

Biostatistik, SS 2016

**Grundlagen aus der
Wahrscheinlichkeitstheorie**

Matthias Birkner und Frederik Klement

<http://www.staff.uni-mainz.de/birkner/Biostatistik16/>

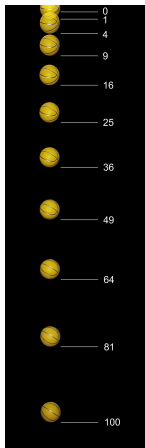
27.5.2016



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

- 1 Deterministische und zufällige Vorgänge
- 2 Zufallsvariablen und Verteilung
- 3 Die Binomialverteilung
- 4 Erwartungswert
- 5 Varianz und Korrelation
- 6 Ein Anwendungsbeispiel
- 7 Die Normalverteilung
- 8 Normalapproximation
- 9 Der z-Test

Was können wir vorhersagen?



(c) by Michael Maggs

- Freier Fall: Falldauer eines Objektes bei gegebener Fallhöhe läßt sich vorhersagen (falls Luftwiderstand vernachlässigbar)

Deterministische Vorgänge laufen immer gleich ab. Aus Beobachtungen lassen sich künftige Versuche vorhersagen.

Was können wir vorhersagen?

- Würfelnwurf: Das Ergebnis eines einzelnen Würfelnwurfes lässt sich nicht vorhersagen.



(c) public domain

- Wiederholter Würfelnwurf:
Würfelt man 600 mal, so würde man gerne darauf wetten, dass die Anzahl an Einsern zwischen 85 und 115 liegt.
Die genaue Anzahl lässt sich wieder nicht vorhersagen.
Aber: **Eine Aussage über die Verteilung ist möglich**
(die besser ist als reines Raten.)

Empirisch stellt man fest:

Bei Wiederholung eines Zufallsexperiments stabilisieren sich die relativen Häufigkeiten der möglichen Ergebnisse.

Beispiel:

Beim Würfelwurf
stabilisiert sich die relative Häufigkeit
jeder der Zahlen $\{1, 2, \dots, 6\}$ bei $\frac{1}{6}$.

Fazit:

Das Ergebnis eines einzelnen zufälligen Vorgangs
läßt sich nicht vorhersagen.

Aber: Eine Aussage über die Verteilung ist möglich
(die besser ist als reines Raten).

Abstraktionsschritt:

Verwende empirisch ermittelte Verteilung
als Verteilung jedes Einzelexperiments!

Beispiel:

Wir nehmen an,
daß bei einem einzelnen Würfelwurf
jede der Zahlen $\{1, 2, \dots, 6\}$
die **Wahrscheinlichkeit** $\frac{1}{6}$ hat.

Als **Zufallsgröße oder Zufallsvariable** bezeichnet man das (Mess-)Ergebnis eines zufälligen Vorgangs.

Der **Wertebereich \mathcal{S}** (engl. state space) einer Zufallsgröße ist die Menge aller möglichen Werte.

Die **Verteilung einer Zufallsgröße X** weist jeder Menge $A \subseteq \mathcal{S}$ die **Wahrscheinlichkeit $\mathbb{P}(X \in A)$** zu, dass X einen Wert in A annimmt

Für Zufallsgrößen werden üblicherweise Großbuchstaben verwendet (z.B. X, Y, Z), für konkrete Werte Kleinbuchstaben.

Beispiel: Würfelwurf W = Augenzahl des nächsten Würfelwurfs.

$$\mathcal{S} = \{1, 2, \dots, 6\}$$

$$\mathbb{P}(W = 1) = \dots = \mathbb{P}(W = 6) = \frac{1}{6}$$

$$(\mathbb{P}(W = x) = \frac{1}{6} \text{ für alle } x \in \{1, \dots, 6\})$$

Die Verteilung erhält man aus einer Symmetrieüberlegung
oder aus einer langen Würfelreihe.

Beispiel: Geschlecht X bei Neugeborenen.

$$\mathcal{S} = \{„männlich“, „weiblich“\}$$

Die Verteilung erhält man aus einer langen Beobachtungsreihe.

Beispiel: Körpergrößenverteilung in Deutschland.

Die Verteilung erhält man aus einer langen Messreihe.

Rechenregeln:

Beispiel Würfelwurf W :

$$\begin{aligned}\mathbb{P}(\{W = 2\} \cup \{W = 3\}) &= \mathbb{P}(W \in \{2, 3\}) \\ &= \frac{2}{6} = \frac{1}{6} + \frac{1}{6} = \mathbb{P}(W = 2) + \mathbb{P}(W = 3)\end{aligned}$$

$$\begin{aligned}\mathbb{P}(W \in \{1, 2\} \cup \{3, 4\}) &= \frac{4}{6} = \frac{2}{6} + \frac{2}{6} \\ &= \mathbb{P}(W \in \{1, 2\}) + \mathbb{P}(W \in \{3, 4\})\end{aligned}$$

Vorsicht:

$$\mathbb{P}(W \in \{2, 3\}) + \mathbb{P}(W \in \{3, 4\}) \neq \mathbb{P}(W \in \{2, 3, 4\})$$

Beispiel zweifacher Würfelwurf (W_1, W_2):

Sei W_1 (bzw W_2) die Augenzahl des ersten (bzw zweiten) Würfels.

$$\begin{aligned}\mathbb{P}(W_1 \in \{4\}, W_2 \in \{2, 3, 4\}) \\ &= \mathbb{P}((W_1, W_2) \in \{(4, 2), (4, 3), (4, 4)\}) \\ &= \frac{3}{36} = \frac{1}{6} \cdot \frac{3}{6} \\ &= \mathbb{P}(W_1 \in \{4\}) \cdot \mathbb{P}(W_2 \in \{2, 3, 4\})\end{aligned}$$

Sei S die Summe der Augenzahlen, d.h. $S = W_1 + W_2$.
 Was ist die Wahrscheinlichkeit, daß $S = 5$ ist,
 wenn der erste Würfel die Augenzahl $W_1 = 2$ zeigt?

$$\begin{aligned} \mathbb{P}(S = 5 | W_1 = 2) &\stackrel{!}{=} \mathbb{P}(W_2 = 3) \\ &= \frac{1}{6} = \frac{1/36}{1/6} = \frac{\mathbb{P}(S = 5, W_1 = 2)}{\mathbb{P}(W_1 = 2)} \end{aligned}$$

Was ist die Ws von $S \in \{4, 5\}$ unter der Bedingung $W_1 \in \{1, 6\}$?

$$\begin{aligned} \mathbb{P}(S \in \{4, 5\} | W_1 \in \{1, 6\}) &= \frac{\mathbb{P}(S \in \{4, 5\}, W_1 \in \{1, 6\})}{\mathbb{P}(W_1 \in \{1, 6\})} \\ &= \frac{\mathbb{P}(W_2 \in \{3, 4\}, W_1 \in \{1\})}{\mathbb{P}(W_1 \in \{1, 6\})} = \frac{\mathbb{P}((W_1, W_2) \in \{(1, 3), (1, 4)\})}{\mathbb{P}(W_1 \in \{1, 6\})} \\ &= \frac{\frac{2}{36}}{\frac{2}{6}} = \frac{1}{6} \quad (\neq \frac{2}{6} = \mathbb{P}(W_2 \in \{3, 4\})) \end{aligned}$$

(Bem.: Es ist $\mathbb{P}(S \in \{4, 5\} | W_1 = 1) = \frac{\mathbb{P}(S \in \{4, 5\}, W_1 = 1)}{\mathbb{P}(W_1 = 1)} = \frac{\mathbb{P}((W_1, W_2) \in \{(1, 3), (1, 4)\})}{\mathbb{P}(W_1 = 1)} = \frac{2/36}{1/6} = \frac{2}{6}$ und

$\mathbb{P}(S \in \{4, 5\} | W_1 = 6) = \frac{\mathbb{P}((W_1, W_2) \in \{(1, -2), (1, -1)\})}{\mathbb{P}(W_1 = 6)} = \frac{0}{1/6} = 0$ sowie

$\mathbb{P}(W_1 = 1 | W_1 \in \{1, 6\}) = \frac{1/6}{2/6} = \frac{1}{2} = \mathbb{P}(W_1 = 6 | W_1 \in \{1, 6\})$, d.h. man kann durch Bedingen in zwei Schritten sehen:

$\mathbb{P}(S \in \{4, 5\} | W_1 \in \{1, 6\})$

$= \mathbb{P}(W_1 = 1 | W_1 \in \{1, 6\}) \cdot \mathbb{P}(S \in \{4, 5\} | W_1 = 1) + \mathbb{P}(W_1 = 6 | W_1 \in \{1, 6\}) \cdot \mathbb{P}(S \in \{4, 5\} | W_1 = 6).$

Rechenregeln:

Sei X Zufallsgröße mit Wertebereich \mathcal{S} .

- $0 \leq \mathbb{P}(X \in A) \leq 1$ für jede Teilmenge $A \subseteq \mathcal{S}$
- $\mathbb{P}(X \in \mathcal{S}) = 1$
- Sind $A, B \subseteq \mathcal{S}$ disjunkt, d.h. $A \cap B = \emptyset$,

$$\mathbb{P}(X \in A \cup B) = \mathbb{P}(X \in A) + \mathbb{P}(X \in B),$$

insbesondere $\mathbb{P}(X \in A^c) = 1 - \mathbb{P}(X \in A)$ mit $A^c = \mathcal{S} \setminus A$

- Allgemein gilt

$$\mathbb{P}(X \in A \cup B) = \mathbb{P}(X \in A) + \mathbb{P}(X \in B) - \mathbb{P}(X \in A \cap B)$$

(„Einschluss-Ausschluss-Formel“)

Bedingte Wahrscheinlichkeit

Ws des Ereignisses $\{Y \in B\}$ unter der Bedingung $\{X \in A\}$

$$\mathbb{P}(Y \in B | X \in A) := \frac{\mathbb{P}(Y \in B, X \in A)}{\mathbb{P}(X \in A)} \quad (*)$$

„bedingte Ws von $\{Y \in B\}$ gegeben $\{X \in A\}$ “

Beachte:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B | X \in A)$$

(*) in Worten ausgedrückt:

Die Ws des Ereignisses $\{X \in A, Y \in B\}$ läßt sich in zwei Schritten berechnen:

- Zunächst muss das Ereignis $\{X \in A\}$ eintreten.
- Die Ws hiervon wird multipliziert mit der Ws von $\{Y \in B\}$, wenn man schon weiß, daß $\{X \in A\}$ eintritt.

Die Formel von Bayes

Seien X, Y Zufallsgrößen mit Wertebereichen \mathcal{S}_X bzw. \mathcal{S}_Y ,
 $A \subset \mathcal{S}_X$, $B \subset \mathcal{S}_Y$, dann gilt

$$\begin{aligned} & \mathbb{P}(Y \in B \mid X \in A) \\ &= \frac{\mathbb{P}(X \in A \mid Y \in B) \cdot \mathbb{P}(Y \in B)}{\mathbb{P}(X \in A \mid Y \in B) \cdot \mathbb{P}(Y \in B) + \mathbb{P}(X \in A \mid Y \in B^c) \cdot \mathbb{P}(Y \in B^c)} \end{aligned}$$

Denn

$$\text{Zähler} = \mathbb{P}(X \in A, Y \in B)$$

$$\begin{aligned} \text{Nenner} &= \mathbb{P}(X \in A, Y \in B) + \mathbb{P}(X \in A, Y \in B^c) \\ &= \mathbb{P}(X \in A, Y \in B \cup B^c) = \mathbb{P}(X \in A) \end{aligned}$$

Beispiel: Medizinische Reihenuntersuchung

Eine Krankheit komme bei 2% der Bevölkerung vor („Prävalenz 2%“),

ein Test schlage bei 95% der Kranken an („Sensitivität 95%“),
aber auch bei 10% der Gesunden („Spezifität 90%“)

Eine zufällig gewählte Person werde mit positivem Resultat getestet.

Wie wahrscheinlich ist es, dass sie tatsächlich krank ist?

Modell: X = Testergebnis ($S_X = \{\text{positiv, negativ}\}$),

Y = Gesundheitszustand ($S_Y = \{\text{gesund, krank}\}$) der Person

Gesucht $\mathbb{P}(Y = \text{krank} \mid X = \text{positiv}) = ?$

Wir wissen: $\mathbb{P}(Y = \text{krank}) = 0.02$, $\mathbb{P}(Y = \text{gesund}) = 0.98$,

$\mathbb{P}(X = \text{positiv} \mid Y = \text{krank}) = 0.95$,

$\mathbb{P}(X = \text{positiv} \mid Y = \text{gesund}) = 0.1$,

also $\mathbb{P}(Y = \text{krank} \mid X = \text{positiv}) = \frac{0.02 \cdot 0.95}{0.02 \cdot 0.95 + 0.98 \cdot 0.1} \doteq 0.162$

Stochastische Unabhängigkeit

Definition

Zwei Zufallsgrößen X und Y heißen (stochastisch) unabhängig, wenn für alle Ereignisse $\{X \in A\}$, $\{Y \in B\}$ gilt

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$$

Beispiel:

- Werfen zweier Würfel:
 $X =$ Augenzahl Würfel 1, $Y =$ Augenzahl Würfel 2.

$$\mathbb{P}(X = 2, Y = 5) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathbb{P}(X = 2) \cdot \mathbb{P}(Y = 5)$$

Stochastische Unabhängigkeit

In der Praxis wendet man häufig Resultate an, die Unabhängigkeit einer Stichprobe voraussetzen.

Beispiele:

- Für eine Studie wird eine zufällige Person in München und eine zufällige Person in Hamburg befragt. Die Antworten dürfen als unabhängig voneinander angenommen werden.
- Befragt man zwei Schwestern oder nahe Verwandte (getrennt voneinander), so werden die Antworten nicht unabhängig voneinander sein.

Bernoulli-Verteilung

Als **Bernoulli-Experiment** bezeichnet man jeden zufälligen Vorgang mit exakt zwei möglichen Werten.

Diese werden üblicherweise mit 1 und 0 bezeichnet , beziehungsweise als „Erfolg“ und „Misserfolg“.

Bernoulli-Zufallsgröße X :

Zustandsraum $\mathcal{S} = \{0, 1\}$.

Verteilung:

$$\mathbb{P}(X = 1) = p$$

$$\mathbb{P}(X = 0) = 1 - p$$

Der Parameter $p \in [0, 1]$ heißt **Erfolgswahrscheinlichkeit**.

Bernoulli-Verteilung

Beispiele:

- Münzwurf: mögliche Werte sind „Kopf“ und „Zahl“.
- Hat die gesampelte Drosophila eine Mutation, die weiße Augen verursacht?
Mögliche Antworten sind „Ja“ und „Nein“.
- Das Geschlecht einer Person hat die möglichen Werte „männlich“ und „weiblich“.

Angenommen, ein Bernoulli-Experiment (z.B. Münzwurf zeigt Kopf) mit Erfolgsws p , wird n mal *unabhängig* wiederholt.

Wie groß ist die Wahrscheinlichkeit, dass es...

- ① ...immer gelingt?

$$p \cdot p \cdot p \cdots p = p^n$$

- ② ...immer scheitert?

$$(1 - p) \cdot (1 - p) \cdots (1 - p) = (1 - p)^n$$

- ③ ...erst k mal gelingt und dann $n - k$ mal scheitert?

$$p^k \cdot (1 - p)^{n-k}$$

- ④ ...insgesamt k mal gelingt und $n - k$ mal scheitert?

$$\binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

Erläuterung

$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$ ist die Anzahl der Möglichkeiten, die k Erfolge in die n Versuche einzusortieren.

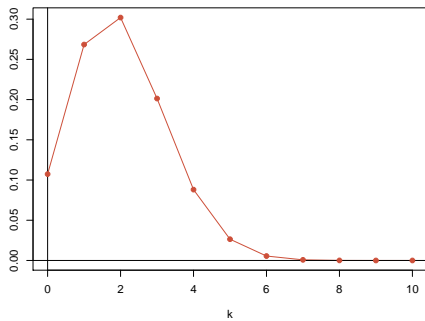
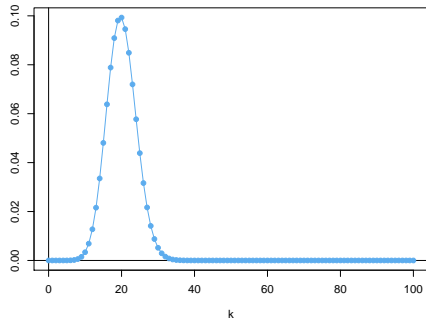
Binomialverteilung

Sei X die Anzahl der Erfolge bei n unabhängigen Versuchen mit Erfolgswahrscheinlichkeit von jeweils p . Dann gilt für $k \in \{0, 1, \dots, n\}$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k \cdot (1 - p)^{n-k}$$

und X heißt *binomialverteilt*, kurz:

$$X \sim \text{bin}(n, p).$$

probabilities of bin($n=10, p=0.2$)probabilities of bin($n=100, p=0.2$)

Definition (Erwartungswert)

Sei X eine Zufallsvariable mit endlichem oder abzählbarem Wertebereich $\mathcal{S} = \{a_1, a_2, a_3 \dots\} \subseteq \mathbb{R}$. Dann ist der *Erwartungswert* von X definiert durch

$$\mathbb{E}X = \sum_{a \in \mathcal{S}} a \cdot \mathbb{P}(X = a)$$

Manchmal schreibt man auch μ_X statt $\mathbb{E}X$.

Ersetzt man in der Definition die Wahrscheinlichkeit durch relative Häufigkeiten, so erhält man die bekannte Formel

$$\text{Erwartungswert} = \frac{\text{Summe der Werte}}{\text{Anzahl der Werte}} :$$

Sei k_a die Häufigkeit des Wertes a in einer Gesamtheit der Größe n , so schreibt sich der Erwartungswert als

$$\mathbb{E}X = \sum_a a \cdot \frac{k_a}{n} = \frac{\sum_a a \cdot k_a}{n} = \frac{\text{Summe der Werte}}{\text{Anzahl der Werte}} .$$

Definition (Erwartungswert)

Sei X eine Zufallsvariable mit endlichem oder abzählbarem Wertebereich $S = \{a_1, a_2, a_3 \dots\} \subseteq \mathbb{R}$. Dann ist der *Erwartungswert* von X definiert durch

$$\mathbb{E}X = \sum_{a \in S} a \cdot \mathbb{P}(X = a)$$

Beispiele:

- Sei X Bernoulli-verteilt mit Erfolgswahrscheinlichkeit $p \in [0, 1]$. Dann gilt

$$\mathbb{E}X = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = \mathbb{P}(X = 1) = p$$

- Sei W die Augenzahl bei einem Würfelwurf. Dann gilt

$$\begin{aligned} \mathbb{E}W &= 1 \cdot \mathbb{P}(W = 1) + 2 \cdot \mathbb{P}(W = 2) + \dots + 6 \cdot \mathbb{P}(W = 6) \\ &= \frac{1 + \dots + 6}{6} = \frac{21}{6} = 3.5 \end{aligned}$$

Erwartungswert und Funktionen

Sei X eine Zufallsvariable mit endlichem Wertebereich $\mathcal{S} \subseteq \mathbb{R}$. Sei $f: \mathcal{S} \rightarrow \mathbb{R}$ eine Funktion. Dann ist der *Erwartungswert* von $f(X)$ gegeben durch

$$\mathbb{E}[f(X)] = \sum_{a \in \mathcal{S}} f(a) \cdot \mathbb{P}(X = a)$$

(Details ggfs. an der Tafel)

Beispiel:

Sei W die Augenzahl bei einem Würfelwurf. Dann gilt

$$\begin{aligned} \mathbb{E}[W^2] &= 1^2 \cdot \mathbb{P}(W = 1) + 2^2 \cdot \mathbb{P}(W = 2) + \dots + 6^2 \cdot \mathbb{P}(W = 6) \\ &= \frac{1^2 + \dots + 6^2}{6} = \frac{91}{6} = 15.\overline{16} \end{aligned}$$

Rechnen mit Erwartungswerten

Satz (Linearität der Erwartung)

Sind X und Y Zufallsvariablen mit Werten in \mathbb{R} und ist $a \in \mathbb{R}$, so gilt:

- $\mathbb{E}(a \cdot X) = a \cdot \mathbb{E}X$
- $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$

Satz (Nur für Unabhängige!)

Sind X und Y **stochastisch unabhängige** Zufallsvariablen mit Werten in \mathbb{R} , so gilt

- $\mathbb{E}(X \cdot Y) = \mathbb{E}X \cdot \mathbb{E}Y.$

(Beweise ggfs. an der Tafel.)

Erwartungswert der Binomialverteilung

Seien Y_1, Y_2, \dots, Y_n die Indikatorvariablen der n unabhängigen Versuche d.h.

$$Y_i = \begin{cases} 1 & \text{falls der } i\text{-te Versuch gelingt} \\ 0 & \text{falls der } i\text{-te Versuch scheitert} \end{cases}$$

Dann ist $X = Y_1 + \dots + Y_n$ binomialverteilt mit Parametern (n, p) , wobei p die Erfolgswahrscheinlichkeit der Versuche ist.

Wegen der Linearität der Erwartung gilt

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}(Y_1 + \dots + Y_n) \\ &= \mathbb{E}Y_1 + \dots + \mathbb{E}Y_n \\ &= p + \dots + p = np \end{aligned}$$

Erwartungswert der Binomialverteilung

Wir halten fest:

$$X \sim \text{bin}(n, p) \Rightarrow \mathbb{E}X = n \cdot p$$

Definition (Varianz, Kovarianz und Korrelation)

Die *Varianz* einer \mathbb{R} -wertigen Zufallsgröße X ist

$$\text{Var}X = \sigma_X^2 = \mathbb{E} [(X - \mathbb{E}X)^2] .$$

$\sigma_X = \sqrt{\text{Var} X}$ ist die *Standardabweichung*.

Ist Y eine weitere reellwertige Zufallsvariable, so ist

$$\text{Cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}X) \cdot (Y - \mathbb{E}Y)]$$

die *Kovarianz* von X und Y .

Die *Korrelation* von X und Y ist

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} .$$

Die Varianz

$$\text{Var}X = \mathbb{E} [(X - \mathbb{E}X)^2]$$

ist die mittlere quadrierte Abweichung vom Mittelwert.

Die Korrelation

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

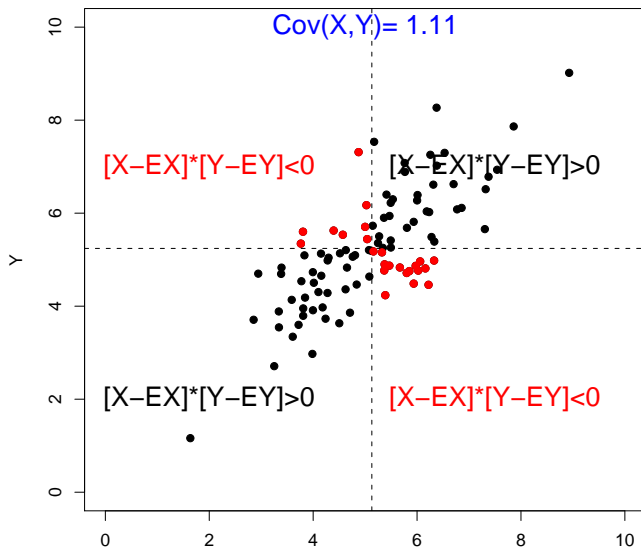
liegt immer im Intervall $[-1, 1]$. Die Variablen X und Y sind

- **positiv korreliert**, wenn X und Y tendenziell entweder beide überdurchschnittlich große Werte oder beide unterdurchschnittlich große Werte annehmen.
- **negativ korreliert**, wenn X und Y tendenziell auf verschiedenen Seiten ihrer Erwartungswerte liegen.

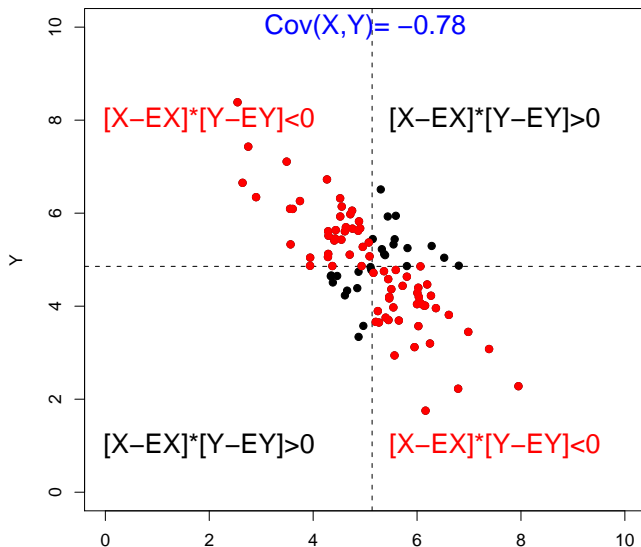
Sind X und Y unabhängig, so sind sie auch **unkorreliert**, d.h.

$$\text{Cor}(X, Y) = 0.$$

Wieso $\text{Cov}(X, Y) = \mathbb{E}([X - \mathbb{E}X][Y - \mathbb{E}Y])$?



Wieso $\text{Cov}(X, Y) = \mathbb{E}([X - \mathbb{E}X][Y - \mathbb{E}Y])$?



Beispiel: Die empirische Verteilung

Sind $x_1, \dots, x_n \in \mathbb{R}$ Daten und entsteht X durch rein zufälliges Ziehen aus diesen Daten, so gilt:

$$\mathbb{E}X = \bar{x}$$

und

$$\text{Var } X = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Entsprechend kann man auch für Daten (x_i, y_i) die empirischen Kovarianzen und Korrelationen ausrechnen, siehe nächste Seite...

$$\sigma_X = 0.95, \sigma_Y = 0.92$$

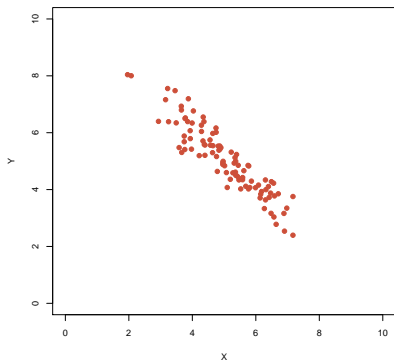
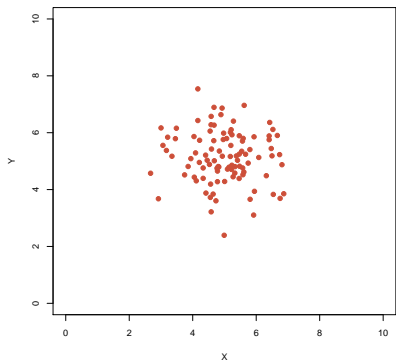
$$\text{Cov}(X, Y) = -0.06$$

$$\text{Cor}(X, Y) = -0.069$$

$$\sigma_X = 1.13, \sigma_Y = 1.2$$

$$\text{Cov}(X, Y) = -1.26$$

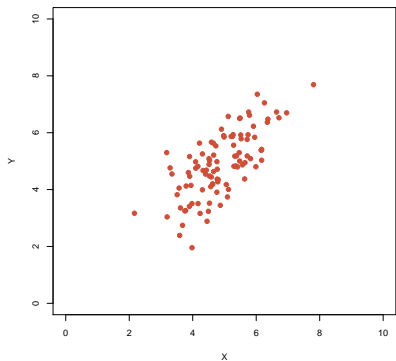
$$\text{Cor}(X, Y) = -0.92$$



$$\sigma_X = 1.14, \sigma_Y = 0.78$$

$$\text{Cov}(X, Y) = 0.78$$

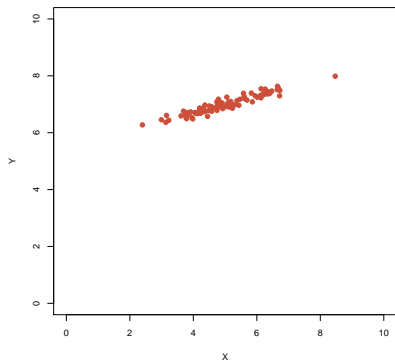
$$\text{Cor}(X, Y) = 0.71$$



$$\sigma_X = 1.03, \sigma_Y = 0.32$$

$$\text{Cov}(X, Y) = 0.32$$

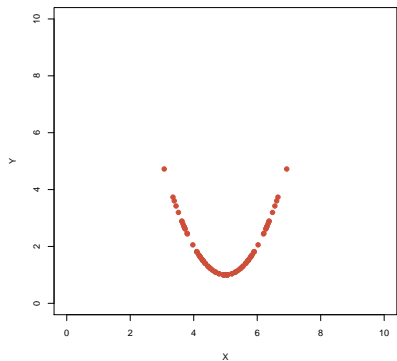
$$\text{Cor}(X, Y) = 0.95$$



$$\sigma_X = 0.91, \sigma_Y = 0.88$$

$$\text{Cov}(X, Y) = 0$$

$$\text{Cor}(X, Y) = 0$$



Rechenregeln für Varianzen

$$\text{Var}X = \mathbb{E}[(X - \mathbb{E}X)^2]$$

- $\text{Var}X = \text{Cov}(X, X)$
- $\text{Var}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2$
- $\text{Var}(a \cdot X) = a^2 \cdot \text{Var}X$
- $\text{Var}(X + Y) = \text{Var}X + \text{Var}Y + 2 \cdot \text{Cov}(X, Y)$
- $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \cdot \sum_{j=1}^n \sum_{i=1}^{j-1} \text{Cov}(X_i, X_j)$
- Sind (X, Y) stochastisch unabhängig, so folgt:

$$\text{Var}(X + Y) = \text{Var}X + \text{Var}Y$$

Rechenregeln für Kovarianzen

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X) \cdot (Y - \mathbb{E}Y)]$$

- Sind X und Y unabhängig, so folgt $\text{Cov}(X, Y) = 0$ (die Umkehrung gilt nicht!)
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}X \cdot \mathbb{E}Y$
- $\text{Cov}(a \cdot X, Y) = a \cdot \text{Cov}(X, Y) = \text{Cov}(X, a \cdot Y)$
- $\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$
- $\text{Cov}(X, Z + Y) = \text{Cov}(X, Z) + \text{Cov}(X, Y)$

Die letzten drei Regeln beschreiben die Bilinearität der Kovarianz.

Rechenregeln für die Korrelation

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

- $-1 \leq \text{Cor}(X, Y) \leq 1$
- $\text{Cor}(X, Y) = \text{Cor}(Y, X)$
- $\text{Cor}(X, Y) = \text{Cov}(X/\sigma_X, Y/\sigma_Y)$
- $\text{Cor}(X, Y) = 1$ genau dann wenn Y eine wachsende, affin-lineare Funktion von X ist, d.h. falls es $a > 0$ und $b \in \mathbb{R}$ gibt, so dass $Y = a \cdot X + b$
- $\text{Cor}(X, Y) = -1$ genau dann wenn Y eine fallende, affin-lineare Funktion von X ist, d.h. falls es $a < 0$ und $b \in \mathbb{R}$ gibt, so dass $Y = a \cdot X + b$

Mit diesen Rechenregeln können wir nun beweisen:

Satz

Sind X_1, X_2, \dots, X_n unabhängige \mathbb{R} -wertige Zufallsgrößen mit Mittelwert μ und Varianz σ^2 , so gilt für $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$:

$$\mathbb{E}\bar{X} = \mu$$

und

$$\text{Var } \bar{X} = \frac{1}{n} \sigma^2,$$

d.h.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Beweis: Linearität des Erwartungswertes impliziert

$$\begin{aligned}\mathbb{E}\bar{X} &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \mu.\end{aligned}$$

Die Unabhängigkeit der X_i vereinfacht die Varianz zu

$$\begin{aligned}\text{Var } \bar{X} &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n} \sigma^2\end{aligned}$$

Bernoulli-Verteilung

Eine Bernoulli-verteilte Zufallsvariable Y mit Erfolgsws $p \in [0, 1]$ hat Erwartungswert

$$\mathbb{E}Y = p$$

und Varianz

$$\text{Var } Y = p \cdot (1 - p)$$

Beweis: Aus $\mathbb{P}(Y = 1) = p$ und $\mathbb{P}(Y = 0) = (1 - p)$ folgt

$$\mathbb{E}Y = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Varianz:

$$\begin{aligned}\text{Var } Y &= \mathbb{E}(Y^2) - (\mathbb{E}Y)^2 \\ &= 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p \cdot (1 - p)\end{aligned}$$

Binomialverteilung

Seien nun Y_1, \dots, Y_n unabhängig und Bernoulli-verteilt mit Erfolgsws p . Dann gilt

$$\sum_{i=1}^n Y_i =: X \sim \text{bin}(n, p)$$

und es folgt:

$$\text{Var } X = \text{Var} \left(\sum_{i=1}^n Y_i \right) = \sum_{i=1}^n \text{Var } Y_i = n \cdot p \cdot (1 - p)$$

Binomialverteilung

Satz (Erwartungswert und Varianz der Binomialverteilung)

Ist X binomialverteilt mit Parametern (n, p) , so gilt:

$$\mathbb{E}X = n \cdot p$$

und

$$\text{Var } X = n \cdot p \cdot (1 - p)$$

In einem Genom werde die Aminosäure Prolin
101844 mal durch CCT und 106159 mal durch CCA codiert.

Ist es nur vom reinen Zufall abhängig,
welches Codon verwendet wird?

Dann wäre die Anzahl X der CCT binomialverteilt mit $p = \frac{1}{2}$ und
 $n = 101844 + 106159 = 208003$.

$$\mathbb{E}X = n \cdot p = 104001.5$$

$$\sigma_X = \sqrt{n \cdot p \cdot (1 - p)} \approx 228$$

$$104001.5 - 101844 = 2157.5 \approx 9.5 \cdot \sigma_X$$

Sieht das nach Zufall aus?

Die Frage ist:

Wie groß ist die Wahrscheinlichkeit
einer Abweichung vom Erwartungswert
von mindestens $\approx 9.5 \cdot \sigma_X$, wenn alles Zufall ist?

Wir müssen also

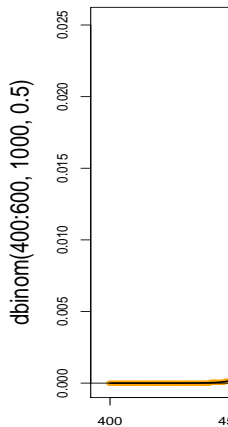
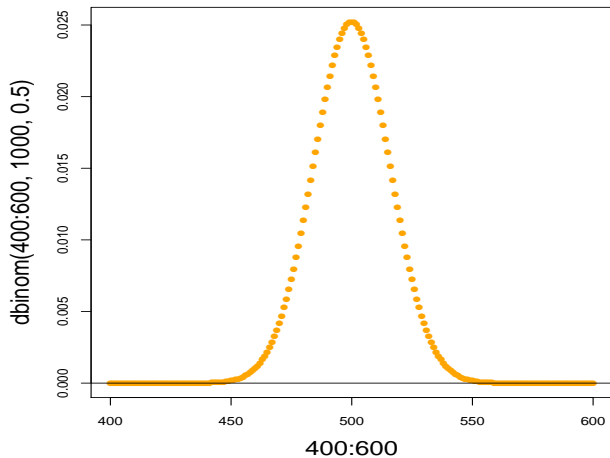
$$\mathbb{P}(|X - \mathbb{E}X| \geq 9.5\sigma_X)$$

berechnen.

Das Problem bei der Binomialverteilung ist: $\binom{n}{k}$ exakt zu berechnen, ist für große n sehr aufwändig. Deshalb:

Die Binomialverteilung wird oft
durch andere Verteilungen approximiert.

Die Binomialverteilung mit großer Versuchszahl n sieht (nahezu) aus wie die Normalverteilung:

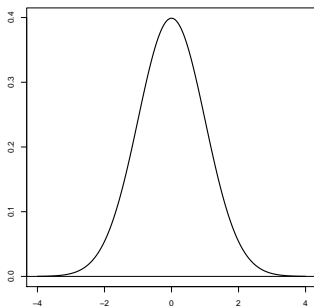


Dichte der Standardnormalverteilung

Eine Zufallsvariable Z mit der Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

“Gauß-Glocke”



kurz:

$$Z \sim \mathcal{N}(0, 1)$$

$$\mathbb{E}Z = 0$$

$$\text{Var } Z = 1$$

heißt *standardnormalverteilt*.

Ist $Z \mathcal{N}(0, 1)$ -verteilt, so ist $X = \sigma \cdot Z + \mu$ normalverteilt mit Mittelwert μ und Varianz σ^2 , kurz:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

X hat dann die Dichte

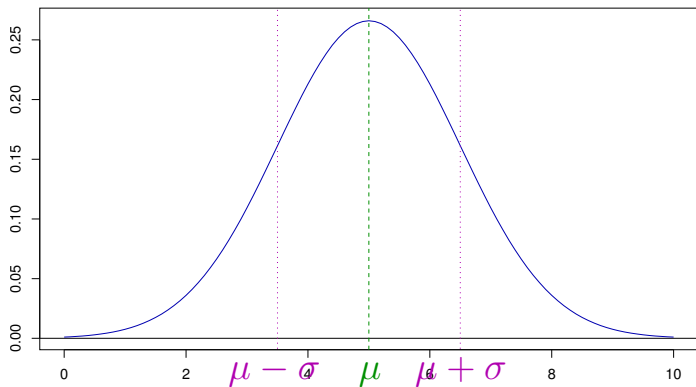
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Merkregeln

Ist $Z \sim \mathcal{N}(\mu, \sigma^2)$, so gilt:

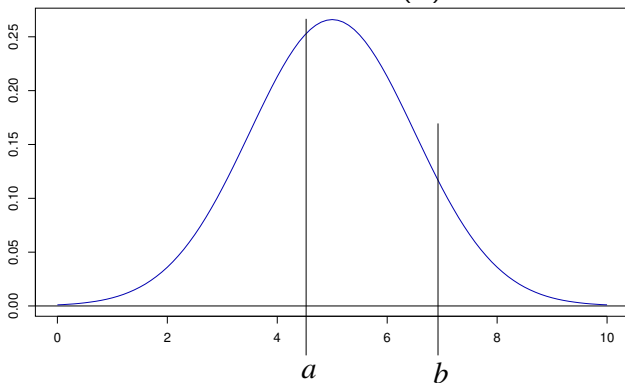
- $\mathbb{P}(|Z - \mu| > \sigma) \approx 33\%$
- $\mathbb{P}(|Z - \mu| > 1.96 \cdot \sigma) \approx 5\%$
- $\mathbb{P}(|Z - \mu| > 3 \cdot \sigma) \approx 0.3\%$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Dichten brauchen Integrale

Sei Z eine Zufallsvariable mit Dichte $f(x)$.

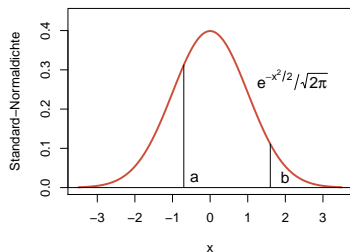


Dann gilt

$$\mathbb{P}(Z \in [a, b]) = \int_a^b f(x) dx.$$

Die Standardnormal-Dichte und das Gauß'sche Fehlerintegral Φ :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



Sei Z standard-normalverteilt:

$$\Phi(a) = \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx, \text{ also für } a \in \mathbb{R}$$

$$\mathbb{P}(Z \leq a) = \Phi(a), \text{ und für } a < b$$

$$\mathbb{P}(Z \in [a, b]) = \mathbb{P}(Z \leq b) - \mathbb{P}(Z \leq a)$$

Die Funktion Φ läßt sich nicht durch elementare Integration ausdrücken, sie wird numerisch bestimmt; man entnimmt in der Praxis den Wert einem Computerprogramm oder einer Tabelle.

Die Normalverteilung in R

Die Normalverteilung hat in R das Kürzel 'norm'.

Es gibt 4 R-Befehle:

`dnorm()`: Dichte der Normalverteilung (**d**ensity)

`rnorm()`: Ziehen einer Stichprobe (**r**andom sample)

`pnorm()`: Verteilungsfunktion der Normalverteilung (**p**robability)

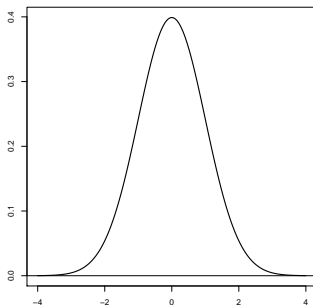
`qnorm()`: Quantilfunktion der Normalverteilung (**q**uantile)

Beispiel: Dichte der Standardnormalverteilung:

```
> dnorm(0)
```

```
[1] 0.3989423
```

```
> plot(dnorm,from=-4,to=4)
```



```
> dnorm(0,mean=1,sd=2)
```


Beispiel: Ziehen einer Stichprobe

Ziehen einer Stichprobe der Länge 6 aus einer Standardnormalverteilung:

```
> rnorm(6)
[1] -1.24777899 0.03288728 0.19222813 0.81642692
-0.62607324 -1.09273888
```

Ziehen einer Stichprobe der Länge 7 aus einer Normalverteilung mit Mittelwert 5 und Standardabweichung 3:

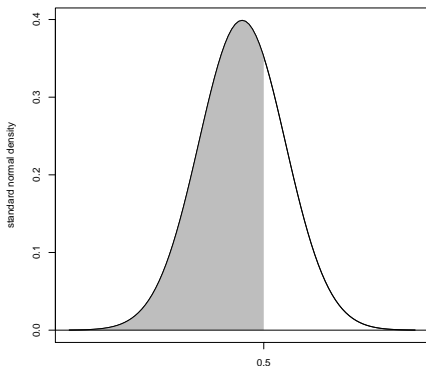
```
> rnorm(7,mean=5,sd=3)
[1] 2.7618897 6.3224503 10.8453280 -0.9829688 5.6143127
0.6431437 8.123570
```

Beispiel: Berechnung von Wahrscheinlichkeiten:

Sei $Z \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$, also standardnormalverteilt.

$\mathbb{P}(Z < a)$ berechnet man in R mit `pnorm(a)`

```
> pnorm(0.5) [1] 0.6914625
```



Beispiel: Berechnung von Wahrscheinlichkeiten:

Sei $Z \sim \mathcal{N}(\mu = 5, \sigma^2 = 2.25)$.

Berechnung von $\mathbb{P}(Z \in [3, 4])$:

$$\mathbb{P}(Z \in [3, 4]) = \mathbb{P}(Z < 4) - \mathbb{P}(Z < 3)$$

```
> pnorm(4, mean=5, sd=1.5) - pnorm(3, mean=5, sd=1.5)
[1] 0.1612813
```

Sei $Z \sim \mathcal{N}(\mu, \sigma^2)$.

Frage: Wie berechnet man $\mathbb{P}(Z = 5)$?

Antwort: Für jedes $x \in \mathbb{R}$ gilt $\mathbb{P}(Z = x) = 0$

Was wird dann aus $\mathbb{E}Z = \sum_{a \in \mathcal{S}} a \cdot \mathbb{P}(Z = a)$?

Muss reformiert werden:

$$\mathbb{E}Z = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Aber zum Glück kennen wir schon das Ergebnis $\mathbb{E}Z = \mu$.

Normalapproximation

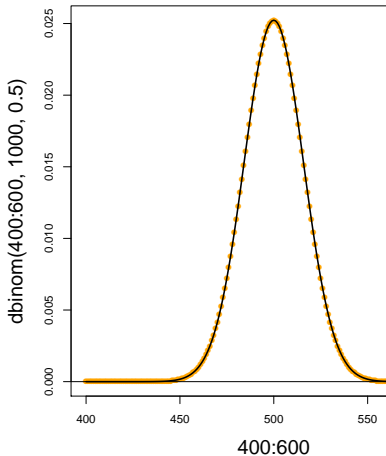
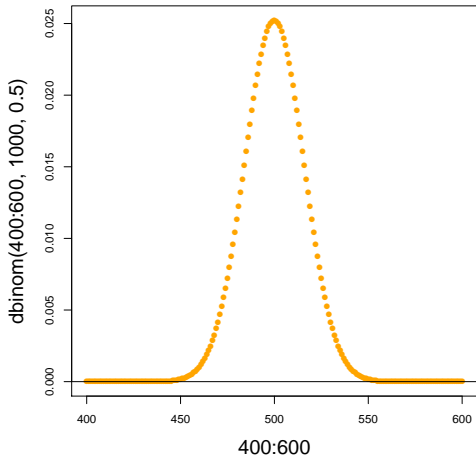
Für große n und p , die nicht zu nahe bei 0 oder 1 liegen, kann man die Binomialverteilung durch die Normalverteilung mit dem entsprechenden Erwartungswert und der entsprechenden Varianz approximieren:

Ist $X \sim \text{bin}(n, p)$ und $Z \sim \mathcal{N}(\mu, \sigma^2)$
mit $\mu = n \cdot p$ und $\sigma^2 = n \cdot p \cdot (1 - p)$
so gilt

$$\mathbb{P}(X \in [a, b]) \approx \mathbb{P}(Z \in [a, b])$$

(eine Faustregel: für den Hausgebrauch meist okay, wenn $n \cdot p \cdot (1 - p) \geq 9$)

$n = 1000, p = 0.5, n \cdot p \cdot (1 - p) = 250$
 $n = 10, p = 0.2,$
 $n \cdot p \cdot (1 - p) = 1.6$



Zentraler Grenzwertsatz

Ein anderer Ausdruck für *Normalapproximation* ist
Zentraler Grenzwertsatz.

Der zentrale Grenzwertsatz besagt,
dass die Verteilung von Summen
unabhängiger und identisch verteilter
Zufallsvariablen in etwa
die Normalverteilung ist.

Zentraler Grenzwertsatz

Die \mathbb{R} -wertigen Zufallsgrößen X_1, X_2, \dots seien unabhängig und identisch verteilt mit endlicher Varianz $0 < \sigma^2 = \text{Var } X_i < \infty$. Sei außerdem

$$Z_n := X_1 + X_2 + \dots + X_n$$

die Summe der ersten n Variablen. Dann ist die zentrierte und reskalierte Summe im Limes $n \rightarrow \infty$ standardnormalverteilt, d.h.

$$Z_n^* := \frac{Z_n - \mathbb{E}Z_n}{\sqrt{\text{Var } Z_n}} \text{ ist ungefähr } \sim \mathcal{N}(0, 1)$$

für große n . Formal: Es gilt für alle $-\infty \leq a < b \leq \infty$

$$\lim_{n \rightarrow \infty} \mathbb{P}(a \leq Z_n^* \leq b) = \mathbb{P}(a \leq Z \leq b),$$

wobei Z eine standardnormalverteilte Zufallsvariable ist.

Anders formuliert: Für große n gilt (approximativ)

$$Z_n \sim \mathcal{N}(\mu, \sigma^2) \quad \text{mit } \mu = \mathbb{E}Z_n, \sigma^2 = \text{Var } Z_n$$

Die Voraussetzungen „unabhängig“ und „identisch verteilt“ lassen sich noch deutlich abschwächen.

Für den Hausgebrauch:

Ist Y das Resultat von vielen kleinen Beiträgen, die großteils unabhängig voneinander sind, so ist Y in etwa normalverteilt,

d.h.

$$Y \sim \mathcal{N}(\mu, \sigma^2), \quad \text{mit } \mu = \mathbb{E}Y, \sigma^2 = \text{Var } Y$$

Zurück zu dem Beispiel mit den
Prolin-Codons im Genom-Beispiel

CCT kommt $k = 101844$ mal vor
CCA kommt $n - k = 106159$ mal vor

Frage: Kann dies Zufall sein?

Wir meinen: Nein.

Die Skeptiker sagen:

„Nur Zufall.“

Die Hypothese

Reiner Zufall
Kein Unterschied

nennt man die
Nullhypothese.

Um die Skeptiker zu überzeugen,
müssen wir die
Nullhypothese entkräften
d.h. zeigen, dass unter der Nullhypothese
die Beobachtung sehr unwahrscheinlich ist.

CCT kommt $k = 101844$ mal vor
CCA kommt $n - k = 106159$ mal vor

Unter der Nullhypothese „alles nur Zufall“
ist die Anzahl X der CCT $\text{bin}(n, p)$ -verteilt
mit $n = 208003$ und $p = 0.5$.

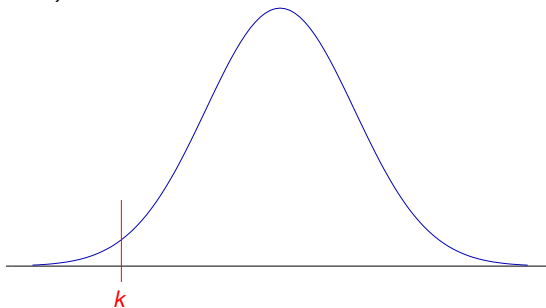
Normalapproximation: X ist ungefähr $\mathcal{N}(\mu, \sigma^2)$ -verteilt mit

$$\mu = n \cdot p = 104001.5 \approx 104000$$

und

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)} = 228.0367 \approx 228$$

Frage: Ist es plausibel, dass eine Größe X , die den Wert $k = 101844$ angenommen hat, ungefähr $\mathcal{N}(104000, 228^2)$ -verteilt ist?

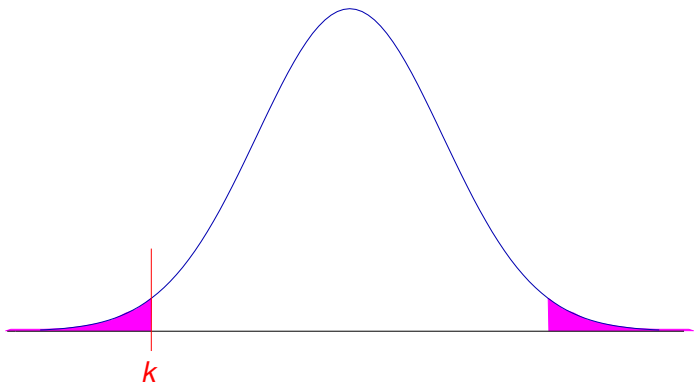


Wenn diese Nullhypothese H_0 gilt, dann folgt

$$\mathbb{P}(X = 101844) = 0$$

Aber das bedeutet nichts, denn $\mathbb{P}(X = k) = 0$ gilt für jeden Wert k !

Entscheidend ist die Wahrscheinlichkeit,
dass X (unter Annahme der H_0) einen
mindestens so extremen Wert wie k annimmt:



$$\mathbb{P}(|X - \mu| \geq |k - \mu|) = \mathbb{P}(|X - \mu| \geq 2156) \approx \mathbb{P}(|X - \mu| \geq 9.5 \cdot \sigma)$$

Wir wissen bereits:

$$Pr(|X - \mu| \geq 3 \cdot \sigma) \approx 0.003 \quad (\text{siehe Merkgeln!})$$

Also muss $\mathbb{P}(|X - \mu| \geq 9.5 \cdot \sigma)$ extrem klein sein.

In der Tat:

Mit Normalapproximation ist

$$\mathbb{P}(|X - \mu| \geq 9.5 \cdot \sigma) \cong \mathbb{P}(|Z| \geq 9.5) \doteq 3 \cdot 10^{-21} \quad (\text{wo } Z \sim \mathcal{N}(0, 1))$$

(Der exakte Wert für die Binomialverteilung mit $n = 208003$, $p = 1/2$ ist $3,1 \cdot 10^{-21}$.)

Wir können also argumentieren,
dass eine derartig starke Abweichung vom Erwartungswert
nur durch einen extremen Zufall zu erklären ist.

Wir werden also die
Nullhypothese “alles nur Zufall” verwerfen
und nach alternativen Erklärungen suchen,
etwa unterschiedliche Effizienz von CCA und CCT
oder unterschiedliche Verfügbarkeit von A und T.

Zusammenfassung z-Test

Nullhypothese H_0 (möchte man meistens verwerfen): der beobachtete Wert x kommt aus einer Normalverteilung mit Mittelwert μ und **bekannter** Varianz σ^2 .

p -Wert $= \mathbb{P}(|X - \mu| \geq |x - \mu|)$, wobei $X \sim \mathcal{N}(\mu, \sigma^2)$, also die Wahrscheinlichkeit einer *mindestens* so großen Abweichung wie der beobachteten.

Signifikanzniveau α : oft 0.05. Wenn der p -Wert kleiner ist als α , verwerfen wir die Nullhypothese auf dem Signifikanzniveau α und suchen nach einer alternativen Erklärung.

Grenzen des z-Tests

Der z-Test kann nur angewendet werden, wenn die Varianz der Normalverteilung bekannt ist oder zumindest in der Nullhypothese als bekannt angenommen wird.

Das ist meistens nicht der Fall, wenn die Normalverteilung beim statistischen Testen verwendet wird:

Meistens wird die Varianz aus den Daten geschätzt. Dann wird anstelle des z-Tests der (wesentlich berühmtere)

t-Test

angewendet.

Tabelle: Verteilungsfunktion Φ der Standard-Normalverteilung

x	0	1	2	3	4	5	6	7	8	9
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.40	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.50	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.60	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.70	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.80	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8079	.8106	.8133
0.90	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.00	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.10	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.20	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.30	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.40	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.50	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.60	.9452	.9463	.9474	.9485	.9495	.9505	.9515	.9525	.9535	.9545
1.70	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.80	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.90	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9762	.9767
2.00	.9773	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.10	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.20	.9861	.9865	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.30	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.70	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.80	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9980	.9980	.9981
2.90	.9981	.9982	.9983	.9983	.9984	.9984	.9985	.9985	.9986	.9986

$Z \sim \mathcal{N}(0, 1)$, so ist $\mathbb{P}(Z \leq x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(x)$

Beispiel: $\Phi(1.74) = 0.9591$. Für $x < 0$ verwende $\Phi(-x) = 1 - \Phi(x)$