Die Gausssche Normalverteilung

Contents

9.1. Approximation der Binomialverteilung

二项分布的近似值

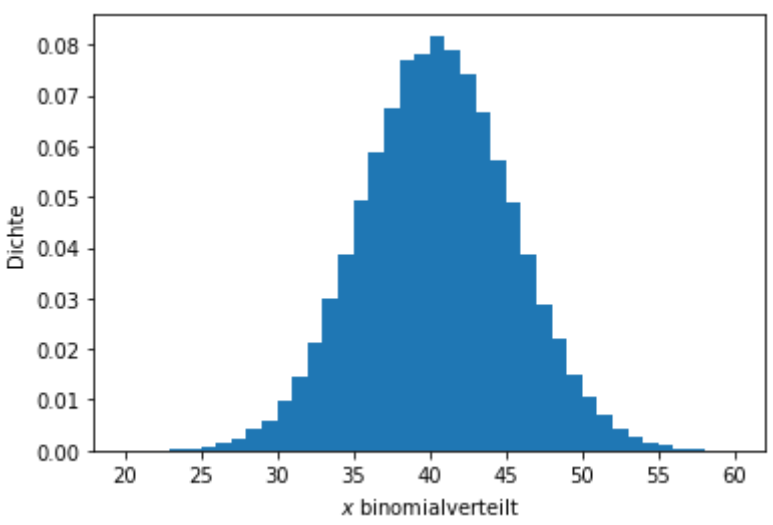
Betrachten wir eine Folge von Bernoulli-verteilten Zufallsvariablen X_i mit gleichem Parameter p

und $S_n := \sum_{i=1}^{n} X_i$ die Summe über die ersten n davon, so wissen wir bereits, dass $S_n \sim Bin(n,p)$, also dass diese Summe Binomialverteilt ist.

如果我们考虑一连串的伯努利分布的随机变量xi，具有相同的参数p和 在其中的前n个上的和sn，我们已经知道，即这个和是二项分布的。

Für die konkreten Werte $n = 100,\ p = 0.4$ sieht ein Histogramm von 100000 Samples etwa so aus:

```
import numpy as np
from scipy.stats import binom, norm
from matplotlib import pyplot as plt
np.random.seed(123123)
binom_samples = binom.rvs(n=100,
                          p=0.4,
                          size=100000)
plt.hist(binom_samples, bins=40, density=True)
plt.xlabel("$x$ binomialverteilt")
plt.ylabel("Dichte")
plt.show()
```



Da diese Form für große Stichproben stets so aussieht, unabhängig von p (probieren Sie das aus!), könnte man sich vorstellen, dass die Verteilung sich approximieren lässt mit Hilfe einer Funktion, die leichter zu berechnen ist als Binomialkoeffizienten.

由于这个形状对于大样本来说总是这样的，与（试试吧！）无关，可以想象，可以用一个比二项式系数更容易计算的函数来近似分布。

Tatsächlich gibt es zur Approximation von Binomialkoeffizienten auch die *Stirling-Formel*

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

事实上，也有斯特林公式用于近似二项式系数的计算

Damit lässt sich nun (auch wenn es nicht ganz einfach ist) eine Approximation zeigen:

$$P(S_n = k) \sim \frac{1}{\sqrt{np(1-p)}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{fuer } n \to \infty$$

wobei für x auf der rechten Seite die Folge $x = x_n$ mit $x_n = x_n(k) = \frac{k-np}{\sqrt{np(1-p)}}$ eingesetzt werden muss. Das Symbol $\sim$ bedeutet: asymptotisch gleich, d.h. der Quotient konvergiert gegen 1.

Diese Approximation ist insofern hilfreich, als dass wir nun die Funktion $e^{-\frac{x^2}{2}}$ tabellieren können und konkrete Werte für gewisse n, p ablesen können. Das ist wesentlich effizienter, als Binomialkoeffizienten auszurechnen. Die ganze Beobachtung heißt auch **Satz** von deMoivre-Laplace und soll uns zunächst als Motivation dienen, die Funktion $e^{-\frac{x^2}{2}}$ näher zu untersuchen.

这个近似值很有帮助，因为我们现在可以用表格表示出函数，，，这比计算二项式系数要有效得多。这比计算二项式系数要有效得多。整个观察也被称为deMoivre-Laplace定理，首先应作为我们更仔细地研究函数的动机。

9.2. Ein Integral

Um aus der Funktion $e^{-\frac{x^2}{2}}$ eine Wahrscheinlichkeitsdichte zu machen, muss das Integral 1 ergeben. Wir berechnen, weil es danach sehr nützlich wird, gleich ein etwas allgemeineres Integral:

> 🎓 **Lemma**
>
> Sei $v \in \mathbb{R}$ und $v > 0$. Dann gilt
>
> $$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2v}} dx = \sqrt{2\pi v}.$$

> **Beweis**
>
> Da auf der rechten Seite eine Quadratwurzel steht, bietet es sich an, beide Seiten der Gleichung zu quadrieren. Wir formen die linke Seite dann weiter um, bis wir ein Integral über $\mathbb{R}^2$ in Polarkoordinaten transformieren können und dann leicht Stammfunktionen bestimmen können:
>
> $$\begin{aligned}
> &\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2v}} dx\right)^2 \\
> =&\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2v}} dx\right)\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2v}} dy\right) \\
> =&\int_{-\infty}^{\infty} e^{-\frac{x^2}{2v}} \left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2v}} dy\right) dx \\
> =&\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2v}} e^{-\frac{y^2}{2v}} dy\right) dx \\
> =&\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2v}} dy\right) dx \\
> =&\int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2v}} d(x,y) \\
> =&\int_0^{\infty} \int_0^{2\pi} r e^{-\frac{r^2}{2v}} d\phi dr \\
> =&\int_0^{\infty} \left[r e^{-\frac{r^2}{2v}} \phi\right]_{\phi=0}^{\phi=2\pi} dr \\
> =&\int_0^{\infty} r e^{-\frac{r^2}{2v}} (-2\pi) dr \\
> =&-2\pi\left[-v e^{-\frac{r^2}{2v}} dr\right]_{r=0}^{r=\infty} \\
> =&2\pi v
> \end{aligned}$$

Da Integrale in der Statistik häufiger vorkommen, ist es nicht ganz verkehrt, sich bei dieser Rechnung klar zu machen, was genau warum in jedem Schritt passiert.

由于积分在统计学中比较常见，所以在这个计算中明确每一步到底发生了什么原因，也不是完全错误。

9.3. Die eindimensionale Normalverteilung

9.3 一维正态分布

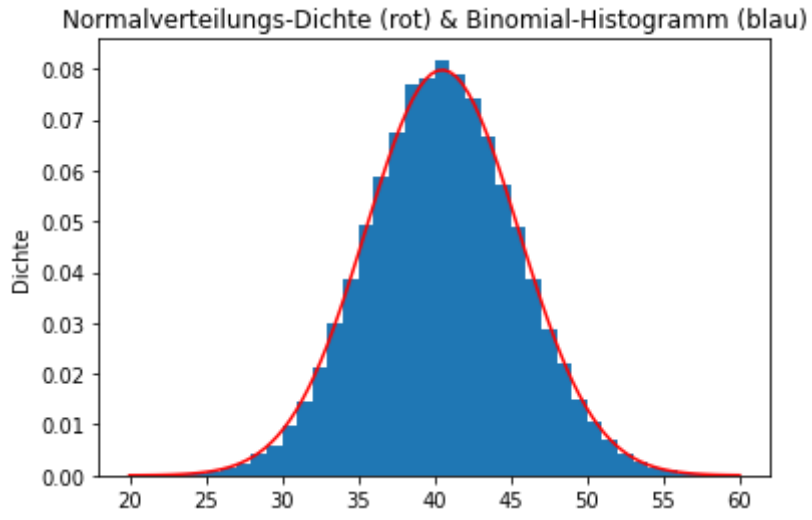
Definition

Seien $\mu, \sigma \in \mathbb{R}$ mit $\sigma > 0$. Dann heißt die Verteilung mit Dichtefunktion

$$\phi(x) = \phi(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normalverteilung und wir schreiben für eine reelle Zufallsvariable X mit dieser Verteilung auch $X \sim \mathcal{N}(\mu, \sigma^2)$.

```
x = np.linspace(20,60,100)
plt.hist(binom_samples, bins=40, density=True)
plt.plot(x, norm(loc=40.5, scale=5).pdf(x), color="red")
plt.ylabel('Dichte')
plt.title('Normalverteilungs-Dichte (rot) & Binomial-Histogramm (blau)')
plt.show()
```



Proposition

Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist $\mathbb{E}(X) = \mu$ und $\mathbb{V}(X) = \sigma^2$.

Beweis

Wir setzen die Dichte für X in die Formel für $\mathbb{E}(X)$ ein und transformieren das Integral über x in ein Integral über $x+\mu$:

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} x\phi(x)dx \\
&= \int_{-\infty}^{\infty} x\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\
&= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\
&= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} (x+\mu)e^{-\frac{x^2}{2\sigma^2}}dx \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left(\left(\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2\sigma^2}}dx\right) + \mu\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}dx\right)\right) \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left(\left(\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2\sigma^2}}dx\right) + \mu\sigma\sqrt{2\pi}\right) \\
&= \mu + \frac{1}{\sigma\sqrt{2\pi}}\left(\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2\sigma^2}}dx\right) = \mu
\end{aligned}$$

Der Ausdruck nach dem $\mu+$ ist einfach 0, denn die Funktion ist punktsymmetrisch um 0 (der Faktor x ist offensichtlich eine ungerade Funktion, der andere Faktor hängt nur von $|x|$ ab).

Genau so kann man bei der Varianz verfahren, indem man diesen Ausdruck vereinfacht:

$$\begin{aligned}
\mathbb{V}(X) &= \int_{-\infty}^{\infty} x^2\phi(x)dx - (\mathbb{E}(X))^2 \\
&= \int_{-\infty}^{\infty} x^2\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx - \mu^2
\end{aligned}$$

Definition

Man nennt eine Zufallsvariable mit $X \sim \mathcal{N}(0,1)$ auch *standardnormalverteilt*

Proposition

Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Beweis

Idee: Es ist sofort klar, dass $\mathbb{E}(X-\mu) = 0$ und dass $\mathbb{V}\left(\frac{X-\mu}{\sigma}\right) = 1$. Weniger klar ist, dass die neue Zufallsvariable tatsächlich normalverteilt ist. Das lässt sich z.B. mit der Momenterzeugendenfunktion beweisen.

Alternativ nutzen wir, dass die Summe von unabhängigen Zufallsvariablen mit Dichten f, g selbst wieder eine Dichte hat, nämlich die *Faltung* $f * g$. Anstatt das nun rigoros einzuführen und zu beweisen, benutzen wir es ein weiteres Mal, damit lässt sich nämlich zeigen:

Proposition

Wenn $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ und $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, so ist $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

或者，我们使用这样一个事实：密度为**f*g**的独立随机变量之和本身就有一个密度，即卷积**f*g**。我们不严格介绍和证明这一点，而是再使用一次，因为它可以被证明。

Satz

Unter allen Verteilungen reeller Zufallsvariablen X mit festem Erwartungswert $\mu = \mathbb{E}(X)$ und Varianz $\sigma^2 = \mathbb{V}(X)$ ist die Normalverteilung diejenige mit der maximalen Entropie.

为了能够准确地提出这个定理，我们需要一个连续分布的熵的概念。与其现在这样做，不如在以后我们讨论相对熵、其连续版本和可能性时再处理这个问题。

Um diesen Satz überhaupt präzise formulieren zu können, benötigen wir einen Entropiebegriff für stetige Verteilungen. Anstatt das jetzt zu tun, wollen wir uns später damit beschäftigen, wenn wir auch relative Entropie, die stetige Version davon und Likelihood diskutieren.

Wichtig ist aber die Take-Home-Message des Satzes: Wenn über eine stetige Verteilung reeller Zahlen außer Erwartungswert und Varianz nichts bekannt ist, dann ist die entprechende Normalverteilung die vernünftigste Annahme. In diesem Sinne ist die Normalverteilung ein guter stetiger Ersatz für die diskrete Gleichverteilung, aber nun auf ganz $\mathbb{R}$ (die Normalverteilung hat Träger $\mathbb{R}$).

但该定理的启示是很重要的。如果除了期望值和方差之外，对实数的连续分布一无所知，那么相应的正态分布就是最合理的假设。在这个意义上，正态分布是离散均匀分布的一个很好的连续替代品，但现在是在所有的**R**上（正态分布有支持**R**）。

10 中心极限定理

Satz

Wenn X_i eine Folge von identisch verteilten, voneinander unabhängigen (iid = independent identically distributed) Zufallsvariablen mit Erwartungswert $\mathbb{E}(X_i) = 0$ und Varianz $\mathbb{V}(X_i) = \sigma^2$ ist, und wir die skalierte Summe $S_N := \frac{\sqrt{N}}{N} \sum_{i=1}^{N} X_i$ betrachten, dann gilt die Konvergenz in Verteilung:

$$S_N \xrightarrow{N\to\infty} \mathcal{N}(0,1)$$

分布收敛是指分布函数向标准正态分布的分布函数收敛。

Konvergenz in Verteilung heißt, dass die Verteilungsfunktion von S_N gegen die Verteilungsfunktion einer Standardnormalverteilung konvergiert.

Von diesem Satz gibt es auch Abschwächungen, die gewisse Abhängigkeiten zwischen den X_i erlauben.

该定理对我们很重要，因为它允许我们用正态分布对独立随机变量之和进行近似。在对物理或技术社会系统进行测量时，这种情况经常发生，这些系统通常受到一长串基本独立的扰动的影响。这个误差源的总和（对于足够多的独立误差源）是近似正态分布。

Bemerkung

Wichtig ist der Satz für uns, weil er erlaubt eine Summe unabhängiger Zufallsvariablen mit einer Normalverteilung zu approximieren. Das tritt in der Praxis häufig auf, wenn man Messungen an physikalischen oder technosozialen Systemen vornimmt, die in der Regel von einer langen Liste weitgehend unabhängiger Störeinflüsse beeinträchtigt werden. Diese Summe an Fehlerquellen ist (für eine hinreichend große Zahl an unabhängigen Fehlerquellen) etwa normalverteilt.

举两个具体的例子：用GPS等GNSS系统确定位置的误差是正态分布的（即使使数值不那么准确的大气干扰也不是正态分布）。在部件的工业生产中，整个生产过程中也有一些误差源，最终加起来就是一个正态分布--其方差必须保持足够小，以便部件能够完成其任务。例如，一块乐高砖不能与一块相同的乐高砖有太大的偏差，否则以后的结构就不能正常保持。

Um zwei konkrete Beispiele zu nennen: der Fehler bei der Ortsbestimmung mit GNSS-Systemen wie GPS ist normalverteilt (auch wenn die atmosphärischen Störungen, die den Wert ungenauer machen, nicht normalverteilt sind). Bei der industriellen Fertigung von Bauteilen gibt es ebenfalls im gesamten Produktionsprozess Fehlerquellen, die am Ende zu einer Normalverteilung aufsummieren - deren Varianz man hinreichend klein halten muss, damit das Bauteil seine Aufgabe erfüllen kann. So darf ein Legostein nicht zu stark von einem baugleichen Legostein abweichen, sonst hält das Bauwerk hinterher nicht richtig.

Man kann außerdem die Geschwindigkeit der Konvergenz abschätzen, und damit in Erfahrung bringen, wie gut die Approximation durch eine Normalverteilung ist:

我们还可以估计收敛的速度，从而找出正态分布的近似程度。

🎓 Satz

Es existiert eine Konstante $C > 0,4409$ mit der folgenden Eigenschaft:

Wenn X_i eine Folge von iid Zufallsvariablen mit Erwartungswert 0 und Varianz σ^2 ist, und außerdem die dritten absoluten Momente $\mathbb{E}|X_i^3| = \rho < \infty$ existieren, wir F_N für die kumulative Verteilungsfunktion von $\frac{\sqrt{N}}{N\sigma}\sum_{i=1}^{N} X_i$ schreiben und Φ für die kumulative Verteilungsfunktion der Standardnormalverteilung $\mathcal{N}(0,1)$, so gilt für alle x und alle N:

$$|F_n(x) - \Phi(x)| \;\leq\; \frac{C\rho}{\sigma^3\sqrt{n}}.$$

所以对于具体已知的，我们总是可以确定**N**来估计左手边的任意小。换句话说：可以按任何所需的精度计算出要把多少个**iid**随机变量加起来（按比例）才能用标准正态分布来近似。

Für konkret bekannte σ, ρ können wir also stets das n bestimmen, um die linke Seite beliebig klein abzuschätzen. Mit anderen Worten: es lässt sich zu jeder gewünschten Genauigkeit berechnen, wie viele der iid Zufallsvariablen man (skaliert) aufaddieren muss, um es mit einer Standardnormalverteilung zu approximieren.

Klar: Wenn der Erwartungswert nicht 0 ist, lässt sich eine Variante als Korollar beweisen (wie auch beim zentralen Grenzwertsatz), sodass die Konvergenz gegen $\mathcal{N}(\mu, 1)$ geht. Wenn man anders reskaliert, auch gegen $\mathcal{N}(\mu, \sigma^2)$. In der Praxis sieht es eher so aus, dass man alle unbekannten Fehlerterme mit einer Normalverteilung modelliert, deren Erwartungswert und Varianz man empirisch schätzt.

显然：如果期望值不是，可以作为一个推论证明一个变体（如中心极限定理），这样收敛性就会与**N,1**相反。在实践中，它看起来就像是用正态分布来模拟所有的未知误差项，其期望值和方差是根据经验估计的。