Wahrscheinlichkeit und die Normalverteilung

Jonathan Harrington

Der Bevölkerungs-Mittelwert

99 Stück Papier nummeriert 0, 1, 2, …99

Ich ziehe 10 davon und berechne den Mittelwert.

Was ist der Mittelwert der von mir gezogenen Zahlen **im theoretischen Fall**? 49.5

Wir nennen diesen theoretischen Mittelwert den **Bevölkerungs-mittelwert** (population mean) und verwenden dafür das griechische Symbol μ.

$$\mu = 49.5$$

$\mu = 49.5$ bedeutet u.a.: ich bekomme diesen Wert bei diesem Vorgang **mit größter Wahrscheinlichkeit**.

Noch ein Beispiel…

Ich werfe einen Würfel *k* Mal (oder *k* Würfel gleichzeitig ein Mal). Ich berechne den Mittelwert der *k* Zahlen. Was ist μ?

μ = 3.5 `mean(1:6)`

Stichprobenmittelwert

Ich werfe einen Würfel *k* Mal (oder *k* Würfel gleichzeitig ein Mal). Ich berechne den Mittelwert der *k* Zahlen.

Wenn ich den obigen Vorgang tatsächlich für k = 10 durchführe, bekomme ich 10 **Zufallswerte,** z.B.

6 2 5 4 2 3 5 1 1 3

Der Mittelwert dieser **Stichprobe** wird (fast immer) etwas von μ abweichen: wir nennen diesen Durchschnitt den **Stichprobenmittelwert (sample mean)**, **m**

Fuer diesen Fall, **m** = 3.2 (und μ = 3.5)

(Zufalls)Stichproben in R

Eine Würfel werfen

```
sample(1:6, 1, replace=T)
```

10 Würfel werfen

```
sample(1:6, 10, replace=T)
```

Der Stichprobenmittelwert davon

```
mean(sample(1:6, 10, replace=T))
```

Ich will 50 solcherStichprobenmittelwerte bekommen

```
wuerfel <- NULL
for(j in 1:50){
ergebnis = mean(sample(1:6, 10, replace=T))
wuerfel = c(wuerfel, ergebnis)
}
```

```
wuerfel
```

```
3.1 3.9 3.6 4.2 2.8 3.3 4.6 2.9 4.2 3.1 3.7 4.3 4.1 4.5 4.0
4.9 2.6 3.3 3.6 4.2 3.6 4.0 2.9 3.6 3.1 3.3 4.9 3.2 2.9 2.7
3.5 3.2 1.9 4.2 4.6 3.7 3.9 4.4 3.5 3.4 3.2 3.5 3.5 3.1
3.4 4.3 3.0 3.3 3.7 3.0
```

Der **Mittelwert der Stichprobenmittelwerte** ist ziemlich nah an μ

```
mean(wuerfel)
[1] 3.588
```

Je mehr Stichprobenmittelwerte, umso mehr nähert sich dessen Mittelwert μ

```
# 5000 Stichprobenmittelwerte
wuerfel <- NULL
for(j in 1:5000){
ergebnis = mean(sample(1:6, 10, replace=T))
wuerfel = c(wuerfel, ergebnis)
}

mean(wuerfel)
[1] 3.50812
```

sodass wenn wir **unendlich viele** Stichprobenmittelwerte hätten, wäre der Mittelwert davon **genau** μ

Stichprobenmittelwerte in R erzeugen

Vier Variablen:

A. Die Reichweite der ganzen Zahlen (zB beim Würfel 1, 6). unten, oben

B. k: Wieviele Würfel werfen wir zusammen (oder wieviel Stück Papier ziehen wir aus dem Hut)?

C. N: wie oft wiederholen wir Vorgang B?

```
proben <- function(unten=1, oben = 6, k = 10, N = 50)
{
# default: wir werfen 10 Wuerfel 50 Mal
alle <- NULL
for(j in 1:N){
ergebnis = mean(sample(unten:oben, k, replace=T))
alle = c(alle, ergebnis)
}
alle
}
```

100 Stück Papier nummeriert 0, 1, 2, …99 in einem Hut.

A. Ich ziehe 8 davon und berechne den Mittelwert, und tue sie wieder in den Hut rein.

Was ist μ? 49.5

B. Die Funktion proben() verwenden, um für A. 50 Stichprobenmittelwerte zu bekommen. Diese 50 Werte in einem Vektor Speichern. Den Mittelwert davon berechnen.

C. Nochmals A und B wiederholen, diesmal um 500 Stichprobenmittelwerte zu bekommen.

Ist die zweite Berechnung näher an 49.5?

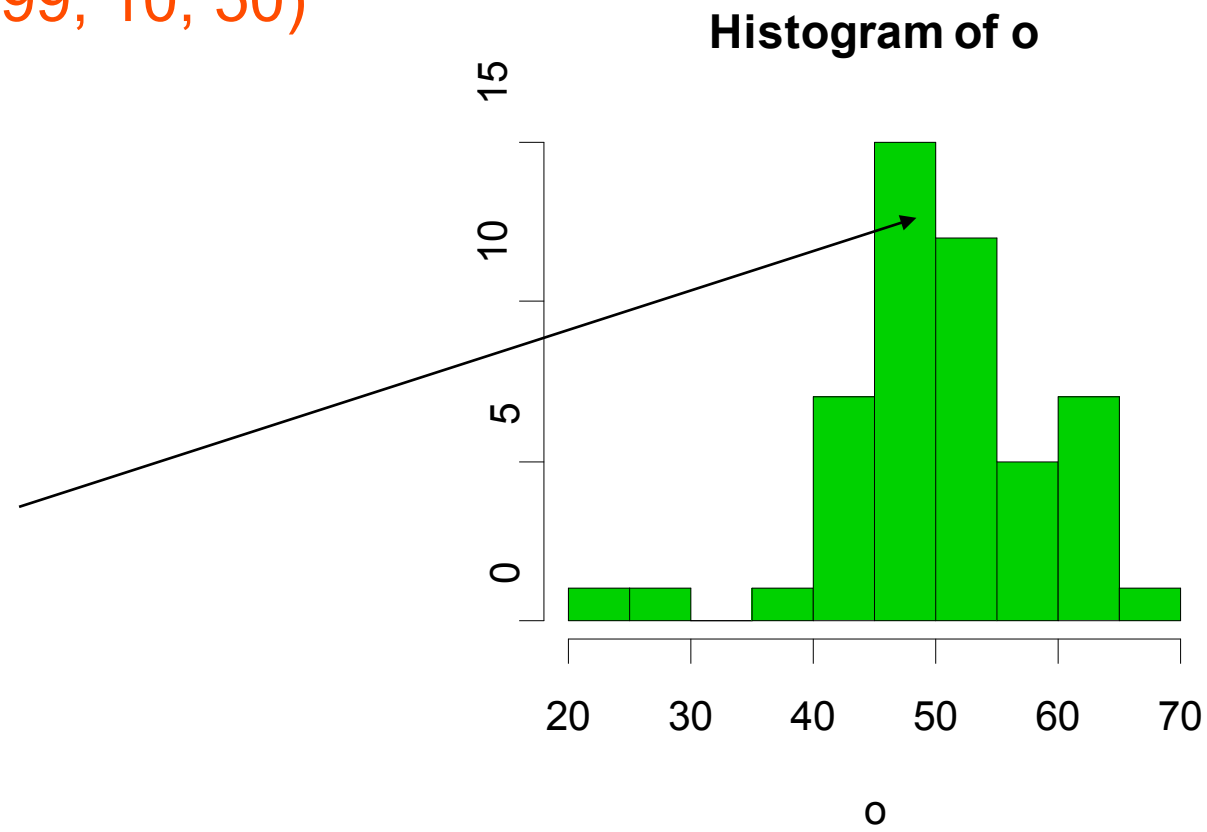
Die Verteilung der Stichprobenmittelwerte

kann man grob mit einem **Histogramm** sehen.

Hut mit Zahlen, 0-99; ich ziehe 10, berechne den Stichprobenmittelwert, wiederhole das 50 Mal.

```
o = proben(0, 99, 10, 50)
hist(o, col=3)
```

15 **m** Werte lagen zwischen 45 und 50

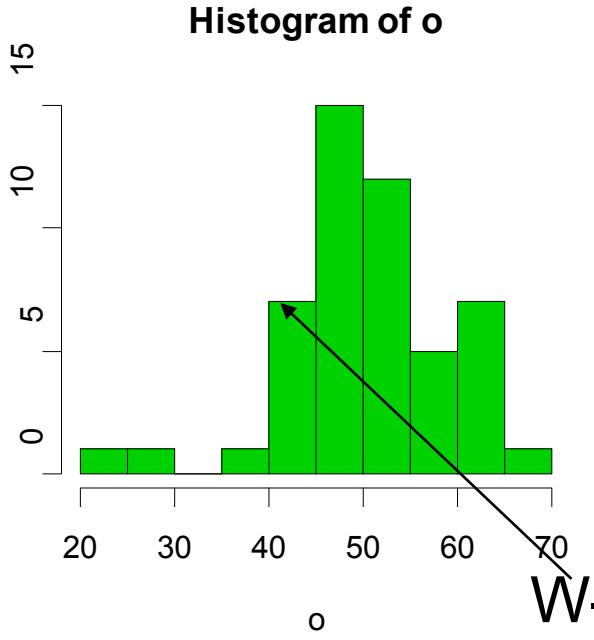


Die Wahrscheinlichkeitsdichte

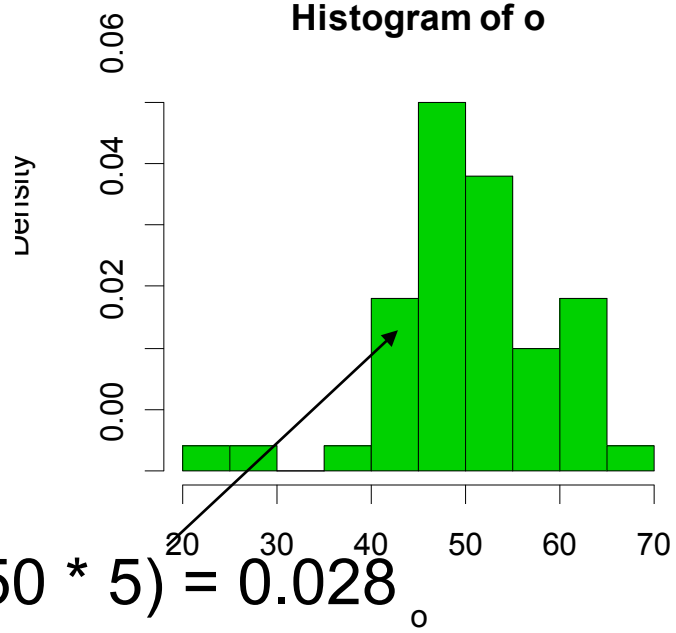
Die **Wahrscheinlichkeitsdichte** (probability density) ist eine Umstellung der Häufigkeit, sodass die **Balken-Flächensumme** im Histogramm 1 (eins) ist.

W-Dichte = Häufigkeit/(N x Balkenbreite)

hist(o, col=3)



hist(o, col=3, freq=F)

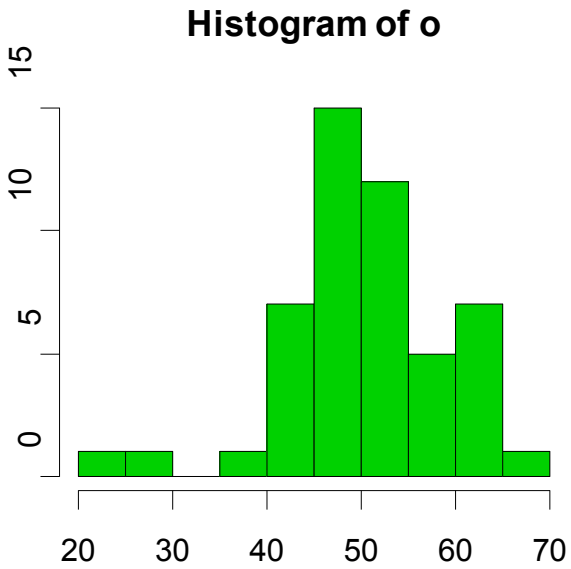


W-Dichte = 7/(50 * 5) = 0.028

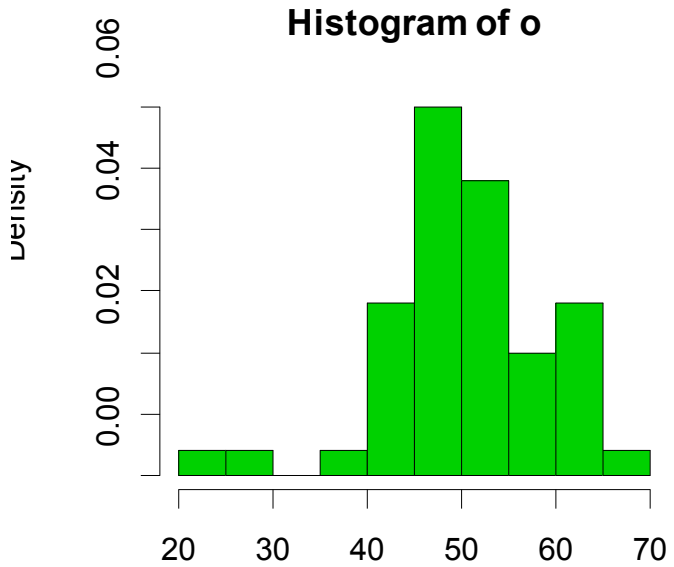
Die Fläche von diesem Balken ist 5 * 0.028 = 0.14. Daher liegen 14% der Werte zwischen 40 und 45.

Die Wahrscheinlichkeitsdichte

hist(o, col=3)



hist(o, col=3, freq=F)



Σ Wahrscheinlichkeitsdichten x Balkenbreiten = 1

```
h =hist(o, col=3, freq=F)
sum(h$density * 5)
```

```
[1] 1
```

Die Normalverteilung

ist ein 'Histogramm' (mit W-Dichten auf der y-Achse), der unter zwei Bedingungen erstellt wird:

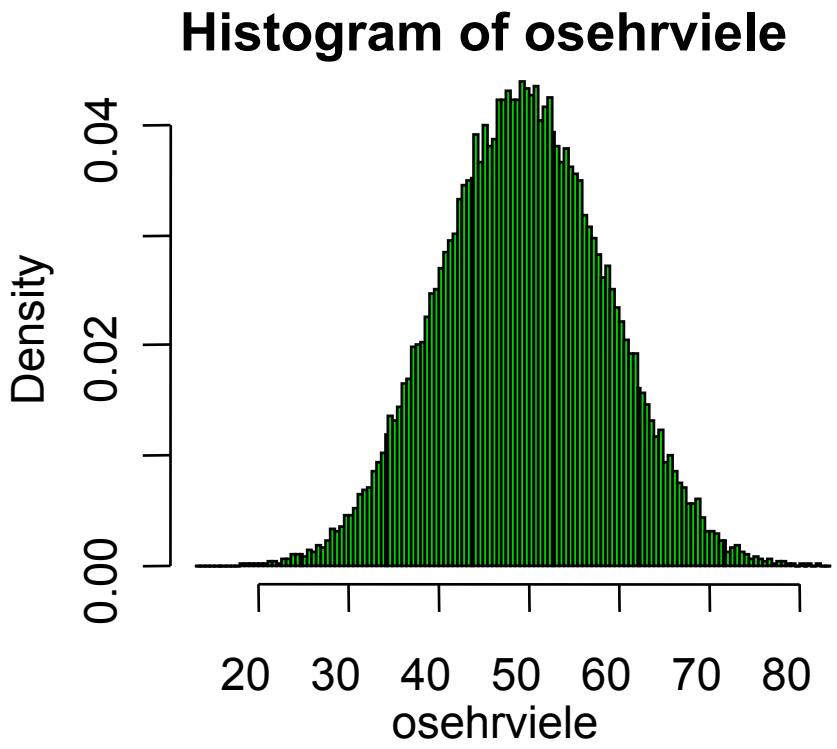
(a) der Vorgang (um Stichprobenmittelwerte zu bekommen) wiederholt sich nicht 50 sondern **unendlich viel** Mal.

(b) wir lassen mit zunehmenden Stichproben **die Balkenbreite immer kleiner werden**, sodass im unendlichen Fall die Balkenbreite unendlich klein ist (= 0 also wird die Balkenfläche zu einer Linie). Daher haben wir keine Stufen mehr (von einem Balken zum nächsten) sondern **eine glatte Kurve**.

Normalverteilung simulieren

Wir können das teilweise mit der proben() Funktion simulieren. Hier haben wir 50000 Stichprobenmittelwerte und 200 Balken und eine Balkenbreite von 0.5*

```
osehrviele = proben(0, 99, 10, 50000)
h4 = hist(osehrviele, col=3, freq=F, breaks=200)
```



*(wird durch 1/sum(h4$density) ermittelt)

Die Normalverteilung berechnen

Die Normalverteilung kann mit einer Formel (die wir später besprechen werden) berechnet werden, in der nur zwei Variablen gesetzt werden müssen.

- Der Bevölkerungs-mittelwert, μ
- Die Bevölkerungs-Standardabweichung, σ

Die Bevölkerungs-Standardabweichung, σ

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \mu^2}$$

zB für den Würfel ist x 1, 2, 3, 4, 5, 6 und n = 6

Was ist σ? (in R berechnen)

```
unten = 1
oben = 6
 x =     unten:oben
 n =     length(x)
 mu =    mean(x)
sigma =      sqrt((sum(x^2)/n - mu^2))
 sigma
```

```
[1] 1.707825
```

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \mu^2}$$

in eine Funktion sigma(x) umsetzen.

```
sigma <- function(unten=1, oben=6)
{
x = unten:oben
n = length(x)
m = mean(x)
sqrt((sum(x^2)/n - m^2))
}
    sigma()
    [1] 1.707825
```

Die Bevölkerungs-Standardabweichung, σ

Dies ist σ wenn wir **einen** Würfel werfen.

```
sigma()
[1] 1.707825
```

Bedeutung: dies ist die Standardabweichung von den Zahlen (1-6) eines unendlich viel Mal geworfenen Würfels.

Die Bevölkerungs-Standardabweichung, σ

Wichtig!! Wenn wir k Würfel werfen, und den Mittelwert der Zahlen berechnen, dann ist die Bevölkerungsstandardabweichung (genannt auch 'the standard error of the mean') dieselbe wie für einen Würfel **aber durch $\sqrt{k}$ dividiert.**

sigma()/sqrt(7)

Bevölkerungs-Standardabweichung (Standard error of the mean) in R wenn wir 7 Würfel werfen, und davon den Mittelwert berechnen.

Bedeutung: dies ist die Standardabweichung der (unendlich vielen) Mittelwerte von 7 Zahlen, die ich bekomme, wenn ich unendlich viel Mal 7 Würfel werfe (und bei jedem Wurf den Mittelwert berechne).

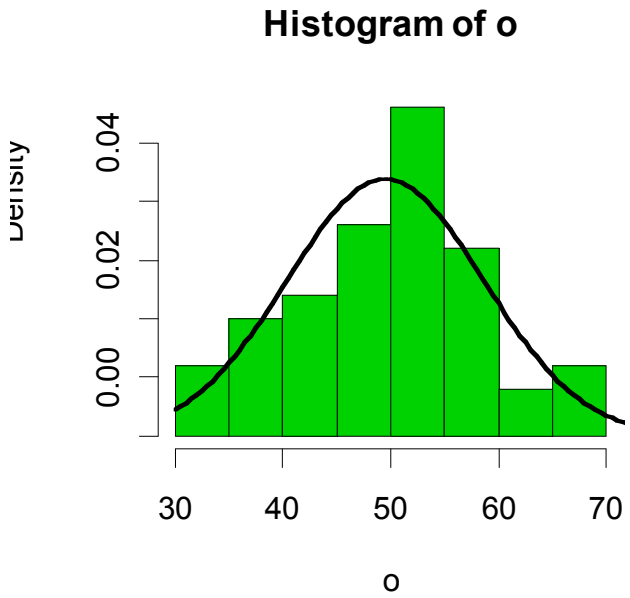
Ich ziehe 10 Stück Papier aus einem Hut mit Zahlen 0 bis 99. σ (standard error of the mean) in R = ….

```
sigma(0, 99)/sqrt(10)          [1] 9.128253
```

Normalverteilung auf Histogramm überlagern

Hut mit Zahlen, 0-99; ich ziehe 10, berechne den Stichprobenmittelwert, wiederhole das 50 Mal.

```
o = proben(0, 99, 10, 50)
hist(o, col=3, freq=F)
```



Normalverteilung überlagern

μ

```
mu = mean(0:99)
```

σ

```
sig = sigma(0,99)/sqrt(10)
```

```
curve(dnorm(x, mu, sig), 30, 80, add=T)
```

Je mehr Stichproben, umso besser die Anpassung an die Normalverteilung

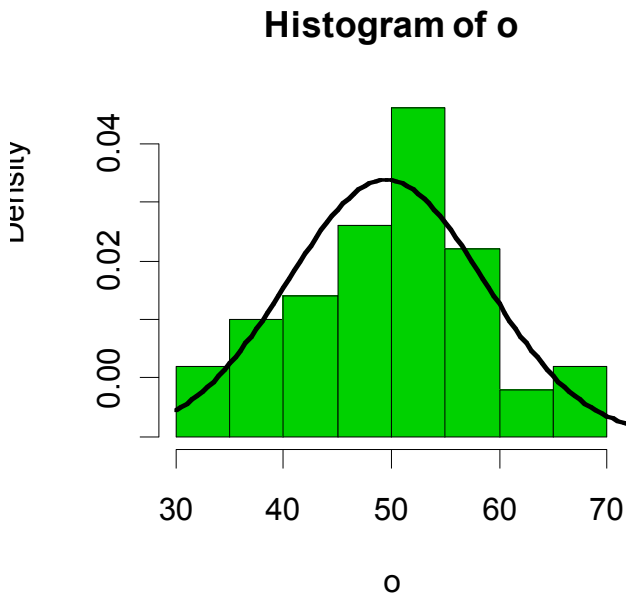
o = proben(0, 99, 10, 50)

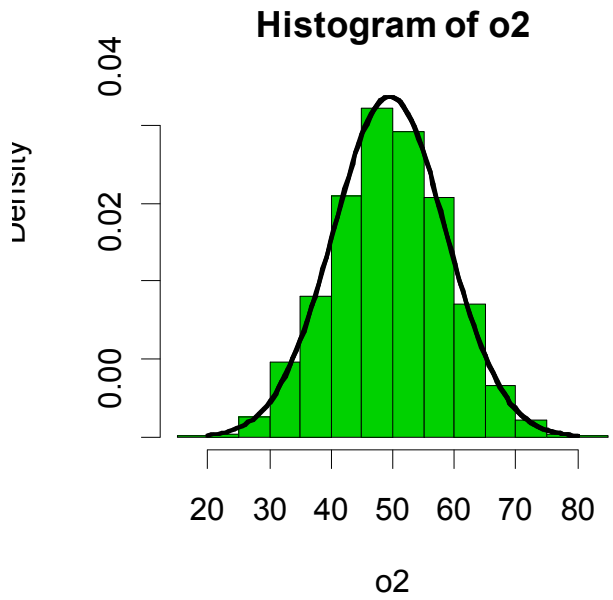
hist(o, col=3, freq=F)

o2 = proben(0, 99, 10, 5000)

hist(o2, col=3, freq=F)

curve(dnorm(x, mu, sig), 30, 80, add=T)

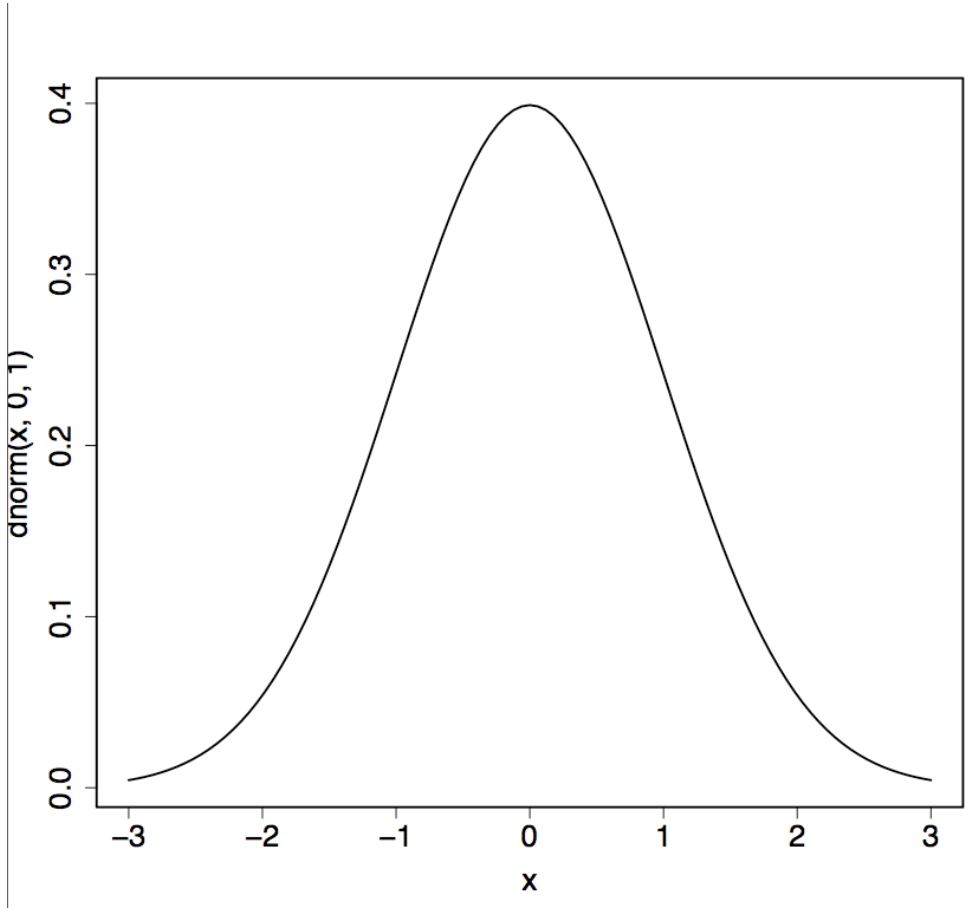




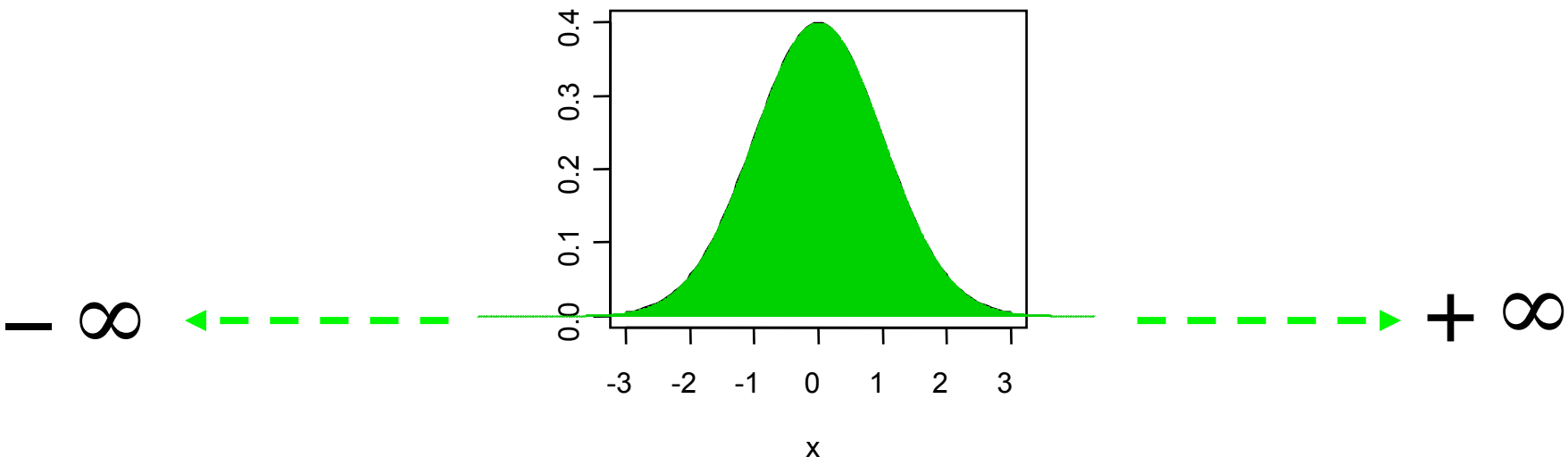
Normalverteilung abbilden

zB $\mu = 0$, $\sigma = 1$, zwischen -3 und +3

curve(dnorm(x, 0, 1), -3, 3)



Einige Merkmale der Normalverteilung



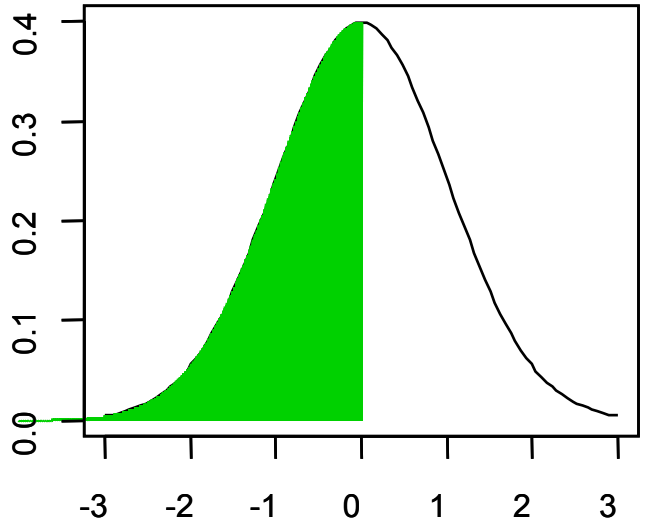
Die maximale W-Dichte liegt bei μ (in diesem Fall bei 0)

Es ist wichtig zu bemerken, dass es W-Dichten-Werte gibt (die immer kleiner werden) **bis ins Unendliche** in beiden Richtungen.

Normalverteilungen und Flächen

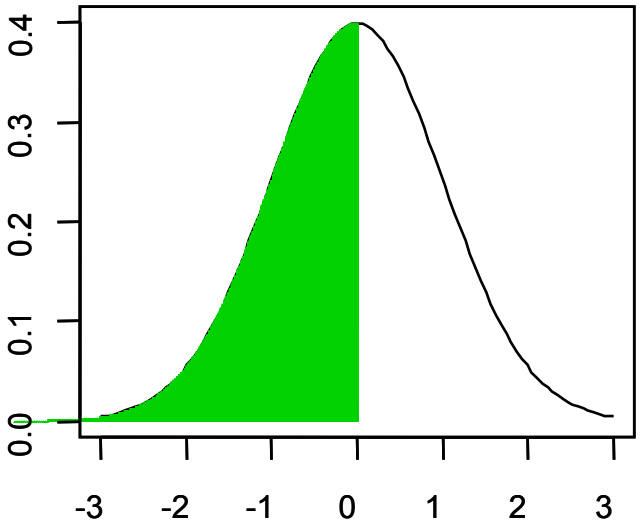
Die Gesamtfläche unter einer Normalverteilung ist 1

Die Fläche zwischen $-\infty$ und μ ist daher immer: 0.5



Die Bedeutung davon: wenn wir eine Stichprobe aus einer Normalverteilung mit $\mu = 0$ und $\sigma = 1$ entnehmen, dann ist die Wahrscheinlichkeit 0.5 (50%), dass unsere Stichprobe unter 0 liegt

Flächensummierung einer Normalverteilung in R



In R erfolgt die Flächensummierung zwischen $-\infty$ und einem Wert, w, fuer eine Normalverteilung mit Parametern (μ, σ) durch pnorm(w, μ, σ)

Daher ist die Fläche bis μ für den Fall oben

```
pnorm(0, 0, 1)
[1] 0.5
```

($\mu = 0, \sigma = 1$)

Noch zwei Beispiele…

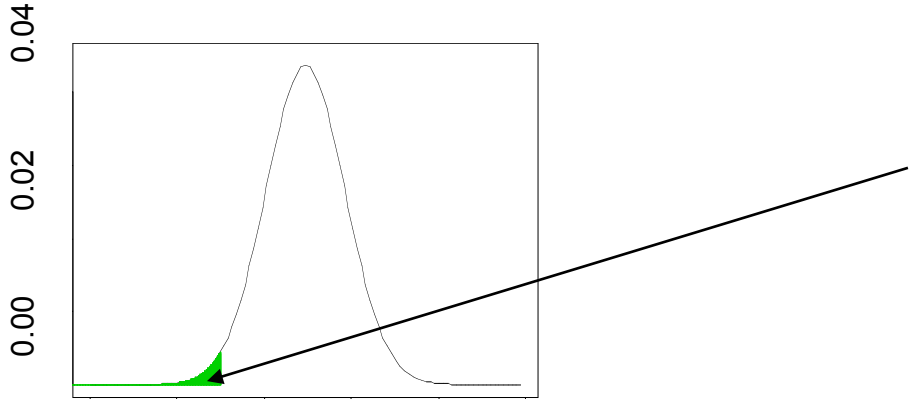
Ich ziehe 10 Stück Papier aus einem Hut mit Zahlen 0 bis 99. Ich berechne den Mittelwert davon. Was ist die Wahrscheinlichkeit, dass dieser Mittelwert (a) unter 30 (b) unter 60 liegt?

μ mu = mean(0:99)

σ sig = sigma(0,99)/sqrt(10)

Normalverteilung abbilden zwischen 0 und 99

curve(dnorm(x,mu, sig), 0, 99)



(a) pnorm(30, mu, sig)

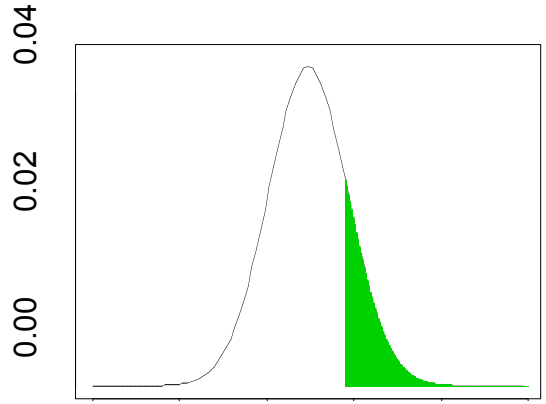
[1] 0.01633055

(b) pnorm(60, mu, sig)

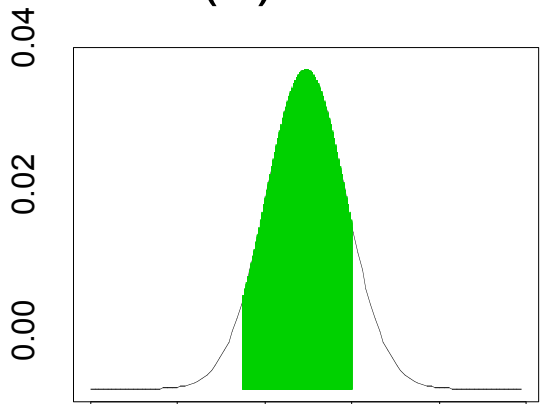
[1] 0.8749847

…Was ist die Wahrscheinlichkeit, dass dieser Mittelwert (c) über 58 (d) zwischen 35 und 60 liegt?

(c)



(d)




```
1 - pnorm(58, mu, sig)
```

```
[1] 0.1758815
```

```
pnorm(60, mu, sig) - pnorm(35, mu, sig)
```

```
[1] 0.8188952
```

Eine Normalverteilung und das Vertrauensintervall

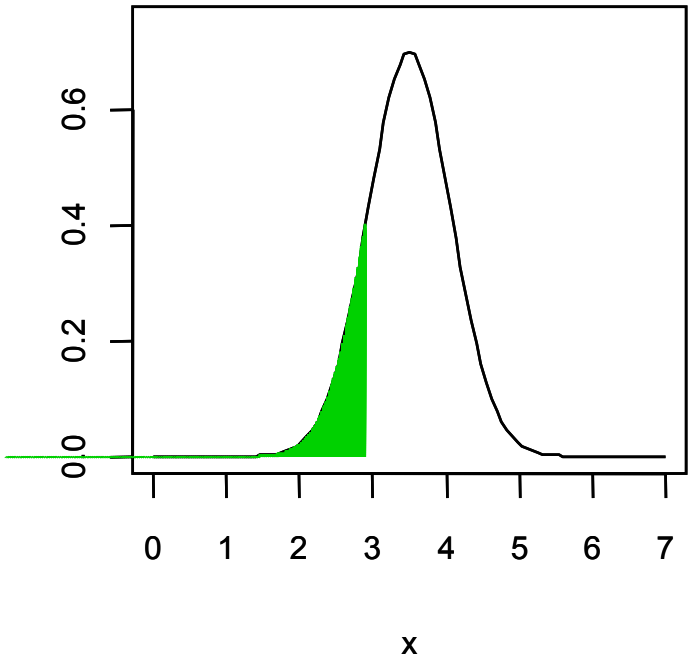
- In einer Normalverteilung weichen die Werte ab **im Verhältnis zur Standardabweichung**.

- Wahrscheinlichkeiten (Flächen unter der Normalverteilung) können in **Standardabweichungen vom Mittelwert** umberechnet werden). In R: qnorm()

- Mit qnorm() können wir ein **Vertrauensintervall (Konfidenzintervall)** setzen.

qnorm() und Standardabweichungen

Was ist die Wahrscheinlichkeit, dass ich einen Mitelwert von 2.9 oder weniger bekomme, wenn ich 9 Würfel werfe?




```
mu =      mean(1:6)
SE =      sigma(1, 6)/sqrt(9)
pnorm(2.9, mu, SE)
[1] 0.1459479
```

Das sind wieviele Standardabweichungen von μ?

```
qnorm(0.1459479)
-1.053972
```

Daher bekommen wir wieder 2.9 durch:

```
mu + qnorm(0.1459479) * SE
[1] 2.9
```

Das Vertrauensintervall

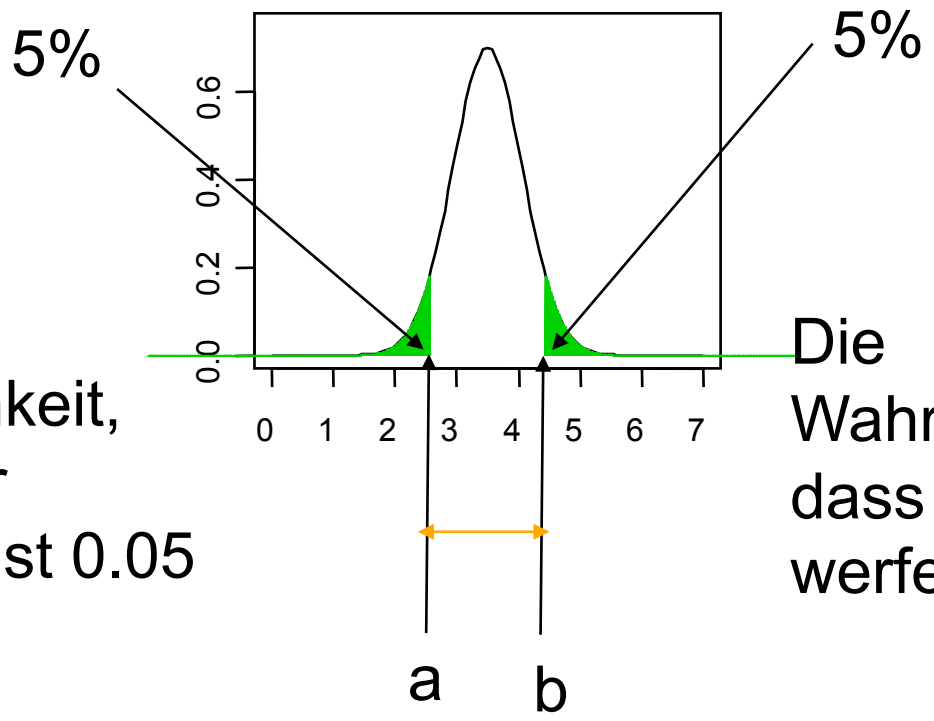
Ich kaufe 9 Würfel in einem Spielgeschäft. Ich werfe die Würfel, und berechne den Mittelwert. Sollte die Wahrscheinlichkeit dieses Mittelwertes unter 0.05 (5%) liegen, dann klage ich den Händler an (weil er gezinkte Würfel verkauft).

Innerhalb von welchem Bereich muss der Zahlenmittelwert liegen, damit der Händler nicht angeklagt wird?

μ = 3.5

Standard-Abweichung von μ (Standard error) = sigma(1,6)/sqrt(9)

Die Wahrscheinlichkeit, dass ich *a* oder weniger werfe ist 0.05



Die Wahrscheinlichkeit, dass ich *b* oder mehr werfe ist 0.05

oder Die Wahrscheinlichkeit, dass ein Wert zwischen *a* und *b* liegt = 0.90. (**Ein 90% Vertrauensintervall**)

Was ist (a)?

```
qnorm(0.05, mu, SE)
2.563626
```

Was ist (b)?

```
qnorm(0.95, mu, SE)
4.436374
```

oder

```
mu + qnorm(0.05) * SE
```