Rachunek prawdopodobieństwa i statystyka

Statystyka matematyczna

Prof. UEK dr hab. Paweł Ulman

Statystyka matematyczna – zagadnienia wstępne

W rachunku prawdopodobieństwa zakłada się znajomość rozkładu prawdopodobieństwa zmiennej losowej

W statystyce matematycznej nie zakłada się pełnej znajomości rozkładu – poznaje się go w ramach tzw. wnioskowania statystycznego, czyli:

Estymacji parametrycznej (punktowej lub przedziałowej) lub nieparametrycznej.

Weryfikacji hipotez parametrycznych lub nieparametrycznych

Statystyka matematyczna – zagadnienia wstępne

Wnioskowanie statystyczne oparte jest na częściowej informacji.

Próba statystyczna – podzbiór populacji generalnej

Próba jest reprezentatywna, gdy jej struktura ze względu na badane cechy jest co najmniej zbliżona do struktury populacji generalnej, z której ona pochodzi.

Próba reprezentatywna: losowa i odpowiednio liczna.

Statystyka matematyczna – zagadnienia wstępne

Próba jest losowa, gdy dobór jednostek do próby został dokonany w drodze doboru losowego (losowania).

Schematy losowania:

1. losowanie niezależne i zależne
2. losowanie indywidualne i zespołowe
3. losowanie jednostopniowe i wielostopniowe
4. losowanie nieograniczone i ograniczone (tu warstwowe i systematyczne)

Losowanie proste: indywidualne, nieograniczone i niezależne

Statystyka matematyczna – zagadnienia wstępne

Próba prosta – ciąg niezależnych zmiennych losowych $X_1, \ldots, X_n$ o jednakowym rozkładzie takim jaki ma cecha X w populacji.

Model statystyczny to przestrzeń próby $\boldsymbol{\chi}$ wraz z rodziną rozkładów P.

Niech rozkład badanej cechy X zależy od nieznanego parametru $\boldsymbol{\theta}$, który będzie szacowany o n-elementową próbę prostą pobraną z populacji.

Każdą funkcję $\mathbf{g}(X_1, \ldots, X_n)$, będącą funkcją próby losowej $X_1, \ldots, X_n$ nazywamy statystyką

Statystyka matematyczna – zagadnienia wstępne

Rozkład statystyki zależy od:

- postaci funkcji **g,**
- rozkładu zmiennych losowych X_1, …, X_n,
- liczebności próby.

Rozkład statystyki $\widehat{\Theta}_n$ - rozkład z próby

- dokładny
- graniczny

Statystyka matematyczna – rozkłady statystyk z próby

1. Rozkład średniej arytmetycznej z próby dla zmiennej X o rozkładzie normalnym

X~N(μ, σ) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

$$D^2(\bar{X}) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Statystyka matematyczna – rozkłady statystyk z próby

2. Rozkład średniej arytmetycznej z próby dla zmiennej X o rozkładzie normalnym z nieznanym odchyleniem standardowym σ

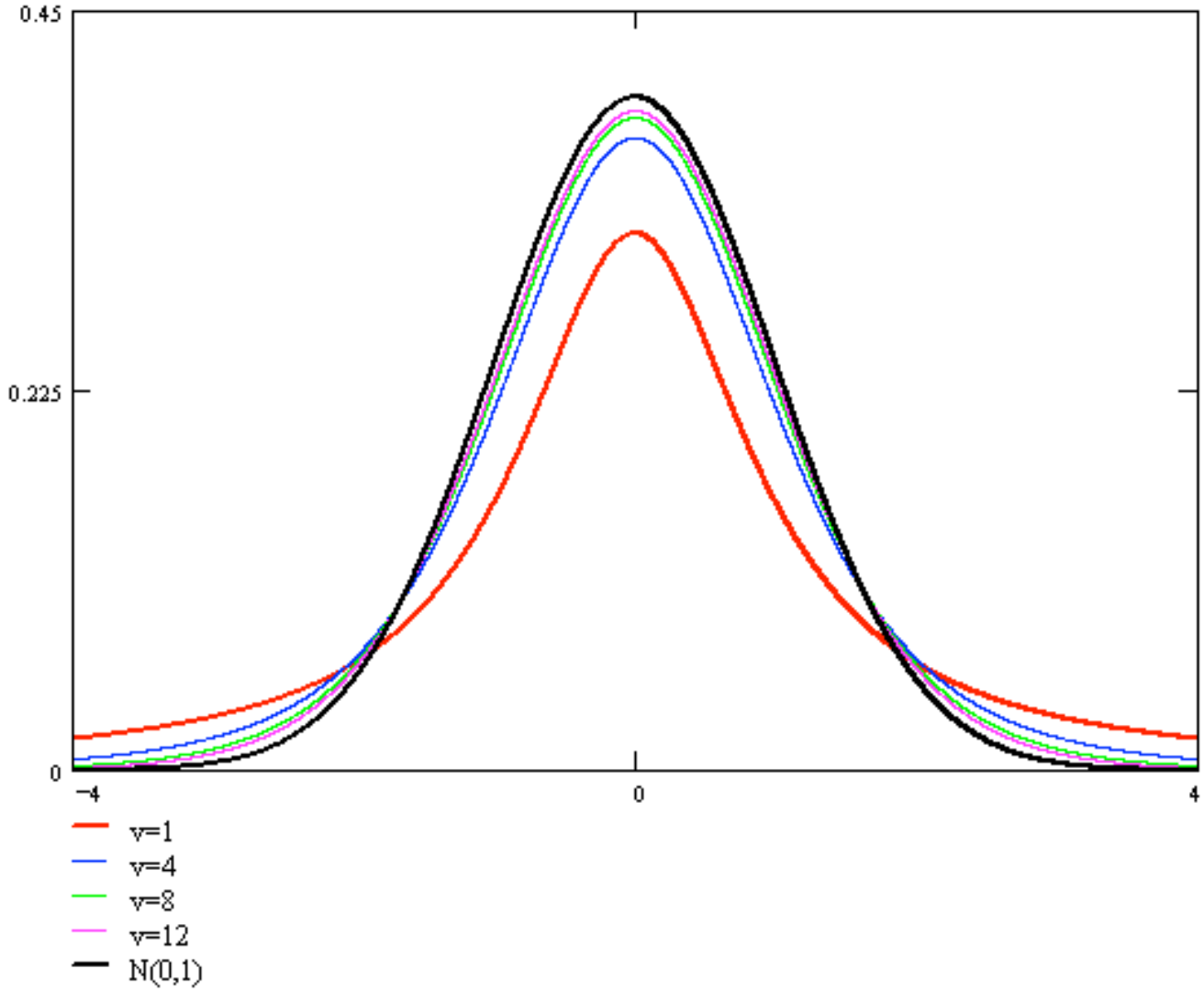
X~N(μ, σ) σ - ? $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$t = \frac{\bar{X} - \mu}{S}\sqrt{n-1} = \frac{\bar{X} - \mu}{\hat{S}}\sqrt{n} \sim t - Studenta$$

$$o\ n-1\ stopniach\ swobody$$

$$\mathrm{E(t)} = 0 \qquad D^2(t) = \sqrt{\frac{n-1}{n-3}}$$

Statystyka matematyczna – rozkłady statystyk z próby



źródło: Wikipedia

Statystyka matematyczna – rozkłady statystyk z próby

Wartości krytyczne rozkładu t-Studenta

$X \sim t_\nu$ - X zmienna losowa o rozkładzie t-Studenta z liczbą stopni swobody ν,
α - poziom istotności,
$t_{\nu,\alpha}$ - wartość krytyczna - liczba taka, że $P(|X| > t_{\nu,\alpha}) = \alpha$

ν \ a	**0,400**	**0,300**	**0,200**	**0,100**	**0,050**	**0,025**	**0,025**	**0,010**	**0,005**	**0,001**
1	1,3764	1,9626	3,0777	6,3137	12,7062	25,4519	25,4519	63,6559	127,3211	636,5776
2	1,0607	1,3862	1,8856	2,9200	4,3027	6,2054	6,2054	9,9250	14,0892	31,5998
3	0,9785	1,2498	1,6377	2,3534	3,1824	4,1765	4,1765	5,8408	7,4532	12,9244
4	0,9410	1,1896	1,5332	2,1318	2,7765	3,4954	3,4954	4,6041	5,5975	8,6101
5	0,9195	1,1558	1,4759	2,0150	2,5706	3,1634	3,1634	4,0321	4,7733	6,8685
6	0,9057	1,1342	1,4398	1,9432	2,4469	2,9687	2,9687	3,7074	4,3168	5,9587
7	0,8960	1,1192	1,4149	1,8946	2,3646	2,8412	2,8412	3,4995	4,0294	5,4081
8	0,8889	1,1081	1,3968	1,8595	2,3060	2,7515	2,7515	3,3554	3,8325	5,0414
9	0,8834	1,0997	1,3830	1,8331	2,2622	2,6850	2,6850	3,2498	3,6896	4,7809
10	0,8791	1,0931	1,3722	1,8125	2,2281	2,6338	2,6338	3,1693	3,5814	4,5868
11	0,8755	1,0877	1,3634	1,7959	2,2010	2,5931	2,5931	3,1058	3,4966	4,4369
12	0,8726	1,0832	1,3562	1,7823	2,1788	2,5600	2,5600	3,0545	3,4284	4,3178
13	0,8702	1,0795	1,3502	1,7709	2,1604	2,5326	2,5326	3,0123	3,3725	4,2209
14	0,8681	1,0763	1,3450	1,7613	2,1448	2,5096	2,5096	2,9768	3,3257	4,1403
15	0,8662	1,0735	1,3406	1,7531	2,1315	2,4899	2,4899	2,9467	3,2860	4,0728
16	0,8647	1,0711	1,3368	1,7459	2,1199	2,4729	2,4729	2,9208	3,2520	4,0149
17	0,8633	1,0690	1,3334	1,7396	2,1098	2,4581	2,4581	2,8982	3,2224	3,9651
18	0,8620	1,0672	1,3304	1,7341	2,1009	2,4450	2,4450	2,8784	3,1966	3,9217
19	0,8610	1,0655	1,3277	1,7291	2,0930	2,4334	2,4334	2,8609	3,1737	3,8833
20	0,8600	1,0640	1,3253	1,7247	2,0860	2,4231	2,4231	2,8453	3,1534	3,8496
21	0,8591	1,0627	1,3232	1,7207	2,0796	2,4138	2,4138	2,8314	3,1352	3,8193
22	0,8583	1,0614	1,3212	1,7171	2,0739	2,4055	2,4055	2,8188	3,1188	3,7922

Statystyka matematyczna – rozkłady statystyk z próby

3. Rozkład różnicy średnich arytmetycznych z próby dla X o rozkładzie normalnym

$X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$

$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$, $\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}\right)$$

Statystyka matematyczna – rozkłady statystyk z próby

4. Rozkład różnicy średnich arytmetycznych z próby dla zmiennej X o rozkładzie normalnym z nieznanymi, ale jednakowymi odchyleniami standardowymi

$X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$ $\sigma_1 = \sigma_2$ - ?

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t - Studenta \; o \; n_1 + n_2 - 2$$

stopniach swobody

Statystyka matematyczna – rozkłady statystyk z próby

5. Rozkład wariancji z próby dla zmiennej X o rozkładzie normalnym

X~N(μ, σ)

$$\chi^2 = \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum(X_i - \mu)^2}{\sigma^2} = \frac{nS_*^2}{\sigma^2} \sim \chi^2 o\ n\ st.swob.$$

gdy $\boldsymbol{\mu}$ - ?

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2 o\ v = n - 1\ st.swob. \qquad E(\chi^2) = v;\ D^2(\chi^2) = 2v$$

gdy v $\rightarrow \infty$ to $\sqrt{2\chi^2} \sim N\left(\sqrt{2v-1}, 1\right)$

Statystyka matematyczna – rozkłady statystyk z próby

Wartości krytyczne rozkładu chi-kwadrat

$X \sim \chi^2_\nu$ - X zmienna losowa o rozkładzie chi-kwadrat z liczbą stopni swobody ν,
α - poziom istotności,
$\chi^2_{\alpha,\nu}$ - wartość krytyczna - liczba taka, że $P(X > \chi^2_{\alpha,\nu}) = \alpha$

ν \ a	**0,995**	**0,990**	**0,975**	**0,950**	**0,900**	**0,100**	**0,050**	**0,025**	**0,010**	**0,005**
1	$0,0^4393$	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2104	10,5965
3	0,0717	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449	12,8381
4	0,2070	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767	14,8602
5	0,4118	0,5543	0,8312	1,1455	1,6103	9,2363	11,0705	12,8325	15,0863	16,7496
6	0,6757	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119	18,5475
7	0,9893	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,3444	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902	21,9549
9	1,7349	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,1558	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093	25,1881
11	2,6032	3,0535	3,8157	4,5748	5,5778	17,2750	19,6752	21,9200	24,7250	26,7569
12	3,0738	3,5706	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	26,2170	28,2997
13	3,5650	4,1069	5,0087	5,8919	7,0415	19,8119	22,3620	24,7356	27,6882	29,8193
14	4,0747	4,6604	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	29,1412	31,3194
15	4,6009	5,2294	6,2621	7,2609	8,5468	22,3071	24,9958	27,4884	30,5780	32,8015
16	5,1422	5,8122	6,9077	7,9616	9,3122	23,5418	26,2962	28,8453	31,9999	34,2671
17	5,6973	6,4077	7,5642	8,6718	10,0852	24,7690	27,5871	30,1910	33,4087	35,7184
18	6,2648	7,0149	8,2307	9,3904	10,8649	25,9894	28,8693	31,5264	34,8052	37,1564
19	6,8439	7,6327	8,9065	10,1170	11,6509	27,2036	30,1435	32,8523	36,1908	38,5821
20	7,4338	8,2604	9,5908	10,8508	12,4426	28,4120	31,4104	34,1696	37,5663	39,9969
21	8,0336	8,8972	10,2829	11,5913	13,2396	29,6151	32,6706	35,4789	38,9322	41,4009
22	8,6427	9,5425	10,9823	12,3380	14,0415	30,8133	33,9245	36,7807	40,2894	42,7957
23	9,2604	10,1957	11,6885	13,0905	14,8480	32,0069	35,1725	38,0756	41,6383	44,1814

Statystyka matematyczna – rozkłady statystyk z próby

6. Rozkład ilorazu wariancji z próby dla zmiennej X o rozkładzie normalnym dla dwóch populacji

$X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$

$$F = \frac{\frac{n_1 S_1^2}{\sigma_1^2 (n_1 - 1)}}{\frac{n_2 S_2^2}{\sigma_2^2 (n_2 - 1)}} = \frac{n_1 S_1^2}{\sigma_1^2 (n_1 - 1)} * \frac{\sigma_2^2 (n_2 - 1)}{n_2 S_2^2} =$$

$$= \frac{S_1^2}{S_2^2} * \frac{\sigma_2^2}{\sigma_1^2} * \frac{(n_2 - 1)}{(n_1 - 1)} * \frac{n_1}{n_2} \sim rozkład\ F - Snedecora\ o\ n_1 - 1\ i\ n_2 - 1\ st.swob.$$

$$E\left(\frac{v_2}{v_2 - 2}\right) \qquad D^2\left(\frac{2v_2^2 (v_1 + v_2 - 2}{v_1 (v_2 - 2)^2 (v_2 - 4)^2}\right)$$

Statystyka matematyczna – rozkłady statystyk z próby

Wartości krytyczne rozkładu F-Snedecora

$X \sim F_{v1, v2}$ - X zmienna losowa o rozkładzie F- Snedecora z liczbami stopni swobody (v1, v2)

poziom istotności α =0,05,

$F_{\alpha, v1, v2}$ - wartość krytyczna - liczba taka, że $P(X > F_{\alpha, v1, v2}) = \alpha$

	v1														
v2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	161,446	199,499	215,707	224,583	230,160	233,988	236,767	238,884	240,543	241,882	242,981	243,905	244,690	245,363	245,949
2	18,513	19,000	19,164	19,247	19,296	19,329	19,353	19,371	19,385	19,396	19,405	19,412	19,419	19,424	19,429
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785	8,763	8,745	8,729	8,715	8,703
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,936	5,912	5,891	5,873	5,858
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,704	4,678	4,655	4,636	4,619
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,027	4,000	3,976	3,956	3,938
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,603	3,575	3,550	3,529	3,511
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388	3,347	3,313	3,284	3,259	3,237	3,218
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,102	3,073	3,048	3,025	3,006
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,943	2,913	2,887	2,865	2,845
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,818	2,788	2,761	2,739	2,719
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,717	2,687	2,660	2,637	2,617
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,635	2,604	2,577	2,554	2,533
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,565	2,534	2,507	2,484	2,463
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,507	2,475	2,448	2,424	2,403
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,456	2,425	2,397	2,373	2,352
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,413	2,381	2,353	2,329	2,308
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,374	2,342	2,314	2,290	2,269
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378	2,340	2,308	2,280	2,256	2,234
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,310	2,278	2,250	2,225	2,203

Statystyka matematyczna – rozkłady statystyk z próby

6. Graniczny rozkład frakcji z próby dla zmiennej X o rozkładzie Bernoulliego

$X_n \sim B(n, p)$,

$$X_n = \sum X_i \qquad Y_n = \frac{X_n}{n}$$

Z twierdzenia Moivre'a-Laplace'a

$$Y_n \sim N\left(p, \sqrt{\frac{pq}{n}}\right) \qquad X_n \sim N\left(np, \sqrt{npq}\right)$$

Statystyka matematyczna – rozkłady statystyk z próby

7. Rozkład średniej arytmetycznej z próby dla zmiennej X o dowolnym rozkładzie

X~rozkład dowolny $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Z twierdzenia Lindeberga-Levy'ego wynika, że:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Podobnie dla różnicy średnich

Statystyka matematyczna – rozkłady statystyk z próby

8. Rozkład różnicy frakcji z próby dla zmiennej X o rozkładzie Bernoulliego

$X_{n1} \sim B(n_1, p_1)$, $X_{n2} \sim B(n_2, p_2)$

$$X_{n1} = \sum X_{i1} \qquad Y_{n1} = \frac{X_{n1}}{n_1} \qquad X_{n2} = \sum X_{i2} \qquad Y_{n2} = \frac{X_{n2}}{n_2}$$

$$Y_{n1} - Y_{n2} \sim N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Statystyka matematyczna – estymacja

Estymacja – szacowanie, ocenianie, przybliżanie

Estymator – każda statystyka $\widehat{\Theta}_n(X_1, \dots, X_n)$, która służy do oszacowania danego parametru.

Estymator – zmienna losowa, jego rozkład jest zależny od rozkładu zmiennej losowej X w populacji oraz od szacowanego parametru.

Ocena parametru – wartość estymatora obliczona na bazie jednej próby

Statystyka matematyczna – estymacja

Błąd estymacji (szacunku) parametru $\boldsymbol{\theta}$ – różnica między wartością estymatora a wartością parametru: $d = \widehat{\Theta}_n - \Theta$.

Błąd szacunku jest zmienną losową, więc: $\Delta = E\left(\widehat{\Theta}_n - \Theta\right)^2$.

Jeśli $E\left(\widehat{\Theta}_n\right) = \Theta$ to $\Delta = E\left(\widehat{\Theta}_n - E\left(\widehat{\Theta}_n\right)\right)^2 = D^2\left(\widehat{\Theta}_n\right)$.

$D\left(\widehat{\Theta}_n\right)$ - standardowy błąd szacunku,

$D\left(\widehat{\Theta}_n\right)/\ \Theta$ – względny błąd szacunku

Statystyka matematyczna – estymacja

Własności estymatora

1. Nieobciążoność

$$E\left(\widehat{\Theta}_n\right) = \Theta$$

Obciążenie: $B_n(\Theta) = E\left(\widehat{\Theta}_n\right) - \Theta$

Estymator asymptotycznie nieobciążony

$$\lim_{n\to\infty} B_n(\Theta) = 0$$

Statystyka matematyczna – estymacja

2. Zgodność estymatora

$$\lim_{n\to\infty} P\left(\left|\widehat{\Theta}_n - \Theta\right| < \varepsilon\right) = 1, \boldsymbol{\varepsilon} > 0.$$

Jeśli estymator jest zgodny to jest asymptotycznie nieobciążony

Jeśli estymator jest nieobciążony lub asymptotycznie nieobciążony oraz jego wariancja spełnia warunek $\lim_{n\to\infty} D^2\left(\widehat{\Theta}_n\right) = 0$ to estymator ten jest estymatorem zgodnym

Statystyka matematyczna – estymacja

3. Efektywność estymatora

Estymatorem efektywnym (najefektywniejszym) nazywamy nieobciążony estymator $\widehat{\Theta}_n$ parametru Θ, który ma najmniejszą wariancję spośród wszystkich nieobciążonych estymatorów tego parametru wyznaczonych z prób n-elementowych.

Nierówność Rao-Cramera

$$D^2\left(\widehat{\Theta}_n\right) \geq \frac{1}{nE\left(\frac{\partial}{\partial\Theta} lnf(X,\Theta)\right)^2}$$

Statystyka matematyczna – estymacja

3. Efektywność estymatora

Miarą efektywności estymatora $\widehat{\Theta}_n$ jest liczba

$$ef\left(\widehat{\Theta}_n\right) = \frac{D^2\left(\widetilde{\Theta}_n\right)}{D^2\left(\widehat{\Theta}_n\right)}$$

$$0 < ef\left(\widehat{\Theta}_n\right) < 1$$

Estymator asymptotycznie efektywny

$$\lim_{n\to\infty} ef\left(\widehat{\Theta}_n\right) = 1$$

Statystyka matematyczna – estymacja

Metody wyznaczania estymatorów

1. Metoda momentów – przyrównanie momentów z próby do momentów rozkładu

Własności takich estymatorów:

- Na ogół niska efektywność
- Są na ogół zgodne

Statystyka matematyczna – estymacja

2. Metoda największej wiarygodności (MNW)

Niech rozkład zmiennej X zależy od k nieznanych parametrów $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, …, $\boldsymbol{\theta}_k$, które chcemy oszacować na podstawie n-elementowej próby.

$$L = \prod_{i=1}^{n} f(X_i, \Theta_1, \dots, \Theta_k)$$

Własności takich estymatorów:

- zgodne
- asymptotycznie nieobciążone
- asymptotycznie efektywne
- mają asymptotyczne rozkłady normalne

Statystyka matematyczna – estymacja

Przegląd estymatorów

Szanowanie wartości przeciętnej $\boldsymbol{\mu}$:

Estymator - średnia arytmetyczna z próby: nieobciążony, zgodny dla rozkładu dowolnego, efektywny dla rozkładu normalnego,

Estymator – mediana z próby: asymptotycznie nieobciążony, zgodny dla rozkładu dowolnego, efektywność równa $2/\boldsymbol{\pi}$ = 0,64 dla rozkładu normalnego

Statystyka matematyczna – estymacja

Przegląd estymatorów

Szanowanie wariancji $\boldsymbol{\sigma}^2$:

Estymator - $S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$: nieobciążony, zgodny dla rozkładu dowolnego, efektywny dla rozkładu normalnego,

Estymator – $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$: asymptotycznie nieobciążony, zgodny dla rozkładu dowolnego

Estymator – $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$: nieobciążony, zgodny dla rozkładu dowolnego, efektywność równa (n-1)/n dla rozkładu normalnego

Statystyka matematyczna – estymacja

Przegląd estymatorów

Szanowanie odchylenia standardowego $\boldsymbol{\sigma}$

Estymator - S_* , S , $\hat{S}$: zgodny dla rozkładu dowolnego

Estymator – $b_n S$, $c_n \hat{S}$: zgodny, asymptotycznie nieobciążony, efektywny dla rozkładu normalnego

$$b_n = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma(n)}\sqrt{\frac{n}{2}} \qquad c_n = \sqrt{\frac{n-1}{n}}\, b_n$$

dla n = 20 b_n = 1,04, a c_n = 1,014

Statystyka matematyczna – estymacja

Przegląd estymatorów

Szanowanie wskaźnika struktury p

Estymator - $\hat{p} = \frac{m}{n}$: nieobciążony, zgodny i efektywny dla rozkładu Bernoulliego

Statystyka matematyczna – estymacja

Estymacja przedziałowa

Przedziałem ufności dla parametru $\boldsymbol{\theta}$ na poziomie ufności $1 - \alpha$ nazywamy przedział (θ_1, θ_2) spełniający warunki:

- jego końce $\theta_1(X_1, \ldots, Xn)$ i $\theta_2(X_1, \ldots, X_n)$ są funkcjami próby losowej i nie zależą od szacowanego parametru,
- prawdopodobieństwo pokrycia przez ten przedział nieznanego parametru $\boldsymbol{\theta}$ jest równe $1 - \alpha$, tzn.

$$P\big(\theta_1(X_1, \ldots, X_n) < \theta < \theta_2(X_1, \ldots, X_n)\big) = 1 - \alpha$$

Statystyka matematyczna – estymacja

Estymacja przedziałowa

Dokładność estymacji – różnica między górną i dolną granicą przedziału – długość przedziału ufności.

Zależy ona od:

- współczynnika ufności – im wyższy tym długość przedziału większa a mniejsza dokładność,
- liczebności próby – im większa tym długość przedziału mniejsza a dokładność większa

Statystyka matematyczna – estymacja

Konstrukcja przedziału ufności dla wartości przeciętnej μ w populacji, w której badana cecha ma rozkład N(μ, σ), gdy σ jest znane.

1. Wiadomo, że statystyka $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ma rozkład $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

2. Standaryzujemy: $U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$

3. Ponieważ rozkład U nie zależy od szacowanego parametru możemy wykorzystać go do konstrukcji przedziału ufności

Statystyka matematyczna – estymacja

Konstrukcja przedziału ufności dla wartości przeciętnej μ w populacji, w której badana cecha ma rozkład N(μ, σ), gdy σ jest znane.

4. Dla danego α można znaleźć takie wartości u_1 i u_2 , aby:

$$P(u_1 < U < u_2) = \Phi(u_2) - \Phi(u_1) = 1 - \alpha$$

$\alpha_1 + \alpha_2 = \alpha$; $\alpha_1 > 0$; $\alpha_2 < \alpha$

$u_1 = U(\alpha_1)$; $u_2 = U(1-\alpha_2)$

Statystyka matematyczna – estymacja

Konstrukcja przedziału ufności dla wartości przeciętnej μ w populacji, w której badana cecha ma rozkład N(μ, σ), gdy σ jest znane.

5. Podstawiamy

$$P\left(u(\alpha_1) < \frac{\bar{X} - \mu}{\sigma}\sqrt{n} < u(1-\alpha_2)\right) = 1 - \alpha$$

6. Rozwiązując nierówność wewnątrz nawiasu względem μ mamy:

$$P\left(\bar{X} - u(1-\alpha_2)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - u(\alpha_1)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Statystyka matematyczna – estymacja

Konstrukcja przedziału ufności dla wartości przeciętnej μ w populacji, w której badana cecha ma rozkład N(μ, σ), gdy σ jest znane.

7. Przyjmując, że $\alpha_1 = \alpha_2 = ½\alpha$ to przedział ufności wygląda następująco:

$$P\left(\bar{X} - u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - u\left({}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

8. Ze względu na symetrię rozkładu N(0, 1), gdzie –u(α/2) = u(1-α/2):

$$P\left(\bar{X} - u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Statystyka matematyczna – estymacja

Wyznaczanie minimalnej liczebności próby

Przyjmijmy, że estymować chcemy wartość przeciętną μ w populacji, w której badana cecha ma rozkład N(μ, σ), gdy σ jest znane. Wtedy przedział ufności jest następujący:

$$P\left(\bar{X} - u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Długość tego przedziału wynosi:

$$2d = 2u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}$$

Statystyka matematyczna – estymacja

Wyznaczanie minimalnej liczebności próby

Po przekształceniu

$$n = \left[\frac{u^2\left(1 - {}^{\alpha}/_{2}\right)\sigma^2}{d^2}\right] + 1$$

Analogiczny wzór uzyskujemy dla w przypadku estymacji wskaźnika struktury:

$$n = \left[\frac{u^2\left(1 - {}^{\alpha}/_{2}\right)\hat{p}(1-\hat{p})}{d^2}\right] + 1$$

Statystyka matematyczna – estymacja

Wyznaczanie minimalnej liczebności próby – przykład

Jaka liczebność próby dla estymacji procentu poparcia partii politycznych

$$n = \left[\frac{u^2\left(1 - {}^{\alpha}/_{2}\right)\hat{p}(1-\hat{p})}{d^2}\right] + 1$$

Zakładając, że α = 0,05; $\hat{p} = 0{,}5$; d = 0,03 mamy:

$$n = \left[\frac{(1{,}96)^2 * 0{,}5 \ * (1 - 0{,}5)}{(0{,}03)^2}\right] + 1 = [1067{,}11] + 1 = 1068$$

Statystyka matematyczna – weryfikacja hipotez

Hipoteza statystyczna – każde przypuszczenie dotyczące nieznanego rozkładu badanej cechy populacji o prawdziwości lub fałszywości, którego wnioskuje się na podstawie próby losowej

Hipotezy:

- parametryczne
- nieparametryczne

- proste
- złożone

Statystyka matematyczna – weryfikacja hipotez

Test statystyczny – metoda postępowania (procedura), która każdej możliwej realizacji próby losowej X_1, …, X_n przyporządkowuje (z ustalonym prawdopodobieństwem) decyzję przyjęcia lub odrzucenia sprawdzanej hipotezy.

Konstrukcja testu statystycznego

1. Formułuje się hipotezę, którą weryfikujemy (hipotezę zerową – H_0)
2. Formułuje się hipotezę alternatywną
3. Wybieramy odpowiednią statystykę testową
4. Konstruuje się tzw. zbiór krytyczny

Statystyka matematyczna – weryfikacja hipotez

Błędy przy weryfikacji hipotezy statystycznej

Decyzja	Hipoteza H_0	
	prawdziwa	fałszywa
Przyjąć H_0	Decyzja poprawna	Błąd II rodzaju
Odrzucić H_0	Błąd I rodzaju	Decyzja poprawna

$$P(\delta(X_1, \dots, X_n) \in W | H_0) = \alpha$$

$$P(\delta(X_1, \dots, X_n) \in W' | H_1) = \beta$$

Statystyka matematyczna – weryfikacja hipotez

Testy konstruuje się tak, aby zminimalizować prawdopodobieństwo popełnienia błędu II rodzaju przy ustalonym poziomie prawdopodobieństwie popełnienia błędu pierwszego rodzaju (α).

Takie testy nazywamy testami najmocniejszymi, ponieważ przy ustalonym α odpowiada im największa moc, tzn. prawdopodobieństwo odrzucenia fałszywej hipotezy H_0 i przyjęcia hipotezy H_1.

Test jednostajnie najmocniejszy – najmocniejszy względem każdej hipotezy H_1

Statystyka matematyczna – weryfikacja hipotez

Moc testu

$$M(W,\Theta) = P(\delta(X_1,\dots,X_n) \in W|\Theta) = 1 - P(\delta(X_1,\dots,X_n) \in W'|\Theta)$$

$$M(W,\Theta_0) = P(\delta(X_1,\dots,X_n) \in W|\Theta_0) = P(\delta(X_1,\dots,X_n) \in W|H_0) = \alpha$$

$$M(W,\Theta_1) = P(\delta(X_1,\dots,X_n) \in W|\Theta_1) = P(\delta(X_1,\dots,X_n) \in W|H_1)$$
$$= 1 - P(\delta(X_1,\dots,X_n) \in W'|H_1) = 1 - \beta$$

Statystyka matematyczna – weryfikacja hipotez

Funkcja operacyjno-charakterystyczna (charakterystyka testu)

$$L(W,\Theta) = P(\delta(X_1,\dots,X_n) \in W'|\Theta) = 1 - M(W,\Theta)$$

$$L(W,\Theta_0) = 1 - \alpha$$

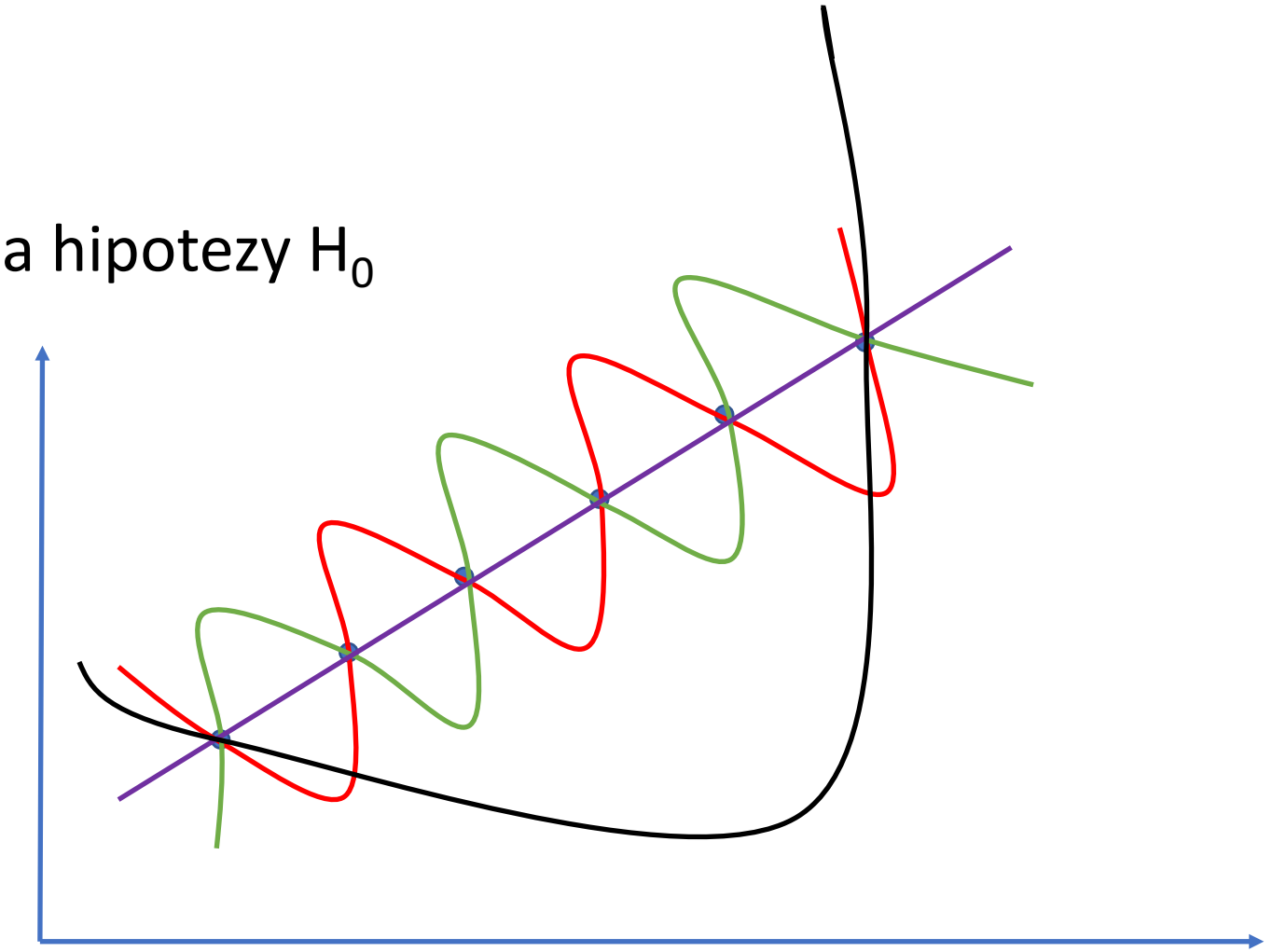
$$L(W,\Theta_1) = \beta$$

Statystyka matematyczna – weryfikacja hipotez

Test istotności

Brak możliwości przyjęcia hipotezy H_0

Istotność różnicy



Statystyka matematyczna – weryfikacja hipotez

Przykład

Załóżmy, że wydajność pracy ma rozkład N(μ, 2). Czy można twierdzić, że wydajność jest wyższa niż 20 szt./h, jeśli na podstawie 16 elementowej próby uzyskano wynik $\bar{x} = 21$, α = 0,05.

H_0: μ = 20 H_1: $\mu > 20$ H_1: $\mu \neq 20$

$$U = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} = \frac{21 - 20}{2} * 4 = 2{,}00$$

$P(U \geq 1{,}64) = 0{,}05$ $P(|U| \geq 1{,}96) = 0{,}05$

Statystyka matematyczna – weryfikacja hipotez

Przykład

A co jeśli σ jest nieznane, a na podstawie próby obliczono s = 2

H_0: $\mu = 20$ H_1: $\mu > 20$ H_1: $\mu \neq 20$

$$t = \frac{\bar{X} - \mu}{s}\sqrt{n-1} = \frac{21-20}{2} * \sqrt{15} = 1{,}94$$

$P(t \geq 1{,}753) = 0{,}05$ $P(|t| \geq 2{,}131) = 0{,}05$

Statystyka matematyczna – weryfikacja hipotez

Testy nieparametryczne

- testy zgodności
- testy niezależności

Test zgodności χ^2 – Pearsona

X ~ dowolny rozkład o określonej postaci dystrybuanty

Duża próba prosta

$H_0: F(x) = F_0(x)$ $\quad\quad$ $H_1: F(x) \neq F_0(x)$

Statystyka matematyczna – weryfikacja hipotez

Statystyka testowa

$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2$ o k-1 stopniach swobody przy hipotezie prostej,

natomiast przy hipotezie złożonej liczba stopni swobody wynosi k-s-1.

Warunki stosowalności:

- Taki podział na klasy, by $np_i \geq 5$
- Gdy liczba stopni swobody jest > 5, to np_i w dwóch klasach może być mniejsze od 5, ale większe bądź równe od 1
- W rozkładach jednomodalnych o klasach tej samej długości w skrajnych klasach mogą być mniejsze liczebności teoretyczne

Statystyka matematyczna – weryfikacja hipotez

Test niezależności χ^2

Cel: weryfikacja hipotezy o związkach między dwiema zmiennymi

H_0: $P(X = x_i, Y = y_j) = P(X = x_i) * P(Y = y_j)$

H_1: $P(X = x_i, Y = y_j) \neq P(X = x_i) * P(Y = y_j)$

Statystyka testowa: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \widehat{n_{ij}})^2}{\widehat{n_{ij}}}$

$$P\left(\chi^2 \geq \chi^2_{\alpha;(r-1)*(s-1)}\right) = \alpha$$

Statystyka matematyczna – weryfikacja hipotez

Test niezależności χ^2

H_0: rodzaj preferowanego programu nie zależy od poziomu wykształcenia

Wykształcenie (X)	Rodzaj programu (Y)				Ogółem
	film	teatr	programy rozrywkowe	programy publicystyczne	
Podstawowe	105	10	75	10	200
Średnie	120	60	80	40	300
Wyższe	35	30	15	20	100
Ogółem	260	100	170	70	600

$\chi^2 = 62{,}04$; **r = 3; s = 4**

$$P\left(\chi^2 \geq \chi^2_{0{,}05;(3-1)*(4-1)}\right) = P\left(\chi^2 \geq \chi^2_{0{,}05,6}\right) = P(\chi^2 \geq 12{,}592) = 0{,}05$$