MATH 180A (Lecture A00)

mathweb.ucsd.edu/~ynemish/teaching/180a

Today: Confidence intervals

Next: ASV 4.4

Week 7:

- no homework
- Midterm 2 on Wednesday, March 1 (lectures 8-18)
- Homework 5 due Friday, March 3

Summary

CLT for Bernoulli distribution:

Let $S_n \sim \text{Bin}(n,p)$, let $a<b$. Then

$$\lim_{n\to\infty} P\left(a < \frac{S_n - np}{\sqrt{np(1-p)}} < b\right) = \Phi(b) - \Phi(a)$$

Rule of thumb

$np(1-p) > 10$

For the average $\overline{S_n} := \frac{S_n}{n}$

$$P\left(a\frac{\sqrt{p(1-p)}}{\sqrt{n}} < \overline{S_n} - p < b\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \approx \Phi(b) - \Phi(a)$$

LLN for Bernoulli

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) = 1$$

Confidence intervals. Motivation

Consider n independent trials, success rate p (unknown)

S_n = number of successes after n trials, $S_n \sim \mathrm{Bin}(n,p)$

By the LLN $\frac{S_n}{n} \to p$, $n \to \infty$ (in probability)

If n is big, then $\frac{S_n}{n}$ is close to p

random variable ← observable, estimate of p

number

Usually we do not know p, but we can get a realization of $\frac{S_n}{n}$ (flipping a coin) for finite n.

What can we say about p?

Confidence intervals. Set-up

Denote $\hat{p} := \frac{S_n}{n}$ and use the CLT for the interval $(-a, a)$

$$P\left(\frac{-a\sqrt{p(1-p)}}{\sqrt{n}} \le \hat{p} - p \le \frac{a\sqrt{p(1-p)}}{\sqrt{n}} \right) \approx$$

$$P\left(|\hat{p} - p| \le \underbrace{\frac{a\sqrt{p(1-p)}}{\sqrt{n}}}_{=:\,\varepsilon} \right) \approx$$

Questions:

1) For fixed n, find $\varepsilon > 0$ such that $P(|\hat{p} - p| \le \varepsilon) \ge \gamma$

2) For fixed ε, find $n \in \mathbb{N}$ such that $P(|\hat{p} - p| \le \varepsilon) \ge \gamma$

Confidence intervals

$$P\left(p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]\right) \geq \gamma$$

Notice that $\hat{p}$ is a random variable, so the interval is random

Take some realization of

Then $[\hat{p}_* - \varepsilon, \hat{p}_* + \varepsilon]$ is the

Confidence intervals. Computations

Problem: To find an estimate of p, we use

$$P\left(|\hat{p}-p| \le \underbrace{\frac{a\sqrt{p(1-p)}}{\sqrt{n}}}_{=:\varepsilon} \right) \approx 2\Phi(a)-1 =: \gamma ,$$

for which ε depends on (unknown) p.

Solution: notice that

$$P\left(|\hat{p}-p| \le \frac{a\sqrt{p(1-p)}}{\sqrt{n}} \right) \approx \gamma$$

and the γ-confidence interval can be taken as

$[\hat{p}-\varepsilon, \hat{p}+\varepsilon]$ with

-

-

Confidence intervals. Example 1 (fixed n)

Flip a coin 10000 times. Number of heads is 5370.

Compute a 99%-confidence interval for $p = P(\text{Heads})$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Conclusion:

Confidence intervals. Example 2 (fixed accuracy)

Flip a (possibly biased) coin. How many times should we repeat the experiment to be able to compute a 95%-confidence interval for $p = P(\text{Heads})$ of length 0.01?

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Conclusion:

Confidence intervals. Polling

Without going into details (see example 4.14, Remark 4.16)

Part of the population (unknown p) prefers product A / supports candidate B / etc.... We interview n individuals and k of them give a positive answer (about product/candidate)

What can we say about p?

Fix confidence level γ (often $\gamma = 0.95$).

Again, $\hat{p}_* =$ is the estimate of p, and

gives a -confidence interval,

so from $2\Phi(2\varepsilon\sqrt{n}) - 1 \geq \gamma$ we find ε for given n,

or find n for given ε (how many people we have to interview to achive the derised margin error)

Confidence intervals. Example (rock vs rap)

You ask 400 randomly chosen San Diegans whether they prefer rock or rap. 230 choose rock.

Give a 99% confidence interval for the part of the population that prefers rock.

CLT for Bernoulli / Binomial

Let $S_n \sim \text{Bin}(n,p)$, $a<b$. Then

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) = \Phi(b) - \Phi(a)$$

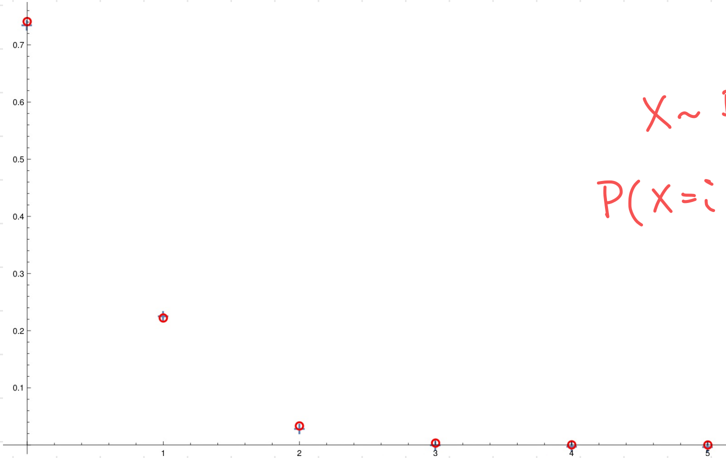
- n independent trials
- probability of success p
- rule of thumb: $np(1-p) > 10$

If this holds, then $N(np, np(1-p))$ gives a good approximation of $\text{Bin}(n,p)$

What about other regimes? $np(1-p) < 10$

Example

$S_{10} \sim \text{Bin}(10, 0.03)$

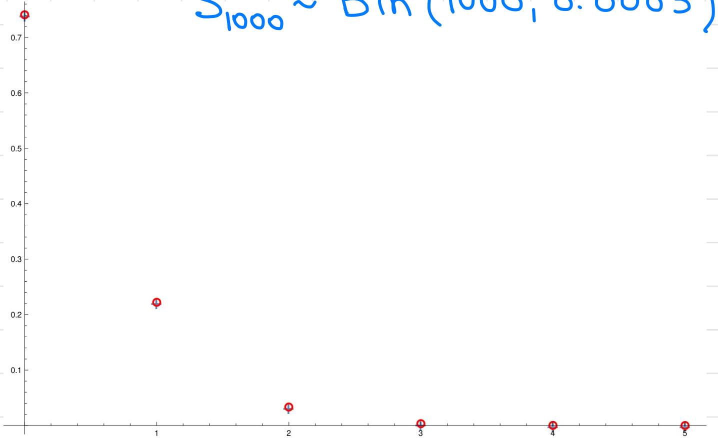


$P(S_{10} = i)$

$X \sim \text{Poisson}(0.3)$

$P(X = i)$

$S_{1000} \sim \text{Bin}(1000, 0.0003)$



$P(S_{1000} = i)$

Poisson approximation

$\lambda > 0$, $X \sim \text{Poisson}(\lambda)$, for $k = 0, 1, 2, \ldots$

$$P(X = k) =$$

Lecture 13: $E(X) =$

$Var(X) =$ (similarly as $E(X)$)

Lecture 12: Let $\lambda > 0$ and n positive integer.

Let $S_n \sim Bin(n, \frac{\lambda}{n})$ (assume $\frac{\lambda}{n} < 1$).

Then for any fixed $k \in \{0, 1, 2, \ldots\}$

Poisson approximation for finite n

When to approximate by Poisson ?

<u>Proposition</u> Let $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Poisson}(np)$.

Then for any $A \subset \{0,1,2,\ldots\}$

If np^2 is small $\to$ use Poisson approximation.

<u>Remark.</u> If $p = \frac{\lambda}{n}$, then one can have $np(1-p) = \lambda(1-\frac{\lambda}{n})$ and at the same time $np^2 = \frac{\lambda^2}{n}$

Then both approximations, normal by $N(\lambda, \lambda)$ and Poisson (λ), are close to $\mathrm{Bin}(n, \frac{\lambda}{n})$

Example. Approximating Bin(n,p)

John flips a coin repeatedly until the tails comes up and counts the number of flips.

Approximate the probability that in a year there are at least 3 days when he needs more that 10 flips.

Approximating probabilities of rare events

Poisson distribution is used to model the occurances of rare events. Examples:

customers arriving in a store ⟷ all potential customers decide independently to come or not

number of emergency calls ⟷ all people in the city have an emergency or not "independently" of each other

number of car accidents ⟷ all drivers in the county have accidents (or not) "independently"

number of goals scored in a hockey game ⟷ a lot of "independent" shots

Example

Number of phone calls in a day can be modeled by Poisson random variable. We know that on average 0.5% of the time the call center receives no calls at all. What is the average number of calls per day?

Exercise

10% of households earn > 80000 \$

0.25% of households earn > 450000 \$

Choose 400 households at random. Denote

$X = \#$ households > 80000 , $Y = \#$ households > 450000 \$

Estimate $P(X \geq 48)$ and $P(Y \geq 2)$

Question

A fair 20-sided die is tossed 400 times. We want to calculate the probability that a 13 came up at least 25 times. We should use

(a) Poisson approximation
(b) Normal approximation
(c) Neither
(d) Both