Distribution

1. Bernoulli $X \sim \text{Bernoulli}(p)$

$P(X=1) = p$, success

$\mu = p$, $\sigma^2 = p(1-p)$

2. Binomial $X \sim \text{Bin}(n, p)$

n - number of trials

p - success probability

$\mu = np$, $\sigma^2 = np(1-p)$

$$P(X=x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

- Uncertainty in sample proportion $\hat{p} = \frac{x}{n}$

$$\begin{cases} \mu_{\hat{p}} = p, \text{ unbiased} \\ \sigma_{\hat{p}} = \frac{\sigma_x}{n} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}} \end{cases}$$

3. Poisson $X \sim \text{poisson}(\lambda)$

an approximation of Binomial when n 大 p 小

$\lambda = np$, $P(X=x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$

$\mu = \lambda$, $\sigma_x^2 = \lambda$

λ: 浓度 conc X 个体数，单位 unit 内的量；

number of events occur in one unit of time / space

- A rate estimate:

  if $X \sim \text{Poisson}(\lambda t)$, $\hat{\lambda} = \frac{X}{t}$

  Uncertainty in $\hat{\lambda}$

  $\mu_{\hat{\lambda}} = \lambda$, unbiased

  $\sigma_{\hat{\lambda}} = \frac{\sigma_X}{t} = \frac{\sqrt{\lambda t}}{t} = \sqrt{\frac{\lambda}{t}}$

4. Hypergeometric $X \sim H(N, R, n)$

   N: size of finite population.

   R: number of items classified as success

   n: sample size

   X: number of success in the sample

$$P(X=x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}$$

$$\mu = \frac{nR}{N},$$

$$\sigma^2 = n\left(\frac{R}{N}\right)\left(1-\frac{R}{N}\right)\left(\frac{N-n}{N-1}\right)$$

5. Geometric $X \sim \text{Geom}(p)$.

   X: number of trials up to and including the 1st success

$$P(X=x) = p(1-p)^{x-1}$$

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}$$

6. Negative Binomial $X \sim NB(r, p)$

   X: number of trials up to and including rth success

r = rth success

$$P(X=x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

$$\mu = \frac{r}{p}, \quad \sigma^2 = \frac{r(1-p)}{p^2}$$

$X \sim NB(r,p) \iff X = Y_1 + \cdots + Y_r$, $Y \sim Geom(p)$

7. Multinomial $\quad X_1, \cdots, X_k \sim MN(n, p_1, p_2, \cdots, p_k)$

$$p(x_1, \cdots, x_k) = P(X=x_1, X=x_2, \cdots, X_k=x_k) = \frac{n!}{x_1! x_2! x_3! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

each $X_i \sim Bin(n_i, p_i)$

8. Normal $\quad X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$68\% = \mu \pm \sigma$

$95\% = \mu \pm 2\sigma$

$99.7\% = \mu \pm 3\sigma$

z-score:

$$z = \frac{x-\mu}{\sigma} \Rightarrow \text{standard normal}$$

sample mean $\bar{X}$:

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

Linear Function:

① $X \sim N(\mu, \sigma^2) \Rightarrow aX+b \sim N(a\mu+b, a^2\sigma^2)$

② $c_1X_1 + c_2X_2 \sim N(c_1\mu_1 + c_2\mu_2, c_1^2\sigma_1^2 + c_2^2\sigma_2^2)$

⑤ $X+Y \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$

$X-Y \sim N(\mu_X-\mu_Y, \sigma_X^2+\sigma_Y^2)$

9. Lognormal

$Y=e^X$, $X \sim N(\mu, \sigma^2)$; $\ln Y \sim N(\mu, \sigma^2)$

$\mu_Y = e^{\mu+\frac{1}{2}\sigma^2}$, $\sigma_Y^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$

10. Exponential $X \sim Expon(\lambda)$

waiting time

$\mu = \frac{1}{\lambda}$, $\sigma^2 = \frac{1}{\lambda^2}$

$f(x) = \lambda e^{-\lambda x}$

$F(x) = P(X \le x) = 1 - e^{-\lambda x}$

- Rate parameter λ from Poisson & $Expon(\lambda)$ (T)
- Memoryless property.

$P(T > t+s \mid T > s) = P(T > t)$

- Uncertainty:

$\mu_{\frac{1}{X}} \neq \frac{1}{\mu}$, bc $\frac{1}{\mu}$ is not a linear function of μ.

① $\hat{\lambda} = \frac{1}{\bar{x}}$, uncertainty $= \frac{\lambda}{n}$

② $\sigma_{\hat{\lambda}} = \frac{1}{\bar{x}\sqrt{n}}$ $n > 20$ 时，estimate 准确

Point Estimation

① Measure the goodness : mean squared error (MSE)

$\hat{\theta}$: estimator

$\mu_{\hat{\theta}} - \theta$: bias of estimator

$\sigma_{\hat{\theta}}$: uncertainty = standard error.

$$MSE_{\hat{\theta}} = (\mu_{\hat{\theta}} - \theta)^2 + \sigma^2_{\hat{\theta}} = \mu_{(\hat{\theta}-\theta)^2}$$

$$(\text{bias}^2 + \text{variance})$$

② Maximum Likelihood Estimate (MLE)

$n \uparrow$, bias of MLE $\to 0$; variance of MLE $\to$ Min

Central Limit Theorem

$X \sim N(\mu, \sigma)$

$\bar{X}$ — sample mean $\Rightarrow$ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

S — sample sum $S_n \sim N(n\mu, n\sigma^2)$

Confidence Intervals

probability ↔ random events ↔ computing method.

confidence ↔ fixed numerical value

1. Large sample CI for population mean

$n \geq 30$, $X \sim N(\mu, \sigma^2)$, 则 $\bar{X}$ 近似 normal

$100(1-\alpha)\%$ CI for μ:

- Two-sided:

$$\bar{X} \pm z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{X}}, \text{ 其中 } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

- One-sided:

lower CI: $\bar{X} - z_\alpha \cdot \sigma_{\bar{X}}$

upper CI: $\bar{X} + z_\alpha \cdot \sigma_{\bar{X}}$

2. CI for proportions

由 CLT 可知, Bin 的正态近似下, $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

已知 $X \sim Bin(n, p)$.

令 $\tilde{n} = n+4$, $\tilde{p} = \frac{X+2}{\tilde{n}}$, 则 $100(1-\alpha)\%$ CI for p:

- Two-sided $\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$ $[0,1]$

- One-sided $\underset{(\text{upper})}{\text{lower}}\ \tilde{p} \underset{(+)}{-} z_\alpha \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$

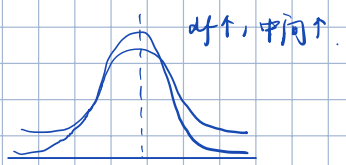
- Traditional method: (要求 succ & fail 均 >10)

  令 $\hat{p}$ — Large number n 次 trials 中的成功次数.

  则 CI for p 为: $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

3. Small sample CI for population mean μ

$n<30$, $df = n-1$, $\left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right) \sim t_{n-1}$



要求: ① population 是 normal ② 无 outlier

则: 100(1-α)% CI for μ:

- Two-sided: $\bar{x} \pm t_{n-1, \frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$
- One-sided: lower $\bar{x} - t_{n-1,\alpha} \cdot \frac{s}{\sqrt{n}}$ (upper) (+)

- If σ 已知, 用 z-distribution 不用 t: $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
- Single value $n=1$ 时, $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \sigma$

4. CI for difference between 2 means

已知 $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, 则

100(1-α)% CI for $(\mu_X - \mu_Y)$ 是:

$$(\bar{X} - \bar{Y}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \quad (\sigma_X = s_X, \sigma_Y = s_Y)$$

5. CI for difference between 2 proportions.

已知 $X \sim Bin(n_X, p_X)$, $Y \sim Bin(n_Y, p_Y)$

令 $\tilde{n}_X = n_X + 2$, $\tilde{n}_Y = n_Y + 2$, $\tilde{p}_X = \frac{X+1}{\tilde{n}_X}$, $\tilde{p}_Y = \frac{Y+1}{\tilde{n}_Y}$

则 $100(1-\alpha)\%$ CI for $(p_X - p_Y)$ 为:

$$(\tilde{p}_X - \tilde{p}_Y) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\tilde{p}_X(1-\tilde{p}_X)}{\tilde{n}_X} + \frac{\tilde{p}_Y(1-\tilde{p}_Y)}{\tilde{n}_Y}} \quad [-1, 1]$$

- Traditional method:

$$(\hat{p}_X - \hat{p}_Y) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}$$

(要求 suc & fail 均 >10)

6. Small sample CI for difference between 2 means

$100(1-\alpha)\%$ CI for $(\mu_X - \mu_Y)$:

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}}$$

, round down to nearest integer

$$\bar{X} - \bar{Y} \pm t_{\nu, \frac{\alpha}{2}} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

若 X 与 Y 的 σ^2 相同, 则:

$$\bar{X} - \bar{Y} \pm t_{n_X + n_Y - 2, \frac{\alpha}{2}} \cdot s_p \cdot \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} .$$

其中 $S_p = \sqrt{\frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{n_x+n_y-2}}$, pooled std.

7. CI with paired data.

$100(1-\alpha)\%$ CI for difference of pairs μ_D:

- small sample ($n < 30$)

$$\bar{D} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}}$$

- Large Sample

$$\bar{D} \pm z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{D}} = \bar{D} \pm z_{\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}}$$

Hypothesis Testing

Steps:

① Define H_0 和 H_1 ② Assume H_0 to be true

③ Compute a test statistic

(used to assess the strength of evidence against H_0)

④ Compute P-value (observed significance level)

(the prob that assuming H_0 is true, the test statistic would have a value whose disagreement with H_0 is as great or greater than the actually observed)

⑤ Conclusion

rule of thumb: $p \le 0.05$, significant

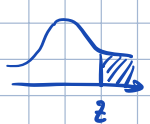
H_0 is rejected at 100α level

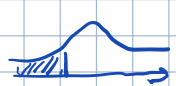
1. Large Sample, Population mean μ.

$n \ge 30$, z test,

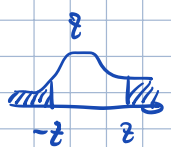
$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \rightarrow$$

p-value 对应于 H_1 那一侧

e.g. $H_1: \mu > \mu_0$, 则 p-value 

$\mu < \mu_0$, 

2. Population Proportion

$\mu \ne \mu_0$, 

$X \sim Bin(n, p)$. 当 np 和 $n(1-p)$ 均 > 10

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \rightarrow \text{p-value}.$$

3. Small Sample, population mean ($n<30$)

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}, \quad df=n-1.$$

- 当 σ 已知, 用 z test 而不是 t test: $z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

4. Large sample, Difference between 2 means

$H_0: \mu_X-\mu_Y \le \Delta_0$ ...

$$z=\frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_X}}}, \quad (\sigma_X=s_X,\ \sigma_Y=s_Y)$$

5. Difference between 2 proportions

$H_0: p_X-p_Y \le 0$

$$\hat{P}_X=\frac{X}{n_X}, \quad \hat{P}_Y=\frac{Y}{n_Y}, \quad \hat{P}=\frac{X+Y}{n_X+n_Y}.$$

$$z=\frac{\hat{P}_X-\hat{P}_Y}{\sqrt{\hat{P}(1-\hat{P})/\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}$$

6. Small sample, differeme between 2 means

$H_0: \mu_X-\mu_Y \le \Delta_0$

$$\nu = \frac{\left[\left(s_x^2/n_x\right) + \left(s_y^2/n_y\right)\right]^2}{\left[\left(s_x^2/n_x\right)^2/(n_x-1)\right] + \left[\left(s_y^2/n_y\right)^2/(n_y-1)\right]},$$

$$t = \frac{(\bar{x}-\bar{y}) - \Delta_0}{\sqrt{s_x^2/n_x + s_y^2/n_y}}$$

- Equal Variance:

$$s_p = \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}}$$

$$t = \frac{(\bar{x}-\bar{y})-\Delta_0}{s_p\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}, \quad df = n_x+n_y-2$$