

levels of Measure: RichardFang

1. Nominal: 名义尺度, 给变量赋值但值之间无关系

$$S = \frac{\text{Range}}{4} \quad \text{Mean: } E(x) \quad \text{variance: } \sigma^2 = E(x^2) - (E(x))^2$$

2. Ordinal: 顺序尺度, 值之间可比较

$$\text{Markov's inequality: } P(Y \geq a) \leq \frac{E(Y)}{a}$$

Continuous Uniform: $X \sim U(a, b)$

3. Interval: 等距尺度: 可以进行计算, 只有相对零点

$$P(|x - \mu| \geq k\sigma) = P((x - \mu)^2 \geq k^2\sigma^2) \leq \frac{E[(x - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad E(x) = \frac{a+b}{2} \quad \text{var}(x) = \frac{(b-a)^2}{12}$$

4. Ratio: 比率尺度: 具有绝对零点

$$\text{so } P(|x - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

$$\text{Exponential } x \sim \text{Exp}(\lambda) \quad E(x) = \frac{1}{\lambda} \quad \text{var}(x) = \frac{1}{\lambda^2}$$

simple random sample: 等可能抽 n 个, 不放回

$$z\text{-score: } z = \frac{x - \mu}{\sigma}$$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

systematic sample: 分组, 在第一组中 SRS, $k = \frac{N}{n}$, 每 k 步取一个

$$\text{Percentile} = \frac{(\# \text{ of values below } x) + 0.5}{\text{total } \# \text{ of values}} \cdot 100\%$$

stratified random sample: 分组, 每组内 SRS

$$C = \frac{n \cdot P}{100} \rightarrow \text{round up}$$

$$\text{Hypergeometric } (N, D, n) \quad P(X=x) = \frac{C_D^x \cdot C_{N-D}^{n-x}}{C_N^n}$$

cluster random sample: 分组, 选组

Decile + 分位数, Quartile 四分位数

$$E(x) = \frac{nD}{N} \quad \text{var}(x) = \frac{nD}{N} \left(1 - \frac{D}{N}\right) \left(\frac{N-n}{N-1}\right)$$

Lower/upper class limits: the smallest / largest possible data

$$\text{Interquartile Range IQR} = Q_3 - Q_1$$

$$\text{Poisson: } P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad e = 2.718281828$$

Class boundary: average of upper of one class and lower of next class

outlier: less than $Q_1 - 1.5 \text{ IQR}$ greater than $Q_3 + 1.5 \text{ IQR}$

$$E(x) = \lambda \quad D(x) = \lambda$$

classwidth: Range / # of class, round up

Boxplot: Low Q_1 , MD Q_3 High (Five number summary)

$$\text{if } n \geq 100, np \leq 10 \quad x \sim \text{Bin}(n, p) = x \cdot P(x)$$

classmidpoint: average upper and lower class limit

$$P(E) = P(A_1)P(E|A_1) + P(A_2)P(E|A_2) + \dots + P(A_n)P(E|A_n) \quad \lambda = np$$

histogram: height $\rightarrow$ frequencies, class boundaries $\rightarrow$ horizontal axis

$$\text{Bayes: } P(A_i|E) = \frac{P(E|A_i)P(A_i)}{P(E)}$$

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

frequency polygon: height $\rightarrow$ frequencies, class midpoints

$$\text{Discrete Uniform } P(X=n) = \frac{1}{N} \quad E(x) = \frac{1+p}{2} \quad \text{var}(x) = \frac{(1-p)^2}{12}$$

ogive cumulative frequencies for the class, upper class boundaries

$$E(x) = \frac{N+1}{2}, \quad \text{var}(x) = \frac{N^2-1}{12}$$

normal: unimodal 只有一个众数

relative frequency: 把上面的纵坐标换成此

$$\text{Bernoulli: } E(x) = p \quad \text{var}(x) = (1-p)p$$

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

Bell shape uniform J-shape Reverse J

Right skewed left-skew Bimodal U-shape

frequency midpoint

$$\bar{x} = \frac{\sum f \cdot x_m}{n} \quad \bar{x} = \arg \min \frac{1}{n} \sum_{i=1}^n (x_i - a)^2$$

$$\text{Binomial } (n, p) \quad P(x) = C_n^x p^x (1-p)^{n-x}$$

$$E(x) = np \quad \text{var}(x) = np(1-p)$$

$$\text{Pearson index: } PI = \frac{3(\bar{x} - \text{MD})}{s}$$

$-1 \leq PI \leq 1 \Rightarrow$ 分布 is not significantly skewed

Median = $\arg \min \frac{1}{n} \sum_{i=1}^n |x_i - a|$

$$\text{Negative Binomial: } P(X=x) = C_{n-1}^x (1-p)^x p^{n-x} \quad \text{Multinomial distribution:}$$

$$\text{median} = \arg \min \frac{1}{n} \sum_{i=1}^n |x_i - a| \quad \text{midrange} = \frac{\text{Lowest} + \text{Highest}}{2}$$

$$E(x) = \frac{r(1-p)}{p} \quad \text{var}(x) = \frac{r(1-p)}{p^2} \quad P(X) = \frac{n!}{x_1! x_2! \dots x_r!} p_1^{x_1} p_2^{x_2} \dots p_r^{x_r}$$

$$\text{range} = \text{Highest} - \text{Lowest} \quad \text{coefficient of variation} = \frac{\text{CVAR}}{\bar{x}} \cdot 100\%$$

样方差 $S^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$ Z值代表 $X \sim \chi^2$ 的分布 Central Limit Theorem: sample distribution of sample means $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ sample error $\bar{x} - \mu$ $E[\bar{x}] = E[\frac{1}{n} \sum_{i=1}^n x_i] = \frac{1}{n} \sum_{i=1}^n E[x_i] = \mu$ $Var[\bar{x}] = Var[\frac{1}{n} \sum_{i=1}^n x_i] = \frac{1}{n^2} Var[\sum_{i=1}^n x_i] = \frac{1}{n} \sigma^2$ $\mu_{\bar{x}} = \mu$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ $P(x_1, \dots, x_n) = \sum P(x_i = k) P(x_i = n-k)$ continuous: $f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(u) f_Y(t-u) du$ $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ correction factor: $\frac{\sqrt{n-1}}{\sqrt{n}}$ 修正偏差 for $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \frac{\sqrt{n-1}}{\sqrt{n}}}$ if $np \geq 5$ $nq \geq 5$ $X \sim Bin(n, p) \Rightarrow X \sim NC(n, p, (\frac{1}{np}))^2$ Binomial Normal $P(X=a)$ $P(a-0.5 < X < a+0.5)$ $P(X \geq a)$ $P(X > a-0.5)$ $P(X > a)$ $P(X > a+0.5)$ $P(X \leq a)$ $P(X < a+0.5)$ $P(X < a)$ $P(X < a-0.5)$ consistent: 测量量增加 估计量收敛于参数 unbiased: 估计量的期望为参数 置信区间: Given $\alpha \in (0,1)$, the confidence interval of sample mean with confidence level $1-\alpha$ $\mu \in (\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) \Leftrightarrow P(-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}) = 1-\alpha$ Minimum sample size Needed for an Interval Estimate of μ: $n = (\frac{Z_{\frac{\alpha}{2}} \sigma}{E})^2$ E is margin error	卡方分布: $X_1, \dots, X_n \sim N(0,1)$ $Q = \sum_{i=1}^n X_i^2 \sim \chi^2(n)$ 自由度为 n $Q, X_1, \dots, X_n \sim N(0,1)$ F 分布: $U \sim \chi^2(n_1)$ $V \sim \chi^2(n_2)$ U, V 独立 $F = \frac{\frac{U}{n_1}}{\frac{V}{n_2}}$ 服从自由度为 (n_1, n_2) F 分布 $Q_1 \sim \chi^2(n_1)$ $Q_2 \sim \chi^2(n_2)$ Q_1, Q_2 相互独立, 则 $Q_1 + Q_2 \sim \chi^2(n_1+n_2)$ $F \sim F(n_1, n_2)$ $\frac{1}{F} \sim F(n_2, n_1)$ 定理一: 设 $X_1, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本 $D(X^*) = \sum D(X_i) = n[E(X^*) - E(X^*)^2] = n[\int_{-\infty}^{\infty} x^* \frac{1}{\sigma} e^{-\frac{x^{*2}}{2\sigma^2}} dx - 1] = n[1] = n$ $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ $\bar{x}$ 与 S^2 相互独立 定理二: 设 $X_1, \dots, X_n$ 来自 $N(\mu_1, \sigma_1^2)$ 的样本 $Y_1, \dots, Y_n$ 来自 $N(\mu_2, \sigma_2^2)$, 两样本相互独立 $S_1^2 = \frac{1}{n_1-1} \sum (x_i - \bar{x})^2$ $S_2^2 = \frac{1}{n_2-1} \sum (y_i - \bar{y})^2$ (1) $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$ (2) 当 $\sigma_1^2 = \sigma_2^2$ $\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2 + S_2^2}{2} \frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$ $p =$ 目标在总体中的比例 $\hat{p}$ 样本中目标占比 $\hat{p} = \frac{x}{n}$ and $\hat{q} = \frac{n-x}{n}$ 则 $p \in (\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}})$ when $np \geq 5$ and $nq \geq 5$, $n = \hat{p}\hat{q} (\frac{Z_{\frac{\alpha}{2}}}{E})^2$ 1. 均值 μ 的置信区间 σ^2 已知: $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ $\mu \in (\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$ σ^2 未知: $E(S^2) = \sigma^2 \Rightarrow \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t(n-1) \Rightarrow \mu \in (\bar{x} - Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}})$ $\sigma_1^2 = \sigma_2^2 = \sigma^2$, σ^2 未知 2. σ^2 的置信区间 $\frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2 + S_2^2}{2} \frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$ $E(S^2) = \sigma^2 \Rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ $\sigma^2 \in (\frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2}, \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2})$ $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$
--	--

Hypothesis Test

example:

$H_0: \mu = 82$ $H_1: \mu \neq 82$ two tailed

$H_0: \mu = 36$ $H_1: \mu > 36$ right-tailed

$H_0: \mu = 78$ $H_1: \mu < 78$ left-tailed

$P(\text{type I error}) = \alpha$ H_0 is correct but reject

α is the level of significance $P(\text{type II error}) = \beta$

right-tail 看上 α 位置 left-tail 看下 α 位置

Z-Test $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

solutions for z-test

① step 1: $H_0: \mu = 29$ and $H_1: \mu > 29$ (claim)

step 2: $\alpha = 0.05$ and due to right-tailed test $Z = 1.65$

step 3: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{30.1 - 29}{2.81 / \sqrt{50}} = 1.59$

step 4: $Z < Z_0$ so we don't reject H_0

② step 1: $H_0: \mu = 80$ and $H_1: \mu < 80$

step 2: $\alpha = 0.1$ due to left-tailed test $Z_0 = -1.28$

step 3: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = -1.56$

step 4: $-1.56 < -1.28$ so we reject the H_0

③ step 1: $H_0: \mu = 24672$ and $H_1: \mu \neq 24672$

step 2: $\alpha = 0.01$ and two-tailed test, $Z_0 = \pm 2.58$

step 3: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 3.04$

step 4: $3.04 \notin (-2.58, 2.58)$ reject the H_0

p-value

③ step 1: $H_0: \mu = 5700$ and $H_1: \mu > 5700$ (claim) if P-value ≤ 0.01 , reject H_0

step 2: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 2.28$ 证据强烈支持 H_1

step 3: $P(z) = 0.9887$ due to right-tailed if P-value > 0.01 but $P \leq 0.05$

$P = 1 - P(z) = 0.0113$ reject H_0 证据支持 H_1

step 4: $P < \alpha = 0.05$ so we reject H_0 if P-value > 0.05 but $P \leq 0.1$

④ step 1: $H_0: \mu = 8$ (claim) $H_1: \mu \neq 8$ 可能有第一类错误

step 2: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 1.89$ if P-value > 0.1 不要拒绝 H_0

step 3: $P(z) = 0.9706$ 若 σ 已知选 Z-test

$P = 2 \times (1 - P(z)) = 0.0588$ 若 σ 未知选 t-test

due to two-tailed, P should $\times 2$ 相关系数 $r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$

step 4: $P > 0.05$ so we do not reject H_0 t-test for r: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

solutions for t-test $t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ 若 σ 未知自由度为 $n-1$ step 1: $H_0: \mu = 0$ and $H_1: \mu \neq 0$

① step 1: $H_0: \mu = 60$ $H_1: \mu < 60$ (claim) step 2: CV: ± 2.776

step 2: $\alpha = 0.1$ and $d.f. = 7$ the critical value is $-1.415 = t_0$ step 3: $t = r \sqrt{\frac{n-2}{1-r^2}} = 10.4$

step 3: $\bar{x} = 58.88$ and $s = 5.08$ step 4: reject the H_0

$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = -0.624$ $y = (y' \pm t_{\alpha/2} \text{se } \sqrt{1/n + \frac{(\bar{x} - \bar{x})^2}{n \sum x^2 - (\sum x)^2}})$ $d.f. = n - 2$

step 4: $t > t_0$, do not reject H_0

② step 1: $H_0: \mu = 36.7$ $H_1: \mu > 36.7$ (claim) so $P \in (0.01, 0.025)$

step 2: $t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = 2.517$ step 4: $P < 0.05$

step 3: Find P 2.517 falls between 2.145 reject the H_0

and 2.624 corresponding to $\alpha = 0.025$ and $\alpha = 0.01$

z-test for proportion

np ≥ 5 nq ≥ 5

$\hat{p} = \frac{P}{n}$
 $Z = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}}$

step 1: $H_0: p = 0.6$ (claim) $H_1: p \neq 0.6$
 step 2: $\alpha = 0.05 \Rightarrow Z_0 = \pm 1.96$
 step 3: $\hat{p} = \frac{x}{n} = 0.64 \Rightarrow Z = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}} = 1.15$
 step 4: $Z \notin (-Z_0, Z_0)$ so do not reject H_0

from two independent

z-test for comparing two means

$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

step 1: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
 step 2: cv $Z_0 = \pm 1.96$
 step 3: $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 7.45$
 step 4: $7.45 \notin (-1.96, 1.96)$ reject H_0

Test between Proportion

$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\hat{P}\hat{Q}(\frac{1}{n_1} + \frac{1}{n_2})}}$ $\bar{P} = \frac{x_1 + x_2}{n_1 + n_2}$ $\hat{P}_1 = \frac{x_1}{n_1}$

step 1: $H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 \neq 0$
 step 2: $\alpha = 0.05 \Rightarrow Z_0 = \pm 1.96$
 step 3: $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\hat{P}\hat{Q}(\frac{1}{n_1} + \frac{1}{n_2})}} = -2.7$
 step 4: reject H_0

Test for a variance or a standard deviation

$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ 查表用 $n-1$
 step 1: $H_0: \sigma^2 = 225$ $H_1: \sigma^2 > 225$ (claim)
 step 2: The cv $\chi^2_{0.05} = 12.338$
 step 3: $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = 19.36$
 step 4: $19.36 > Z_0$, do not reject H_0

$F = \frac{s_1^2}{s_2^2}$ 查表用 $d.f. (n_1-1, n_2-1)$
 step 1: $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$
 step 2: cv $F_0 = 3.31$
 step 3: $F = \frac{s_1^2}{s_2^2} = 1.31$
 step 4: do not reject H_0

step 1: $H_0: \sigma^2 = 0.644$ (claim) $H_1: \sigma^2 \neq 0.644$
 step 2: cv $\chi^2 = 32.852$ and 8.907
 step 3: $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = 29.5$
 step 4: $29.5 \notin (8.907, 32.852)$ do not reject H_0

step 1: $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$
 step 2: cv $F_0 = 3.31$
 step 3: $F = \frac{s_1^2}{s_2^2} = 1.31$
 step 4: do not reject H_0

step 1: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
 step 2: cv $t_0 = 12.36$
 step 3: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -0.57$
 step 4: do not reject H_0

step 1: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
 step 2: cv $t_0 = 12.36$
 step 3: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -0.57$
 step 4: do not reject H_0

step 1: $H_0: \mu = 5$ and $H_1: \mu \neq 5$
 step 2: cv $t = \pm 2.010$
 step 3: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = -4.04$
 step 4: $t > t_0$ so reject H_0
 step 5: $\mu \notin (\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$

step 1: $H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$
 step 2: cv $t = 12.015$...

step 1: $H_0: \mu = 5$ and $H_1: \mu \neq 5$
 step 2: cv $t = \pm 2.010$
 step 3: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = -4.04$
 step 4: $t > t_0$ so reject H_0
 step 5: $\mu \notin (\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$

step 1: $H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$
 step 2: cv $t = 12.015$...