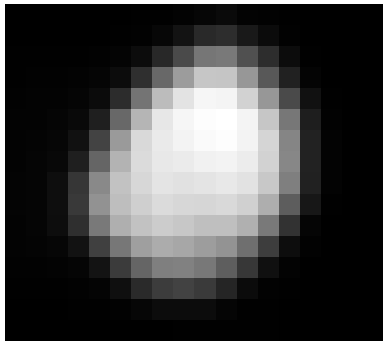


Differentielle Genexpression

Genomische Datenanalyse – Kapitel 17

Genexpression

Aus dem Kapitel über Diagnostische Markergene kennen wir **Microarrays und Genexpressionsdaten**.



- Damit haben wir Patienten klassifiziert.

- Wir haben ein Diagnosemodell entwickelt, das einen Patienten aufgrund seiner Genexpression einer Klasse zuordnet (z.B. Therapiekategorie A).

- Dazu haben wir vorausgesetzt, dass sich die Klassen in ihrer Genexpression unterscheiden (sonst könnten wir nicht klassifizieren). **Können wir diese Annahme überprüfen?**

Differentiell expremierte Gene

Ein Gen, dessen Genexpression sich zwischen zwei Klassen unterscheidet, nennt man **differentiell expremiert**. Es ist von klinischer Bedeutung, denn es scheint ja ein besonderes Merkmal der Klassen(einteilung) zu sein.



Differentielles Gen → das Gen unterliegt einem Regulierungsmechanismus

→ Verstehen der biologischen Prozesse



→ Therapie?

Fragen:

- 1. Wie finden wir heraus, ob sich die Genexpression unterscheidet?**
- 2. Wie stark muss sie sich unterscheiden, damit wir uns auf das Ergebnis verlassen können?**

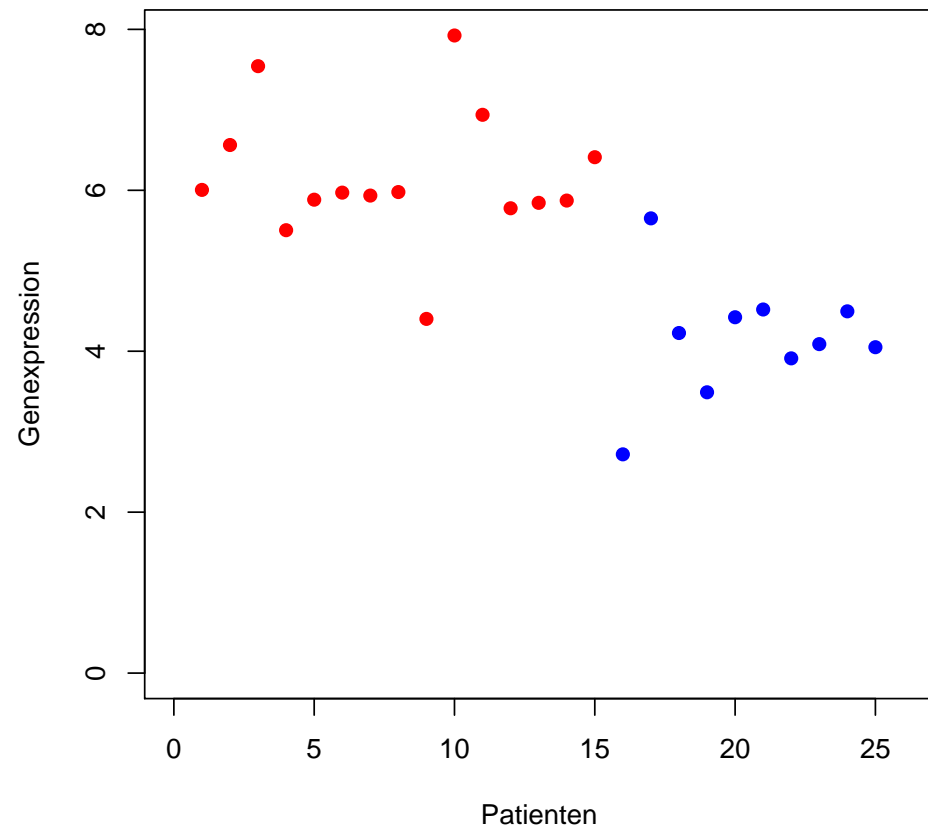


Unterscheidet sich die Genexpression?

Wir haben ein Sample von 25 Patienten.

15 gehören in die Risikoklasse A und 10 in die Risikoklasse B.

Für jeden Patienten haben wir die Genexpression X eines einzelnen Genes gemessen.



1. Versuch: Vergleich der Mittelwerte

Wenn sich die Genexpressionslevel zwischen den beiden Klassen unterscheiden, dann müssen sich auch die Klassenmittelwerte unterscheiden!

Wir berechnen die **Differenz der Klassenmittelwerte**:

$$\bar{x}_A - \bar{x}_B = \frac{1}{15} \sum_{i=1}^{15} x_i - \frac{1}{10} \sum_{i=16}^{25} x_i \approx 2.$$



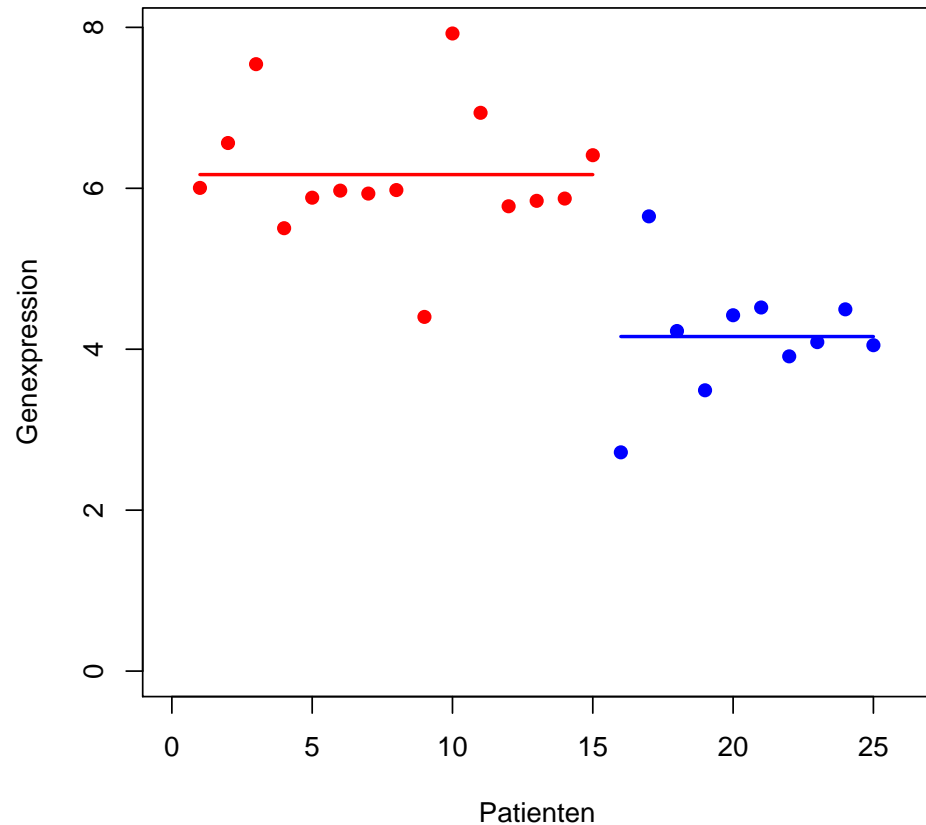
Ist der Unterschied groß genug?



Ist der Unterschied groß genug?

Wir beobachten einen Unterschied von ca. 2.

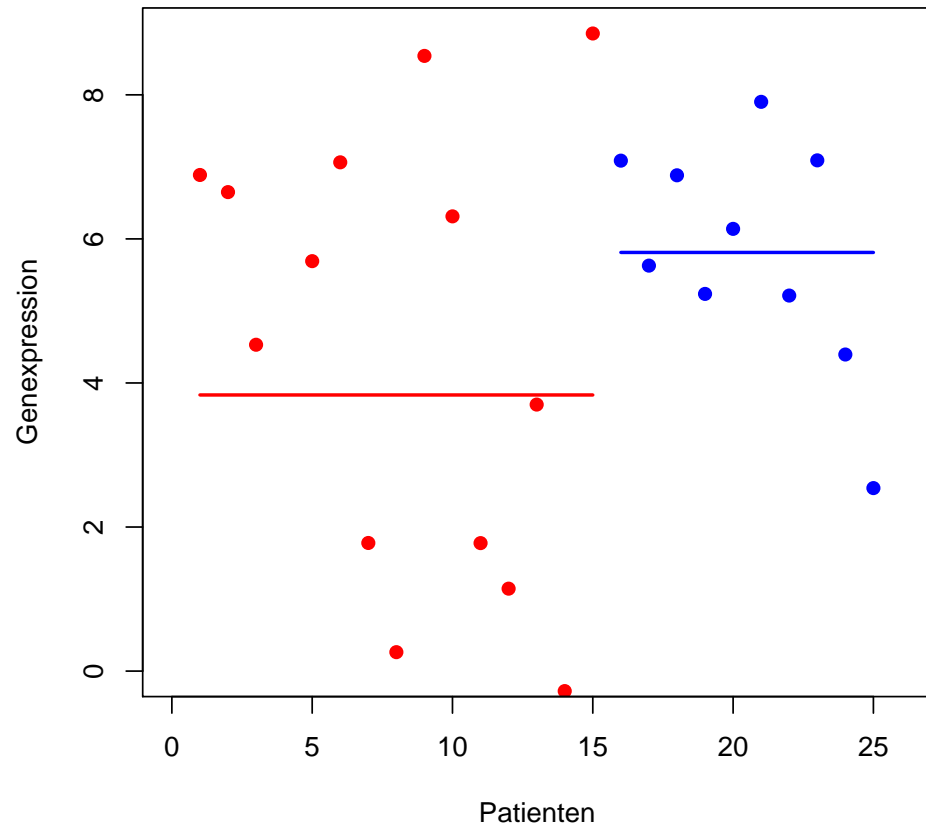
Die Datenpunkte streuen wenig um den Mittelwert. Die Vermutung liegt nahe, dass das Gen differentiell expremiert ist.



Was kann schiefgehen?

Bei den Genexpressionswerten eines anderen Gens beobachten wir ebenfalls einen Unterschied von ca. 2.

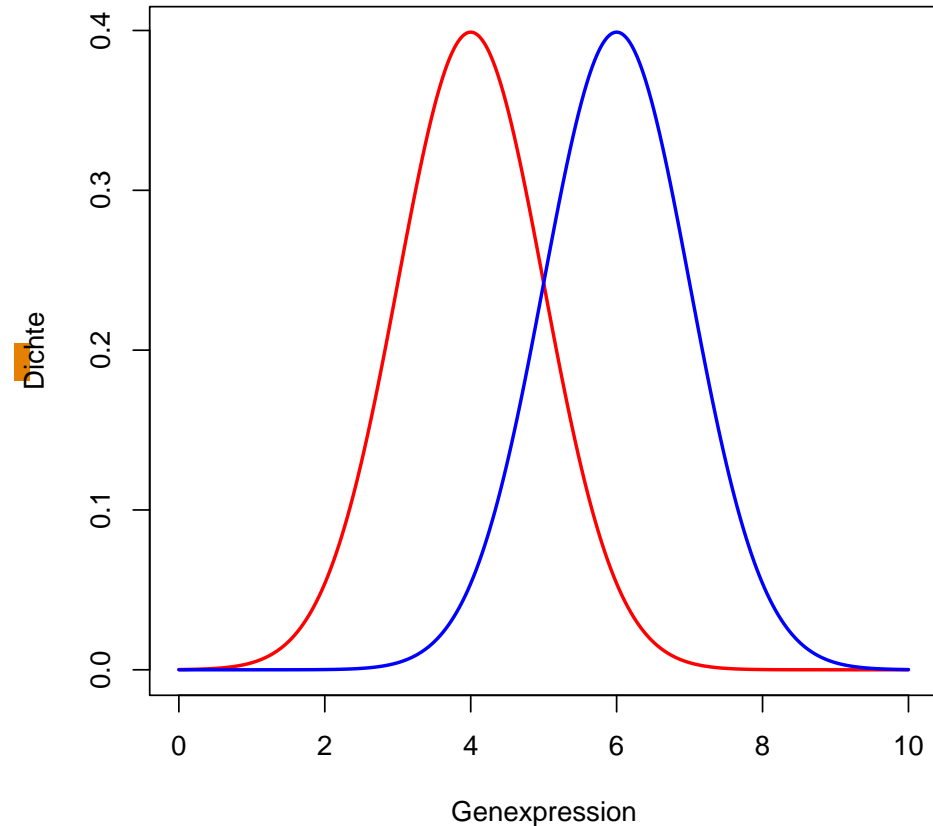
Die Datenpunkte streuen stärker um den Mittelwert. Würden Sie dieses Gen als differentiell exprimiert bezeichnen?



Wir sollten also die **Varianzen in den Klassen** beachten. Wie geht das?

■
Nehmen wir mal an, die Genexpression ist in den beiden Klassen jeweils i.i.d. normalverteilt:

$$X_1, \dots, X_{15} \sim N(\mu_A, \sigma_A^2) \text{ und} \\ X_{16}, \dots, X_{25} \sim N(\mu_B, \sigma_B^2).$$



$$X_1, \dots, X_{15} \sim N(\mu_A, \sigma_A^2) \quad \text{und} \quad X_{16}, \dots, X_{25} \sim N(\mu_B, \sigma_B^2)$$

μ_A und μ_B sind die **unbekannten wahren Klassenmittel**, über die wir mit Hilfe der empirischen Klassenmittel \bar{x}_A und \bar{x}_B eine Aussage treffen wollen.

■
 σ_A^2 und σ_B^2 sind die **wahren Klassenvarianzen**. Über die wollen wir keine Aussagen treffen, trotzdem müssen wir uns gleich noch mehr Gedanken über sie machen.

■
Nehmen wir erstmal an, diese Varianzen seien uns **bekannt**. Bei Genexpressionsdaten wissen wir normalerweise nicht, wie groß die Varianzen sind.

Über die **Verteilung des Mittelwertes** wissen Sie schon einiges . . .



. . . auch der Mittelwert ist normalverteilt:

$$\bar{X}_A \sim N\left(\mu_A, \frac{\sigma_A^2}{15}\right) \quad \text{und} \quad \bar{X}_B \sim N\left(\mu_B, \frac{\sigma_B^2}{10}\right)$$



Gut, wir betrachten aber die Differenz der Mittelwerte . . .



Summen i.i.d. normalverteilter Zufallsvariablen sind wieder normalverteilt. Das haben wir gerade bei den Mittelwerten ausgenutzt.



Differenzen sind im Prinzip auch Summen, nur eben mit speziellen Vorzeichen. Daher ist auch die **Differenz der Klassenmittel** normalverteilt mit:

$$\begin{aligned} E [\bar{X}_A - \bar{X}_B] &= E [\bar{X}_A] - E [\bar{X}_B] \\ &= \mu_A - \mu_B \end{aligned}$$

$$\begin{aligned} \mathbf{Var} [\bar{X}_A - \bar{X}_B] &= \mathbf{Var} [\bar{X}_A + (-1) \cdot \bar{X}_B] \\ &= \mathbf{Var} [\bar{X}_A] + (-1)^2 \mathbf{Var} [\bar{X}_B] \\ &= \frac{\sigma_A^2}{15} + \frac{\sigma_B^2}{10} \end{aligned}$$



Also $\bar{X}_A - \bar{X}_B \sim N\left(\mu_A - \mu_B, \frac{\sigma_A^2}{15} + \frac{\sigma_B^2}{10}\right)$.

Gut, dann wissen wir, wie die Differenz verteilt ist.

Wir **erwarten**, dass die Klassenmittel gleich sind und setzen daher $\mu_A - \mu_B = 0$.

Im nächsten Schritt **standardisieren** wir die Differenz. Das hat den Vorteil, dass wir mit der Standardnormalverteilung weiterarbeiten können.

$$\bar{X}_A - \bar{X}_B \sim N\left(0, \frac{\sigma_A^2}{15} + \frac{\sigma_B^2}{10}\right) \quad \Rightarrow \quad T := \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\sigma_A^2}{15} + \frac{\sigma_B^2}{10}}} \sim N(0, 1)$$

T ist wieder eine Zufallsvariable. Schauen wir uns den Nenner mal genauer an:



Wenn die Varianzen σ_A^2 und σ_B^2 **beide klein** sind, dann wird T **groß**.



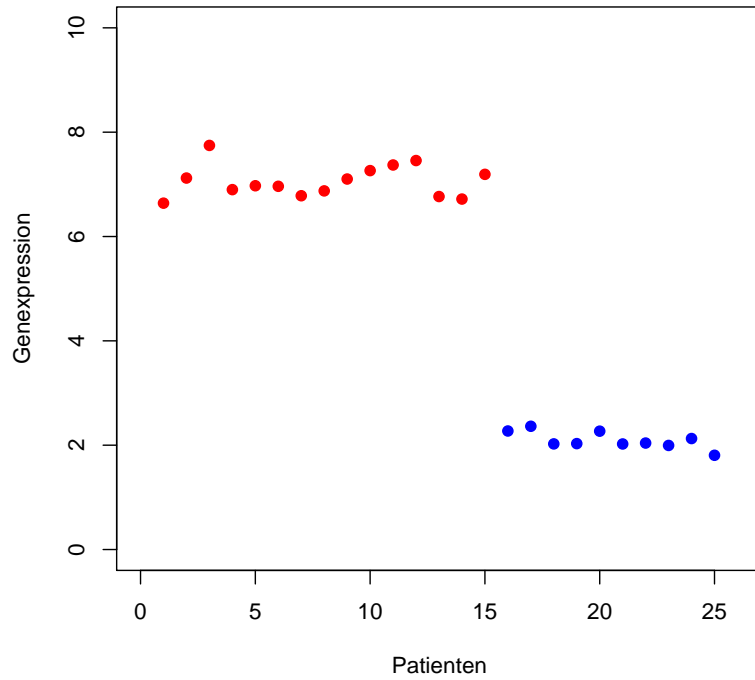
Wenn die Varianzen **beide groß** sind, dann wird T **klein**.



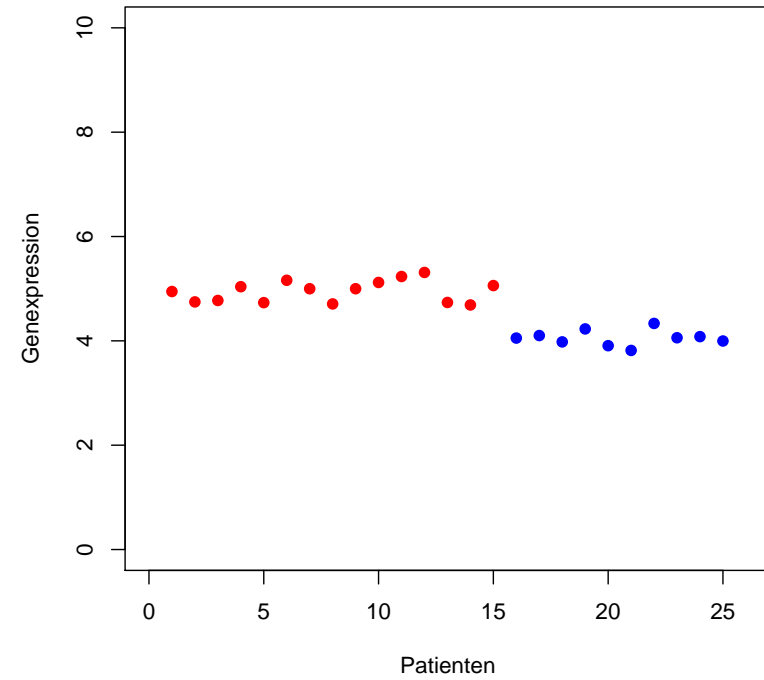
Haben wir damit die Varianzen sinnvoll eingebaut?

Ja, denn wir wollen ja entscheiden, ob sich die Genexpression zwischen den beiden Klassen “signifikant” unterscheidet. Folgende Situationen sind denkbar . . .



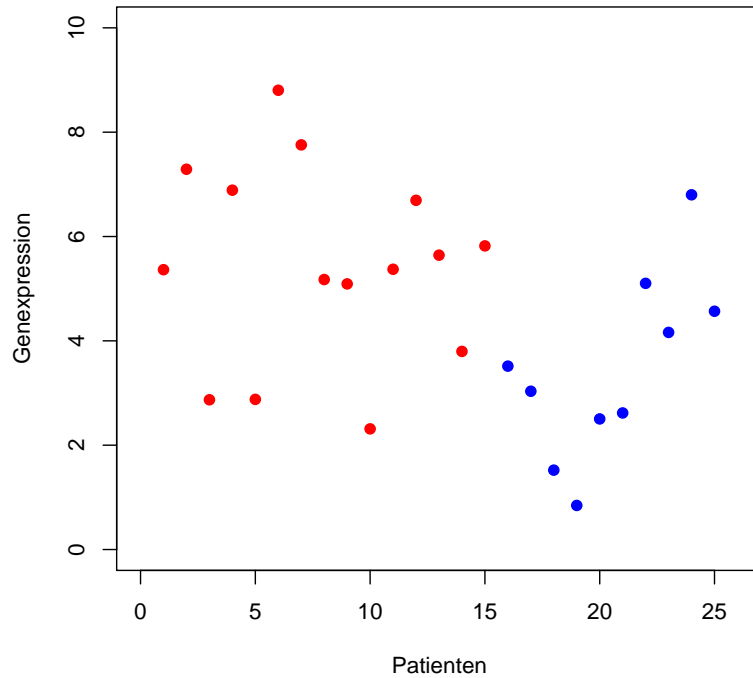


Große Differenz, kleine Varianz
 $\Rightarrow T$ "riesig"



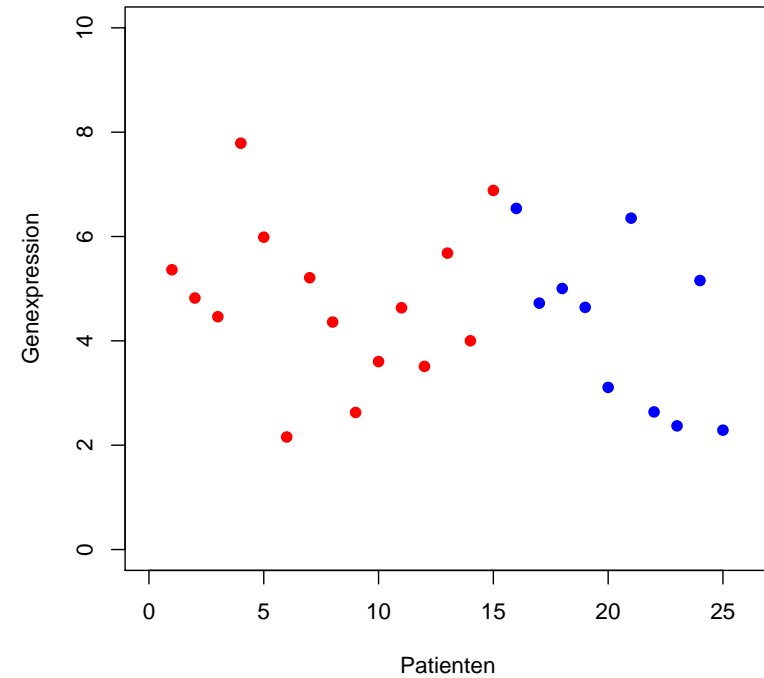
Kleine Differenz, kleine Varianz
 $\Rightarrow T$ groß





Große Differenz, große Varianz

⇒ T klein



Kleine Differenz, große Varianz

⇒ $T \approx 0$

Die interessanten Fälle sind also die, in denen T groß wird.

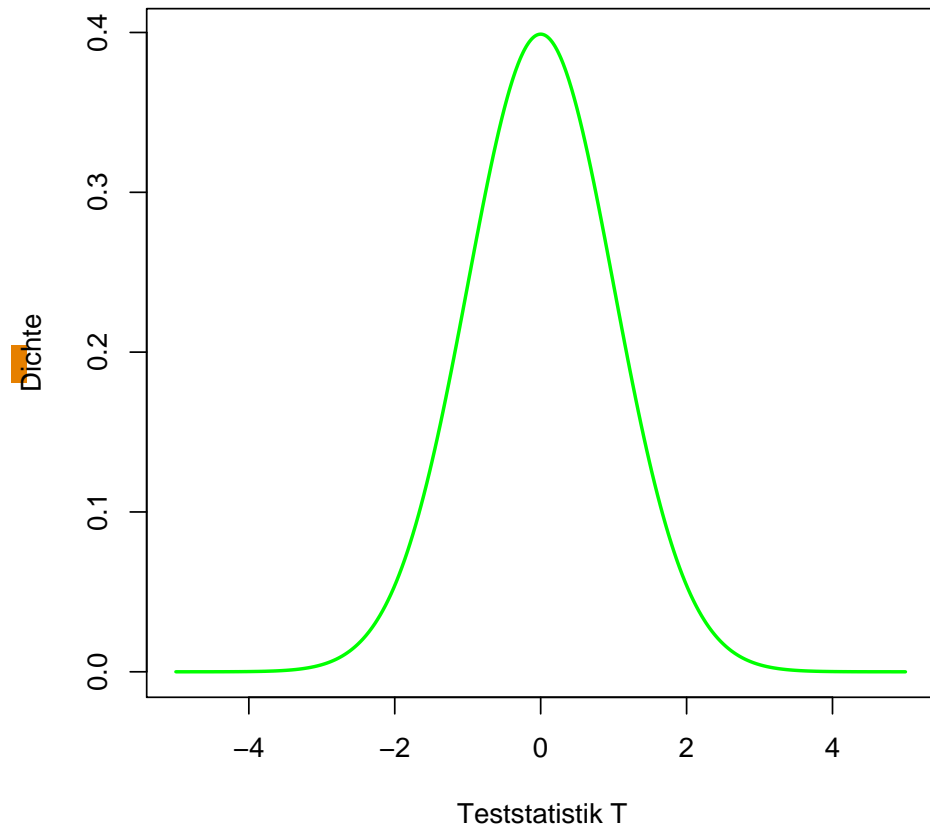


T ist eine **Teststatistik**, deren Verteilung wir kennen.



Was wir hier betrachten ist das **Nullmodell**:

Wir nehmen an, dass in Wahrheit *keine* Mittelwertunterschiede vorliegen. Dann ist die Teststatistik standardnormalverteilt.

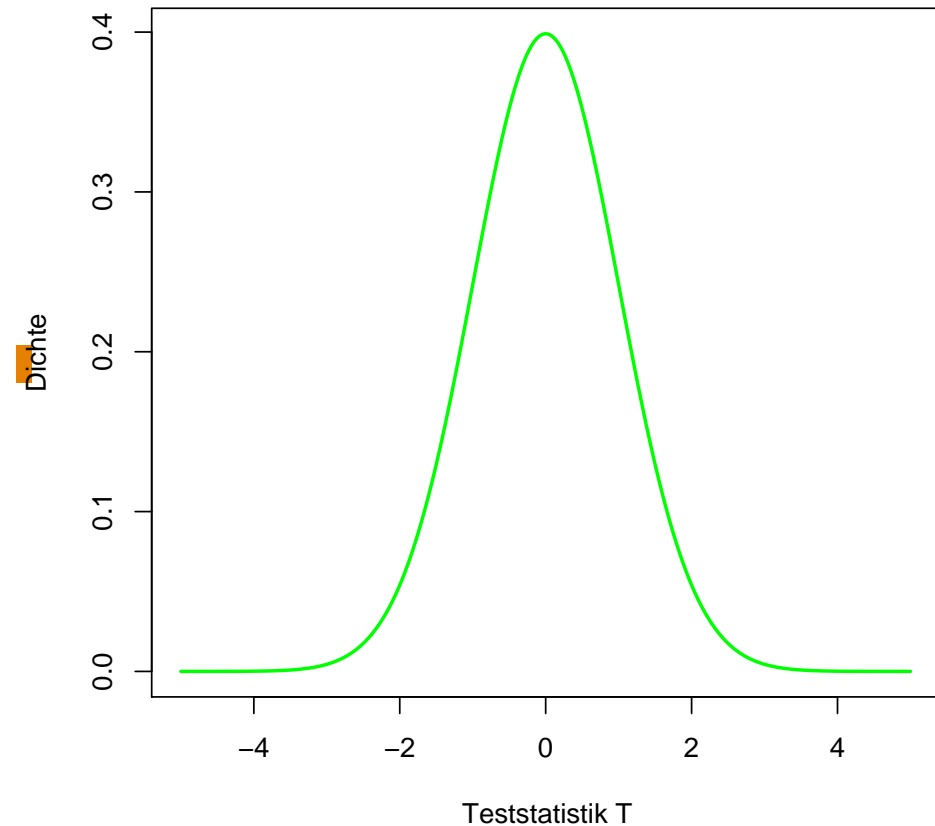


In Wahrheit liegen keine Mittelwertunterschiede vor, daher streut T um Null.

■ Sei t eine Beobachtung von T . Ein **positives** t spricht dafür, dass der Mittelwert in A größer ist als in B, ein **negatives** t für den umgekehrten Fall.

■ T ist eine Zufallsvariable. Ein hoher Wert von T kann also zufällig vorkommen. Die Frage ist nur, **wie wahrscheinlich ist es, dass er zufällig ist?**

■



Nehmen wir an, unsere Daten führen zu einem Wert von $t = 2$. Was ist die Wahrscheinlichkeit, dass T einen Wert größer als 2 annimmt?

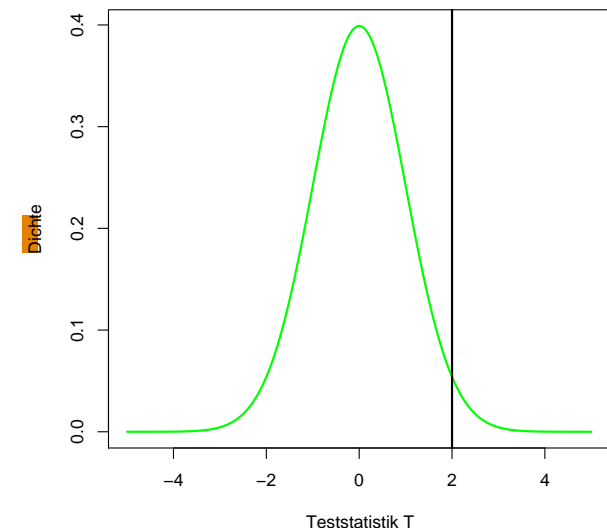


Das errechnet sich aus der **Verteilungsfunktion $\Phi(\cdot)$ der Standardnormalverteilung** und die ist in jedem Allround-Statistikbuch vertafelt bzw in R implementiert:

$$\blacksquare P(T > 2) = 1 - P(T \leq 2) = 1 - \Phi(2) = 1 - \text{pnorm}(2) = 1 - 0.977 = 0.023$$

Es ist also eher unwahrscheinlich, **zufällig** einen Wert größer als 2 zu beobachten.

Damit scheint der beobachtete Unterschied signifikant zu sein.



Damit haben wir alle Zutaten, um einen ordentlichen Test zu beschreiben: Eine Teststatistik, eine bekannte Verteilung und die daraus resultierende Wahrscheinlichkeit. Jetzt wagen wir uns formal an einen **Test** heran:



Wir wollen entscheiden, ob sich die Klassenmittel signifikant unterscheiden und nicht zufällig. ■

Das führt zur **Nullhypothese** $H_0 : \mu_A - \mu_B = 0$

und zur **Alternative** $H_1 : \mu_A - \mu_B \neq 0$

■ Die Nullhypothese spiegelt die Situation wider, dass es aufgrund der Patientenklassen **keinen** Unterschied in der Genexpression gibt. Wir unterstellen aber, dass die Unterschiede auf den Klassen beruhen (die Alternative).



Die Teststatistik lautet dann:

$$|T| = \frac{|\bar{X}_A - \bar{X}_B|}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

wobei n_A und n_B die Anzahlen der Beobachtungen pro Klasse sind.

Den Betrag bauen wir ein, weil es im Moment egal ist, welche Klasse den höheren Mittelwert hat.



$|T|$ ist **nicht** standardnormalverteilt, weil nur noch positive Werte angenommen werden können. Wir rechnen trotzdem mit $N(0, 1)$ weiter und wenden gleich einen Trick an.



Aufgrund der Daten berechnen wir $|T| = t$. Dann ist:

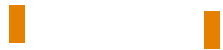
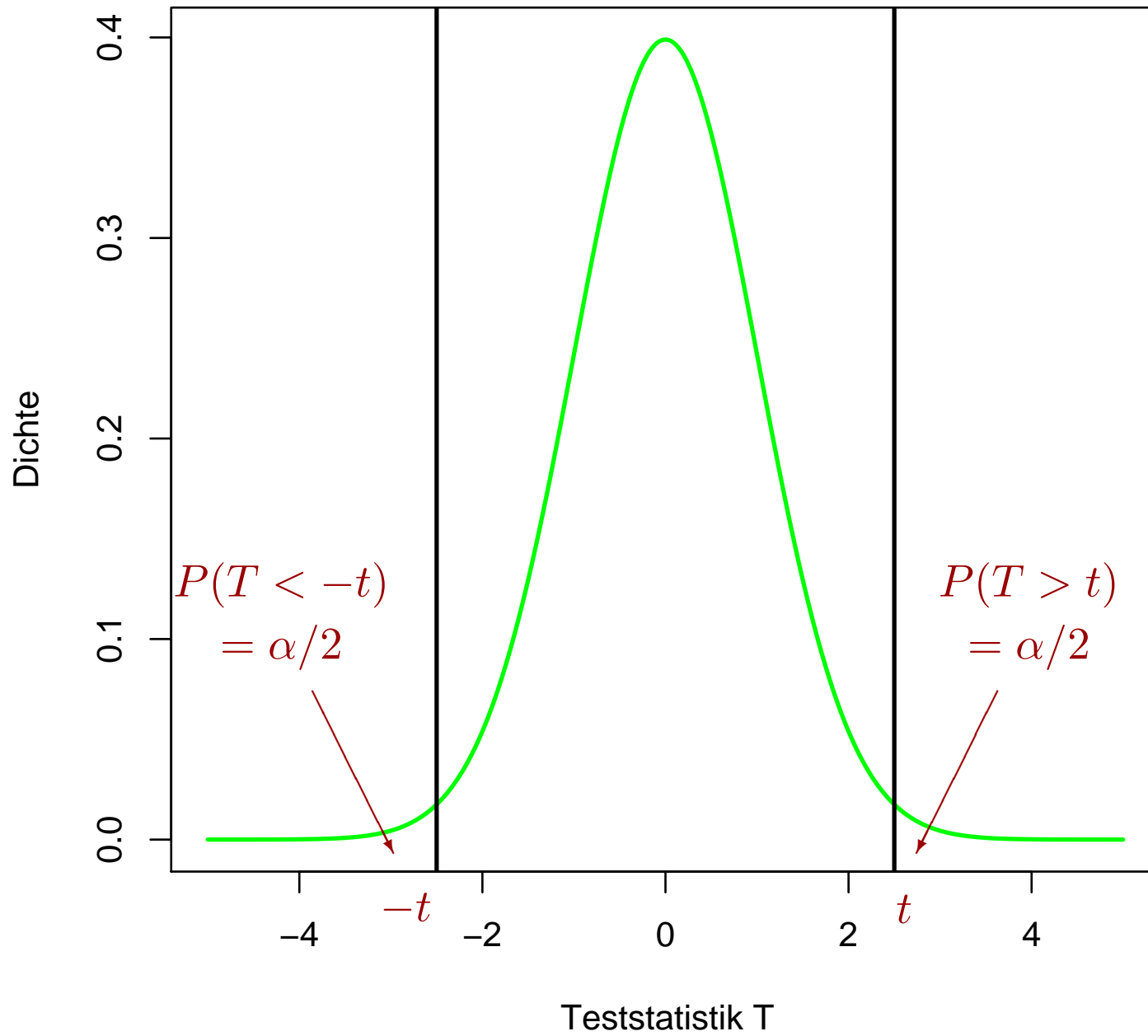
$$\begin{aligned} p &= P(|T| > t) = P(T > t) + P(T < -t) \\ &= 1 - P(T \leq t) + P(T < -t) \\ &= 1 - \Phi(t) + \Phi(-t) \\ &= 2 [1 - \Phi(t)] \end{aligned}$$

p heißt **p-value** oder p-Wert von t . Der p-value gibt die Wahrscheinlichkeit an, mit der zufällig ein Wert größer als t oder kleiner als $-t$ auftritt. Mit diesem zufälligen Fehler müssen wir leben.



Je kleiner der p-value ist, desto unwahrscheinlicher ist es, dass das beobachtete t zufällig ist. Anders: Desto **signifikanter** ist das Ergebnis.





Man kann sich ein **Signifikanzniveau** α vorgeben. Setzen wir $\alpha = 0.05$, so akzeptieren wir einen Fehler von 5%. Dann vergleichen wir den p-value mit α :

$p < \alpha$: Signifikanz \Rightarrow **Lehne die Nullhypothese ab**

$p \geq \alpha$: Keine Signifikanz \Rightarrow **Lehne die Nullhypothese **nicht** ab**

■
 α nennt man auch die Wahrscheinlichkeit für den **Fehler 1. Art (Type I error)**. Dies ist die Wahrscheinlichkeit mit der wir die Nullhypothese ablehnen, obwohl sie *in Wahrheit* zutrifft.

Bis jetzt haben wir angenommen, die Varianzen σ_A^2 und σ_B^2 seien **bekannt**.

Das war zu einfach! Was machen wir also, wenn uns *echte Genexpressionsdaten* begegnen, bei denen wir die Varianzen nicht kennen?



Bei **unbekannten** Varianzen müssen wir diese schätzen:

$$S_A^2 = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} (X_i - \bar{X}_A)^2$$

und S_B^2 analog.



Jetzt lässt sich die Teststatistik aber nicht mehr auf eine Normalverteilung zurückführen. Da muss was Neues her . . .



Die t-Verteilung

Es gilt:

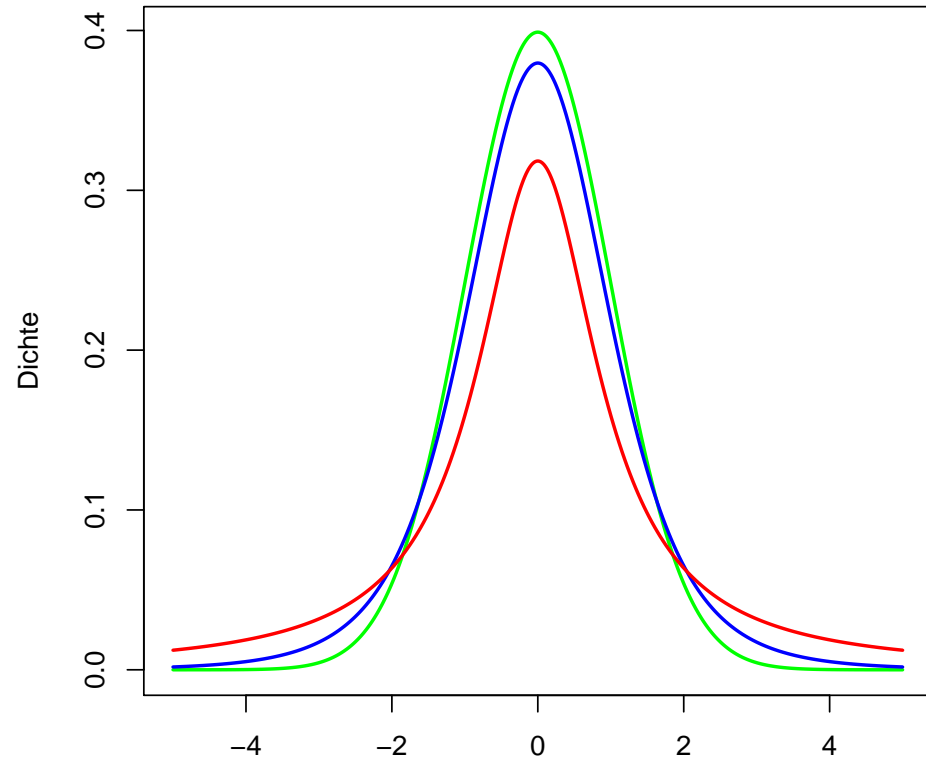
$$Y_0, Y_1, \dots, Y_n \sim N(0, 1) \quad \Rightarrow \quad T = \frac{Y_0}{\sqrt{\frac{1}{n} \sum_{i=1}^n Y_i^2}} \sim t_n$$

t_n bezeichnet die **t-Verteilung** mit n **Freiheitsgraden (degrees of freedom)**, $n \in \mathbb{N}$.
Bei der t-Verteilung beeinflusst n die **Form** der Dichte.

Die t-Verteilung ist symmetrisch um Null und hat schwerere Ränder als die Standardnormalverteilung.

Die Verteilungsfunktion ist vertafelt bzw in R implementiert.

Für $n > 30$ ist sie annähernd gleich der Standardnormalverteilung.



t_1 t_5 $N(0, 1)$

1. Fall: Varianzen unbekannt, aber gleich

Nehmen wir an, die **wahren** Varianzen seien gleich: $\sigma_A^2 = \sigma_B^2$. Dann lässt sich eine t-verteilte Teststatistik bilden und wir sind beim **t-Test** angelangt, *dem* Standardtest der Statistik.



t-Test mit gleichen Varianzen:

1. Hypothesen:

$$H_0 : \mu_A - \mu_B = 0 \quad \text{vs.} \quad H_1 : \mu_A - \mu_B \neq 0$$



2. Teststatistik:

$$|T| = \frac{|\bar{X}_A - \bar{X}_B|}{S}, \quad T \sim t_{n_A+n_B-2}$$

$$\text{mit } S = \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}}$$

S heißt auch gepoolte Varianz.



3. p-value:

$$p = P(|T| > t) = 2 [1 - F(t)]$$

wobei $F(\cdot)$ die Verteilungsfunktion von $t_{n_A+n_B-2}$ ist.



4. Mit vorgegebenem Signifikanzniveau α :

$p < \alpha$: Signifikanz \Rightarrow Lehne die Nullhypothese ab

$p \geq \alpha$: Keine Signifikanz \Rightarrow Lehne die Nullhypothese **nicht** ab



t-Verteilung und t-Test wurden von William S. Gosset (1876-1937) unter dem Pseudonym "Student" entwickelt, daher auch "Student's t-Test".



2. Fall: Varianzen unbekannt und nicht gleich

Nehmen wir nun an, die wahren Varianzen unterscheiden sich: $\sigma_A^2 \neq \sigma_B^2$. Auch dann lässt sich eine t-verteilte Teststatistik bilden. Jetzt sieht die Teststatistik etwas einfacher aus, aber die Berechnung der Freiheitsgrade ist schwieriger. ■

t-Test mit verschiedenen Varianzen:

1. Hypothesen:

$$H_0 : \mu_A - \mu_B = 0 \quad \text{vs.} \quad H_1 : \mu_A - \mu_B \neq 0$$

2. Teststatistik:

$$|T| = \frac{|\bar{X}_A - \bar{X}_B|}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}, \quad T \sim t_f$$

■ Berechnung der Freiheitsgrade: $k = \frac{S_A^2/n_A}{(S_A^2/n_A) + (S_B^2/n_B)}$

$$\Rightarrow f = \left\lceil \left(\frac{k^2}{n_A - 1} + \frac{(1 - k)^2}{n_B - 1} \right)^{-1} \right\rceil \in \mathbb{N}$$

■ Obige Formel heißt **Welch-Approximation** der Freiheitsgrade, daher heißt dieser t-Test auch **“Welch’s t-Test”**.

3. p-value:

$$p = P(|T| > t) = 2 [1 - F(t)]$$

wobei $F(\cdot)$ die Verteilungsfunktion von t_f ist.



4. Mit vorgegebenem Signifikanzniveau α :

$p < \alpha$: Signifikanz \Rightarrow Lehne die Nullhypothese ab

$p \geq \alpha$: Keine Signifikanz \Rightarrow Lehne die Nullhypothese **nicht** ab



Das waren viele Formeln. Zum Schluss wenden wir den t-Test mal an . . .



t-Test in R

Der Befehl `t.test(x,y,var.equal=TRUE)` führt einen Two-sample t-Test mit gleicher Varianz durch.

Die zwei Samples `x` und `y` sind dabei unsere Beobachtungen aus Klasse A und Klasse B.

Two Sample t-test

data: A and B

$t = 3.0144$, $df = 23$, $p\text{-value} = 0.006178$

alternative hypothesis: true difference in means is not equal to 0

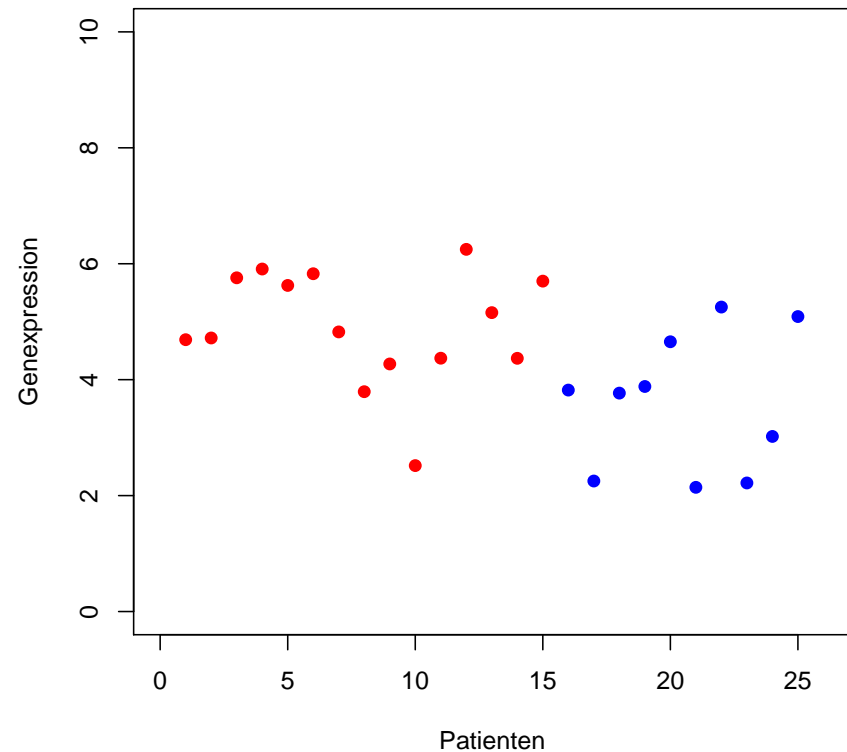
95 percent confidence interval:

0.4105315 2.2064912

sample estimates:

mean of x mean of y

4.918845 3.610334



Two Sample t-test

data: A and B

$t = 2.1355$, $df = 23$, $p\text{-value} = 0.04359$

alternative hypothesis: true difference in means is not equal to 0

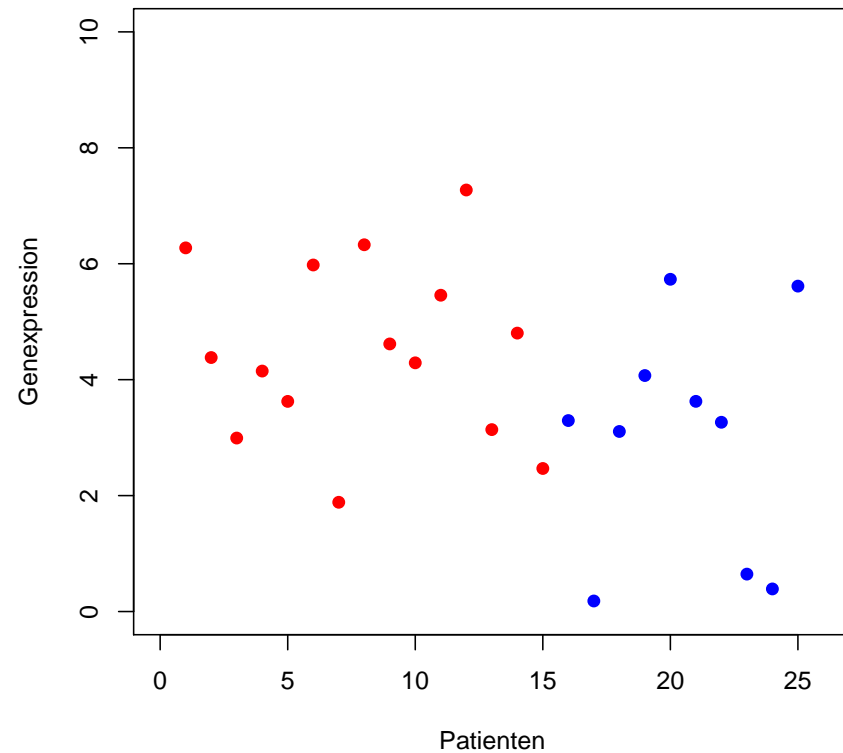
95 percent confidence interval:

0.04754084 2.98823222

sample estimates:

mean of x mean of y

4.510930 2.993043



Two Sample t-test

data: A and B

$t = 0.25$, $df = 23$, $p\text{-value} = 0.8048$

alternative hypothesis: true difference in means is not equal to 0

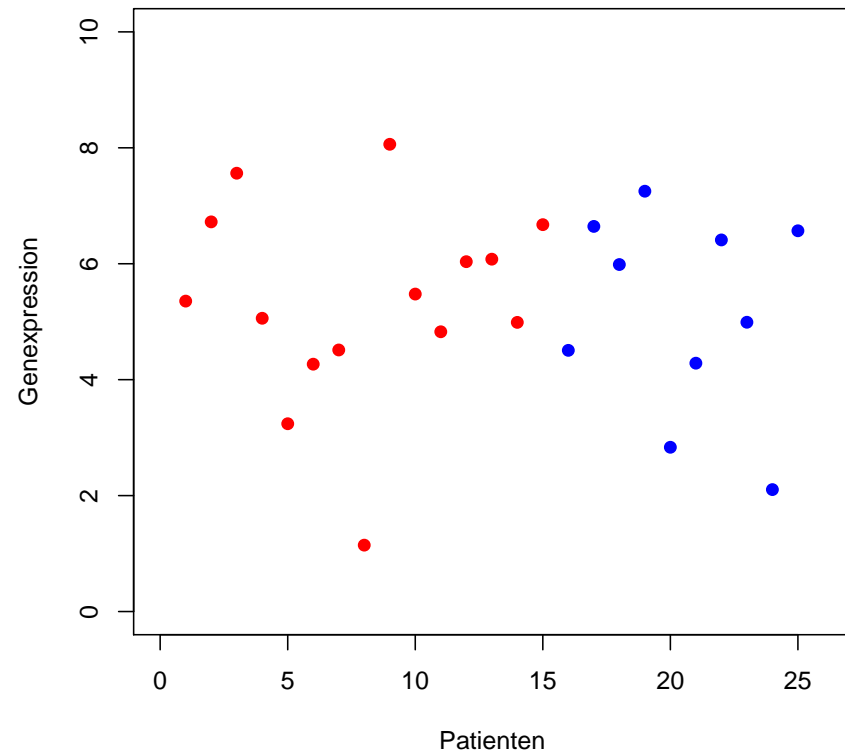
95 percent confidence interval:

-1.278209 1.629579

sample estimates:

mean of x mean of y

5.333996 5.158311



Zusammenfassung

t-Test

t-Verteilung

Nullhypothese

Alternative

Teststatistik

p-value

Signifikanzniveau

Fehler 1. Art

