

Wykład 6

Przedział ufności dla frakcji w populacji

Dyskusja:

- Estymujemy p za pomocą $\hat{p}$
- Chcemy skonstruować przedział ufności dla p
- Moglibyśmy skorzystać z rozkładu dwumianowego (uciążliwe)
- Korzystamy z przybliżenia rozkładu Bernoulliego rozkładem normalnym: gdy Y ma rozkład dwumianowy (n, p) oraz n jest duże, to Y ma w przybliżeniu rozkład normalny

$$N(np, \sqrt{np(1-p)})$$

- $\hat{p} = Y/n$ ma wartość oczekiwaną $\mu = \dots$ i odchylenie standardowe $\sigma = \dots$

- Zatem $\hat{p}$ ma w przybliżeniu rozkład: ...

Przedział ufności dla p

- Będziemy korzystali z *przybliżonego przedziału ufności Agrestiego-Coula* (patrz np. Brown, Cai i DasGupta, Annals of Statistics, 2002):

$$\tilde{p} \pm Z_{\alpha/2} SE_{\tilde{p}}$$

- Środkiem przedziału jest $\tilde{p}$, pewna modyfikacja $\hat{p}$
- Jak poprzednio, $Z_{\alpha/2}$ jest taką liczbą, że $\Pr(Z < -Z_{\alpha/2}) = \Pr(Z > Z_{\alpha/2}) = \alpha/2$
- Np. dla 95% PU, $\alpha = 0.05$ i $Z_{\alpha/2} = 1.96$

- Definiujemy
$$\tilde{p} = \frac{y + 0.5(Z_{\alpha/2}^2)}{n + Z_{\alpha/2}^2}$$

oraz
$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + Z_{\alpha/2}^2}}$$

- Np. dla 95% PU wstawiamy $Z_{0.025} = 1.96$ i dostajemy

$$\tilde{p} = \frac{y + 0.5(1.96^2)}{n + 1.96^2} = \frac{y + 1.92}{n + 3.84} \approx \frac{y + 2}{n + 4}$$

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + 1.96^2}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + 3.84}} \approx \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + 4}}$$

TABLE A Standard normal probabilities (continued)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Przedział ufności dla p (cd.)

- Użyjemy kwantyli z rozkładu normalnego $Z_{\alpha/2}$
- Dla 90% PU użyjemy $Z_{0.05} = 1.65$; dla 99% PU użyjemy $Z_{0.005} = 2.58$.
- Dla 95% PU użyjemy $Z_{0.025} = 1.96$
- Zatem przybliżony 95% PU dla p wynosi

$$\tilde{p} \pm 1.96 \times SE_{\tilde{p}} = \tilde{p} \pm 1.96 \times \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

Przykład:

- Złapano 125 myszy i 6 z nich ma brzuski nakrapiane na biał
- p = frakcja myszek w całej populacji, które mają nakrapiane na biał brzuski
- 95% PU dla p:

90% PU dla p

- Sformułowanie konkluzji:

Mamy 90% pewności że frakcja myszek w tej populacji, które mają brzuszki nakrapiane na białło zawiera się w przedziale między ... a

- Zauważmy, że 90% PU jest ... niż 95% PU i że przedziały te mają różne środki.

Klasyczny przybliżony przedział ufności
(informacja porządkująca)

- Klasyczny przedział ufności uzyskuje się biorąc $\hat{p}$ za jego środek i zastępując p przez $\hat{p}$ we wzorze na σ i błąd standardowy.
- Klasyczne przedziały ufności zachowują się źle, gdy liczba sukcesów (Y) jest bliska zeru lub n. Może się wtedy zdarzyć, że klasyczny PU zawiera ujemne wartości!

Jak duża powinna być próba ?

- Chcemy aby 95% PU miał długość nie większą od zadanej. Jak ustalić rozmiar próby?
- Idea: szerokość przedziału ufności zależy od n, stąd warunek na n. Trudność: szerokość zależy też od $\hat{p}$, którego na etapie planowania próby jeszcze nie znamy.
- Jeżeli wiemy w przybliżeniu, jakie jest p, to możemy tej przybliżonej wartości użyć w równaniu na długość przedziału (skąd wyznaczymy n).
- Jeżeli brak wstępnych informacji, to użyjemy $p = 0.5$. Prowadzi to do *ostrożnego/konserwatywnego* wyboru n: szerokość PU skonstruowanego w oparciu o próbę o tak wyznaczonym rozmiarze będzie na ogół dużo mniejsza. (Czy to dobrze?)

Przykład

- Chcemy aby SE było równe 0.005 (wtedy 95% PU będzie miał długość około 0.02).
- Przypuszczamy, że prawdziwe p jest bliskie 0.05. Rachunki:

- Konkluzja: „Proszę złapać ... myszy.”

- Obliczenia na okoliczność, gdy nie wiemy nic o p :

- Konkluzja: „Proszę złapać ... myszy.”