Examinator: Zhenxia Liu (Tel: 0700895208). You are permitted to bring: a calculator, and “formel -och tabellsamling i matematisk statistik”. Scores rating: 8-11 points giving rate 3; 11.5-14.5 points giving rate 4; 15-18 points giving rate 5.

English Version

1 (3 points)

Assume that the distribution of lifetimes (unit: year) of a certain type of electronic components is $Exp(1/\mu)$ where the true average lifetime μ is unknown. One chose 400 such electronic components, and after one year 109 components still worked (namely, the other 291 components were broken after one year). Based on this information, use the method of moments to find a point estimate of μ.

Solution. Lethal X = number of components which still work after one year. Then $X \sim Bin(400, p)$, where

$$p = P(Exp(1/\mu) > 1) = \int_1^\infty frac1\mu e^{-x/\mu} dx = e^{1-/\mu}.$$

It follows from the method of moments that $E(X) = \bar{x}$ (here $\bar{x} = x_1/1 = 109$), therefore $400 \cdot p = 109$ implying

$$e^{1-/\mu} = 109/400 \Rightarrow \hat{\mu} = -1/\ln(109/400) = 0.77.$$

□

2 (3 points)

A random sample $\{X_1, \dots, X_n\}$ is taken from a population $N(\mu, \sigma)$ with unknown μ and known σ.
(2.1). (1p) Find a point estimator of μ using Maximum-Likelihood method.
(2.2). (1p) Is this point estimator in (2.1) unbiased? Why?
(2.3). (1p) Is this point estimator in (2.1) consistent? Why?

Solution. (2.1). The likelihood function is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2}.$$

Taking the logarithm gives

$$\ln L(\mu) = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2.$$

In order to find the maximal point, we take the first derivative

$$0 = \ln' L(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) \quad \Rightarrow \quad \hat{\mu} = \bar{X}.$$

The second derivative rule verifies that $\hat{\mu} = \bar{X}$ is indeed a maximum.
(2.2). Yes, since $E(\hat{\mu}) = E(\bar{X}) = E(\frac{1}{n}(X_1 + \ldots + X_n)) = \frac{1}{n}(E(X_1) + \ldots + E(X_n)) = \frac{1}{n} \cdot n\mu = \mu$.
(2.3). Yes, since $\hat{\mu}$ is unbiased and

$$V(\hat{\mu}) = V(\frac{1}{n}(X_1 + \ldots + X_n)) = \frac{1}{n^2}(V(X_1) + \ldots + V(X_n)) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \to 0, \text{ as } n \to \infty.$$

□

3 (3 points)

One wants to collect a random sample of n values from a population $Po(\mu)$. Using the sample, one intends to test the null hypothesis $H_0 : \mu = 4$ against the alternative hypothesis $H_1 : \mu > 4$ such that the probability of the first type error is 0.05 and the probability of the second type error is 0.01 with the true $\mu = 5$. How should n be chosen?

Solution. Let's pretend that n is large so that we can use normal approximations, that is

$$X_1 + \ldots + X_n \sim Po(n\mu) \approx N(n\mu, \sqrt{n\mu}).$$

It then follows that $\frac{\bar{X}-\mu}{\sqrt{\mu/n}} \approx N(0,1)$. The fact of the first type error is 0.05 gives that

$$H_0 \text{ is rejected when } \frac{\bar{X}-\mu_0}{\sqrt{\mu_0/n}} > z_{0.05} = 1.645.$$

Therefore,

$$\begin{aligned}
0.01 &= \text{the second type error} = P(\text{don't reject } H_0 \text{ when } H_0 \text{ is false and } \mu = 5) \\
&= P(\frac{\bar{X}-\mu_0}{\sqrt{\mu_0/n}} \le 1.645 \text{ when } \mu = 5) \\
&= P(\bar{X} \le 4 + 1.645 \cdot \sqrt{\frac{4}{n}} \text{ when } \mu = 5) \\
&= P(\frac{\bar{X}-\mu}{\sqrt{\mu/n}} \le \frac{(4 + 1.645 \cdot \sqrt{\frac{4}{n}}) - 5}{\sqrt{5/n}}) \approx P(N(0,1) \le \frac{(4 + 1.645 \cdot \sqrt{\frac{4}{n}}) - 5}{\sqrt{5/n}}).
\end{aligned}$$

Therefore,

$$\frac{(4 + 1.645 \cdot \sqrt{\frac{4}{n}}) - 5}{\sqrt{5/n}} = -2.33 \quad \Rightarrow \quad n = 72.25 \text{ (i.e. } n = 73).$$

□

4 (3 points)

Assume that $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 2 \\ 5 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}\right)$. One wants to make a linear combination $Y = aX_1 + bX_2$ such that the mean $E(Y) = 8$ and the variance $V(Y)$ is minimized. Determine a and b.

Solution. It follows from $8 = E(Y) = aE(X_1) + bE(X_2) = 2a + 5b$ that $2a = 8 - 5b$. Then the variance is computed as

$$\begin{aligned}
V(Y) &= V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2abcov(X_1, X_2) \\
&= 4a^2 + 3b^2 + 2ab \\
&= (8-5b)^2 + 3b^2 + (8-5b)b = 23b^2 - 72b + 64.
\end{aligned}$$

To find the minimal vale of $V(X)$ we just take the first derivative

$$0 = dV(Y)/db = 46b - 72 \quad \Rightarrow \quad b = 72/46 = 1.565 \qquad \Rightarrow \qquad a = (8-5b)/2 = 0.087.$$

□

5 (3 points)

The number of cars passing a bridge can be assumed to be Poisson distributed with a mean μ_1 cars per minute from North and a mean μ_2 cars per minute from South. Suppose that the number of cars from North is independent of the number of cars from South. During an hour there were 160 cars passed of which 90 cars were from North. Find a 95% confidence interval for $\mu_1 - \mu_2$.

Solution. Let

$$\begin{aligned} X &= \text{number of cars from North in an hour } \sim Po(60\mu_1) \approx N(60\mu_1, \sqrt{60\mu_1}), \\ Y &= \text{number of cars from South in an hour } \sim Po(60\mu_2) \approx N(60\mu_2, \sqrt{60\mu_2}). \end{aligned}$$

Then

$$X - Y \approx N(60\mu_1 - 60\mu_2, \sqrt{60\mu_1 + 60\mu_2}) \quad \Rightarrow \quad \frac{(\frac{X}{60} - \frac{Y}{60}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\mu_1+\mu_2}{60}}} \approx N(0,1).$$

Therefore, the confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned} I_{\mu_1-\mu_2} &= (\frac{x}{60} - \frac{y}{60}) \mp z_{\alpha/2} \cdot \sqrt{\frac{\hat{\mu}_1 + \hat{\mu}_2}{60}} \\ &= (\frac{90}{60} - \frac{70}{60}) \mp 1.96 \cdot \sqrt{\frac{90/60 + 70/60}{60}} \\ &= 0.333 \mp 1.96 \cdot 0.211 = 0.333 \mp 0.413 = (-0.08, 0.746). \end{aligned}$$

□

6 (3 points)

One wishes to investigate whether or not the check out frequency in a certain library varies with the day of the week. During a randomly chosen week one counts the number of books checked out at the individual days:

weekday	Monday	Tuesday	Wednesday	Thursday	Friday
# books checked out	108	135	114	146	120

Test on a significance level $\alpha = 0.01$ whether or not the check out frequency varies with the day of the week.

Solution. In this case,

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = 0.2 \quad \text{against} \quad H_1 : \text{ some } p_i \neq 0.2.$$

Then the test statistic is

$$TS = \sum_{i=1}^{5} \frac{(N_i - np_i)^2}{np_i} = 7.8266,$$

and the rejection region is

$$C = (\chi^2_\alpha(5-1), +\infty) = (13.28, +\infty).$$

It is clear that $TS \notin C$, so we don't reject H_0 (i.e. there is no evidence that frequency varies with the day of the week.

□

Svensk Version

1 (3 poäng)

Antag att fördelningen av livslängden (enhet: år) för en viss typ av elektroniska komponenter är $Exp(1/\mu)$ där den sanna genomsnittliga livslängden μ är okänd. Man valde 400 sådana elektroniska komponenter, och efter ett år arbetade 109 komponenter fortfarande (nämligen de andra 291 komponenterna bröts efter ett år). Basera på denna information, använd moment-metoden för att beräkna en punktskattning av μ.

2 (3 poäng)

Ett slumpmässigt stickprov $\{X_1, \ldots, X_n\}$ tas från en population $N(\mu, \sigma)$ med okänd μ och känd σ.
(2.1). (1p) Beräkna en punktskattning av μ genom att använda Maximum Likelihood-metoden.
(2.2). (1p) Är denna punktskattning i (2.1) väntevärdesriktig? Varför?
(2.3). (1p) Är denna punktskattning i (2.1) konsistent? Varför?

3 (3 poäng)

Man önskar insamla ett slumpmässigt stickprov om n värden från en population $Po(\mu)$. Med hjälp av stickprovet avser man att testa nollhypotesen $H_0 : \mu = 4$ mot den alternativa hypotesen $H_1 : \mu > 4$ på sådant sätt att sannolikheten för fel av första slaget är 0.05 och sannolikheten för fel av andra slaget är 0.01 med den sanna $\mu = 5$. Hur skall n väljas?

4 (3 poäng)

Antag att $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 2 \\ 5 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}\right)$. Man vill göra en linjärkombination $Y = aX_1 + bX_2$ sådan att väntevärdet $E(Y) = 8$ och variansen $V(Y)$ minimeras. Bestäm a och b.

5 (3 poäng)

Antalet bilar som passerar en bro kan antas vara Poissonfördelat med ett väntevärde μ_1 bilar per minut norrut och ett väntevärde μ_2 bilar per minut söderut. Antag att antalet bilar norrut är oberoende av antalet bilar söderut. Under en timme passerade 160 bilar varav 90 var norrut. Bilda ett approximativt 95% konfidensintervall för $\mu_1 - \mu_2$.

6 (3 poäng)

Man vill undersöka om utlåningsfrekvensen för ett bibliotek varierar med veckodag. Under en slumpmässigt vald vecka erhölls följande resultat:

veckodag	måndag	tisdag	onsdag	torsdag	fredag
# utlånade böcker	108	135	114	146	120

Testa på signifikansnivån $\alpha = 0.01$ huruvida utlåningen varierar med veckodag.

TAMS24: Notations and Formulas

— by Xiangfeng Yang

1 Basic notations and definitions

X: random variable (stokastiska variabel);

Mean (Väntevärde):

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

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 $N(\mu,\sigma)$

$$\begin{cases} I_\mu = \begin{cases} \bar{x} \mp \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \text{ if } \sigma \text{ is known; } \left[\text{fact } \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\right] \\ \bar{x} \mp t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \text{ if } \sigma \text{ is unknown; } \left[\text{fact } \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)\right] \end{cases} \\ I_{\sigma^2} = \left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \quad \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right); \left[\text{fact } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)\right] \\ \text{Unknown } \sigma^2 \text{ can be estimated by the sample variance } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{cases}$$

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

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

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 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 $N\hat{p} > 10$ and $N\hat{p}(1-\hat{p}) > 10$)

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

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

(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*:

$$\underline{\text{reject } H_0} \ \textit{if and only if} \ \underline{p\text{-value} < \alpha}.$$

For notational simplicity, we employ

$$\text{TS} := \text{“test statistic”; and C} := \text{“critical region”}.$$
$$\underline{\text{reject } H_0} \ \textit{if} \ \underline{\text{TS} \in \text{C}};$$
$$\underline{\text{reject } H_0} \ \textit{if and only if} \ \underline{p\text{-value} < \alpha}.$$

4.2 Hypothesis testing for population mean(s)

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

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

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 $H_0 : \mu_1 = \mu_2$.

$$\left\{ \begin{array}{l} \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} \left\{ \begin{array}{l} 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), \\ \qquad\qquad 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), \\ \qquad\qquad 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), \\ \qquad\qquad p\text{-value} = 2P(N(0,1) \geq |\text{TS}|). \end{array} \right. \\ \\ \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} \left\{ \begin{array}{l} 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)), \\ \qquad\qquad 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), \\ \qquad\qquad 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\qquad\qquad\qquad \cup \ (t_{\alpha/2}(n_1+n_2-2), +\infty), \\ \qquad\qquad p\text{-value} = 2P(t(n_1+n_2-2) \geq |\text{TS}|). \end{array} \right. \\ \\ \sigma_1 \neq \sigma_2 \text{ both unknown: similarly as in the tree of confidence intervals.} \end{array} \right.$$

4.3 Hypothesis testing for population variance(s)

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

4.4 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 $n \geq 30$). The idea is exactly the same as the one used in confidence intervals. **One example** is: a sample $\{X_1,\ldots,X_n\}, n \geq 30$, 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

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

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,

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

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

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

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

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

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

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

6 (Simple and multiple) Linear regressions

Simple linear regression: $Y_j=\beta_0+\beta_1x_j+\varepsilon_j, \quad \varepsilon_j\sim N(0,\sigma), j=1,\ldots,n.$

Multiple linear regression: $Y_j=\beta_0+\beta_1x_{j1}+\beta_2x_{j2}+\ldots+\beta_kx_{jk}+\varepsilon_j, \quad \varepsilon_j\sim N(0,\sigma), j=1,\ldots,n.$

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 $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$.

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

Estimated line is: $\hat{\mu}_j=\hat{\beta}_0+\hat{\beta}_1x_{j1}+\hat{\beta}_2x_{j2}+\ldots+\hat{\beta}_kx_{jk}$.

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

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

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:

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

For a given/fixed $\mathbf{x} = (1, x_1, \ldots, x_k)'$, 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: $I_\mu = \hat{\mu} \mp t_{\alpha/2}(n-k-1) \cdot s \cdot \sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$.

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

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

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

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

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

$$\begin{cases} H_0 : & \text{different rows have a same pattern (in terms of columns)}; \\ H_1 : & \text{different rows have different patterns (in terms of columns)}. \end{cases}$$

Equivalently,

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

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.