Lecture 13

§1 Confidence interval

1. 样本数量与准确性

1° point estimator 越高 ≠ true parameter 越高

2° data 越多，estimator 就越接近 true parameter.

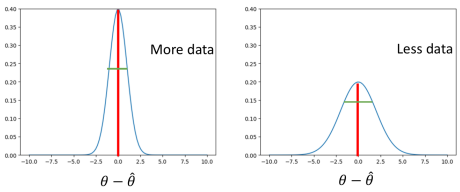
Experiments

Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$

- Drug 1: $\hat{p}_1 = 90\%$. 10 experiments.
- Drug 2: $\hat{p}_2 = 80\%$. 10000 experiments.

Which drug do you think is more effective?

- Number of samples can affect the accuracy!!!
- With more data, we believe the estimator is closer to the true parameter.

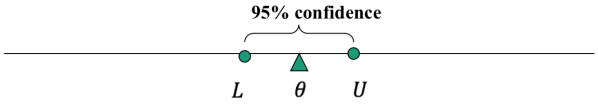


2. Confidence interval (置信区间/可靠区间)

1° 是一个关于参数 θ 的 interval estimate [L, U]

2° 当 P(L≤θ≤U) 达到 0.95 时，认为 θ 的 true value 落在 [L,U] 上的概率很大

- **If we want to find the interval estimation such that**
  $P(L \le \theta \le U) = 0.95$
- **This means that if we generate many such intervals, then we expect 95% of those intervals will include the true parameter θ**

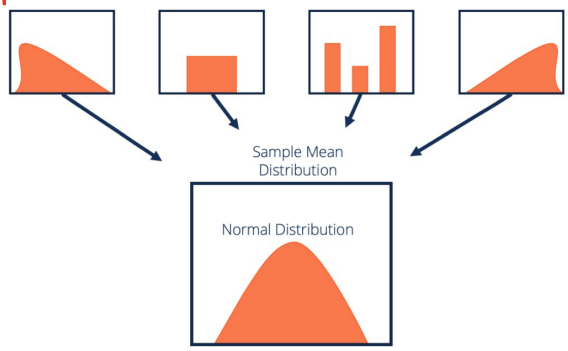


3° L与U是从数据中取的，thus are random

§2 Central limit theorem

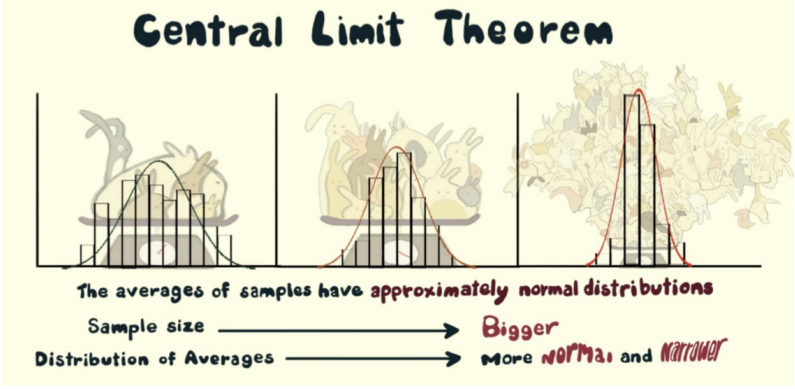
1. Central limit theorem (中心极限定理)

1° 对于任意概率分布，只要每轮取样的样本空间足够大，轮数足够多，则各轮取样的 sample mean 服从正态分布



No matter what the true distribution is, the **sample mean** will be very close to the **normal distribution**, as long as the sample size is **large**.

2° 样本空间越大，sample means 的正态分布曲线就越高、越窄.
(每轮取样的均值更接近 true mean)



* 3° 若 true mean 为 μ，true variance 为 σ^2，每轮的样本空间为 n，共 N 轮

则对于每一轮而言：

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

若令 $Y_i = \bar{X}_i$，则对于整体而言：

$$\text{mean} = \frac{Y_1 + Y_2 + \cdots + Y_N}{N} = \frac{X_1 + X_2 \cdots + X_{n\cdot N}}{N \cdot n} \to \mu \text{ as } n\cdot N \to \infty$$

$$\text{variance} = \frac{(Y_1-\mu)^2 + \cdots + (Y_N-\mu)^2}{N}$$

$$= \frac{[(X_1-\mu)+(X_2-\mu)+\cdots+(X_n-\mu)]^2 + \cdots}{n^2 \cdot N}$$

$$\approx \frac{(X_1-\mu)^2 + (X_2-\mu)^2 \cdots + (X_{nN}-\mu)^2}{n^2 \cdot N}$$

$$\to \frac{\sigma^2}{n} \text{ as } n \to \infty$$

因此，$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

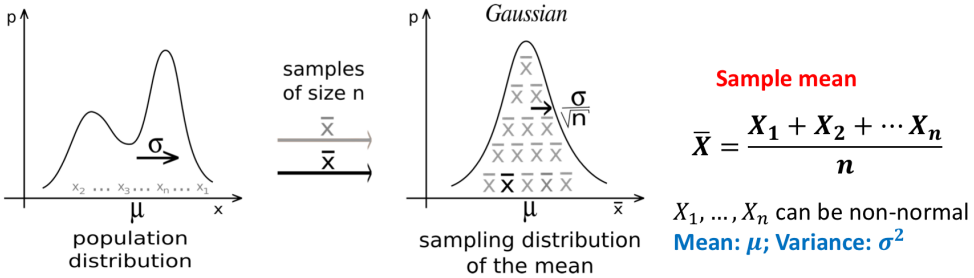
2、正态分布标准化处理

1° 均值与方差变换

若 $X \sim N(\mu, \sigma^2)$，则 $3X \sim N(3\mu, 9\sigma^2)$

2° 由此可求得：

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$



Sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$$

$X_1, \ldots, X_n$ can be non-normal

Mean: μ; Variance: σ^2

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Or write as: $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$

standard normal density

§3 Interval estimation

- We have data $X_1, X_2, \ldots, X_n$ that are sampled from some distribution
- Their mean is μ, which we want to estimate
- We can easily give a point estimate: $\bar{X}$ (sample mean)
- How to get an interval estimate??
  - Use Central Limit Theorem!

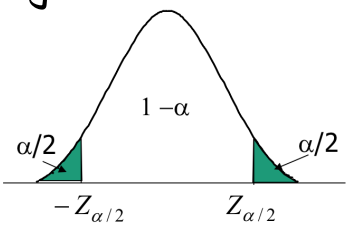
已知 $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$，Confidence level 为 $1-\alpha$ $(0 \le \alpha \le 1)$

1、步骤一：在标准正态分布中选取合适的区间

1° 对于标准正态分布曲线，取 $z_{\frac{\alpha}{2}}$，使得在 $z_{\frac{\alpha}{2}}$ 右侧的曲线下方的面积总和为 $\frac{\alpha}{2}$，则 $P(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha$

2° 此时求出的最终区间会关于 μ 对称

e.g. 若要 $P(L \le \theta \le U) = 0.95$，则 $L = -z_{0.025}$，$U = z_{0.025}$



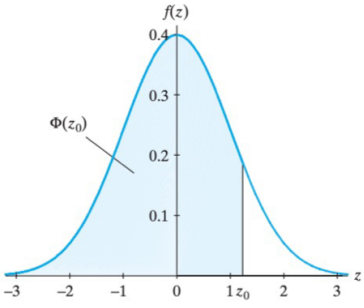
let $z_{\alpha/2}$ be the number such that the area under the **standard normal density function** to the right of $z_{\alpha/2}$ is $\alpha/2$.

Then if Z~N(0,1)

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$$

* 正态分布表

Normal Table



$$P(Z \le z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw$$

$$\Phi(-z) = 1 - \Phi(z)$$

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	0.7703	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	0.9925	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	0.9959	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
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

2、步骤二：转化为要求的区间

$1-\alpha$ Confidence interval：$[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$

3. 分析

- 样本容量↑，CI长度↓
- 方差σ^2↑，CI长度↑
- confidence level $(1-\alpha)$↑，CI长度↑

例：

- n patients use the new drug, whether the drug can cure the disease is a Bernoulli RV
- We have data $X_1, X_2, \dots, X_n$ that are sampled from this Bernoulli distribution with unknown cure rate p to be estimated
- Clearly, the mean of Bernoulli(p) is $\boldsymbol{p}$
- We can easily give a point estimate: $\boldsymbol{\hat{p} = \bar{X}}$ (sample cure rate)
- How to get an interval estimate??
  - Use Central Limit Theorem!

- $Z = \frac{\sqrt{n}\cdot(\hat{p}-p)}{\sqrt{p(1-p)}} \sim N(0,1)$
- $P\left(-z_{\frac{\alpha}{2}} \le \frac{\sqrt{n}\cdot(\hat{p}-p)}{\sqrt{p(1-p)}} \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$
- $P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$
- 用$\hat{p}$来近似p，得置信区间

  $\left[\hat{p} - z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

Experiments

Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

95% Confidence Interval

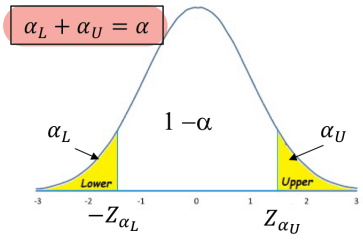
- Drug 1: $\hat{p}_1 = 90\%$. 10 experiments. [71.41%, 100%]
- Drug 2: $\hat{p}_2 = 80\%$. 10000 experiments. [79.22%, 80.78%]

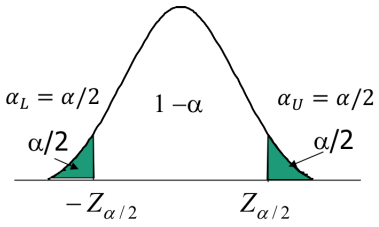
Now which drug do you think is more effective??

补：

* 区间的选择不一定要关于原点对称，只要满足区间以外的图像面积为α即可

The best interval?





$$P(-Z_{\alpha_L} \le \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le Z_{\alpha_U}) = 1-\alpha$$

$$\left[\bar{X} - Z_{\alpha_U}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha_L}\frac{\sigma}{\sqrt{n}}\right]$$

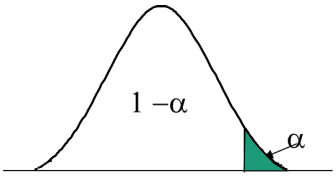
- Fix α, there are many choices of α_L and α_U to construct the interval.

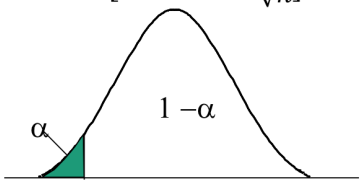
- We can also only consider **one-sided** confidence intervals!
- A **lower confidence bound** for μ

$$P(-\infty \le \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le Z_\alpha) = 1-\alpha \quad \Rightarrow \quad \left[\bar{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}, \ +\infty\right]$$

- A **upper confidence bound** for μ

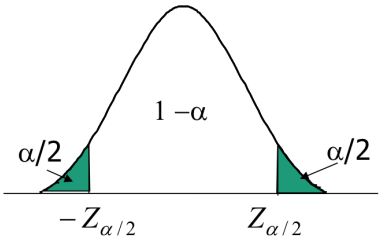
$$P(-Z_\alpha \le \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le +\infty) = 1-\alpha \quad \Rightarrow \quad \left[-\infty, \bar{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$$





* 最短区间

- Does there exists the shortest interval?



$$[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

The **shortest** one among all confidence intervals with confidence level $1-\alpha$

Why?
-Obvious from the shape of the normal density

* 正态分布的"3σ"原则