Spis treści

1 Rozkłady używane w statystyce

1.1 Rozkład normalny

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

gdzie μ to wartość oczekiwana, a σ^2 to wariancja. Rozkład normalny jest rozkładem ciągłym, który jest symetryczny względem średniej. Wartość oczekiwana i mediana są równe.

Listing 1: gęstość w punkcie x

```
dnorm(x, mean = 0, sd = 1)
```

Listing 2: dystrybuanta

```
pnorm(q, mean = 0, sd = 1)
```

Listing 3: kwantyl p-tego percentyla

```
qnorm(p, mean = 0, sd = 1)
```

Listing 4: n losowych zmiennych z rozkładu normalnego

```
rnorm(n, mean = 0, sd = 1)
```

1.2 Rozkład χ^2

Jeśli $X_1, \ldots, X_n$ są niezależne i $X_{1..n} \sim N(0,1)$ to:

$$Z = X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$$

$$\mathbb{E}Z = n \quad Var(Z) = 2n$$

Jeśli zmienne niezależne $X_1, \ldots, X_n \sim N(m_n, 1)$ to zmienna losowa też ma rozkład χ^2, lecz z parametrem niecentralności $m = \sqrt{m_1^2 + \ldots m_n^2}$. Wtedy $\mathbb{E}Z = n + m \quad Var(Z) = 2(k + 2m)$.

Listing 5: gęstość w punkcie x

```
dchisq(x, df = 1)
```

Listing 6: dystrybuanta

```
pchisq(q, df = 1)
```

Listing 7: kwantyl p-tego percentyla

```
qchisq(p, df = 1)
```

Listing 8: n losowych zmiennych z rozkładu chi-kwadrat

```
rchisq(n, df = 1)
```

1.3 Rozkład t-Studenta

Jeżeli $Z \sim N(0,1), X \sim \chi^2(k)$ to wtedy zmienna:

$$Y = \frac{Z}{\sqrt{\frac{X}{k}}} \sim t(k)$$

Listing 9: gęstość w punkcie x

```
dt(x, df = 1)
```

Listing 10: dystrybuanta

```
pt(q, df = 1)
```

Listing 11: kwantyl p-tego percentyla

```
qt(p, df = 1)
```

Listing 12: n losowych zmiennych z rozkładu t-Studenta

```
rt(n, df = 1)
```

1.4 Rozkład F-Snedecora

Jeżeli $X \sim \chi^2(k_1)$ i $Y \sim \chi^2(k_2)$ to wtedy zmienna:

$$Z = \frac{X/k_1}{Y/k_2} \sim F(k_1, k_2)$$

$$\mathbb{E}Z = \frac{k_2}{k_2 - 2} \quad Var(Z) = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2-2)^2(k_2-4)}$$

Listing 13: gęstość w punkcie x

```
df(x, df1 = 1, df2 = 1)
```

Listing 14: dystrybuanta

```
pf(q, df1 = 1, df2 = 1)
```

Listing 15: kwantyl p-tego percentyla

```
qf(p, df1 = 1, df2 = 1)
```

Listing 16: n losowych zmiennych z rozkładu F

```
rf(n, df1 = 1, df2 = 1)
```

2 Metoda estymacji punktowej

Jak dopasować rozkład i parametry do danych?

1. Wybieramy n próbek z danych $(X_1 \dots X_n)$
2. Patrzymy na histogram i oceniamy vibe
3. Dla parametrów wybieramy estymator, i przy pomocy estymatorów obliczamy parametry rozkładu
4. Sprawdzamy, czy rozkład pasuje do danych

Każdy estymator ma swój zakres ufności, z reguły określany przy pomocy wzoru. Mając obliczony zakres ufności, możemy określić poziom ufności, czyli błąd estymacji. Jeśli realna wartość parametru leży poza przedziałem ufności, to sugeruje wadę w estymacji. Poziom ufności to procent prób, w których przedział ufności zawiera prawdziwą wartość parametru.

3 Metoda Monte Carlo

Metoda Monte Carlo to metoda numeryczna, która polega na symulacji losowych próbek z rozkładu i obliczeniu wartości funkcji na podstawie tych próbek. W estymacji, na przykład, wystarczająco duża liczba próbek danych, pozwala stworzyć wykres wartości estymatora, który przybliża dystrybucje estymatora, co może pozwolić na lepsze oszacowanie wartości parametru.

Każda metoda, w której wykorzystujemy losowe próbki do obliczenia wartości funkcji, to metoda Monte Carlo.

4 Bootstrapping

Mając próbkę danych, możemy stworzyć wiele próbek z tej samej populacji, z reguły poprzez losowanie z zwracaniem. Dla wystarczająco dużej próbki początkowej w ten sposób możemy stworzyć wiele próbek, które będą miały podobny rozkład do oryginalnej próbki. Próba stworzona w ten sposób nazywana jest próbą bootsrapową.

5 Testowanie hipotez

Hipoteza statystyczne, to przypuszczenie dotyczące danych. Do weryfikacji hipotez korzystamy z testów. Dla konkretnego testu wyznacza się poziom istotności α, który określa prawdopodobieństwo odrzucenia hipotezy zerowej. Wielkość $1 - \beta$ dla układu hipotez prostych nazywa się mocą testu hipotezy zerowej wobec (prostej) hipotezy alternatywnej. Testy również są parametryzowane przez c czyli wartość krytyczną testu. W obecnie używanych implementacjach komputerowych wartości krytyczne zastępowane są tzw. p-wartościami (p-value), według pomysłu Ronalda Fishera. Jest to prawdopodobieństwo wylosowania próby takiej lub bardziej skrajnej, jak zaobserwowana przy założeniu, że hipoteza zerowa jest prawdziwa. Inaczej mówiąc, jest to prawdopodobieństwo, że zależność, jaką otrzymaliśmy w próbie z populacji mogła wystąpić przypadkowo, wskutek losowej zmienności, chociaż w populacji nie występuje.