1、3种 testing

① Fisher null hypothesis test

必须先找到我们想验证事件的相反情况(H_0)，然后证明 H_0 错误。(因为证明一件事正确很难，但证明它错误容易)

One-tailed test：形式为 $H_0: \theta \leq \theta_0$ 或 $H_0: \theta \geq \theta_0$

P-value：评判标准，越小可信度越高 (level of significance)

若 $P[D|H_0] \leq$ P-value (D为实验数据)，则可以 reject H_0 (reject H_0 at the [P-value] level of significance)

关于D的选取：对测试样本取平均值(记为 $\bar{x}$，样本数量n)，

若 H_0 为 $\theta \leq \theta_0$，则D为n个样本在 H_0 中平均值大于 $\bar{x}$ 的概率

.....≥.....小于.....

若 H_0 为 $\theta = \theta_0$，先按照 $\bar{x}$ 偏向的一边计算，然后 ×2 (two-tail)

具体的概率计算，依赖正态分布的累积值，需参考mid资料第9页①部分

拒绝的P-value：小于0.05

② Neyman-Pearson decision theory

考虑 H_0 与 H_1，H_1 即为我们想要证明的情况(与 H_0 相反)(这里的 H_0 一般为 $\theta = \theta_0$)

Type I error: $\alpha = P[\text{reject } H_0 | H_0 \text{ true}]$

Type II error: $\beta = P[\text{accept } H_0 | H_1 \text{ true}]$

Power: $1-\beta = P[\text{accept } H_1 | H_1 \text{ true}]$ (high power: rejection of H_0 is likely if H_1 true)

Critical region: for α, $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$, $\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$ ($\bar{x}$ 的范围为 critical region)

需要注意是两边，不是中间

若实际处于 critical region 内，则 reject H_0

假设 H_1 为 $\mu = \mu_0 + \delta$

$$\beta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2} - \delta\sqrt{n}/\sigma} e^{-\frac{1}{2}t^2} dt = \int_{-\infty}^{-z_\beta} e^{-\frac{1}{2}t^2} dt$$

$$-z_\beta \approx z_{\alpha/2} - \delta\sqrt{n}/\sigma$$

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2}$$

也就是说，Neyman-Pearson 是教你怎样选取样本数量与判定标准的(根据目标的 α 与 β)

Operating characteristic curves (OC curves): 在非正态分布下，不能把绝 H_0 的概率与实验所得平均数的关系。(α 和 n 提前确定)

(接受 H_0)

怎样知道 β：计算出 $d(\delta)$ 在横坐标上找到，再找到数量为 n 的曲线，横坐标为 d 处的纵坐标值即为 β (课件 446~447页)

OC curve 的最大值(即 μ_0 处)为 $1-\alpha$ $d=\frac{|\mu-\mu_0|}{\sigma}$

注意，OC curve 为选择方法，即在获得平均值后，根据曲线上的概率进行选择，选择 H_1 的概率即为曲线上的 β 值。与真实情况是 H_0 还是 H_1 无关。

③ NHST, Tea test

④ T-test (原始方差未知)

$T_{n-1}=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ (s 为样本标准差) (注意：是样本标准差，前面的都是本身已知的标准差)

Reject at significance level α:

$H_0: \mu=\mu_0$ if $|T_{n-1}|>t_{\alpha/2,n-1}$

$H_0: \mu\leq\mu_0$ if $T_{n-1}>t_{\alpha,n-1}$

$H_0: \mu\geq\mu_0$ if $T_{n-1}<-t_{\alpha,n-1}$

OC curve: 课件 442~443页 $d=\frac{|\mu-\mu_0|}{\sigma}$

⑤ Chi-squared test (用来测试标准差)

s 为样本标准差，σ_0 为假设的，σ 为实际标准差

$\chi^2_{n-1}=\frac{(n-1)s^2}{\sigma_0^2}$

Reject at significance level α:

$H_0: \sigma=\sigma_0$ if $\chi^2_{n-1}>\chi^2_{\alpha/2,n-1}$ or $\chi^2_{n-1}<\chi^2_{1-\alpha/2,n-1}$

$H_0: \sigma\leq\sigma_0$ if $\chi^2_{n-1}>\chi^2_{\alpha,n-1}$

$H_0: \sigma\geq\sigma_0$ if $\chi^2_{n-1}<\chi^2_{1-\alpha,n-1}$

OC curve: 课件 448~449页 $\lambda=\frac{\sigma}{\sigma_0}=\frac{\text{true}}{\text{hyphsis}}$

OC curve: 只取决于分布本身，α，n，不取决于具体的测试

描述的是 H_0 为假时无法否决 H_0 的概率(即 β)

横坐标为真实分布的平均值

例如，若真实为 μ_0+t，则测试结果会不同，可能在拒绝 H_0 的区间内，也有可能在该区间外，这些不同情况即对应纵坐标值(错误概率)

横坐标为实际值在假设情况中的分布(转换，类似于标准化)，纵坐标为对应的真实情况下原始假设被接受的概率

2. Non-parameter statistics

non-parameteric statistics：不依赖任何参数

distribution-free statistics：不服从常规的分布

通常，这两种合并起来称为 non-parametric methods

用中位数取代平均值，用 $q_1 q_2 q_3$ 取代方差

Fisher test

想要验证 M_0，$Q_+ = \#\{X_k : X_k > M_0\}$（样本中比 M_0 大的数量），$Q_- = \#\{X_k : X_k < M_0\}$

$$P[Q_- \le k \mid M = M_0] = \sum_{x=0}^{k} \binom{n}{x} \frac{1}{2^n}$$

Reject at significance level α:

$H_0: M \le M_0$ if $P[Q_- \le k \mid M = M_0] < \alpha$

$H_0: M \ge M_0$ if $P[Q_+ \le k \mid M = M_0] < \alpha$

$H_0: M = M_0$ if $P[\min(Q_-, Q_+) \le k \mid M = M_0] < \alpha/2$

Wilcoxon signed rank test（课件46（~462页）（要求关于某个数对称才可使用）

将与 M_0 差的绝对值从小到大排序，比 M_0 小的乘 -1

W_+ 为比 M_0 大的序号之和，W_- 为比 M_0 小的序号之和（计算序号时有重复要取对应区间平均值）

$$E[W] = \frac{n(n+1)}{4} \quad Var[W] = \frac{n(n+1)(2n+1)}{24} - \sum \frac{a^3 - a}{48}$$

（重复项，每组都要减，a 为一组中的数量）

①可直接比较 W 和 critical value（查表）

Reject at significance level α:

$H_0: M \le M_0$ if $|W_-| <$ critical value for α

$H_0: M \ge M_0$ if $W_+ <$ critical value for α

$H_0: M = M_0$ if $W = \min(W_+, |W_-|) <$ critical value for $\alpha/2$

②计算概率（与①是两种方法，不是共同使用）

$\mu = E[W]$ $\sigma^2 = Var[W]$

$$z = \frac{|W| - \mu}{\sigma}$$

calculate $P[Z < z]$ 与 P-Value 比较

3. Proportion

考虑一个群体中，每个个体只有两种情况(0和1)

$X=0$ or 1

$p=\frac{1}{N}\sum_{N} X_i$

平均数 $\hat{p}=\bar{X}=\frac{1}{n}\sum_{n} X_i$

$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ 为正态分布

$100(1-\alpha)\%$ 置信区间：$\hat{p}\pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

为使@可信度大于 $1-\alpha$，样本数量 $n\geq\frac{z_{\alpha/2}^2\hat{p}(1-\hat{p})}{d^2}$

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

Reject at significance level α:

$H_0: p=p_0$ if $|z|>z_{\alpha/2}$

$H_0: p\leq p_0$ if $z>z_\alpha$

$H_0: p\geq p_0$ if $z<-z_\alpha$

考虑两个群体，p_1 和 p_2，要研究 p_1-p_2

$\widehat{p_1-p_2}=\hat{p_1}-\hat{p_2}$

$100(1-\alpha)\%$ 置信区间：$\hat{p_1}-\hat{p_2}\pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$ (pooled)

Reject at significance level α: ($z=\frac{\hat{p_1}-\hat{p_2}-(p_1-p_2)_0}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1+\hat{p_2}(1-\hat{p_2})/n_2}}$)

$H_0: p_1-p_2=(p_1-p_2)_0$ if $|z|>z_{\alpha/2}$

.... $\leq$ ..... $z>z_\alpha$

.... $\geq$ ..... $z<-z_\alpha$

Pooled estimator for proportion (比较两个大小)

$H_0: p_1=p_2$

$$\hat{p} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1+n_2}$$

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Reject at significiance level α:

$H_0: p_1 = p_2$ if $|Z| > z_{\alpha/2}$

$H_0: p_1 \leq p_2$ if $Z > z_\alpha$

$H_0: p_1 \geq p_2$ if $Z < -z_\alpha$

4. Comparisons

① Variances

$X_1 : N(\mu_1, \sigma_1^2)$ (正态分布) $X_2 : N(\mu_2, \sigma_2^2)$ A

$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$ A

F-distribution: $F_{\gamma_1,\gamma_2} = \frac{\chi^2_{\gamma_1}/\gamma_1}{\chi^2_{\gamma_2}/\gamma_2}$ (F-distribution with γ_1 and γ_2 自由度)

$P[F_{\gamma_1,\gamma_2} < x] = 1 - P[F_{\gamma_2,\gamma_1} < \frac{1}{x}]$

density function: $f_{\gamma_1,\gamma_2}(x) = \gamma_1^{\gamma_1/2}\gamma_2^{\gamma_2/2}\frac{\Gamma(\frac{\gamma_1+\gamma_2}{2})}{\Gamma(\frac{\gamma_1}{2})\Gamma(\frac{\gamma_2}{2})}\frac{x^{\gamma_1/2-1}}{(\gamma_1 x+\gamma_2)^{(\gamma_1+\gamma_2)/2}}$

$f_{\alpha,\gamma_1,\gamma_2}$: $P[F_{\gamma_1,\gamma_2} > f_{\alpha,\gamma_1,\gamma_2}] = \alpha$

$f_{1-\alpha,\gamma_1,\gamma_2} = \frac{1}{f_{\alpha,\gamma_2,\gamma_1}}$

考虑A处，若 $\sigma_1^2 = \sigma_2^2$，则 $\frac{S_1^2}{S_2^2}$ 服从 n_1-1 和 n_2-1 自由度的F分布

$F_{n_1-1,n_2} = \frac{S_1^2}{S_2^2}$

F-test, reject at significance level α:

$H_0: \sigma_1 \leq \sigma_2$ if $\frac{S_1^2}{S_2^2} > f_{\alpha,n_1-1,n_2-1}$

... 又 ... $\frac{S_2^2}{S_1^2} > f_{\alpha,n_2-1,n_1-1}$

$H_0: \sigma_1 = \sigma_2$ if $\frac{S_1^2}{S_2^2} > f_{\alpha/2,n_1-1,n_2-1}$ or $\frac{S_2^2}{S_1^2} > f_{\alpha/2,n_2-1,n_1-1}$

OC curve: $\lambda = \frac{\sigma_1}{\sigma_2}$，课件502~503页

② Means

$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_2$

$\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$ $\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$

i) 方差已知

$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ 为标准正态分布

$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$

Reject at level α:

$H_0: \mu_1 \le \mu_2$ if $Z > z_\alpha$

$H_0: \mu_1 \ge \mu_2$ if $Z < -z_\alpha$

$H_0: \mu_1 = \mu_2$ if $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

OC curve: $d = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ $n = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ (equivalent sample size)（课件510~511）

ii) 方差未知但相等 ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}$

$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$ (用来预测方差, pooled estimator for variance)

$\chi^2_{n_1+n_2-2} = \frac{(n-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2}$ ①，为 n_1+n_2-2 自由度的 chi-squared 分布

$T_{n_1+n_2-2} = \frac{Z}{\sqrt{\chi^2_{n_1+n_2-2}/(n_1+n_2-2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$ 为 n_1+n_2-2 自由度的 T 分布

$\mu_1 - \mu_2$ 的 $100(1-\alpha)\%$ 置信区间: $(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2}\sqrt{S_p^2(1/n_1 + 1/n_2)}$

Student's T test: $T_{n_1+n_2-2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$

~~Reject at α:~~

Reject at α:

$H_0: \mu_1-\mu_2 = (\mu_1-\mu_2)_0$ if $|T_{n_1+n_2-2}| > t_{\alpha/2, n_1+n_2-2}$

$H_0: \mu_1-\mu_2 \leq (\mu_1-\mu_2)_0$ if $T_{n_1+n_2-2} > t_{\alpha, n_1+n_2-2}$

$H_0: \mu_1-\mu_2 \geq (\mu_1-\mu_2)_0$ if $T_{n_1+n_2-2} < -t_{\alpha, n_1+n_2-2}$

OC curves: $d=\frac{|\mu_1-\mu_2|}{\sigma}$ ($n^*=2n-1$), 课件520-521

iii) 方差未知且不同

$\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right) = \frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}$

Welch-Satterthwaite Approximation:

$\gamma := \frac{(\lambda_1 S_1^2+\cdots+\lambda_k S_k^2)^2}{\sum_{i=1}^{k} \frac{(\lambda_i S_i)^2}{n_i-1}}$, 则 $\gamma \cdot \frac{\lambda_1 S_1^2+\cdots+\lambda_k S_k^2}{\lambda_1\sigma_1^2+\cdots+\lambda_k\sigma_k^2}$ 近似为 γ 自由度的卡方分布

Let $k=2$, $\lambda_1=\frac{1}{n_1}$, $\lambda_2=\frac{1}{n_2}$

则 $\gamma = \frac{(S_1^2/n_1+S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1-1)+(S_2^2/n_2)^2/(n_2-1)}$

$\gamma \cdot \frac{S_1^2/n_1+S_2^2/n_2}{\sigma_1^2/n_1+\sigma_2^2/n_2}$ 为 γ 自由度的 chi-squared 分布

$T_\gamma = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)_0}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$ 为 γ 自由度的 T 分布 (正常需将 γ 向下取整)

(pooled)

这是 Welch's test, Reject at α:

$H_0: \mu_1-\mu_2 = (\mu_1-\mu_2)_0$ if $|T_\gamma| > t_{\alpha/2,\gamma}$

$H_0: \mu_1-\mu_2 \leq (\mu_1-\mu_2)_0$ if $T_\gamma > t_{\alpha,\gamma}$

$H_0: \mu_1-\mu_2 \geq (\mu_1-\mu_2)_0$ if $T_\gamma < -t_{\alpha,\gamma}$

5-Non-parametric comparisons

有 X、Y 两个分布，比较 X > Y 的概率

$X_1 \sim X_m$ 和 $Y_1 \sim Y_n$ ($m \leq n$)；从小到大排序，记为 R_i ($i=1 \sim m+n$)，重复的记为序号平均值

令 $W_m = X_1 \sim X_m$ 的序号之和

$E[W_m] = \frac{m(m+n+1)}{2}$ $Var[W_m] = \frac{mn(m+n+1)}{12}$

若有重复项，则 $Var[W_m] = \frac{mn(m+n+1)}{12 - \sum \frac{t^3+t}{12}}$ （t为重复的项数）（#0果为负就不要减）

$H_0: P[X>Y] = \frac{1}{2}$

$Z = \frac{W_m - E[W_m]}{\sqrt{Var[W_m]}}$，若小于$Z_\alpha$则reject $H_0: P[X>Y] > \frac{1}{2}$，大于$Z_\alpha$则reject $H_0: P[X<Y] > \frac{1}{2}$

Paired test：有X、Y两个分布，研究 $D = X - Y$（要求X、Y正态分布）

$H_0: \mu_D = 0$

$T_{n-1} = \frac{\bar{D} - \mu_D}{\sqrt{S_D^2/n}}$（paired T test），reject规则和前面都一样，不再赘述

pooled vs paired: pooled（池子，所有数一起），paired（一对一对比较）

$T_{pooled} = \frac{\bar{X} - \bar{Y}}{\sqrt{2S_p^2/n}}$ critical value $= t_{\alpha/2, 2n-2}$

$T_{paired} = \frac{\bar{X} - \bar{Y}}{\sqrt{S_D^2/n}}$ critical value $= t_{\alpha/2, n-1}$

Pooled test 可以可容易 reject H_0，更好

$\frac{\sigma_D^2}{n} = \frac{2\sigma^2}{n}(1 - \rho_{XY})$

若$\rho_{XY} > 0$，paired更好，否则pooled更好

Correlation

$\widehat{Var}[X] = \frac{1}{n-1}\sum_n (X_i - \bar{X})^2$

$\widehat{Cov}[X,Y] = \frac{1}{n-1}\sum_n (X_i - \bar{X})(Y_i - \bar{Y})$

correlation efficient（相当于ρ_{XY}）：$R = \hat{\rho} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$

更多内容：课件558～563页

6. Categorical data

Categorical random variable: X can take on values 1～k with probabilities $p_1 \sim p_k$ $(p_1 + \cdots + p_k = 1)$

Multinomial trial (with parameters $p_1 \sim p_k$): result in exactly one from k possible outcomes, possibility is p_i respectively

Multinomial random variables: n次 M... trial中出现结果i的次数

Multinomial distribution: $f_{X_1 X_2 \cdots X_k}(x_1, x_2, \cdots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$

n次的 M ... dis...: $E[X_i] = np_i$, $Var[X_i] = np_i(1-p_i)$

$Cov[X_i, X_j] = -np_i p_j$

Pearson statistic: $\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ 近似为 k-1自由度的chi-squared分布

Cochran's rule: Pearson statistic要求n起码多大，即 $E[X_i] \geq 1$ 对所有 i，且 $E[X_i] \geq 5$ 对80%以上的 i

Pearson's Chi-squared Goodness-of-fit Test

$H_0: p_i = p_{i0}$ $(i = 1, \cdots, k)$

$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{i0})^2}{np_{i0}}$ (reject H_0 at α if $\chi^2_{k-1} > \chi^2_{\alpha, k-1}$)

Goodness-of-fit Test for discrete distribution

根据对应分布的自身特性，分出n类，得到每类的数量和其期望，对其使用上面方法（不满足Cochran's rule的类要去掉），看能否reject（大于χ_α）

课件579~583页

For continuous distribution

先分类（分段），$p = p_i = P[a_{i-1} \leq X \leq a_i] = \int_{a_{i-1}}^{a_i} f(x)\,dx$ （每段概率都相等）

注意：是根据 H_0 分段，不是根据测试数据

计算出每类的 O_i 和 E_i，计算 χ^2，看能否reject（大于χ_α）

课件584~588页

Contingency Table

	<6h	6-9h	>9h
low	n_{11}	n_{12}	n_{13}
avg	n_{21}	n_{22}	n_{23}
high	n_{31}	n_{32}	n_{33}

marginal row and column sums:

$n_{i\cdot} = \sum_{j=1}^{c} n_{i,j}$ $n_{\cdot j} = \sum_{i=1}^{r} n_{i,j}$

$H_0: p_{ij} = p_{i\cdot} p_{\cdot j}$ (完全独立)

$E_{ij} = n \cdot \hat{p_{ij}} = n \cdot \hat{p_{i\cdot}} \hat{p_{\cdot j}} = n \cdot \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}$

$X^2_{(r-1)(c-1)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ 为 chi-squared 分布，看能否 reject (大于 χ_α)

例子：课件597~600页

Comparing proportions (test for homogeneity)

每一总体的概率设为 π，各画都一样，自由度仍为 $(r-1)(c-1)$

7. linear regression

Independent variable X，选定的，predictor variable，regressor

Dependant variable Y，随机的，要研究 Y 与 X 的关系，response variable

① Simple linear regression

$Y|X = \beta_0 + \beta_1 X + E$ ($E[E]=0$)

需要计算：b_0 : estimator for β_0

b_1 : estimator for β_1

会有多组 (X, Y) 数据，写作 $Y_i = Y|X_i$ ($i = 1 \sim n$)，则有 $(X_1, Y_1) \sim (X_n, Y_n)$

$Y_i = b_0 + b_1 X_i + e_i$ (e_i 称为 residual)，目标是使 e_i 小

最小二乘法 (Least-Squares Estimation)：$e_1^2 + e_2^2 + \cdots + e_n^2 \to$ minimum

Error sum of squares: $SS_E = e_1^2 + \cdots + e_n^2 = \sum_n (Y_i - (b_0 + b_1 X_i))^2$

$\frac{\partial SS_E}{\partial b_0} = -2 \sum_n (Y_i - b_0 - b_1 X_i) = 0$, $\frac{\partial SS_E}{\partial b_1} = -2 \sum_n (Y_i - b_0 - b_1 X_i) X_i = 0$

Then $n b_0 + b_1 \sum_n X_i = \sum_n Y_i$ $b_0 \sum_n X_i + b_1 \sum_n X_i^2 = \sum_n X_i Y_i$

Then $b_1 = \frac{\sum_n X_i Y_i - \sum_n X_i \cdot \sum_n Y_i}{n \sum_n X_i^2 - (\sum_n X_i)^2}$ $b_0 = \frac{1}{n} \sum_n Y_i - \frac{b_1}{n} \sum_n X_i$

$S_{XX} = \sum_n (X_i - \bar{X})^2$ $S_{YY} = \sum_n (Y_i - \bar{Y})^2$ $S_{XY} = \sum_n (X_i - \bar{X})(Y_i - \bar{Y})$

则 $b_0 = \bar{Y} - b_1 \bar{X}$, $b_1 = \frac{S_{XY}}{S_{XX}}$, $SS_E = S_{YY} - b_1 S_{XY}$

$\frac{b_1-\beta_1}{\sigma/\sqrt{\Sigma(x_i-\bar{x})^2}}$ 和 $\frac{b_0-\beta_0}{\sigma\sqrt{\frac{\Sigma x_i^2}{n\Sigma(x_i-\bar{x})^2}}}$ 服从标准正态分布（其中σ为error的标准差，基于理论而不是计算）

Estimator for variance

$s^2=\frac{SS_E}{n-2}=\frac{1}{n-2}\sum_n(Y_i-\hat{\mu}_{Y|x_i})^2$ 为unbiased estimator

$(n-2)s^2/\sigma^2=\frac{SS_E}{\sigma^2}$，为$n-2$自由度的chi-squared分布

$100(1-\alpha)\%$ 置信区间：$\beta_1=b_1\pm t_{\alpha/2,n-2}\frac{s}{\sqrt{S_{xx}}}$，$\beta_0=b_0\pm t_{\alpha/2,n-2}\frac{s\sqrt{\Sigma x_i^2}}{\sqrt{nS_{xx}}}$

Test for regression parameters

$H_0:\beta_0=\beta_0^*$ and $H_0:\beta_1=\beta_1^*$

Significant: if a regression has evidence that slope $\beta_1\neq 0$

$H_0:\beta_1=0$

$T_{n-2}=\frac{b_1}{s/\sqrt{S_{xx}}}$，reject at α if $|T_{n-2}|>t_{\alpha/2,n-2}$

$\mu_{Y|x}=\beta_0+\beta_1x=\bar{Y}-B_1\bar{x}+\beta_1x=\bar{Y}+B_1(x-\bar{x})$ 这里是根据新的x计算对应的Y

$\frac{\hat{\mu}_{Y|x}-\mu_{Y|x}}{s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$ 为$n-2$自由度的T分布

$100(1-\alpha)\%$ 置信区间：$\hat{\mu}_{Y|x}\pm t_{\alpha/2,n-2}s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$ (confidence band)

Prediction band: $T_{n-2}=\frac{\hat{Y}|x-Y|x}{s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$

band is $Y|x\pm t_{\alpha/2,n-2}s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$ $(100(1-\alpha)\%)$

区别：confidence band 针对的是我们本身存在误差，给的区间，

prediction band 针对数据本身具有随机性，给的空间

SS_T (Total Sum of squares): $SS_T=S_{yy}=\sum_n(Y_i-\bar{Y})^2$

Coefficient of determination: $R^2=\frac{SS_T-SS_E}{SS_T}$ (越接近100%越好)

$$R^2 = \frac{S_{xY}^2}{S_{xx} S_{YY}}$$

$$T_{n-2} = \frac{B_1}{\sqrt{S^2/S_{xx}}} = \frac{R}{\sqrt{1-R^2}} \sqrt{n-2}$$

$$F_{1,n-2} = (n-2)\frac{R^2}{1-R^2} = (n-2)\frac{SS_T - SS_E}{SS_E}$$

F Test for significance

Reject $H_0: \beta_1 = 0$ at α if $F_{1,n-2} > f_{\alpha,1,n-2}$

Test for correlation

$H_0: \rho = 0$, reject at α if $\left| \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \right| > t_{\alpha/2,n-2}$

假设检验

Y_{ij} 为第 j 次测出 x_i 的对应 Y 值

Sample mean $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$

Internal sum of squares: $\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$

Error sum of squares due to pure error: $SS_{E;pe} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{j=1}^{n} Y_{ij} \right)^2$$

~~$\frac{1}{\sigma^2} SS_{E;pe} = \sum_{i=1}^{k} \frac{1}{\sigma^2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$~~

选取 $z = \bar{Y}_i$，误差最小

Error sum of squares due to lack of fit: $SS_{E;lf} = SS_E - SS_{E;pe}$

Test for lack of fit

$$F_{k-2,n-k} = \frac{SS_{E;lf}/(k-2)}{SS_{E;pe}/(n-k)}$$, H_0: model appropriate

Reject at α: $F_{k-2,n-k} > f_{\alpha,k-2,n-k}$

② Multiple linear regression (太难，略过)