# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

*Marcel Dettling*

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

Winterthur, 10. Mai 2017

# *Zufallsvariablen: Diskret oder Stetig?*

- *Eine diskrete Zufallsvariable X liegt dann vor, wenn sie endlich oder unendlich abzählbar viele Werte annimmt.*

  **Beispiele**: - überall, wo gezählt wird
  - Fahrzeuge, Krankheitsfälle, Ausschuss, ...

- *Eine stetige Zufallsvariable X liegt dann vor, wenn sie jeden beliebigen Wert eines Intervalls annehmen kann.*

  **Beispiele**: - überall, wo gemessen wird
  - Wartezeit, Hämatokrit, Regenmenge, ...

# *Wahrscheinlichkeitsverteilungen*

Eine Wahrscheinlichkeitsverteilung gibt an, welche Werte eine Zufallsvariable mit welcher Wahrscheinlichkeit annimmt.

Je nach Ausprägung der Zufallsvariablen (diskret/stetig) gibt es die entsprechenden Wahrscheinlichkeitsverteilungen

- ***Diskrete Wahrscheinlichkeitsverteilungen***
- ***Stetige Wahrscheinlichkeitsverteilungen***

Die diskrete Wahrscheinlichkeitsfunktion ist bekanntlich das Stabdiagramm. Es stellt sich nun die Frage, wie sich das Konzept auf stetige Zufallsvariablen verallgemeinern lässt.

# *Wartezeiten am Flughafen*

An einem Schalter (z.B. am Flughafen) können pro Minute 6 Personen abgefertigt werden. Es treffen im langfristigen Schnitt nur 4 Personen pro Minute ein.

Bisher: $X$ = „Anzahl eintreffender Personen“ $\sim Poi(\lambda)$

Neu: $Y$ = „Zeitdauer bis nächster Kunde kommt“ $\sim ???$

Während z.B. die Anzahl der eintreffenden Personen nur ganzzahlige Werte annehmen kann, ist dies für die Wartezeit nicht der Fall. Diese ist von kontinuierlichem Charakter.

→ $Y$ ist eine stetige Zufallsvariable mit stetiger Verteilung

# *Wie könnte diese Verteilung aussehen?*

Wir betrachten zu einem beliebigen Zeitpunkt die nächste Sekunde. Die Anzahl Kunden $X$, welche in der nächsten Sekunde am Schalter eintrifft, hat eine Poisson-Verteilung:

$$X \sim Pois(\lambda = 4/60)$$

Wir bestimmen die W'keit, dass in der nächsten Sekunde kein Kunde eintrifft, die Wartezeit also länger als 1s dauert:

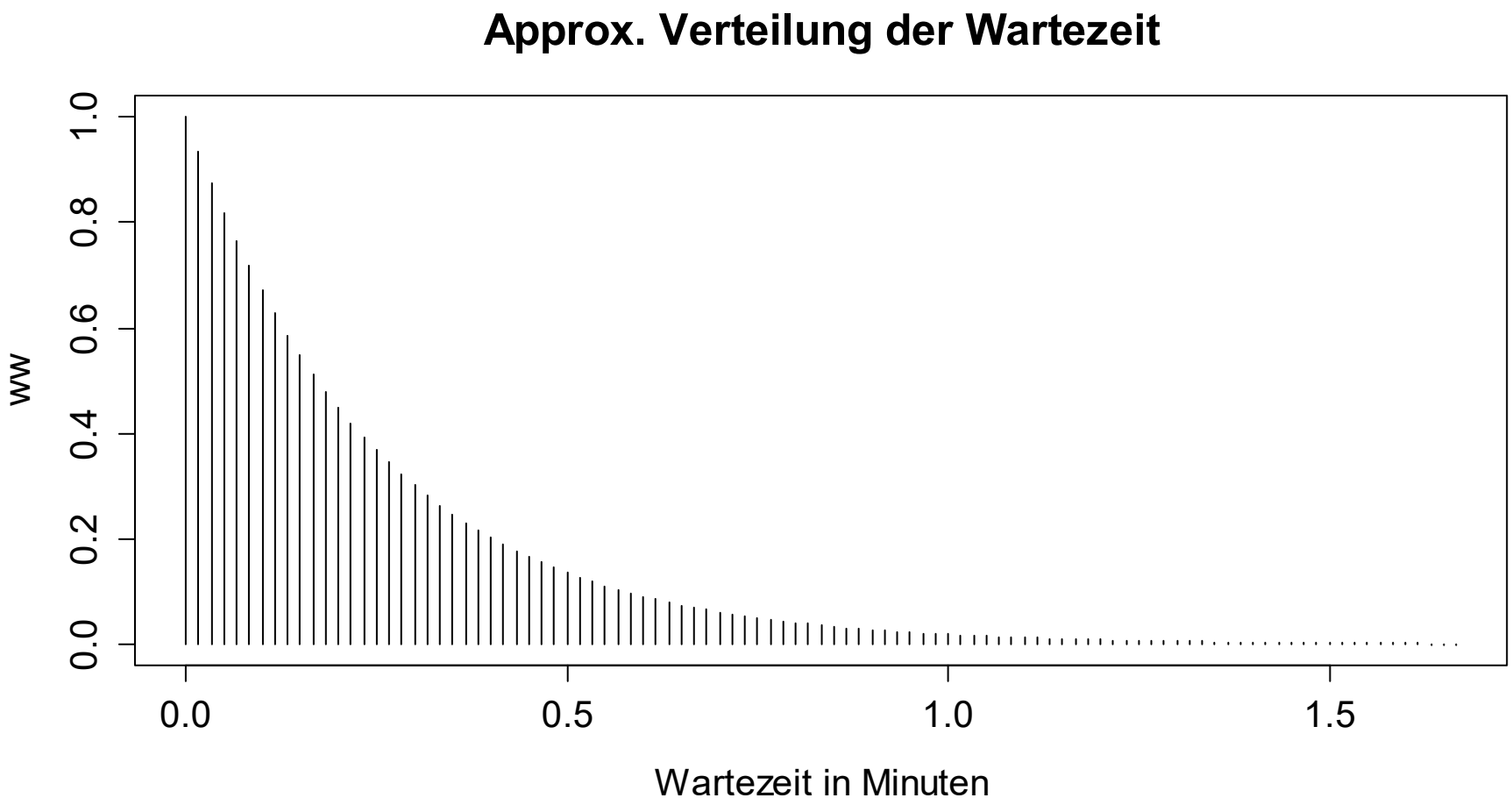
$$P(X = 0) = \exp(-4/60) \approx 0.936$$

Die W'keit, dass nach $y$ Sekunden noch kein Kunde da ist:

$$P(Y > y) \approx P(X = 0)^k \approx 0.936^k$$

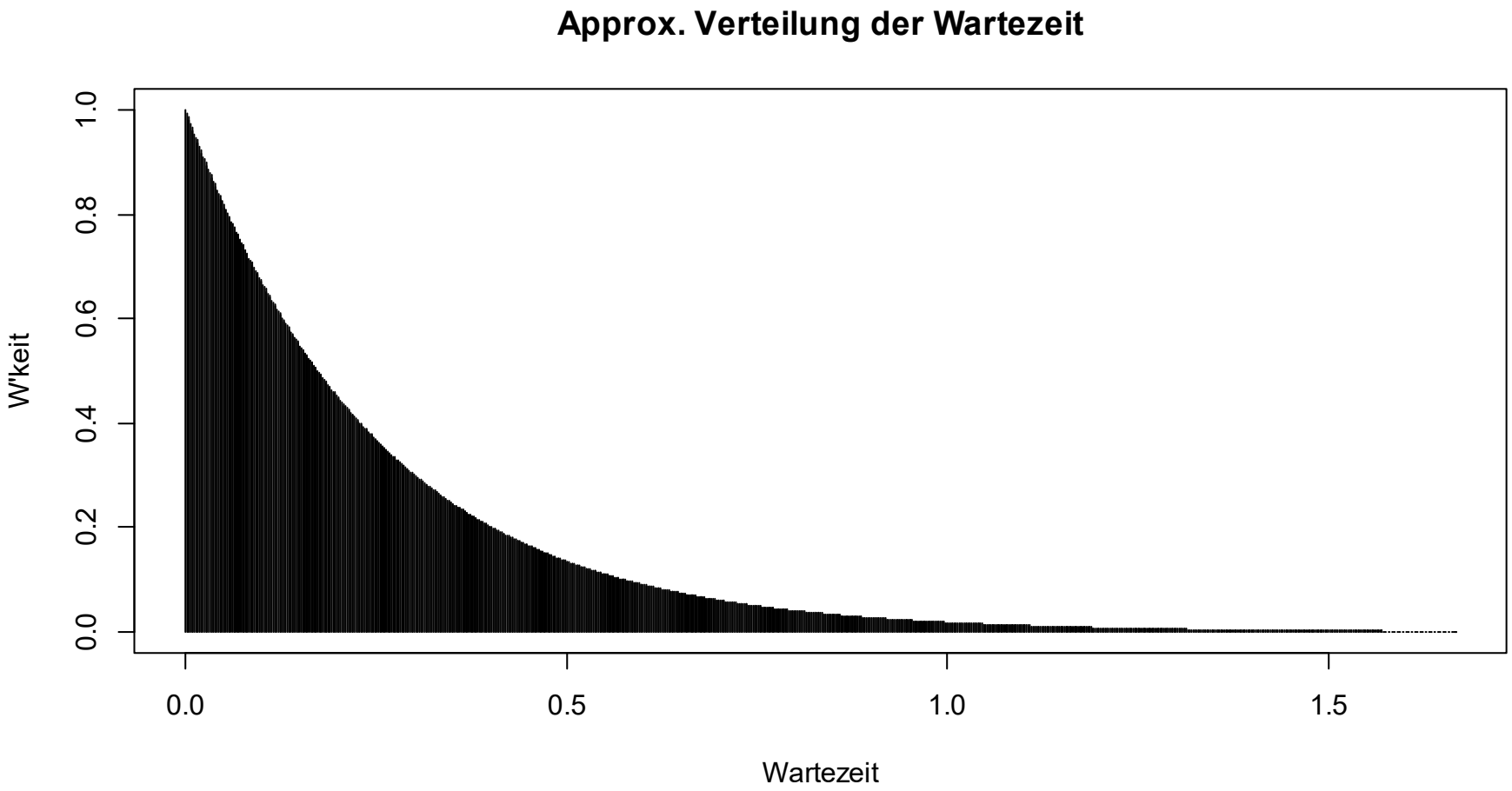
# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

# *Wie könnte diese Verteilung aussehen?*



# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

# *Wie könnte diese Verteilung aussehen?*



# *Dichtefunktion*

→ Im Grenzübergang erhalten wir eine Dichtefunktion.

Für die Wartezeiten zwischen dem unabhängigen Eintreffen von Personen ist die Exponentialverteilung ein geeignetes Modell. Deren Dichtefunktion ist wie folgt definiert:

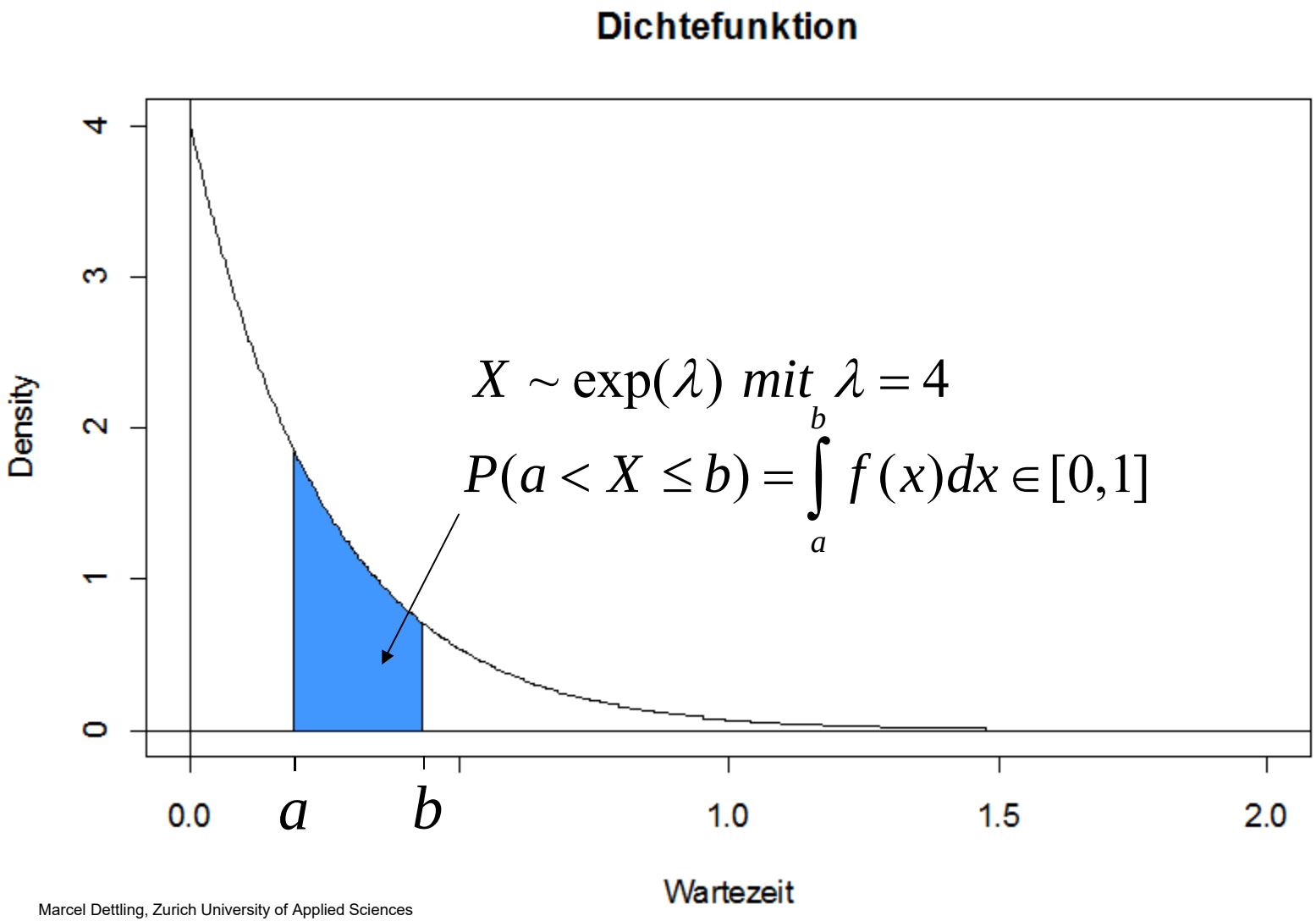
$$f(x) = \lambda \cdot e^{-\lambda x} \quad \textit{für} \quad x \geq 0$$

$$f(x) = 0 \quad \textit{für} \quad x < 0$$

Die Exponentialverteilung bezeichnet man als *"gedächtnislos"*. Das Eintreffen des letzten Kunden hat keinen Einfluss auf das Eintreffen des nächsten und alle folgenden Kunden.

# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

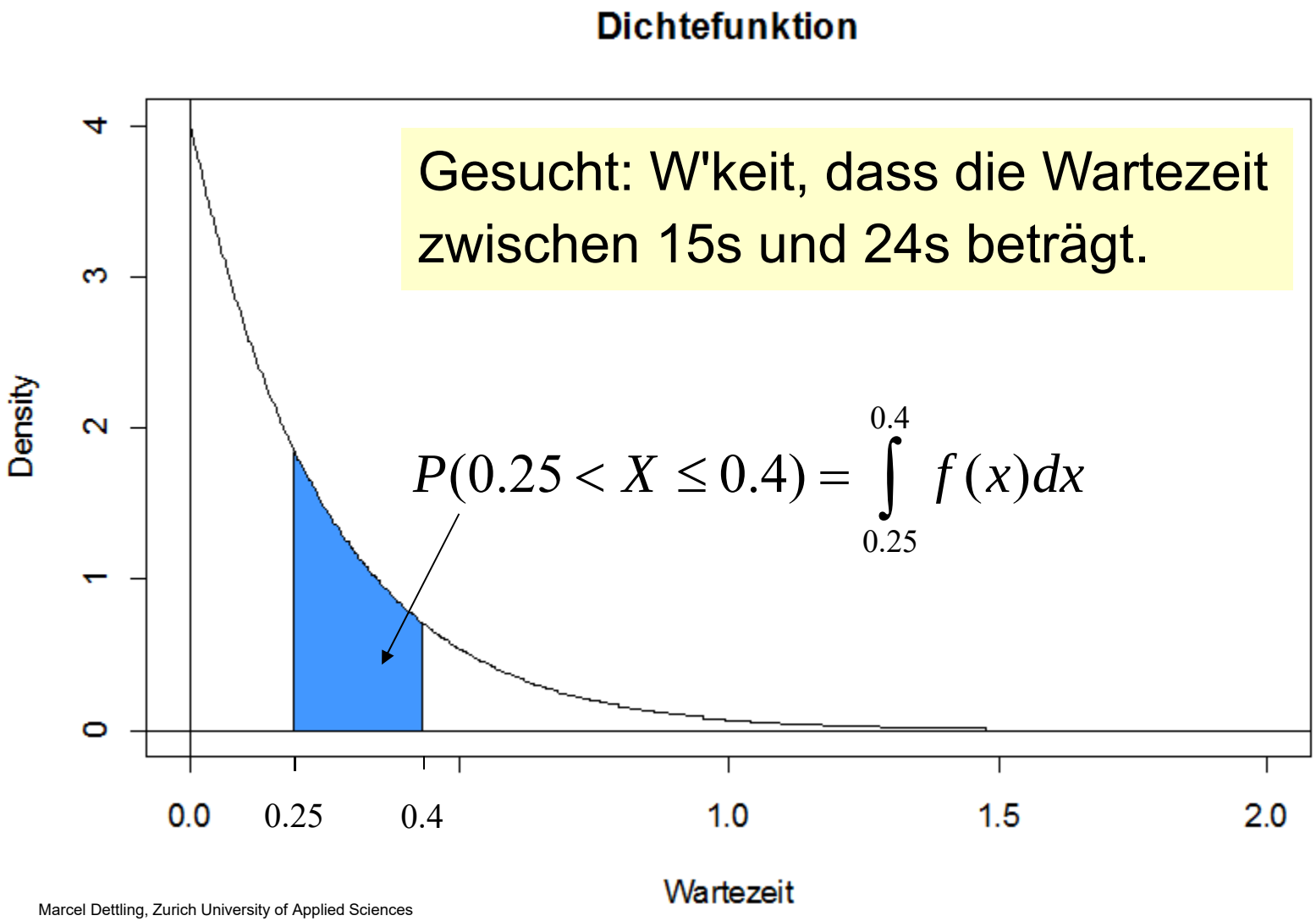
# *Bedeutung der Dichtefunktion*




$$X \sim \exp(\lambda) \; mit \; \lambda = 4$$

$$P(a < X \leq b) = \int_a^b f(x)dx \in [0,1]$$

# *Bedeutung der Dichtefunktion*



# *Kumulative Verteilungsfunktion*

Die kumulative Verteilungsfunktion $F(\cdot)$ ist definiert durch:

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

Es ist die Wahrscheinlichkeit, dass die Zufallsvariable $X$ einen Wert kleiner als $x$ annimmt. Sie ist bestimmt durch das Integral über die Dichtefunktion vom linken Rand bis zum Wert $x$.

*Konkrete Durchführung für die Exponentialverteilung:*

→ **siehe Wandtafel...**

# *Allgemeine Definition der Dichtefunktion*

Die Dichte $f(\cdot)$ ist eine stückweise stetige Funktion mit

$$f(x) \geq 0$$

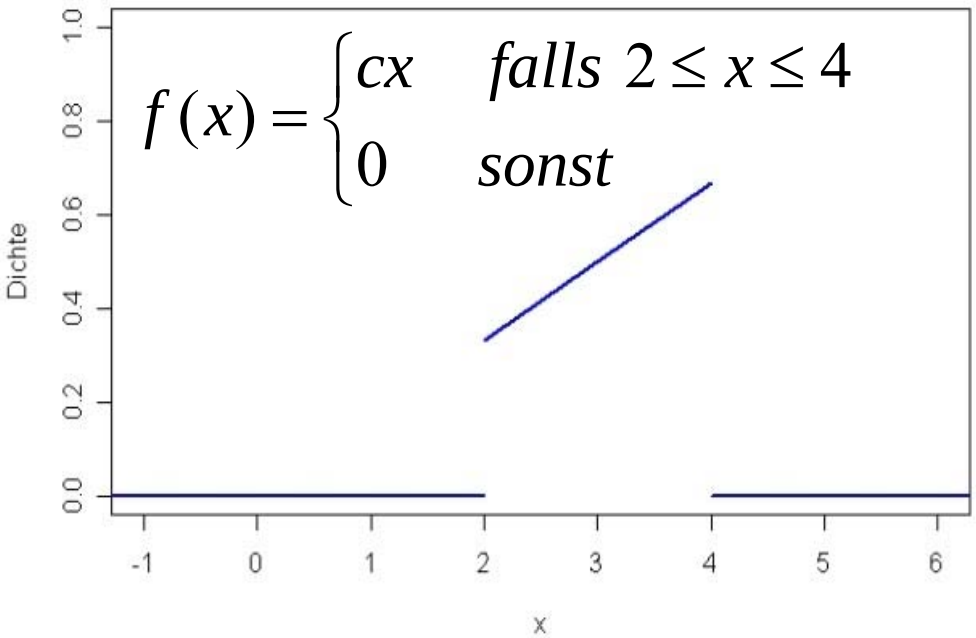
und

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

Oder in Worten: die Dichtefunktion nimmt keine negativen Werte an, und das Integral über den ganzen Bereich der reellen Zahlen ist gleich 1.

# *Rechnen mit stetigen Verteilungen*

In einem Betrieb verdienen Aushilfskräfte zwischen 2000 und 4000 CHF brutto. Es sei $X$ der Lohn, für welchen folgende Dichtefunktion gilt:



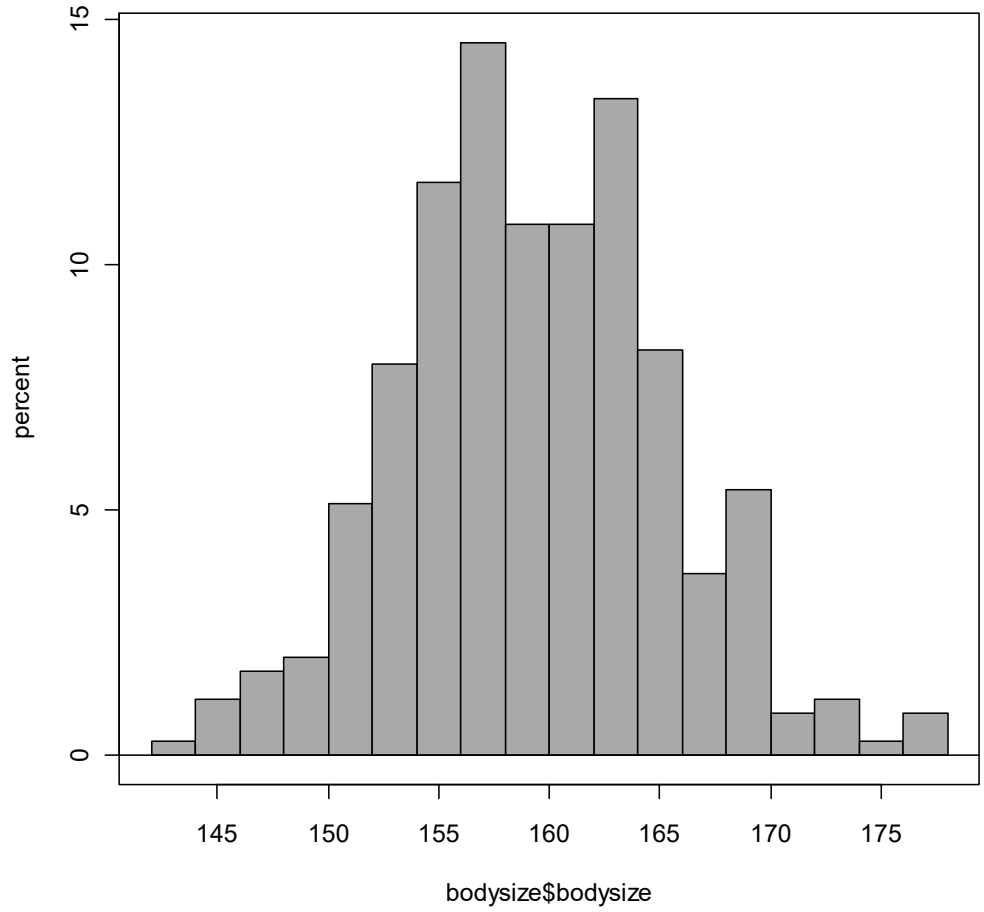

a) Gibt es mehr Leute, die zwischen 3000 und 4000 als unter 3000 verdienen?

b) Wie ist die Konstante $c$ ?

c) Wie gross ist die W'keit für einen Lohn zwischen 2500 und 3000?

# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

# *Welche Verteilung haben Messdaten?*




Histogramm der Körper-grösse von 350 zufällig ausgewählten Frauen

# *Normalverteilung*

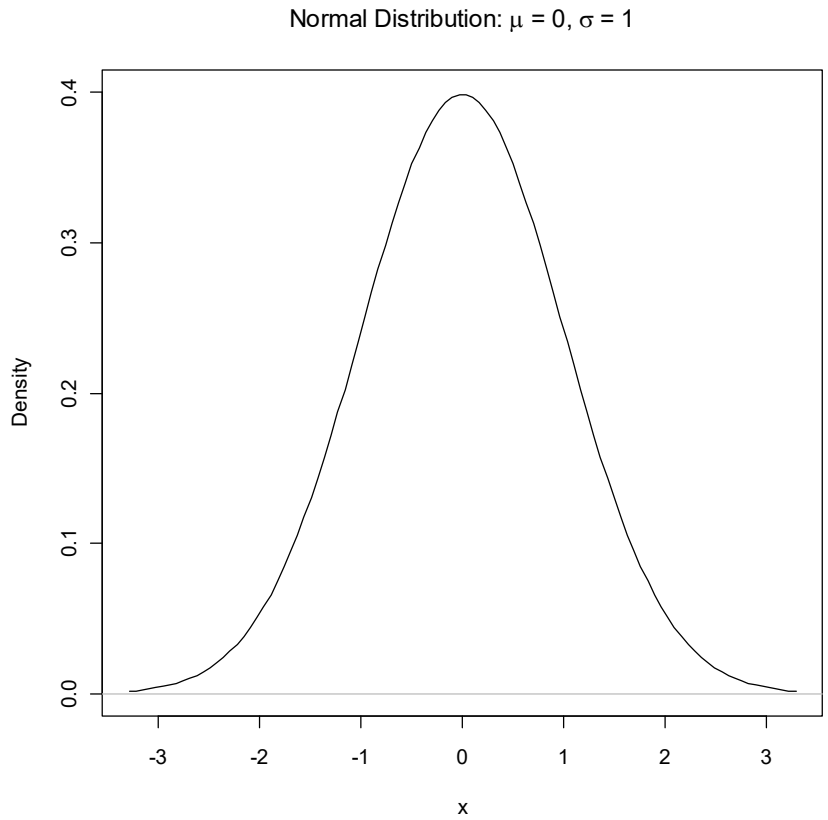
- Die Normalverteilung kommt dann zum Einsatz, wenn stetige Variablen einer symmetrischen, unimodalen Verteilung folgen.
- Dies ist typischerweise der Fall, wenn sich das Resultat eines Zufallsexperiments aus der additiven Überlagerung von vielen kleinen Einflüssen ergibt.
- Die Dichtefunktion der Normalverteilung lautet:

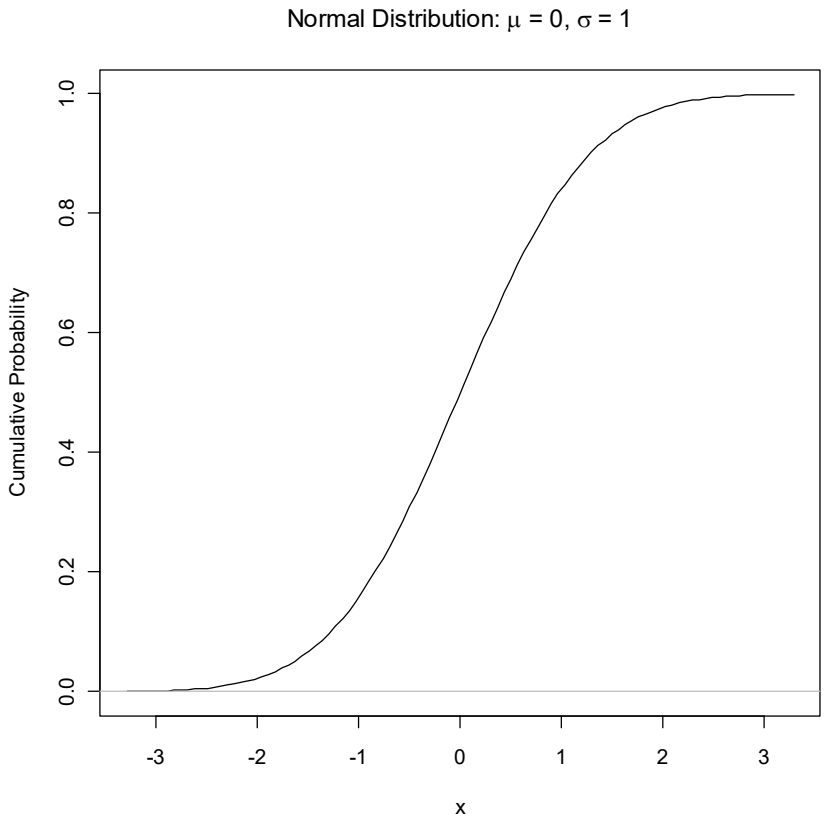
$$f(x) := \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left\{ -\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2} \right\}$$

Die Parameter der Verteilung sind $\mu$ und $\sigma^2$. Sie beschreiben Lage und Streuung. Wir schreiben kurz $X \sim N(\mu, \sigma^2)$.

# *Dichte und Verteilungsfunktion*

Wir betrachten die Zufallsvariable $X \sim N(0,1)$

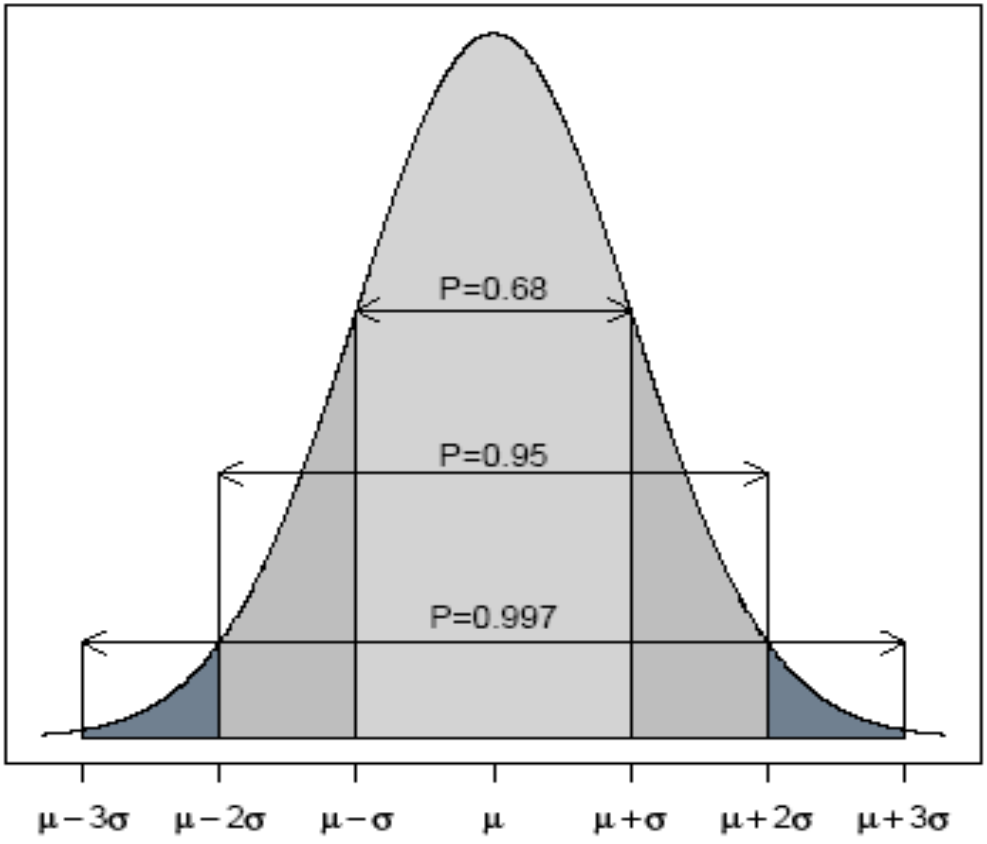


# *Eigenschaften der Normalverteilung*

- Die Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ nennt man Standard-Normalverteilung. Wir schreiben $X \sim N(0,1)$.
- Das Integral über die Dichtefunktion existiert, aber es kann nicht in geschlossener Form angegeben werden.
- Es gibt also keine "Formel" für die Verteilungsfunktion. Für die Berechnung von Wahrscheinlichkeiten muss man deshalb auf Tabellen oder Computerprogramme zurückgreifen.
- Es gibt unendlich viele Normalverteilungen, je nach Wahl der beiden Parameter. Die Glockenform haben alle gemeinsam.
- Parameter: $\mu$ verschiebt die Glocke, $\sigma^2$ streckt/staucht sie.

# *Schwankungsbereich der Normalverteilung*



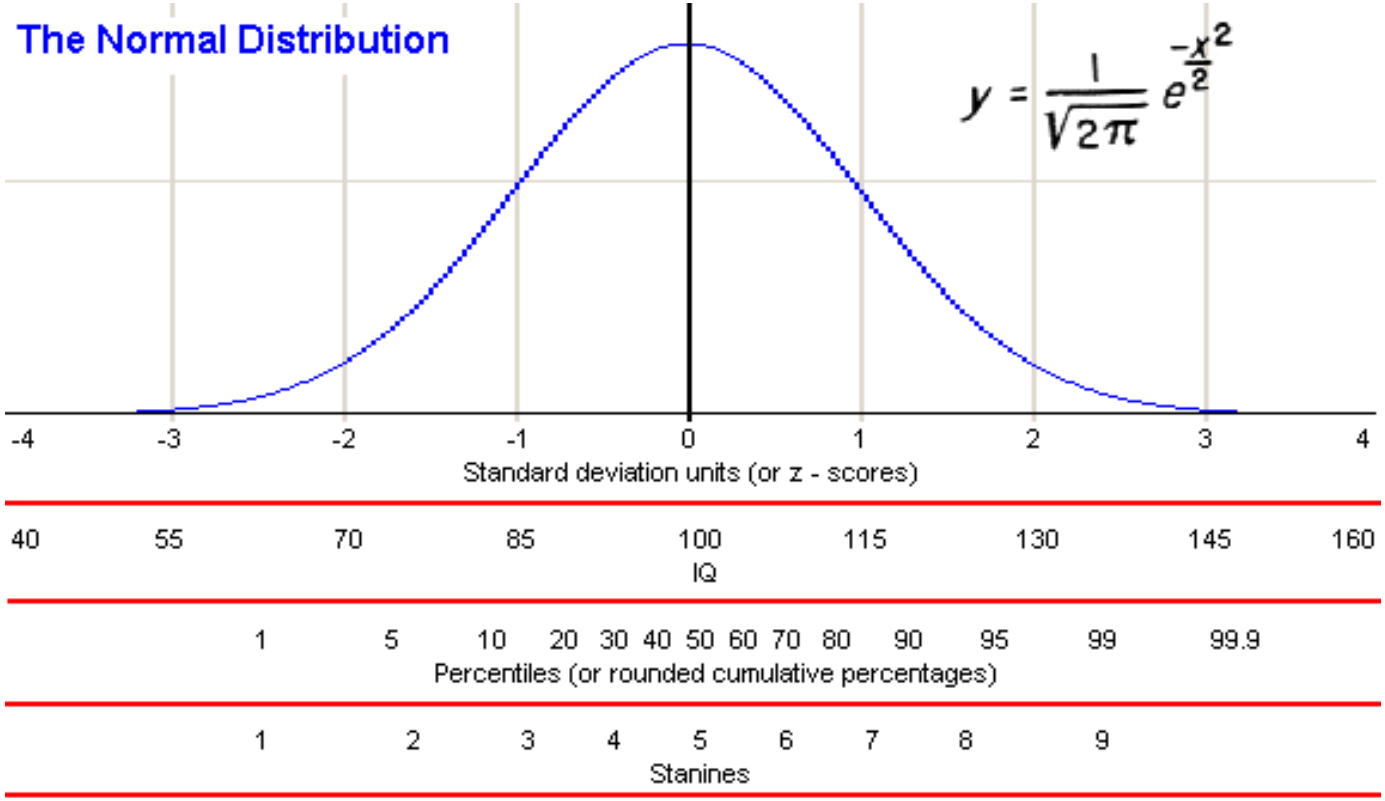

Diese Grafik liefert uns eine nützliche Faust-regel: das Intervall

$$[\mu - 2\sigma, \mu + 2\sigma]$$

enthält rund 95% der Wahrscheinlichkeit, d.h.

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$$

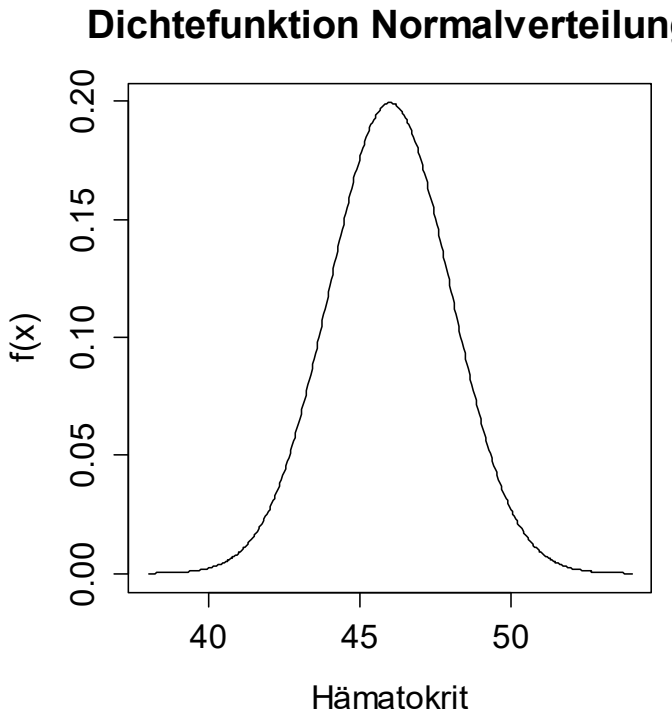
# *Quantile und Anwendungsbeispiel*

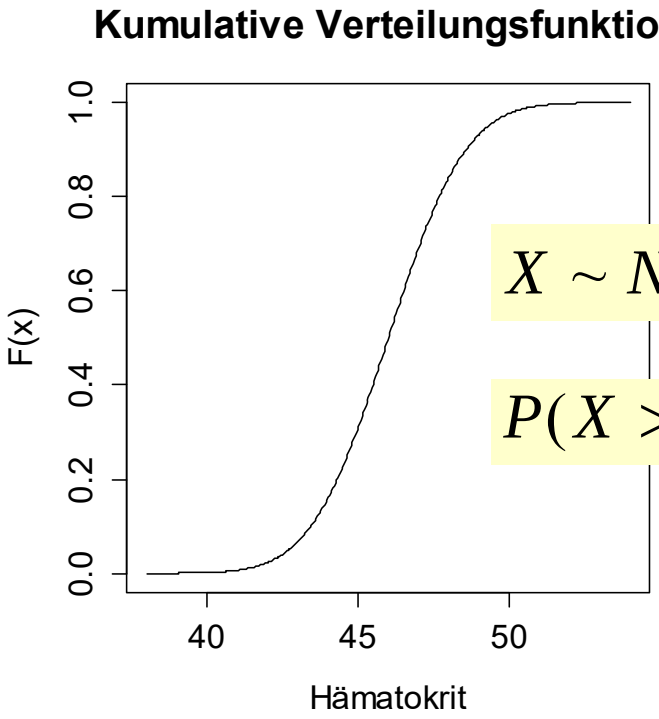



→ Der IQ über alle Personen hat die Verteilung $N(100,\ 225)$

# *Beispiel: Hämatokrit*

Anteil der Erythrozyten am Volumen des Bluts. Diese stellen rund 99% des Gesamtvolumen der Blutzellen dar. Normale Werte beim Menschen liegen zwischen 42% und 50%.




$$X \sim N(\mu = 46, \sigma^2 = 2^2)$$

$$P(X > 51.5) = 0.3\%$$

# *Schätzung der Parameter*

I.A. sind die Parameter einer Normalverteilung noch nicht bekannt und müssen aufgrund von Daten geschätzt werden.

Wir verwenden dazu den Stichproben-Mittelwert für $\mu$ und die Stichproben-Standardabweichung für $\sigma^2$.

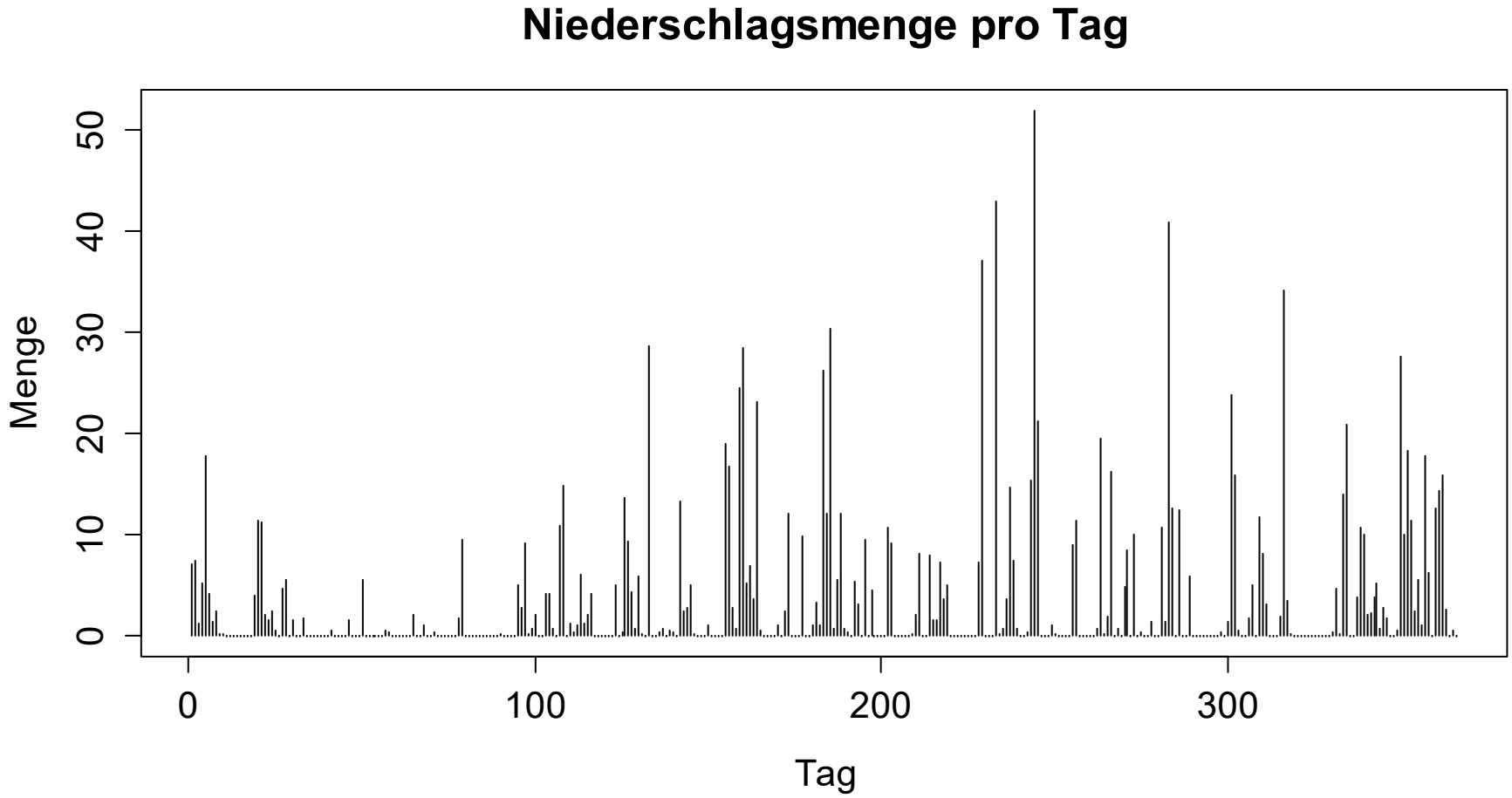
$$\hat{\mu} = \overline{x} = \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

# *Lognormal-Verteilung*

$Z =$ **"Regenmenge an einem Tag mit Niederschlag"**



# *Lognormal-Verteilung*

- Die Lognormal-Verteilung kommt für stetige Variablen zum Einsatz, welche nur positive Werte annehmen können und die eine rechtsschiefe Verteilung aufweisen.

- Charakteristisch ist es, dass für solche Grössen Unterschiede besser durch Verhältnisse als durch Differenzen ausgedrückt werden, bzw. sich das Resultat aus einer multiplikativen Überlagen von vielen kleinen Einflüssen ergibt.

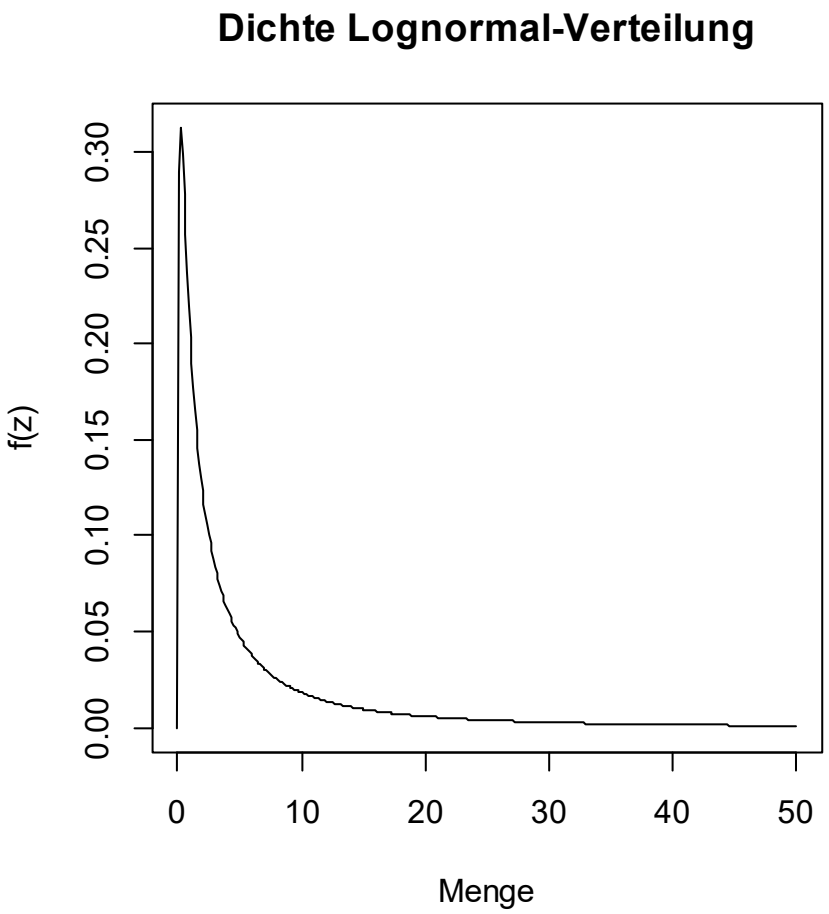
- Zur Definition/Rechnung benützen wir folgende Eigenschaft:

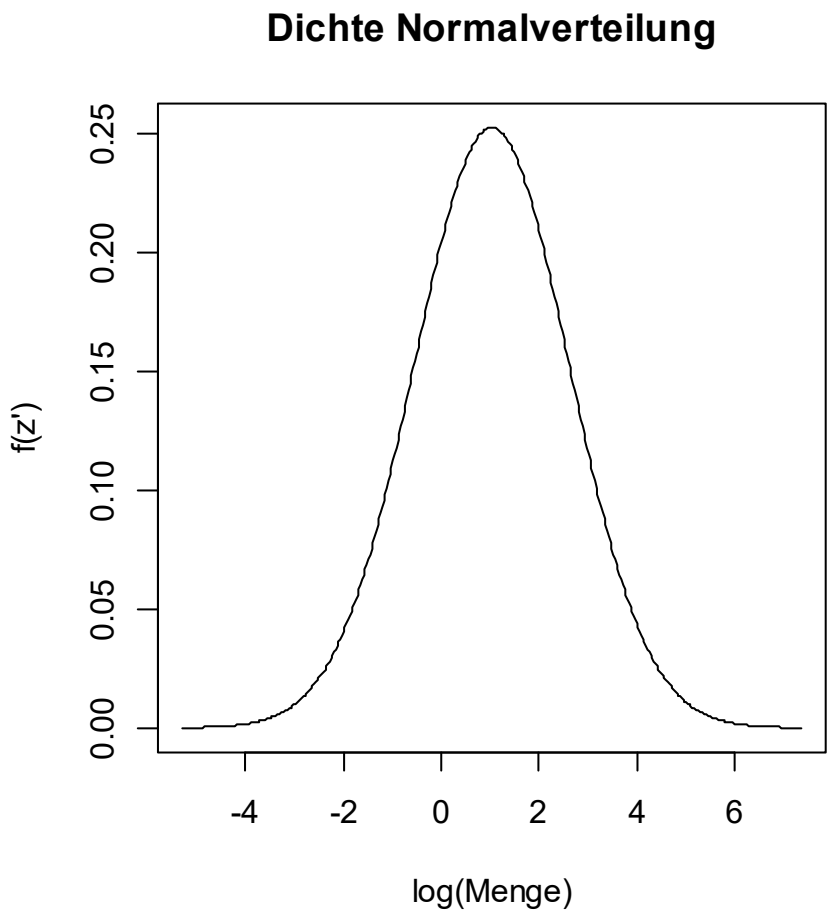
  Es gilt $Z \sim \log N(\mu, \sigma^2)$ falls $\log(Z) \sim N(\mu, \sigma^2)$

  Wir schreiben in diesen Fällen kurz: $Z \sim \log N(\mu, \sigma^2)$.

# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

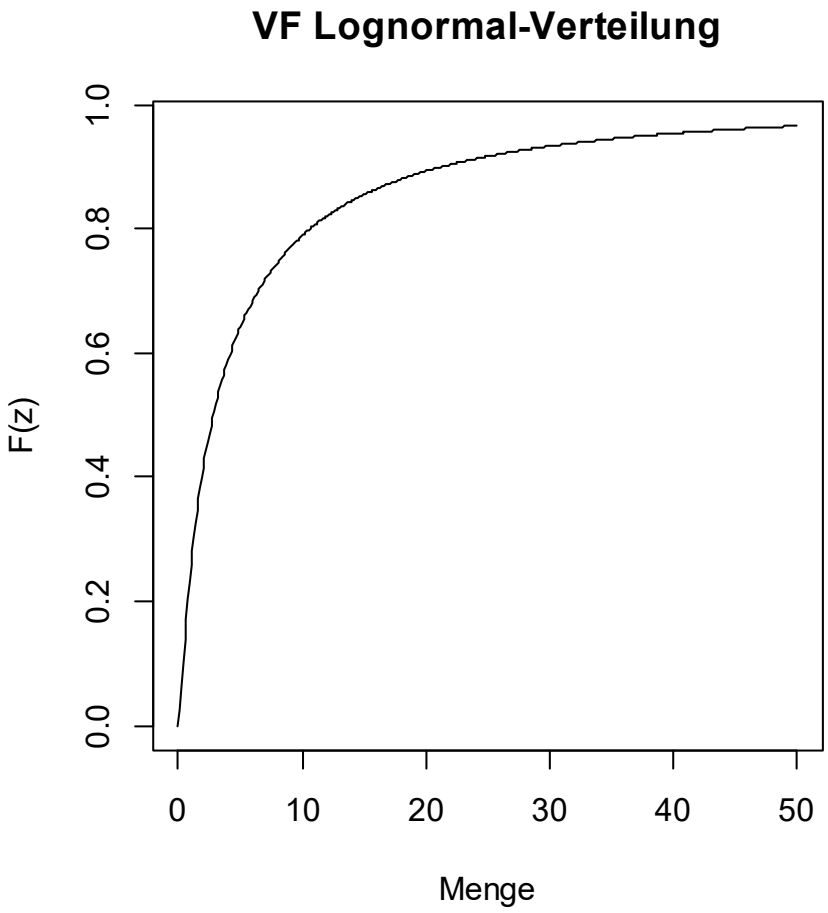
# *Dichtefunktionen im Beispiel*

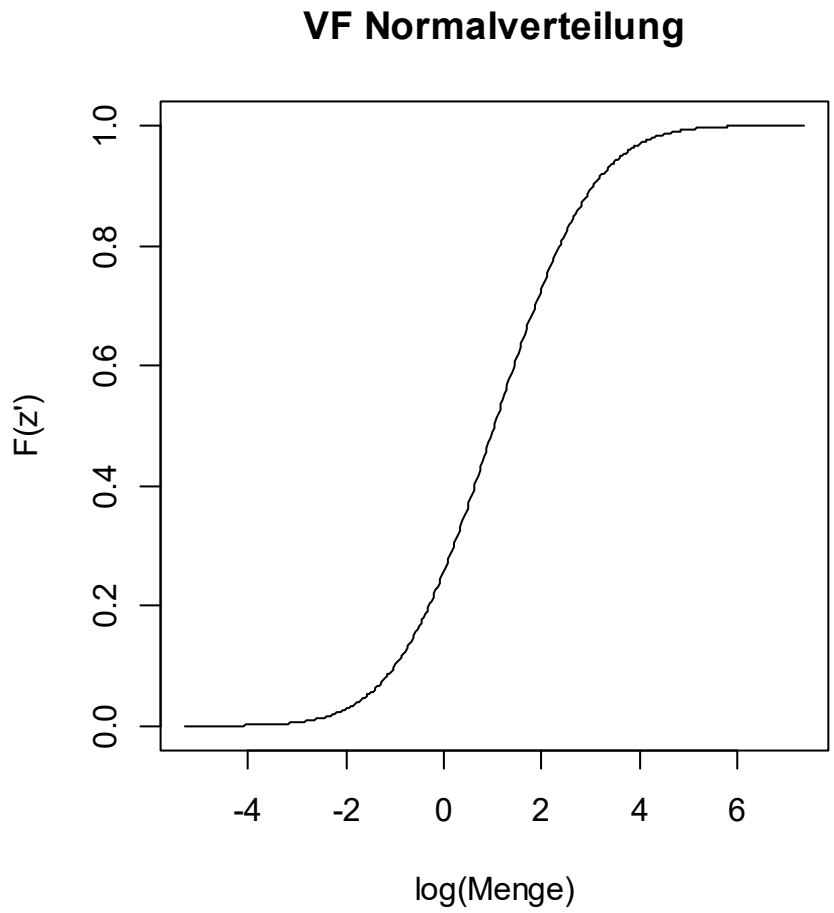


# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

# *Dichtefunktionen im Beispiel*



# *Anpassen und Rechnen*

Die Parameter $\mu, \sigma^2$ einer Lognormal-Verteilung werden durch Bilden von Mittelwert und Stichproben-Varianz auf den mit dem Logarithmus transformierten Daten bestimmt:

```
mw <- mean(log(menge))
sv <- var(log(menge))
```

In R benützt man `dlnorm()`, `plnorm()`, `qlnorm()`:

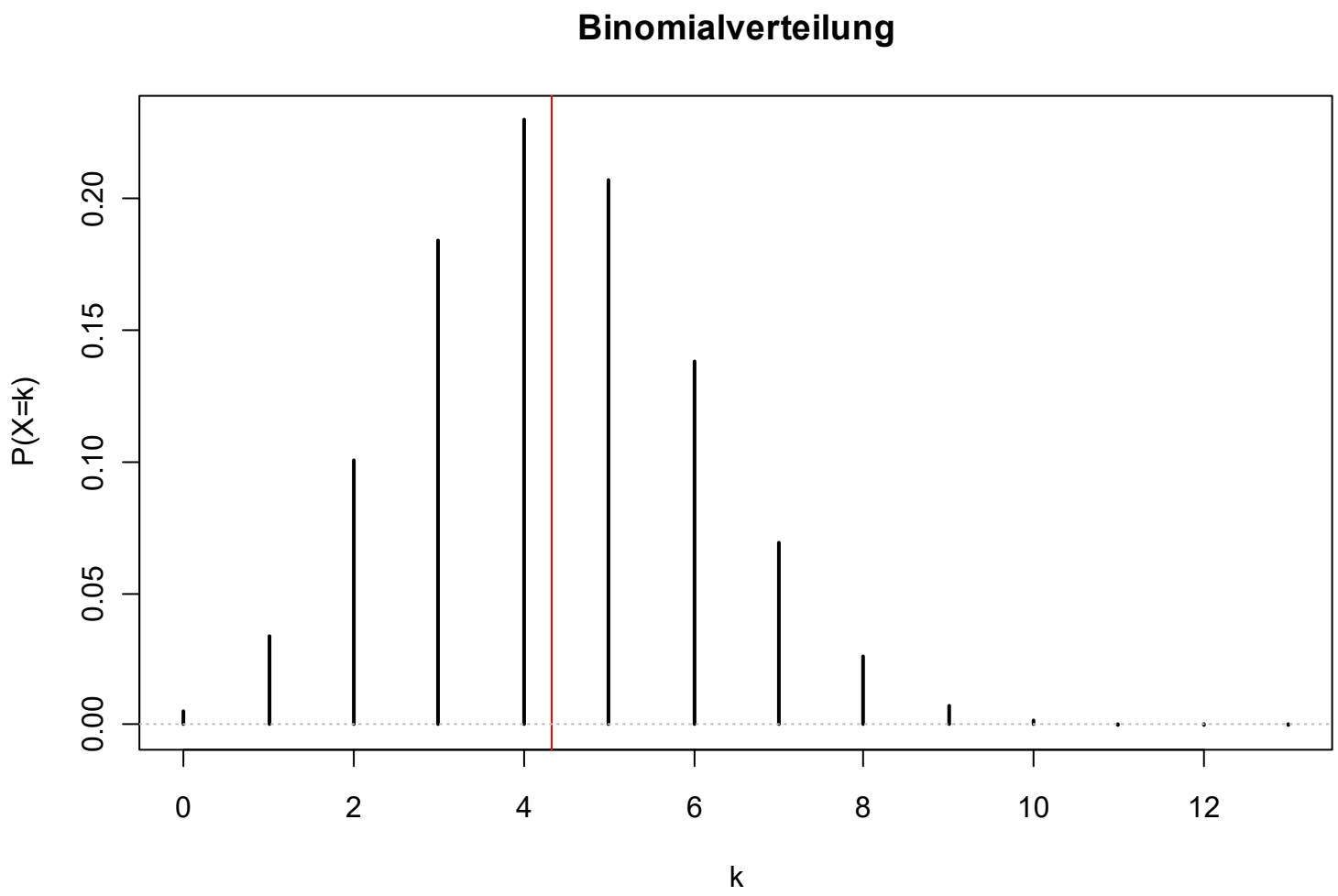
```
> plnorm(10, mw, sqrt(sv))
[1] 0.7829
> pnorm(log(10), mw, sqrt(sv))
[1] 0.7829
```

# *Erwartungswert*

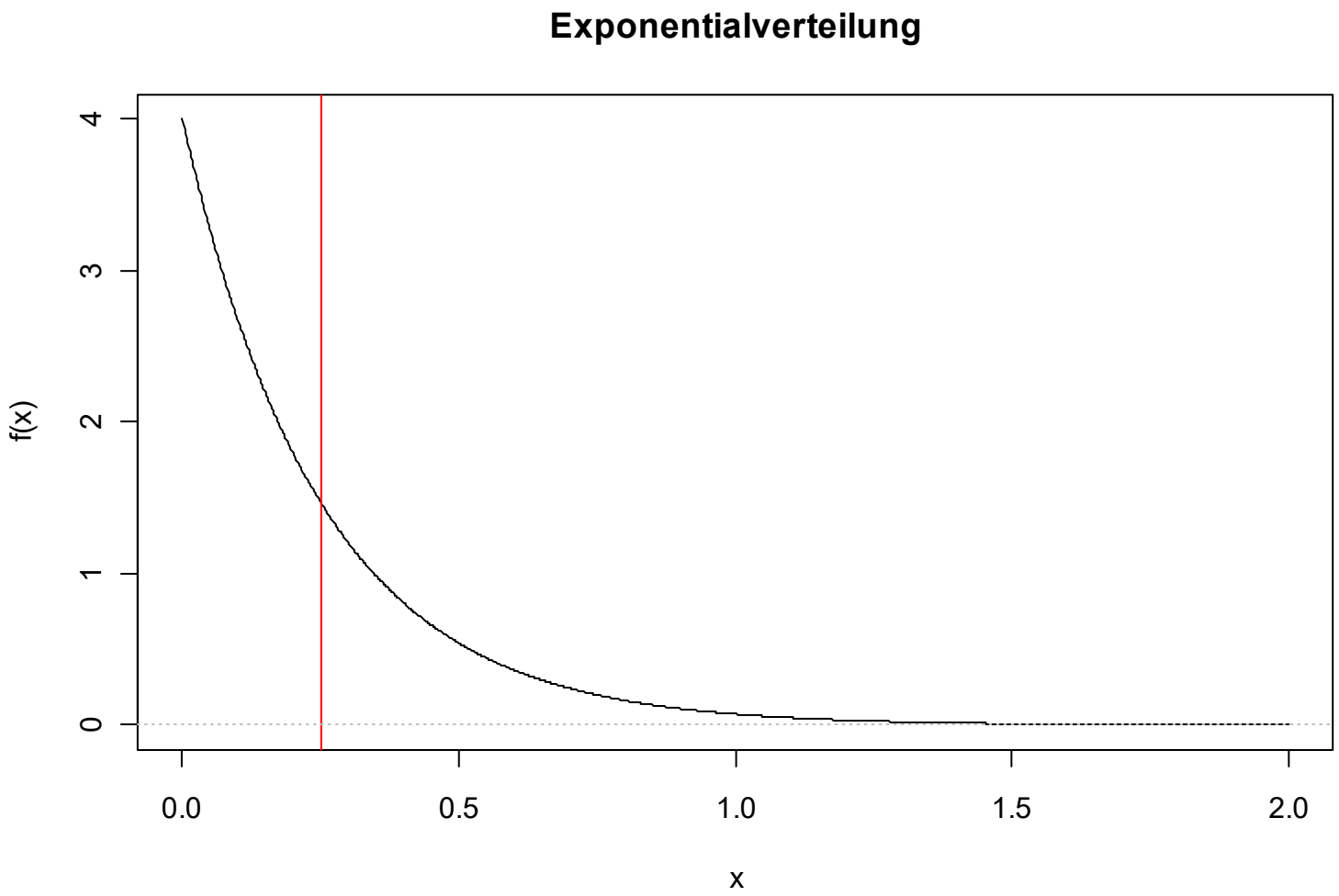
- Der Erwartungswert einer Zufallsvariablen ist das, was man im Schnitt, bei unendlich vielen Realisierungen, erhält.
- Er ist ein Lagemass für eine Zufallsvariable. Bildlich handelt es sich um die Position der Verteilung auf der x-Achse
- Man kann den Erwartungswert geometrisch bestimmen: Es ist der Schwerpunkt (in x-Richtung) der Wahrscheinlichkeits- bzw. Dichtefunktion
- Natürlich kann man den Erwartungswert einer Verteilung auch mittels Formeln rechnerisch bestimmen

# GdM 2: LinAlg & Statistik
## FS 2017 – Woche 11

# *Schwerpunkt/EW der Binomialverteilung*



# *Schwerpunkt/EW der Exponentialverteilung*



# *Berechnung des Erwartungswerts*

1) **Für diskrete Zufallsvariablen** $X$

Es sei $X$ eine diskrete Zufallsvariable mit W'keitsfunktion $p(\cdot)$. Der Erwartungswert $E[X]$ ist definiert als:

$$E(X) = \sum_i x_i \cdot P(X = i) = \sum_i x_i \cdot p(x_i) = \sum_i x_i \cdot p_i$$

falls diese Summe existiert. Wir schreiben oft auch $\mu = E(X)$.

2) **Für stetige Zufallsvariablen** $Z$

Sei $Z$ eine stetige Zufallsvariable mit Dichtefunktion $f(\cdot)$. Dann ist, falls das Integral existiert:

$$\mu = E(Z) = \int_{-\infty}^{+\infty} z \cdot f(z)\, dz$$

# *Varianz und Standardabweichung*

Sei $X$ eine Zufallsvariable mit Erwartungswert $E[X]$.
Dann ist die Varianz von $X$ gegeben durch:

$$Var(X) = E\left[(X - E(X))^2\right]$$

falls dieser Erwartungswert existiert. Die Standardabweichung ist die Wurzel aus der Varianz. Wir schreiben auch

$$\sigma^2, \ bzw. \ \sigma$$

für die Varianz, bzw. die Standardabweichung. Die Varianz ist ein Streuungsmass der Zufallsvariablen. Sie bestimmt grob gesagt die Breite der Verteilung, bzw. wird durch die Breite der Verteilung bestimmt.

# *Berechnung der Varianz*

1) **Für diskrete Zufallsvariablen** $X$

   Es sei $X$ eine diskrete Zufallsvariable mit W'keitsfunktion $p(\cdot)$ und Erwartungswert $\mu = E[X]$. Die Varianz berechnet sich als:

$$Var(X) = \sum_i (x_i - \mu)^2 \cdot p(x_i)$$

   falls diese Summe existiert.

2) **Für stetige Zufallsvariablen** $Z$

   Sei $Z$ eine stetige Zufallsvariable mit Dichtefunktion $f(\cdot)$ und Erwartungswert $\mu = E[Z]$. Dann ist, falls das Integral existiert:

$$Var(Z) = \int_{-\infty}^{+\infty} (z - \mu)^2 \, f(z) dz$$

# *Erwartungswerte bestimmter Verteilungen*

Für alle parametrischen Verteilungsfamilien sind Erwartungswert und Varianz bereits als Funktion der Parameter bekannt:

| Verteilung | | E[X] bzw. E[Z] | Var(X) bzw. Var(Z) |
|---|---|---|---|
| **Bernoulli** | $X \sim Bernoulli(p)$ | $p$ | $p(1-p)$ |
| **Binomial** | $X \sim Bin(n,p)$ | $np$ | $np(1-p)$ |
| **Poisson** | $X \sim Pois(\lambda)$ | $\lambda$ | $\lambda$ |
| **Exponential** | $Z \sim Exp(\lambda)$ | $1/\lambda$ | $1/\lambda^2$ |
| **Uniform** | $Z \sim Unif(u,o)$ | $(u+o)/2$ | $(o-u)^2/12$ |
| **Normal** | $Z \sim N(\mu,\sigma^2)$ | $\mu$ | $\sigma^2$ |
| **Lognormal** | $Z \sim \log N(\mu,\sigma^2)$ | $e^{(\mu+0.5\sigma^2)}$ | $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$ |

# *Rechenregeln für Erwartungswert/Varianz*

- Es ist nicht selten der Fall, dass wir an der Summe $S$ von zwei beliebigen Zufallsvariablen $X$ und $Y$ interessiert sind.
- Selbst wenn wir die Verteilung von $X$, $Y$ kennen, so ist es nur in wenigen Ausnahmefällen (z.B. Poissonverteilung, Normalverteilung) möglich, die Verteilung von $S$ anzugeben.
- Einfacher ist die Sache für die Kenngrössen Erwartungswert und Varianz. Diese lassen sich für lineare Transformationen und für Summen einfach angeben.

→ siehe nächste Folie…

# *Rechenregeln für Erwartungswert/Varianz*

Es seien $a, b$ skalare Grössen und $S = X + Y$ Zufallsvariablen. Dann gelten die folgenden Rechenregeln ***immer***:

## 1) Für den Erwartungswert

- $E[aX + b] = aE[X] + b$
- $E[S] = E[X] + E[Y]$

## 2) Für die Varianz

- $Var(aX + b) = a^2 Var(X)$
- $Var(S) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$

Der Term $Cov(X, Y)$ beschreibt den linearen Zusammenhang der beiden Zufallsvariablen. Er ist nur in Ausnahmefällen gleich null, z.B. wenn $X, Y$ unabhängig sind.

# *Beispielaufgabe*

Beispiel: Jasskarten – wir studieren die Zufallsvariable

$$X = "Punktewert\ einer\ zufällig\ herausgegriffenen\ Jasskarte"$$

**Wir studieren Bedeutung, Form und Wert:**

- der Wahrscheinlichkeitsverteilung
- des Erwartungswerts $E[X]$
- der Varianz $Var(X)$
- der Summe $S = X + X$

→ *Siehe Wandtafel…*