

## Statistische Software (R-Vertiefung)

Paul Fink, M.Sc.

Institut für Statistik  
Ludwig-Maximilians-Universität München

*Pseudo Zufallszahlen, Dichten, Verteilungsfunktionen, etc.*



Funktion	Maß
mean()	arithmetische Mittel
median()	Median
exp(mean(log( )))	Geometrisches Mittel
quantile()	Quantile
var()	Varianz (Version mit $1/(n-1)$ )
sd()	Standardabweichung (Version mit $1/(n-1)$ )
range()	Minimum und Maximum
diff(range())	Spannweite
cov()	Kovarianz (Version mit $1/(n-1)$ )
cor()	Korrelation (Spearman, Bravais–Pearson)
density()	Kerndichteschätzer
ecdf()	Empirische Verteilungsfunktion

Fink: Statistische Software (R) SoSe 2013

1

## Nützliche Funktionen

Weitere nützliche Funktionen im Zusammenhang der deskriptiven Beschreibung von Daten:

- Sortieren eines Vektors:

```
> x <- c(1, 3, 2, 5)
> sort(x)
[1] 1 2 3 5
> sort(x, decreasing = TRUE)
[1] 5 3 2 1
```

- Bestimmung der Ränge:

```
> x <- c(1, 3, 2, 5, 2)
> rank(x)
[1] 1.0 4.0 2.5 5.0 2.5
```

- Indizierung mehrfach vorkommender Werte in einem Vektor:

## Nützliche Funktionen

```
> x <- c(1, 3, 2, 5, 2)
> duplicated(x)
[1] FALSE FALSE FALSE FALSE TRUE
```

- Entfernung von Duplikaten (z.B. Bestimmung aller vorkommenden Merkmalsausprägungen in einer Stichprobe):

```
> x <- c(1, 3, 2, 5, 2)
> unique(x)
[1] 1 3 2 5
```

- Diskretisierung einer (quasi-)stetigen Variable:

## Nützliche Funktionen

```
> x <- c(1.3, 1.5, 2.5, 3.8, 4.1, 5.9, 7.1, 8.4, 9.0)
> xdiscrete <- cut(x, breaks=c(-Inf, 2, 5, 8, Inf) )
> is.factor(xdiscrete)
[1] TRUE
> xdiscrete
[1] (-Inf,2] (-Inf,2] (2,5] (2,5] (2,5] (5,8] (5,8] (8, Inf]
[9] (8, Inf]
Levels: (-Inf,2] (2,5] (5,8] (8, Inf]
> table(xdiscrete)
xdiscrete
(-Inf,2] (2,5] (5,8] (8, Inf]
      2      3      2      2
```

Nützliche Funktionen, die im Zusammenhang mit Statistik I und Statistik II bereits zum Teil verwendet wurden:

- Kumulierte Summe und Produkt:

```
> x <- c(1, 3, 2, 5)
> cumsum(x)      # 1, 1+3, 1+3+2, 1+3+2+5
[1] 1 4 6 11
> cumprod(x)     # 1, 1*3, 1*3*2, 1*3*2*5
[1] 1 3 6 30
```

## Nützliche Funktionen

```
> gamma(5) * gamma(3) / gamma(5 + 3)
[1] 0.00952381
> beta(5, 3)
[1] 0.00952381
```

- Der Hilfe ?union ist folgendes Beispiel zu Operationen auf Mengen entnommen:

## Nützliche Funktionen

- Fakultät:

```
> factorial(5)
[1] 120
```

- Binomialkoeffizient  $\binom{n}{k}$ :

```
> choose(4,2)
[1] 6
```

- Gammafunktion. Für natürliche Zahlen  $n$  gilt:  $\Gamma(n) = (n-1)!$

```
> gamma(5)
[1] 24
> gamma(0.5)
[1] 1.772454
> sqrt(pi)
[1] 1.772454
```

- Betafunktion  $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ :

## Nützliche Funktionen

```
> (x <- c(sort(sample(1:20, 9)),NA))
[1] 1 5 9 15 16 17 18 19 20 NA
> (y <- c(sort(sample(3:23, 7)),NA))
[1] 4 10 11 14 15 17 21 NA
> union(x, y)
[1] 1 5 9 15 16 17 18 19 20 NA 4 10 11 14 21
> intersect(x, y)
[1] 15 17 NA
> setdiff(x, y)
[1] 1 5 9 16 18 19 20
> setdiff(y, x)
[1] 4 10 11 14 21
> setequal(x, y)
[1] FALSE
> ## True for all possible x & y :
> setequal(union(x,y),
+ c(setdiff(x,y), intersect(x,y), setdiff(y,x)))
[1] TRUE
> is.element(x, y)# length 10
[1] FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE TRUE
> is.element(y, x)# length 8
[1] FALSE FALSE FALSE FALSE TRUE TRUE FALSE TRUE
```

## Verteilungen und Zufallszahlen

- R bietet Funktionen zur Berechnung von Dichten, Verteilungsfunktionen und Quantilen, sowie zur Erzeugung von (Pseudo-) Zufallszahlen gebräuchlicher Verteilungen.

- Der Funktionsname folgt dabei folgendem Schema:

Anfangsbuchstabe	Art der Funktion
d	Dichte (density)
p	Verteilungsfunktion (probability)
q	Quantilsfunktion (quantiles)
r	Zufallszahl (random number)

- Beispiele:

```
> dnorm(x = 0)
[1] 0.3989423
> 1 / sqrt(2*pi)
[1] 0.3989423
```

## Verteilungen und Zufallszahlen

- Diese Funktionen stehen u.a. für folgende Verteilungen zur Verfügung:

### Modellverteilungen

Funktion	Verteilung
beta	Beta
binom	Binomial
cauchy	Cauchy
exp	Exponential
gamma	Gamma
geom	Geometrische
hyper	Hypergeometrische
lnorm	Log-Normal
norm	Normal
pois	Poisson
unif	Gleichverteilung, Rechteckverteilung
mvnorm	Multivariate Normalverteilung, package mvtnorm

## Verteilungen und Zufallszahlen

liefert die Dichte der  $N(0,1)$ -Verteilung an der Stelle  $x = 0$  ( $1/\sqrt{2\pi}$ ).

```
> pnorm(q = 0)
[1] 0.5
> pnorm(q = 1.96)
[1] 0.9750021
```

liefert die Verteilungsfunktion der  $N(0,1)$ -Verteilung an der Stelle  $q$ , also  $\Phi(q) = P(X \leq q)$ .

```
> qnorm(p=0.95)
[1] 1.644854
```

liefert das  $p$ -Quantil der  $N(0,1)$ -Verteilung, also  $z_p$ .

```
> X <- rnorm(n = 5)
> X
[1] -0.1934486 -1.7696142 -2.4949481 -1.0112126 0.9839869
```

schliesslich liefert eine Stichprobe vom Umfang  $n = 5$  einer  $N(0,1)$ -Verteilung, sogenannte Pseudo-Zufallszahlen.

## Verteilungen und Zufallszahlen

### Prüfverteilungen

Funktion	Verteilung
chisq	$\chi^2$
f	$F$
signrank	Wilcoxon Vorzeichen-Rangsummen (1 Stichprobe)
t	$t$
wilcox	Wilcoxon Rangsummen (2 Stichproben)

- Eine nützliche Funktion, zum Beispiel für *Bootstrap-Verfahren*, ist `sample()`. Sie erlaubt das Ziehen einer Stichprobe aus einem Vektor vom Umfang `size` mit oder ohne Zurücklegen (`replace=TRUE` oder `replace=FALSE`) und optional mit bestimmten Wahrscheinlichkeiten (Parameter `prob`).

Beispiel:

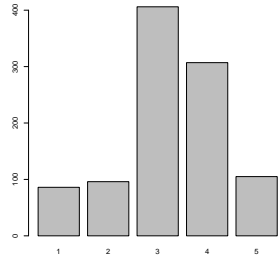
```
> sample(x = c(1, 2, 3, 4, 5), size = 10, replace = TRUE)
[1] 1 3 3 3 5 3 2 3 1 1
```

# Verteilungen und Zufallszahlen

Der `replace`-Parameter ist auf `FALSE` voreingestellt.

Beispiel zur Verwendung des `prob()`- Parameters:

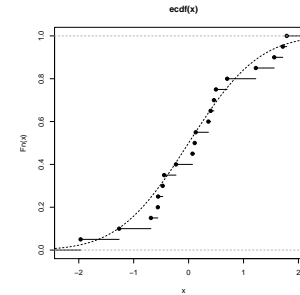
```
> y <- sample(x = c(1, 2, 3, 4, 5),
+           size = 1000,
+           replace = TRUE,
+           prob = c(0.1, 0.1, 0.4, 0.3, 0.1))
> barplot(table(y))
```



# Verteilungen und Zufallszahlen

- Empirische und theoretische Verteilungsfunktion der Standardnormalverteilung:

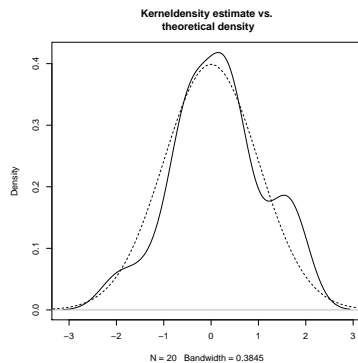
```
> set.seed(123)
> x <- rnorm(20)
> Fn <- ecdf(x)
> plot(Fn)
> lines(seq(-4, 4, by = .1), pnorm(seq(-4, 4, by = .1)), lty = "dashed")
```



# Verteilungen und Zufallszahlen

- Kerndichtschätzer und theoretische Verteilungsfunktion der Standardnormalverteilung:

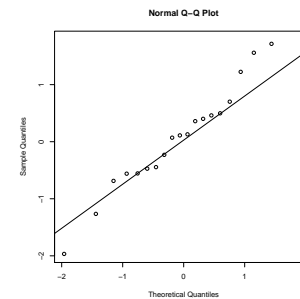
```
> fn <- density(x)
> plot(fn, main = "Kerndensity estimate vs.\n theoretical density")
> lines(seq(-4, 4, by = .1), dnorm(seq(-4, 4, by = .1)), lty = "dashed")
```



# Verteilungen und Zufallszahlen

- Im weiteren können die quantile der Stichprobe mit den theoretischen Quantilen der Standardnormalverteilung verglichen werden. (Mithilfe von `qqplot` können auch beliebige Verteilungen auf Basis ihrer Quantile verglichen werden.)

```
> qqnorm(x)
> qqline(x)
```



# Aufgaben

---

1. Erzeugen Sie Stichproben aus verschiedenen Verteilungen (Poisson, Binomial,  $\chi^2$ , Exponential) mit verschiedenen Parametern und Stichprobenumfängen. Visualisieren Sie die Approximation der standardisierten Summen durch die Normalverteilung (zentraler Grenzwertsatz).
2. Zeigen Sie, dass das Vorgehen wie in 1. für die Cauchy-Verteilung nicht klappt.