Wahrscheinlichkeit und die Normalverteilung

Jonathan Harrington

Der Populations-Mittelwert

100 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 **Populations-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 einmal 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 Mittelwert 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

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

Stichprobenmittelwerte in R erzeugen

Vier Variablen:

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

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 Würfel 50 Mal
alle <- NULL
for(j in 1:N){
ergebnis = mean(sample(unten:oben, k, replace=T))
alle = c(alle, ergebnis)
}
alle
}
```

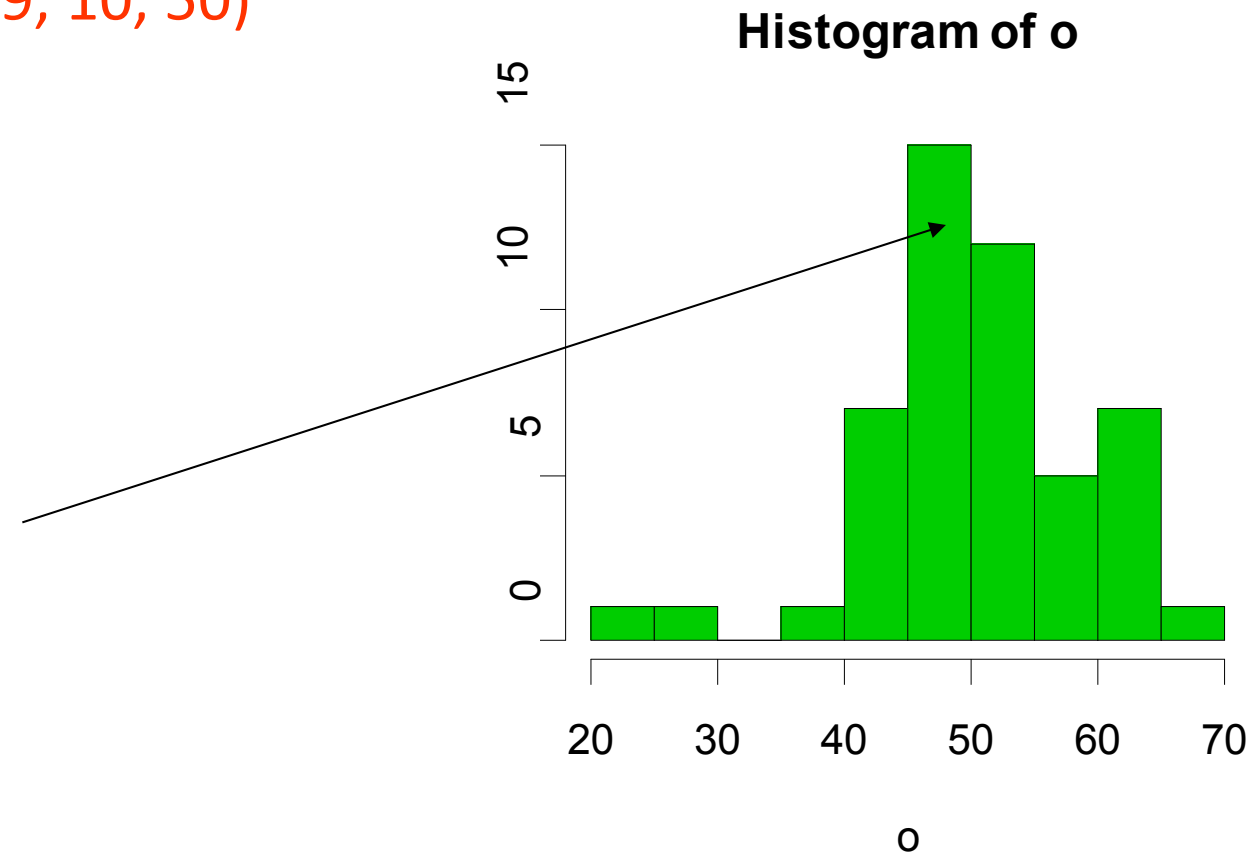
Die Verteilung der Stichprobenmittelwerte

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

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

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

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

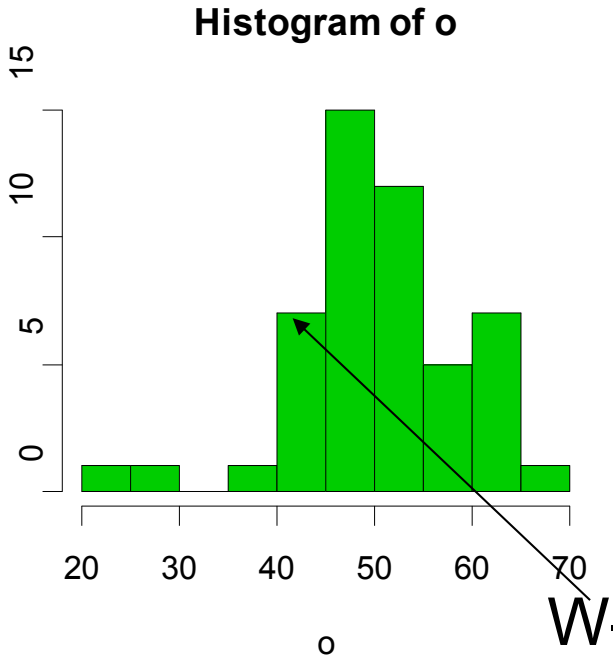


Die Wahrscheinlichkeitsdichte

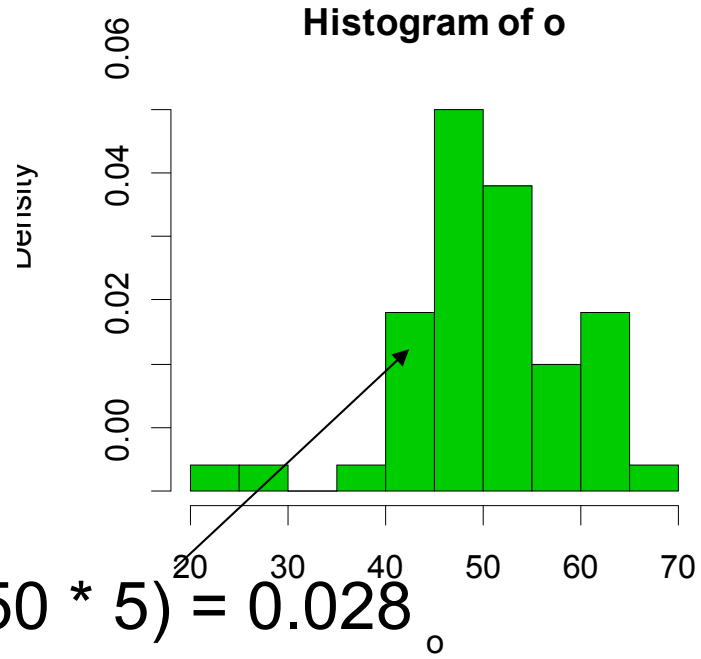
Die **Wahrscheinlichkeitsdichte** (probability density) ist eine Umstellung der Häufigkeit, sodass die **Balken-Flächensumme** im Histogramm 1 (eins) ist.

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 Normalverteilung

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

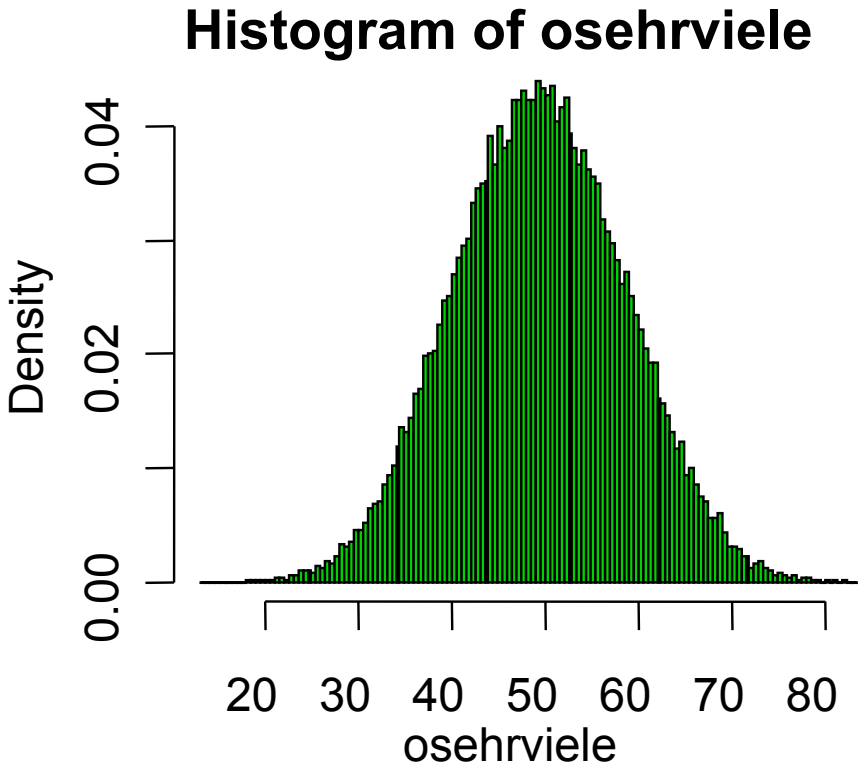
(a) der Vorgang (um Stichprobenmittelwerte zu bekommen) wiederholt sich nicht 50 sondern **unendlich viel** Mal.

(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

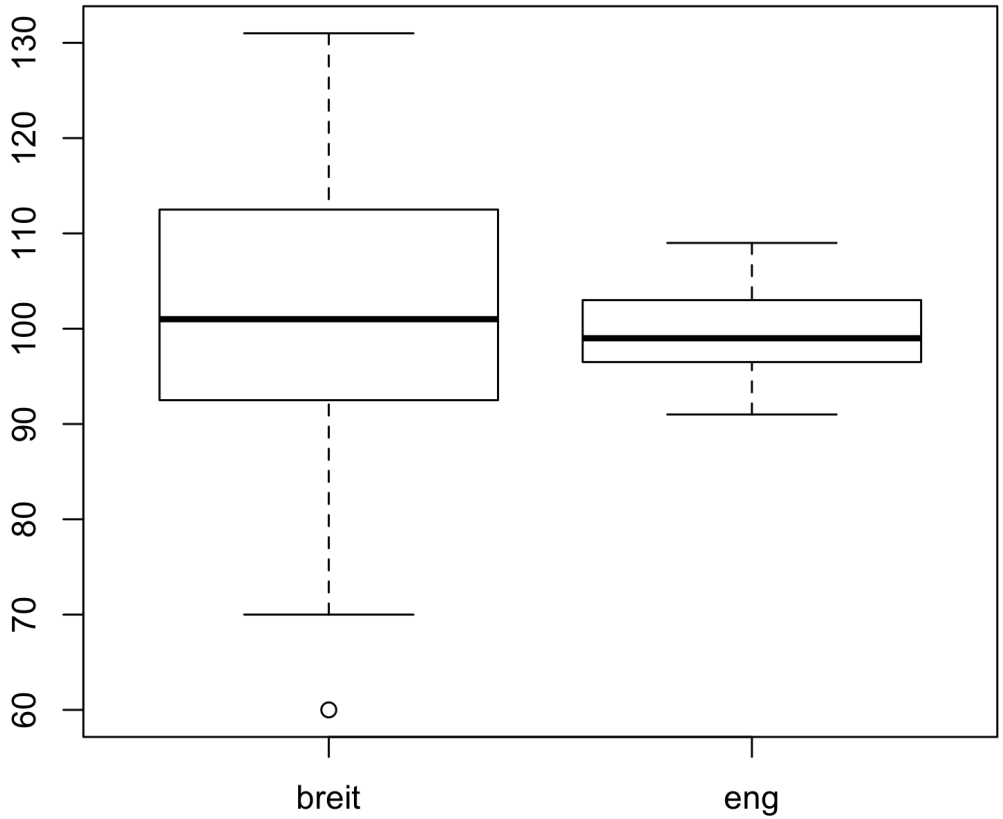
Die Normalverteilung kann mit einer Formel berechnet werden, in der nur zwei Variablen gesetzt werden müssen.

- Der Populations-Mittelwert, μ
- Die Populations-Standardabweichung, σ

  Die Standardabweichung misst wie groß die Streuung um den Mittelwert ist

Die Standardabweichung

```
nex = read.table(file.path(pfadu, "normexample.txt"))
boxplot(werte ~ Verteilung, data=nex)
```



Die Standardabweichung **einer Stichprobe** wird mit sd() in R berechnet:

```
aggregate(nex$werte, list(nex$Verteilung), sd)
```

```
breit      17.3642102
eng         4.739531
```

Die Standardabweichung

Die Populations-Standardabweichung, σ, weicht etwas von der Stichprobenstandardabweichung ab (vor allem wenn *n*, die Anzahl der Stichproben klein ist) und wird mit folgender Formel berechnet:

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

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
```

Die Populations-Standardabweichung, σ

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

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))
 }
```

Die Populations-Standardabweichung, σ

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

Bedeutung: dies ist die Standardabweichung der Werte eines unendlich viel Mal geworfenen Würfels (wenn die Stichprobe unendlich groß ist).

Der Standard-Error (SE)

ist die Populations-Standardabweichung **von Mittelwerten**

k Würfel werfen, den Mittelwert, m_1, berechen.

Diesen Vorgang unendlich viel Mal wiederholen (jedes Mal den Mittelwert der *k* Würfel berechnen).

Wir bekommen dadurch unendlich viele Mittelwerte, m_1, m_2, m_3 ...

Die Standardabweichung dieser undendlich vielen Mittelwerte, genannt SE, wird mit:

sigma()/sqrt(k)

berechnet, wo *k* die Anzahl der Würfel ist, deren Mittelwert wir berechnen.

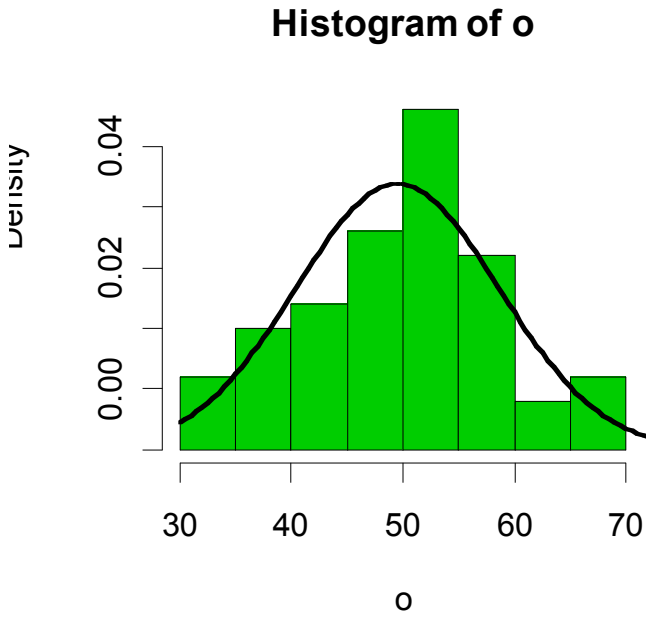
Ich ziehe 10 Stück Papier aus einem Hut mit Zahlen 0 bis 99 berechne den Mittelwert, m_1, wiederhole diesen Vorgang unendlich viel Mal, bekomme daher unendlich viele Mittelwerte. Was ist SE in R?

```
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)
```

SE

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

```
curve(dnorm(x, mu, SE), add=T)
```

Je mehr Stichproben, umso besser die Anpassung an die Normalverteilung

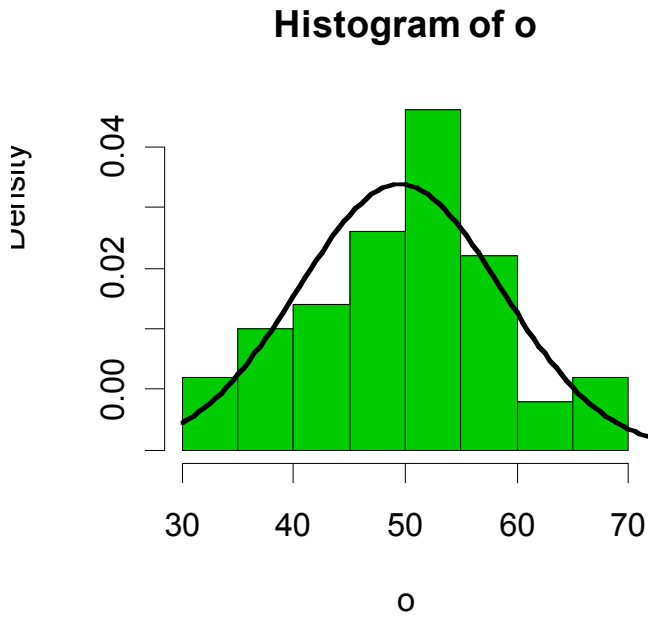
o = proben(0, 99, 10, 50)

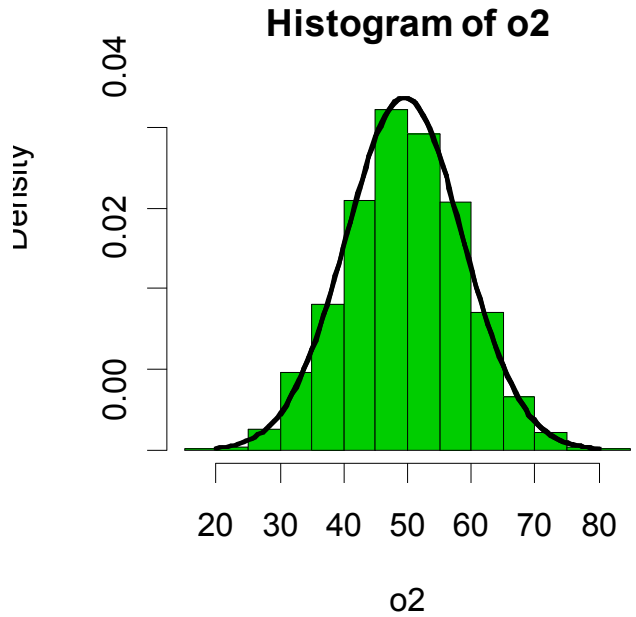
hist(o, col=3, freq=F)

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

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

curve(dnorm(x, mu, SE), add=T)





Berechnung von Wahrscheinlichkeiten

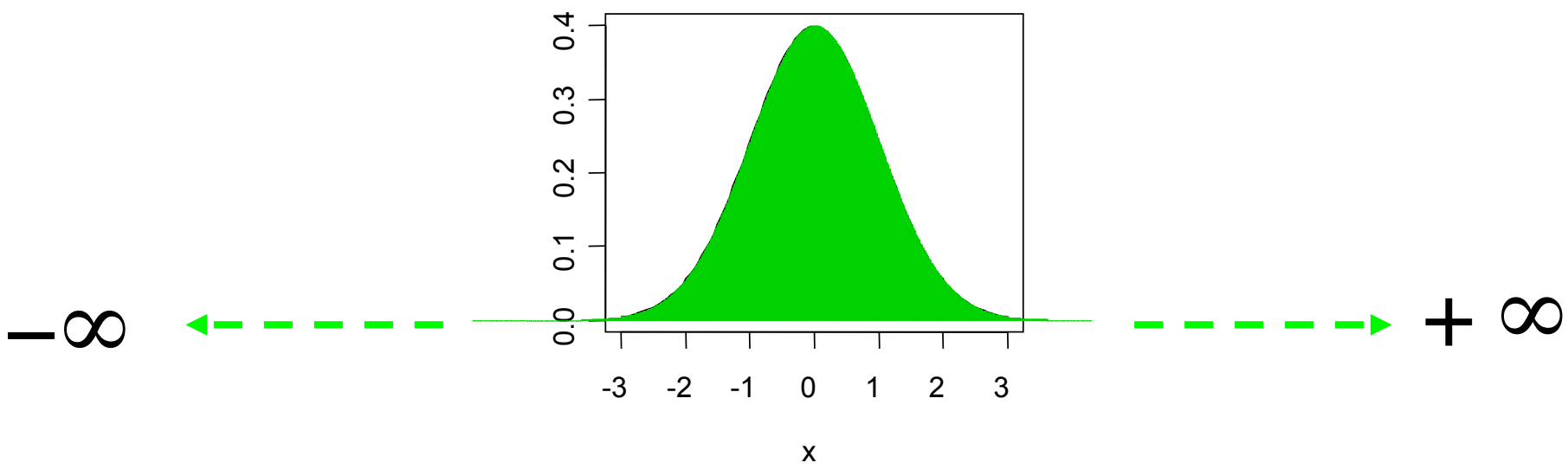
z.B.

Wenn ich 10 Stück Papier aus einem Hut mit Zahlen 0-99 ziehe, was ist die Wahrscheinlichkeit, dass der Mittelwert z.B. unter 38 liegt, über 76, zwischen 30-65 usw.

Solche Wahrscheinlichkeiten werden durch **die proportionale Fläche unter der Normalverteilung** berechnet.

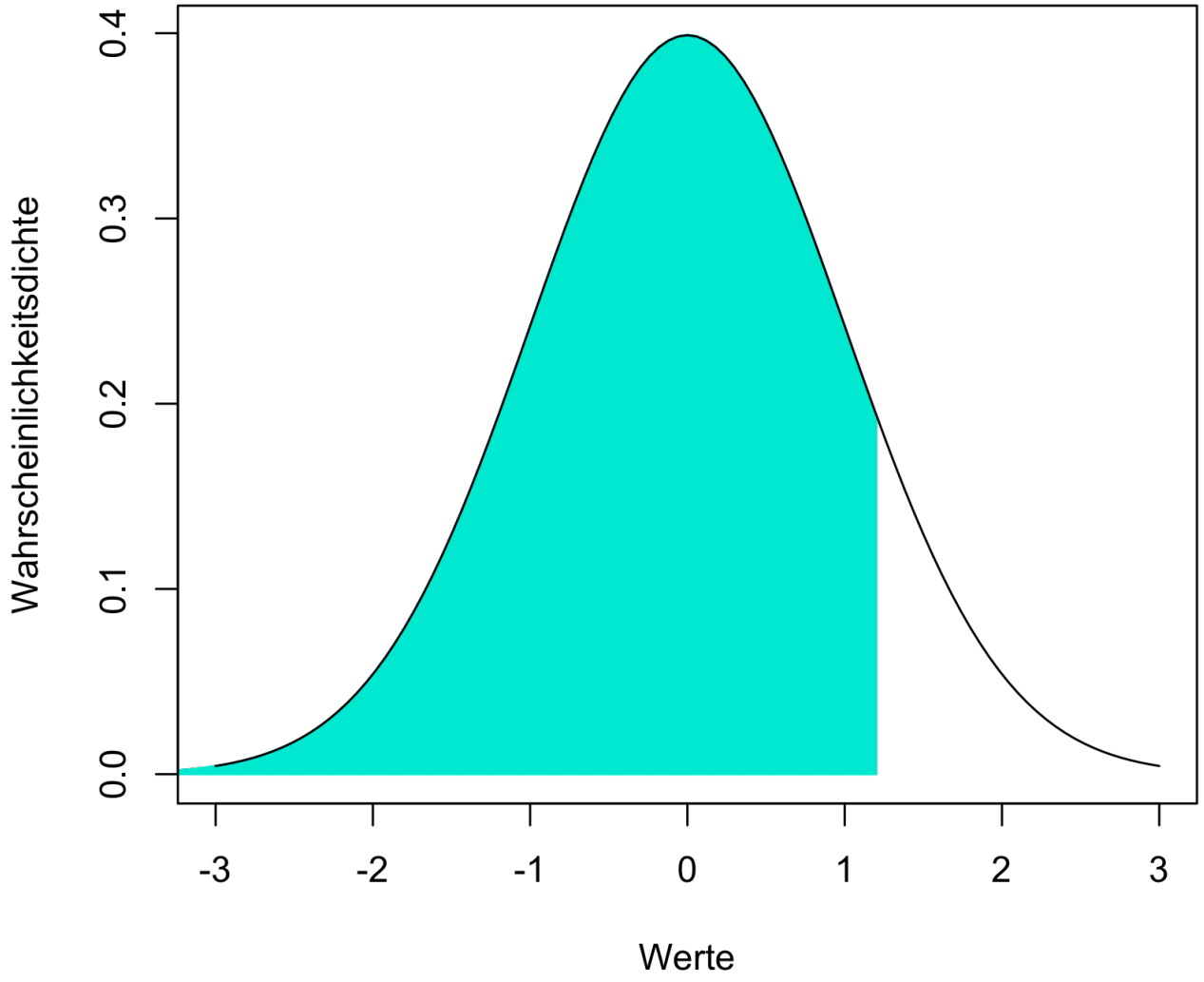
Die Fläche unter einer Normalverteilung

Die Fläche unter jeder Normalverteilung zwischen ±∞ ist immer 1 (eins)

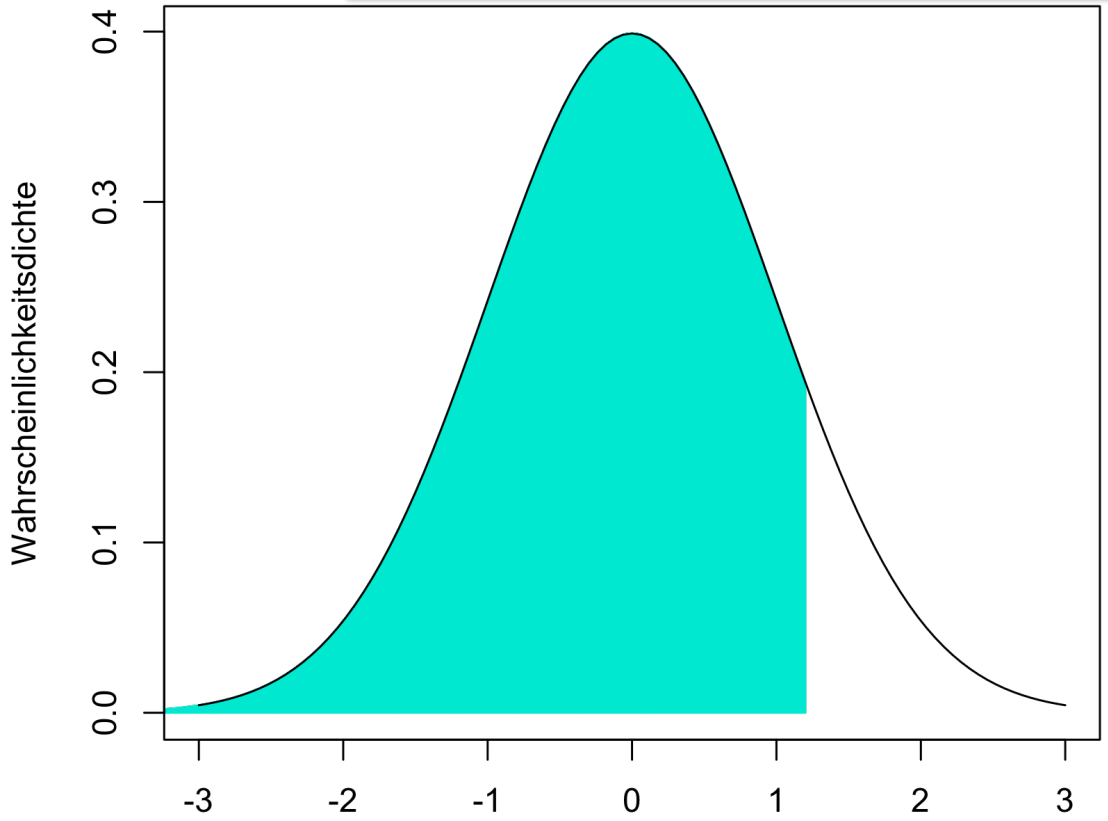


Die Flächensummierung

pnorm(x) summiert die Fläche unter einer Normalverteilung zwischen -∞ und einen Wert x (per Default in einer Normalverteilung mit $\mu = 0$ und $\sigma = 1$).



Die Flächensummierung



pnorm(1.2)

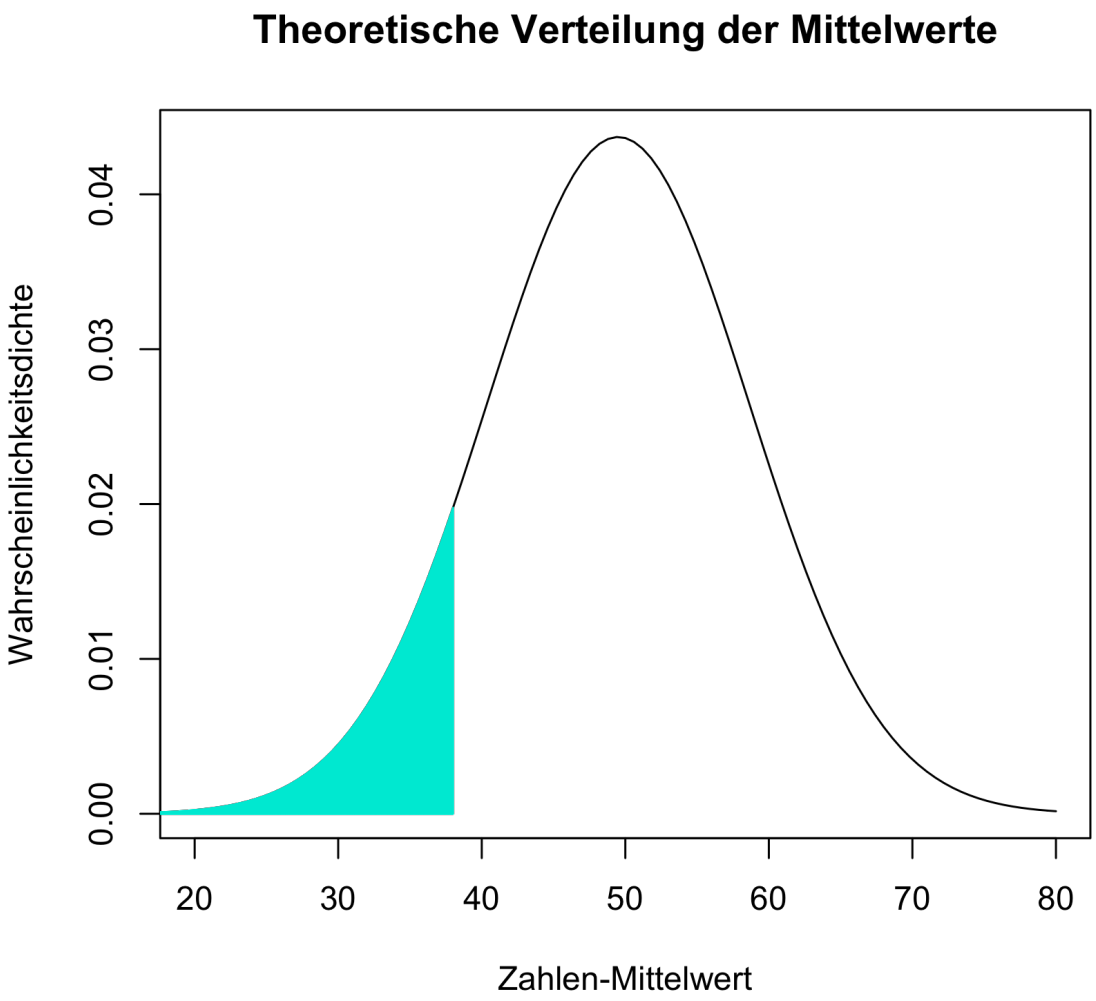
```
[1] 0.8849303
```

Die Bedeutung: die Wahrscheinlichkeit, dass ich einen Wert weniger als 1.2 bekomme in dieser Normalverteilung ist 0.8849303

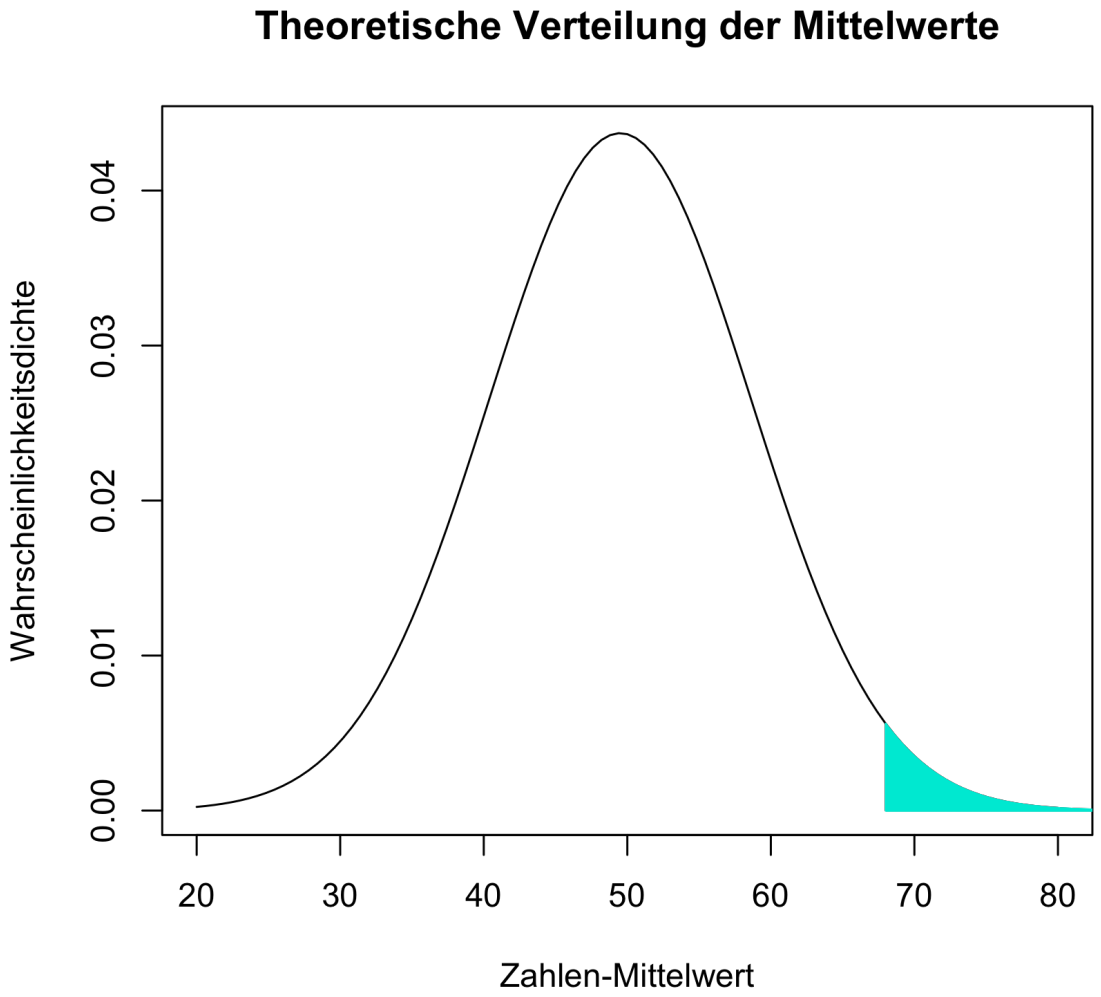
Wenn ich 10 Stück Papier aus einem Hut mit Zahlen 0-99 ziehe, was ist die Wahrscheinlichkeit, dass der Mittelwert unter 38 liegt?

```
mu = mean(0:99)

SE = sigma(0, 99)/sqrt(10)

pnorm(38, mu, SE)

[1] 0.1038663
```

ca. 10%. (kommt ca. 1/10 Mal vor).

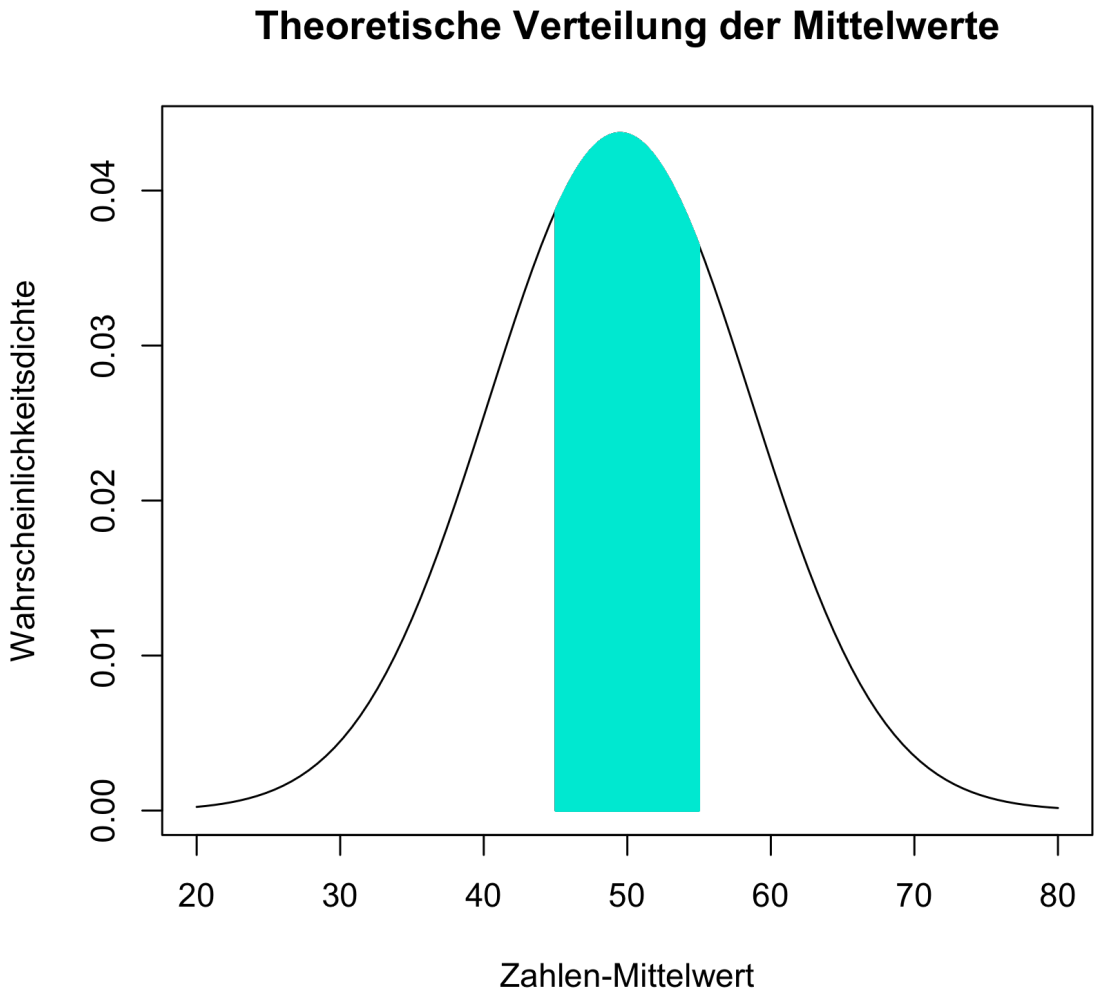


Wenn ich 10 Stück Papier aus einem Hut mit Zahlen 0-99 ziehe, was ist die Wahrscheinlichkeit, dass der Mittelwert über 68 liegt?




```
1 - pnorm(68, mu, SE)
[1] 0.02134784
```

Wenn ich 10 Stück Papier aus einem Hut mit Zahlen 0-99 ziehe, was ist die Wahrscheinlichkeit, dass der Mittelwert zwischen 45 und 55 liegt?




```
pnorm(55, mu, SE) - pnorm(45, mu, SE)
```

```
0.4155725
```

Konfidenzintervall

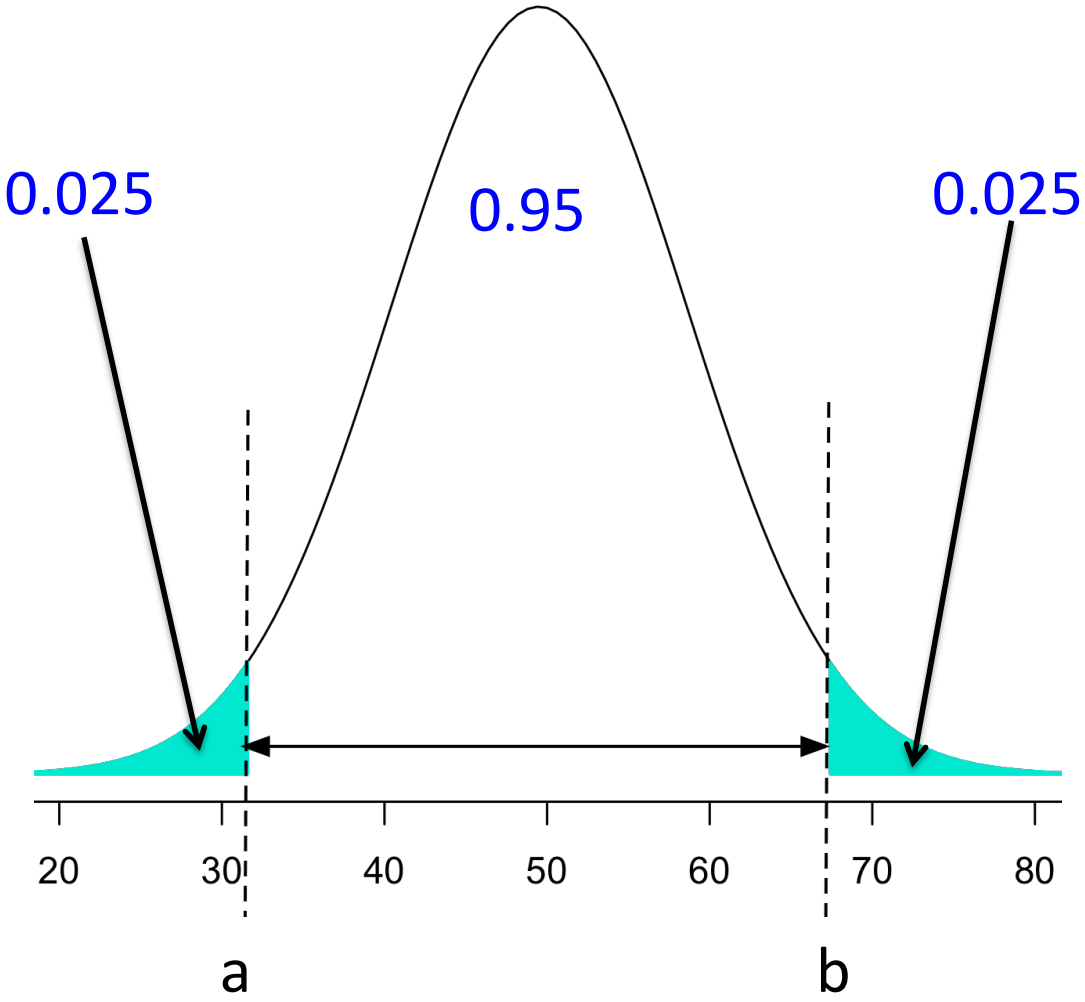
Ich ziehe 10 Stück Papier aus einem Hut mit Zahlen 0-99 und berechne den Mittelwert.

Ich will zwei Werte berechnen, *a* und *b*, sodass der Stichprobenmittelwert zwischen *a* und *b* mit einer Wahrscheinlicheit von z.B. 0.95 liegt (95% Konfidenzintervall).

Konfidenzintervall und qnorm()

Die Wahrscheinlichkeit, dass der Mittelwert zwischen *a* und *b* fällt ist 0.95. Was sind *a* und *b*?

Flächen unter der Normalverteilung




```
mu = mean(0:99)

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

```
qnorm(0.025, mu, SE)
 [1] 31.60895
qnorm(0.975, mu, SE)

 [1] 67.39105
```

Konfidenzintervall und qnorm()

Ich ziehe 10 Stück Papier aus einem Hut mit Zahlen 0-99 und berechne den Mittelwert. Was ist der 95% Konfidenzintervall für den Stichprobenmittelwert?

95% Konfidenzintervall: $31.6 < m < 67.4$

= der Stichprobenmittelwert liegt zwischen 31.6 und 67.4 mit einer Wahrscheinlichkeit von 0.95.

99% Konfidenzintervall: $26.0 < m < 73.0$