1 Der Begriff der Wahrscheinlichkeit

Stochastik befasst sich mit *Zufallsexperimenten*. Deren Ergebnisse sind unter „Versuchsbedingungen“verschieden.

Bsp.
- *Kartenziehen, Würfeln, Roulette*
- *Simulation*
- *Komplexe Phänomene (zumindest approximativ): Börse, Data-Mining, Genetik, Wetter*

Ergebnisse von Zufallsexperimenten werden in *Ereignisse* zusammengefasst.

- *Ereignisraum* (Grundraum) Ω: Menge aller möglichen Ergebnisse des Zufallsexperiments
- *Elementarereignisse* ω: Elemente von Ω, also die möglichen Ergebnisse des Zufallsexperiments
- *Ereignis*: Teilmenge von Ω
- Operationen der Mengenlehre haben natürliche Interpretation in der Sprache der Ereignisse:

Durchschnitt	$A \cap B$	*A und B*
Vereinigung	$A \cup B$	*A oder B*
Komplement	A^c	*Nicht A*
Differenz	$A \setminus B$	*A ohne B*

Das Vorgehen der Stochastik zur Lösung eines Problems kann in drei Schritte unterteilt werden:

1. Man bestimmt die Wahrscheinlichkeiten gewisser Ereignisse A_i. Dabei sind Expertenwissen, Daten und Plausibilitäten wichtig.
2. Man berechnet aus den Wahrscheinlichkeiten $\mathbb{P}(B_j)$ die Wahrscheinlichkeiten von gewissen anderen Ereignissen B_j gemäss den Gesetzen der Wahrscheinlichkeitstheorie (oft vereinfachend unter Unabhängigkeitsannahme).
3. Man interpretiert die Wahrscheinlichkeiten $\mathbb{P}(B_j)$ im Hinblick auf die Problemstellung.

Das *Bestimmen von Wahrscheinlichkeiten* (siehe Schritt 1) wird oft konkreter formalisiert.

Bsp. (Laplace-Modell)
Die Wahrscheinlichkeit von einem Ereignis A ist gegeben durch

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \quad \left(= \frac{\#\text{günstige Fälle}}{\#\text{mögliche Fälle}}\right).$$

Dem Laplace-Modell liegt die uniforme Verteilung *von Elemntarereignissen ω zugrunde:*

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}$$

Andere Wahrscheinlichkeitsverteilungen werden mit Hilfe des Konzepts von *Zufallsvariablen* (siehe Kapitel 2) eingeführt. Es sei aber bereits hier festgehalten: die Stochastik geht weit über das Laplace-Modell hinaus. Für viele Anwendungen ist das Laplace-Modell ungeeignet.

1.1 Rechenregeln für Wahrscheinlichkeiten

Die drei Axiome sind:

(A1) $\mathbb{P}(A) \geq 0$: Wahrscheinlichkeiten sind immer nicht-negativ.

(A2) $\mathbb{P}(\Omega) = 1$: sicheres Ereignis Ω hat Wahrscheinlichkeit eins.

(A3) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ $\forall$ Ereignisse A, B, die sich gegenseitig ausschliessen (d.h. $A \cup B = \varnothing$).

Weitere (abgeleitete) Regeln:

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

für jedes Ereignis A,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

für je zwei Ereignisse A und B,

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$$

für je n Ereignisse $A_1, \ldots, A$

$$\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$$

für je zwei Ereignisse A und B mit $A \subset B$.

1.2 Unabhängigkeit von Ereignissen

Wenn zwischen zwei Ereignissen A und B kein kausaler Zusammenhang besteht (d.h. es gibt keine gemeinsamen Ursachen oder Ausschliessungen), dann werden sie *unabhängig* genannt, genauer: Zwei Ereignisse A und B heissen (*stochastisch*) *unabhängig* wenn für jedes $k \leq n$ und all $1 \leq i_1 < \ldots < i_k \leq n$ gilt

$$\mathbb{P}(A_{i_1} \cap \cdots \cup A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Achtung: Zwei *unabhängige* Ereignisse A und B sind *nicht disjunkt* (und umgekehrt), vorausgesetzt die W'keiten $\mathbb{P}(A), \mathbb{P}(B) \neq 0$.

1.3 Interpretation von Wahrscheinlichkeiten

Die beiden wichtigsten Interpretationen sind:

- frequentistisch: „Idealisierung der relative Häufigkeiten bei vielen unabhängigen Wiederholungen“
- subjektive: „Mass für den Glauben, dass ein Ereignis eintreten wird“(Bayes'sch)

2 Zufallsvariable und Wahrscheinlichkeitsverteilung

Ergebnisse eines physikalischen Versuchs (Zufallsexperiment) sind oft Zahlen (Messungen). Diese werden als Beobachtung von so genannten Zufallsvariablen interpertiert, d.h. beobachtet wird nicht das ω welches bei einem Zufallsexperiment herauskommnt, sondern die Werte aller beobachtetetn Zufallsvariablen.

2.1 Definition einer Zufallsvariablen

Eine Zufallsvariable X ist ein Zuffalsexperiment mit möglichen Werten in $\mathbb{R}$, bzw. in einer Teilmenge von $\mathbb{R}$, z.B. $\mathbb{N}_0 = \{0, 1, \ldots\}$. Deren Wert ist im Voraus nicht bekannt, sondern hängt vom Ergebnis eines Zufallsexperiments ab. Mathematisch ist eine Zufallsvariable einfach nur eine Abbildung von Ω nach $\mathbb{R}$:

$$X : \Omega \to \mathbb{R},$$
$$\omega \mapsto X(\omega).$$

Das heisst wenn das Ergebnis ω herauskommt, nimmt die Zufallsvariable den Wert $X(\omega)$ an.

Bsp. (Wert einer zuf. gez. Jasskarte.)
Sei $\Omega = \{Jasskarten\}$; ein $\omega \in \Omega$ ist z.B. ein Schilten-As; Zufallsvariable X:

$$\textit{As irgendeiner Farbe} \mapsto 11$$
$$\textit{König irgendeiner Farbe} \mapsto 4$$
$$\textit{„Brettchen“ irgendeiner Farbe} \mapsto 0.$$

2.2 Wahrscheinlichkeitsverteilung auf $\mathbb{R}$

Eine Zufallsvariable X legt eine Wahrscheinlichkeit Q auf $\mathbb{R}$ fest, die segenannte *Verteilung* von X:

$$Q(B) = \mathbb{P}\Big(\big\{\omega; X(\omega) \in B\big\}\Big) = \mathbb{P}(X \in B)$$

Bsp. (Wert einer zuf. gez. Jk. Forts.)
In obigem Beispiel ist beispielsweise

$$Q(11) = \mathbb{P}(\textit{As irgendeiner Farbe}) = \frac{4}{36}.$$

Die *kumulative Verteilungsfunktion* ist definiert als

$$F(b) = \mathbb{P}(X \leq b) = Q\big((-\infty, b]\big).$$

Sie enthält dieselbe Information wie die Verteilung $Q(\cdot)$, ist aber einfacher darzustellen. Die Umkehrung der Verteilungsfunktion stellen di esogenannten Quantile dar, für $\alpha \in (0, 1)$ ist das *α-Quantil von X* definiert als das kleinste $x \in \mathbb{R}$ für welches $F(x) \geq \alpha$ gilt, also

$$q_\alpha \coloneqq q(\alpha) \coloneqq \min\{x \in \mathbb{R} \mid F(x) \geq \alpha\}$$

Es gilt

$$F(q_\alpha) = \alpha$$

bzw. äquivalent dazu

$$q_\alpha = F^{-1}(\alpha).$$

Das $\frac{1}{2}$-Quantil von X heisst auch *Median von X*.

2.3 Diskrete und stetige Zufallsvariablen

Eine Zufallsvariable X heisst *diskret*, falls die Menge W der möglichen Werte von X endlich oder abzählbar ist. Zum Beispiel $W = \{0, 1, 2, \ldots, 100\}$ oder $W = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$. Die Verteilung einer diskreten

Zufallsvariablen ist festgelegt durch die Angabe der sogenannten *Wahrscheinlichkeitsfunktion*:

$$p(x) := \mathbb{P}(X = x), \qquad x \in W.$$

Offensichtlich ist die kumulative Verteilungsfunktion eine Treppenfunktion mit Sprüngen an den Stellen $x \in W$ mit Sprunghöhen $p(x)$, also nicht stetig. Ferner gilt

$$Q(B) = \sum_{x \in B} p(x).$$

Eine Zufallsvariable X heisst *stetig* falls die Menge der möglichen Werte W ein Intervall enthält. Zum Beispiel $W = [0, 1]$ oder $W = \mathbb{R}$.

2.4 Erwartungswert und Varianz

Eine Verteilung einer Zufallsvariablen X kann durch mindestens zwei Kennzahlen zusammengefasst werden, eine für die Lage (der Erwartungswert $E(X) = \mu_X$) und eine für die Streuung (die Standardabweichung σ_X). Der *Erwartungswert* einer diskreten Zufallsvariable X ist definiert durch

$$\begin{aligned}\mu_X &= E(X) \\ &:= \sum_{x \in W} xp(x)\end{aligned}$$

Für den Erwartungswert einer tranformierten diskreten Zufallsvariable $Y = f(X)$ ergibt sich daraus:

$$\begin{aligned}E(Y) &= E(f(X)) \\ &= \sum_{x \in W} f(x)p(x)\end{aligned}$$

Die *Varianz* einer diskreten Zufallsvariable X ist definiert durch:

$$\begin{aligned}V(X) &:= E((X - E(X))^2) \\ &= \sum_{x \in W} (x - \mu_X)^2 p(x)\end{aligned}$$

Die *Standardabweichung* ist die Wurzel aus der Varianz, d.h.

$$\sigma_X = \sqrt{V(X)}.$$

Folgende Rechenregeln sind nützlich:

$$\begin{aligned}E(a + bX) &= a + bE(X),\ a, b \in \mathbb{R} \\ V(X) &= E(X^2) - E(X)^2 \\ V(a + bX) &= b^2 V(X)\end{aligned}$$

In der frequentistischen Interpretation ist der Erwartungswert eine Idealisierung des arithmetischen Mittels der Werte einer Zufallsvariablen bei vielen Wiederholungen.

2.5 Die wichtigsten diskreten Verteilungen

2.5.1 Binomialverteilung

Die *Binomialverteilung* $\mathrm{Bin}(n, p)$ ist die Verteilung der Anzahl Erfolge bei n unabhängigen Wiederholungen eines Experiments mit Erfolgswahrscheinlichkeit p.

$$\begin{aligned}W &= 0, 1, \cdots, n \\ p(x) &= \binom{a}{b} p^x (1 - p)^{n-x} \\ E(X) &= np \\ \sigma_X &= \sqrt{np(1 - p)}\end{aligned}$$

2.5.2 Poissonverteilung

Die *Poissonverteilung* $\mathrm{Poi}(\lambda)$ ist eine Approximation der Binomialverteilung für grosses n und kleines p, mit $np = \lambda$. Die Anzahl Ausfälle einer Komponente oder eines Systems in einem interval der Länge t ist oft in erster Näherung Poisson-verteilt mit Parameter λt.

$$\begin{aligned}W &= 0, 1, \cdots \\ p(x) &= e^{-\lambda} \frac{\lambda^x}{x!} \\ E(X) &= \lambda \\ \sigma_X &= \sqrt{\lambda}\end{aligned}$$

2.5.3 Geometrische Verteilung

Die *geometrische Verteilung* $\mathrm{Geo}(p)$ ist die Verteilung der Anzahl Wiederholungen bis ein Ereignis mit Wahrscheinlichkeit p eintritt.

$$\begin{aligned}W &= 1, 2, \cdots \\ p(x) &= p(1 - p)^{x-1} \\ E(X) &= \frac{1}{p} \\ \sigma_X &= \frac{\sqrt{1 - p}}{p}\end{aligned}$$

3 Stetige Wahrscheinlichkeitsverteilung

Bei einer stetigen Zufallsvariablen X ist $\mathbb{P}(X = x) = 0$ für jedes feste x. Wir betrachten nur Fälle, wo $\mathbb{P}(x \le X \le x + h)$ für kleine h ungefähr proportional zu h ist. Die Proportionalitätskonstante heisst die *Dichte* f von X.

3.1 Wahrscheinlichkeitsdichte

Die Dichte von einer stetigen Verteilung P ist definiert als

$$f(x) = \lim_{h \downarrow 0} \frac{\mathbb{P}(x \le X \le x + h)}{h}$$

Zwischen der Dichte f und der kumulativen Verteilungsfunktion F bestehen die folgenden Beziehungen:

$$f(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} \mathrm{d}x\, f(x)$$

Erwartungswert und Varianz berechnen sich gemäss

$$\begin{aligned}\mathbb{E}(X) = \mu_X &= \int_{-\infty}^{\infty} \mathrm{d}x\, x f(x) \\ \mathbb{V}(X) = \sigma_X^2 &= \int_{-\infty}^{\infty} \mathrm{d}x\, (x - \mu_X)^2 f(x)\end{aligned}$$

und es gelten die gleichen Rechenregeln wie im diskreten Fall (TODO: Referenz hier?).

3.2 Die wichtigsten stetigen Verteilungen

3.2.1 Uniforme Verteilung

Die *uniforme Verteilung* $\mathrm{Uni}[a, b]$ tritt auf bei Rundungsfehlern und als Formalisierung der völligen "Ignoranz".

$$\begin{aligned}W &= [a, b] \\ f(x) &= \frac{1}{b - a} 1_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{sonst} \end{cases} \\ \mathbb{E}(X) &= \frac{a + b}{2} \\ \sigma_X &= \frac{b - a}{\sqrt{12}}\end{aligned}$$

3.2.2 Exponentialverteilung

Die *Exponentialverteilung* $\exp(\lambda)$ ist das einfachste Modell für Wartezeiten auf Ausfälle und eine stetige Version der geometrischen Verteilung.

$$\begin{aligned}W &= [0, \infty) \\ f(x) &= \lambda e^{-\lambda x}, \qquad \text{für } x > 0 \\ F(x) &= 1 - e^{-\lambda x} \\ \mathbb{E}(X) &= \sigma_X = \frac{1}{\lambda}\end{aligned}$$

Wenn die Zeiten zwischen den Ausfällen eines Systems Exponential(λ)-verteilt sind, dann ist die Anzahl Ausfälle in einem Intervall der Länge t $\mathrm{Poi}(\lambda t)$-verteilt.

3.2.3 Normal- oder Gaussverteilung

Die *Normal- oder Gaussverteilung* $\mathcal{N}(\mu, \sigma^2)$ ist die häufigste Verteilung für Messwerte.

$$\begin{aligned}W &= \mathbb{R} \\ f(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right) \\ \mathbb{E}(X) &= \mu \\ \sigma_X &= \sigma\end{aligned}$$

Die $\mathcal{N}(0, 1)$Verteilung, auch *Standardverteilung* bezeichnet, ist ein wichtiger Sonderfall, weshalb es für dessen Verteilungsfunktion sogar ein eigenes Symbol gibt: $\Phi(x) := F_{\mathcal{N}(0,1)}$, $x \in \mathbb{R}$ und deren Umkehrfunktionen (d.h. die Quantile) kürzen wir ab mit $z_\alpha := \Phi^{-1}(\alpha)$, $\alpha \in [0, 1]$. Die Verteilungsfunktion F einer $\mathcal{N}(\mu, \sigma^2)$-verteilten Zufallsvariable ist nicht geschlossen darstellbar, sie wird aus der Verteilungsfunktion Φ der Standardnormalverteilung (welche tabelliert ist) berechnet mittels der Formel:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad x \in \mathbb{R} \qquad (1)$$

3.3 Transformationen

Bei stetigen Verteilungen spielen Transformationen $Y = g(X)$ eine wichtige Rolle. Falls g *linear* ist: $g(x) = a + bx$ mit $b > 0$, dann gilt $\mathbb{E}(Y) = a+b\,\mathbb{E}(X), \sigma_Y = b\sigma_X, F_Y(x) = F_X((x-a)/b)$ und $f_Y(x) = f_X((x-a)/b)/b$. Durch Skalenänderungen kann man also alle Exponentialverteilungen ineinander überführen, und ebenso durch lineare Transformationen alle Normalverteilungen ineinander. Für beliebiges g gilt

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \int_{-\infty}^{\infty} \mathrm{d}x\, g(x) f(x) \qquad (2)$$

Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$ normalverteilt ist, dann heisst $Y = e^X$ *lognormal-verteilt*. Es gilt z.B. $\mathbb{E}(Y) = \exp(\mu + \sigma^2/2)$.

3.4 Simulation von Zufallsvariablen

Wenn U uniform auf $[0, 1]$ verteilt ist und F eine beliebige kumulative Verteilungsfunktion, dann ist die Verteilungsfunktion von $X = F^{-1}(U)$ gleich F. Dies ist ein wichtiges Faktum um Verteilungen, respektive Realisierungen von Zufallsvariabeln, zu *simulieren*:

1. Erzeuge Realisation u von uniform verteilter Zufallsvariable $U \sim \mathrm{Uni}[0, 1]$. Dies wird mittels einem "Standard-Paket" gemacht.
2. Berechne $x = F^{-1}(u)$. Gemäss obigem Faktum ist dann x eine Realisation einer Zufallsvariablen X mit kumulativer Verteilungsfunktion F.

Diese Methode ist nicht immer rechentechnisch effizient.

4 Mehrere Zufallsvariablen und Funktionen davon

Das Ziel ist hier, Genaueres über den Unterschied von $\mathbb{P}(A)$ und der relative Häufigkeit $f_N[A]$ respektive von A, respektive von $\mathbb{E}(X)$ und dem arithmetischen Mittel, bei n Wiederholungen zu sagen.

4.1 Die i.i.d. Annahme

Dabeid müssen wir präzisieren was eine „Wiederholung" ist. Die n-fache Wiederholung eines Zufallsexperimentes ist selber wieder ein Zufallsexperiment. Wenn A ein Ereignis im urpsrünglichen Experiment ist, bezeichnen wir mit A_i das Ereignis „A tritt bei der i-ten Wiederholung ein". Dann ist es sinnvoll, folgendes anzunehmen:

- $A_1, \ldots, A_n$ sind unabhängig: Unabhängigkeit der Ereignisse
- $\mathbb{P}(A_1) = \ldots = \mathbb{P}(A_n) = \mathbb{P}(A)$: gleiche Wahrscheinlichkeiten

Ebenso, wenn X die ursprüngliche Zufallsvariable ist, dann soll X_i die Zufallsvariable der i-ten Wiederholung bezeichnen. Die *i.i.d.* Annahme verlangt folgendes:

- $X_1, \ldots, X_n$ sind unabhängig
- alle X_i haben dieselbe Verteilung

Die Abkürzung „i.i.d." kommt vom Englischen: *i*ndependent and *i*dentically *d*istributed. Unabhängigkeit von Zufallsvariablen heisst, dass zum Beispiel

$$\mathbb{P}(X_i \in A \text{ und } X_j \in B) = \mathbb{P}(X_i \in A)\mathbb{P}(X_j \in B)$$

für alle $i \neq j$ und für all $A \subseteq \mathbb{R}, B \subseteq \mathbb{R}$, und analog für Trippel etc. Die i.i.d. ANnahme ist ein „Postulat", welches in der Praxis in vielen Fällen vernünftig scheint. Die Annahme bringt erhebliche Vereinfachungen um mit mehreren Zufallsvariablen zurechnen.

4.2 Funktionen von Zufallsvariablen

Ausgehend von $X_1, \ldots, X_n$ kann man neue Zufallsvariablen

$$Y = g(X_1, \ldots, X_n)$$

bilden. Hier betrachten wir die wichtigen Spezialfälle Summe

$$S_n = X_1 + \cdots + X_n$$

und arithmetisches Mittel

$$\overline{X}_n = \frac{S_n}{n}$$

Wir nehmen stets an, dass $X_1, \ldots, X_n$ i.i.d. sind. Wenn $X_i = 1$ falls ein bestimmtes Ereignis bei der i-ten Wiederholung eintritt und $X_i = 0$ sonst, dann ist $\overline{X}_n$ nichts anderes als die relative Häufigkeit dieses Ereignisses. Die Verteilung von S_n ist im allgemeinen schwierig exakt zu bestimmen, mit den folgenden Ausnahmen:

1. Wenn $X_i \in \{0, 1\}$ wie oben, dann ist $S_n \sim \mathrm{Bin}(n, p)$ mit $p = \mathbb{P}(X_i = 1)$.
2. Wenn $X_i \sim \mathrm{Poi}(\lambda)$, dann ist $S_n \sim \mathrm{Poi}(n\lambda)$.
3. Wenn $X_i \sim \mathcal{N}(\mu, \sigma^2)$, dann ist $S_n \sim$ TODO (normal$n\mu n\sigma^2$)

Einfacher sind die Berechnungen von Erwartungswert, Varianz und Standardabweichung, allgemein gilt

$$\mathbb{E}(S_n) = n\,\mathbb{E}(X_i), \qquad \mathbb{E}(\overline{X}_n) = \mathbb{E}(X_i);$$
$$\mathbb{V}(S_n) = n\,\mathbb{V}(X_i), \qquad \mathbb{V}(\overline{X}_n) = \frac{1}{n}\,\mathbb{V}(X_i);$$
$$\sigma_{S_n} = \sqrt{n}\sigma_{X_i}, \qquad \sigma_{\overline{X}_n} = \frac{1}{\sqrt{n}}\sigma_{X_i}.$$

Die Streuung der Summe wächst also, aber langsamer als die Anzahl Beobachtungen, während die Streuung des arithmetischen Mittels abnimmt, aber ebenfalls langsamer als die Anzahl Beobachtungen. Um die Genauigkeit des arithmetischen Mittels zu verdoppeln (d.h. die Standardabweichung zu halbieren), braucht man viermal soviele Beobachtungen. Die zufälligen Abweichungen von $\overline{X}_n$ zum Erwartungswert $\mathbb{E}(X)$ kompensieren sich in dem Sinne, dass $\sigma_{\overline{X}_n}$ abnimmt mit der Ordnung $1/\sqrt{n}$ wenn n wächst.

4.3 Das Gesetz der Grossen Zahlen und der Zentrale Grenzwertsatz

Von den obigen Formeln über Erwartungswert und Varianz wissen wir, dass:

1. $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X_i)$: das heisst $\overline{X}_n$ hat denselben Erwartungswert wie die einzelnen Variablen X_i.
2. $\mathbb{V}(\overline{X}_n) \xrightarrow{n\to\infty} 0$: das heisst, $\overline{X}_n$ besitzt keine Variabilität mehr im Limes.

Diese beiden Punkte implizieren den folgenden Satz.

Satz (Gesetz der Grossen Zahlen)
Seien $X_1, \ldots, X_n$ i.i.d. mit Erwartungswert μ. Dann

$$\overline{X}_n \xrightarrow{n\to\infty} \mu.$$

Als Spezialfall davon gilt:

$$f_n[A] \xrightarrow{n\to\infty} \mathbb{P}(A).$$

(Der Begriff der Konvergenz muss für Zufallsvariablen geeignet definiert werden).

Zur Berechnung der genäherten *Verteilung* von S_n und $\overline{X}_n$ (dies ist ein bedeutend präziseres Resultat als das GGZ) stützt man sich auf den folgenden berühmten Satz.

Satz (Zentraler Grenzwertsatz (ZGS))
Seien $X_1, \ldots, X_n$ i.i.d. mit Erwartungswert μ und Varianz σ^2, dann ist

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2),$$
$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

für grosse n.

Wie gut diese Approximationen für ein gegebenes n sind, hängt von der Verteilung der X_i ab. Mit der sogenannten *Chebychev-Ungleichung*

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > c\right) \leq \frac{\sigma^2}{nc)}$$

ist man stets auf der sicheren Seit. Dafür ist diese aber meistens ziemlich grob.

Immer wenn eine Zufallsvariable als eine Summe von vielen kleinen Effekten aufgefasst werden kann, ist sie wegen des Zentralen Grenzwertsatzes in erster Näherung normalverteilt. Das wichtigse Beispiel dafür sind Messfehler. Wenn sich die Effekte eher multiplizieren als addieren lassen, kommt man zur lognormal-Verteilung (Beispiel Teilchengrößen).

5 Gemeinsame und bedingte Wahrscheinlichkeiten

Oft besteht ein Zufallsexperiment aus verschiedenen Stufen, und amn erfährt das Resultat auch entsprechend diesen Stufen. Im einfachsten Fall erfährt man in der ersten Stufe, ob ein bestimmtes Ereignis B eingetreten ist oder nicht, und in der zweiten Stufe erfährt man, welches Ergebnis ω eingetreten ist.

5.1 Bedingte Wahrscheinlichkeit

Im allgemeinen wird die information aus der ersten Stufe die Unsicherheit über die zweite Stufe verändern, Dies modifizierten Unsicherheit wird gemessen durch die *bedingte Wahrscheinlichkein von A gegeben B*

bzw. B^c definiert durch

$$\mathbb{P}(A \mid B) \coloneqq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

bzw.

$$\mathbb{P}(A \mid B^c) \coloneqq \frac{\mathbb{P}(A \cap B^c)}{\mathbb{P}(B^c)}.$$

Dass diese Definition sinnvoll ist, kann man anhand der Entsprechung von Wahrscheinlichkeiten und relativen Häufigkeiten sehen. Insbesondere gilt für 2 Ereignisse A, B mit $\mathbb{P}(A) \neq 0$ und $\mathbb{P}(B) \neq 0$:

$$\begin{aligned} & A, B \text{ unabhängig} \\ \Leftrightarrow \quad & \mathbb{P}(A \mid B) = \mathbb{P}(A) \\ \Leftrightarrow \quad & \mathbb{P}(B \mid A) = \mathbb{P}(B). \end{aligned}$$

5.2 Satz der totalen Wahrscheinlichkeit und Satz von Bayes

Die obige Definition kann man aber auch als

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \mid B)P(B)$$

lesen, d.h. $\mathbb{P}(A \cup B)$ ist bestimmt durch $\mathbb{P}(A \mid B)$ und $\mathbb{P}(B)$. In vielen Anwendungen wird dieser Weg beschritten. Man legt die Wahrscheinlichkeiten für die erste Stufe $\mathbb{P}(B)$ und die bedingten Wahrscheinlichkeiten $\mathbb{P}(A \mid B)$ und $\mathbb{P}(A \mid B^c)$ für die zweite Stufe gegeben die erste fest (aufgrund von Daten, Plausibilität und subjektiven Einschätzungen). Dann lassen sich die übrigen Wahrscheinlichkeiten berechnen. Es gilt zum beispiel der folgende Satz:

Satz (Satz der tot. Wahrsch.keit I)

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cup B) + \mathbb{P}(A \cup B^c) \\ &= \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c) \end{aligned}$$

Dieses Vorgehen wird besonders anschaulich, wenn man das Experiment als Baum darstellt. Wenn man dagegen von den Wahrscheinlichkeiten der Durchschnitte ausgeht, wählt man besser eine Matrixdarstellung.

Wenn die einzelnen Stufen komplizierter sisdn, geht alles analog. Betrachte den Fall mit k Ereignissen auf der ersten Stufe $B_1, \dots, B_k$, wobei $B_i \cup B_j = \varnothing$ falls $i \neq j$ und $B_1 \cup \dots \cup B_k = \Omega$.

Satz (Satz der tot. Wahrsch.keit II)

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

In manchen Situationen erhält man die Information über die verschiedenen Stufen aber nicht in der ursprünglichen Reihenfolge, d.h. man kennt zuerst das Ergebnis der zweiten Stufe, weiss also z.B. dass A eingetreten ist. in einem solchen Fall will man die bedingten Wahrscheinlichkieten der ersten Stufe gegeben die zweite Stufe $\mathbb{P}(B_i \mid A)$ berechnen. Das Ergebnis liefert der folgende Satz:

Satz (Satz von Byes)

$$\begin{aligned} & \mathbb{P}(B_i \mid A) \\ &= \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\mathbb{P}(A \mid B_1)\mathbb{P}(B_1) + \dots + \mathbb{P}(A \mid B_k)\mathbb{P}(B_k)}. \end{aligned}$$

Oft ist das numerische Resultat einer solchen Berechnung stark verschieden von dem, was man naiverweise erwartet. Der Satz von Bayes ist vor allem in der subjektiven Wahrscheinlichkeitstheorie sehr wichtig: Wenn man für die verschiedenen Möglichkeiten $B_1, \cdots, B_k$ subjektive Wahrscheinlichkeiten festlegt und danach erfährt, dass A eingetreten ist, dann muss man die subjektiven Wahrscheinlichkeiten gemäss diesem Satz modifizieren.

5.3 Gemeinsame und bedingte diskrete Verteilungen

Die beiden, obig beschriebenen Stufen können auch ducrh Zufallsvariablen X und Y gegeben sein. Die gemeinsame Verteilung zweier diskreter Zufallsvariablen X und Y ist eindeutig charakterisiert durch ihre *gemeinsame Wahrscheinlichkeitsfunktion von X und Y*, d.h. die Werte

$$\mathbb{P}(X = x, Y = y), \qquad x \in W_X, y \in W_Y$$

weshalb diese dann auch (eigentlich fälschlicherweise) die gemeinsame Verteilung von X und Y genannt wird. In diesem „gemeinsamen" Zusammenhang nennt man die einzelnen Verteilungen $\mathbb{P}(X = x), x \in W_X$ von X und $\mathbb{P}(Y = y), y \in W_y$ von Y die *Randverteilungen (der gemeinsamen Zufallsvariable (X, Y))*, sie lassen sich aus der gemeinsamen Verteilung berechnen durch

$$\mathbb{P}(X = x) = \sum_{y \in W_y} \mathbb{P}(X = x, Y = y)$$

und analog für Y. Aus den randverteilungen auf die gemeinsame Verteilung zu schliessen geht *nur* im Falle der Unabhängigkeit von X und Y, denn es gilt: Zwei diskrete Zufallsvariablen X und Y sind unabhängig genau dann wenn

$$\begin{aligned} \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y), \\ x \in W_X, y \in W_y. \end{aligned}$$

In diesem Fall ist die gemeinsame Verteilung durch die Randverteilungen vollständig bestimmt und man erhält sie einfach durch Multiplikation. Schlussendlich defiiniert man noch die *bedingte Verteilung von X gegeben $Y = y$* durch die Werte

$$\mathbb{P}(X = x \mid X = y), \qquad x \in W_y.$$

Der Satz von der totalen Wahrscheinlichkeit lässt sich dann schreiben als

$$\mathbb{P}(X = x) = \sum_{y \in W_Y} \mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y)$$

und kommt immer dann zum Einsatz wenn man die Verteilung von X berechnen will, aber nur dessen bedingte Verteilung gegeben Y und die Verteilung von Y kennt.

Bei mehr als zwei STufen bzw. Zufallsvariablen geht alles analog. Die Bäume werden einfach länger und die Matrizen werden zu Feldern in höheren Dimensionen. Das Ganze wird aber sehr rasch unübersichtlich, und es gibt sehr viele Wahrscheinlichkeiten, die man zu Beginn festlegen muss. Eine wesentliche Vereinfachung erhält man, wenn man annimmt, dass jede STu zwar von der unmittelbar vorangehenden, aber nicht von den weiter zurückliegenden Stufen abhängt. Das fühlt auf die sogenannten *Markovketten*, deren Verhalten durch eine Startverteilung und eine Übergangsmatrix gegeben ist.

6 Gemeinsame und bedingte stetige Verteilungen

Bei zwei oder mehreren stetigen Zufallsvariablen kann die gemeinsame und bedingte Verteilung nicht mehr mit Bäumen oder Matrizen dargestellt werde wie in Kapitel 5. Bei zwei oder mehreren stetigen Zufallsvariablen kann die gemeinsame und bedingte Verteilung nicht mehr mit Bäumen oder Matrizen dargestellt werden wie in Kapitel 5.

6.1 Gemeinsame Dichte

Die gemeinsame Dichte $f_{x,Y}(\cdot, \cdot)$ von zwei stetigen Zufallsvariablen X und Y ist gegeben, in „Ingenieurnonation", durch

$$\begin{aligned} \mathbb{P}(x \leq X \leq x + \mathrm{d}x, y \leq Y \leq y + \mathrm{d}y) \\ = f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y. \end{aligned}$$

(Die Darstellugn als Ableitung einer geeigneten kumulativen Verteilungsfunktion ist nicht sehr instruktiv.) daraus kann man allgemein Wahrscheinlichkeiten durch Integration berechnen:

$$\mathbb{P}\big((X, Y) \in A\big) = \iint_A \mathrm{d}x\, \mathrm{d}y\, f_{X,Y}(x, y)$$

6.2 Randdichte und bedingte Dichte

Aus der gemeinsamen Dichte erhält man insbesondere die Randdichte von X bzw. Y

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} \mathrm{d}y\, f_{X,Y}(x, y), \\ f_Y(y) &= \int_{-\infty}^{\infty} \mathrm{d}x\, f_{X,Y}(x, y). \end{aligned}$$

Für die bedingte Verteilung von Y gegeben $X = x$ wird die bedingte Dichte benützt, definiert durch

$$\begin{aligned} f_{Y|X=x}(y) &\coloneqq f_Y(y \mid X = x) \\ &\coloneqq \frac{f_{X,Y}(x, y)}{f_X(x)} \end{aligned}$$

Aus den obigen Definitionen ist klar, dass all wahrscheinlichkeitstheoretischen Aspekte von 2 Zufallsvariablen X und Y durch deren gemeinsame Dichte $f_{X,Y}$ vollständig bestimmt sind. X und Y sind unabhängig genau dann wenn

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \qquad x, y \in \mathbb{R}^2. \quad (3)$$

In diesem Fall genügt das Konzept von 1-dimensionalen Dichten: die gemeinsame Dichte kann dann sehr einfach mittels Multiplikation berechnet werden.

6.3 Erwartungswert bei mehreren Zufallsvariablen

Der Erwartungswert macht nur Sinn für eine $\mathbb{R}$-wertige Gröss (oder Teilmenge von $\mathbb{R}$). Den Erwartungswert einer transformierten Zufallsvariable $Z = g(X, Y)$ mit $g : \mathbb{R}^2 \to \mathbb{R}$ können wir berechnen

als

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}y\, g(x,y) f_{X,Y}(x,y)$$

Im diskreten Fall lautet die entrpechende Formel:

$$\mathbb{E}(g(X,Y)) = \sum_{i,j} g(x_i,y_j)\mathbb{P}(X=x_i, Y=y_j).$$

Der Erwartungswert von der einen Zufallsvariablen Y gegen $X=x$ ist gegeben durch

$$\mathbb{E}(Y \mid X=x) = \int_{-\infty}^{\infty} \mathrm{d}y\, y f_{Y|X=x}(y)$$

6.4 Kovarianz und Korrelation

Da die gemeinsame Verteilung von abhängigen Zufallsvariablen i.A. kompliziert ist, begnügt man sich oft mit einer *vereinfachenden* Kennzahl zur Beschreibung der Abhängigkeit. Die *Kovarianz von X und Y* sowie *Korrelation von X und Y* sind wie folgt definiert:

$$\begin{aligned}\mathrm{Cov}(X,Y) &\coloneqq \mathbb{E}\big((X-\mu_X)(Y-\mu_Y)\big)\,,\\ \mathrm{Corr}(X,Y) &\coloneqq \rho_{XY}\\ &\coloneqq \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}\end{aligned}$$

Unmittelbar aus der Definition folgt sofort

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y),$$

sowie di wichtige Formel

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y),$$

zur praktischen Berechnung der Kovarianz. Weiter ist die Kovarianz *bilinear*, d.h. es gilt

$$\begin{aligned}\mathrm{Cov}&\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right)\\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j\,\mathrm{Cov}(X_i,Y_j)\end{aligned}$$

und symmetrisch, d.h.

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

womit wir nun auch in der Lage sind die Varianz von Linearkombinationen von Zufallsvariablen elegant auszudrücken, es gilt nämlich

$$\begin{aligned}\mathbb{V}\left(\sum_{i=1}^{m} X_i\right) &= \sum_{i=1}^{m}\mathbb{V}(X_i)\\ &\quad + 2\sum_{\substack{i,j=1\\ i<j}}^{m}\mathrm{Cov}(X_i,X_j).\end{aligned}$$

falls die X_i unabhängig (oder noch allgemeiner unkorreliert) sind vereinfacht sich diese Formel zu

$$\mathbb{V}(X_1+\cdots+X_m) = \mathbb{V}(X_1)+\cdots+\mathbb{V}(X_m)$$

$X_1,\ldots,X_n$ *unabhängig*.

$$\begin{aligned}\mathrm{Cov}(a+bX, c+dY) &= bd\,\mathrm{Cov}(X,Y),\\ \mathrm{Corr}(a+bX, c+dY) &= \mathrm{sgn}(b)\,\mathrm{sgn}(d)\\ &\quad\cdot \mathrm{Corr}(X,Y);\\ V(X+Y) &= \mathbb{V}(X)+\mathbb{V}(Y)\\ &\quad + 2\,\mathrm{Cov}(X,Y)\end{aligned}$$

Die Korrelation misst Stärke und Richtung der *linearen Abhängigkeit zwischen X und Y*. Es gilt

$$\mathrm{Corr}(X,Y) = +1 \quad\Leftrightarrow\quad Y = a+bX$$

$\forall a\in\mathbb{R}, b>0$

$$\mathrm{Corr}(X,Y) = -1 \quad\Leftrightarrow\quad Y = a+bX$$

$\forall a\in\mathbb{R}, b<0$.
Überdies gilt

$$X \text{ und } Y \text{ unabhängig} \quad\Rightarrow\quad \mathrm{Corr}(X,Y)=0. \tag{4}$$

Die Umkehrung gilt i.A. nicht. Ein Spezialfall, wo auch die Umkehrung gilt, wird in Kapitel 6.6 diskutiert.

6.5 Linear Prognose

Bei der linearen Prognose von Y gestützt auf X macht man den Ansatz

$$\hat{Y} = a + bX$$

und bestimmt die Koeffizienten so, dass der mittlere quadratische Prognosefehler $\mathbb{E}\left(\left(Y-\hat{Y}\right)^2\right)$ minimal wird. Man kann zeigen dass die Lösung dieses Optimierungsproblems gegeben ist durch

$$\hat{Y} = \mu_Y + \frac{\mathrm{Cov}(X,Y)}{\mathbb{V}(X)}(X-\mu_X)$$

$$\mathbb{E}\left(\left(Y-\hat{Y}\right)^2\right) = \left(1-\rho_{XY}^2\right)\mathbb{V}(Y).$$

6.6 Zwei-dimensionale Normalverteilung

Die wichtigste zweidimensionale Verteilung ist die Normalverteilung mit Erwartungswerten (μ_X,μ_Y) und Kovarianzmatrix Σ, wobei

$$\Sigma = \begin{pmatrix}\mathbb{V}(X) & \mathrm{Cov}(X,Y)\\ \mathrm{Cov}(X,Y) & \mathbb{V}(Y)\end{pmatrix}.$$

Sie hat die Dichte

$$\begin{aligned}f_{X,Y}(x,y) &= \frac{1}{2\pi\sqrt{\det(\Sigma)}}\\ \cdot\exp&\left(-\frac{1}{2}(x-\mu_X, y-\mu_Y)\,\Sigma^{-1}\begin{pmatrix}x-\mu_X\\ y-\mu_Y\end{pmatrix}\right)\end{aligned}$$

Wir sehen von dieser Formel: Wenn

$$\mathrm{Cov}(X,Y)=0$$

wird Σ eine Diagonalmatrix und man kann nachrechnen dass dann die Bedingung Gl. 3 gilt. Das heisst: Im Falle der zwei-dimensionalen Normalverteilung gilt auch die Umkehrung von Gl. 4. Zudem: Die Rand- und bedingten Verteilungen sind wieder (1-dimensional) normal.

6.7 Mehr als 2 Zufallsvariablen

Alle diese Begriffe und Definitionen lassen sich natürlich auf mehr als zwei Zufallsvariablen verallgemeinern. Die Formeln sehen im wesentlichen gleich aus, vor allem wenn man die Sprache der Linearen Algebra verwendet.

Ausblick: Wenn man eine dynamische Grösse während eines Zeitintervalls misst, erhält man einen stochastischen Prozess $\{X(t); t\in[a,b]\}$. Die linearen Abhängigkeiten zwischen den Werten zu verschiedenen Zeitpunkten werden dann durch die sogenannte *Autokovarianzfunktion* beschrieben.

7 Deskriptive Statistik

In der *Statistik* will man aus beobachtetet Daten Schlüsse ziehen. Meist nimmt man an, dass die Daten Realisierungen von Zufallsvariablen sind (siehe Kapitel 8.1), deren Verteilung man aufgrund der Daten bestimmen möchte. Als erste Schritt geht es aber zunächst einmal darum, die vorhandenen Daten übersichtlich darzustellen und zusammenzufassen. Dies ist das Thema der *beschreibenden* oder *deskriptiven Statistik*.

7.1 Kennzahlen

Für die numerische Zusammenfassung von Daten gibt es diverse Kennzahlen. das *arithmetische Mittel* ist

$$\overline{x} = \frac{1}{n}(x_1+\cdots+x_n)$$

als Kennzahl für die Lage der Daten. Die *empirische Standardabweichung* ist die Wurzel aus der *empirischen Varianz*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\,,$$

als Kennzahl für die Streuung der Daten. (Eine Begründung für den Nenner $n-1$ statt n folgt später.) Um weitere Kennzahlen zu definieren, führen wir die geordneten Werte

$$x_{(1)} \le x_{(2)} \le x_{(n)}$$

ein. Das *empirische α-Quantil* $(0<\alpha<1)$ ist

$$x_{(k)}, k \text{ die kleinste ganze Zahl} > \alpha n.$$

Wenn αn eine ganze Zahl ist, nimmt man $\frac{1}{2}\left(x_{(\alpha n)}+x_{(\alpha n+1)}\right)$. Der *empirische Median* ist ads 50%-Quantil und ist eine Kennzahl für die Lage. Die Quartilsdifferenz ist das empirische 75%-QUantil minus empirisches 25%-Quantil und ist eine Kennzahl für die Streuung.

Einen ganz anderen Aspekt erfasst man, wenn man die werte gegen den Beobachtungszeitpunkt aufträgt. Damit kann man Trends und andere Arten von systematischen Veränderungen in der Zeit erkennen.

7.2 Histogramm und Boxplot

Wenn man n Werte $x_1,\ldots,x_n$ einer Variablen hat, dann gibt es als grafische Darstellungen das *Histogramm*, den *Boxplot* und die empirische *kumulative Verteilungsfunktion*.

Beim Histogramm bilden wir Klassen (c_{k-1},c_k) und berechnen die Häufigkeiten h_k =#Werte in diesem Intervall. Dann trägt man über den Klassen Balken auf, deren Höhe *proportional* ist zu $h_k/(c_k-c_{k-1})$ ist.

Beim Boxplot hat man ein Rechteck, das von 25%- und vom 75%-Quantil begrenzt ist, und Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten „normalen "Wert gehen (per Definition ist ein normaler wert höchstens 1.5 mal die QUartilsdifferenz von einem der beiden Quartile). Zusätzlich gibt man noch Ausreisser durch Sterne und den Median durch einen Strich an. Der Boxplot ist vor allem dann geeignet, wenn man die Verteilungen einer Variablen

in verschiedenen Gruppen (die im allgemeinen verschiedenen Versuchsbedingungen entsprechen) vergleichen will. springt. Für eine glattere Version verbindet man die Punkte $(x_{(i)}, (i-0.5)/n)$ Die empirische Verteilungsfunktion ist eine Treppenfunktion, die an den Stellen $x_{(i)}$ von $(i-1)/n$ auf i/n durch Strecken.

7.3 Normal- und QQ-Plot

Der Normal- und QQ-Plot („Quantil-Quantil-Plot“) sind of viel geeignetere grafische Mittel als die empirische kumulative Verteilungsfunktion.

Die empirische Quantile für $\alpha_k = (k-0.5)/n$ sind gerade die geordneten Beobachtungen $x_{(k)}$ $(x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)})$. Der QQ-Plot trägt die Punkte $F^{-1}(\alpha_k)$ (die theoretischen Quantile einer kumulativen Verteilung F) gegen $x_{(k)}$ (die empirischen Quantile) auf. Wir interpretieren die Daten $x_1, \ldots, x_n$ als Realisierungen von $X_1, \ldots, X_n$ i.i.d. $\sim \tilde{F}$, siehe 8.1. Falls nun die wahre Verteilung $\tilde{F}$ mit der gewählten Verteilung F in QQ-Plot übereinstimmt, so liefert der QQ-Plot approximativ eine Gerade durch Null mit Steigung 45 Grad. Man kann also Abweichungen der Daten von einer gewählten Modell-Verteilung so grafisch überprüfen.

Der Normal-Plot ist ein QQ-Plot wo die Modell-Verteilung F die Standard-Normalverteilung $\mathcal{N}(0,1)$ ist. Es gilt dann das folgende: Wenn die wahre Verteilung $\tilde{F}$ eine Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ ist, so liefert der Normal-Plot approximativ eine Gerade, welche jedoch im allgemeinen nicht durch Null und nicht Steigung 45 Grad hat. Das heisste: Der Normal-Plot liefert eine gute Überprüfung für irgendeine Normalverteilung, auch wenn die Model-Verteilung als Standard-Normal gewählt wird.

In Normal- und QQ-Plot kann man insbesondere sehen, ob eine Transformation der Daten angebracht ist, oder ob es Ausreisser gibt, die man besonders behandeln sollte.

Im Grunde genommen ist der QQ-Plot bereits mehr als bloss deskriptive Statistik: Es ist ein grafisches Werkzeug um eventuelle Abweichungen von einem Modell festzustellen: Vergleiche mit dem Formalismus des statistischen Tests in Kapitel 8.3.

8 Schliessende Statistik

8.1 Daten als Realisierungen von Zufallsvariablen

In der schliessenden Statistik wollen wir anhand von Daten (Beobachtungen) Aussagen über ein Wahrscheinlichkeitsmodell machen. Dass man dies tun kann ist zunächst erstaunlich: Man benützt die induktive Logik um probilistische Aussagen (d.h. Aussagen, welche mit typischerweise hoher Wahrscheinlichkeit gelten) zu machen.

Grundlegend für die schliessende statistik ist die Annahme, dass Daten Realisierungen von Zufallsvariablen sind. Das heisst: eine Beobachtung (oder „Messung“) x ist entstanden in dem ein $\omega \in \Omega$ zufällig gezogen wurde, so dass die Zufallsvariable X den Wert $X(\omega) = x$ annimmt. Bei mehreren Daten geht alles analog: n Beobachtungen $x_1, \ldots, x_n$ werden aufgefasst als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$, welche die Werte $X_i = x_i$ mit $i = 1, \ldots, n$ angenammon haben.

8.2 Erste Konzepte

Wir betrachten folgende Situation. Gegeben ist eine Beobachtung x (eine Realisierung) einer $\mathrm{Bin}(n,p)-$ oder einer $\mathrm{Poi}(\lambda)$-verteilten Zufallsvariablen X: z.B. Anzahl Ausfälle bei n Wiederholungen oder während eine Beobachtungsdauer t, wobei dann $\lambda = \mu t$ und μ die erwartete Anzahl Ausfälle pro Zeiteinheit ist. Wir möchten daraus Rückschlüsse auf den unbekannten Parameter p bzw. λ ziehen. Genauer geht es um folgende drei Fragestellungen:

1. Welchen ist der plausibelste Wert des unbekannten Parameters (*Punktschätzung*)?
2. Ist ein bestimmter vorgegebener Parameterwert p_0, bzw. λ_0 (z.B. ein Sollwert) mit der Beobachtung verträglich (*Test*)?
3. was ist der Bereich von plausiblen parameterwerten (*Vertrauensintervall*)?

8.3 Das Testproblem

Beim Testproblem beschränken wir uns hier zur Vereinfachung der Notation auf die Binomialverteilung. Wir nehmen an wir haben eine beobachtung $x \in \{0, \ldots, n\}$, die wir als Realisierung einer $\mathrm{Bin}(n,p)$-verteilten Zufallsvariable X interpretieren, wobei n fix und bekannt ist, und p der unbekannte Parameter, den wir mit einem (von der Problemstellung abhängigen) Wert p_0 vergleichen (testen) wollen, genauer gesagt wollen wir überprüfen ob die Beobachtung x damit verträglich ist. Die Annahme $p = p_0$ wird *Nullhypothese* genannt, notiert durch

$$H_0 : p = p_0,$$

die entsprechende Vermutung wird dann *Alternativhypothese* genannt, notiert durch

$$\begin{aligned} H_A : p \neq p_0, &\quad \text{zweiseitig;} \\ p > p_0, &\quad \text{einseitig nach oben, rechtseitig;} \\ p < p_0, &\quad \text{einseitig nach unten, linksseitig.} \end{aligned}$$

Wir beschränken uns im Weiteren auf den Fall $H_A : p > p_0$. Dann lehnen wir die Nullhypothese $H_0 : p = p_0$ ab, falls $x \geq c$. Das ist qualitativ betrachtet plausibel:Die quantitative Wahl von c wird wie folgt gemacht. Wir nehem einmal an, dass die Nullhypothese stimmt. Dann ist die Wahrscheinlichkeit die Nullhypothees fälschlicherweise abzulehnen (d.h. ein *Fehler 1. Art*)

$$\mathbb{P}_{p_0}(X \geq c) = \sum_{k=c}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}.$$

Wir sollten also c nicht zu klein wählen. Umgekehrt möchten wir aber auch c nicht zu gross wählen, weil wir sonst zu häufig einen *Fehler 2. Art* begehen: Kein Verwerfen der Nullhypothese H_0, obwohl sie falsch ist. Man schliesst einen Kompromiss, indem man das kleinste $c = c(\alpha)$ nimmt, so dass

$$\mathbb{P}_{p_0}(X \geq c) \leq \alpha$$

dabei ist α eine im voraus festgelegte (kleine) Zahl, das sogenannte *Signifikanzniveau*. Obige (Un-)Gleichung besagt, dass die Wahrscheinlichkeit eines Fehlers 1.Art mit dem Signifikanzniveau α kontrolliert ist. Die Wahrscheinlichkeit für einen Fehler 2. Art ist nicht explizit kontrolliert, deswegen, weil man nur einen - und hier wählt man den schlimmeren Fehler 1. Art - direkt kontrollieren kann. Nach all diesen Überlegungen kommt man zum Rezept, dass H_0 verworfen wird, falls $x \geq c_\alpha$.

Im Fall, wo man nach Abweichungen nach unten interessiert ist, d.h. $H_A : p < p_0$, geht alles analog, Bei zwei-seitiger Alternative $H_A : p \neq p_0$, verwerfen wir die Nullhypothese $H_0 : p = p_0$, wenn $x \leq c_1$ oder $x \geq c_2$. Hier wählt man c_1 möglichst gross und c_2 möglichst klein unter den Einschränkungen dass

$$\sum_{k=0}^{c_1} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2}$$

$$\sum_{k=c_2}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2}$$

8.3.1 Zusammenfassung eines statistischen Tests

Die Durchführung eines statistischen Tests kann, zumindest teilweise, „rezeptartig“erfolgen.

1. Lege Nullhypothese $H_0 : \theta = \theta_0$ fest. (θ bezeichnet hier allgemein einen Parameter in einem wahrscheinlichkeitstheoretischen Modell).
2. Anhand der Problemstellung, spezifiziere vernünftige Alternative $H_A : \theta \neq \theta_0$ (zweiseitig) oder $H_A : \theta > \theta_0$ (einseitig nach unten).
3. Wähle Signifikanzniveau α, z.B. $\alpha = 0.05$ oder 0.01.
4. Konstruiere Verwerfungsbereich für H_0, so dass
$$\mathbb{P}_{\theta_0}(\text{Fehler 1. Art}) \leq \alpha.$$
5. Erst jetzt: Betrachte ob die Beobachtung x (oder eine Funktion von mehreren Beobachtungen) in den Verwerfungsbereich fällt: Falls ja, so verwerfe H_0 (die Alternative ist dann „signifikant“). Falls x nicht in den Verwerfungsbereich fällt, so belassen wir H_0 (was noch lange nicht heisst, dass deswegen H_0 statistisch bewiesen ist).

8.4 P-Wert

Viele Computer-Pakete liefern obigen Punkt 4 insofern, dass der sogennante P-Wert gegeben wird.

Falls das Signifikanzniveau α kleiner gewählt wird, so wird der Verwerfungsbereich kleiner. Eine intuititive Begründung dafür ist: kleines

$$\begin{aligned} \alpha &\gtrapprox \mathbb{P}(\text{Fehler 1. Art}) \\ &= \mathbb{P}(\text{fälschliches Verwerfen} \\ &\qquad \text{von } H_0 \text{ obschon } H_0 \text{ stimmt}) \end{aligned}$$

bedeutet, dass es (generell) schwieriger wird H_0 zu verwerfen, oder anders ausgedrückt: Der Verwerdungsbereich (für Verwerfen von H_0) wird kleiner. Dies illustriert auch die Tatsache, dass amn bei extrem klein gewähltem α die Null-Hypothese (fast) nie verwerfen kann.

Es gibt also ein Niveau, wo H_0 „gerade noch“ verworfen wird. Der *P-Wert ist das kleinste Signifikanzniveau wo H_0 verworfen wird*. Die Beobachtung X liegt dann gerade auf der Grenze des Verwerfungsbereichs. Man entscheidet dann in obigem Punkt 5 so, adss H_0 verworfen wird, falls der P-Wert kleiner als α ist.

8.5 Vertrauensintervalle

Ein Vertrauensintervall I zum Niveau $1-\alpha$ (oft auch *Konfidenzintervall genannt*) besteht aus allen Parameterwerten, die im Sinne eines statistischen Tests zum Signifikanzniveau α mit der Beobachtung verträglich sind (üblicherweise nimmt man den zweiseitigen Test). Mathematisch heisst das:

$$I = \{\theta_0; \text{ Nullhypothese } H_0 : \theta = \theta_0 \text{ wird belassen}\}$$

Diese Beziehung stellt eine Dualität zwischen Tests und Vertrauensintervall dar.

Die Berechnung kann grafisch, oder mit einer Tabelle, oder basierend auf der Normalapproximation erfolgen. Letztere ergibt

$$\frac{x}{n} \pm z_{1-\alpha/2}\sqrt{\frac{x}{n}\left(1-\frac{x}{n}\right)\frac{1}{n}}$$

ist Vertrauensintervall für p, falls $X \sim \mathrm{Bin}(n,p)$,

$$x \pm z_{1-\alpha/2}\sqrt{x}$$

ist Vertrauensintervall für λ, falls $X \sim \mathrm{Poi}(\lambda)$. Das Vertrauensintervall ist zufällig: Es fängt den unbekannten wahren Parameter mit Wahrscheinlichkeit $1-\alpha$ ein.

8.5.1 Begründung von Formel für Vertrauensintervall bei Binomial-Verteilung

Betrachte $X \sim \mathrm{Bin}(n,p)$, das heisst

$$X = \sum_{i=1}^{n} Y_i, \qquad Y_1,\dots,Y_n \text{ i.i.d.} \sim \text{Bernoulli}(p)$$

Der Zentrale Grenzwertsatz liefert dann

$$\begin{aligned}\hat{p} &= \frac{X}{n} \\ &= \overline{Y}_n \approx \mathcal{N}\left(\mathbb{E}(Y_1), \frac{1}{n}\mathbb{V}(Y_1)\right) \\ &= \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)\end{aligned}$$

Ansatz: konfidenzintervall

$$I = \left[\text{“}p\right] - c_u, \hat{p} + c_o$$

, so dass $\mathbb{P}_p(p \in I) \geq 1-\alpha$. Man rechnet jetzt

$$\begin{aligned}\mathbb{P}_p(p \in I) &= \mathbb{P}(\hat{p} - c_u \leq p \leq \hat{p} + c_o) \\ &= \mathbb{P}(p - c_o \leq \hat{p} \leq p + c_u) \\ &= \mathbb{P}(-c_o \leq \hat{p} - p \leq c_u) \\ &\overset{\spadesuit}{=} \mathbb{P}\left(-\frac{c_o\sqrt{n}}{\sqrt{p(1-p)}} \leq \underbrace{\frac{(\hat{p}-p)\sqrt{n}}{\sqrt{p(1-p)}}}_{\approx\mathcal{N}(0,1)}\right)\end{aligned}$$

♠: standardisieren. Da $\mathcal{N}(0,1)$ eine symmetrische Verteilung ist, wähle

$$\begin{aligned}\text{oberes Quantil} &= \frac{c_u\sqrt{n}}{\sqrt{p(1-p)}} \\ &\approx z_{1-\alpha/2} \\ \text{unteres Quantil} &= -\frac{c_o\sqrt{n}}{\sqrt{p(1-p)}} \\ &= z_{\alpha/2}\end{aligned}$$

daraus folgt

$$c_u \approx z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}},$$

da p unbekannt ist, ersetzen wir es durch den Schätzer $\hat{p} = X/n$

$$c_u \approx z_{1-\alpha/2}\sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)\frac{1}{n}}$$

und wegen Symmetrie $c_o = c_u$ (oder analoge Heirleitung für c_o). Damit folgt die approximative Formel.

8.6 Mehrere Beobachtungen

Wenn man n Beobachtungen hat, geht man oft zu den Summen über: $x_1 + \cdots + x_n$. Für die zugehörigen Summen von Zufallsvariablen, welche als i.i.d. angenommen werden, kennen wir dann die Verteilung der Summen approximativ (ZGS) oder auch in einigen Fällen exakt. Damit lassen sich Verwerfungsbereiche und konfidenzintervalle konstruieren, analog zu Kapitel 8.5.1.

9 Statistik bei normalverteilten Daten

Wir betrachten folgende Situation. Gegeben sind n Beobachtungen (Realisierungen) $x_1,\dots,x_n$ von Zufallsvariablen $X_1,\dots,X_n$ i.i.d. $\sim \mathcal{N}(\mu,\sigma^2)$. Typischerweise sind dies n Messungen einer unbekannten Grösse μ, und σ gibt an, wie genau die Messungen sind. Die Annahme der Normalverteilung wird meist mit dem Zentralen Grenzwertsatz begründet. Ausserdem nehmen wir noch an, dass es keine Beeinflussungen zwischen den einzelnen beobachtungen gibt, so dass die Unabhängigkeit der Zufallsvariablen gerechtfertigt erscheint.

Wir möchten aus $x_1,\dots,x_n$ Rückschlüsse auf die unbekannten Parameter μ und σ ziehen. Wie zuvor geht es um die drei Fragestellungen Punktschätzung, Test für einen vorgegebenen Wert (Sollwert) und Vertrauensintervall (Bereich von plausiblen Werten). Weil μ meist von grösserem Interesse ist als σ, behandeln wir die letzten beiden fragestellungen nur für μ.

9.1 Schätzungen

Die Punktschätzungen sind:

$$\hat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} \tag{5}$$

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \tag{6}$$

Diese Schätzungen sind Funktionen von Zufallsvariablen, also wieder zufällig und im allgemeinen verschieden vom unbekannten wahren Wert. Die realisierten Werte dieser Schätzung (den Wert den man berechnen kann) erhält man indem die Zufallsvariable X_i mit deren realisiertem Wert x_i ersetzt wird.

Der Erwartungswert der Schätzer ist

$$\mathbb{E}(\hat{\mu}) = \mu,$$
$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2.$$

(Dies ist der Grund für den Nenner $n-1$).

9.2 Testen

Wir haben n voneinander unabhängige Beobachtungen $x_1,\dots,x_n$ einer Zufallsvariable $X \sim \mathcal{N}(\mu,\sigma^{)})$, interpretiert als eine Beobachtung der i.i.d. Zufallsvariablen $X_1,\dots,X_n$ mit der selben Verteilung wie X. Wir fixieren je nach Problemstellung ein $\mu_0 \in \mathbb{R}$ und wollen die Nullhypothese $H_0 : \mu = \mu_0$ gegen eine der möglichen Alternativen $H_A : \mu \neq \mu_0$, $H_A : \mu < \mu_0$ bzw. $H_A : \mu > \mu_0$ testen. Dabei unterscheiden wir zwei Fälle: Die Streuung σ ist bekannt, dann verwenden wir den sogenannten z-Test, oder die Streuung σ ist unbekannt. (dann muss sie aus den beobachtetetn Daten geschätzt werden) , in diesem fall ergibt sich der sogenannte t-test.

9.2.1 z-Test (σ bekannt)

Wir nehmen an, dass σ bekannt ist. Die Teststatastik ist definiert als die Zufallsvariable

$$T \coloneqq \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma}$$

wobei $\overline{X}_n$ den Schätzer für den Mittelwert aus Gl. 6 darstellt. Wir lehnen für eine gegebene Realisierung $t = \sqrt{n}(\overline{x}_n - \mu_0)/\sigma$ von T je nach Alternative H_A die Nullhypothese $H_0 : \mu = \mu_0$ ab,

$$\begin{aligned}&|t| \geq z_{1-\alpha/2}, \\ &\Leftrightarrow t \in VB_\alpha = (-\infty, z_{\alpha/2}] \cup [z_{1-\alpha/2}, \infty), \\ &H_A : \mu \neq \mu_0.\end{aligned} \tag{7}$$

$$\begin{aligned}&t \geq z_{1-\alpha}, \Leftrightarrow t \in [z_{1-\alpha}, \infty), \\ &H_A : \mu > \mu_0.\end{aligned} \tag{8}$$

$$\begin{aligned}&t \leq z_{\alpha}, \Leftrightarrow t \in (-\infty, z_{\alpha}], \\ &H_A : \mu < \mu_0.\end{aligned} \tag{9}$$

Die Begründung für 15 ist wie folgt, die Teststatistik T ist unter der Nullhypothese $\mathcal{N}(0,1)$-verteilt, woraus sich der Fehler 1. Art, sofort zu

$$\mathbb{P}_{\mu=\mu_0}(|T| \geq z_{1-\alpha/2})$$

ergibt, also genau wie es sein sollte. Das $(1-?\,o)$-Vertrauensintervall für den Parameter μ ist gegeben durch

$$I_{1-\alpha} = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\right],$$

da es von den $X_1,\dots,X_n$ abhängt ist es zufällig, für eine gegebene Beobachtung $\overline{x}_n$ wird dann $\overline{X}_n$ einfach durch $\overline{x}_n$ ersetzt um die Realisierung des Vertrauensintervalls zur Stichprobe $x_1,\dots,x_n$ zu erhalten. Dass Vertrauensintervall ist so definiert, dass es den wahren Wert von μ mit Wahrscheinlichkeit $1-\alpha$ „einfängt“: Eine kurze Umformung ergibt, dass das Ereignis

$$\{\mu \in I\} = \left\{\omega \in \Omega \mid \mu \in I(\omega)\right\}$$

gegeben ist durch

$$\{\mu \in I\} = \left\{\left|\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}\right| \leq z_{1-\alpha/2}\right\}$$

und wegen

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

ergibt sich sofort

$$\begin{aligned}\mathbb{P}(\mu \in I) &= \mathbb{P}\left(\left|\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}\right| \le z_{1-\alpha/2}\right)\\ &= 1-\alpha\end{aligned}$$

Der P-Wert zu einer gegebenen Realisierung t der Teststatistik T errechnet sich im zweiseitigen Fall zu

$$\begin{aligned}\mathbb{P}_{\mu=\mu_0}(|T| \ge |t|) &= 2\mathbb{P}_{\mu=\mu_0}\\ &= 2\left(1-\Phi(|t|)\right)\end{aligned}$$

9.2.2 t-Test (σ unbekannt)

Kenntnis von σ, welche für den z-Test benötigt wird, ist in der Praxis meist unrealistisch. Wenn wir σ nicht kennen, ersetzen wir es durch den Schätzer S_n aus Gleichung 6. Die Teststatistik ist dass

$$T := \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{S_n}$$

im Gegensatz zur Teststatistik beim z-Test ist sie hier t-verteilt mit Freiheitsgrad $n-1$, das α-Quantil der t-Verteilung mit Freiheitsgrad n wird mit $t_{n,\alpha}$ bezeichnet. Wir lehnen für eine gegebene Realisierung

$$t = \sqrt{n}\,\frac{\overline{x}_n - \mu}{s_n}$$

von T je nach Alternative H_A die Nullhypothese $H_0 : \mu = \mu_0$ ab, falls

$$\begin{aligned}&|t| \ge t_{n-1,1-\alpha/2},\\ &\Leftrightarrow t \in VB_\alpha = (-\infty, t_{n-1,\alpha/2}] \cup [t_{n-1,1-\alpha/2}, \infty),\\ &H_A : \mu \ne \mu_0.\end{aligned} \tag{10}$$

$$\begin{aligned}&t \ge t_{n-1,1-\alpha}, \Leftrightarrow t \in [t_{n-1,1-\alpha}, \infty),\\ &H_A : \mu > \mu_0.\end{aligned} \tag{11}$$

$$\begin{aligned}&t \le t_{n-1,\alpha}, \Leftrightarrow t \in (-\infty, t_{n-1,\alpha}],\\ &H_A : \mu < \mu_0.\end{aligned} \tag{12}$$

Das $(-\alpha)$)-Vertrauensintervall ist gegeben durch

$$I_{1-\alpha} = \left[\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,1-\alpha/2}, \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,1-\alpha/2}\right].$$

Ausblick auf eine verwandte Methode in der Praxis: In der *statistischen Qualitätskontrolle* wird in regelmässigen Abständen eine kleine Stichprobe vom Umfang n aus dem Produktionsprozess gezogen, die Zielgrösse gemessen, gemittelt und gegen die Zeit aufgetragen. Fällt ein Mittelwert ausserhalb der Kontrollgrenzen „Sollwert $\pm 3\sigma/\sqrt{n}$“ oder sind 9 aufeinandererfolgende Mittelwerte alle grösser oder alle kleiner als der Sollwert, dann ist der Produktionsprozess ausser Kontrolle.

9.2.3 Macht eines Tetst

Ein statistischer Test kontrolliert direkt die Wahrscheinlichkeit eines Fehlers 1. Art via dem Signifikanzniveau α:

$$\begin{aligned}\mathbb{P}(\text{Fehler 1. Art}) = \mathbb{P}(&\text{Test verwirft } H_0 \text{ obschon}\\ &H_0 \text{ stimmt })\\ \lessapprox\ &\alpha.\end{aligned}$$

Die Wahrscheinlichkeit eines Fehlers 2. Art ist eine Funktion des Parameterwerts $\mu \in H_A$:

$$\begin{aligned}\beta(\mu) = \mathbb{P}(&\text{Test akzeptiert } H_0\\ &\text{obschon ein } \mu \in H_A \text{ stimmt}).\end{aligned}$$

Die Macht eines Tests ist definiert als

$$\begin{aligned}1-\beta(\mu) = \mathbb{P}(&\text{Test verwirft richtigerweise}\\ &H_0 \text{ für ein } \mu \in H_A).\end{aligned}$$

Die Macht (englisch power) eines Tests beschreibt die Kapazität wie gut ein Test einen Parameter im bereich der Alternative richtigerweise entdecken kann. Die Macht kann deshalb als Gütemass gebraucht werden, um optimale Tests zu charakterisieren.

Man kann zeigen, dass der t-Test der optimale Test (bzgl. der Macht) unter allen möglichen Tests ist, *falls* die Beobachtungen normalverteilt sind. Bei nichtnormalverteilten Beobachtungen können andere Tests (siehe Kaptiel 11.2) sehr viel besser sein als der t-Test.

10 Punktschätzungen

Wir betrachten folgendes Problem: Gegeben sei wieder eine Zufallsvariable X von der wir zwar die Art der Vertilung kennen, aber nicht den dazugehörigen Parameter (Vektor) θ. basierend auf n unabhängigen Beobachtungen $x_1, \dots, x_n$ von X, die wir wie immer als eine realisierung der i.i.d. Zufallsvariablen $X_1, \dots, X_n$ (mit derselben Verteilung wie X) interpretieren soll nun der unbekannte Parameter θ geschätzt werden. Sei θ ein reeller Parameter einer Wahrscheinlichkeitsverteilung auf $\mathbb{R}$, ein *Schätzer für* θ (zur Stichprobengrösse n) ist eine Funktion $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}$ von den $\mathbb{R}^n$

$$\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$$

oder als Zufallsvariable interpretiert

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n).$$

Man beachte dass bei der Verwendung von griechischen buchstaben keine Unterscheidung mehr zwischen Realisierung (Kleinbuchstaben) und Zufallsvariable (Grossbuchstaben) gemacht wird.

10.1 Momentenmethode

Das *k-te Moment von* X ist definiert durch

$$\mu_k := \mathbb{E}(X^k),$$

das erste Moment entspricht also dem Erwartungswert. Die Momentenmethode nimmt an, dass wir den unbekannten Parametervetor $(\theta_1, \dots, \theta_r)$ durch die ersten p Momente von X ausdrücken können, d.h. allgemein

$$\theta_j = g_j(\mu_1, \dots, \mu_p), \qquad j = 1, \dots, r, \tag{13}$$

die j-te Komponente von θ ist also gegeben durch die Funktion $g_j : \mathbb{R}^p \to \mathbb{R}$.

Ausgehend von Gleichung 13 werden nun die wahren μ_k ersetzt durch deren Schätzungen

$$\hat{\mu}_k := \frac{1}{n}\sum_{i=1}^n x_i^k,$$

wenn man diese noch als Zufallsvariable schreibt, dann erhält man einen Momentenschätzer für die einzelnen Komponenten von θ durch

$$\begin{aligned}\hat{\theta}_j &= g_j(\hat{\mu}_(, \dots, \hat{\mu}_p), \qquad j = 1, \dots, r;\\ \hat{\mu}_k &= \frac{1}{n}\sum_{i=1}^n X_i^k, \qquad k = 1, \dots, p.\end{aligned}$$

Rezeptatrige Vorgagnsweise: Gegeben ist eine Zufallsvariable X deren Verteilung von einem unbekannten (zu schätzenden) Parameter θ abhängt. Man berechnet zunächst das erste Moment von X, wenn dieses von θ abhängt, dann löst man die Gleichung nach θ auf, d.h. man schreibt θ als Funktion des ersten Momentes, dann wird das erste Moment durch dessen Schätzer ersetzt und man erhält einen Momentenschätzer für θ. Falls das erste Moment nicht von θ abhängt berechnet man das zweite Moment usw. Es ist aber auch möglich, dass sämtliche Momente von θ abhängen, man erhält dann mehrere (unterschiedliche) Momentenschätzer für θ.

Bsp.
Sei $X \sim \mathrm{Poi}(\lambda)$ *mit unbekanntem Parameter* λ. *Das erste Moment von* X *ist*

$$\mathbb{E}(X) = \lambda,$$

oder anders geschrieben,

$$\lambda = \mu_1,$$

nun wird μ_1 *durch dessen Schätzung*

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n x_i$$

ersetzt, als Zufallsvariable geschrieben erhält man einen Momentenschätzer für λ *durch*

$$\hat{\lambda} = \hat{\mu}_1 = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

Einen weiteren Momentenschätzer fur λ erhalten wir durch Berechnung des zweiten Momentes von

$$X : \mu_2 = \mathbb{E}(X^2) = \mathbb{V}(X) + \mathbb{E}(X)^2 = \lambda + \mu_1^2$$

woraus nun

$$\lambda = \mu_2 - \mu_1^2$$

folgt und wir erhalten den Momentenschätzer

$$\begin{aligned}\hat{\lambda} &= \hat{\mu}_2 - \hat{\mu}_1^2\\ &= \frac{1}{n}\sum_{i=1}^n X_i^2\\ &= \frac{n-1}{n} S_n^2\end{aligned}$$

Man zieht aber $\hat{\lambda} = \overline{X}_n$ vor, denn dies ist auch der sogenannte Maximum-Likelihood Schätzer, welcher im Allgemeinen genauer ist.

Der Momentenschätzer ist einfach, aber nicht immer die optimale (im Sinne einer zu definierenden besten Genauigkeit für den unbekannten Parameter) Methode. Überdies ist der Momentenschätzer nicht eindeutig, wie wir am obigen Beispiel gesehen haben.

10.2 Maximum-Likelihood Schätzer

Sei zunächst X diskret. Um die Abhängigkeit vom unbekannten Parameter θ zu betonen bezeichnen wir die Wahrscheinlichkeitsfunktion p_X von X mit p_θ. Die Wahrscheinlichkeit, dass tatsächlich das Ereignis $\{X_1 = x_1, \ldots, X_n = x_n\}$ eintritt ist wegen der Unabhängigkeit und Gleichverteilung gegeben durch

$$\begin{aligned} L(\theta) &\coloneqq p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \\ &= \prod_{i=1}^{n} p_\theta(x_i) \\ &= p_\theta(x_1)\cdot\cdots\cdot p_\theta(x_n), \end{aligned}$$

dies ist die sogenannte *Likelihoodfunktion (zur gegebenen Stichprobe $x_1,\ldots,x_n$)*. Die *Maximum-Likelihood Methode* basiert nun darauf diese Wahrscheinlichkeit zu maximieren, also jenen Parameter θ zu finden für den die Wahrscheinlichkeit dass die gegebene Stichprobe $x_1,\ldots,x_n$ eintritt am grössten (maximal) ist, daher der Name Maximum-Likelihood.

Da der Logarithmus monoton wachsend ist, kann man äquivalent zu obiger Maximierungsaufgabe auch den Logarithmus maximieren, was meist (aber nicht unbedingt immer!) einfacher ist. Die log-Likelihoodfunktion ist definiert durch

$$\begin{aligned} l(\theta) &\coloneqq \log L(\theta) \\ &= \sum_{i=1}^{n} \log p_\theta(x_i) \\ &= \log p_\theta(x_1),\ldots,\log p_\theta(x_n). \end{aligned}$$

Die Maximierungsaufgabe lößt man wie aus der Analysis bekannt durch Ableiten (nach dem Parameter θ) und Nullsetzen. Um die Abhängigkeit von der Stichprobe $x_1,\ldots,x_n$ zu betonen schreibt man auch

$$l(\theta; x_(,\ldots,x_n) \coloneqq \log p_\theta(x_1) + \cdots + \log p_\theta(x_n),$$

man muss dann die Gleichung

$$\frac{\partial}{\partial\theta} l(\theta; x_1,\ldots,x_n) = 0$$

nach θ auflösen und erhält das Ergebnis

$$\hat\theta = \hat\theta(x_1,\ldots,x_n)$$

bzw. als Zufallsvariable ausgedrückt

$$\hat\theta = \hat\theta(X_1,\ldots,X_n)$$

als allgemeinen Maximum-Likelihood Schätzer von θ (zum Stichprobenumfang n).

Im stetigen Fall geht im Wesentlichen alles analog und man braucht nur den Buchstaben p durch f und „Wahrscheinlichkeit“ durch „Wahrscheinlichkeitsdichte“ zu ersetzen, d.h. statt Wahrscheinlichkeiten hat man dann Dichten, und es wird jener Parameter θ gesucht für den die gemeinseme Dichte der $X_1,\ldots,X_n$ an der Stelle $x_1,\ldots,x_n$ am grössten ist.

Bsp. (Fortsetzung)
$X \sim \mathrm{Poi}(\lambda)$. *Die Wahrscheinlichkeitsfunktion ist*

$$p_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x \in \mathbb{N},$$

die log-Likelihoodfunktion ist somit

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} (x_i \log(\lambda) - \log(x_i!) - \lambda) \\ &= \sum_{i=1}^{n} (x_i \log(\lambda) - \lambda) - C, \\ C &= \sum_{i=1}^{n} \log(x_i!) \end{aligned}$$

Beacht, das die Konstante C keinen Einfluss auf die Maximierung von $\ell(\lambda)$ hat. Leitet man $\ell(\lambda)$ ab und setzt $\ell'(\lambda) = 0$, so erhält man den Maximum-Likelihood Schätzer

$$\begin{aligned} \hat\lambda &= \overline{X}_n \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i. \end{aligned}$$

Dies ist derselbe Schätzer wie bei der Momentenmethode wo g_1 die Identität ist.

10.3 Eigenschaften von Schätzern

Wir haben gesehen, dass ein Schätzer für den Parameter θ als eine aus den Zufallsvariablen $X_1,\ldots,X_n$ zusammengesetzte Zufallsvariable interpretiert werden kann, also $\hat\theta = \hat\theta(X_1,\ldots,X_n)$ wobei der zu schätzende Parameter $\wedge$ meist eine Kenngrößse der Verteilung der X_i ist. Man kann sich nun fragen wie eigentlich der Erwartungsswert dieser Zufallsvariable (also des Schätzers) aussieht, intuitiv wünschenswert wäre es wann dieser am besten gleich dem zu schätzenden Parameter entspricht, woraus sich folgende Eigenschaft ergibt: Ein Schätzer $\hat\theta$ für den Parameter θ heisst *erwartungstreu* wenn

$$\mathbb{E}(\hat\theta) = \mathbb{E}\left(\hat\theta(X_1,\ldots,X_n)\right) = \theta.$$

Im Allgemeinen kann der Erwartungswert aber auch von der Stichprobengrösse n abhängen, woraus sich noch weiter Eigenschaft ergeben: z.B. heisst ein Schätzer *asymptotisch erwartungstreu*, wenn sein Erwartungswert für wachsende Stichprobengrösse n gegen den zu schätzenden Paremeter konvergiert.

11 Vergleich zweier Stichproben

Wichtige Anwendungen der Statistik liegen im Vergleich verschiedener Versuchsbedingungen, oder allgemeiner bei der Bestimmung der Auswirkung verschiedener erklärender Variablen auf eine Zielgrösse. Als einfachsten Fall behandeln wir jetzt den Vergleich zweier Methoden (Gruppen, Versuchsbedingungen, Behandlungen) hinschtlich des Erwartungswertes.

11.1 Geparrte und ungeparrte Stichproben

In allen Anwendungen ist neben der Auswertung auch die korrekte Planung des Versuches wichtig. Man muss sicherstellen, dass eventuelle Unterschiede tatsächlich durch die verschiedenen Methoden und nicht durch eine andere Störgrösse verursacht sind. Die beiden wichtigtsten Prinzipien dazu sind *Blockbildung* und *Randomisierung* .

Randomisierung bedeutet hier, dass man die Reihenfolge der Versuche und die Zuordnung von Versuchseinheit zu Versuchsbedingung zufällig wählt: Man hat den Beobachtungen (realisierte Zufallsvariablen)

$x_1, x_2, \ldots, x_n$ unter Versuchsbedingung 1,
$y_1, y_2, \ldots, y_n$ unter Versuchsbedingung 2.

Im Allgemeinen ist $m \neq n$, aber nicht notwendigerweise. Bei solch zufälliger Zuordnung von verschiedenen Versuchseinheiten zu zwei verschiedenen Versuschsbedingungen spricht man von einer *ungepaarten Stichprobe*.

Bsp.
zufällige Zuordnung von 100 Testpatienten zu Gruppe der Grösse 60 mit Medikamenten-Behandlung und zu anderer Gruppe der Grösse 40 mit Placebo-Behandlung

Andererseits liegt eine *gepaarte Stichprobe* vor, wenn beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt werden:

$x_1,\ldots,x_n$ unter Versuchsbedingung 1,
$y_1,\ldots,y_n$ unter Versuchsbedingung 2.

// (Nochmals die gleiche Paare? Kann irgendwie nicht sein…)
Notwendigerweise ist dann: Die Stichprobengrösse n ist für beide Versuchsebedingungen dieselbe.

Bsp.
Vergleich zweier Reifentypen, wo bei jedem Testfahrzeug und jedem Fahrer beide Reifentypen verwendet werden.

11.2 Gepaarte Vergleiche

Bei der Analyse von gepaarten Vergleichen arbeitet man stets mit den Differenzen innerhalb der Paare,

$$u_i = x_i - y_i, \qquad i = 1,\ldots,n$$

welche wir als Realisierungen von i.i.d. Zufallsvariablen $U_1,\ldots,U_n$ auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $\mathbb{E}(U_i) = 0$. Dies kann man formal testen mit der Nullhypothese $H_0 : \mathbb{E}(U_i) = 0$ und mit der zwei-eitigen (oder auch einseitigen) Alternative $H_A : \mathbb{E}(U_i) \neq 0$ Die folgenden Tests bieten sich dazu an:

1. der t-Test, siehe Kapitel 9.2;
2. der sogenannte *Vorzeichen-Test*, falls die Normalverteilung nicht gerechtfertigt scheint: betrachte Anzahl positiver U_i und benütze die Methoden für die Binomialverteilung um die Nullhypothese $H_0 : p = p_0 = 0.5$ zu testen, siehe Kapitel 8.3;
3. der sogenannte *Wilcoxon-Test*, siehe unten.

Der Wilcoxon-Test ist ein Kompromiss, der weniger veraussetzt als der t-Test und die Information der Daten besser ausnützt als der Vorzeichen-Test. Dazu bildet man die Ränge der Differenzen bezüglich des Absolutwertes: $\mathrm{Rang}(|U_i|) = k$ heisst, dass $|U_i|$ den k-ten kleinsten Wert hat unter $|U_1|,\ldots,|U_n|$. Wenn einzelne $|U_i|$ zusammenfallen, teilt man die Ränge auf durch Mittelung. Ausserdem sei noch V_i der indikator dafür, ob U_i positiv ist, d.h. $V_i = 1$ falls $U_i > 0$ ist und $V_i = 0$ sonst. Dann verwirft man die Nullhypothese, falls

$$W = \sum_{i=1}^{n} \mathrm{Rang}\big(|U_i|\big)\, V_i$$

zu gross oder zu klein oder beides ist (je nach Spezifikation der Alternative). Die Schranken für zu gross oder zu klein entnimmt man aus Tabellen oder Statistikpaketen für den Computer.

Man kann zeigen, adss dieser Test das Niveau exakt einhält (d.h. Wahrscheinlichkeit für einen Fehler 1.Art

ist gleihc α), wenn die U_i i.i.d. sind und eine 0 symmetrische Dichte haben. Beim t-Test wird das Niveau zwar auch ungefähr eingehalten bei vielen nichtnormalen Verteilungen (wegen dem ZGS), aber unter Umständen ist die Wahrscheinlichkeit eines Fehlers 2. ARt beim t-Test *viel grösser* als beim Wilcoxon-Test.

In der Praxis ist der Wilcoxon-Test allermeist dem t- oder Vorzeicehn-Test vorzuziehen. Nur falls die Daten sehr gut mit einer Normalverteilung beschrieben werden ist der t-Test für gute Datenanalyse „vollumfänglich tauglich“: diese Annahme oder Bedingung kann man z.B. mit dem Normal-Plot (siehe Kapitel 7.3) grafisch überprüfen.

11.3 Zwei-Stichproben Tests

Wie bereits beschrieben gibt es Fälle (ungepaarte Stichproben), wo man keine Paare bilden kann. Dann hat man i.i.d. Zufallsvariablen $X_1, \dots, X_n$ für die eine Versuchsbedingung und $Y_1, \dots, 1_m$ für die andere, und man nimmt an, dass alle Zufallsvariablen unabhängig sind. Die effektiv gemachten Beobachtungen sind wie üblich als realisierungen von diesen Zufallsvariablen zu interpretieren. das einfachste Problem lässt sich unter folgender Annahme lösen:

$$X_1, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma^2),$$
$$Y_1, \dots, Y_m \sim \mathcal{N}(\mu_Y, \sigma^2).$$

Beim *Zwei-Stichproben t-Test* ist die Teststatistik definiert durch

$$T \coloneqq \frac{\overline{X}_n - \overline{Y}_m}{S_{\text{pool}}\sqrt{\frac{1}{n}+\frac{1}{m}}},$$

wobei

$$S^2_{\text{pool}} = \frac{1}{n+m-2} \cdot \left(\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 + \sum_{i=1}^{m}\left(Y_i - \overline{Y}_m\right)^2\right) \tag{14}$$

die gepoolte Schätzung für die gemeinsame Varianz σ^2 ist. Die Wahl des Nenners in Gl. 14 ergibt sich aus

$$\mathbb{V}(\overline{X}_n - \overline{Y}_m) = \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right).$$

Die Teststatistik ist unter der Nullhypothese t_{n+m-2}-verteilt. Wir lehnen daher für eine gegebene Realisierung

$$t = \frac{\overline{x}_n - \overline{y}_n}{s_{\text{pool}}\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

von T je nach Alternative H_A die Nullhypothese $H_0 : \mu_X = \mu_y$ ab, falls

$$\begin{aligned} &|t| \geq t_{n+m-2,1-\alpha/2}, \\ \Leftrightarrow\ &t \in VB_\alpha = (-\infty, t_{n+m-2,\alpha/2}] \\ &\qquad \cup [t_{n+m-2,1-\alpha/2}, \infty), \\ &H_A : \mu \neq \mu_0. \end{aligned} \tag{15}$$

$$\begin{aligned} &t \geq t_{n+m-2,1-\alpha}, \Leftrightarrow t \in [t_{n+m-2,1-\alpha}, \infty), \\ &H_A : \mu > \mu_0. \end{aligned} \tag{16}$$

$$\begin{aligned} &t \leq t_{n+m-2,1-\alpha}, \Leftrightarrow t \in (-\infty, t_{n+m-2,\alpha}], \\ &H_A : \mu < \mu_0. \end{aligned} \tag{17}$$

Die Verallgemeinerungen des Zwei-Stichproben t-Tests bei ungleichen Varianzen

$$\sigma^2_X \neq \sigma^2_Y$$

ist in der Literatur zu finden. Ebenfalls in der Literatur zu finden ist der Zwei-Stichproben Wilcoxon-Test, welcher ein für die Praxis sehr guter Test für ungepaarte Stichproben ist.

Verteilung	$p(x)$ bzw. $f(x)$	W_X	$\mathbb{E}(X)$	$\mathbb{V}(X)$
$\text{Bin}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,\dots,n\}$	np	$np(1-p)$
$\text{Geometrisch}(p)$	$p(1-p)^{x-1}$	$\{1,2,\dots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\text{Poi}(\lambda)$	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\{0,(,\dots\}$	λ	λ
$\text{Uni}[a,b]$	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\exp(\lambda)$	$\lambda e^{-\lambda x}$	$\mathbb{R}_+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\text{Gamma}(\alpha,\lambda)$	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$\mathbb{R}_+$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}^2\right)^2}$	$\mathbb{R}$	μ	σ^2