Podstawy biostatystyki

Podstawy wnioskowania statystycznego

Wojciech Zieliński

http:\\wojtek.zielinski.statystyka.info

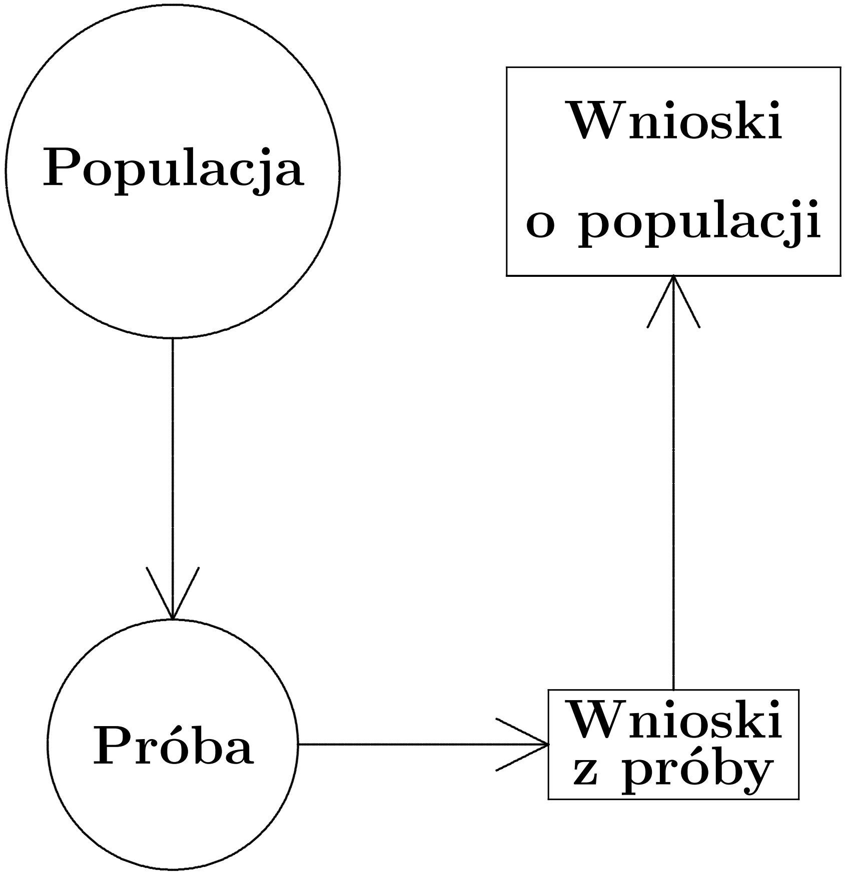
http:\\biostatystykanzc.wum.edu.pl

STATYSTYKA: nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych oraz przedstawianiu wyników w postaci zestawień tabelarycznych, wykresów, itp.; posługuje się rachunkiem prawdopodobieństwa.

STATYSTYKA MATEMATYCZNA: dział matematyki stosowanej oparty na rachunku prawdopodobieństwa; zajmuje się badaniem zbiorów na podstawie znajomości własności ich części.

Encyklopedia Popularna PWN, Warszawa 1982

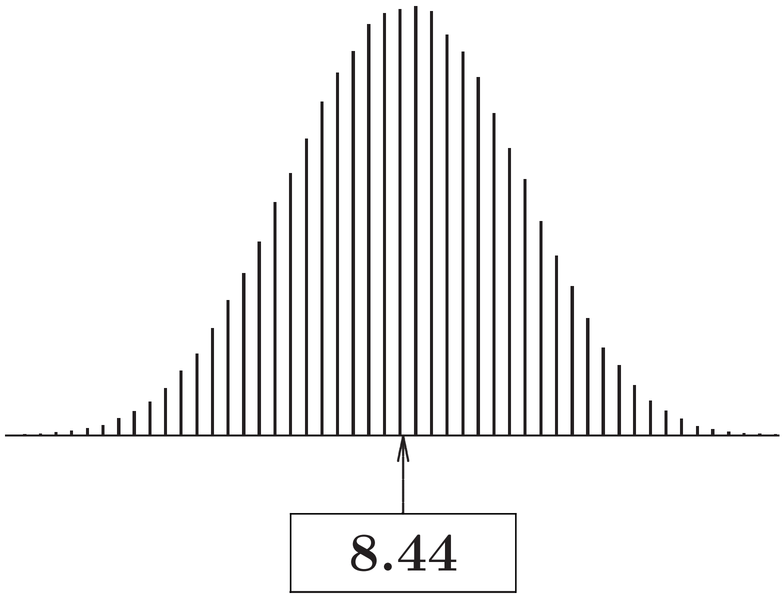
BIOSTATYSTYKA (**biometria**): nauka z pogranicza biologii i statystyki, adaptacja metod statystycznych na potrzeby prac badawczych w dziedzinie biologii, związanych przede wszystkim z medycyną, genetyką, fizjologią, antropologią, ekologią i rolnictwem.



F 5 27	F 2 29	F 4 12	F 1 8	F 5 33
M 8 40	M 8 52	M 6 33	F 6 38	F 8 22
M 9 35	M 10 73	F 7 30	M 11 50	M 9 67
M 14 68	M 12 75	F 8 40	M 14 64	M 11 69
F 9 54	F 8 40	M 11 51	M 10 55	M 15 66

Próba 1:	5	2	4	1	5	Średnia z próby:	3.40
Próba 2:	8	8	6	6	8	Średnia z próby:	7.20
Próba 3:	9	10	7	11	9	Średnia z próby:	9.20
Próba 4:	14	12	8	14	11	Średnia z próby:	11.80
Próba 5:	9	8	11	10	15	Średnia z próby:	10.60

Średnia populacji: 8.44



Pytania

Czy mając do dyspozycji tylko jedną próbę można ocenić na ile dobrze średnia z tej próby przybliża prawdziwą średnią?

Co zrobić, by być „pewniejszym” wyniku?

Populacja
Zbiór obiektów z wyróżnioną cechą (cechami)

Próba
Wybrana część populacji podlegająca badaniu

Cecha
Wielkość losowa charakteryzująca obiekty danej populacji

Cecha jakościowa
Cecha przyjmująca wartości nie będące liczbami (np. *kolor*, *płeć*, *smakowitość*)

Cecha (ilościowa) skokowa
Cecha przyjmująca pewne wartości liczbowe i nie przyjmująca wartości pośrednich (np. *ilość bakterii*, *ilość pracowników*, *ilość pasażerów*). Cechy te nazywane są również **dyskretnymi**.

Cecha (ilościowa) ciągła
Cecha przyjmująca wartości z pewnego przedziału liczbowego (np. *wzrost*, *waga*, *plon*)

Jakość wnioskowania statystycznego

Oceniamy parametr θ cechy na podstawie próby $X_1, X_2, \ldots, X_n$. Niech $\hat{\theta}(X_1, X_2, \ldots, X_n)$ będzie „jakąś" oceną parametru θ

Nieobciążoność
Jeżeli średnia wartość oceny $\hat{\theta}$ jest równa wartości parametru θ, to ocenę $\hat{\theta}$ nazywamy nieobciążoną

Minimalna wariancja

Z dwóch różnych nieobciążonych ocen $\hat{\theta}$ oraz $\hat{\hat{\theta}}$ tego samego parametru θ za lepszą uznajemy tę, która „średnio" przyjmuje wartości bliższe parametrowi θ

Minimalny błąd średniokwadratowy
Jeżeli ocena $\hat{\theta}$ nie jest nieobciążona, to wówczas jako miernik jakości stosuje się błąd średniokwadratowy. Jest to „uśrednienie" obciążenia oraz wariancji

Rozkład dwupunktowy

Zmienna losowa X ma rozkład $D(p)$, jeżeli

$$P\{X = 1\} = p = 1 - P\{X = 0\}$$

$$EX = p \qquad D^2X = p(1-p)$$

Doświadczenie Bernoulliego
Wykonujemy dwuwynikowe doświadczenie. Wyniki nazywane są umownie *sukces* oraz *porażka*. Prawdopodobieństwo sukcesu wynosi p (porażki: $1-p$). Niech zmienną losową X będzie uzyskanie sukcesu.

Zmienna losowa X ma rozkład $D(p)$.

Przykłady.
Płeć osoby.
Wadliwość produktu.

Rozkład dwumianowy

Zmienna losowa X ma rozkład $B(n, p)$, jeżeli

$$P_{n,p}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \; k = 0, 1, \ldots, n.$$

$$EX = np \qquad D^2X = np(1-p)$$

Schemat Bernoulliego
Zmienną losową o rozkładzie $D(p)$ obserwujemy n krotnie w sposób niezależny. Niech zmienną losową X będzie ilość sukcesów.

Zmienna losowa X ma rozkład $B(n, p)$.

Przykłady.
Ilość nasion, z których wzeszły rośliny.
Ilość wadliwych produktów.
„Popularność” danej osobistości publicznej.

Rozkład normalny

Zmienna losowa X ma rozkład normalny $N(\mu, \sigma^2)$ o wartości średniej μ i wariancji σ^2, jeżeli jej funkcja gęstości wyraża się wzorem

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \; -\infty < x < \infty.$$

$$EX = \mu \qquad D^2X = \sigma^2.$$

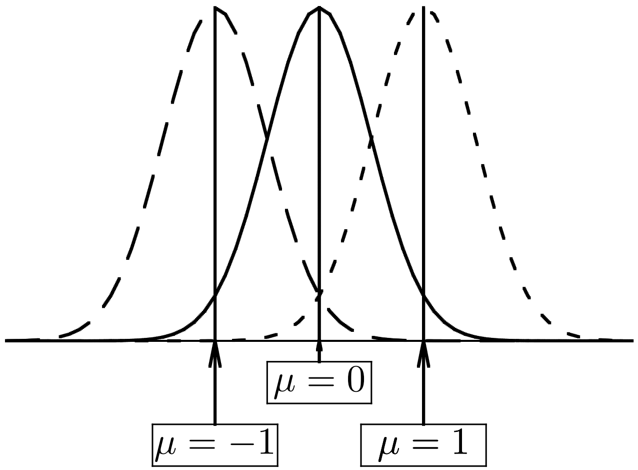
Przykłady.
Błędy pomiarowe.
Ciężar ciała.
Zawartość białka w mięsie.

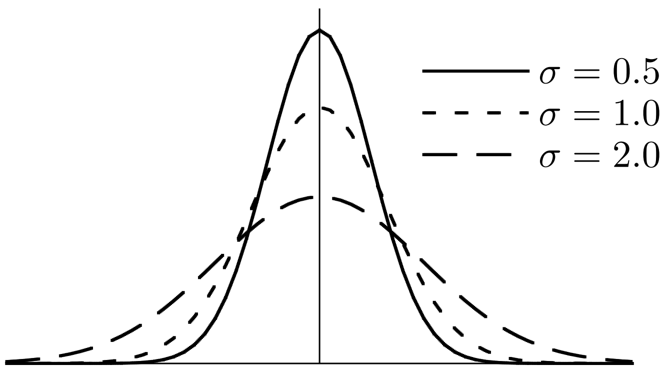
Standardowy rozkład normalny: $N(0, 1)$

Dystrybuanta $F(x)$ standardowego rozkładu normalnego ($N(0, 1)$) jest stablicowana.

$$F(x) = 1 - F(-x)$$

Rozkład normalny



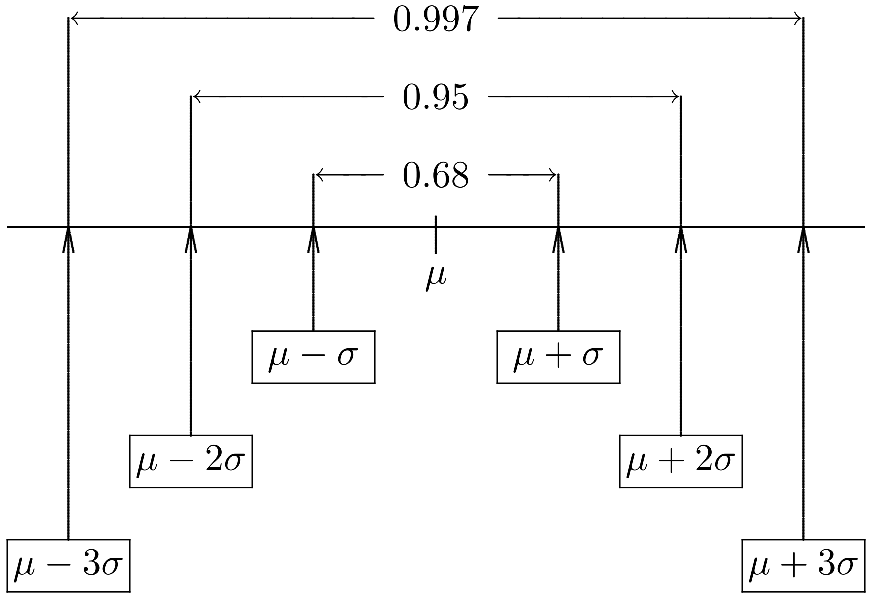


Prawo trzech sigm

$$P\{|X - \mu| < \sigma\} = 0.68268 \approx 0.68$$

$$P\{|X - \mu| < 2\sigma\} = 0.95450 \approx 0.95$$

$$P\{|X - \mu| < 3\sigma\} = 0.99730 \approx 0.997$$



Estymacja parametrów rozkładu cechy

Estymujemy parametr θ rozkładu cechy X
Próba: $X_1, X_2, \ldots, X_n$

Estymator (punktowy) jest funkcją próby

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

przybliżającą wartość parametru θ

Przedział ufności (estymator przedziałowy) jest przedziałem o końcach zależnych od próby, który z pewnym z góry zadanym prawdopodobieństwem pokrywa nieznaną wartość parametru θ

$$P\{\theta \in (\underline{\theta}(X_1, \ldots, X_n),\ \overline{\theta}(X_1, \ldots, X_n))\} = 1 - \alpha$$

Poziom ufności: prawdopodobieństwo $1 - \alpha$

Co wpływa na długość d przedziału ufności?
1. Liczność próby ($n \nearrow \Longrightarrow d \searrow$)
2. Poziom ufności ($1 - \alpha \nearrow \Longrightarrow d \nearrow$)
3. Wariancja cechy ($\sigma^2 \searrow \Longrightarrow d \searrow$)

Rozkład normalny Estymacja parametrów

Próba (prosta): $X_1, X_2, \ldots, X_n$

Estymator średniej μ — średnia arytmetyczna

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

Estymator wariancji σ^2 — wariancja próbkowa

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

Suma kwadratów odchyleń od średniej

$$\mathrm{var}X = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Estymator odchylenia standardowego σ

$$S = \sqrt{S^2}$$

Przedział ufności dla średniej

Wariancja σ^2 jest nieznana

Poziom ufności: $1-\alpha$

$$\left(\bar{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}}\ , \quad \bar{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}}\right)$$

$t(\alpha; n-1)$: wartość krytyczna rozkładu t (Studenta) z ν stopniami swobody

Długość przedziału: $d = 2t(\alpha; n-1)\dfrac{S}{\sqrt{n}}$

Przedziały jednostronne

$$(-\infty, \quad \bar{X} + t(2\alpha; n-1)\frac{S}{\sqrt{n}})$$

$$(\bar{X} - t(2\alpha; n-1)\frac{S}{\sqrt{n}}, \quad \infty)$$

Przykład.

Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować wartość średnią rozkładu obserwowanej cechy.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\mathrm{var}X = (1.1 - 1.0)^2 + \cdots + (1.0 - 1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10-1} = 0.04, \qquad s = \sqrt{s^2} = 0.2$$

Poziom ufności $1-\alpha = 0.95$, czyli $\alpha = 0.05$.

$$t(0.05; 9) = 2.2622$$

$$t(0.05; 9)\frac{s}{\sqrt{n}} = 2.2622\frac{0.2}{\sqrt{10}} = 0.14$$

$$(1 - 0.14, 1 + 0.14) = (0.86, 1.14)$$

Wniosek. Średnia wartość cechy jest jakąś liczbą z przedziału $(0.86, 1.14)$. Zaufanie do tego wniosku wynosi 95%.

Przykład.
Oszacować przeciętną ilość punktów uzyskiwanych na klasówce.

$$n = 300 \quad \sum x_i = 176.566 \quad \sum x_i^2 = 107.845302$$

Populacja:
Słuchacze podstawowego kursu statystyki

Cecha X:
ilość punktów zdobytych na klasówce

Założenie:
cecha X ma rozkład normalny $N(\mu, \sigma^2)$

Zadanie: oszacować parametr μ

Technika statystyczna:
przedział ufności dla średniej
poziom ufności $1 - \alpha = 0.95$

Obliczenia

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{176.566}{300} = 0.589$$

$$\mathrm{var}X = \sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2$$

$$= 107.845302 - \frac{176.566^2}{300} = 3.92679$$

$$s^2 = \frac{3.92679}{300 - 1} = 0.01313, \qquad s = \sqrt{s^2} = 0.11460$$

$$t(0.05; 299) \approx 1.96$$

$$t(0.05; 299) \frac{s}{\sqrt{n}} = 1.96 \frac{0.11460}{\sqrt{300}} = 0.01297$$

$$(0.589 - 0.013, 0.589 + 0.013) = (0.576, 0.602)$$

Odpowiedź: $\mu \in (0.576, 0.602)$

Wniosek. Przeciętna liczba punktów zdobywana na klasówce jest liczbą z przedziału $(0.576, 0.602)$. Zaufanie do tego wniosku wynosi 95%.

Przedział ufności dla wariancji

Średnia μ jest nieznana

Poziom ufności: $1-\alpha$

$$\left(\frac{\mathrm{var}X}{\chi^2\left(\frac{\alpha}{2}; n-1\right)}\,, \quad \frac{\mathrm{var}X}{\chi^2\left(1-\frac{\alpha}{2}; n-1\right)}\right)$$

$\chi^2(\alpha; n-1)$ jest stablicowaną wartością krytyczną rozkładu chi–kwadrat z ν stopniami swobody.

Przedziały jednostronne

$$\left(0, \quad \frac{\mathrm{var}X}{\chi^2(\alpha; n-1)}\right)$$

$$\left(\frac{\mathrm{var}X}{\chi^2(1-\alpha; n-1)}, \quad \infty\right)$$

Przykład.

Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować zróżnicowanie rozkładu obserwowanej cechy.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\mathrm{var}X = (1.1-1.0)^2 + \cdots + (1.0-1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10-1} = 0.04, \qquad s = \sqrt{s^2} = 0.2$$

Poziom ufności $1-\alpha = 0.95$, czyli $\alpha = 0.05$.

$$\chi^2\left(\frac{\alpha}{2}; n-1\right) = \chi^2(0.025; 9) = 19.0228$$

$$\chi^2\left(1-\frac{\alpha}{2}; n-1\right) = \chi^2(0.975; 9) = 2.7004$$

$$\left(\frac{0.36}{19.0228}, \frac{0.36}{2.7004}\right) = (0.019, 0.133)$$

Wniosek. Wariancja cechy jest jakąś liczbą z przedziału $(0.019, 0.133)$. Zaufanie do tego wniosku wynosi 95%.

Przedział ufności dla odchylenia standardowego

Średnia μ jest nieznana

Poziom ufności: $1-\alpha$

$$\left(\sqrt{\frac{\mathrm{var}X}{\chi^2(\frac{\alpha}{2};n-1)}}\ ,\quad \sqrt{\frac{\mathrm{var}X}{\chi^2(1-\frac{\alpha}{2};n-1)}}\right)$$

Przedziały jednostronne

$$\left(0,\sqrt{\frac{\mathrm{var}X}{\chi^2\left(\alpha;n-1\right)}}\right)$$

$$\left(\sqrt{\frac{\mathrm{var}X}{\chi^2\left(1-\alpha;n-1\right)}},\infty\right)$$

....

Przykład (cd).

Przedział ufności dla odchylenia standardowego:

$$(\sqrt{0.019},\sqrt{0.133})=(0.136,0.365)$$

Przykład.

Oszacować zróżnicowanie ilości punktów uzyskiwanych na klasówce.

$$n=300 \qquad \sum x_i=176.566 \qquad \sum x_i^2=107.845302$$

Populacja:
Słuchacze podstawowego kursu statystyki

Cecha X:
ilość punktów zdobytych na klasówce

Założenie:
cecha X ma rozkład normalny $N(\mu,\sigma^2)$

Zadanie: oszacować parametr σ

Technika statystyczna:
przedział ufności dla odchylenia standardowego
poziom ufności 0.95

Obliczenia

$$\bar{x} = 0.589 \qquad \mathrm{var}X = 3.92679$$

$$\chi^2\left(\frac{\alpha}{2}; n-1\right) = \chi^2(0.025; 299) = 348.79420$$

$$\chi^2\left(1-\frac{\alpha}{2}; n-1\right) = \chi^2(0.975; 299) = 252.99251$$

$$\left(\sqrt{\frac{3.92679}{348.79420}}, \sqrt{\frac{3.92679}{252.99251}}\right) = (0.10610, 0.12458)$$

Odpowiedź: $\sigma \in (0.10610, 0.12458)$

Wniosek. Odchylenie standardowe liczby punktów zdobywanych na klasówce jest liczbą z przedziału $(0.106, 0.125)$. Zaufanie do tego wniosku wynosi 95%.

Rozkład dwupunktowy Estymacja parametru

p — frakcja, wskaźnik struktury

Próba: $X_1, \ldots, X_n$ ($X_i = 0$ lub $= 1$)

$k = \sum_{i=1}^n X_i$ — ilość jedynek (sukcesów)

Estymator punktowy:

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

Przedział ufności na poziomie ufności $1-\alpha$

$$\left(p_1\left(1-\frac{\alpha}{2}; k, n-k\right), \quad 1-p_1\left(1-\frac{\alpha}{2}; n-k, k\right)\right)$$

Jednostronne przedziały ufności

$$\left(p_1\left(1-\alpha; k, n-k\right), 1\right)$$

$$\left(0, 1-p_1\left(1-\alpha; n-k, k\right)\right)$$

Przykład.
Wśród 20 zbadanych detali znaleziono dwa braki. Ocenić na tej podstawie wadliwość produkcji.

Cecha X — jakość detalu (dobry, zły).
Sukces — detal wybrakowany
Pytanie: $p =$?

$$n = 20,\ k = 2 \Longrightarrow \hat{p} = 2/20 = 0.1$$

Poziom ufności $1 - \alpha = 0.9$, czyli $\alpha = 0.1$

$$p_1\left(1 - \frac{\alpha}{2}; k, n-k\right) = p_1(0.95; 2, 18) = 0.0123$$

$$p_1\left(1 - \frac{\alpha}{2}; n-k, k\right) = p_1(0.95; 18, 2) = 0.6830$$

$$(0.0123, 1 - 0.6830) = (0.0123, 0.3170)$$

Wniosek. Wadliwość produkcji wyraża się liczbą z przedziału $(1.23\%, 31.70\%)$. Zaufanie do wniosku wynosi 90%.

Przybliżony przedział ufności

$$\left(\hat{p} - u_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + u_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

u_α jest kwantylem rzędu α rozkładu $N(0,1)$.

. . . .

Przykład. (cd)

$$n = 200,\ k = 20 \Longrightarrow \hat{p} = 20/200 = 0.1$$

Poziom ufności $1 - \alpha = 0.9$, czyli $\alpha = 0.1$

$$u_{1-\alpha/2} = u_{0.95} = 1.6449$$

$$0.1 - 1.6449\sqrt{\frac{0.1(1-0.1)}{200}} = 0.0651$$

$$0.1 + 1.6449\sqrt{\frac{0.1(1-0.1)}{200}} = 0.1349$$

Wniosek. Wadliwość produkcji wyraża się liczbą z przedziału $(6.51\%, 13.49\%)$. Zaufanie do wniosku wynosi 90%.

Przykład.
Oszacować odsetek ocen dostatecznych otrzymywanych na klasówce.

$$n = 300 \quad k = 88$$

Populacja:
Słuchacze podstawowego kursu statystyki

Cecha X:
ocena dostateczna z klasówki

Założenie:
cecha X ma rozkład $D(p)$

Zadanie: oszacować parametr p

Technika statystyczna:
przybliżony przedział ufności dla prawdopodobieństwa
poziom ufności 0.95

Obliczenia

$$\widehat{p} = \frac{88}{300} = 0.29$$

$$u_{1-\alpha/2} = u_{0.975} = 1.96$$

$$0.29 - 1.96\sqrt{\frac{0.29(1-0.29)}{300}} = 0.2387$$

$$0.29 + 1.96\sqrt{\frac{0.29(1-0.29)}{300}} = 0.3413$$

Odpowiedź: $p \in (0.2387, 0.3413)$

Wniosek.
Odsetek ocen dostatecznych zdobywanych na klasówce jest liczbą z przedziału $(23.87\%, 34.13\%)$. Zaufanie do tego wniosku wynosi 95%.

Porównanie dwóch rozkładów normalnych

Założenia:
1. $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$
2. X_1, X_2 są niezależne

Ocena $\mu_1 - \mu_2$ oraz σ_1^2/σ_2^2.

Próby: $X_{11}, \ldots, X_{1n_1}$; $X_{21}, \ldots, X_{2n_2}$

$$\bar{X}_1, \quad \mathrm{var}X_1, \quad s_1^2 = \frac{\mathrm{var}X_1}{n_1 - 1}$$

$$\bar{X}_2, \quad \mathrm{var}X_2, \quad s_2^2 = \frac{\mathrm{var}X_2}{n_2 - 1}$$

Ocena różnicy między średnimi $\mu_1 - \mu_2$

Ocena punktowa: $\bar{X}_1 - \bar{X}_2$

Przedział ufności (poziom ufności $1 - \alpha$)

1. Założenie $\sigma_1^2 = \sigma_2^2$

$$(\bar{X}_1 - \bar{X}_2 - t(\alpha; n_1 + n_2 - 2)s_r, \quad \bar{X}_1 - \bar{X}_2 + t(\alpha; n_1 + n_2 - 2)s_r)$$

$$s_e^2 = \frac{\mathrm{var}X_1 + \mathrm{var}X_2}{n_1 + n_2 - 2}, \quad s_r^2 = s_e^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

2. Bez założenia $\sigma_1^2 = \sigma_2^2$

$$(\bar{X}_1 - \bar{X}_2 - V(\alpha; n_1 - 1, n_2 - 1, c)s_r, \quad \bar{X}_1 - \bar{X}_2 + V(\alpha; n_1 - 1, n_2 - 1, c)s_r)$$

$$s_r^2 = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right) \qquad c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}$$

$V(\alpha; n_1 - 1, n_2 - 1, c)$ — wartość krytyczna testu Behrensa–Fishera

Przykład. Ocenić różnicę między średnimi wynikami klasówki pań i panów.

Panowie:

$n_1 = 138, \quad \sum x_{1i} = 82.833, \quad \mathrm{var} x_1 = 1.65841$

Panie:

$n_2 = 162, \quad \sum x_{2i} = 93.733, \quad \mathrm{var} x_2 = 2.23348$

Populacja 1:
Słuchacze podstawowego kursu statystyki
Populacja 2:
Słuchaczki podstawowego kursu statystyki

Cecha X: ilość punktów zdobytych na klasówce

Założenie:
cecha X ma w populacji 1 rozkład $N(\mu_1, \sigma_1^2)$
cecha X ma w populacji 2 rozkład $N(\mu_2, \sigma_2^2)$
$\sigma_1^2 = \sigma_2^2$

Zadanie: oszacować różnicę $\mu_1 - \mu_2$

Technika statystyczna:
przedział ufności t dla różnicy średnich
poziom ufności 0.95

Obliczenia

$$\bar{x}_1 = 0.60024, \quad \bar{x}_2 = 0.57860,$$

$$s_r^2 = \frac{1.65841 + 2.23348}{138 + 162 - 2} \left(\frac{1}{138} + \frac{1}{162} \right) = 0.000175255$$

$$t(0.05; 298) \approx 1.96; \quad t(0.05; 298) s_r = 0.02595$$

$$(0.60024 - 0.57860 \pm 0.00034) = (-0.00431, 0.04759)$$

Odpowiedź: $\mu_1 - \mu_2 \in (-0.00431, 0.04759)$

Wniosek.
Różnica średnich ilości punktów zdobywanych na klasówce przez panie i panów jest liczbą z przedziału $(-0.00431, 0.04759)$. Zaufanie do tego wniosku wynosi 95%.

Sugestia. Ponieważ przedział „obejmuje" zero, więc można uznać, że $\mu_1 = \mu_2$.

Ocena ilorazu wariancji σ_1^2/σ_2^2

Ocena punktowa: S_1^2/S_2^2

Przedział ufności (poziom ufności $1-\alpha$)

$$\left(\frac{S_1^2}{S_2^2}\cdot F\left(1-\frac{\alpha}{2};n_1-1,n_2-1\right),\right.$$
$$\left.\frac{S_1^2}{S_2^2}\cdot F\left(\frac{\alpha}{2};n_1-1,n_2-1\right)\right)$$

$F(\alpha;u,v)$ jest stablicowaną wartością krytyczną rozkładu F–Snedecora (Fishera–Snedecora)

$$F(1-\alpha;u,v)=\frac{1}{F(\alpha;v,u)}$$

Przykład. Porównać zróżnicowanie ocen wyników klasówek pań i panów.
Panowie:
$n_1=138, \quad \sum x_{1i}=82.833, \quad \mathrm{var}x_1=1.65841$

Panie:
$n_2=162, \quad \sum x_{2i}=93.733, \quad \mathrm{var}x_2=2.23348$

Populacja 1:
Słuchacze podstawowego kursu statystyki
Populacja 2:
Słuchaczki podstawowego kursu statystyki

Cecha X: ilość punktów zdobytych na klasówce

Założenie:
cecha X ma w populacji 1 rozkład $N(\mu_1,\sigma_1^2)$
cecha X ma w populacji 2 rozkład $N(\mu_2,\sigma_2^2)$

Zadanie: oszacować iloraz σ_1^2/σ_2^2

Technika statystyczna:
przedział ufności dla ilorazu wariancji
poziom ufności 0.90

Obliczenia

$$s_1^2 = \frac{1.65841}{138-1} = 0.01211, \quad s_2^2 = \frac{2.23348}{162-1} = 0.01387,$$

$$F(0.05; 137, 161) = 1.30936$$

$$F(0.95; 137, 161) = \frac{1}{F(0.05; 161, 137)}$$

$$= \frac{1}{1.31386} = 0.76111$$

$$\left(\frac{0.01211}{0.01387} \cdot 0.76111, \frac{0.01211}{0.0138} \cdot 1.30936\right)$$

$$= (0.66415, 1.14255)$$

Odpowiedź: $\sigma_1^2/\sigma_2^2 \in (0.66415, 1.14255)$

Wniosek.
Iloraz wariancji ilości punktów zdobywanych na klasówce jest liczbą z przedziału $(0.66415, 1.14255)$. Zaufanie do tego wniosku wynosi 90%.

Sugestia. Ponieważ przedział „obejmuje” jedynkę, więc można uznać, że $\sigma_1^2 = \sigma_2^2$.

Porównanie dwóch rozkładów dwupunktowych

Założenia:
1. $X_1 \sim D(p_1), X_2 \sim D(p_2)$
2. X_1, X_2 są niezależne

Ocena $p_1 - p_2$.

Próby: $X_{11}, \ldots, X_{1n_1}; X_{21}, \ldots, X_{2n_2}$ ($X_{ij} = 0$ lub 1)

$$k_1 = \sum_{i=1}^{n_1} X_{1i} \qquad k_2 = \sum_{i=1}^{n_2} X_{12}$$

$$\hat{p}_1 = k_1/n_1 \quad \hat{p}_2 = k_2/n_2 \quad \hat{p} = (k_1+k_2)/(n_1+n_2)$$

Przedział ufności (poziom ufności $1-\alpha$)

$$\left(\hat{p}_1 - \hat{p}_2 - u_{1-\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},\right.$$

$$\left.\hat{p}_1 - \hat{p}_2 + u_{1-\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Przykład.
Oszacować różnicę między „niezaliczalnością” klasówki ze statystyki przez panie i panów. Na podstawie dotychczasowych danych wiadomo, że na 162 pań nie zaliczyło klasówki 46 pań oraz na 138 panów 30 uzyskało ocenę negatywną.

Populacja 1:
Słuchacze podstawowego kursu statystyki
Populacja 2:
Słuchaczki podstawowego kursu statystyki

Cecha X: uzyskanie z klasówki oceny negatywnej

Założenie:
cecha X ma w populacji 1 rozkład $D(p_1)$
cecha X ma w populacji 2 rozkład $D(p_2)$

Zadanie: oszacować różnicę $p_1 - p_2$

Technika statystyczna:
przybliżony przedział ufności dla różnicy prawdopodobieństw
poziom ufności 0.95: $u_{0.975} = 1.96$

Obliczenia

$$n_1 = 162 \quad k_1 = 46 \qquad n_2 = 138 \quad k_2 = 30$$

$$\hat{p}_1 = \frac{k_1}{n_1} = \frac{46}{162} = 0.2840 \quad \hat{p}_2 = \frac{k_2}{n_2} = \frac{30}{138} = 0.2174$$

$$\hat{p} = \frac{(k_1 + k_2)}{(n_1 + n_2)} = \frac{(46 + 30)}{(162 + 138)} = 0.2533$$

$$1.96 \cdot \sqrt{\frac{0.2533(1 - 0.2533)}{300}} \left(\frac{1}{162} + \frac{1}{138} \right) = 0.0987$$

$$(0.2840 - 0.2174 - 0.0987, 0.2840 - 0.2174 + 0.0987)$$

$$(-0.0321,\ 0.1653)$$

Wniosek. Różnica prawdopodobieństw jest liczbą z przedziału $(-0.0321,\ 0.1653)$.

Sugestia. Ponieważ przedział „obejmuje” zero, więc odsetki pań i panów niezaliczających klasówki można traktować jako porównywalne.

Weryfikacja hipotez statystycznych

Hipotezą statystyczną nazywamy dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa cechy w populacji.
Oznaczenie H_0

Testem hipotezy statystycznej nazywamy postępowanie mające na celu odrzucenie lub nie odrzucenie hipotezy statystycznej.

Statystyką testową nazywamy funkcję próby na podstawie której wnioskuje się o odrzuceniu lub nie hipotezy statystycznej.

Rzeczywistość: hipoteza H_0	Wniosek o hipotezie H_0: nie odrzucać	Wniosek o hipotezie H_0: odrzucić
prawdziwa	prawidłowy	nieprawidłowy
nieprawdziwa	nieprawidłowy	prawidłowy

Błędem I rodzaju nazywamy błąd wnioskowania polegający na odrzuceniu hipotezy, gdy w rzeczywistości jest ona prawdziwa.

Błędem II rodzaju nazywamy błąd wnioskowania polegający na nieodrzuceniu hipotezy, gdy w rzeczywistości jest ona fałszywa.

Poziomem istotności nazywamy dowolną liczbę z przedziału $(0, 1)$ określającą prawdopodobieństwo popełnienia błędu I rodzaju.
Oznaczenie: α

Mocą testu nazywamy prawdopodobieństwo odrzucenia testowanej hipotezy, gdy jest ona nieprawdziwa, czyli prawdopodobieństwo nie popełnienia błędu II rodzaju.
Oznaczenie: $1 - \beta$

Rozkład normalny
Porównanie z normą

$$H_0 : \mu = \mu_0$$

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$
Średnia μ oraz wariancja σ^2 są nieznane

Test Studenta (poziom istotności α)

Próba: $X_1, \ldots, X_n$
Statystyka testowa

$$t_{\text{emp}} = \frac{\bar{X} - \mu_0}{S} \sqrt{n}\ .$$

Wartość krytyczna $t(\alpha; n-1)$

Jeżeli $|t_{\text{emp}}| > t(\alpha; n-1)$, to hipotezę $H_0 : \mu = \mu_0$ odrzucamy.

Przykład. Przypuszczenie: maszyna pakująca kostki masła nastawiona na jednostkową masę 250 g uległa po pewnym czasie rozregulowaniu. W celu weryfikacji tego przypuszczenia z bieżącej produkcji pobrano próbę otrzymując wyniki 254, 269, 254, 248, 263, 256, 258, 261, 264, 258. Czy można na tej podstawie sądzić, że maszyna uległa rozregulowaniu?

Populacja:
paczkowane kostki masła

Cecha X:
masa kostki masła

Założenie:
cecha X ma rozkład normalny $N(\mu, \sigma^2)$

Formalizacja:
Rozregulowanie maszyny może być interpretowane jako odejście od nominalnej wagi. Zatem należy zbadać, czy średnia μ wynosi 250, czyli weryfikujemy hipotezę $H_0 : \mu = 250$

Technika statystyczna:
test Studenta (test t)
poziom istotności $\alpha = 0.05$

Obliczenia

$$\bar{x} = 258.5,\ s^2 = 36.05,\ t_{\mathrm{emp}} = 4.47$$

Wartość krytyczna: $t(0.05; 9) = 2.2622$

Odpowiedź: hipotezę odrzucamy

Wniosek: maszyna uległa rozregulowaniu

Moc testu

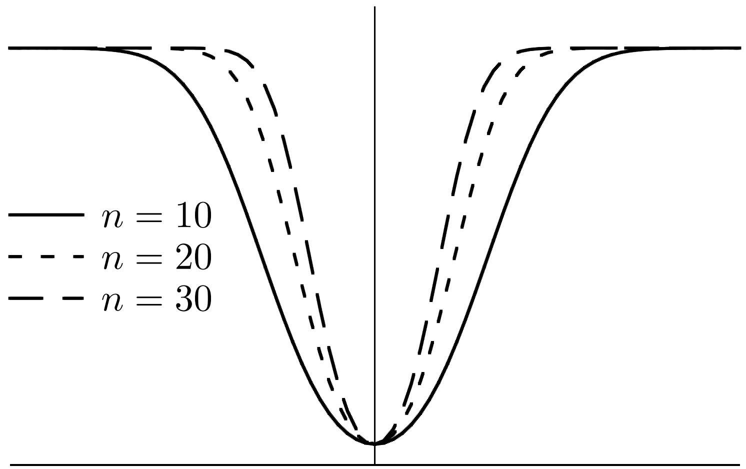
$$\text{Moc testu} = 1 - P\{\text{błąd II rodzaju}\}$$

$$\text{Moc testu} = P\{\text{odrzucenie nieprawdziwej } H_0\}$$

Moc testu Studenta hipotezy $H_0 : \mu = \mu_0$

$$\mathcal{M}(\mu) = P\{|t_{\mathrm{emp}}| > t(\alpha; n-1) | X \sim N(\mu, \sigma^2)\}$$

$$\mathcal{M}(\mu_0) = \alpha$$



Przedział ufności a test hipotezy $H_0 : \mu = \mu_0$

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

$$H_0 : \mu = \mu_0$$

H_0 nie odrzucamy na poziomie istotności α

$$\Updownarrow$$

$$|t_{\text{emp}}| < t(\alpha; n-1)$$

$$\Updownarrow$$

$$-t(\alpha; n-1) < \frac{\bar{X} - \mu_0}{S}\sqrt{n} < t(\alpha; n-1)$$

$$\Updownarrow$$

$$\mu_0 \in \left(\bar{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}}, \bar{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}}\right)$$

$$\Updownarrow$$

μ_0 należy do przedziału ufności

na poziomie ufności $1 - \alpha$

$$H_0 : \sigma^2 = \sigma_0^2$$

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$
Średnia μ oraz wariancja σ^2 są nieznane

Test chi–kwadrat (poziom istotności α)

Próba: $X_1, \ldots, X_n$
Statystyka testowa

$$\chi^2_{\text{emp}} = \frac{\text{var}X}{\sigma_0^2}$$

Wartości krytyczne
$\chi^2\left(1 - \frac{\alpha}{2}; n-1\right)$ oraz $\chi^2\left(\frac{\alpha}{2}; n-1\right)$

Jeżeli

$\chi^2_{\text{emp}} < \chi^2\left(1 - \frac{\alpha}{2}; n-1\right)$
lub
$\chi^2_{\text{emp}} > \chi^2\left(\frac{\alpha}{2}; n-1\right)$,

to hipotezę $H_0 : \sigma^2 = \sigma_0^2$ odrzucamy.

Przykład. Na podstawie obserwacji prowadzonych przez długi okres czasu stwierdzono, że dzienny udój uzyskiwany w pewnym stadzie krów jest wielkością losową, zaś przeciętny dzienny udój mleka wyraża sie liczbą z przedziału $(900, 1200)$. Rachunek finansowy pokazał, że produkcja mleka jest opłacalna, jeżeli całkowity dzienny udój będzie wynosił nie mniej niż $d = 700\ l$ mleka przez co najmniej 280 dni w roku. W jaki sposób można zbadać, czy produkcja mleka jest opłacalna?

Populacja:

Cecha:
całkowity dzienny udój

Założenia:
Cecha X ma rozkład $N(\mu, \sigma^2)$
$\mu_d = 900 \le \mu \le \mu_g = 1200$

Formalizacja problemu

$$P\{X \ge d\} \ge p = \frac{280}{350}$$

$$P\{X \ge d\} = 1 - F\left(\frac{d-\mu}{\sigma}\right) \ge 1 - F\left(\frac{d-\mu_d}{\sigma}\right)$$

$$1 - F\left(\frac{d-\mu_d}{\sigma}\right) \ge 1 - p \Rightarrow F\left(\frac{d-\mu_d}{\sigma}\right) \le 1 - p$$

$$\frac{d-\mu_d}{\sigma} \le F^{-1}(1-p) = u_{1-p}$$

d, μ_d oraz p są ustalone, więc

$$\sigma^2 \ge \sigma_0^2 = \left(\frac{d-\mu_d}{u_{1-p}}\right)^2 = 56472$$

Produkcja mleka jest opłacalna, jeżeli wariancja σ^2 dziennych udojów jest większa niż $\sigma_0^2 = 56472$.

$$H_0 : \sigma^2 \le 56472$$

Rozkład dwupunktowy
Porównanie z normą

$$H_0 : p = p_0$$

Cecha X ma rozkład $D(p)$
Próba: $X_1, \ldots, X_n$ ($X_i = 0$ lub $= 1$)
Test przybliżony (poziom istotności α)
Przypadek: n „duże”

Statystyka testowa

$$u_{\text{emp}} = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}}$$

Wartość krytyczna $u_{1-\alpha/2}$

Jeżeli $|u_{\text{emp}}| > u_{1-\alpha/2}$, to $H_0 : p = p_0$ odrzucamy

Przykład. W swojej ofercie sprzedaży stawu rybnego jego właściciel podaje, iż w stawie żyje co najmniej tysiąc karpi. Potencjalny nabywca zainteresowany jest sprawdzeniem prawdziwości tego twierdzenia. W tym celu wyłowiono sto karpi i po zaobrączkowaniu ich wpuszczono je z powrotem do stawu. Po jakimś czasie ponownie odłowiono sto ryb i stwierdzono, że wśród nich jest piętnaście zaobrączkowanych. Czy w świetle uzyskanych wyników można reklamę uznać za prawdziwą?

Populacja:
ryby w stawie

Cecha:
zaobrączkowanie ryby

Założenia:
Cecha X ma rozkład $D(p)$

Formalizacja problemu
Jeżeli w stawie żyje co najmniej N ryb, to odsetek zaobrączkowanych jest co najwyżej $100/N$. Zgodnie z twierdzeniem właściciela, $N \geq 1000$, czyli odsetek ryb zaobrączkowanych nie przekracza 0.1.

Technika statystyczna
Przybliżony test hipotezy $H_0 : p \leq 0.1$
Poziom istotności: $\alpha = 0.05$

Obliczenia

$$Y = 15 \quad n = 100$$

$$u_{\mathrm{emp}} = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} = \frac{15 - 10}{\sqrt{100 \cdot 0.1 \cdot 0.9}} = 1.6667$$

Wartość krytyczna: $u_{1-0.05} = 1.6449$

Odpowiedź: hipotezę odrzucamy

Wniosek: należy uznać, że ogólna liczb ryb w stawie jest mniejsza niż podana w ofercie

Porównanie dwóch rozkładów normalnych

Założenia:
1. $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$
2. X_1, X_2 są niezależne

Czy $\mu_1 = \mu_2$?

Czy $\sigma_1^2 = \sigma_2^2$?

Próby: $X_{11}, \ldots, X_{1n_1}; X_{21}, \ldots, X_{2n_2}$

$$\bar{X}_1, \quad \mathrm{var}X_1, \quad s_1^2 = \frac{\mathrm{var}X_1}{n_1 - 1}$$

$$\bar{X}_2, \quad \mathrm{var}X_2, \quad s_2^2 = \frac{\mathrm{var}X_2}{n_2 - 1}$$

$$H_0 : \mu_1 = \mu_2$$

Założenie $\sigma_1^2 = \sigma_2^2$

Test Studenta (poziom istotności α)

Statystyka testowa

$$t_{\text{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r}$$

$$S_r = \sqrt{S_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}, \quad S_e^2 = \frac{\text{var} X_1 + \text{var} X_2}{n_1 + n_2 - 2}$$

Wartość krytyczna $t(\alpha; n_1 + n_2 - 2)$

Jeżeli $|t_{\text{emp}}| > t(\alpha; n_1 + n_2 - 2)$,
to hipotezę $H_0 : \mu_1 = \mu_2$ odrzucamy

Przykład. Porównać przeciętne osiągnięcia punktowe pań i panów na klasówce ze statystyki
Panowie:
$n_1 = 138, \quad \sum x_{1i} = 82.833, \quad \text{var} x_1 = 1.65841$

Panie:
$n_2 = 162, \quad \sum x_{2i} = 93.733, \quad \text{var} x_2 = 2.23348$

Populacja 1:
Słuchacze podstawowego kursu statystyki
Populacja 2:
Słuchaczki podstawowego kursu statystyki

Cecha X: ilość punktów zdobytych na klasówce

Założenia:
cecha X ma w populacji 1 rozkład $N(\mu_1, \sigma_1^2)$
cecha X ma w populacji 2 rozkład $N(\mu_2, \sigma_2^2)$
$\sigma_1^2 = \sigma_2^2$

Zadanie: zweryfikować hipotezę $H_0 : \mu_1 = \mu_2$

Technika statystyczna:
test t
poziom istotności 0.05

Obliczenia

$$\bar{x}_1 = 0.60024 \quad \bar{x}_2 = 0.57860$$

$$s_r^2 = \frac{1.65841 + 2.23348}{138 + 162 - 2}\left(\frac{1}{138} + \frac{1}{162}\right)$$
$$= 0.000175255$$

$$t_{\mathrm{emp}} = \frac{0.60024 - 0.57860}{\sqrt{0.000175255}} = 1.634$$

Wartość krytyczna $t(0.05; 298) \approx 1.96$

Odpowiedź: hipotezy nie odrzucamy

Wniosek.
Średnie ilości punktów uzyskiwane przez panie i panów można traktować jako porównywalne.

Przedział ufności a test hipotezy $H_0 : \mu_1 = \mu_2$

Cecha $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, $\sigma_1^2 = \sigma_2^2$

$$H_0 : \mu_1 = \mu_2$$

H_0 nie odrzucamy na poziomie istotności α

$$\Updownarrow$$

$$|t_{\mathrm{emp}}| < t(\alpha; n_1 + n_2 - 2)$$

$$\Updownarrow$$

$$-t(\alpha; n_1 + n_2 - 2) < \frac{\bar{X}_1 - \bar{X}_2}{S_r} < t(\alpha; n_1 + n_2 - 2)$$

$$\Updownarrow$$

$$0 \in \left(\bar{X}_1 - \bar{X}_2 \pm t(\alpha; n_1 + n_2 - 2)S_r\right)$$

$$\Updownarrow$$

0 należy do przedziału ufności

na poziomie ufności $1 - \alpha$

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Test F (poziom istotności α)

Statystyka testowa

$$F_{\text{emp}} = \frac{S_1^2}{S_2^2}$$

Wartości krytyczne
$F\left(1 - \dfrac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$
$F\left(\dfrac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$

Jeżeli
$F_{\text{emp}} < F\left(1 - \dfrac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$
lub
$F_{\text{emp}} > F\left(\dfrac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$
to hipotezę $H_0 : \sigma_1^2 = \sigma_2^2$ odrzucamy

Uwaga

$$F(1 - \alpha; u, v) = \frac{1}{F(\alpha; v, u)}$$

Reguła: większa wariancja do licznika.

Jeżeli $S_1^2 > S_2^2$, to wyznaczana jest statystyka

$$F_{\text{emp}} = \frac{S_1^2}{S_2^2}$$

i hipoteza jest odrzucana, gdy

$$F_{\text{emp}} > F\left(\frac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$$

Jeżeli zaś $S_1^2 < S_2^2$, to wyznaczana jest statystyka

$$F_{\text{emp}} = \frac{S_2^2}{S_1^2}$$

i hipoteza jest odrzucana, gdy

$$F_{\text{emp}} > F\left(\frac{\alpha}{2}; n_2 - 1, n_1 - 1\right)$$

Przykład. Dla sprawdzenia stabilności pracy maszyny pobrano dwie próbki: pierwszą w początkowym okresie eksploatacji oraz drugą po miesięcznym okresie pracy tej maszyny. Wykonano pomiary wylosowanych produktów i otrzymano wyniki: $n_1 = 25$, $\bar{x}_1 = 3.24$, $s_1^2 = 0.1447$ oraz $n_2 = 19$, $\bar{x}_2 = 3.19$, $s_2^2 = 0.1521$. Zbadać na tej podstawie czy maszyna nie rozregulowała się w trakcie pracy.

Populacja 1
produkcja maszyny w początkowym okresie

Populacja 2
produkcja maszyny po miesiącu eksploatacji

Cecha X
pomiar produktu

Założenia
cecha X ma w populacji 1 rozkład $N(\mu_1, \sigma_1^2)$
cecha X ma w populacji 2 rozkład $N(\mu_2, \sigma_2^2)$

Formalizacja
Stabilność pracy maszyny może być mierzona podobieństwem wytwarzanych produktów: im własności produktów są do siebie bardziej zbliżone, tym bardziej stabilna jest praca maszyny. Podobieństwo takie jest wyrażane wariancją cechy. Zatem stabilność pracy można wyrazić liczbowo jako wariancję interesującej cechy produktu, a problem stabilności jako zagadnienie weryfikacji hipotezy $H_0 : \sigma_1^2 = \sigma_2^2$

Technika statystyczna
Test F (poziom istotności $\alpha = 0.10$)

Obliczenia

$$F_{\text{emp}} = \frac{s_2^2}{s_1^2} = 1.051$$

Wartość krytyczna $F\,(0.05; 19, 24) = 2.114$

Odpowiedź: hipotezy nie odrzucamy

Wniosek: można uznać że maszyna nie rozregulowała się w trakcie pracy

Porównanie dwóch rozkładów dwupunktowych

Założenia:
1. $X_1 \sim D(p_1), X_2 \sim D(p_2)$
2. X_1, X_2 są niezależne

$$H_0 : p_1 = p_2$$

Test przybliżony (poziom istotności α)

$$\hat{p}_1 = \frac{k_1}{n_1}, \quad \hat{p}_2 = \frac{k_2}{n_2}, \quad \hat{p} = \frac{(k_1 + k_2)}{(n_1 + n_2)}$$

Statystyka testowa

$$u_{\text{emp}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$|u_{\text{emp}}| \geq u_{1-\alpha/2} \Longrightarrow H_0 : p_1 = p_2$ odrzucamy

Przykład. Celem badania było porównanie przygotowania z matematyki kandydatów na studia będących absolwentami liceów oraz techników. W tym celu spośród kandydatów zdających matematykę wylosowano 400 absolwentów liceów oraz 600 absolwentów techników. W wylosowanej grupie stwierdzono, że 385 absolwentów liceów oraz 501 absolwentów techników rozwiązało test wstępny. Czy można na tej podstawie sądzić, że przygotowanie w obu grupach absolwentów jest jednakowe?

Populacja 1:
absolwenci liceów zdający egzamin wstępny
Populacja 2:
absolwenci techników zdający egzamin wstępny

Cecha X: umiejętność rozwiązania testu (tak/nie)

Założenia:
cecha X ma w populacji 1 rozkład $D(p_1)$
cecha X ma w populacji 2 rozkład $D(p_2)$

Formalizacja
Weryfikacja hipotezy $H_0 : p_1 = p_2$

Technika statystyczna
Test przybliżony (poziom istotności $\alpha = 0.05$)

Obliczenia

$$n_1 = 400 \quad k_1 = 385 \quad \hat{p}_1 = 385/400 = 0.9625$$

$$n_2 = 600 \quad k_2 = 501 \quad \hat{p}_2 = 501/600 = 0.8350$$

$$\hat{p} = (385 + 501)/(400 + 600) = 0.886$$

$$u_{\text{emp}} = \frac{0.9625 - 0.8350}{\sqrt{0.886(1 - 0.886)\left(\frac{1}{400} + \frac{1}{600}\right)}} = 6.215.$$

Wartość krytyczna $u_{0.975} = 1.96$

Odpowiedź: hipotezę $H_0 : p_1 = p_2$ odrzucamy

Wniosek:
przygotowanie absolwentów liceów i techników z matematyki nie jest takie same.

Porównanie wielu rozkładów normalnych

Założenia:
1. $X_i \sim N(\mu_i, \sigma_i^2), \quad i = 1, \ldots, k$
2. $X_1, \ldots, X_k$ są niezależne

Czy $\mu_1 = \cdots = \mu_k$?

Czy $\sigma_1^2 = \cdots = \sigma_k^2$?

Próby: $X_{i1}, \ldots, X_{in_i}, \quad i = 1, \ldots, k$

$$\bar{X}_i, \quad \text{var} X_i, \quad s_i^2 = \frac{\text{var} X_i}{n_i - 1}; \quad i = 1, \ldots, k$$

$$H_0 : \mu_1 = \cdots = \mu_k$$

Założenie $\sigma_1^2 = \cdots = \sigma_k^2$

Test F (poziom istotności α)

Statystyka testowa

$$F_{\mathrm{emp}} = \frac{S_a^2}{S_e^2}$$

$$S_a^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$$

$$S_e^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad \bar{X} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

$$N = \sum_{i=1}^{k} n_i$$

Jeżeli $F_{\mathrm{emp}} > F\,(\alpha; k-1, N-k)$,
to hipotezę $H_0 : \mu_1 = \cdots = \mu_k$ odrzucamy.

Wniosek praktyczny:
przynajmniej jedna ze średnich $\mu_1, \ldots, \mu_k$ jest inna od pozostałych

Model analizy wariancji

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

Błąd losowy $\varepsilon_{ij} \sim N(0, \sigma^2)$

Przykłady
Plenność kilku *odmian* pewnej rośliny uprawnej
Wydajność pracowników kilku *zakładów* pracy
Zarobki kilku *grup* społecznych

Czynnik: odmiana, zakład, grupa
Poziomy czynnika: badane odmiany, badane zakłady, badane grupy

Model analizy wariancji

$$X_{ij} = \mu + a_i + \varepsilon_{ij}$$

a_i — efekt i–tego poziomu czynnika: $\sum_{i=1}^{k} a_i = 0$

$$H_0 : a_1 = \cdots = a_k = 0, \qquad H_0 : \sum_{i=1}^{k} a_i^2 = 0$$

Tabela analizy wariancji

Źródło zmienności	Stopnie swobody	Sumy kwadratów	Średnie kwadraty	F_{emp}
Czynnik	$k-1$	$\text{var}A$	$S_a^2 = \frac{\text{var}A}{k-1}$	S_a^2/S_e^2
Błąd losowy	$N-k$	$\text{var}E$	$S_e^2 = \frac{varE}{N-k}$	
Ogółem	$N-1$	$\text{var}T$		

$$\text{var}A = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \text{var}E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

$$\text{var}T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

$$\text{var}A + \text{var}E = \text{var}T$$

Grupy jednorodne — podzbiory średnich, które można uznać za takie same

Procedury porównań wielokrotnych — postępowanie statystyczne zmierzające do podzielenia zbioru średnich na grupy jednorodne

Procedury: Tukeya, Scheffégo, Bonfferroniego, Duncana, Newmana–Kuelsa i inne.

Ogólna idea procedur porównań wielokrotnych ($n_1 = \cdots = n_k$)

NIR — najmniejsza istotna różnica

Jeżeli $|\bar{X}_i - \bar{X}_j| < NIR$, to uznajemy, że $\mu_i = \mu_j$.

Jeżeli
$|\bar{X}_i - \bar{X}_j| < NIR$
$|\bar{X}_i - \bar{X}_l| < NIR$
$|\bar{X}_l - \bar{X}_j| < NIR$,
to uznajemy, że $\mu_i = \mu_j = \mu_l$.

Badając w ten sposób wszystkie pary średnich próbkowych otrzymujemy podział zbioru średnich na grupy jednorodne.

Procedura Tukeya

Założenie: $n_1 = \cdots = n_k = n$

$$NIR = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{n}}$$

$t(\alpha; k, N-k)$ — wartość krytyczna studentyzowanego rozstępu

Przypadek nierównolicznych prób
Jedna z modyfikacji procedury Tukeya

$$NIR_{ij} = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Przykład. Przeprowadzić analizę porównawczą wyników punktowych klasówki w grupach studenckich.

Populacje
Możemy wyodrębnić dziesięć populacji indeksowanych numerami grup studenckich

Cecha X
Ilość punktów uzyskanych na klasówce

Założenia
cecha X ma w i–tej populacji rozkład $N(\mu_i, \sigma_i^2)$
$(i = 1, \ldots, 10)$
$\sigma_1^2 = \cdots = \sigma_{10}^2$

Formalizacja
weryfikacja hipotezy $H_0 : \mu_1 = \cdots = \mu_{10}$

Techniki statystyczna
- Jednoczynnikowa analiza wariancji
- Porównania szczegółowe

Poziom istotności 0.05

Obliczenia

i	n_i	$\sum x_i$	$\sum x_i^2$
1	30	18.230	11.375950
2	30	16.672	9.596790
3	30	14.292	7.087458
4	30	18.879	12.069655
5	30	18.200	11.355982
6	30	19.568	13.172884
7	30	16.522	9.420960
8	30	19.134	12.514874
9	30	18.548	11.945964
10	30	16.521	9.304785
	300	176.566	107.845302

i	n_i	$\bar{x}_i$	$n_i(\bar{x}_i - \bar{x})^2$	$\mathrm{var}x_i$
1	30	0.607667	0.010960	0.298187
2	30	0.555733	0.032315	0.331604
3	30	0.476400	0.377351	0.278749
4	30	0.629300	0.049809	0.189100
5	30	0.606667	0.009843	0.314649
6	30	0.652267	0.121782	0.409330
7	30	0.550733	0.042911	0.321744
8	30	0.637800	0.072757	0.311209
9	30	0.618267	0.026486	0.478354
10	30	0.550700	0.042986	0.206670
	N=300	$\bar{x}$=0.588553	var$A$=0.787199	var$E$=3.139595

$$\mathrm{var}T = 107.845302 - 176.566^2/300 = 3.926794$$

Tabela analizy wariancji

Źródło zmienności	Stopnie swobody	Sumy kwadratów	Średnie kwadraty	Femp
Grupa	9	0.787199	0.087467	8.079
Błąd losowy	290	3.139595	0.010826	
Ogółem	299	3.926794		

Wartość krytyczna

$$F\,(0.05; 9, 290) = 1.912$$

Odpowiedź:
hipotezę $H_0 : \mu_1 = \cdots = \mu_{10}$ odrzucamy

Wniosek:
przynajmniej jedna grupa uzyskała inną średnią liczbę punktów niż pozostałe

Wyznaczenie grup jednorodnych

Procedura Tukeya ($\alpha = 0.05$)

Wartość krytyczna: $t(0.05; 10, 290) = 4.474$

$$NIR = 4.474 \cdot \sqrt{0.010826} \cdot \sqrt{\frac{1}{30}} = 0.084990$$

i	$\bar{x}_i$				
3	0.476400	$*$			
10	0.550700	$*$	$*$		
7	0.550733	$*$	$*$		
2	0.555733	$*$	$*$	$*$	
5	0.606667		$*$	$*$	$*$
1	0.607667		$*$	$*$	$*$
9	0.618267		$*$	$*$	$*$
4	0.629300		$*$	$*$	$*$
8	0.637800			$*$	$*$
6	0.652267				$*$