Ergänzende Unterlagen
Statistische Mustererkennung
SS 2021

V 1.00

Thomas Melzer

10. Januar 2022

Inhaltsverzeichnis

1 Dichte- und Wahrscheinlichkeitsfunktionen **5**
1.1 Charakteristische Funktion 5

2 Multivariate Verteilungen **7**
2.1 Fehlerfortpflanzung 7
2.1.1 Beispiel *risk ratio* 8
2.1.2 Beispiel *odds ratio* 10

Kapitel 1

Dichte- und Wahrscheinlichkeitsfunktionen

1.1 Charakteristische Funktion

Bei der charakteristischen Funktion CF der DF $p(x)$ einer Zufallsvariable X handelt es sich um deren (inverse) Fourier-Transformation bzg. der Kreisfrequenz t:

$$\varphi_X(t) = \int \exp(itx)\, p(x)\, dx = \mathcal{E}[\exp(itx)]. \tag{1.1}$$

CFen erlauben es, die DFen transformierter Zufallsgrößen elegant zu berechne, was wir anhand der DF der Summe zweier unabhängiger, normalverteilter Größen demonstrieren wollen. Für die DF der Summe zweier unabhängiger Zufallsvariablen gilt zunächst allgemein

$$\begin{aligned}
\varphi_{X+Y}(t) &= \int\int \exp(it(x+y))\, p(x,y)\, dx\, dy \\
&= \int\int \exp(itx) \exp(ity)\, p(x)\, p(y)\, dx\, dy \\
&= \int \exp(itx)\, p(x)\, dx \int \exp(ity)\, p(y)\, dy \\
&= \varphi_X(t)\, \varphi_Y(t).
\end{aligned} \tag{1.2}$$

Mit dem Faltungstheorem (die Transformierte des Produkts zweier Funktionen ist gleich der Faltung der Transformierten) folgt, dass man die DF der Summe zweier unabhängiger Zufallsgrößen als **Faltung** deren DFen erhält.

Wir nehmen nun an, dass X normalverteilt ist. Gemäß (1.1) berechnet

sich die CF mit

$$\begin{aligned}\varphi_X(t) &= \int \exp(itx)\,\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,dx &(1.3)\\ &= \frac{1}{\sqrt{2\pi}\sigma}\int \exp\left(-\frac{x^2-2x\mu+\mu^2+itx2\sigma^2}{2\sigma^2}\right)\,dx. &(1.4)\end{aligned}$$

Wir formen nun den Nenner des Exponenten so um, dass wir eine quadratische Funktion in der Integrationsvariablen x (erste drei Terme ZQ) und einen nicht von x abhängigen Rest (die letzten zwei Terme ZR) erhalten:

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}\sigma}\int \exp\left(-\frac{ZQ+ZR}{2\sigma^2}\right)\,dx \tag{1.5}$$

mit

$$ZQ + ZR = x^2 - 2x(\mu+\sigma^2 it) + (\mu+it\sigma^2)^2 - (\mu+it\sigma^2)^2 + \mu^2. \tag{1.6}$$

Dank dieser "Vervollständigung des Quadrats" können wir das Integral über eine vollständige Normal-DF (welches 1 ergibt) abspalten

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}\sigma}\int \exp\left(-\frac{ZQ}{2\sigma^2}\right)\,dx\,\exp\left(-\frac{ZR}{2\sigma^2}\right) = \exp\left(-\frac{ZR}{2\sigma^2}\right) \tag{1.7}$$

und erhalten mit

$$ZR = t^2\sigma^4 - 2it\sigma^2\mu \tag{1.8}$$

schließlich

$$\varphi(X) = \exp\left(-\frac{ZR}{2\sigma^2}\right) = \exp\left(it\mu - \frac{1}{2}t^2\sigma^2\right). \tag{1.9}$$

Da die Fourier-Transformation invertierbar ist, entspricht jeder CF obiger Form in eindeutiger Weise eine Normal-DF. Betrachten wir nun das Produkt zweier Normal-CFen

$$\begin{aligned}\varphi_X(t)\varphi_Y(t) &= \exp\left(it\mu_1 - \frac{1}{2}t^2\sigma_1^2\right)\exp\left(it\mu_2 - \frac{1}{2}t^2\sigma_2^2\right)\\ &= \exp\left(it(\mu_1+\mu_2) - \frac{1}{2}t^2(\sigma_1^2+\sigma_2^2)\right), &(1.10)\end{aligned}$$

so ist unmittelbar ersichtlich, dass die Summe $Z = X+Y$ mit $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ ebenfalls normalverteilt ist mit $Z \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

Kapitel 2

Multivariate Verteilungen

Wir unterscheiden im folgenden nicht mehr in der Notation zwischen Schätzfunktionen und Schätzwerten. Es sollte aus dem Kontext heraus klar sein, worum es sich handelt. Geben wir z.B. eine Punktschätzung und deren Standardfehler an

$$\hat{p} \pm \mathrm{se}(\hat{p}), \tag{2.1}$$

so handelt es sich beim ersten Summanden um einen Schätzwert (eine Realisierung), das Argument von se(.) hingegen ist klarerweise der korrespondierende Schätzer (also eine Zufallsvariable).

2.1 Fehlerfortpflanzung

Sei $\vec{X} \in \mathbb{R}^p$ eine p-dimensionale Zufallsvariable mit Mittelwert $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$. Dann berechnen sich Mittelwert und Varianz der unter der affinen Transformation

$$\vec{Y} = \mathbf{F}\vec{X} + \mathbf{t}, \tag{2.2}$$

$\mathbf{F} \in \mathbb{R}^{q \times p}, \mathbf{t} \in \mathbb{R}^q, q \leq p$, erhaltenen Zufallsvariablen $\vec{Y}$ wie folgt

$$\mathcal{E}[\vec{Y}] = \mathbf{F}\boldsymbol{\mu} + \mathbf{t} \tag{2.3}$$

$$Var[\vec{Y}] = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T. \tag{2.4}$$

Für nichtlineare Abbildungen $\vec{Y} = f(\vec{X})$ erhält man ausgehend von der Taylor-Entwicklung um den Punkt $\mathbf{x}_0$

$$\vec{Y} - f(\mathbf{x}_0) = \frac{df}{d\mathbf{x}}(\vec{X} - \mathbf{x}_0) \tag{2.5}$$

eine lineare Approximation erster Ordnung mit

$$\mathbf{t} = f(\mathbf{x}_0) - \frac{df}{d\mathbf{x}}(\mathbf{x}_0) \tag{2.6}$$

$$\mathbf{F} = \frac{df}{d\mathbf{x}}(\mathbf{x}_0), \tag{2.7}$$

d.h. für $\mathbf{F}$ in (2.2) wählen wir die Jakobi-Matrix von f.

Wird die durch $\mathbf{F}$ bzw. f vermittelte Abbildung als Schätzer aufgefasst, so entsprechen die Quadratwurzeln der Diagonalelemente von $Cov[\vec{Y}]$ den Standardfehlern. Man spricht im Zusammenhang mit (2.4) auch von (statistischer) Fehlerfortpflanzung, da diese Gleichung beschreibt, wie sich die Fehler (hier: Unsicherheit, beschrieben durch die Kovarianzmatrix) der Eingangsgrößen unter der Transformation (Schätzung) f auf die Fehler der Ausgangsgrößen auswirken.

2.1.1 Beispiel *risk ratio*

Seien p_1, p_2 die bedingten Wahrscheinlichkeiten für das Eintreten eines Ereignisses $p_i = P(Y = 1|\omega_i)$ innerhalb der beiden Klassen ω_1, ω_2. Die beobachteten Ereignisse werden in der Epidemiologie auch als Fälle (*cases*), die Klassen je nach Kontext als Expositions- (*exposure*) oder Behandlungs-(*treatment*) Kategorien bezeichnet; wir werden im folgenden die erstere Bezeichnung verwenden. Die p_i bezeichnet man in diesem Kontext als Risiko (*risk*) (nicht zu verwechseln mit dem Erwartungswert des *loss* in der Entscheidungstheorie).

Beispiele:

- Erkrankung an Lungenkrebs (Y), Raucher vs. Nichtraucher (ω_1, ω_2)
- Tod bei Autounfall (Y), angeschnallt vs. nicht angeschnallt (ω_1, ω_2)
- Erkrankung an Infektionskrankheit (Y), geimpft vs. nicht geimpft (ω_1, ω_2)

Man interessiert sich nun dafür, wie stark die relative Anzahl der Fälle von der Exposition abhängt. Ein Maß für diese Effektstärke ist das relative Risiko (*risk ratio*) RR, definiert als

$$RR = p_1/p_2. \tag{2.8}$$

Ein Wert von 1 bedeutet, dass kein Effekt (Zusammenhang zwischen der Häufigkeit der Fälle und der Exposition) besteht.

Sei c_i die Anzahl der Fälle in einer Stichprobe vom Umfang N_i mit Exposition ω_i. p_i schätzt man als relativen Anteil

$$\hat{p}_i = c_i/N_i. \tag{2.9}$$

Der korrespondierende Schätzer ist binomialverteilt mit Standardfehler

$$\sqrt{p_i\,(1-p_i)/N_i}, \tag{2.10}$$

wobei anstelle der unbekannten p_i in der Praxis der Schätzwert $\hat{p}_i$ verwendet wird. Unter der Annahme, dass $\hat{p}_1, \hat{p}_2$ unabhängig sind, erhalten wir für die Eingangsgrößen die Kovarianzmatrix

$$\mathbf{\Sigma} = \begin{pmatrix} \frac{p_1\ (1-p_1)}{N_1} & 0 \\ 0 & \frac{p_2\ (1-p_2)}{N_2} \end{pmatrix} \tag{2.11}$$

Schwieriger verhält es sich mit dem Quotienten $\hat{RR} = \hat{p}_1/\hat{p}_2$. Dieser weist eine sehr schiefe Verteilung auf; allerdings ist der Logarithmus dieser Größe approximativ normalverteilt, was es nahelegt, Quantile, Konfidenzintervalle etc, in der Log-Domäne zu berechnen und anschließend zurückzutransformieren. Dazu benötigen wir allerdings den Standardfehler der Größe $\log \hat{RR}$. Wir berechnen zunächst die Jakobi-Matrix

$$\mathbf{F} = \frac{d \log RR(p_1, p_2)}{d(p1, p2)} = \left(\frac{1}{RR}\frac{\partial RR}{\partial p_1}, \frac{1}{RR}\frac{\partial RR}{\partial p_2} \right) \tag{2.12}$$

$$= \left(\frac{p_2}{p_1}\frac{1}{p_2}, -\frac{p_2}{p_1}\frac{p_1}{p_2^2} \right) = \left(\frac{1}{p_1}, -\frac{1}{p_2} \right) \tag{2.13}$$

(2.4) liefert nun die gesuchte Varianz des Logarithmus des Quotienten

$$Var[\log \hat{RR}] = \mathbf{F}\mathbf{\Sigma}\mathbf{F}^T = \frac{1-\hat{p}_1}{\hat{p}_1\ N_1} + \frac{1-\hat{p}_2}{\hat{p}_2\ N_2}, \tag{2.14}$$

wobei wir die p_i durch ihre korrespondierenden $\hat{p}_i$ ersetzt haben (was für (2.12) einer rein formalen Ersetzung der Argumente entspricht, für die Eingabekovarianz (2.11) allerdings, wie oben erwähnt, eine Ersetzung der wahren Wahrscheinlichkeiten durch deren Schätzwerte).

Führen wir abschließend die Substitution (2.9) durch, so erhalten wir

$$Var[\log \hat{RR}] = \frac{N_1 - c_1}{c_1\ N_1} + \frac{N_2 - c_2}{c_1\ N_2}, \tag{2.15}$$

bzw. folgende äquivalente, gebräuchliche Darstellung

$$Var[\log \hat{RR}] = \frac{1}{c_1} - \frac{1}{N_1} + \frac{1}{c_2} - \frac{1}{N_2}. \tag{2.16}$$

2.1.2 Beispiel *odds ratio*

Unter Chance bzw. Quote (*odds*) versteht man das Verhältnis der Wahrscheinlichkeit p zu ihrer Komplementärwahrscheinlichkeit $1-p$

$$o = \frac{p}{1-p}. \tag{2.17}$$

Chancen werden häufig als Bruch angegeben, z.B. $5:1$ (sprich: fünf zu eins) für die Wahrscheinlichkeit, keinen Sechser zu würfeln; so entsprechen von $a:b$ einer Wahrscheinlichkeit von $p = a/(a+b)$. Durch Verwendung von Chancen wird der Wertebereich der Wahrscheinlichkeit von $[0\ldots1]$ auf $[0,+\infty]$ aufgespreizt, bzw. für deren Logarithmus (*log odds*) $\log o$ auf $[-\infty,+\infty]$, was man sich z.B. bei der logistischen Regression zunutze macht.

Während das epidemiologische Risiko den Anteil der Fälle innerhalb einer Gruppe mit identischer Exposition angibt (Wahrscheinlichkeit, als Raucher Lungenkrebs zu bekommen), werden Chancen verwendet, um das Verhältnis der Expositionsklassen innerhalb der Fälle bzw Nichtfälle auszudrücken. Kommen z.B. auf einen an Lungenkrebs erkrankten Nichtraucher sechs Raucher, d.h. Chancen von $6:1$, während in der Gesamtbevölkerung zwei Raucher auf einen Nichtraucher kommen $(2:1)$, so wären die Raucher unter den Krebskranken um den Faktor $6/2 = 3$ überrepräsentiert.

Den obigen Faktor

$$OR = o_1/o_2 = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \tag{2.18}$$

bezeichnet man als Chancenverhältnis (*odds ratio*).

Chancen o und Risiko p sind nicht äquivalent, stimmen aber für kleine Werte von p gut überein. Die *odds ratio* wird insbesondere als Effektgröße für Fall-Kontroll-Studien (*case control studies*) verwendet.

Die Herleitung der Varianz von $\log OR$ erfolgt analog zum relativen Risiko $\log RR$. Die Kovarianzmatrix ist in beiden Fällen dieselbe, und für die Jacobi-Matrix erhalten wir

$$\begin{aligned}
\mathbf{F} &= \frac{d\log OR(p_1,p_2)}{d(p1,p2)} && (2.19)\\
&= \left(\frac{\partial\log p_1 - \log(1-p_1)}{\partial p_1}, -\frac{\partial\log p_2 - \log(1-p_2)}{\partial p_2}\right) && (2.20)\\
&= \left(\frac{1}{p_1\,(1-p_1)}, -\frac{1}{p_2\,(1-p_2)}\right). && (2.21)
\end{aligned}$$

(2.4) liefert wieder die gesuchte Varianz des Logarithmus des Quotienten

$$Var[\log\hat{OR}] = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T = \frac{1}{\hat{p}_1\,(1-\hat{p}_1)\,N_1} + \frac{1}{\hat{p}_2\,(1-\hat{p}_2)\,N_2}, \tag{2.22}$$

und mit der Substitution (2.9) erhalten wir schließlich

$$Var[\log \hat{OR}] = \frac{N_1}{(N_1 - c_1)\, c_1} + \frac{N_2}{(N_2 - c_2)\, c_2} \tag{2.23}$$

$$= \frac{1}{N_1 - c_1} + \frac{1}{c_1} + \frac{1}{N_2 - c_2} + \frac{1}{N_2}. \tag{2.24}$$