x) &= \frac{f(x)}{S(x)} = \frac{1}{10 - x} \end{aligned}$$

Exponentialv. Hazardrate und Survivorfunktion



Weibullverteilung

$\mathbf{X} \sim \mathbf{Wb}(\mathbf{c}, \alpha)$

1. **Modell:** Verteilung für Bruchfestigkeit von Materialien. Die Verteilung ist auch durch ihre Hazardrate charakterisiert und wird daher auch als Lebensdauerverteilung benutzt.
2. **Dichte, Verteilungsfunktion und Momente**

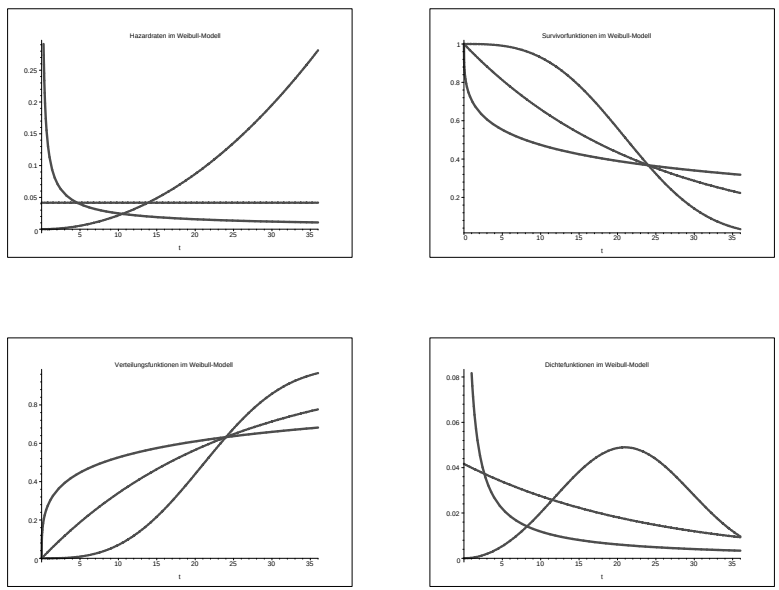
$$\begin{aligned} f(x) &= cx^{c-1}/\alpha^c \cdot \exp\left(- \left(\frac{x}{\alpha}\right)^c \right) \\ F(x) &= 1 - \exp\left(- \left(\frac{x}{\alpha}\right)^c \right) \end{aligned}$$

3. **Hazardrate**

$$\lambda(x) = \frac{f(x)}{1 - F(x)} = \frac{c}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}$$

4. Für c= 1 erhält man die Exponentialverteilung

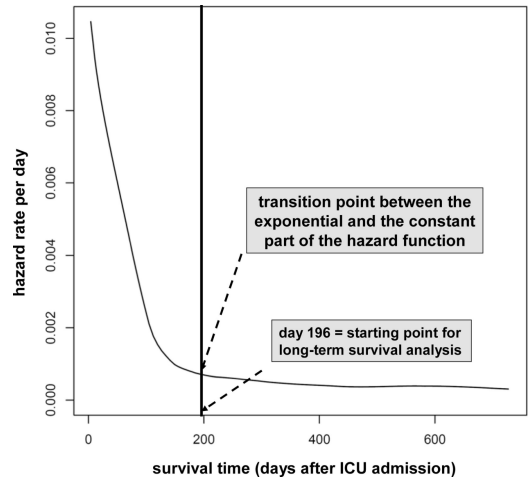
Beispiele Weibullverteilungen



Beispiel: Überleben von Intensivpatienten

- Studie in Kooperation mit W. Hartl (Klinikum Großhadern)
- 1462 Patienten, die mehr als 4 Tage auf der Intensivstation waren
- Fragestellung: Wie ist der Risikoverlauf (Hazrad) für Intensivpatienten
- Wie lange dauert es bis die Hazarrate konstant wird ?
- Modell mit Weibullverteilung in zwei Phasen

Schätzung des Verlaufs

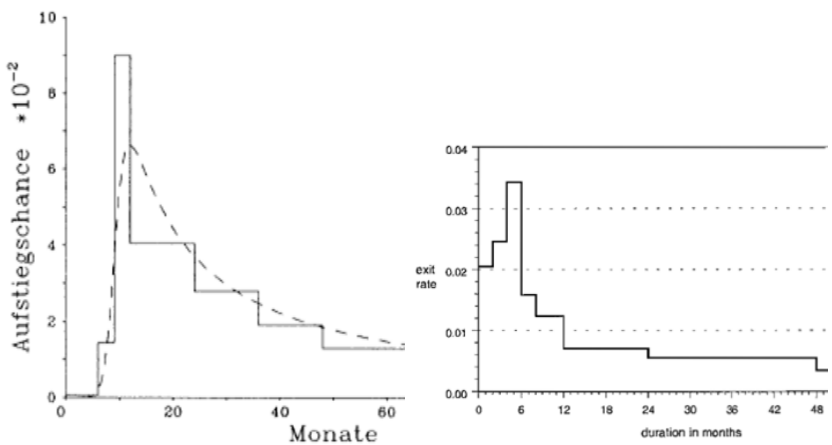


Beispiel: Mobilität in Betrieben

- J. Brüderl (1990): Mobilitätsprozesse in Betrieben
- Personaldaten 1976-1984 der Arbeiter eines großen süddeutschen Maschinenbauunternehmens
- Analyse von Zeitdauern bis zur Beförderung bzw. Verlassen des Betriebs

Das Honeymoon-Modell nach J. Brüderl

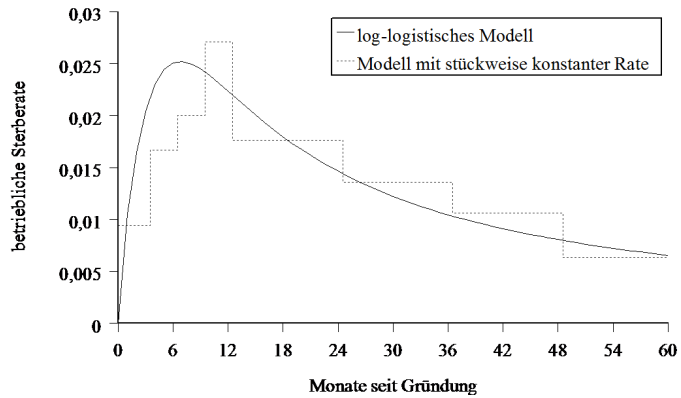
Hazardrate Aufstieg und Verlassen des Betriebs



Die Münchener Gründerstudie

- Brüderl/Preisendörfer/Ziegler (1996) Der Erfolg neugegründeter Betriebe. Duncker & Humblot.
- Gewerbemeldedaten der IHK München/Oberbayern 1985/86
- Mündliche Befragung von 1.849 Unternehmensgründern im Jahr 1990

Betriebliche Sterberaten



Modellierung mit der Log-logistischen Verteilung

Grenzwertsätze: Einführung

Gerade in der Soziologie beobachtet man häufig *große* Stichprobenumfänge.

- Was ist das Besondere daran?
- Vereinfacht sich etwas und wenn ja was?
- Kann man „Wahrscheinlichkeitsgesetzmäßigkeiten“ durch Betrachten vielfacher Wiederholungen erkennen?

Das i.i.d.-Modell

Betrachtet werden diskrete oder stetige Zufallsvariablen $X_1, \ldots, X_n$, die *i.i.d.* (independently, identically distributed) sind, d.h. die

1) unabhängig sind und
2) die gleiche Verteilung besitzen.

Ferner sollen der Erwartungswert μ und die Varianz σ^2 existieren. Die Verteilungsfunktion werde mit F bezeichnet.
Dies bildet insbesondere die Situation ab in der $X_1, \ldots, X_n$ eine Stichprobe eines Merkmals $\tilde{X}$ bei einer einfachen Zufallsauswahl sind.

Beispiel:
$\tilde{X}$ Einkommen, n Personen zufällig ausgewählt

X_1	Einkommen der	ersten	zufällig ausgewählten Person
X_2	Einkommen der	zweiten	zufällig ausgewählten Person
$\vdots$		$\vdots$	
X_n	Einkommen der	n-ten	zufällig ausgewählten Person

Stichprobenvariable

Jede Funktion von $X_1, \ldots, X_n$ ist wieder eine Zufallsvariable, z.B. das arithmetische Mittel oder die Stichprobenvarianz

$$\frac{1}{n}\sum_{i=1}^{n} X_i \qquad \tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Wahrscheinlichkeitsaussagen möglich $\Longrightarrow$ Wahrscheinlichkeitsrechnung anwenden

- Gerade bei diesen Zufallsgrößen ist die Abhängigkeit von n oft wichtig, man schreibt dann $\bar{X}_n$, $\tilde{S}_n^2$
- Sind $X_1, \ldots, X_n$ jeweils $\{0,1\}$-Variablen, so ist $\bar{X}_n$ gerade die empirische *relative Häufigkeit* von Einsen in der Stichprobe vom Umfang n. Notation: H_n

Erwartungswert und Varianz von $\bar{X}_n$

$X_1, X_2, \ldots, X_n$ seien unabhängig und identisch verteilt.

$$X_1, X_2, \ldots, X_n \qquad i.i.d.$$

Ist $\mathbb{E}(X_i) = \mu$ und $Var(X_i) = \sigma^2$, so gilt:

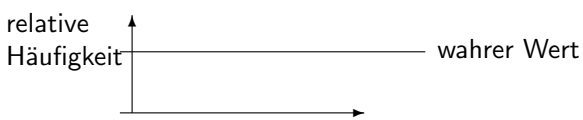
$$\begin{aligned}
\mathbb{E}(X_1 + X_2 + \ldots + X_n) &= n\mu \\
Var(X_1 + X_2 + \ldots + X_n) &= n\sigma^2 \\
\mathbb{E}\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) &= \mu \\
Var\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) &= \frac{\sigma^2}{n}
\end{aligned}$$

Diese Eigenschaften bilden die Grundlage für die folgenden Sätze.

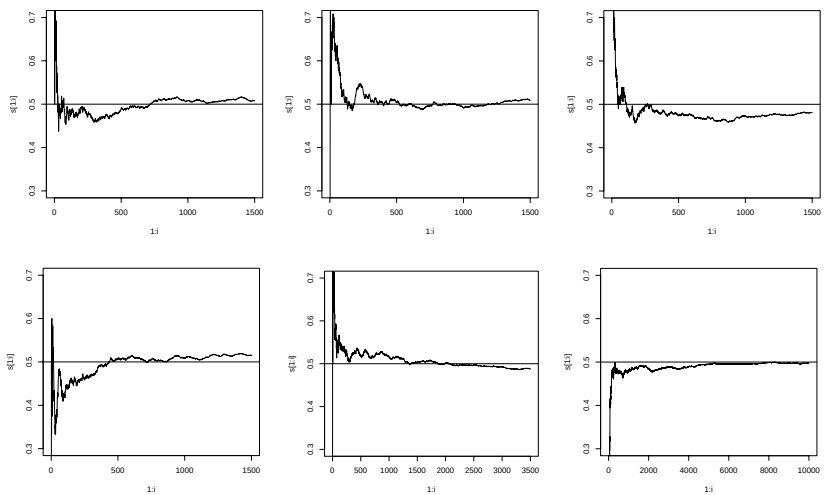
Das schwache Gesetz der großen Zahlen

Betrachte für wachsenden Stichprobenumfang n:

- $X_1, \ldots, X_n$ i.i.d.
- $X_i \in \{0,1\}$ binäre Variablen mit $\pi = P(X_i = 1)$
  Beispiele: Pro/Contra, Kopf/Zahl, A tritt ein/A tritt nicht ein
- H_n = die relative Häufigkeit der Einsen in den ersten n Versuchen.



Simulationen



Beobachtungen

1. Am Anfang sehr unterschiedlicher, unregelmäßiger Verlauf der Pfade.
2. Mit wachsendem n pendeln sich die Pfade immer stärker um π herum ein, d.h. mit wachsendem Stichprobenumfang konvergiert die relative Häufigkeiten eines Ereignisses gegen seine Wahrscheinlichkeit.
3. Formalisierung von 2.: Legt man sehr kleine Korridore/Intervalle um π, so ist bei sehr großem n der Wert von H_n fast sicher in diesem Korridor.

   Das Ereignis „Die relative Häufigkeit H_n liegt im Intervall der Breite 2ϵ um π„ lässt sich schreiben als:

$$\begin{aligned} \pi - \varepsilon \leq \quad & H_n & \leq \pi + \varepsilon \\ -\varepsilon \leq \quad & H_n - \pi & \leq \varepsilon \\ & |H_n - \pi| & \leq \varepsilon \end{aligned}$$

Theorem von Bernoulli

Seien $X_1, \ldots, X_n$, i.i.d. mit $X_i \in \{0, 1\}$ und $P(X_i = 1) = \pi$. Dann gilt für

$$H_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(relative Häufigkeit der „Einsen“) und beliebig kleines $\epsilon > 0$

$$\lim_{n \to \infty} P(|H_n - \pi| \leq \epsilon) = 1$$

Anschauliche Interpretation: Die relative Häufigkeit eines Ereignisses nähert sich praktisch sicher mit wachsender Versuchszahl an die Wahrscheinlichkeit des Ereignisses an.

Zwei wichtige Konsequenzen

1) **Häufigkeitsinterpretation von Wahrscheinlichkeiten:**

   $P(A)$, die Wahrscheinlichkeit eines Ereignisses A, kann man sich vorstellen als Grenzwert der relativen Häufigkeit des Eintretens von A in einer unendlichen Versuchsreihe identischer Wiederholungen eines Zufallsexperiments.

2) **Induktion:** Man kann dieses Ergebnis nutzen, um Information über eine unbekannte Wahrscheinlichkeit ($\pi \mathrel{\hat{=}}$ Anteil in einer Grundgesamtheit) zu erhalten.

   Sei z.B. π der (unbekannte) Anteil der SPD Wähler, so ist die relative Häufigkeit in der Stichprobe eine „gute Schätzung für π“. Je größer die Stichprobe ist, umso größer ist die Wahrscheinlichkeit, dass die relative Häufigkeit sehr nahe beim wahren Anteil π ist.

Gesetz der großen Zahl (allgemein)

Das Ergebnis lässt sich verallgemeinern auf Mittelwerte beliebiger Zufallsvariablen:

Gegeben seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit (existierendem) Erwartungswert μ und (existierender) Varianz σ^2. Dann gilt für

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

und beliebiges $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| \leq \epsilon) = 1$$

Schreibweise:

$$\bar{X}_n \xrightarrow{P} \mu$$

(„Stochastische Konvergenz“, „X_n konvergiert in Wahrscheinlichkeit gegen μ“.)

Konsequenz

- **Interpretation des Erwartungswerts:** μ kann in der Tat interpretiert werden als Durchschnittswert in einer unendlichen Folge von Wiederholungen des Zufallsexperiments.
- **Spiele.** Wenn ein Spiel mit negativem Erwartungswert häufig gespielt wird, verliert man mit sehr hoher Wahrscheinlichkeit (Grund für Rentabilität von Spielbanken und Wettbüros)

Die Verteilungsfunktion

Jetzt betrachten wir die empirische Verteilungsfunktion: In jedem Punkt x ist $F_n(x)$ vor der Stichprobe eine Zufallsvariable, also ist F_n eine zufällige Funktion

Wie vergleicht man die zufällige Funktion $F_n(x)$ mit der Funktion $F(x)$? Der Abstand hängt ja von dem Punkt x ab, in dem gemessen wird!

Idee: Maximaler Abstand

$$\max_{x \in R} |F_n^{X_1,\ldots,X_n}(x) - F(x)|$$

Existiert nicht immer; formal muss man das sogenannte Supremum betrachten.

Hauptsatz der Statistik

Seien $X_1, \ldots, X_n$ i.i.d. mit Verteilungsfunktion F und sei $F_n(x)$ die empirische Verteilungsfunktion der ersten n Beobachtungen. Mit

$$D_n := \sup_x |F_n(x) - F(x)|,$$

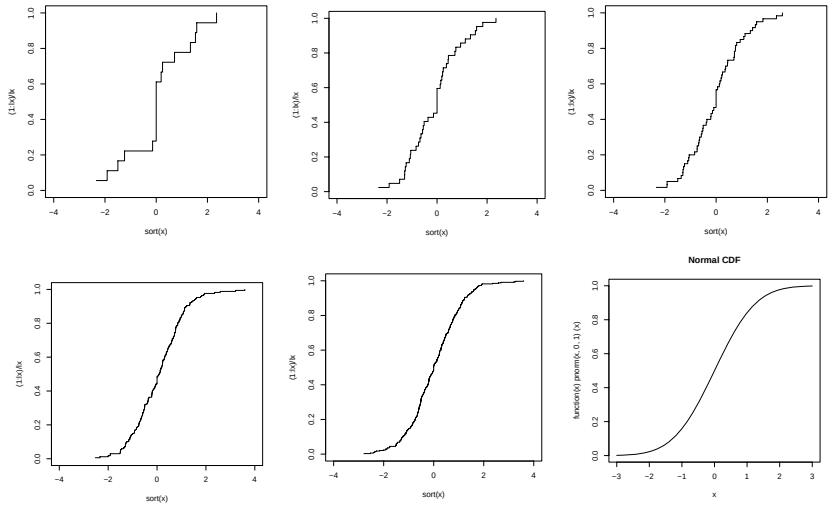
gilt für jedes $c > 0$

$$\lim_{n\to\infty} P(D_n > c) = 0.$$

Interpretation

- „Erträglichkeitsschranke“ c vorgegeben. Wsk, dass maximaler Abstand größer c ist geht für hinreichend großes n gegen 0 $\Longrightarrow$ überall kleiner Abstand. Man kann $\{D_n > c\}$ interpretieren als „Die Stichprobe führt den Betrachter hinter das Licht.“. Dann ist also die Wahrscheinlichkeit mit hinreichend großem n praktisch null.
- Anschaulich: Praktisch sicher spiegelt die empirische Verteilungsfunktion einer unendlichen Stichprobe die wahre Verteilungsfunktion wider.
- Falls die Stichprobe groß genug ist, so wird letztendlich immer representativ für die Grundgesamtheit, d.h. man kann Verteilungsgesetzmäßigkeiten durch Beobachtungen erlernen (grundlegend für die Statistik) $\rightarrow$ „Hauptsatz “.

Beispiele



Der zentrale Grenzwertsatz I

- Gibt es für große Stichprobenumfänge Regelmäßigkeiten im Verteilungstyp?
- Gibt es eine Standardverteilung, mit der man oft bei großen empirischen Untersuchungen rechnen kann?

Der zentrale Grenzwertsatz II

Seien $X_1, \ldots, X_n$ i.i.d. mit $\mathbb{E}(X_i) = \mu$ und $\text{Var}(X_i) = \sigma^2 > 0$ sowie

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right).$$

Dann gilt: Z_n ist *asymptotisch standardnormalverteilt*, in Zeichen: $Z_n \overset{a}{\sim} \text{N}(0; 1)$, d.h. es gilt für jedes z

$$\lim_{n \to \infty} P(Z_n \leq z) = \Phi(z).$$

Für die Eingangsfragen gilt also:
Ja, wenn man die Variablen geeignet mittelt und standardisiert, dann kann man bei großem n näherungsweise mit der Normalverteilung rechnen. Dabei ist für festes n die Approximation umso besser, je „symmetrischer“ die ursprüngliche Verteilung ist.

Standardisieren

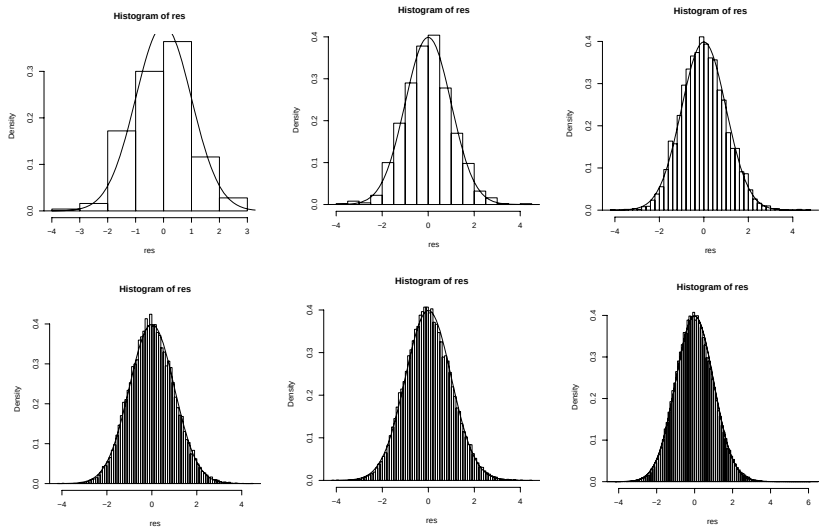
Die Funktion kommt durch *Standardisieren* und durch *geeignetes mitteln* zustande.

Dabei ist es wichtig, durch $\sqrt{n}$ (und nicht durch n) zu teilen.

$\sum X_i \quad \longrightarrow$ verliert sich; $\text{Var}(\sum X_i) \to \infty$

$\frac{1}{n} \sum x_i \quad \longrightarrow \text{Var}\left(\frac{1}{n} \sum X_i\right) \to 0$

Beispiele



Anwendung des zentralen Grenzwertsatz auf $\bar{X}$ I

Gemäß dem Gesetz der großen Zahlen weiß man: $\bar{X}_n \longrightarrow \mu$

Für die Praxis ist es aber zudem wichtig, die konkreten Abweichungen bei großem aber endlichem n zu quantifizieren, etwa zur Beantwortung folgender Fragen:

- Gegeben eine Fehlermarge ε und Stichprobenumfang n: Wie groß ist die Wahrscheinlichkeit, dass $\bar{X}$ höchstens um ε von μ abweicht?
- Gegeben eine Fehlermarge ε und eine „Sicherheitswahrscheinlichkeit“ γ: Wie groß muss man n mindestens wählen, damit mit mindestens Wahrscheinlichkeit γ das Stichprobenmittel höchstens um ε von μ abweicht *(Stichprobenplanung)*?

Anwendung des zentralen Grenzwertsatz auf $\bar{X}$ II

Aus dem zentralen Grenzwertsatz folgt:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right) = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n} \cdot \sigma}$$

$$= \frac{n\bar{X}_n - n\mu}{\sqrt{n} \cdot \sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} \mathrm{N}(0,1)$$

oder auch

$$\bar{X}_n \overset{a}{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n} \right).$$

$\frac{\sigma^2}{n}$ wird mit wachsendem n immer kleiner

* Schwankung im richtigen Wert (μ)
* Ausschläge werden kleiner

Approximation der Binomialverteilung I

Sei $X \sim B(n, \pi)$. Kann man die Verteilung von X approximieren?

Hier hat man zunächst nur ein X. Der zentrale Grenzwertsatz gilt aber für eine Summe vieler Glieder.
Idee: Schreibe X als Summe von binären Zufallsvariablen.
X ist die Anzahl der Treffer in einer $i.i.d.$ Folge $Y_1, \ldots, Y_n$ von Einzelversuchen, wobei

$$Y_i = \begin{cases} 1 & \text{Treffer} \\ 0 & \text{kein Treffer} \end{cases}$$

Derselbe Trick wurde bei der Berechnung von Erwartungswerten angewendet.
Die Y_i sind i.i.d. Zufallsvariablen mit $Y_i \sim Bin(1, \pi)$ und es gilt

$$X = \sum_{i=1}^{n} Y_i, \quad \mathbb{E}(Y_i) = \pi, \quad \mathrm{Var}(Y_i) = \pi \cdot (1 - \pi).$$

Approximation der Binomialverteilung II

Damit lässt sich der zentrale Grenzwertsatz anwenden:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{Y_i - \pi}{\sqrt{\pi(1-\pi)}} \right) = \frac{1}{\sqrt{n}} \frac{\sum Y_i - n \cdot \pi}{\sqrt{\pi(1-\pi)}}$$
$$= \frac{\sum Y_i - n \cdot \pi}{\sqrt{n \cdot \pi(1-\pi)}} \overset{a}{\sim} \mathrm{N}(0,1)$$

und damit

$$\frac{X - \mathbb{E}(X)}{\sqrt{\mathrm{Var}(X)}} \overset{a}{\sim} \mathrm{N}(0,1)$$

so dass

$$P(X \leq x) \approx \Phi \left(\frac{x - n \cdot \pi}{\sqrt{n \cdot \pi(1-\pi)}} \right)$$

falls n groß genug.

Faustregeln

Es gibt verschiedene Faustregeln, ab wann diese Approximation gut ist, z.B.

$$n \cdot \pi \geq 5 \text{ und } n \cdot (1-\pi) \geq 5$$
$$n \cdot \pi(1-\pi) \geq 9$$

Wichtig: Ob die Approximation hinreichend genau ist, hängt insbesondere ab vom substanzwissenschaftlichen Kontext ab.

Stetigkeitskorrektur

Durch die Approximation der *diskreten* Binomialverteilung durch die *stetige* Normalverteilung geht der diskrete Charakter verloren. Man erhält als Approximation $P(X = x) \approx 0$ für jedes $x \in N$, was gerade für mittleres n unerwünscht ist.
Benutze deshalb

$$P(X \leq x) = P(X \leq x + 0.5)$$

bei ganzzahligem $x \in N$.

Man erhält als bessere Approximation

$$P(X \leq x) \approx \Phi \left(\frac{x + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} \right)$$

$$P(X = x) \approx \Phi \left(\frac{x + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} \right) - \Phi \left(\frac{x - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} \right)$$

Beispiel

Ein Politiker ist von einer gewissen umstrittenen Maßnahme überzeugt und überlegt, ob es taktisch geschickt ist, zur Unterstützung der Argumentation eine Mitgliederbefragung zu dem Thema durchzuführen. Er wählt dazu 200 Mitglieder zufällig aus und beschließt, eine Mitgliederbefragung zu „riskieren“, falls er in der Stichprobe mindestens 52% Zustimmung erhält.

Wie groß ist die Wahrscheinlichkeit, in der Stichprobe mindestens 52% Zustimmung zu erhalten, obwohl der wahre Anteil nur 48% beträgt?

Lösung

- X Anzahl der Ja-Stimmen
- X ja/nein $\Rightarrow$ Binomialmodell
- $X \sim B(n, \pi)$ mit $n = 200$ und $\pi = 48\%$
- $n \cdot \pi = 96$ und $n \cdot (1 - \pi) = 104$: Faustregel erfüllt, die Normalapproximation darf also angewendet werden.

Gesucht: W'keit dass mind. 52%, also 104 Mitglieder, zustimmen, d.h.

$$\begin{aligned} P(X \geq 104) &= 1 - P(X \leq 103) \\ &= 1 - \Phi\left(\frac{x + 0.5 - n\pi}{\sqrt{n \cdot \pi(1 - \pi)}}\right) \\ &= 1 - \Phi\left(\frac{103.5 - 200 \cdot 0.48}{\sqrt{200 \cdot 0.48(1 - 0.48)}}\right) \\ &= 1 - \Phi(1.06) \\ &= 1 - 0.8554 = 14.5\% \end{aligned}$$