TAMS24: Notations and Formulas

— by Xiangfeng Yang

1 Basic notations and definitions

X: random variable (stokastiska variabel);

Mean (Väntevärde):

$$\mu = E(X) = \begin{cases} \sum k p_X(k), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x f_X(x)dx, & \text{if } X \text{ is continuous;} \end{cases}$$

Variance (Varians): $\sigma^2 = V(X) = E((X-\mu)^2) = E(X^2) - (E(X))^2$;

Standard deviation (Standardavvikelse): $\sigma = D(X) = \sqrt{V(X)}$;

Population X;

Random sample (slumpmässigt stickprov): $X_1, \ldots, X_n$ are independent and have the same distribution as the population X. Before observe/measure, $X_1, \ldots, X_n$ are random variables, and after observe/measure, we use $x_1, \ldots, x_n$ which are numbers (not random variables);

Sample mean (Stickprovsmedelvärde): Before observe/measure, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, and after observe/measure, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$;

Sample variance (Stickprovsvarians): Before observe/measure, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$, and after observe/measure, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$;

Sample standard deviation (Stickprovsstandardavvikelse): Before observe/measure, $S = \sqrt{S^2}$, and after observe/measure, $s = \sqrt{s^2}$;

$$E(\sum_{i=1}^n c_i X_i) = \sum_{i=1}^n c_i E(X_i),$$
$$V(\sum_{i=1}^n c_i X_i) = \sum_{i=1}^n c_i^2 V(X_i), \text{ if } X_1, \ldots, X_n \text{ are independent (oberoende);}$$

If $X \sim N(\mu, \sigma)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$;

If $X_1, \ldots, X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i)$, then

$$d + \sum_{i=1}^n c_i X_i \sim N(d + \sum_{i=1}^n c_i \mu_i, \sqrt{\sum_{i=1}^n c_i^2 \sigma_i^2});$$

For a population X with an unknown parameter θ, and a random sample $\{X_1, \ldots, X_n\}$:

Estimator (Stickprovsvariabeln): $\hat{\Theta} = g(X_1, \ldots, X_n)$, a random variable;

Estimate (Punktskattning): $\hat{\theta} = g(x_1, \ldots, x_n)$, a number;

Unbiased (Väntevärdesriktig): $E(\hat{\Theta}) = \theta$;

Effective (Effektiv): Two estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are unbiased, we say that $\hat{\Theta}_1$ is more effective than $\hat{\Theta}_2$ if $V(\hat{\Theta}_1) < V(\hat{\Theta}_2)$;

Binomial distribution $X \sim Bin(N,p)$: there are N independent and identical trials, each trial has a probability of success p, and X = the number of successes in these N trials. The random variable $X \sim Bin(N,p)$ has a probability function (sannolikhetsfunktion)

$$p(k) = P(X=k) = \binom{N}{k} p^k (1-p)^{N-k};$$

Exponential distribution $X \sim Exp(1/\mu)$: when we consider the waiting time/lifetime... The random variable $X \sim Exp(1/\mu)$ has a density function (täthetsfunktion)

$$f(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geq 0.$$

2 Point estimation

Method of moments (Momentmetoden): # of equations depends on # of unknown parameters,

$$E(X) = \bar{x}, \quad E(X^2) = \frac{1}{n}\sum_{i=1}^n x_i^2, \quad E(X^3) = \frac{1}{n}\sum_{i=1}^n x_i^3, \quad \ldots\ldots$$

Consistent (Konsistent): An estimator $\hat{\Theta} = g(X_1, \ldots, X_n)$ is consistent if

$$\lim_{n\to\infty} P(|\hat{\Theta} - \theta| > \varepsilon) = 0, \text{ for any constant } \varepsilon > 0.$$

(This is called "*convergence in probability*").

Theorem: If $E(\hat{\Theta}) = \theta$ and $\lim_{n\to\infty} V(\hat{\Theta}) = 0$, then $\hat{\Theta}$ is consistent.

Least square method (minsta-kvadrat-metoden): The least square estimate $\hat{\theta}$ is the one minimizing

$$Q(\theta) = \sum_{i=1}^n (x_i - E(X))^2.$$

Maximum-likelihood method (Maximum-likelihood-metoden): The maximum-likelihood estimate $\hat{\theta}$ is the one maximizing the likelihood function

$$L(\theta) = \begin{cases} \prod_{i=1}^n f(x_i;\theta), & \text{if } X \text{ is continuous,} \\ \prod_{i=1}^n p(x_i;\theta), & \text{if } X \text{ is discrete.} \end{cases}$$

Remark 1 on ML: In general, it is easier/better to maximize $\ln L(\theta)$;

Remark 2 on ML: If there are several random samples (say m) from different populations with a same unknown parameter θ, then the maximum-likelihood estimate $\hat{\theta}$ is the one maximizing the likelihood function defined as $L(\theta) = L_1(\theta) \ldots L_m(\theta)$, where $L_i(\theta)$ is the likelihood function from the i-th population.

Estimates of population variance σ^2: If there is only one population with an unknown mean, then method of moments and maximum-likelihood method, in general, give an estimate of σ^2 as follows

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{(NOT unbiased)}.$$

An adjusted (or corrected) estimate would be the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{(unbiased)}.$$

If there are m different populations with unknown means and a same variance σ^2, then an adjusted (or corrected) ML estimate is

$$s^2 = \frac{(n_1-1)s_1^2 + \ldots + (n_m-1)s_m^2}{(n_1-1)+\ldots+(n_m-1)} \qquad \text{(unbiased)}$$

where n_i is the sample size of the i-th population, and s_i^2 is the sample variance of the i-th population.

Standard error (medelfelet) of an estimator $\hat{\Theta}$: $\sim$ is an estimate of the standard deviation $D(\hat{\Theta})$.

3 Interval estimation

One sample $\{X_1, \ldots, X_n\}$ from $N(\mu, \sigma)$

$$I_\mu = \begin{cases} \bar{x} \mp \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}, & \text{if } \sigma \text{ is known; } \left[\text{fact } \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\right] \\ \bar{x} \mp t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, & \text{if } \sigma \text{ is unknown; } \left[\text{fact } \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)\right] \end{cases}$$

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \quad \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right); \left[\text{fact } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)\right]$$

Unknown σ^2 can be estimated by the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Two samples $\{X_1, \ldots, X_{n_1}\}$ from $N(\mu_1, \sigma_1)$; $\{Y_1, \ldots, Y_{n_2}\}$ from $N(\mu_2, \sigma_2)$; $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ are independent

$$I_{\mu_1-\mu_2} = \begin{cases} (\bar{x}-\bar{y}) \mp \lambda_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \text{ if } \sigma_1 \text{ and } \sigma_2 \text{ are known;} \\ \qquad \left[\text{fact } \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)\right] \\ (\bar{x}-\bar{y}) \mp t_{\alpha/2}(n_1+n_2-2)\cdot s \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \text{ if } \sigma_1 = \sigma_2 = \sigma \text{ is unknown;} \\ \qquad \left[\text{fact } \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)\right] \\ \approx (\bar{x}-\bar{y}) \mp t_{\alpha/2}(f)\cdot\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, \text{ if } \sigma_1 \neq \sigma_2 \text{ both are unknown;} \\ \qquad \left[\text{fact } \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \approx t(f)\right] \\ \qquad \left[\text{degrees of freedom } f = \frac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}\right] \end{cases}$$

$$I_{\sigma^2} = \left(\frac{(n_1+n_2-2)s^2}{\chi^2_{\frac{\alpha}{2}}(n_1+n_2-2)}, \quad \frac{(n_1+n_2-2)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n_1+n_2-2)}\right), \text{ if } \sigma_1 = \sigma_2 = \sigma;$$

$$\left[\text{fact } \frac{(n_1+n_2-2)S^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)\right]$$

Unknown σ^2 can be estimated by the samples variance $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

m samples: The unknown $\sigma_1^2 = \ldots = \sigma_m^2 = \sigma^2$ can be estimated by $s^2 = \frac{(n_1-1)s_1^2+\ldots+(n_m-1)s_m^2}{(n_1-1)+\ldots+(n_m-1)}$.

Remark: The idea of using fact (**sampling distribution**) to find confidence intervals is very important. There are a lot more different confidence intervals besides above. For instance, we consider two independent samples: $\{X_1, \ldots, X_{n_1}\}$ from $N(\mu_1, \sigma)$ and $\{Y_1, \ldots, Y_{n_2}\}$ from $N(\mu_2, \sigma)$. In this case, we can easily prove that

$$c_1\bar{X} + c_2\bar{Y} \sim N\left(c_1\mu_1 + c_2\mu_2, \quad \sigma\sqrt{\frac{c_1^2}{n_1}+\frac{c_2^2}{n_2}}\right).$$

- If σ is known, then fact $\frac{(c_1\bar{X}+c_2\bar{Y})-(c_1\mu_1+c_2\mu_2)}{\sigma\sqrt{\frac{c_1^2}{n_1}+\frac{c_2^2}{n_2}}} \sim N(0,1)$. So we can find $I_{c_1\mu_1+c_2\mu_2}$;
- If σ is unknown, then fact $\frac{(c_1\bar{X}+c_2\bar{Y})-(c_1\mu_1+c_2\mu_2)}{S\sqrt{\frac{c_1^2}{n_1}+\frac{c_2^2}{n_2}}} \sim t(n_1+n_2-2)$. So we can find $I_{c_1\mu_1+c_2\mu_2}$.

3.1 ***Confidence intervals from normal approximations.***

$$X \sim Bin(N,p): \; I_p = \hat{p} \mp \lambda_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}, \text{ fact } \frac{\hat{P}-p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{N}}} \approx N(0,1).$$

(we require that $N\hat{p}(1-\hat{p}) > 10$)

$$X \sim Hyp(N,n,p): \; I_p = \hat{p} \mp \lambda_{\alpha/2}\sqrt{\frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\hat{p}(1-\hat{p})}, \text{ fact } \frac{\hat{P}-p}{\sqrt{\frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\hat{P}(1-\hat{P})}} \approx N(0,1).$$

$$X \sim Po(\mu): \; I_\mu = \bar{x} \mp \lambda_{\alpha/2}\sqrt{\frac{\bar{x}}{n}}, \text{ fact } \frac{\bar{X}-\mu}{\sqrt{\frac{\mu}{n}}} \approx N(0,1).$$

(we require that $n\bar{x} > 15$)

$$X \sim Exp(\frac{1}{\mu}): \; \bullet\, I_\mu = \left(\frac{\bar{x}}{1+\frac{\lambda_{\alpha/2}}{\sqrt{n}}}, \frac{\bar{x}}{1-\frac{\lambda_{\alpha/2}}{\sqrt{n}}}\right), \text{ fact } \frac{\bar{X}-\mu}{\mu/\sqrt{n}} \approx N(0,1),$$

$$\bullet\, I_\mu = \bar{x} \mp \lambda_{\alpha/2}\frac{\bar{x}}{\sqrt{n}}, \text{ fact } \frac{\bar{X}-\mu}{\bar{X}/\sqrt{n}} \approx N(0,1).$$

(we require that $n \geq 30$)

Remark: Again there are more confidence intervals besides above. For instance, we consider two independent samples: X from $Bin(N_1, p_1)$ and Y from $Bin(N_2, p_2)$, with unknown p_1 and p_2. As we know

$$\hat{P}_1 \approx N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \text{ and } \hat{P}_2 \approx N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right),$$

so $\hat{P}_1 - \hat{P}_2 \approx N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}\right)$. Therefore, fact is $\frac{(\hat{P}_1-\hat{P}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1}+\frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}} \approx N(0,1)$,

$$I_{p_1-p_2} = (\hat{p}_1 - \hat{p}_2) \mp \lambda_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

3.2 ***Confidence intervals from the ratio of two population variances.***

Suppose there are two independent samples $\{X_1, \ldots, X_{n_1}\}$ from $N(\mu_1, \sigma_1)$, and $\{Y_1, \ldots, Y_{n_2}\}$ from $N(\mu_2, \sigma_2)$. Then $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$ and $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$, therefore

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1), \quad \text{fact.}$$

Thus

$$I_{\sigma_2^2/\sigma_1^2} = \left(\frac{s_2^2}{s_1^2} \cdot F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1), \quad \frac{s_2^2}{s_1^2} \cdot F_{\frac{\alpha}{2}}(n_1-1, n_2-1) \right).$$

3.3 ***Large sample size ($n \geq 30$, population may be completely unknown).***

If there is no information about the population(s), then we can apply Central Limit Theorem (usually with a large sample $n \geq 30$) to get an approximated normal distributions. Here are two examples:

Example 1: Let $\{X_1, \ldots, X_n\}, n \geq 30$, be a random sample from a population, then (no matter what distribution the population is)

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx N(0,1).$$

Example 2: Let $\{X_1, \ldots, X_{n_1}\}, n_1 \geq 30$, be a random sample from a population, and $\{Y_1, \ldots, Y_{n_2}\}, n_2 \geq 30$, be a random sample from another population which is independent from the first population, then (no matter what distributions the populations are)

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0,1).$$

4 Hypothesis testing

4.1 *One sample* and the *general theory* of hypothesis testing

Suppose there is a random sample $\{X_1, \ldots, X_n\}$ from a population X with an unknown parameter θ,

$$H_0: \ \theta = \theta_0 \qquad vs. \qquad H_1: \ \theta < \theta_0, \ \text{or } \theta > \theta_0, \ \text{or } \theta \neq \theta_0$$

	H_0 is true	H_0 is false and $\theta = \theta_1$
reject H_0	(type I error or significance level) α	(power) $h(\theta_1)$
don't reject H_0	$1 - \alpha$	(type II error) $\beta(\theta_1) = 1 - h(\theta_1)$

Regarding the *p-value*:

reject H_0 *if and only if* p-value $< \alpha$.

For notational simplicity, we employ

TS := "test statistic"; and C := "critical region".

reject H_0 *if* TS $\in$ C;

reject H_0 *if and only if* p-value $< \alpha$.

4.2 Hypothesis testing for population mean(s)

One sample: $\{X_1, \ldots, X_n\}$ from $N(\mu, \sigma)$. Null hypothesis $H_0 : \mu = \mu_0$.

$$\begin{cases} \begin{array}{c}\sigma \text{ is known:} \\ \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\end{array} \begin{cases} H_1: \mu < \mu_0: & \text{TS} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}, \ \text{C} = (-\infty, -\lambda_\alpha), \\ & p\text{-value} = P(N(0,1) \leq \text{TS}); \\ H_1: \mu > \mu_0: & \text{TS} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}, \ \text{C} = (\lambda_\alpha, +\infty), \\ & p\text{-value} = P(N(0,1) \geq \text{TS}); \\ H_1: \mu \neq \mu_0: & \text{TS} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}, \ \text{C} = (-\infty, -\lambda_{\alpha/2}) \cup (\lambda_{\alpha/2}, +\infty), \\ & p\text{-value} = 2P(N(0,1) \geq |\text{TS}|). \end{cases} \\ \begin{array}{c}\sigma \text{ is unknown:} \\ \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)\end{array} \begin{cases} H_1: \mu < \mu_0: & \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}, \ \text{C} = (-\infty, -t_\alpha(n-1)), \\ & p\text{-value} = P(t(n-1) \leq \text{TS}); \\ H_1: \mu > \mu_0: & \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}, \ \text{C} = (t_\alpha(n-1), +\infty), \\ & p\text{-value} = P(t(n-1) \geq \text{TS}); \\ H_1: \mu \neq \mu_0: & \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}, \ \text{C} = (-\infty, -t_{\alpha/2}(n-1)) \cup (t_{\alpha/2}(n-1), +\infty), \\ & p\text{-value} = 2P(t(n-1) \geq |\text{TS}|). \end{cases} \end{cases}$$

Two samples: $\{X_1, \ldots, X_{n_1}\}$ from $N(\mu_1, \sigma_1)$; $\{Y_1, \ldots, Y_{n_1}\}$ from $N(\mu_2, \sigma_2)$; Null hypothesis $H_0 : \mu_1 = \mu_2$.

$$\begin{cases} \begin{array}{c}\sigma_1, \sigma_2 \text{ are known:} \\ \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)\end{array} \begin{cases} H_1: \mu_1 < \mu_2: & \text{TS} = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \ \text{C} = (-\infty, -\lambda_\alpha), \\ & p\text{-value} = P(N(0,1) \leq \text{TS}); \\ H_1: \mu_1 > \mu_2: & \text{TS} = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \ \text{C} = (\lambda_\alpha, +\infty), \\ & p\text{-value} = P(N(0,1) \geq \text{TS}); \\ H_1: \mu_1 \neq \mu_2: & \text{TS} = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \ \text{C} = (-\infty, -\lambda_{\alpha/2}) \cup (\lambda_{\alpha/2}, +\infty), \\ & p\text{-value} = 2P(N(0,1) \geq |\text{TS}|). \end{cases} \\ \begin{array}{c}\sigma_1 = \sigma_2 \text{ is unknown:} \\ \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)\end{array} \begin{cases} H_1: \mu_1 < \mu_2: & \text{TS} = \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \ \text{C} = (-\infty, -t_\alpha(n_1+n_2-2)), \\ & p\text{-value} = P(t(n_1+n_2-2) \leq \text{TS}); \\ H_1: \mu_1 > \mu_2: & \text{TS} = \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \ \text{C} = (t_\alpha(n_1+n_2-2), +\infty), \\ & p\text{-value} = P(t(n_1+n_2-2) \geq \text{TS}); \\ H_1: \mu_1 \neq \mu_2: & \text{TS} = \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \ \text{C} = (-\infty, -t_{\alpha/2}(n_1+n_2-2)) \\ & \qquad \cup (t_{\alpha/2}(n_1+n_2-2), +\infty), \\ & p\text{-value} = 2P(t(n_1+n_2-2) \geq |\text{TS}|). \end{cases} \\ \sigma_1 \neq \sigma_2 \text{ both unknown: similarly as in the tree of confidence intervals.} \end{cases}$$

4.3 Hypothesis testing for population variance(s)

$$\begin{cases} \begin{matrix}\{X_1,\dots,X_{n_1}\} \text{ from } N(\mu,\sigma) \\ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \\ H_0 : \sigma^2 = \sigma_0^2\end{matrix} & \begin{cases} H_1 : \sigma^2 < \sigma_0^2 : \ \text{TS} = \frac{(n-1)s^2}{\sigma_0^2},\ \text{C} = (0, \chi^2_{1-\alpha}(n-1)), \\ \qquad p\text{-value} = P(\chi^2(n-1) \le \text{TS}); \\ H_1 : \sigma^2 > \sigma_0^2 : \ \text{TS} = \frac{(n-1)s^2}{\sigma_0^2},\ \text{C} = (\chi^2_{\alpha}(n-1), +\infty), \\ \qquad p\text{-value} = P(\chi^2(n-1) \ge \text{TS}); \\ H_1 : \sigma^2 \neq \sigma_0^2 : \ \text{TS} = \frac{(n-1)s^2}{\sigma_0^2},\ \text{C} = (0, \chi^2_{1-\frac{\alpha}{2}}(n-1)) \cup (\chi^2_{\frac{\alpha}{2}}(n-1), +\infty), \\ \qquad p\text{-value} = 2P(\chi^2(n-1) \ge \text{TS}) \text{ or } 2P(\chi^2(n-1) \le \text{TS}). \end{cases} \\ \begin{matrix}\{X_1,\dots,X_{n_1}\} \text{ from } N(\mu_1,\sigma_1) \\ \{Y_1,\dots,Y_{n_2}\} \text{ from } N(\mu_2,\sigma_2) \\ \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1) \\ H_0 : \sigma_1^2 = \sigma_2^2\end{matrix} & \begin{cases} H_1 : \sigma_1^2 < \sigma_2^2 : \ \text{TS} = s_1^2/s_2^2,\ \text{C} = (0, F_{1-\alpha}(n_1-1, n_2-1)), \\ \qquad p\text{-value} = P(F(n_1-1, n_2-1) \le \text{TS}); \\ H_1 : \sigma_1^2 > \sigma_2^2 : \ \text{TS} = s_1^2/s_2^2,\ \text{C} = (F_{\alpha}(n_1-1, n_2-1), +\infty), \\ \qquad p\text{-value} = P(F(n_1-1, n_2-1) \ge \text{TS}); \\ H_1 : \sigma_1^2 \neq \sigma_2^2 : \ \text{TS} = s_1^2/s_2^2,\ \text{C} = (0, F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)) \\ \qquad \cup\ (F_{\frac{\alpha}{2}}(n_1-1, n_2-1), +\infty), \\ \qquad p\text{-value} = 2P(F(n_1-1, n_2-1) \ge \text{TS}) \\ \qquad \text{or } 2P(F(n_1-1, n_2-1) \le \text{TS}). \end{cases} \end{cases}$$

4.4 Large sample size ($n \ge 30$, population may be completely unknown)

If there is no information about the population(s), then we can apply Central Limit Theorem (usually with a large sample $n \ge 30$). The idea is exactly the same as the one used in confidence intervals. **One example** is: a sample $\{X_1, \dots, X_n\}, n \ge 30$, from some population (which is unknown) with a mean μ and standard deviation σ. Null hypothesis $H_0 : \mu = \mu_0$. Then it follows from CLT that $\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0,1)$, therefore

$$\begin{cases} H_1 : \mu < \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}},\ \text{C} = (-\infty, -\lambda_\alpha), \\ \qquad p\text{-value} = P(N(0,1) \le \text{TS}); \\ H_1 : \mu > \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}},\ \text{C} = (\lambda_\alpha, +\infty), \\ \qquad p\text{-value} = P(N(0,1) \ge \text{TS}); \\ H_1 : \mu \neq \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}},\ \text{C} = (-\infty, -\lambda_{\alpha/2}) \cup (\lambda_{\alpha/2}, +\infty), \\ \qquad p\text{-value} = 2P(N(0,1) \ge |\text{TS}|). \end{cases}$$

5 Multi-dimension random variables (or random vectors)

Covariance (Kovarians) of (X,Y)**:** $\sigma_{X,Y} = cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right], (cov(X,X) = V(X))$.

Correlation coefficient (Korrelation) of (X,Y)**:** $\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{V(X)\cdot V(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}$.

A rule: for real constants a, a_i, b and b_j,

$$cov(a + \sum_{i=1}^m a_i X_i, b + \sum_{j=1}^n b_j Y_j) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j cov(X_i, Y_j).$$

X **and** Y **are uncorrelated:** if $cov(X,Y) = 0$.

An important theorem: Suppose that a random vector $\mathbf{X}$ has a mean $\mu_{\mathbf{X}}$ and a covariance matrix $C_{\mathbf{X}}$. Define a new random vector $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, for some matrix A and vector $\mathbf{b}$. Then

$$\mu_{\mathbf{Y}} = A\mu_{\mathbf{X}} + \mathbf{b}, \quad C_{\mathbf{Y}} = AC_{\mathbf{X}}A^T.$$

Standard normal vectors: $\{X_i\}$ are independent and $X_i \sim N(0,1)$,

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \text{ thus } \quad \mu_{\mathbf{X}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad C_{\mathbf{X}} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \text{ density } f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\mathbf{x}^T\mathbf{x}}.$$

General normal vectors: $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, where $\mathbf{X}$ is a standard normal vector, and

$$\mu_{\mathbf{Y}} = \mathbf{b}, \quad C_{\mathbf{Y}} = AA^T, \quad \text{density } f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(C_{\mathbf{Y}})}} e^{-\frac{1}{2}\left[(\mathbf{y}-\mu_{\mathbf{y}})^T C_{\mathbf{Y}}^{-1} (\mathbf{y}-\mu_{\mathbf{y}})\right]}.$$

6 (Simple and multiple) Linear regressions

Simple linear regression: $Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0,\sigma)$.

Multiple linear regression: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0,\sigma)$.

Both 'Simple linear regression' and 'Multiple linear regression' can be written as vector forms:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} : \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n}).$$

$$\mathbf{Y} \sim N(\mu_{\mathbf{Y}}, C_{\mathbf{Y}}), \text{ where } \mu_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} \text{ and } C_{\mathbf{Y}} = \sigma^2 \mathbf{I}_{n\times n}.$$

Estimate of the coefficient $\boldsymbol{\beta}$**:** $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$.

Estimator of the coefficient $\boldsymbol{\beta}$**:** $\hat{\boldsymbol{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$.

Estimated line is: $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$.

Analysis of variance:

$$SS_{TOT} = \sum_{j=1}^n (y_j - \bar{y})^2, \quad \frac{SS_{TOT}}{\sigma^2} = \frac{\sum_{j=1}^n (Y_j - \bar{Y})^2}{\sigma^2} \sim \chi^2(n-1), \text{ if } \beta_1 = \ldots = \beta_k = 0;$$

$$SS_R = \sum_{j=1}^n (\hat{\mu}_j - \bar{y})^2, \quad \frac{SS_R}{\sigma^2} = \frac{\sum_{j=1}^n (\hat{\mu}_j - \bar{Y})^2}{\sigma^2} \sim \chi^2(k), \text{ if } \beta_1 = \ldots = \beta_k = 0;$$

$$SS_E = \sum_{j=1}^n (y_j - \hat{\mu}_j)^2, \quad \frac{SS_E}{\sigma^2} = \frac{\sum_{j=1}^n (Y_j - \hat{\mu}_j)^2}{\sigma^2} \sim \chi^2(n-k-1).$$

$$SS_{TOT} = SS_R + SS_E, \text{ and } R^2 = \frac{SS_R}{SS_{TOT}}.$$

$***$ σ^2 is estimated as $\hat{\sigma}^2 = s^2 = \frac{SS_E}{n-k-1}$.

$***$ For the Hypothesis testing: $H_0 : \beta_1 = \ldots = \beta_k = 0$ vs $H_1 :$ at least one $\beta_j \neq 0$,

$$\begin{cases} \frac{SS_R/k}{SS_E/(n-k-1)} \sim F(k, n-k-1) \\ \text{TS} = \frac{SS_R/k}{SS_E/(n-k-1)} \\ C = (F_\alpha(k, n-k-1), +\infty). \end{cases}$$

$***$ We know $\hat{\boldsymbol{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \sim N\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$, thus if we denote

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{pmatrix} h_{00} & h_{01} & \cdots & h_{0k} \\ h_{10} & h_{11} & \cdots & h_{1k} \\ \vdots & \vdots & & \vdots \\ h_{k1} & h_{k2} & \cdots & h_{kk} \end{pmatrix},$$

then $\hat{B}_j \sim N(\beta_j, \sigma\sqrt{h_{jj}})$ and $\frac{\hat{B}_j - \beta_j}{\sigma\sqrt{h_{jj}}} \sim N(0,1)$. But σ is generally unknown, therefore

$$\frac{\hat{B}_j - \beta_j}{S\sqrt{h_{jj}}} \sim t(n-k-1), \qquad \left[s\sqrt{h_{jj}} \text{ is sometimes denoted as } d(\hat{\beta}_j) \text{ or } se(\hat{\beta}_j)\right].$$

Confidence interval of β_j is: $I_{\beta_j} = \hat{\beta}_j \mp t_{\alpha/2}(n-k-1) \cdot s\sqrt{h_{jj}}$;

Hypothesis testing $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ has

$$\begin{cases} \text{TS} = \frac{\hat{\beta}_j}{s\sqrt{h_{jj}}} \\ C = (-\infty, -t_{\alpha/2}(n-k-1)) \cup (t_{\alpha/2}(n-k-1), +\infty). \end{cases}$$

Rewrite simple and multiple linear regressions as follows:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma), \quad \text{(the model)};$$
$$\mu = E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \quad \text{(the mean)};$$
$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k, \quad \text{(the estimated line)}.$$

For a given/fixed new observation $\mathbf{u} = (1, u_1, \ldots, u_k)^T$, the scalar $\hat{\mu}$ is an estimate of unknown μ (and Y). Then we can talk about 'accuracy' of this estimate in terms of confidence intervals (and prediction intervals).

Confidence interval of μ: $I_\mu = \hat{\mu} \mp t_{\alpha/2}(n-k-1) \cdot s \cdot \sqrt{\mathbf{u}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{u}}$.

Prediction interval of Y: $I_Y = \hat{\mu} \mp t_{\alpha/2}(n-k-1) \cdot s \cdot \sqrt{\mathbf{u}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{u} + 1}$.

Suppose we have two models:

$$\begin{cases} \text{Model 1:} \quad Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon; \\ \text{Model 2:} \quad Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} x_{k+1} + \ldots + \beta_{k+p} x_{k+p} + \varepsilon, \end{cases}$$

and we want to test $H_0 : \beta_{k+1} = \ldots = \beta_{k+p} = 0$ vs $H_1 :$ at least one $\beta_{k+i} \neq 0$,

$$\begin{cases} \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1) \\ \text{TS} = \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \\ C = (F_\alpha(p, n-k-p-1), +\infty). \end{cases}$$

Variable selection. If we have a response variable y with possibly many predictors $x_1, \ldots, x_k$, then how to choose appropriate x's (some x's are useful to Y, and some are not):

Step 1: $corr([x_1, \ldots, x_k], y)$, choose a maximal correlation (say x_i), $Y = \beta_0 + \beta_i x_i + \varepsilon$, test if $\beta_i = 0$?

Step 2: do regression $Y = \beta_0 + \beta_i x_i + \beta_* x_* + \varepsilon$ for $* = 1, \ldots, i-1, i+1, \ldots, k$, choose a minimal SS_E (say x_j), $Y = \beta_0 + \beta_i x_i + \beta_j x_j + \varepsilon$, test if $\beta_j = 0$?

Step 3: repeat Step 2 until the last test for $\beta = 0$ is not rejected.

7 Basic χ^2-test

$$\text{Suppose we want to test } \begin{cases} H_0: & X \sim \text{ distribution (with or without unknown parameters)}; \\ H_1: & X \nsim \text{ distribution} \end{cases}$$

$$\text{Then } \begin{cases} \text{fact is } : \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \sim \chi^2(k - 1 - \#\text{of unknown parameters}); \\ \text{TS} = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}; \\ C = \left(\chi^2_\alpha(k - 1 - \#\text{of unknown parameters}), \ +\infty\right). \end{cases}$$

Independence / Homogeneity. Suppose we have a data with r rows and k columns,

$$\begin{cases} H_0: & \text{the grouping of } r \text{ rows and the grouping of } k \text{ columns are independent;} \\ H_1: & \text{the grouping of } r \text{ rows and the grouping of } k \text{ columns are not independent.} \end{cases}$$

Equivalently,

$$\begin{cases} H_0: & \text{the distributions of } r \text{ rows in each column are the same} \\ H_1: & \text{the distributions of } r \text{ rows in each column are Not the same} \end{cases}$$

Then

$$\begin{cases} \text{fact is } : \sum_{j=1}^k \sum_{i=1}^r \frac{(N_{ij} - np_{ij})^2}{np_{ij}} \sim \chi^2((r-1)(k-1)); \\ \text{TS} = \sum_{j=1}^k \sum_{i=1}^r \frac{(N_{ij} - np_{ij})^2}{np_{ij}}; \\ C = \left(\chi^2_\alpha((r-1)(k-1)), \ +\infty\right), \end{cases}$$

where $p_{ij} = p_i \cdot q_j$ are the theoretical probabilities.