Moc testów statystycznych

Znana wariancja

Moc testu to prawdopodobieństwo odrzucenia hipotezy zerowej, jeśli jest ona nieprawdziwa. Jest to więc pewna miara czułości naszego testu. Mówi nam, jak łatwo będzie nam uzyskać wynik istotny statystycznie w sytuacji, w której nam na takim wyniku zależy. Moc to $1 - \beta$, gdzie β to prawdopodobieństwo popełnienia błędu drugiego rodzaju (czyli nieodrzucenia hipotezy zerowej, kiedy jest ona nieprawdziwa).

Tak, jak do kontrolowania α (czyli prawdopodobieństwa odrzucenia hipotezy zerowej, kiedy jest ona prawdziwa) musimy znać rozkład statystyki z próby przy założeniu, że hipoteza zerowa jest prawdziwa, podobnie żeby oszacować moc i β, musimy skonstruować rozkład statystyki z próby przy założeniu, że hipoteza zerowa jest fałszywa. W praktyce musimy przyjąć rozkład z próby dla jakiejś **punktowej hipotezy alternatywnej**. Skoro zazwyczaj hipoteza zerowa mówi o braku różnic (przy dwóch grupach będzie to $\mu_1 - \mu_2 = 0$), ta hipoteza alternatywna będzie miała postać $\mu_1 - \mu_2 = d$, gdzie $d \neq 0$.

UWAGA: różnicę między średnimi d, zarówno tę oczekiwaną, w kontekście obliczeń związanych z mocą, jak i tę otrzymaną, w kontekście wielkości uzyskanego efektu, często wyraża się również w formie standaryzowanej, jako stosunek do odchylenia standardowego: $d = (\mu_1 - \mu_2)/\sigma$.

Przyjmijmy, że interesuje nas właśnie porównanie średnich dwóch prób niezależnych, i załóżmy takie dane:

```
# liczebność jednej próby (przyjmijmy, że mamy dwie próby równoliczne)
n <- 20
# dla uproszczenia skupmy się na teście jednostronnym
alpha <- 0.025
# odchylenie w obu populacjach (zakładamy homogeniczność wariancji)
sigma <- 5
# efekt niewystandaryzowany, czyli różnica między średnimi w populacjach
 ↪  przewidziana przez hipotezę alternatywną
diff <- 6
```

Na początek obliczmy błąd standardowy różnicy dwóch średnich. Proszę wpisać wyrażenie, które powinno zastąpić znak zapytania poniżej.

```
se <- ? * sqrt(2/n)
```

```
se <- ? * sqrt(2/n)
```

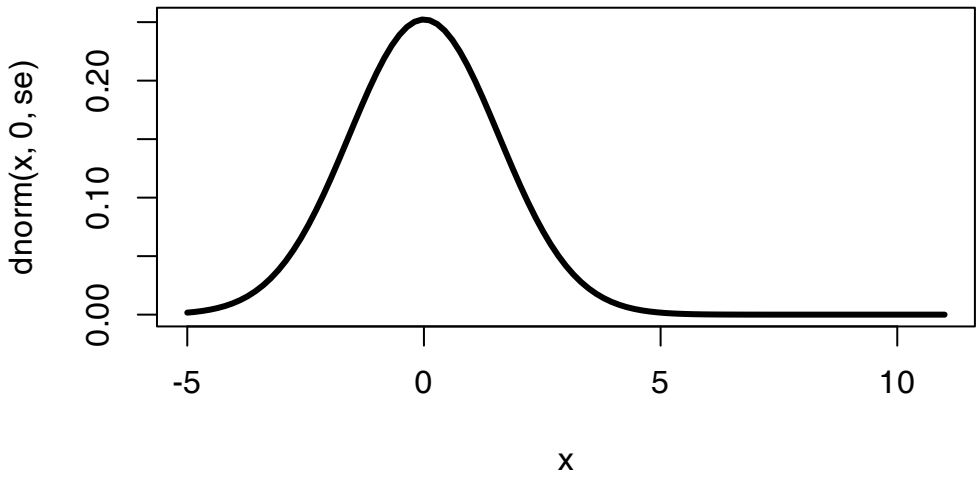
Na wariancję różnicy dwóch średnich składają się wariancje generowane przez obie średnie:

$$\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}$$

Stąd błąd standardowy (czyli pierwiastek z wariancji) w naszym przypadku to `sigma * sqrt(2/n)`.

Znając błąd standardowy, możemy nakreślić rozkład różnicy średnich z próby przy założeniu hipotezy zerowej:

```
se <- sigma * sqrt(2/n)
curve(dnorm(x, 0, se), -5, 11, lwd=3)
```



Żeby dodać do powyższego wykresu rozkład różnicy średnich przy założeniu naszej hipotezy alternatywnej, użyjemy:

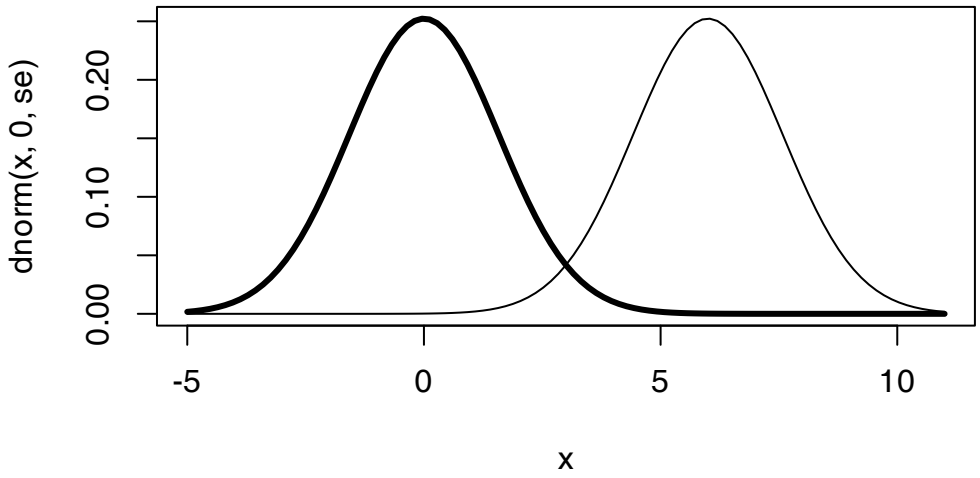
```
curve(dnorm(x, ?, se), add=TRUE)
```

Czym zastąpić znak zapytania w wyrażeniu powyżej?

```
curve(dnorm(x, 0, se), -5, 11, lwd=3)
curve(dnorm(x, ?, se), add=TRUE)
```

Róznica między hipotezą zerową a hipotezą alternatywną sprowadza się do wartości oczekiwanej różnicy między średnimi: w pierwszym przypadku wynosi ona 0, a w drugim — `diff` czyli 6.

```
curve(dnorm(x, 0, se), -5, 11, lwd=3)
curve(dnorm(x, 6, se), add=TRUE)
```



Jeśli statystyka policzona z naszych danych (różnica między średnimi) przekroczy wartość krytyczną, odrzucimy hipotezę zerową. Obliczmy tę wartość krytyczną i dorysujmy ją do wykresu:

Pytanie

Jakim wyrażeniem należy zastąpić znak zapytania?

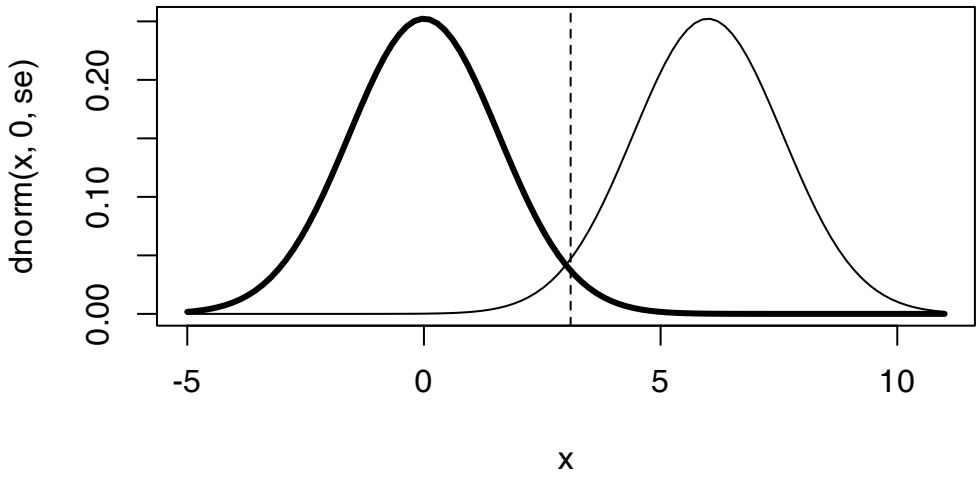
```
curve(dnorm(x, 0, se), -5, 11, lwd=3)
curve(dnorm(x, 6, se), add=TRUE)

cr <- qnorm(?, 0, se, lower.tail=FALSE)
abline(v=cr, lty=2)
```

Wartość krytyczna odcina górne 2.5% (czyli alpha) powierzchni pod tym rozkładem, który zakłada prawdziwość hipotezy zerowej (czyli tym nakreślonym grubszą kreską).

```r
curve(dnorm(x, 0, se), -5, 11, lwd=3)
curve(dnorm(x, 6, se), add=TRUE)

cr <- qnorm(0.025, 0, se, lower.tail=FALSE)
abline(v=cr, lty=2)
```



Pytanie

Pamiętając, że:

- α to prawdopodobieństwo odrzucenia hipotezy zerowej, kiedy jest ona prawdziwa,
- β to prawdopodobieństwo nieodrzucenia hipotezy zerowej, kiedy jest ona fałszywa, a
- moc to $1-\beta$ (prawdopodobieństwo odrzucenia hipotezy zerowej, kiedy jest ona fałszywa),

proszę dopasować odpowiedzi:

- Pole pod wykresem nakreślonym grubszą kreską na prawo od przerywanej linii.
- Pole pod wykresem nakreślonym cienką kreską na lewo od przerywanej linii.
- Pole pod wykresem nakreślonym cienką kreską na prawo od przerywanej linii.

Moc testu, czyli prawdopodobieństwo odrzucenia hipotezy zerowej, kiedy prawdziwa jest hipoteza alternatywna, to powierzchnia pod wykresem nakreślonym cienką kreską na prawo od linii przerywanej. Możemy ją zatem obliczyć tak:

Pytanie Proszę zastąpić znak zapytania właściwym wyrażeniem.

```r
power <- pnorm(?, diff, se, lower.tail=FALSE)
print(power)
```

```r
power <- pnorm(cr, diff, se, lower.tail=FALSE)
print(power)
```

```
[1] 0.9667301
```

Moc testu liczymy więc (1) wyznaczając wartość krytyczną na podstawie rozkładu dla hipotezy zerowej (by odciąć górne 2.5% tego rozkładu), a następnie (2) odnosząc tę wartość do rozkładu dla hipotezy alternatywnej (sprawdzając, jaka część tego rozkładu znajduje się powyżej wartości krytycznej). W naszym przykładzie moc testu wynosi ponad 96%; jest zatem bardzo wysoka.

Pytanie

Proszę zastanowić i zaznaczyć wszystkie czynniki, które mają wpływ na moc testu.

- Liczebność (n).
- Poziom istotności statystycznej (α).
- Wariancja w populacji, σ^2 (albo odchylenie standardowe, σ).
- Spodziewany efekt (czyli różnica między średnimi przy założeniu hipotezy alternatywnej, d czyli `diff` w kodzie).

Czynniki wpływające na moc testu

Im większa różnica, na wykryciu której nam zależy, tym łatwiej ją wykryć. Zatem, im większa różnica między średnimi przewidywana przez hipotezę alternatywną (im większy spodziewany efekt), tym większa moc testu.

Im mniejszy "szum" w naszych danych, tym łatwiej wykryć interesującą nas różnicę. Zatem, im mniejsza wariancja, tym większa moc testu (bo od wariancji zależy błąd standardowy, a im mniejszy błąd standardowy, tym większa moc). Na wariancję wyników mamy niewielki wpływ (bo zależy ona od specyfiki naszej zmiennej i populacji). Warto jednak pamiętać, że precyzja pomiaru zmniejsza wariancję (nieprecyzyjność pomiarów wprowadza dodatkową losowość do ich wyników)!

Im bardziej jesteśmy skłonni akceptować "fałszywe alarmy", tym łatwiej będzie nam wykryć prawdziwe różnice. Zatem, zwiększając poziom α, zwiększamy równocześnie moc testu: większa α oznacza niższą wartość krytyczną, po przekroczeniu której odrzucimy hipotezę zerową. Coś za coś: łatwiej będzie nam odrzucić hipotezę zerową, niezależnie od tego, czy jest ona fałszywa czy nie (czyli zwiększamy równocześnie prawdopodobieństwo popełnienia błędu I rodzaju).

Wreszcie, im więcej pomiarów, tym łatwiej wykryć interesującą nas różnicę. Im większa liczebność, tym większa moc testu (bo od liczebności zależy błąd standardowy, a im mniejszy błąd standardowy, tym większa moc). W praktyce to właśnie liczebnością się manipuluje, kiedy chce się uzyskać zadowalającą moc testu: na spodziewaną różnicę nie mamy wpływu, na wariancję pomiarów — niewielki, a zwiększeniem prawdopodobieństwa błędu I rodzaju nie jesteśmy zainteresowani.

Dla przypomnienia: wariancja w populacji i liczebność próby wpływają na moc, bo od nich zależy błąd standardowy, a błąd standardowy:

- to odchylenie standardowe statystyki z próby.
- jest równy pierwiastkowi wariancji statystyki z próby.
- jest równy wariancji podzielonej przez pierwiastek liczebności.
- jest równy odchyleniu standardowemu podzielonemu przez liczebność.

Nieznana wariancja

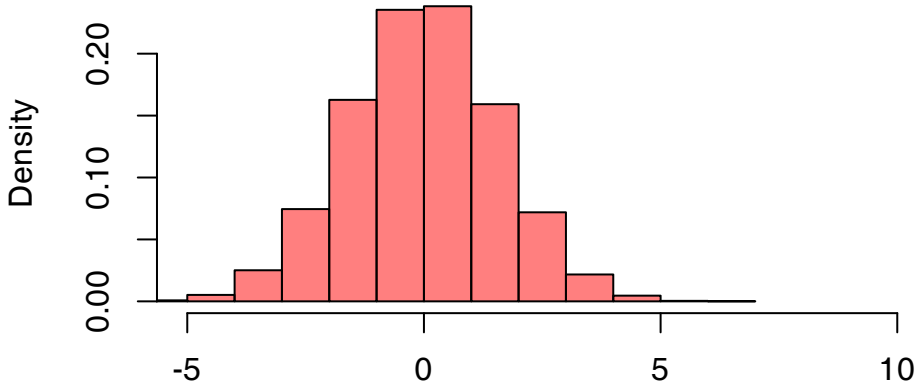
Obliczenie mocy w przykładzie z poprzednich stron było prostsze niż zazwyczaj. Założyliśmy bowiem, że znamy wariancję w populacji. Wtedy korzystamy z faktu, że rozkład różnicy średnich to rozkład normalny. Zmiana hipotezy zerowej na alternatywną powoduje, że jeden parametr tego rozkładu — jego średnia — zmienia się z 0 na d, ale nie wpływa to na drugi parametr — błąd standardowy.

Nie znając wariancji w populacji, a zazwyczaj jej nie znamy, musimy skorzystać z rozkładu t: jeśli błąd standardowy liczymy, biorąc wariancję policzoną z danych, różnica między statystyką a parametrem podzielona przez ten błąd ma taki właśnie rozkład. Ale do tej pory uczyliśmy się, że rozkład t ma zawsze średnią 0 i tylko jeden parametr: stopnie swobody pozostałe po oszacowaniu wariancji.

Jak w takim wypadku wygląda rozkład różnicy średnich przy założeniu hipotezy alternatywnej? Na szczęście znamy już R wystarczająco dobrze, żeby z pomocą prostej symulacji to sprawdzić.

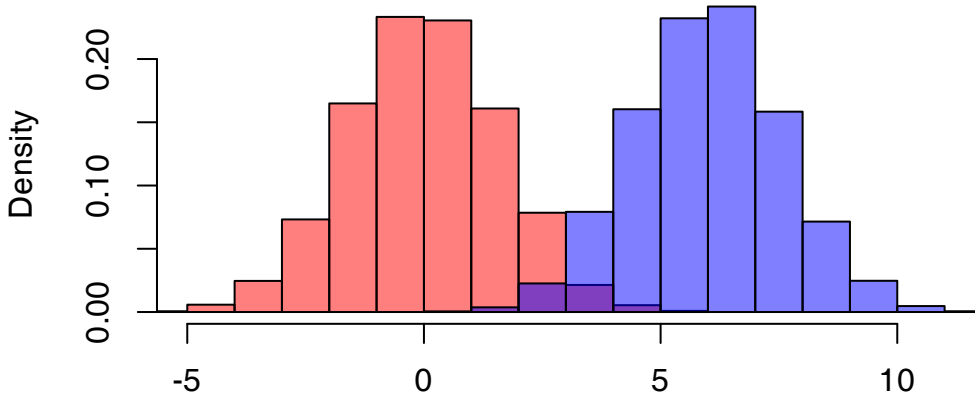
Zacznijmy od dotychczasowej sytuacji. Ten histogram przedstawia przybliżenie rozkładu różnicy średnich przy założeniu hipotezy zerowej (10000 razy liczymy różnicę między średnimi prób wylosowanych z tej samej populacji):

```
hist(
    replicate(
        10000,
        mean(rnorm(n, 0, sigma)) - mean(rnorm(n, 0, sigma))), # różnica
         ↪  między średnimi przy h_0
    main="",
    xlab="",
    xlim=c(-5, 11),
    freq=FALSE,
    col = adjustcolor('red', alpha.f = 0.5))
```



Teraz dorysujemy histogram, który przedstawia przybliżenie rozkładu różnicy średnich przy założeniu hipotezy alternatywnej (10000 razy liczymy różnicę między średnimi prób wylosowanych z populacji, których średnie różnią się o diff):

```
hist(
    replicate(
        10000,
        mean(rnorm(n, 0, sigma)) - mean(rnorm(n, 0, sigma))), # różnica
         ↪ między średnimi przy h_0
    main="",
    xlab="",
    xlim=c(-5, 11),
    freq=FALSE,
    col = adjustcolor('red', alpha.f = 0.5))
hist(
    replicate(
        10000,
        mean(rnorm(n, diff, sigma)) - mean(rnorm(n, 0, sigma))), # różnica
         ↪ między średnimi przy h_a
    add=TRUE,
    freq=FALSE,
    col = adjustcolor('blue', alpha.f = 0.5))
```



Łatwo możemy sprawdzić, że te histogramy odpowiadają krzywym normalnym o odpowiednich parametrach (średnia równa `0` albo `diff`, a odchylenie standardowe równe `se`). Możemy też dorysować naszą wartość krytyczną decydującą o odrzuceniu, bądź nie, hipotezy zerowej:

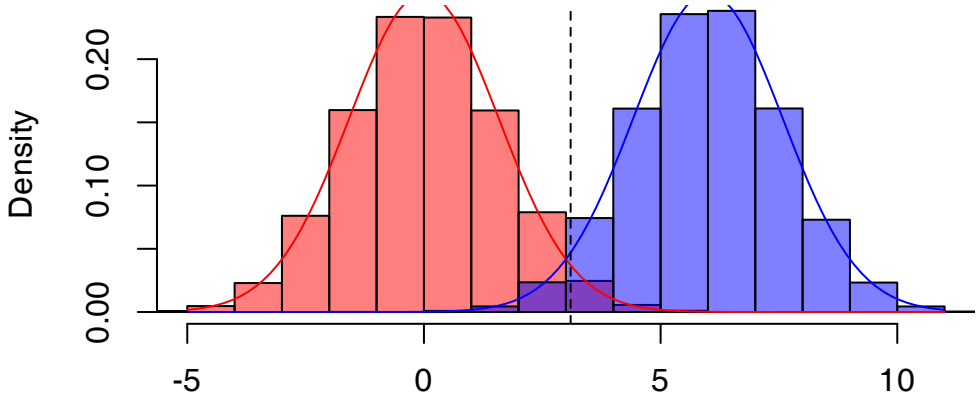
```
hist(
    replicate(
        10000,
        mean(rnorm(n, 0, sigma)) - mean(rnorm(n, 0, sigma))), # różnica
         ↪ między średnimi przy h_0
    main="",
    xlab="",
    xlim=c(-5, 11),
    freq=FALSE,
    col = adjustcolor('red', alpha.f = 0.5))

hist(
    replicate(
        10000,
        mean(rnorm(n, diff, sigma)) - mean(rnorm(n, 0, sigma))), # różnica
         ↪ między średnimi przy h_a
    add=TRUE,
    freq=FALSE,
    col = adjustcolor('blue', alpha.f = 0.5))



curve(dnorm(x, 0, se), add=TRUE, col = 'red')
curve(dnorm(x, diff, se), add=TRUE, col = 'blue')

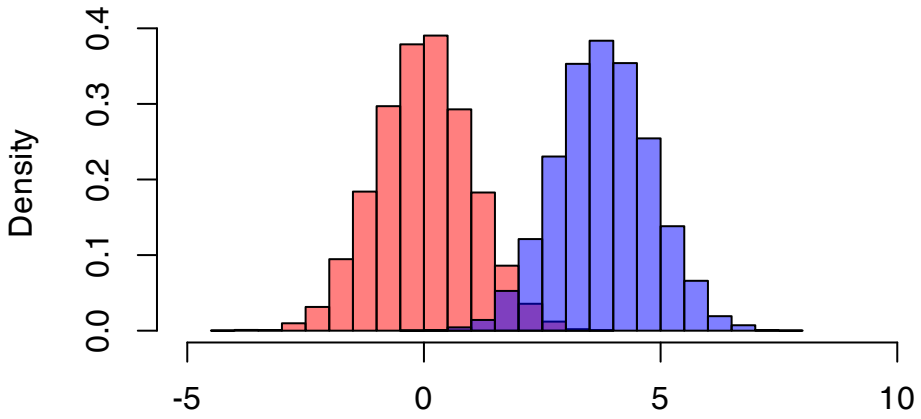
abline(v=cr, lty=2)
```



W kolejnym kroku nakreślmy te same histogramy, ale standaryzując różnice między średnimi, czyli dzieląc je przez błąd standardowy:

```
hist(
    replicate(
        10000,
        (mean(rnorm(n, 0, sigma)) - mean(rnorm(n, 0, sigma))) / se),
    main="",
    xlab="",
    xlim=c(-5, 11),
    freq=FALSE,
    col = adjustcolor('red', alpha.f = 0.5))


hist(
    replicate(
        10000,
        (mean(rnorm(n, diff, sigma)) - mean(rnorm(n, 0, sigma))) / se),
    add=TRUE,
    freq=FALSE,
    col = adjustcolor('blue', alpha.f = 0.5))
```

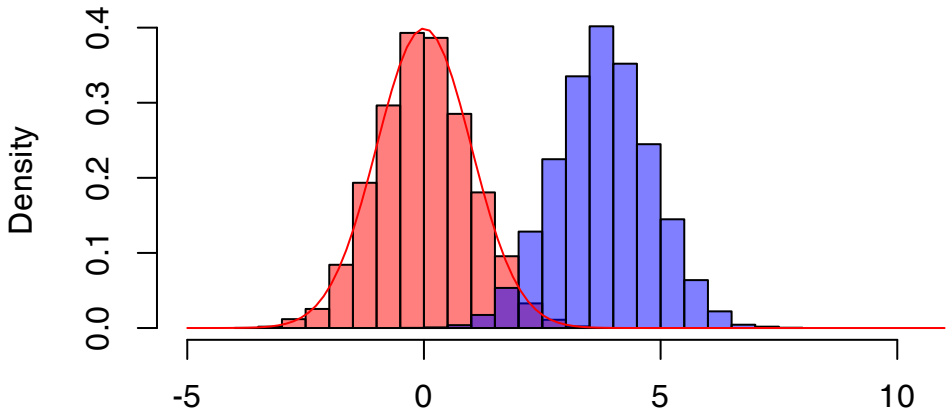


Teraz rozkład normalny odpowiadający hipotezie zerowej dorysujemy tak:

```
hist(
    replicate(
        10000,
        (mean(rnorm(n, 0, sigma)) - mean(rnorm(n, 0, sigma))) / se),
    main="",
    xlab="",
    xlim=c(-5, 11),
    freq=FALSE,
    col = adjustcolor('red', alpha.f = 0.5))


hist(
    replicate(
        10000,
        (mean(rnorm(n, diff, sigma)) - mean(rnorm(n, 0, sigma))) / se),
    add=TRUE,
    freq=FALSE,
    col = adjustcolor('blue', alpha.f = 0.5))


curve(dnorm(x), add=TRUE, col = 'red')
```



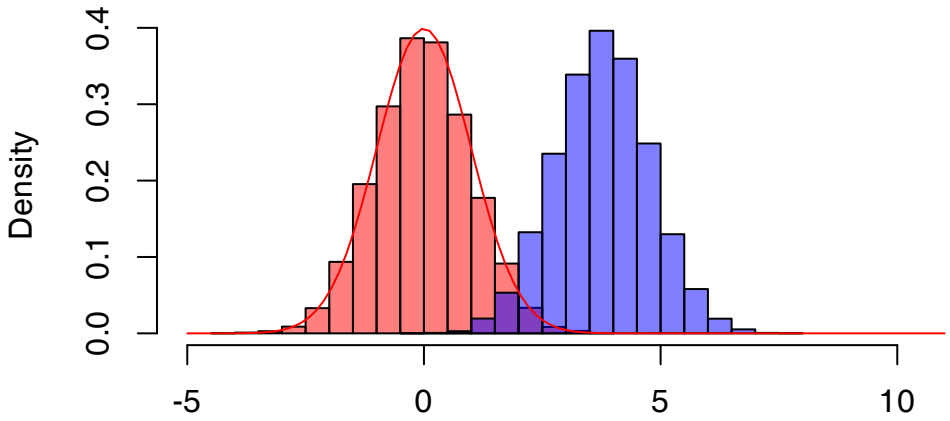
Pytanie

Dlaczego? Kto wie, ten z łatwością zamieni znak zapytania na właściwe wyrażenie w poleceniu rysującym rozkład odpowiadający hipotezie alternatywnej:

```
hist(
    replicate(
        10000,
        (mean(rnorm(n, , sigma)) - mean(rnorm(n, 0, sigma))) / se),
    main="",
    xlab="",
    xlim=c(-5, 11),
    freq=FALSE,
    col = adjustcolor('red', alpha.f = 0.5))


hist(
    replicate(
        10000,
        (mean(rnorm(n, diff, sigma)) - mean(rnorm(n, 0, sigma))) / se),
    add=TRUE,
    freq=FALSE,
    col = adjustcolor('blue', alpha.f = 0.5))


curve(dnorm(x), add=TRUE, col = 'red')
```




```
curve(dnorm(x, ?), add=TRUE, col = 'blue')
```

Ponieważ dzieliliśmy otrzymaną różnicę między średnimi przez błąd standardowy, zmianie ulegają oba rozkłady: zarówno ich średnie, jak i odchylenia standardowe należy podzielić przez tę samą wartość. W przypadku hipotezy zerowej mamy teraz do czynienia ze standardowym $N(0, 1)$, a w przypadku hipotezy alternatywnej — z N(`diff/se`, 1).

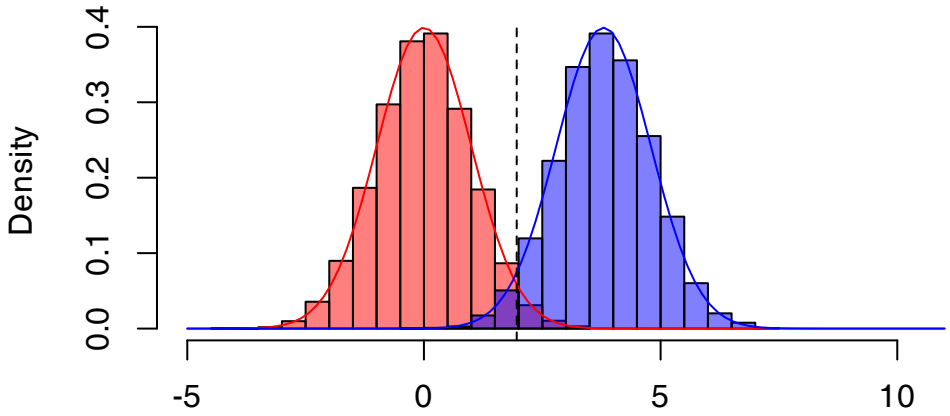
Zmianie ulega również wartość krytyczna:

```
hist(
    replicate(
        10000,
        (mean(rnorm(n, 0, sigma)) - mean(rnorm(n, 0, sigma))) / se),
    main="",
    xlab="",
    xlim=c(-5, 11),
    freq=FALSE,
    col = adjustcolor('red', alpha.f = 0.5))


hist(
    replicate(
        10000,
        (mean(rnorm(n, diff, sigma)) - mean(rnorm(n, 0, sigma))) / se),
    add=TRUE,
    freq=FALSE,
    col = adjustcolor('blue', alpha.f = 0.5)
    )
```

```r
curve(dnorm(x), add=TRUE, col = 'red')
curve(dnorm(x, diff/se), add=TRUE, col = 'blue')

abline(v=cr/se, lty=2)
```




```r
print(cr/se)
```

```
[1] 1.959964
```

Można też sprawdzić, że tę samą wartość zwróci qnorm(.975). Ta wartość to w przybliżeniu z = 1.96.

Które wyrażenia pozwalają zatem obliczyć moc testu przy znanej wariancji w populacji?

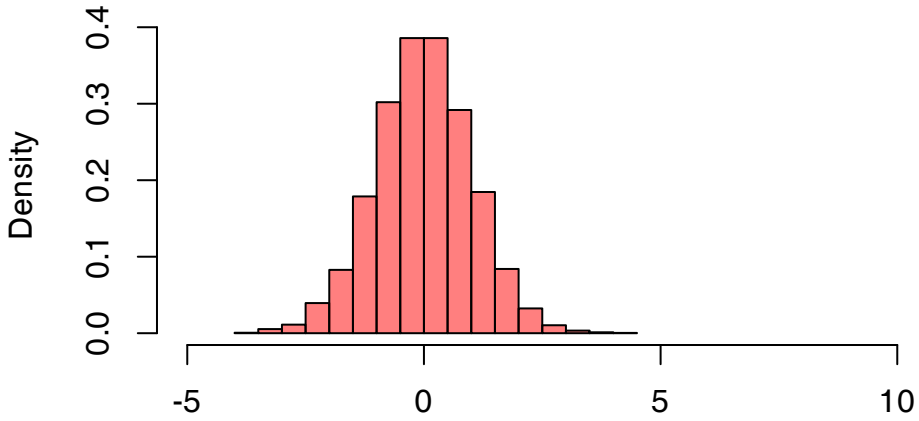
- `pnorm(cr, diff, se, lower.tail=FALSE)`
- `pnorm(cr/se, diff/se, lower.tail=FALSE)`
- `pnorm(cr/se, diff/se, se, lower.tail=FALSE)`

Moc testu *t*

Dla przypomnienia: różnica między średnimi podzielona przez błąd standardowy to statystyka z, jeśli znamy wariancję w populacji, albo statystyka t, jeśli szacujemy wariancję na podstawie danych z próby. W tym drugim przypadku, żeby zasymulować rozkład odpowiadający hipotezie zerowej, przy każdym losowaniu musimy na nowo szacować błąd standardowy, korzystając z wariancji policzonych w wylosowanych próbach:

```r
hist(replicate(10000, {
     s1 <- rnorm(n, 0, sigma)
     s2 <- rnorm(n, 0, sigma)
     (mean(s1) - mean(s2)) / sqrt((var(s1) + var(s2))/n)
}),
```

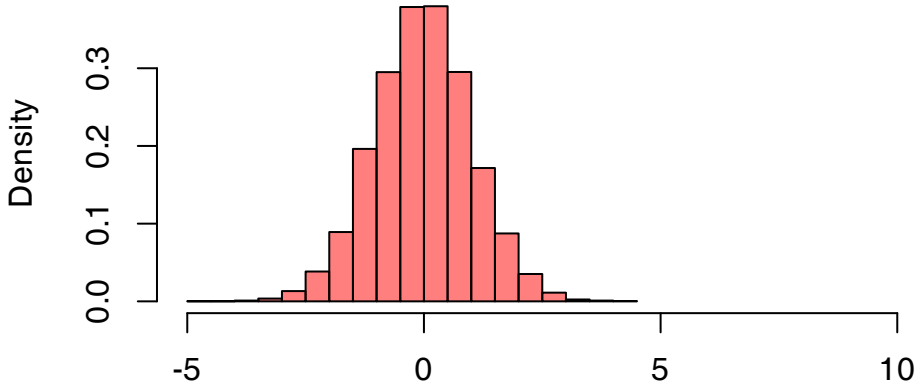
```
     main="",
     xlab="",
     xlim=c(-5, 11),
     freq=FALSE,
     col = adjustcolor('red', alpha.f = .5))
```



Odpowiednią krzywą dorysujemy wtedy tak:

Pytanie Jakiej funkcji musimy użyć?

```
hist(replicate(10000, {
     s1 <- rnorm(n, 0, sigma)
     s2 <- rnorm(n, 0, sigma)
     (mean(s1) - mean(s2)) / sqrt((var(s1) + var(s2))/n)
}),
     main="",
     xlab="",
     xlim=c(-5, 11),
     freq=FALSE,
     col = adjustcolor('red', alpha.f = .5))
```



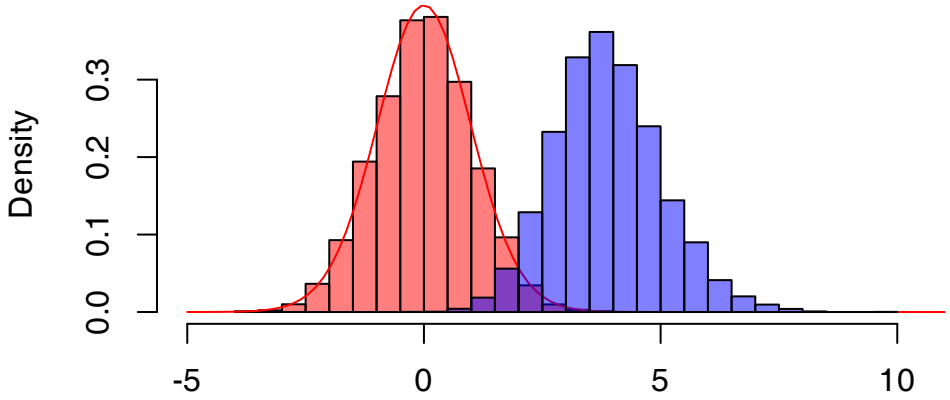

```
#curve(?(x, 2*n-2), add=TRUE, col = 'red')
```

Funkcją potrzebną do nakreślenia odpowiedniej krzywej jest oczywiście `dt()`. Mamy bowiem do czynienia ze zwykłym rozkładem t o 2*n-2 stopniach swobody. Żeby przekonać się teraz, jak wygląda w tym przypadku rozkład dla hipotezy alternatywnej, musimy nieco zmodyfikować naszą symulację:

```
hist(replicate(10000, {
     s1 <- rnorm(n, 0, sigma)
     s2 <- rnorm(n, 0, sigma)
     (mean(s1) - mean(s2)) / sqrt((var(s1) + var(s2))/n)
}),
     main="",
     xlab="",
     xlim=c(-5, 11),
     freq=FALSE,
     col = adjustcolor('red', alpha.f = .5))

curve(dt(x, 2*n-2), add=TRUE, col = 'red')

hist(replicate(10000, {
     s1 <- rnorm(n, diff, sigma)
     s2 <- rnorm(n, 0, sigma)
     (mean(s1) - mean(s2)) / sqrt((var(s1) + var(s2))/n)
}), add=TRUE, freq=FALSE,
    col = adjustcolor('blue', alpha.f = 0.5))
```



Jeśli się bliżej przyjrzeć nowemu rozkładowi, nie jest to po prostu nasz oryginalny rozkład *t*, przesunięty jedynie o wartość `diff` w prawo. Jest on trochę niższy i nieznacznie skośny. Zwiększając wartość `diff`, można lepiej zaobserwować te jego cechy.

Okazuje się, że rozkład t, z jakim mieliśmy do tej pory do czynienia, jest tylko szczególnym przypadkiem ogólniejszego rozkładu. Ten ogólniejszy rozkład t ma oprócz stopni swobody jeszcze jeden parametr — parametr niecentralności. Kiedy ten parametr przyjmuje wartość zerową, mamy do czynienia ze zwykłym, czyli centralnym, rozkładem t. Kiedy parametr niecentralności jest niezerowy, mamy do czynienia z niecentralnym rozkładem t.

Parametr niecentralności dla rozkładu t przy założeniu hipotezy alternatywnej wynosi `diff/se`, czyli dokładnie tyle, o ile przesuwaliśmy rozkład z, kiedy przyjmowaliśmy, że znamy wariancję w populacji. Czyli krzywą naszego rozkładu nakreślimy tak:

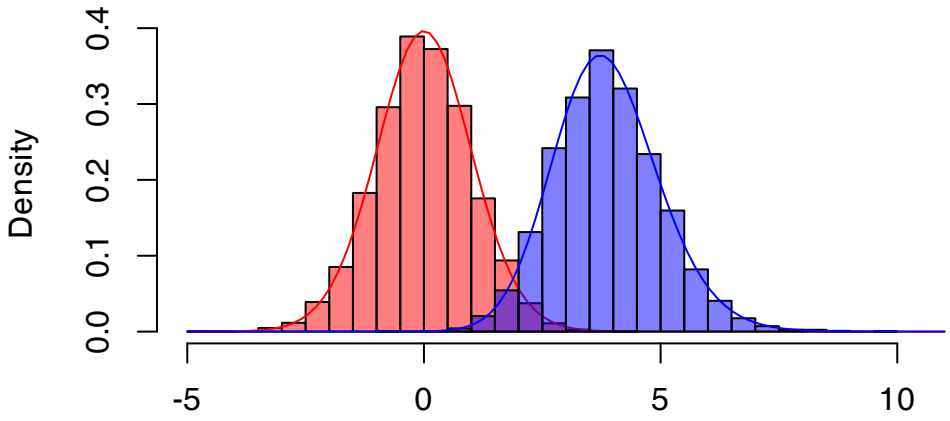
```
hist(replicate(10000, {
     s1 <- rnorm(n, 0, sigma)
     s2 <- rnorm(n, 0, sigma)
     (mean(s1) - mean(s2)) / sqrt((var(s1) + var(s2))/n)
}),
     main="",
     xlab="",
     xlim=c(-5, 11),
     freq=FALSE,
     col = adjustcolor('red', alpha.f = .5))

curve(dt(x, 2*n-2), add=TRUE, col = 'red')

hist(replicate(10000, {
     s1 <- rnorm(n, diff, sigma)
     s2 <- rnorm(n, 0, sigma)
     (mean(s1) - mean(s2)) / sqrt((var(s1) + var(s2))/n)
}), add=TRUE, freq=FALSE,
```

```
    col = adjustcolor('blue', alpha.f = 0.5))

curve(dt(x, 2*n-2, diff/se), add=TRUE, col = 'blue')
```



Żeby obliczyć moc naszego testu t, musimy podobnie jak do tej pory wyznaczyć wartość krytyczną t (korzystając z rozkładu dla hipotezy zerowej), a następnie odnieść tę wartość do rozkładu dla hipotezy alternatywnej (żeby sprawdzić, jaka część tego rozkładu znajduje się powyżej tej wartości).

Pytanie

Które wyrażenie pozwoli poznać nam wartość krytyczną t?

- `cr <- qt(alpha, 2*n-2, lower.tail=FALSE)`
- `cr <- qt(alpha, 2*n-2, lower.tail=TRUE)`
- `cr <- pt(alpha, 2*n-2, lower.tail=FALSE)`
- `cr <- pt(alpha, 2*n-2, lower.tail=TRUE)`

Wartość krytyczną sprawdzamy, korzystając z centralnego rozkładu t o 2*n-2 stopniach swobody:

```
cr <- qt(alpha, 2*n-2, lower.tail=FALSE)
print(cr)
```

```
[1] 2.024394
```

Moc testu liczymy, sprawdzając, jaka część niecentralnego rozkładu t o $2n-2$ stopniach swobody i parametrze niecentralności równym `diff/se` znajduje się na prawo od wartości krytycznej:

```r
power <- pt(cr, 2*n-2, diff/se, lower.tail=FALSE)
print(power)
```

```
[1] 0.9588268
```

Może wydawać się dziwne, że do obliczenia mocy testu t korzystamy z błędu standardowego, skoro test t stosujemy wtedy, kiedy tego błędu nie znamy (a jedynie szacujemy go z próby). Należy jednak pamiętać, jak liczyliśmy błąd standardowy:

```r
se <- sigma * sqrt(2/n)
print(se)
```

```
[1] 1.581139
```

Parametr niecentralności, potrzebny nam do obliczenia mocy testu, możemy zatem podzielić na dwie składowe: jedna wiąże się bezpośrednio z liczebnością, a druga — z oczekiwaną różnicą między średnimi wyrażoną w odchyleniach standardowych: $d = (\mu_1 - \mu_2)/\sigma$. Wtedy ten nieznany nam parametr populacji staje się częścią wielkości przewidywanej przez hipotezę alternatywną:

```r
power <- pt(cr, 2*n-2, diff/sigma * sqrt(n/2), lower.tail=FALSE)
print(power)
```

```
[1] 0.9588268
```

Zadanie

Żeby obliczyć moc testu t, możemy też użyć wbudowanej funkcji `power.t.test()`. Korzystając z tej funkcji, proszę odpowiedzieć na poniższe pytanie:

Załóżmy, że porównujemy wiedzę statystyczną studentów UW i studentów UJ (mierzoną wynikiem jakiegoś testu), ale interesuje nas wykrycie tylko takiej różnicy, która będzie wynosiła co najmniej pół odchylenia standardowego (mniejsza wydaje nam się nieistotna na gruncie teoretycznym). Co najmniej ilu studentów z każdego z tych uniwersytetów musimy wylosować, żeby mieć nie mniej niż 80% szans na wykrycie takiej różnicy?