Statystyka

Martyna Śpiewak

Bootcamp Data Science

Teoria estymacji

Teoria estymacji jest działem statystyki poświęconym szacowaniu wartości parametrów (bądź ich funkcji) rozkładu badanej cechy lub, ewentualnie, postaci rozkładu cechy.

- **estymacja parametryczna** — szacowanie parametrów rozkładu;
- **estymacja nieparametryczna** — szacowanie postaci rozkładu.

Sposoby szacowania wielkości

1. **estymacja punktowa** — dostarcza ocenę liczbową nieznanego parametru w postaci jednej, konkretnej wartości
   - **wady:** brak możliwości oceny dokładności oszacowania;
2. **estymacja przedziałowa** — dostarcza ocenę parametru za pomocą pewnego przedziału liczbowego, tzw. **przedziału ufności**, zawierającego prawdziwą wartość poszukiwanego parametru na z góry zadanym poziomie ufności;

Podstawowe własności estymatorów

Przyjmij, że badana cecha ma rozkład F_θ, gdzie θ jest nieznanym parametrem tego rozkładu.

Wartości parametru θ będziemy szacować na podstawie próby losowej

$$X_1, \ldots, X_n$$

pochodzącej z badanej populacji.

Estymatorem parametru θ rozkładu nazywamy dowolną statystykę z próby

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n).$$

Innymi słowy, estymatorem nazywamy każde narzędzie, za pomocą którego będziemy starali się dokonać oceny owego parametru.

Błąd średniokwadratowy

Dla danego parametru θ można utworzyć wiele estymatorów, to jednak interesować się będziemy wyłącznie takimi estymatorami, które „dobrze" szacują θ.

Najczęściej stosowanym kryterium oceny estymacji jest tzw. **błąd średniokwadratowy**

$$R(\hat{\theta}_n, \theta) = \mathbb{E}(\hat{\theta}_n - \theta)^2.$$

Zgodność

Mówimy, że estymator $\hat{\theta}_n$ parametru θ jest **zgodny**, jeżeli

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1 \quad \forall \varepsilon > 0.$$

Interpretacja: Zgodność estymatora odpowiada postulatowi, aby przy dostatecznie dużej liczebności próby estymator $\hat{\theta}_n$ przyjmował z dużym prawdopodobieństwem wartości bliskie estymowanemu parametrowi θ.

Nieobciążoność

Mówimy, że estymator $\hat{\theta}_n$ parametru θ jest **nieobciążony** jeżeli

$$\mathbb{E}(\hat{\theta}_n) = \theta.$$

W przeciwnym przypadku, gdy $\mathbb{E}(\hat{\theta}_n) \neq \theta$, estymator $\hat{\theta}_n$ nazywamy **obciążonym**, a wielkość

$$b_n(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

nazywamy **obciążeniem estymatora**.

Nieobciążoność

Interpretacja: Nieobciążoność estymatora oznacza, że uzyskiwane dzięki niemu oceny parametru nie są obciążone błędem systematycznym, tzn., że stosując go nie będzie z zasady ani przeszacowywać ani nie doszacowywać θ, ale średnio rzecz biorąc otrzymamy tyle, ile trzeba.

Warto zaznaczyć:

- estymator nieobciążony pozostanie dalej nieobciążony przy zmianie liczności próbki;
- estymator obciążony przy zwiększeniu liczności próbki może zmniejszyć obciążenie.

Nieobciążony estymator wartości oczekiwanej — przykład

Niech $X_1, \ldots, X_n$ będzie ciągiem zmiennych losowych z tego samego rozkładu o wartości oczekiwanej μ, gdzie μ jest nieznane.

Załóżmy, że estymatorem wartości oczekiwanej jest średnia arytmetyczna, tj.

$$\hat{\mu} = \overline{X}.$$

Czy zachodzi równość $\mathbb{E}\hat{\mu} = \mu$?

Nieobciążony estymator wartości oczekiwanej — przykład

Niech $X_1, \ldots, X_n$ będzie ciągiem zmiennych losowych z tego samego rozkładu o wartości oczekiwanej μ, gdzie μ jest nieznane.

Załóżmy, że estymatorem wartości oczekiwanej jest średnia arytmetyczna, tj.

$$\hat{\mu} = \overline{X}.$$

Czy zachodzi równość $\mathbb{E}\hat{\mu} = \mu$?

$$\mathbb{E}\hat{\mu} = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\mu = \mu.$$

Nieobciążony estymator wartości oczekiwanej — przykład

Niech $X_1, \ldots, X_n$ będzie ciągiem zmiennych losowych z tego samego rozkładu o wartości oczekiwanej μ, gdzie μ jest nieznane.

Załóżmy, że estymatorem wartości oczekiwanej jest średnia arytmetyczna, tj.

$$\hat{\mu} = \overline{X}.$$

Czy zachodzi równość $\mathbb{E}\hat{\mu} = \mu$?

$$\mathbb{E}\hat{\mu} = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\mu = \mu.$$

Wniosek: $\hat{\mu} = \overline{X}$ jest estymator nieobciążonym μ.

Nieobciążony estymator wariancji — przykład

Niech $X_1, \ldots, X_n$ będzie ciągiem zmiennych losowych z tego samego rozkładu o wartości oczekiwanej μ i wariancji σ^2.

Jak wyznaczyć nieobciążony estymator wariancji σ^2?

1. Zakładamy, że wartość oczekiwana μ jest znana.

$$\hat{\sigma}^2 = \tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

$$\mathbb{E}\hat{\sigma^2} = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(X_i^2 - 2X_i\mu + \mu^2\Big) = \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}X_i^2 - 2\mu\mathbb{E}X_i + \mathbb{E}\mu^2\Big)$$

$$= \Big|\mathrm{Var}(X_i) = \mathbb{E}X_i^2 - (\mathbb{E}X_i)^2\Big| = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + \mu^2 - 2\mu^2 + \mu^2) = \frac{1}{n}\cdot n\sigma^2 = \sigma^2.$$

Nieobciążony estymator wariancji — przykład

2a. Zakładamy, że wartość oczekiwana μ jest nieznana.

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

$$\begin{aligned}
\mathbb{E}S^2 &= \mathbb{E}\Big(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big) = \mathbb{E}\Big(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2\Big) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\Big(\mathbb{E}(X_i-\mu)^2 - 2\mathbb{E}(X_i-\mu)(\overline{X}-\mu) + \mathbb{E}(\overline{X}-\mu)^2\Big) \\
&= \frac{n\sigma^2}{n-1} - \frac{2\sigma^2}{n-1} + \frac{\sigma^2}{n-1} = \sigma^2
\end{aligned}$$

Obciążony estymator wariancji — przykład

2b. Zakładamy, że wartość oczekiwana μ jest nieznana.

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

$$\mathbb{E}S_n^2 = \frac{n-1}{n}\mathbb{E}S^2 = \frac{n-1}{n}\sigma^2.$$

Wniosek: Estymator S_n^2 jest estymator obciążonym parametru σ^2.

Obciążenie estymatora wynosi

$$b_n(S_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2 < 0,$$

tzn. estymator niedoszacowuje wartości σ^2.

Asymptotyczna nieobciążoność

Mówimy, że estymator $\hat{\theta}_n$ parametru θ jest **asymptotycznie nieobciążony** jeżeli

$$\lim_{n\to\infty} b_n(\hat{\theta}_n) = \lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) - \theta = 0.$$

Interpretacja: Dla dostatecznie dużej próby obciążenie estymatora asymptotycznie nieobciążonego jest pomijalne.

Asymptotyczna nieobciążoność — przykład

Niech

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Wiemy, że obciążenie estymatora S_n^2 wynosi

$$b_n(S_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2 < 0,$$

stąd

$$\lim_{n\to\infty} b_n(S_n^2) = 0.$$

Wniosek: Estymator wariancji S_n^2 jest obciążony, ale jest asymptotycznie nieobciążony.

Twierdzenie

Jeśli estymator jest zgodny, to jest asymptotycznie nieobciążony.

Twierdzenie

Jeśli estymator $\hat{\theta}_n$ jest asymptotycznie nieobciążony oraz jeżeli jego wariancja spełnia warunek

$$\lim_{n \to \infty} \mathrm{Var}(\hat{\theta}_n) = 0,$$

to $\hat{\theta}_n$ jest zgodny.

Dla danego parametru θ może istnieć wiele estymatorów nieobciążonych. **Pozostaje więc kwestia wyboru najlepszego z nich.**

Jeśli więc $\hat{\theta}_n^*$ i $\hat{\theta}_n^{**}$ są dwoma estymatorami nieobciążonymi parametru θ, to powiemy, że $\hat{\theta}_n^*$ jest **estymatorem efektywniejszym**, niż $\hat{\theta}_n^{**}$, gdy

$$\mathrm{Var}(\hat{\theta}_n^*) < \mathrm{Var}(\hat{\theta}_n^{**}).$$

Interpretacja: Oznacza to, że ten estymator jest efektywniejszy, którego wartości są bardziej skupione wokół θ.

Efektywność

Estymator nieobciążony parametru θ, który ma najmniejszą wariancję spośród wszystkich nieobciążonych estymatorów danego parametru, nazywamy **estymatorem efektywnym** (najefektywniejszym).

Przypuśćmy, że nieznany parametr θ można wyrazić za pomocą funkcji kilku momentów rozkładu badanej cechy, tzn.

$$\theta = g(\mathbb{E}X, \mathbb{E}X^2, \ldots, \mathbb{E}X^r).$$

Oznaczmy, że M_k k-ty moment empiryczny z próby $X_1, \ldots, X_n$, gdzie

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Idea **metody momentów** polega na tym, że za estymator poszukiwanego parametru θ przyjmuje się wspomniana funkcję, tyle że momentów empirycznych, a nie teoretycznych.

Estymator $\hat{\theta}_n$ parametru θ wyznaczonym metodą momentów jest wielkość

$$\hat{\theta}_n = g(M_1, M_2, \ldots, M_r).$$

Metoda momentów

Zalety:

- prostota;
- zazwyczaj zgodny.

Wady:

- słabe własności statystyczne: obciążone i nieefektywne;

Metoda momentów — przykład dla rozkładu jednostajnego

Niech $X_1, \ldots, X_n$ oznacza próbkę prostą z rozkładu jednostajnego $\mathrm{U}(0, t)$. Skonstruuj estymator parametru t posługując się metodą momentów.

Wiemy, że

$$\mathbb{E}X = \frac{0 + t}{2} = \frac{t}{2} \implies t = 2 \cdot \mathbb{E}X.$$

Metoda momentów — przykład dla rozkładu jednostajnego

Niech $X_1, \ldots, X_n$ oznacza próbkę prostą z rozkładu jednostajnego $\mathrm{U}(0, t)$. Skonstruuj estymator parametru t posługując się metodą momentów.

Wiemy, że

$$\mathbb{E}X = \frac{0+t}{2} = \frac{t}{2} \implies t = 2 \cdot \mathbb{E}X.$$

Wówczas

$$\mathrm{EMM}: \hat{t} = 2M_1 = 2\overline{X}.$$

Metoda momentów — przykład dla rozkładu normalnego

Niech $X_1, \ldots, X_n$ oznacza próbkę prostą z rozkładu normalnego $\mathcal{N}(\mu, \sigma)$. Skonstruuj estymator parametrów μ i σ^2 posługując się metodą momentów.

Wiemy, że

$$\mu = \mathbb{E}X \quad \text{oraz} \quad \sigma^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Wówczas

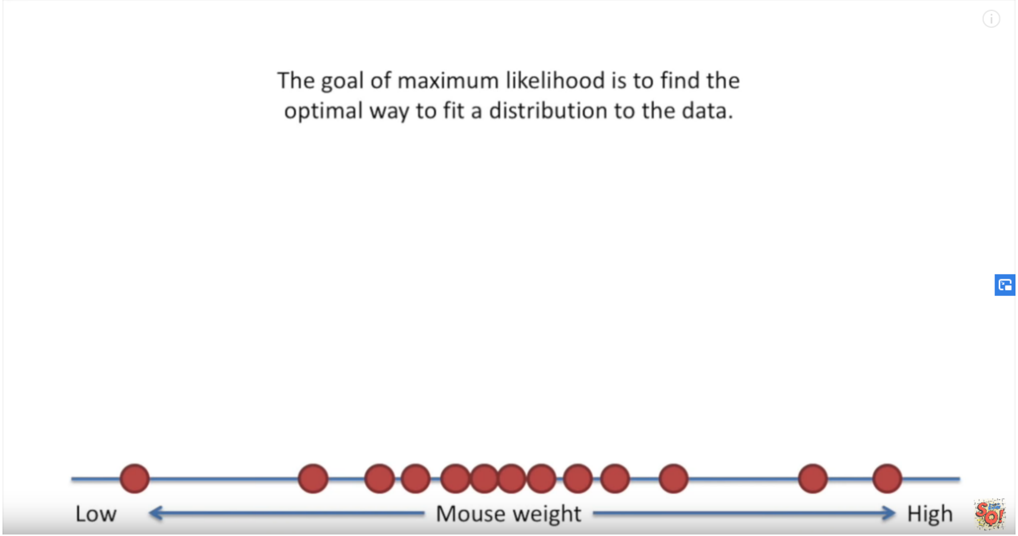
$$\hat{\mu} = M_1 = \overline{X},$$

$$\hat{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\Big) = S_n^2.$$

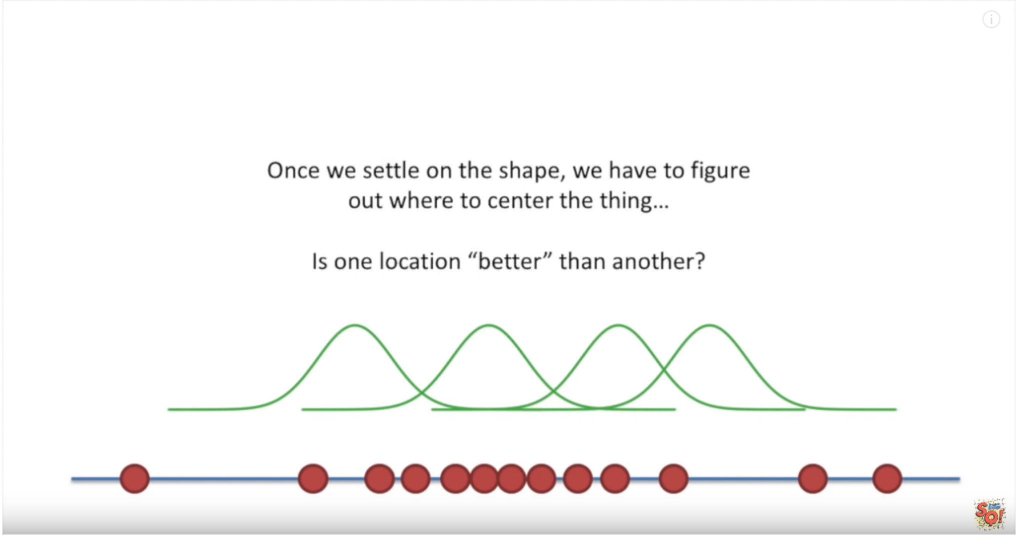
Wniosek:

$$\text{EMM} = \begin{cases} \hat{\mu} = \overline{X}, \\ \hat{\sigma}^2 = S_n^2. \end{cases}$$

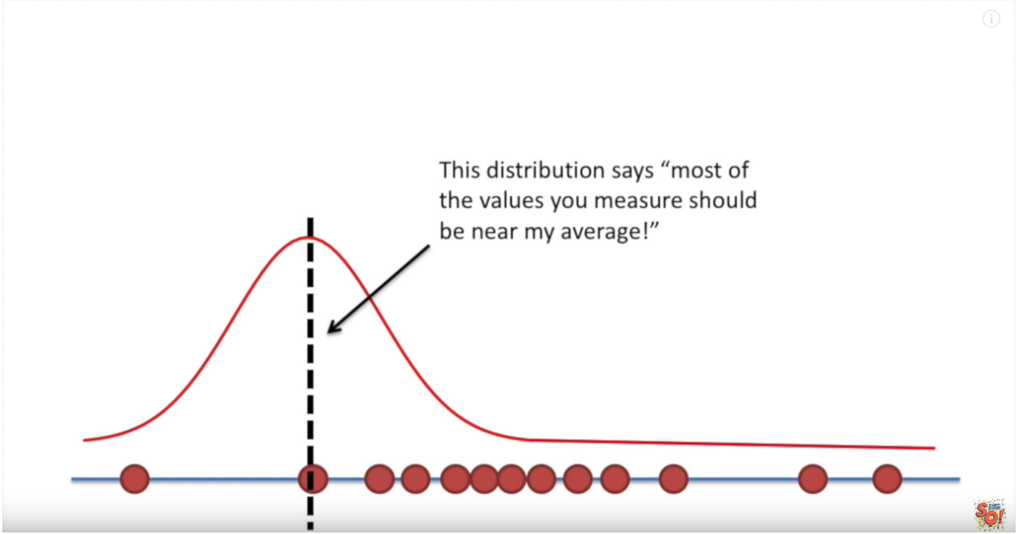
Idea **metody największej wiarogodności** sprowadza się do wyboru takiej wartości $\hat{\theta}_n$, jako estymatora parametru θ, która maksymalizuje prawdopodobieństwo (lub gęstość rozkładu cechy) otrzymania takiej realizacji próby, jaką właśnie otrzymano.



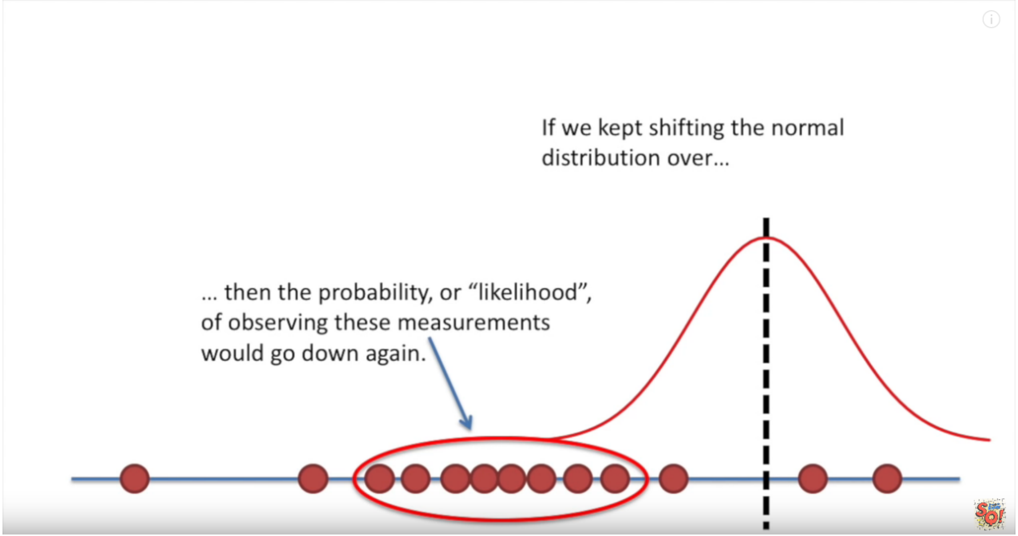
https://www.youtube.com/watch?v=XepXtl9YKwc



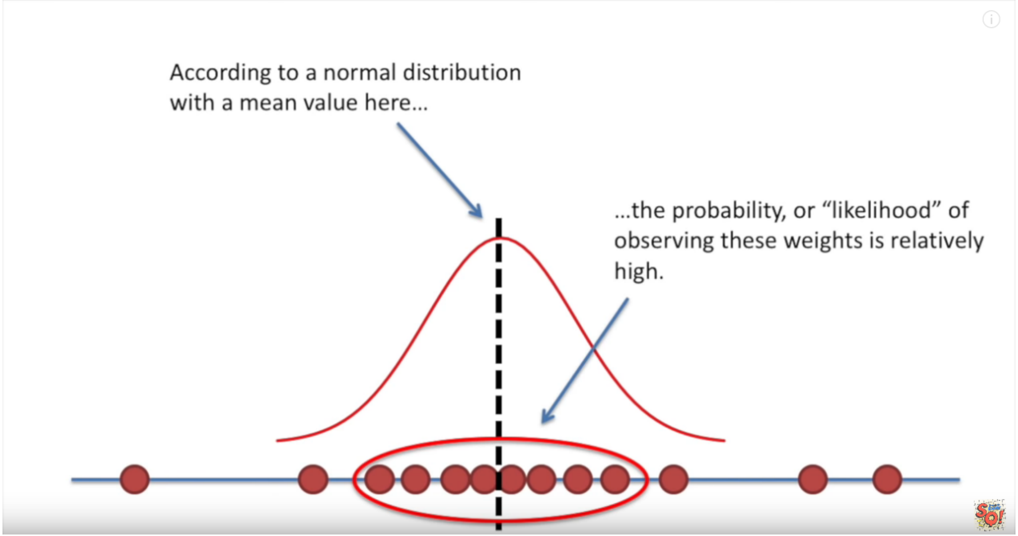
https://www.youtube.com/watch?v=XepXtl9YKwc



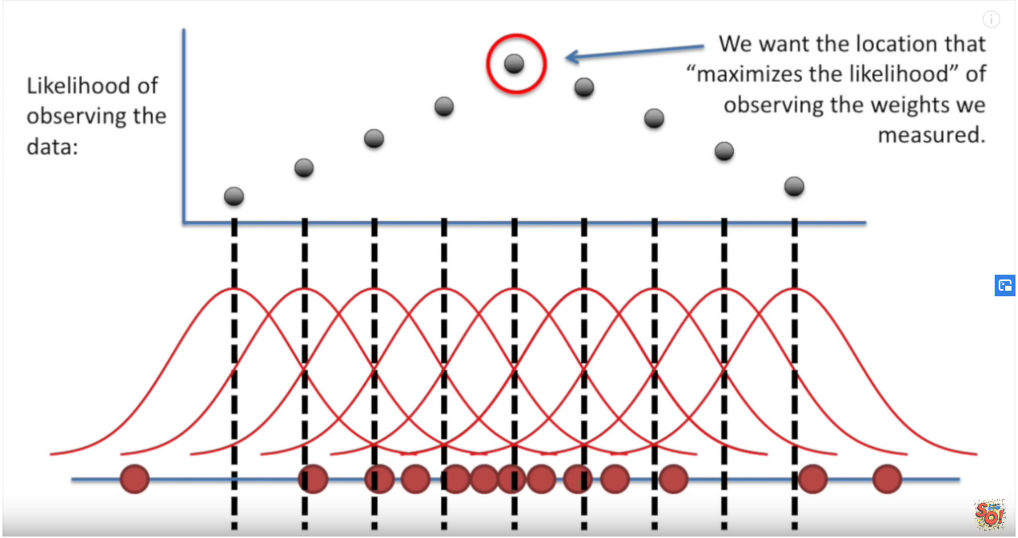
https://www.youtube.com/watch?v=XepXtl9YKwc



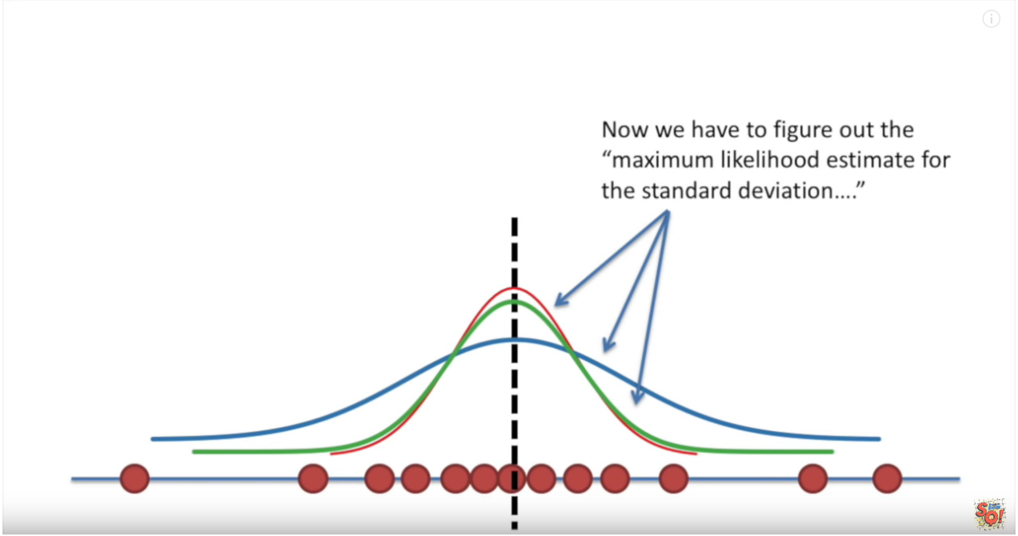
https://www.youtube.com/watch?v=XepXtl9YKwc



https://www.youtube.com/watch?v=XepXtl9YKwc



https://www.youtube.com/watch?v=XepXtl9YKwc



https://www.youtube.com/watch?v=XepXtl9YKwc

Funkcja wiarogodności

Niech $x_1, \ldots, x_n$ będzie realizacją próby $X_1, \ldots, X_n$.

- Jeżeli rozkład badanej cechy jest dyskretny, wówczas **funkcja wiarogodności** dla realizacji próby nazywamy wyrażenie

$$L(x_1, \ldots, x_n; \theta) = p(x_1; \theta) \cdot \ldots \cdot p(x_n; \theta),$$

  gdzie $p(x_i; \theta)$ oznacza prawdopodobieństwo przyjęcia przez zmienną losową X wartości x_i.
- Jeżeli rozkład badanej cechy jest ciągły **funkcja wiarogodności** przyjmuje postać

$$L(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdot \ldots \cdot f(x_n; \theta),$$

  gdzie $f(x_i; \theta)$ oznacza gęstość rozkładu.

Metoda największej wiarogodności

$\hat{\theta}_n$ jest **estymatorem największej wiarogodności** parametru θ, jeżeli maksymalizuje on wartość funkcji wiarogodności

$$L(x_1, \ldots, x_n; \theta).$$

Zalety:

- dobre własności statystyczne: są zgodne, co najmniej asymptotycznie nieobciążone;
- wiadomo, że jeśli w danym przypadku istnieje estymator efektywny, to można go uzyskać metodą największej wiarogodności;

Algorytm wyznaczania estymatora największej wiarogodności

Założenie: Funkcja $\ln L$ jest co najmniej dwukrotnie różniczkowalna względem zmiennej θ.

1. znaleźć funkcję wiarogodności L;
2. znaleźć $\ln L$;
3. obliczyć pochodną cząstkową: $\frac{\partial}{\partial\theta} \ln L$;
4. znaleźć rozwiązanie θ_0 równania $\frac{\partial}{\partial\theta} \ln L = 0$;
5. sprawdzi, czy w θ_0, funkcja $\ln L$ osiąga maksimum

$$\left. \frac{\partial^2}{\partial\theta^2} \ln L \right|_{\theta=\theta_0} < 0.$$

Algorytm wyznaczania estymatora największej wiarogodności

Jeżeli jest spełniony ostatni warunek, oznacza to, że w punkcie θ_0 funkcja $\ln L$, a także funkcja L osiąga maksimum, a więc

$$\hat{\theta}_n = \theta_0$$

jest *estymatorem największej wiarogodności*.

Metoda największej wiarogodności — przykład dla rozkładu normalnego

Niech $X_1, \ldots, X_n$ oznacza próbkę prostą z rozkładu normalnego $\mathcal{N}(\mu, \sigma)$. Skonstruuj estymator parametrów μ i σ^2 posługując się metodą największej wiarogodności.

Przypomnijmy, że gęstość rozkładu normalnego jest postaci

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x-\mu)^2}{2\sigma^2} \quad \text{dla} \quad x \in \mathbb{R}.$$

Metoda największej wiarogodności:

1. $L = L(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x_i-\mu)^2}{2\sigma^2}$
2. $l = \ln L = -\ln\left(\sqrt{2\pi}\sigma\right)^n - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} =$
   $-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$

Metoda największej wiarogodności — przykład dla rozkładu normalnego

3a. Liczymy pochodną cząstkowe dla parametru μ

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

4a. Porównujemy pochodną cząstkową $\frac{\partial l}{\partial \mu}$ do zera

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{1}{\sigma^2} \Big(\sum_{i=1}^{n} x_i - n\mu \Big) = 0 \qquad \Longrightarrow \qquad \mu_0 = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

5a. Liczymy drugą pochodną cząstkowe dla parametru μ

$$\left. \frac{\partial^2 l}{\partial \mu^2} \right|_{\mu = \mu_0} = -\frac{1}{\sigma^2} < 0$$

Wniosek: ENW: $\hat{\mu} = \overline{X}$.

3b. Liczymy pochodną cząstkowe dla parametru σ^2

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

4b. Porównujemy pochodną cząstkową $\frac{\partial l}{\partial \sigma^2}$ do zera

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \qquad \Longrightarrow \qquad \sigma_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = S_n^2.$$

5b. Liczymy drugą pochodną cząstkowe dla parametru σ^2

$$\left.\frac{\partial^2 l}{\partial (\sigma^2)^2}\right|_{\sigma^2 = S_n^2} = -\frac{n}{2S_n^2} < 0$$

Wniosek: ENW: $\hat{\sigma}^2 = S_n^2$.

Estymatory wartości średniej

Średnia arytmetyczna z próby

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- zgodny;
- nieobciążony;
- jeśli badana cecha ma rozkład normalny, jest estymatorem efektywnym.

Mediana z próby

- zgodny;
- asymptotycznie nieobciążony.

Estymatory wariancji

- gdy znana jest wartość oczekiwana μ rozkładu badanej cechy

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

  - zgodny;
  - nieobciążony;
  - jeśli badana cecha ma rozkład normalny, jest estymatorem efektywnym.

Estymatory wariancji

- gdy wartość oczekiwana μ rozkładu badanej cechy nie jest znana

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  - zgodny;
  - nieobciążony;
  - jeśli badana cecha ma rozkład normalny, jest estymatorem efektywnym.

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  - zgodny;
  - asymptotycznie nieobciążony;
  - jeśli badana cecha ma rozkład normalny, jest estymatorem największej wiarogodności.

Przedział ufności

Przedział losowy $(\underline{\theta}, \overline{\theta})$, którego końcami są statystyki

$$\underline{\theta} = \underline{\theta}(X_1, \ldots, X_n) \quad \text{oraz} \quad \overline{\theta} = \overline{\theta}(X_1, \ldots, X_n),$$

gdzie $\underline{\theta} < \overline{\theta}$, nazywamy **przedziałem ufności** dla parametru θ **na poziomie ufności** $0 < 1 - \alpha < 1$, jeżeli

$$P\Big(\underline{\theta}(X_1, \ldots, X_n) < \theta < \overline{\theta}(X_1, \ldots, X_n)\Big) \geq 1 - \alpha.$$

Długość przedziału

W praktyce interesować nas będą przedziały ufności o jak najmniejszej długości, bowiem owa **długość przedziału**

$$l_n = \overline{\theta}(X_1, \ldots, X_n) - \underline{\theta}(X_1, \ldots, X_n)$$

jest **miarą precyzji estymacji.**

Przedział ufności dla wartości średniej — model 1

Niech $X_1, \ldots, X_n$ będzie próbą prostą z populacji o rozkładzie normalnym $\mathcal{N}(\mu, \sigma)$ o znanej wariancji σ^2.

Wtedy dla ustalonego poziomu ufności $1 - \alpha$ najkrótszy przedział ufności dla wartości oczekiwanej ma postać

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right),$$

gdzie $z_{1-\frac{\alpha}{2}}$ oznacza kwantyl rozkładu normalnego standardowego rzędu $1 - \frac{\alpha}{2}$.

Przedział ufności dla wartości średniej — model 2

Niech $X_1, \ldots, X_n$ będzie próbą prostą z populacji o rozkładzie normalnym $\mathcal{N}(\mu, \sigma)$ o nieznanej wariancji σ^2.

Wtedy dla ustalonego poziomu ufności $1 - \alpha$ najkrótszy przedział ufności dla wartości oczekiwanej ma postać

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}}^{[n-1]} \frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}}^{[n-1]} \frac{S}{\sqrt{n}}\right),$$

gdzie $t_{1-\frac{\alpha}{2}}^{[n-1]}$ oznacza kwantyl rzędu $1 - \frac{\alpha}{2}$ rozkładu t-Studenta o $n - 1$ stopniach swobody.

Przedział ufności dla wartości średniej — model 3

Niech $X_1, \ldots, X_n$ będzie dostatecznie dużą próbą ($n \geq 100$) o dowolnym rozkładzie o nieznanej, ale skończonej wartości oczekiwanej i wariancji.

Wtedy dla ustalonego poziomu ufności $1 - \alpha$ najkrótszy przedział ufności dla wartości oczekiwanej ma postać

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right),$$

gdzie $z_{1-\frac{\alpha}{2}}$ oznacza kwantyl rozkładu normalnego standardowego rzędu $1 - \frac{\alpha}{2}$.

Załóżmy, że badana cecha ma rozkład dwupunktowy z nieznanym parametrem p, a liczność próby jest dostatecznie duża ($n \geq 100$).

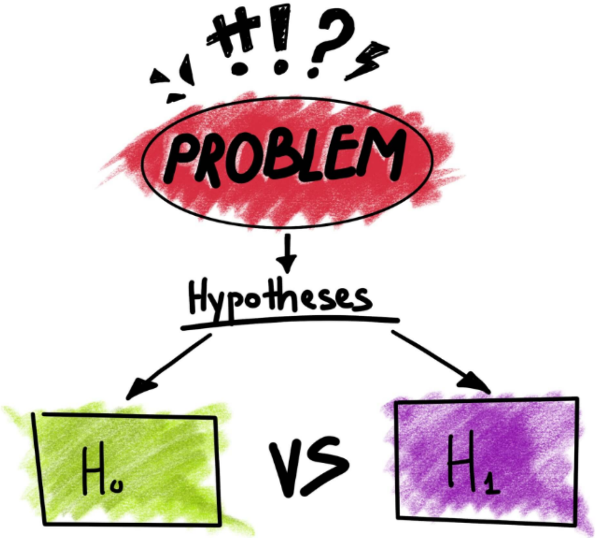
Z centralnego twierdzenia granicznego Moivre'a-Laplace'a wynika, że statystyka

$$\frac{k}{n},$$

gdzie k oznacza liczbę elementów wyróżnionych w próbie ma w przybliżeniu rozkład normalny $\mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}})$.

Przedział ufność dla wskaźnika struktury p przyjmuje postać

$$\left(\frac{k}{n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}, \frac{k}{n} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}\right).$$



https://towardsdatascience.com/demystifying-hypothesis-testing-with-simple-python-examples-4997ad3c5294

Weryfikacja hipotez

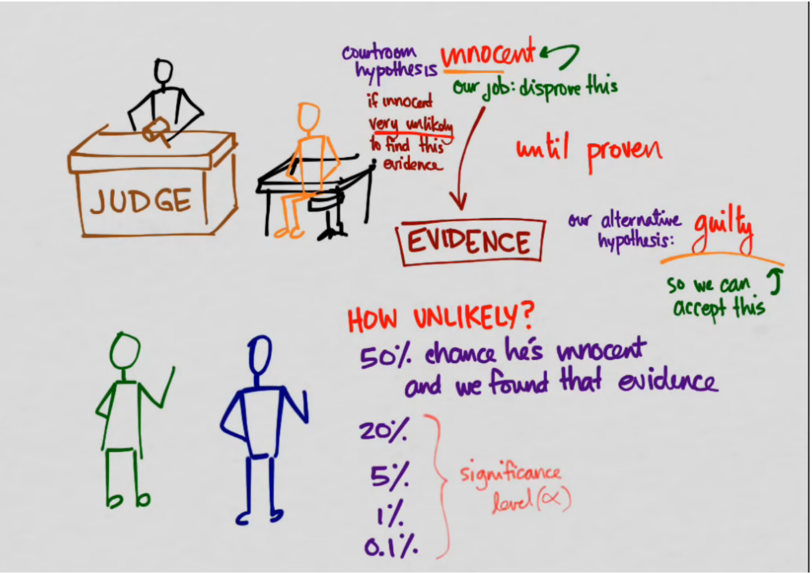
Hipotezą statystyczną nazywamy dowolne przypuszczenie dotyczące rozkładu badanej cechy.

Weryfikacji takiej hipotezy dokonuje się na podstawie pobranej próby losowej. Jej zaś celem jest odpowiedź na pytanie, czy postawiona hipoteza jest prawdziwa czy też fałszywa.

Narzędzia służące do weryfikacji hipotez nazywamy **testami statystycznymi.**

Hipotezy/testy statystyczne dzieli na

- parametryczne;
- nieparametryczne.



https://www.youtube.com/watch?v=z5gPXoRkic

Podstawowe pojęcia

Niech $(\chi, \mathcal{A}, \mathcal{P})$ będzie **przestrzenią statystyczną**, gdzie

- χ oznacza przestrzeń prób (zbiór możliwych wyników obserwacji);
- $\mathcal{A}$ jest σ-ciałem podzbiorów zbioru χ;
- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ jest rodziną rozkładów prawdopodobieństwa na $\mathcal{A}$.

Podstawowe pojęcia

Rozpatrujemy pewną hipotezę H dotyczącą parametru θ.

Rodzinę rozkładów $\mathcal{P}$ można podzielić na dwie rozłączne podrodziny:

- podrodzinę $\{P_\theta : \theta \in \Theta_{H_0}\}$ zawierającą rozkłady, dla których rozważana hipoteza jest prawdziwa,
- podrodzinę $\{P_\theta : \theta \in \Theta_{H_1}\}$ zawierającą rozkłady, dla których rozważana hipoteza jest fałszywa,

gdzie $\Theta_{H_0}, \Theta_{H_1} \in \Theta$ oraz $\Theta_{H_0} \cap \Theta_{H_1} = \emptyset$.

Hipoteza zerowa

$$H_0 : \theta \in \Theta_{H_0}$$

Hipoteza alternatywna

$$H_1 : \theta \in \Theta_{H_1}$$

Na podstawie zaobserwowanej próby losowej $X_1, \ldots, X_n$ możemy podjąć jedna z dwóch decyzji:

- przyjąć H_0 i odrzucić H_1;
- odrzucić H_0 i przyjąć H_1.

Test statystyczny

Testem statystycznym nazywamy regułę decyzyjną, przypisującą możliwym realizacjom próby losowej $X_1, \ldots, X_n$ decyzję odrzucenia lub przyjęcia weryfikowanej hipotezy.

Test hipotezy H_0 będziemy utożsamiali z funkcją $\varphi : \chi \to \{0, 1\}$, gdzie 0 odpowiada **przyjęciu** hipotezy zerowej, natomiast 1 jej **odrzuceniu**.

Każdy test statystyczny rozbija przestrzeń prób χ na dwa rozłączne podzbioru

- $\{(x_1, \ldots, x_n) \in \chi : \varphi(x_1, \ldots, x_n) = 0\}$ — zbiór przyjęć hipotezy H_0;
- $W_\alpha = \{(x_1, \ldots, x_n) \in \chi : \varphi(x_1, \ldots, x_n) = 1\}$ — zbiór odrzuceń hipotezy H_0 nazywany **obszarem krytycznym**.

Test statystyczny

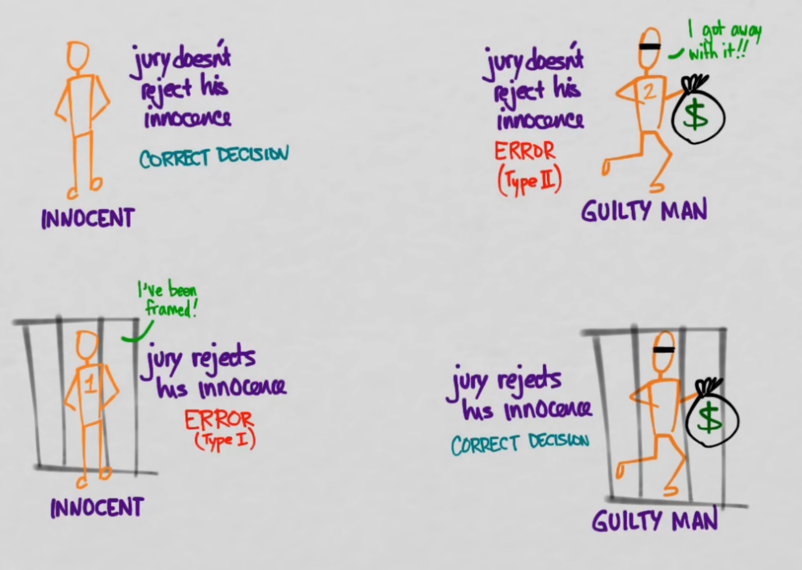
Test statystyczny ma następującą postać

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{gdy} \quad T(X_1, \ldots, X_n) \in W_\alpha \\ 0 & \text{gdy} \quad T(X_1, \ldots, X_n) \notin W_\alpha, \end{cases}$$

gdzie $T = T(X_1, \ldots, X_n)$ jest pewną funkcją próby zwaną **statystyką testową**.

Decyzja

		Sytuacja faktyczna	
		H_0-prawdziwa	H_0-fałszywa
Decyzja	przyjęcie H_0	☺	błąd II rodzaju
	odrzucenie H_0	błąd I rodzaju	☺



https://www.youtube.com/watch?v=z5gPXoRkic

Błąd I i II rodzaju

Błąd pierwszego rodzaju — odrzucenie H_0, gdy jest ona prawdziwa

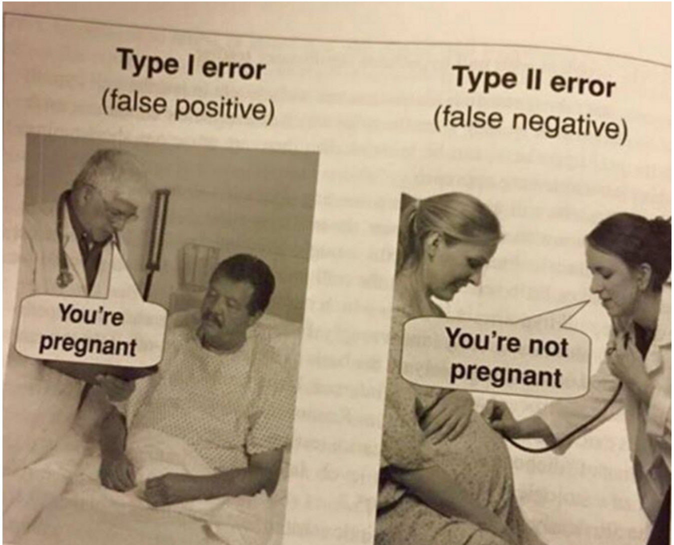
$$\alpha_\varphi = P(\varphi(X_1, \ldots, X_n) = 1 | H_0) = P(T(X_1, \ldots, X_n) \in W_\alpha | H_0)$$

Błąd drugiego rodzaju — przyjęcie H_0, gdy jest ona fałszywa

$$\beta_\varphi = P(\varphi(X_1, \ldots, X_n = 0 | H_1) = P(T(X_1, \ldots, X_n) \notin W_\alpha | H_1)$$

H_0 : You're not pregnant

H_1 : You're pregnant.



Poziom istotności testu

W klasycznej teorii weryfikacji hipotez testy konstruuje się w ten sposób, że

- przyjmuje się górne ograniczenie na prawdopodobieństwo popełnienia błędu pierwszego rodzaju, tzw. **poziom istotności testu** α:

$$\alpha_\varphi \leq \alpha$$

- a następnie poszukuje się takiego testu, który — przy ograniczeniu na błąd pierwszego rodzaju — minimalizuje prawdopodobieństwo popełnienia błędu drugiego rodzaju:

$$\beta_\varphi \to \text{min.}$$

Klasyczny algorytm testowania hipotez

1. postawić hipotezę zerową H_0 i hipotezę alternatywną H_1;
2. wyspecyfikować model matematyczny (np. zakładamy, że próba losowa pochodzi z rozkładu normalnego o nieznanej wariancji);
3. przyjąć poziom istotności α;
4. obliczyć wartość statystyki testowej $T = T(X_1, \ldots, X_n)$
5. wyznaczyć obszar krytyczny W_α (w zależności od przyjętego poziomu istotności oraz hipotezy alternatywnej);
6. podjąć decyzję
    - jeśli $T \in W_\alpha$, wówczas odrzuć hipotezę H_0,
    - jeśli $T \notin W_\alpha$, wówczas nie ma podstaw do odrzucenia hipotezy H_0;

p-**wartością** (istotnością testu) nazywamy najmniejszy poziom istotności, przy którym zaobserwowana wartość statystyki testowej prowadzi do odrzucenia rozważanej hipotezy zerowej.

Algorytm testowania hipotez

1. postawić hipotezę zerową H_0 i hipotezę alternatywną H_1;
2. wyspecyfikować model matematyczny;
3. przyjąć poziom istotności α;
4. obliczyć wartość statystyki testowej $T = T(X_1, \ldots, X_n)$
5. obliczyć p-wartość;
6. podjąć decyzję
    - jeśli $p \leq \alpha$, wówczas odrzuć hipotezę H_0,
    - jeśli $p > \alpha$, wówczas nie ma podstaw do odrzucenia hipotezy H_0;

Testy dla wartości oczekiwanej

Załóżmy, że jesteśmy zainteresowaniu weryfikacją hipotezy dotyczącej wartości oczekiwanej μ:

$$H_0 : \mu = \mu_0,$$

wobec jednej z trzech hipotez alternatywnych

$$H_1 : \mu \neq \mu_0$$

$$H_1' : \mu < \mu_0$$

$$H_1'' : \mu > \mu_0.$$

Test dla wartości oczekiwanej — model 1

Załóżmy, że badana cecha X ma rozkład normalny $\mathcal{N}(\mu, \sigma)$ o znanym odchyleniu standardowym σ.

Statystyka testowa przyjmuje w tym przypadku postać

$$T = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}.$$

Przy założeniu prawdziwości hipotezy H_0 statystyka ma rozkład normalny standardowy $\mathcal{N}(0, 1)$, z związku z czym obszar krytyczny — w zależności od przyjętej hipotezy alternatywnej — ma postać

$$W_\alpha = (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, +\infty),$$
$$W_\alpha' = (-\infty, -z_{1-\alpha}],$$
$$W_\alpha'' = [z_{1-\alpha}, +\infty].$$

Test dla wartości oczekiwanej — model 2

Jeżeli cecha X ma rozkład normalny $\mathcal{N}(\mu, \sigma)$ o nieznanym odchyleniu standardowym σ, to do weryfikacji hipotezy H_0 wykonujemy test zbudowany na statystyce

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n},$$

która przy założeniu prawdziwości hipotezy H_0 na rozkład t-Studenta o $n-1$ stopniach swobody.

W zależności od przyjętej hipotezy alternatywnej obszar krytyczny przybiera postać

$$W_\alpha = (-\infty, -t_{1-\frac{\alpha}{2}}^{[n-1]}] \cup [t_{1-\frac{\alpha}{2}}^{[n-1]}, +\infty),$$
$$W_\alpha' = (-\infty, -t_{1-\alpha}^{[n-1]}],$$
$$W_\alpha'' = [t_{1-\alpha}^{[n-1]}, +\infty).$$

Test dla wartości oczekiwanej — model 3

Jeżeli próba pochodzi z dowolnego rozkładu (posiadającego jednakże skończoną wariancję), ale jest wystarczająco duża ($n \geq 100$), wówczas statystyka testowa przyjmuje postać

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}.$$

Przy założeniu prawdziwości hipotezy H_0 i dla dostatecznie dużej próby statystyka ma w przybliżeniu rozkład normalny standardowy $\mathcal{N}(0, 1)$, w związku z czym obszar krytyczny — w zależności od przyjętej hipotezy alternatywnej — ma postać

$$W_\alpha = (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, +\infty),$$
$$W_\alpha' = (-\infty, -z_{1-\alpha}],$$
$$W_\alpha'' = [z_{1-\alpha}, +\infty].$$

Testy dla dwóch prób niezależnych

W praktyce istotną rolę odgrywają testy statystyczne, za pomocą których można porównywać wartości oczekiwane badanej cechy w dwóch różnych zbiorowościach statystycznych.

W szczególności interesująca jest weryfikacja hipotezy, że obie porównywalne średnie sa jednakowe

$$H_0 : \mu_1 = \mu_2,$$

przy jednej z trzech hipotez alternatywnych:

$$H_1 : \mu_1 \neq \mu_2$$
$$H_1' : \mu_1 < \mu_2$$
$$H_1'' : \mu_1 > \mu_2.$$

Test dla dwóch prób niezależnych — model 1

Załóżmy, że próby $X_1, \ldots, X_{n_1}$ i $Y_1, \ldots, Y_{n_2}$ są niezależne i pochodzą z populacji o rozkładach normalnych, odpowiednio, $\mathcal{N}(\mu_1, \sigma_1)$ i $\mathcal{N}(\mu_2, \sigma_2)$ oraz odchylenia standardowe σ_1 i σ_2 są znane.

Wówczas statystyka testowa ma postać

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Statystyka przy założeniu prawdziwości hipotezy zerowej H_0 ma rozkład normalny $\mathcal{N}(0, 1)$. Obszar krytyczny — w zależności od przyjętej hipotezy alternatywnej — ma postać

$$W_\alpha = (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, +\infty),$$
$$W_\alpha^{'} = (-\infty, -z_{1-\alpha}],$$
$$W_\alpha^{''} = [z_{1-\alpha}, +\infty].$$

Test dla dwóch prób niezależnych — model 2

Załóżmy, że próby $X_1, \ldots, X_{n_1}$ i $Y_1, \ldots, Y_{n_2}$ są niezależne i pochodzą z populacji o rozkładach normalnych, odpowiednio, $\mathcal{N}(\mu_1, \sigma_1)$ i $\mathcal{N}(\mu_2, \sigma_2)$ o nieznanych odchyleniach standardowych σ_1 i σ_2, ale równych, tzn. $\sigma_1 = \sigma_2$.

Wówczas statystyka testowa ma postać

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Statystyka przy założeniu prawdziwości hipotezy zerowej H_0 ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody. Obszar krytyczny — w zależności od przyjętej hipotezy alternatywnej — ma postać

$$W_\alpha = (-\infty, -t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]}] \cup [t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]}, +\infty),$$

$$W_\alpha^{'} = (-\infty, -t_{1-\alpha}^{[n_1+n_2-2]}],$$

$$W_\alpha^{''} = [t_{1-\alpha}^{[n_1+n_2-2]}, +\infty].$$

Test dla dwóch prób niezależnych — model 3

Załóżmy, że próby $X_1, \ldots, X_{n_1}$ i $Y_1, \ldots, Y_{n_2}$ są niezależne i pochodzą z populacji o rozkładach normalnych, odpowiednio, $\mathcal{N}(\mu_1, \sigma_1)$ i $\mathcal{N}(\mu_2, \sigma_2)$ o nieznanych i różnych odchyleniach standardowych σ_1 i σ_2, tzn. $\sigma_1 \neq \sigma_2$. Ponadto, próby są dostatecznie **duże**.

Wówczas statystyka testowa ma postać

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Statystyka przy założeniu prawdziwości hipotezy zerowej H_0 ma rozkład normalny standardowy. Obszar krytyczny ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, +\infty), \\ W_\alpha' &= (-\infty, -z_{1-\alpha}], \\ W_\alpha'' &= [z_{1-\alpha}, +\infty]. \end{aligned}$$

Testy dla wskaźnika struktury

Zakładamy, że próba pochodzi z rozkładu dwupunktowego. Weryfikowana hipoteza dotyczy nieznanego parametru p

$$H_0 : p = p_0,$$

wobec jednej z trzech hipotez alternatywnych

$$H_1 : p \neq p_0$$
$$H_1' : p < p_0$$
$$H_1'' : p > p_0.$$

Do weryfikacji hipotezy H_0 wykorzystujemy wskaźnik struktury z próby

$$\hat{p} = \frac{k}{n},$$

gdzie k jest liczbą elementów wyróżnionych w próbie o liczności n.

Test dla wskaźnika struktury — model 1

Jeżeli dysponujemy liczbą próbką ($n \geq 100$), wówczas statystyka testowa ma postać

$$T = \frac{k - np_0}{\sqrt{np_0(1-p_0)}}.$$

Na podstawie centralnego twierdzenia granicznego Moivre'a-Laplace'a wiemy, że statystyka T ma w przybliżeniu rozkład $\mathcal{N}(0,1)$. Obszar krytyczny — w zależności od przyjętej hipotezy alternatywnej — ma postać

$$W_\alpha = (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, +\infty),$$
$$W'_\alpha = (-\infty, -z_{1-\alpha}],$$
$$W''_\alpha = [z_{1-\alpha}, +\infty].$$

Test dla wskaźnika struktury — model 2

Jeżeli próba nie jest dostatecznie duża korzystamy ze statystyki testowej

$$T = 2(\arcsin\sqrt{\frac{k}{n}} - \arcsin\sqrt{p_0})\sqrt{n}$$

mającej w przybliżeniu rozkład normalny standardowy. Obszar krytyczny — w zależności od przyjętej hipotezy alternatywnej — ma postać

$$W_\alpha = (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, +\infty),$$
$$W'_\alpha = (-\infty, -z_{1-\alpha}],$$
$$W''_\alpha = [z_{1-\alpha}, +\infty].$$

$$H_0 : \text{rozkład badanej cechy jest normalny}$$
$$H_1 : \text{rozkład badanej cechy nie jest normalny}$$

Statystyka testowa **testu Shapiro-Wilka** dana jest wzorem

$$T = \frac{\left(\sum_{i=1}^{[\frac{n}{2}]} a_i(n)(X_{n-i+1:n} - X_{i:n})\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

gdzie $a_i(n)$ są pewnymi stałymi zależnymi od liczności próby, natomiast $[n/2]$ oznacza część całkowitą wyrażenia $n/2$.

Obszar krytyczny ma postać

$$W_\alpha = (0, w(\alpha, n)],$$

gdzie $w(\alpha, n)$ oznacza kwantyl rzędu α rozkładu statystyki.

Materiały na podstawie

- Grzegorzewski P., Bobecka K., Dembińska A., Pusz J., Rachunek prawdopodobieństwa i statystyka, WSISiZ, Warszawa, wyd. V - 2008.
- J. Koronacki, J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, Wydawnictwa Naukowo-Techniczne, Warszawa 2001.