

Eine Einführung in R: Dichten und Verteilungsfunktionen

Bernd Klaus, Verena Zuber

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
Universität Leipzig

25. November 2009

I. Diskrete Daten

Theorie: Wahrscheinlichkeits und Verteilungsfunktion
Diskrete Verteilungen

II. Stetige Daten

Theorie: Dichte und Verteilungsfunktion
Stetige Verteilungen

III. Der Umgang mit Zufallszahlen

Erzeugen von Zufallszahlen
Darstellung von Verteilungen

Einschub: Zufallsvariablen

Eine Variable oder Merkmal X , dessen Werte die Ergebnisse eines Zufallsvorganges sind, heißt Zufallsvariable.

Notation:

- ▶ X : Die Zufallsvariable
- ▶ x : Eine Realisierung oder Beobachtung der Zufallsvariable

I. Diskrete Daten

Eine Zufallsvariable heißt diskret, wenn sie endlich viele Werte x_1, \dots, x_k, \dots annehmen kann.

Die **Wahrscheinlichkeitsfunktion** $f(x)$ einer diskreten Zufallsvariable X ist für $x \in \mathbb{R}$ definiert durch die Wahrscheinlichkeiten p_i :

$$f(x) = \begin{cases} P(X = x_i) = p_i & \text{falls } x = x_i \in \{x_1, \dots, x_k, \dots\} \\ 0 & \text{sonst} \end{cases}$$

Die **Verteilungsfunktion** $F(x)$ einer diskreten Zufallsvariable ist gegeben durch die Summe:

$$F(y) = P(X \leq y) = \sum_{i: x_i \leq y} f(x_i)$$

Eigenschaften

Für die Wahrscheinlichkeitsfunktion $f(x)$ gilt:

$$0 \leq f(x) \leq 1$$

$$\sum_{i \geq 1} p_i = 1$$

Für die Verteilungsfunktion $F(x)$ gilt:

$$F(x) = \begin{cases} 1 & x \geq \max(x) \\ 0 & x \leq \min(x) \end{cases}$$

$F(x)$ ist monoton steigend mit Wertebereich 0 bis 1.

Bernoulli-Experiment

Binäre Zufallsvariable X : Tritt ein Ereignis A ein?

$$X = \begin{cases} 1 & \text{falls } A \text{ eintritt} \\ 0 & \text{falls } A \text{ nicht eintritt} \end{cases}$$

Das Ereignis A tritt mit einer bestimmten Wahrscheinlichkeit $0 < \pi < 1$ ein

$$P(X = 1) = \pi$$

$$P(X = 0) = 1 - \pi$$

Binomialverteilung

Die Binomialverteilung entspricht dem n -maligen Durchführen eines Bernoulli-Experimentes mit Wahrscheinlichkeit π

$$f(x) = \begin{cases} \binom{n}{x} \pi^x (1 - \pi)^{n-x} & \text{falls } x = 0, 1, \dots, n \\ 0 & \text{sonst} \end{cases}$$

Bsp: Ein Schütze schießt $n = 10$ mal auf eine Torwand. Wie groß ist die Wahrscheinlichkeit, dass er genau fünfmal trifft, wenn er eine Trefferwahrscheinlichkeit π von 25 % hat?

$$P(X = 5) = \binom{10}{5} 0.25^5 (1 - 0.25)^{10-5} = 0.058$$

Diskrete Gleichverteilung

Die diskrete Gleichverteilung charakterisiert die Situation, dass k_1, \dots, k_n -verschiedene Werte mit gleicher Wahrscheinlichkeit angenommen werden.

$$f(x) = \begin{cases} \frac{1}{n} & \text{falls } x = 1, \dots, n \\ 0 & \text{sonst} \end{cases}$$

Bsp: Würfeln, jede Zahl hat die gleiche Wahrscheinlichkeit $\frac{1}{6}$

II. Stetige Daten

Eine Zufallsvariable heißt stetig, wenn sie unendlich viele Werte x_1, \dots, x_k, \dots annehmen kann, wie beispielsweise metrische Variablen.

Die Dichte $f(x)$ einer stetigen Zufallsvariable X ist für ein Intervall $[a, b]$ definiert als:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Die Verteilungsfunktion $F(y)$ einer stetigen Zufallsvariable ist gegeben durch das Integral:

$$F(y) = P(X \leq y) = \int_{-\infty}^y f(x) dx$$

Eigenschaften

Für die Dichte $f(x)$ gilt:

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

$$P(X = a) = \int_a^a f(x) dx = 0$$

Für die Verteilungsfunktion $F(x)$ gilt:

$$F(x) = \begin{cases} 1 & \text{für } x \geq \max(x) \\ 0 & \text{für } x \leq \min(x) \end{cases}$$

$$F'(x) = \frac{\partial F(X)}{\partial x} = f(x)$$

Normalverteilung $N(\mu, \sigma)$

Eine der wichtigsten Verteilungen ist die Normal- oder Gauß-Verteilung mit Erwartungswert μ und Varianz σ^2 :

$$f(x|\mu, \sigma) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right)$$

- ▶ Symmetrisch um μ
- ▶ Nur abhängig von μ und σ
- ▶ Beispiele: Klausurnoten, das (logarithmierte) Einkommen, Messfehler, Größe und Gewicht

Stetige Gleichverteilung $U(a, b)$

Gegeben: ein Intervall, definiert durch reelle Zahlen a und b mit $a < b$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in [a, b] \\ 0 & \text{sonst} \end{cases}$$

Die stetige Gleichverteilung spielt eine wichtige Rolle bei statistischen Tests.

Hat man x_1, \dots, x_n Realisierungen einer Variablen X mit Verteilungsfunktion F , so gilt:

$$F(x_1), \dots, F(x_n) \sim U(0, 1)$$

III. Umgang mit Zufallszahlen

R ermöglicht den Umgang mit Zufallszahlen.

Beispiel: (Standard)Normalverteilung

1. Ziehen von n Zufallszahlen: `rnorm(n, mean=0, sd=1)`

2. Dichte im Wert x : `dnorm(x, mean=0, sd=1)`

Beispiel: `dnorm(c(-1,0,1))`

0.24197 0.39894 0.24197

3. Verteilungsfunktion im Wert x :

`pnorm(x, mean=0, sd=1)`

Beispiel: `pnorm(c(-1,0,1))`

0.15866 0.50000 0.84134

4. Quantil für Wahrscheinlichkeit p :

`qnorm(p, mean=0, sd=1)`

Beispiel: `qnorm(c(0.25,0.5,0.75))`

-0.67449 0.00000 0.67449

Beispiel: (Standard)Normalverteilung

1. Dichte im Wert x :

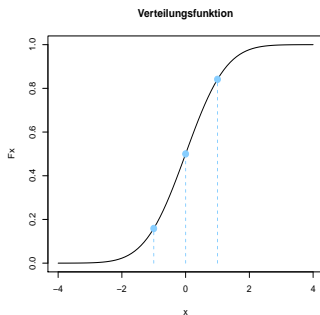
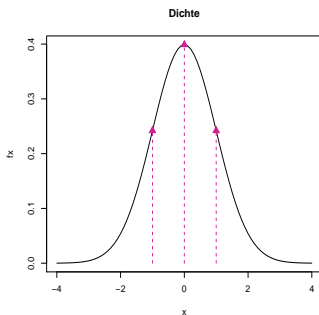
```
dnorm(c(-1,0,1))
```

0.24197 0.39894 0.24197

2. Verteilungsfunktion im Wert x :

```
pnorm(c(-1,0,1))
```

0.15866 0.50000 0.84134



R-Befehle für weitere Verteilungen

- ▶ `rnorm(n, mean=0, sd=1)` Normalverteilung mit Mittelwert `mean` und Standardabweichung `sd`
- ▶ `rexp(n, rate=1)` Exponentialverteilung mit Rate `rate`
- ▶ `rpois(n, lambda)` Poissonverteilung mit Rate `lambda`
- ▶ `rcauchy(n, location=0, scale=1)` Cauchyverteilung mit Lokations- und Skalenparameter
- ▶ `rt(n, df)` (Student)t-verteilung mit Freiheitsgraden `df`
- ▶ `rbinom(n, size, prob)` Binomialverteilung vom Umfang `size` und Wahrscheinlichkeit `prob`
- ▶ `rgeom(n, prob)` Geometrische Verteilung mit Wahrscheinlichkeit `prob`
- ▶ `rhyper(nn, m, n, k)` Hypergeometrische Verteilung
- ▶ `runif(n, min=0, max=1)` Stetige Gleichverteilung im Intervall `[min, max]`

Darstellung: Histogramme und Kerndichteschätzer

1. **Histogramme:** Darstellung von stetigen und diskreten Verteilungen

```
hist(x, breaks = "AnzahlBins", freq = NULL )
```

- ▶ `x`: Daten
- ▶ `breaks = "AnzahlBins"`: Steuerung der Teilintervalle
- ▶ `freq=TRUE`: absolute Häufigkeiten
- ▶ `freq=FALSE`: relative Häufigkeiten ("empirische Dichte")

2. **Kerndichteschätzer:** Darstellung von stetigen Verteilungen

```
plot(density(x, kernel="gaussian", bw))
```

- ▶ `density(x)`: Kerndichteschätzung der Daten
- ▶ `kernel`: Option für spezielle Kerntypen
- ▶ `bw`: Bandbreite

Darstellung: Kerndichteschätzer

Kerndichteschätzer sind aus dem Histogramm abgeleitete Verfahren zur Schätzung von stetigen Dichten

Hat man gegebene Daten x_1, \dots, x_n und eine konstante Bandbreite $h \in \mathbb{R}$ so ist der Kerndichteschätzer gegeben durch:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

Typische Kerne sind:

- ▶ Bisquare Kern:

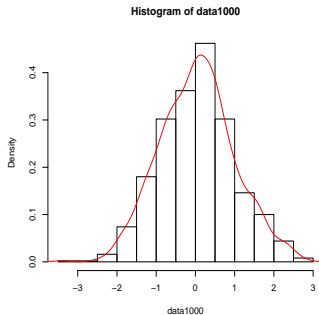
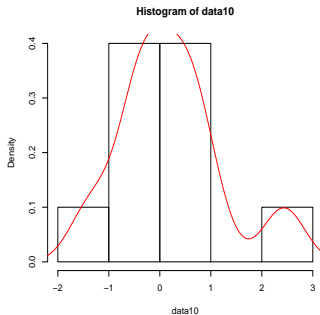
$$K(u) = \frac{15}{16}(1 - u^2)^2 \quad \text{für } u \in [-1, 1] \text{ und } 0 \text{ sonst}$$

- ▶ Gauß Kern: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$ für $u \in \mathbb{R}$

Beispiel: Simulation aus der Normalverteilung

```
data10<-rnorm(10)  
hist(data10, freq=FALSE)  
lines(density(data10), col=2)
```

```
data1000<-rnorm(1000)  
hist(data1000, freq=FALSE)  
lines(density(data1000), col=2)
```



Beispiel: Wie plottet man die Normalverteilung?

```
x<-seq(from=-4, to=4, by=0.1)
```

```
## Dichte
```

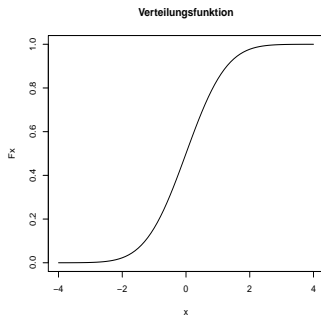
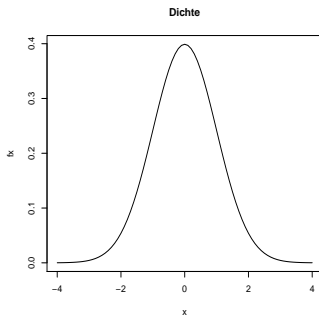
```
fx<-dnorm(x)
```

```
plot(x,fx, type="l")
```

```
## Verteilungsfunktion
```

```
Fx<-pnorm(x)
```

```
plot(x,Fx, type="l")
```



Darstellung: Q-Q-Plot

Quantil-Quantil-Plots tragen die (eventuell empirischen) Quantile zweier Verteilungen gegeneinander ab. Somit können Verteilungen miteinander verglichen werden.

- ▶ `qqplot(x,y)`: Plottet die emp. Quantile von x gegen die von y
- ▶ `qqnorm(y)`: Plottet die emp. Quantile von y gegen die einer Standard-Normalverteilung
- ▶ `qqline(y)`: Fügt dem Quantilplot eine Gerade hinzu die durch das erste und dritte Quartil geht

Bsp: Vergleich von Normal- und t -Verteilung

```
data <- rt(400, df = 2)
qqnorm(data, main = "QQ-Plot", xlab= "Normalverteilung",
ylab = "t-Verteilung")
qqline(data, col = "green")
```

Darstellung: Q-Q-Plot

