1. Introduction to Data

1.1 Empirical Studies

- empirical studies deal with populations & process which are collection of units
- population : a collection of units
- process : a collection of units, 'produced' over time
- unit : an individual which we can take some measurements.
  区分 unit 与 units

- Type of empirical studies
  - sample survey 抽样 : obtained in finite population
  - observational studies 观察 : collect information without change variates.
  - experimental studies 实验 : 实验者进行干预/改变. inference
    ↳ eg. 实验者选择样本

  researchers randomly determined which participants were assigned to which group

区别

	survey	observational
# 研究对象	finite eg. # 选民	infinite eg. 患癌人数
收集数据频率	仅一次	定期
问题设置	specific	passive 被动

variates
- response variates 响应变量 (y)
- explanatory variates 解释变量 (x)

1.2 Data Collection

- Variate 变量 : a characteristic of a unit (由 x, y, z 表示)

  - continuous 连续 — ep. 身高、体重、年龄
  - discrete 离散 — "数量" ep. 产品数
  - categorical 清晰分类 — non-numerical ep. 是/否 ← binary variable 0/1
  - ordinal 模糊分类 — "程度" ep. 大中小
  - complex. 复杂变量 — unusual ep. 不明确的作者、聊天记录、照片…
    ↑ 分析时需转化成其它变量.

- attribute : a function of the variates over the population or process.

  ep. attribute for the completion of assignments:

  - the modal number of completed assignment
  - the proportion of assignments submitted during last hour.

1.3 Data Summaries

numerical summaries

- **location**
  - mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
  - median: $\hat{m}$
  - mode: 众数
- **dispersion / variability** (离散程度)
  - variance: $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$
  - range: max − min
  - interquartile (IQR) = $Q(0.75) - Q(0.25)$
    - → 越接近 1.35σ，越 fit Gaussian
- **shape**
  - skewness (中间) (对称性): $g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{\frac{3}{2}}}$
  - kurtosis (两端): $g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{2}}$ 越接近 3 越 fit Gaussian

- five number summary

min $Q(0.25)$ median $Q(0.75)$ max

Quartile calculation $Q(p)$: 1. Let $k = (n+1)p$ n: sample size

2. $\begin{cases} k \in \mathbb{Z} & Q(p) = y_{(k)} \\ k \notin \mathbb{Z} & Q(p) = \frac{1}{2}(y_{(j)} + y_{(j+1)}) \end{cases}$

- sample correlation (r)

$-1 \le r \le 1$

$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$ $S_{yy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

- close to 1: strong positive linear relationship
- close to 0: weak ~
- close to −1: strong negative linear relationship

不代表因果关系；不代表 x 会导致 y in变化

- relative risk

relative risk of event A in group B as compared to group $\bar{B}$ is:

$$\frac{y_{11} / (y_{11} + y_{12})}{y_{21} / (y_{21} + y_{22})}$$

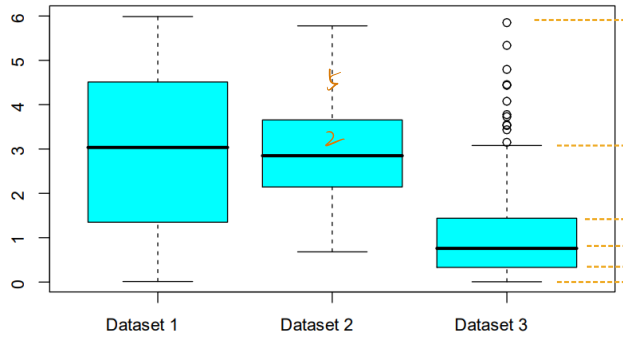
适用于 bivariate categorical data

two-way table →

	A	$\bar{A}$	total
B	y_{11}	y_{12}	$y_{11} + y_{12}$
$\bar{B}$	y_{21}	y_{22}	$y_{21} + y_{22}$
total	$y_{11} + y_{21}$	$y_{12} + y_{22}$	n

graphical summaries

boxplot



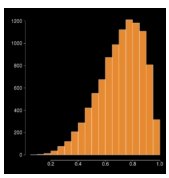
max
outlier
$Q(0.75) + 1.5\,IQR$
$Q(0.75)$ upper quartile
$Q(0.5)$
$Q(0.25)$ lower quartile
$Q(0.25) - 1.5\,IQR$

kurtosis: distribution — < 3 uniform — $= 3$ uniform + bell

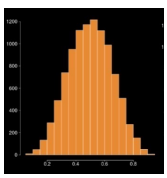
Dataset 1 :
1. sample median divides both the box and the IQRs in half -- indicates relative frequency histogram reasonably symmetric.
2. approximately 25% observations lie in 4 intervals of approximately equal width -- indicates shape of the relative frequency histogram reasonably uniform.

Dataset 2:
1. sample median divides both the box and the IQRs approximately in half -- indicates relative frequency histogram reasonably symmetric.
2. The distance from the sample median to the IQRs is approximately 2:5 times the distance from the sample median to the edge of the box -- indicates relative frequency histogram would be reasonably bell-shaped.

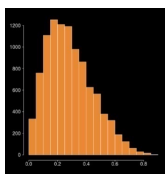
histogram



long left tail
neg skew

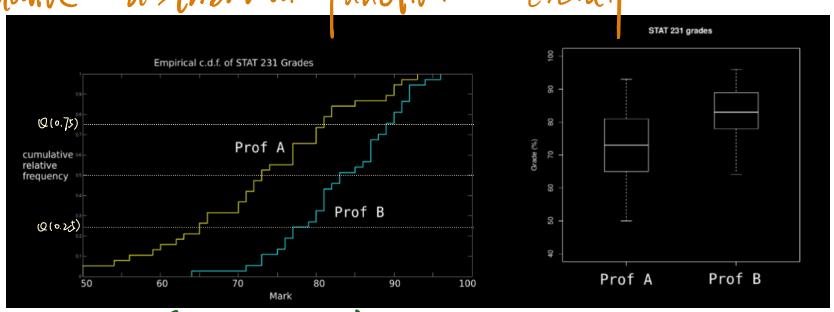


uniform & bell-shaped
skew approximate 0.



long right tail
pos skew

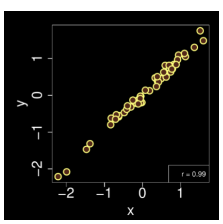
empirical cumulative distribution function e.c.d.f



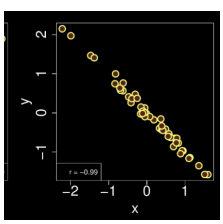
$$\hat{F}(y) = \frac{\#\{y_1, \cdots, y_n\} \le y}{n}$$

more jumps, more data

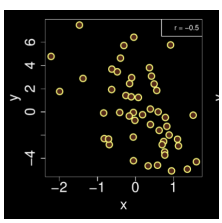
scatter plot 散点图



$r \approx 1$
strong pos linear relationship



$r \approx -1$
strong neg linear relationship



$r = 0$
no linear relationship

run chart 折线

bar / pie chart 条形图

1.4-1.5 Data analysis

- aspects of analysis & interpretation of data 分析、解释数据

- descriptive statistics 描述性统计
  - 对部分数据的描述
  - 以数字和图表的形式
- statistical inference 统计推断
  - 对群体数据的描述
  - 得出一些一般性结论
  - method
    - estimation problems 估算
    - hypothesis testing problems 假设检验
    - prediction problems 预测

2. Statistical Model

2.1 Choose model

- Distribution model

discrete →

Notation and Parameters	Probability Function $f(x)$	Mean $E(X)$	Variance $Var(X)$
Discrete Uniform(a,b) $b \geq a$ a, b integers	$\frac{1}{b-a+1}$ $x = a, a+1, \ldots, b$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Hypergeometric(N,r,n) $N = 1,2,\ldots$ $n = 0,1,\ldots,N$ $r = 0,1,\ldots,N$	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$ $x = max(0, n-N+r), \ldots, min(r,n)$	$\frac{nr}{N}$	$\frac{nr}{N}\left(1-\frac{r}{N}\right)\frac{N-n}{N-1}$
Binomial(n,p) $0 \leq p \leq 1, q = 1-p$ $n = 1,2,\ldots$	$\binom{n}{x}p^x q^{n-x}$ $x = 0,1,\ldots,n$	np	npq
Bernoulli(p) $0 \leq p \leq 1, q = 1-p$	$p^x q^{1-x}$ $x = 0,1$	p	pq
Negative Binomial(k,p) $0 < p \leq 1, q = 1-p$ $k = 1,2,\ldots$	$\binom{x+k-1}{x}p^k q^x$ $= \binom{-k}{x}p^k(-q)^x$ $x = 0,1,\ldots$	$\frac{kq}{p}$	$\frac{kq}{p^2}$
Geometric(p) $0 < p \leq 1, q = 1-p$	pq^x $x = 0,1,\ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$
Poisson(μ) $\mu \geq 0$	$\frac{e^{-\mu}\mu^x}{x!}$ $x = 0,1,\ldots$	μ	μ
Multinomial$(n; p_1, p_2, \ldots, p_k)$ $0 \leq p_i \leq 1$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{k} p_i = 1$	$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$ $x_i = 0,1,\ldots,n$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{n} x_i = n$	$E(X_i) = np_i$ $i = 1,\ldots,k$	$Var(X_i) = np_i(1-p_i)$ $i = 1,2,\ldots,k$

X = 首次S前in#F

X = k次S前in#F

不涉及重复性. 特定时间/空间内事件发生次数

continuous →

Uniform(a,b) $b > a$	$\frac{1}{b-a}$ $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(θ) $\theta > 0$	$\frac{1}{\theta}e^{-x/\theta}$ $x \geq 0$	θ	θ^2
$N(\mu,\sigma^2) = G(\mu,\sigma)$ $\mu \in \mathbb{R}, \sigma > 0$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in \mathbb{R}$	μ	σ^2

waiting time/rate

近似

$Z = \frac{X-\mu}{\sigma}$

- interative process

collect dataset

⇓

choose model

⇓ ⇑

fit model ⇒ check ⟹ draw conclusion

- discrete V.S continuous

	discrete	continuous
c.d.f	$F(x) = \sum_{t \le x} P(X=t)$	$F(x) = \int_{-\infty}^{x} f(t)\,dt$
p.d.f	$f(x) = P(X=x)$	0
$P(A)$	$P(X \in A) = \sum_{x \in A} f(x)$	$P(a<X<b) = \int_a^b f(x)\,dx$
$E(X)$	$E(g(X)) = \sum_{\forall x} g(x) f(x)$	$E(g(x)) = \int_{-\infty}^{\infty} g(x) f(x)\,dx$

2.2 - 2.3 Likelihood

- def. (point) estimate $\hat{}$

the value of a function of the observed data $y_1, \cdots, y_n$ and other known quantities such as sample size n.

eg. $Bi(n, \theta)$ sample proportion $\hat{\theta} = \frac{y}{n}$

$G(\mu, \sigma)$ sample mean $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- def. maximum likelihood function $\hat{\theta}$. m.l. estimate

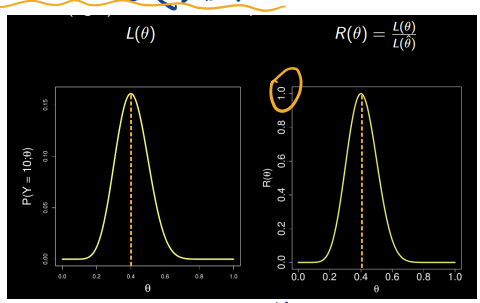
使$L(\theta)$最大 in θ 值 $\frac{d}{d\theta} l(\theta) = 0$

maximize $L(\theta)$. $R(\theta)$. $l(\theta)$. $R(\hat{\theta}) = 1$

- def. likelihood function $L(\theta)$

$L(\theta) = L(\theta; y) = P(Y = y; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$ $\theta \in \Omega$ (θ所有可能in值)

* 化简式不带 constant 项 in 形式 $L(\theta) \times$ constant. $L(\theta)$形状不变
* $L(\theta)$用于描述模型是否合理：$L(\theta)$越大，越接近$L(\hat{\theta})$，越合理



def. relative likelihood function $R(\theta)$

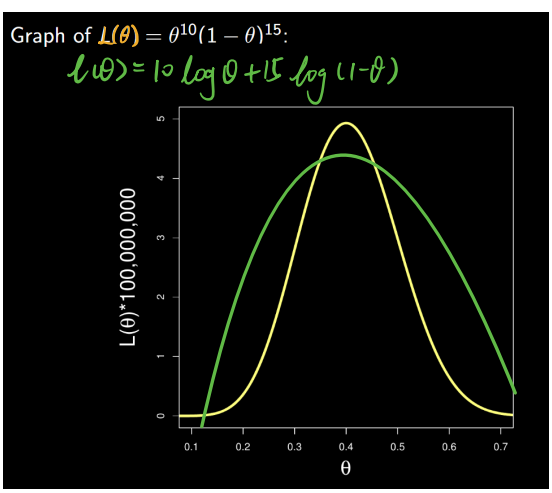
$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$ $\theta \in \Omega$ $0 \le R(\theta) \le 1$

* $\frac{L(\theta_1)}{L(\theta_2)}$ in 值表示：与$L(\theta_2)$比，数据与$L(\theta_1)$ in 值in一致性差异.

- def. log likelihood function $l(\theta)$

$l(\theta) = \log L(\theta)$

* ln无这 相当于 log
* $L(\theta)$与$l(\theta)$ in θ相同



def. log relative function $r(\theta)$

$r(\theta) = \log R(\theta) = l(\theta) - l(\hat{\theta})$

probability function	likelihood function
$P(Y=y;\theta)$ ↑ 变量 抛20次. 15次H 的概率	$P(Y=y;\theta)$ ↑ 变量 抛20次. 15次H. 硬币均匀的可能性

一些推导

constant. 不影响 θ that maximizes $\mathcal{L}(\theta)$

Binomial data: $P(Y=y;\theta)=\binom{n}{y}\theta^y(1-\theta)^{n-y}$

$\mathcal{L}(\theta)=P(Y=y;\theta)=\binom{n}{y}\theta^y(1-\theta)^{n-y}$ $\quad 0<\theta<1$

$l(\theta)=\log\mathcal{L}(\theta)$

Possion data: $\mathcal{L}(\theta)=\prod_{i=1}^{n}P(Y_i=y_i;\theta)$

$=\prod_{i=1}^{n}\frac{\theta^{y_i}e^{-\theta}}{y_i!}$

$=\left(\prod_{i=1}^{n}\frac{1}{y_i!}\right)\theta^{\sum_{i=1}^{n}y_i}e^{-n\theta}$ $\quad (\theta\geq 0)$

$=\theta^{n\bar{y}}e^{-n\theta}$

$l(\theta)=\log\mathcal{L}(\theta)=n(\bar{y}\log\theta-\theta)$

$\frac{d}{d\theta}l(\theta)=n\left(\frac{\bar{y}}{\theta}-1\right)=0$ $\quad \theta=\bar{y}$ $\quad \hat{\theta}=\bar{y}$

Exponential data: $\mathcal{L}(\theta)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{y_i}{\theta}}=\theta^{-n}e^{-\frac{n\bar{y}}{\theta}}$

$l(\theta)=n\left(\log\theta+\frac{\bar{y}}{\theta}\right)$

$\frac{d}{d\theta}l(\theta)=-n\left(\frac{1}{\theta}-\frac{\bar{y}}{\theta^2}\right)=\frac{n}{\theta^2}(\bar{y}-\theta)$

Gaussian data: $\theta=(\mu,\sigma)$

$\mathcal{L}(\theta)=\mathcal{L}(\mu,\sigma)$

$=\prod_{i=1}^{n}f(y_i;\mu,\sigma)$

$=(2\pi)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$

$=\sigma^{-n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2}e^{-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}}$

$l(\theta)=l(\mu,\sigma)=-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}$

2.5 Invariance property — m.l.e.

- Invariance property of m.l.e

If $\hat{\theta}$ is the m.l.e of θ, then $g(\hat{\theta})$ is the m.l.e of $g(\theta)$

通俗解释：所有能写作 function of θ 的 estimate

Q. $Y \sim Bi(n, \theta)$. 已知 $\hat{\theta}$, 求 20 次胜 5 次为 m.l.e.

对 $\binom{20}{5}\theta^5(1-\theta)^{20-5}$ 求导. 得出 m.l.e $= \hat{\theta}$.

Gaussian (μ, σ) $\quad \theta = (\mu, \sigma)$

$$L(\mu,\sigma) = \sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\}\exp\left[-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right]$$

$$l(\boldsymbol{\theta}) = l(\mu,\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2 - \frac{n(\bar{y}-\mu)^2}{2\sigma^2} \quad \text{for } \mu \in \Re \text{ and } \sigma > 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \quad \text{and} \quad \hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2}$$

Exp (θ)

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n}\frac{1}{\theta}e^{-y_i/\theta} = \frac{1}{\theta^n}\exp\left(-\sum_{i=1}^{n} y_i/\theta\right) \\ &= \theta^{-n}e^{-n\bar{y}/\theta} \quad \text{for } \theta > 0 \end{aligned}$$

$$l(\theta) = -n\left(\log\theta + \frac{\bar{y}}{\theta}\right) \quad \text{for } \theta > 0$$

Bi (n, p)

$$P(Y=y;\theta) = f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \quad \text{for } y = 0,1,\ldots,n \text{ and } 0 \le \theta \le 1$$

$$L(\theta) = \theta^y(1-\theta)^{n-y}$$

$$l(\theta) = n\left(\bar{y}\log\theta - \theta\right) \quad \text{for } \theta > 0$$

Po

$$L(\theta) = \theta^{n\bar{y}}e^{-n\theta} \quad \text{for } \theta \ge 0$$

$$l(\theta) = n\left(\bar{y}\log\theta - \theta\right) \quad \text{for } \theta > 0$$

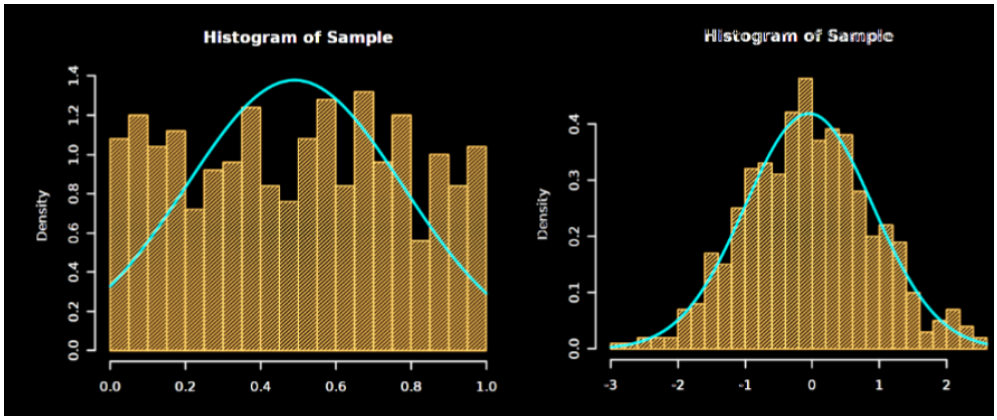
Y_1 & Y_2 indep

$$P(\mathbf{Y}_1=\mathbf{y}_1, \mathbf{Y}_2=\mathbf{y}_2;\theta) = P(\mathbf{Y}_1=\mathbf{y}_1;\theta)P(\mathbf{Y}_2=\mathbf{y}_2;\theta) \qquad L(\theta) = L_1(\theta)L_2(\theta) \quad \text{for } \theta \in \Omega$$

2.6 Check model fit

1. relative frequency histogram V.S p.d.f

cont.

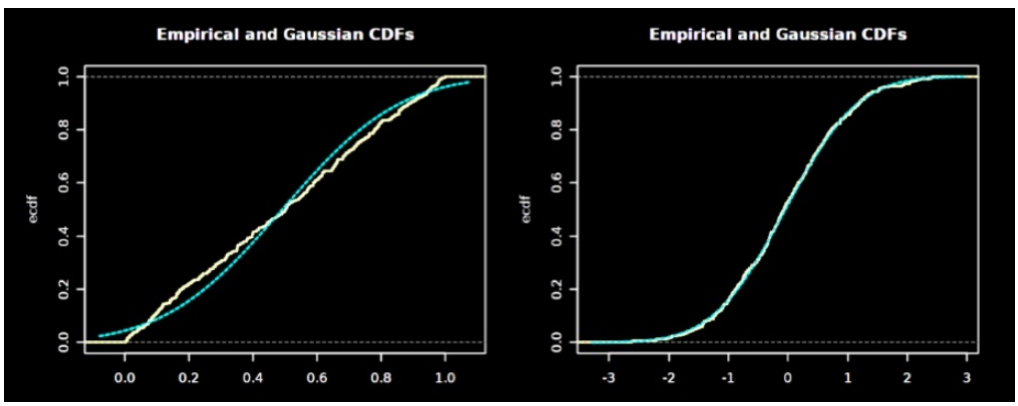


← r.f.h are in reasonable agreement with Gaussian p.d.f.

drawback : the intervals for the relative frequency histogram must be chosen

2. e.c.d.f V.S. cdf.

cont.

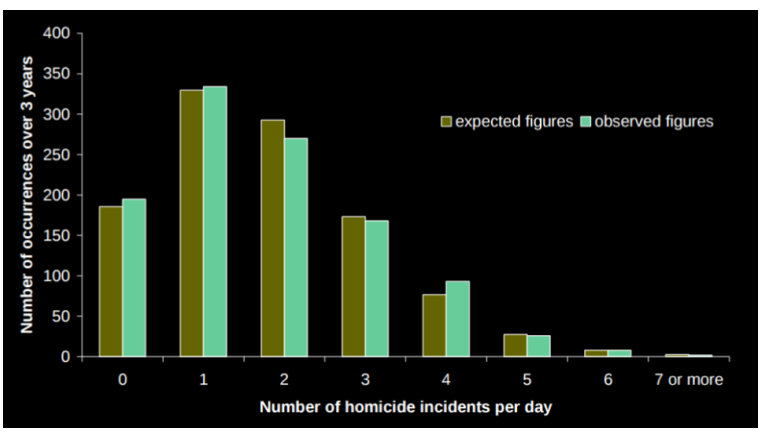


poor agreement between 2 curves

解释：the proposed Exp model disagrees with the observed distribution in both tails of the distribution

3. observed frequency V.S. expected frequency

discrete



4. Gaussian qq-plot cont.

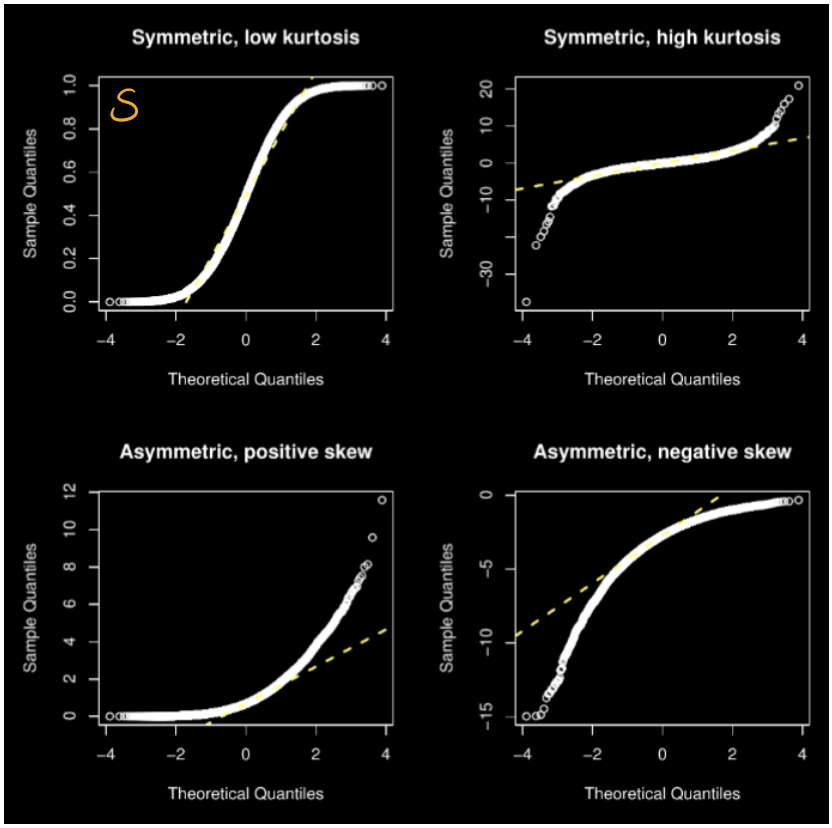
通过比较两个概率 ins probability distribution 来比较两个概率分布 ins 情况

S-Shaped

symmetry : low skewness

判断 more / fewer observations in tails

kurtosis: high low



U-Shaped

asymmetry :

判断 more pts on left / right

skewness : neg pos

points along a straight line : normality

Gaussian model is reasonable for a given data if:

1. The sample mean and median should be approximately equal.
2. The sample skewness should be close to 0.
3. The sample kurtosis should be close to 3.
4. Approximately 95% of the observations should lie in the interval [y − 2s, y + 2s].
5. Histograms and/or empirical cdfs should show agreement between the data and a theoretical distribution.
6. Q-Q plots should show points scattered approximately along a straight line.

Normality checking:

- The empirical cdf/histogram seem to reasonably fit the Gaussian cdf/density function.
- The points in our QQ plot appear to lie reasonably along a straight line.
- The empirical cdf/histogram do not seem to fit the Guassian cdf/density functions well. There is evidence of [positive/negative] skewness [and/or] [low/high] kurtosis.
- The points in our QQ plot do not appear to lie reasonably along a straight line.
- The QQ plot appears U-shaped, suggesting asymmetry, and a long [left/right] tail, suggesting [negative/positive] skewness.
- The QQ plot appears S-shaped, suggesting symmetry, but with [light/heavy] tails suggesting [low/high] kurtosis.

Binomial:
- independent trials
- 2 outcomes (S/F)
- P(S) are same on each trial

Possion
- independent. individuality. homogeneity
- sample mean & σ^2 close to each other. ($P_0: \bar{x} = \sigma^2$)
- Goodness fit test (Chap 7)

Exponential
- Memoryless property (与已经进去的时间无关)
- $\bar{x} \approx \sigma^2$
- $\hat{m} \approx \hat{\theta} \log 2 = \bar{y} \log 2$ $\quad \bar{y} > \hat{m}$
- c.d.f 与 e.c.d.f 重合
- qq-plot : U-shaped

Gaussian:
- $\bar{y} \approx \hat{m}$
- skewness ≈ 0
- kurtosis ≈ 3
- $IQR = Q(0.75) - Q(0.25) \approx 1.35\sigma$
- qq-plot : 点沿一条线分散. 两端对称分散

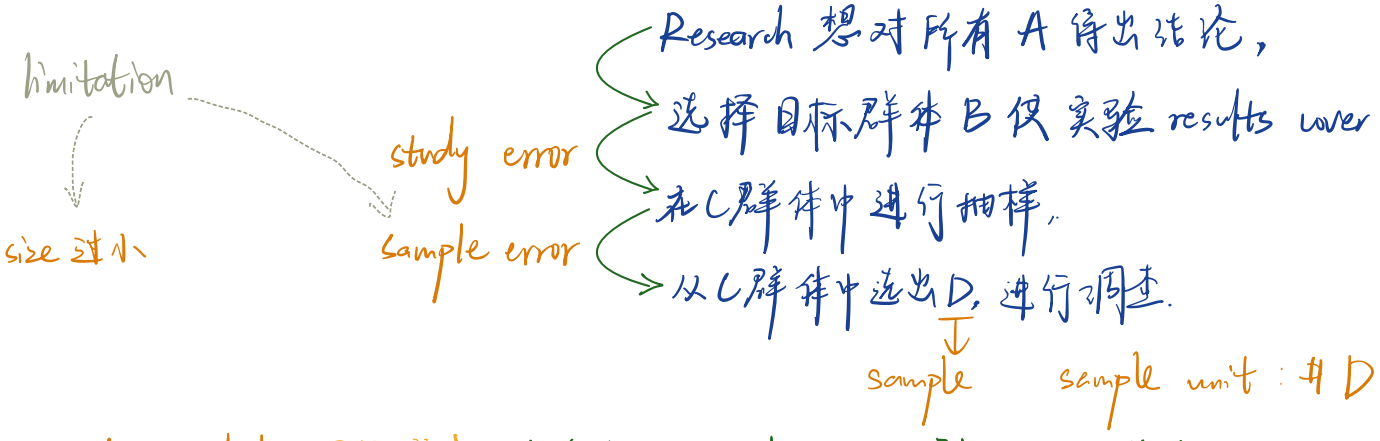
3. Empirical Studies 实证研究

- Step (PPPAC)

Problem clear statement of study objects 通常多个问题

- Steps:
  1. descriptive 确定人口的某一属性 / relationship
     1st step in descriptive: define units & target population / process (目标人群/单位)
  2. causative 确定两变量间是否有因果属性
  3. predictive 预测未来变量的反应
- Variate 变量 characteristic of every units
- Attribute 属性 a function of variates over a population

Plan procedures need to carry out the study including how data is collected.



B target population 目标群体: 所有想让 results cover 到 in 目标群体.

C study population 研究对象: 实验中可能被测试到 in 群体.

D sample protocal 样例: procedure of take (# sample unit) random sample from (study population)

study error: the attributes difference between study population & target population.

sample error: the attributes difference between sample & study population

measurement error: measured value V.S. true value (很严重,不可出现) 测量差错

systematic error: 测试群体选择有误 ep. B无法代表A

↳ most serious limitation for conclusion

Data the physical collection of data

- mistakes can occur in recording
- 随着时间推移，可能偏移研究计划 (# sample units 下降)

Analysis analysis data collected in Problems & Plans

Conclusion drawn about the problem and their limitation

purpose : address the questions posted in Problem

? Importance of control group: 对照组

- Serving as a baseline for comparing A & B.
- Allows researchers to determine the effect of 测量变.
- It helps to establish casuality and ensures any observed changes can attribute to the manipulated factors rather than external variables.

Suppose a polling company has a list of Instagram users, and messages a random group of users to ask them how they'll vote.

- Target population: people who will vote in the election.
- Study population: people on the polling company's list of Instagram users.
- Study error: In this case our target population (people who'll vote in the election) might be older on average than our study population (people with Instagram accounts).

4. Estimation

4.1 Statistical Models & estimation

Suppose the Ontario Ministry of Health wanted to conduct an empirical study to determine the proportion of adults aged 18-25 currently living in Ontario who have had HPV.

attribute of interest

target population : people aged 18-25 living in ON right now. 所有可能成为样本的

study population : We don't know

variate of interest : HPV status of each study unit. 观测啥

attribute of interest : proportion of target population who have had HPV.

How to estimate attribute? A Bi model and method of m.l.e

模型A：研究对象的变化模型 (includes the attributes which are to be estimated)

ep. Assume $Y \sim Bi(n, \theta_T)$

Y : # people in a random sample of n people aged 18-25 in ON had HPV.

θ_T : the proportion of all adults aged 18-25 currently live in ON had HPV.

模型B：数据收集方式的模型 (B模型与A共同构建)

ep. Assume $Y \sim Bi(n, \theta)$

Y : # people in a random sample of size n from the study population who had HPV.

θ : the probability a random chosen member of study population had HPV.

4.2 Estimators & Sampling distribution

- point estimate of θ $\hat{\theta}$

a function $\hat{\theta} = g(y_1, y_2, \cdots, y_n)$ of the observed data used to estimate the unknown parameter θ.

ep. point estimate of μ using sample mean : $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- point estimator $\tilde{\theta}$

a random variable which is a function $\tilde{\theta} = g(Y_1, \cdots, Y_n)$ Y_i : random variables

estimator is a rule that tells how to process data to obtain an estimate θ. (θ: unknown parameter)

ep. y_i = \$ grocery spend by student i. $Y_i \sim G(\mu, \sigma)$

point estimator : $\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

- Sampling distribution

(To determine the uncertainty in an estimate.)

The distribution of an estimator $\tilde{\theta}$

idea: Repeated Sampling

重复选择 random samples 获得不同 $\hat{\theta}$

* each sample gives us a different sample mean (每组 sample mean 都 slightly different)

每个 sample mean 可被当作一个随机变量

抽取样本中 $\hat{\theta}$ 估计值 θ 接近 μ 的概率 p : 1. $n\uparrow$ $p\uparrow$

2. $\sigma\uparrow$ $p\downarrow$

3. 与 μ 无关 (只要知道 σ 和 n, 即可求 p)

n	$P(234 \le \bar{Y} \le 236)$
25	0.057
50	0.080
100	0.114
1000	0.349

σ	$P(234 \le \bar{Y} \le 236)$
10	0.520
50	0.112
70	0.080
100	0.056

μ	$P(\mu - 1 \le \bar{Y} \le \mu + 1)$
100	0.080
235	0.080
500	0.080
1000	0.080

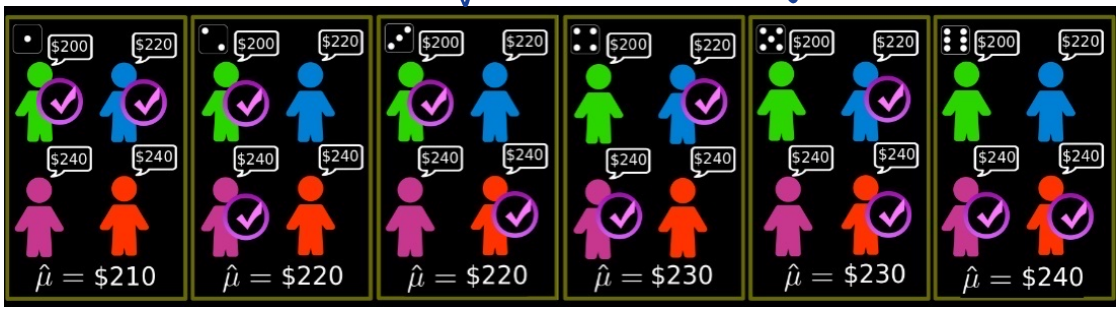
- Gaussian population

$Y \sim G(\mu, \sigma)$ → a linear combination of Gaussian random variables

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ $\quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G(\mu, \frac{\sigma}{\sqrt{n}})$

sampling distribution of $\tilde{\mu} = \bar{Y}$: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G(\mu, \frac{\sigma}{\sqrt{n}})$

Q. estimate μ = average \$ spend on groceries in n student



Sample (i)	Mean ($\hat{\mu}_i$)
1	242
2	243
3	224
4	268
...	...
99	212
100	234

Assume $Y \sim G(235, 70)$. the true mean we r trying to estimate

Sampling distribution of $\tilde{\mu}$: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G(\mu, \frac{\sigma}{\sqrt{n}}) = G(235, \frac{70}{\sqrt{50}})$

→ Gaussian data with known σ

What is the probability the sample will result in a sample mean within \$1 of the true mean monthly spend among Canadian Students?

Y = the monthly grocery spend of a randomly selected member of the study population

$Y_i \sim G(\mu, 70)$

estimator of μ : $\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

sampling distribution: $\tilde{\mu} = \bar{Y} = \frac{1}{50}\sum_{i=1}^{50} Y_i \sim G(\mu, \frac{70}{\sqrt{50}})$

point estimate of μ : $\hat{\mu} = \bar{y} = \frac{1}{50}\sum_{i=1}^{50} y_i$

$$P(\mu - 1 \le \bar{Y} \le \mu + 1) = P\left(\frac{\mu - 1 - \mu}{\frac{70}{\sqrt{50}}} \le \frac{\bar{Y} - \mu}{\frac{70}{\sqrt{50}}} \le \frac{\mu + 1 - \mu}{\frac{70}{\sqrt{50}}}\right)$$

$$= P\left(-\frac{\sqrt{50}}{70} \le Z \le \frac{\sqrt{50}}{70}\right)$$

$$= 0.08$$

```
> pnorm(236, 235, 70/sqrt(50)) - pnorm(234 - 1, 235, 70/sqrt(50))
[1] 0.08046165
```

• Non - Gaussian distribution

Using Gaussian approximation to Possion / Binomial / Exponential. (CLT)

→ $Y_i \sim Po(\theta)$ n large ⇒ $\frac{\bar{Y}-\theta}{\sqrt{\frac{\theta}{n}}} \sim G(0,1)$ $\bar{Y} \sim G(\theta, \sqrt{\frac{\theta}{n}})$

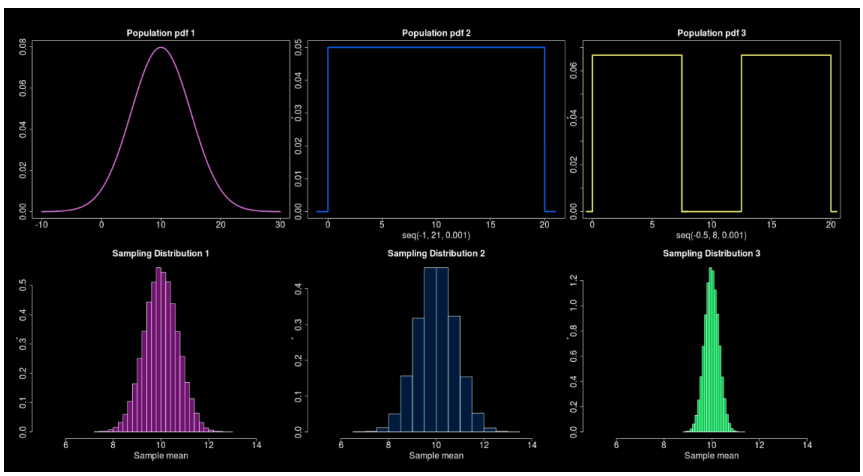
$Y_i \sim Po(\theta)$ $i = 1, 2, \cdots, n.$ $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

$E(Y_i) = \mu$ $Var(Y_i) = \sigma^2$ $Z_n = \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}$ (CLT)

→ $Y_i \sim B_i(n, \theta)$ n large ⇒ $\frac{\frac{Y}{n}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim G(0,1)$ $\bar{Y} \sim G(\theta, \sqrt{\frac{\theta(1-\theta)}{n}})$

→ $Y_i \sim Exp(\theta)$ n large ⇒ $\frac{\bar{Y}-\theta}{\frac{\theta}{\sqrt{n}}} \sim G(0,1)$ $\bar{Y} \sim G(\theta, \frac{\theta}{\sqrt{n}})$

- estimate value 与 real value 的差距
  - n ↑ → $\sigma(\bar{Y})$ ↓ → estimate value 更接近 real value
  - σ in population ↓ → $sd(\bar{Y})$ ↓ → 更接近 real value
  - Shape of population distribution will affect how many of our sample estimates will be close to the true value μ



n ↑ → distribution 接近 Gaussian

4.3 Interval estimation using likelihood

- relative likelihood function $R(\theta)$

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \qquad \theta \in \Omega$$

eg. $Y \sim Bi(n, \theta)$

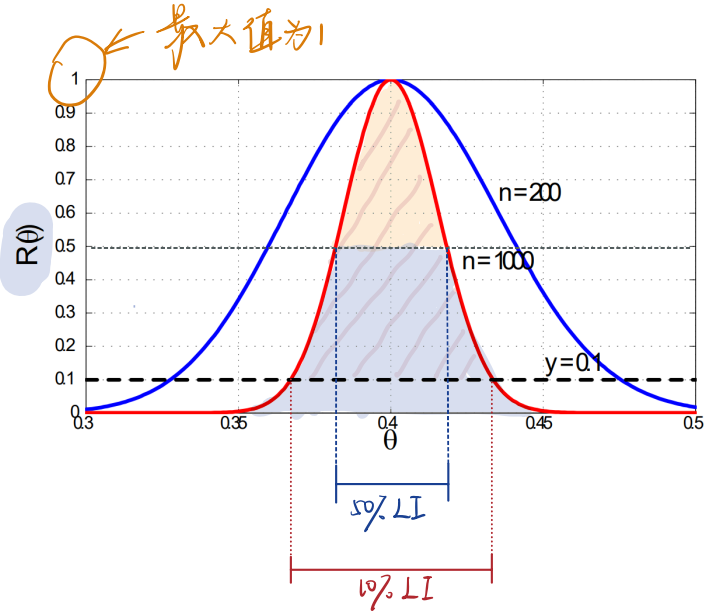
$$R(\theta) = \frac{\theta^y (1-\theta)^{n-y}}{\hat{\theta}^y (1-\hat{\theta})^{n-y}}$$

Poll 1 : $n = 200$ $y = 80$ $\hat{\theta} = \frac{80}{200} = 0.4$
Poll 2 : $n = 1000$ $y = 400$ $\hat{\theta} = \frac{400}{1000} = 0.4$

- $100p\%$ likelihood interval

$$\{\theta : R(\theta) \geq p\}$$

↓ 红色 in sample size 大于 蓝色



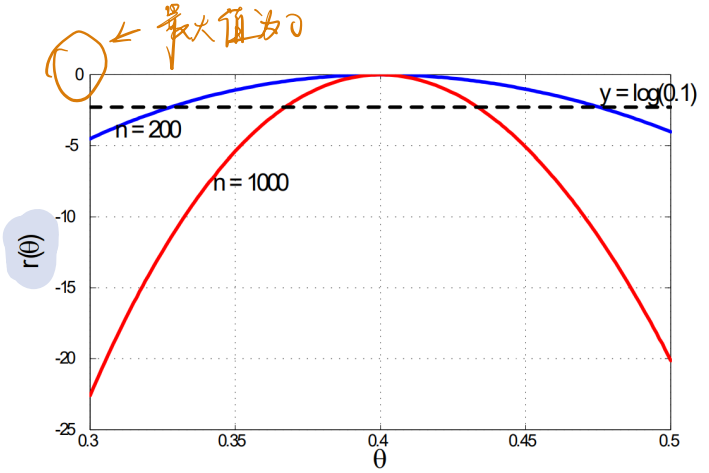
sample size ↑ → $R(\theta)$ graph more concentrate around θ

θ inside 50% likelihood interval are very plausible values

θ inside 10% likelihood interval are plausible values
θ outside 50% likelihood interval are implausible values

A 10% likelihood interval for p is found by using the R commands

```
uniroot(function(x)((x/0.89)^89*((1-x)/0.11)^11-0.1),
lower=0.8,upper=0.9)$root
uniroot(function(x)((x/0.89)^89*((1-x)/0.11)^11-0.1),
lower=0.9,upper=0.99)$root
```



$$r(\theta) = \log(R(\theta)) = \ell(\theta) - \ell(\hat{\theta})$$

- Coverage probability $P(\theta \in [L(Y), U(Y)])$

The coverage probability for the interval estimator $[L(Y), U(Y)]$ is :

$$P(\theta \in [L(Y), U(Y)]) = P[L(Y) \le \theta \le U(Y)]$$

($L(Y)$, $U(Y)$: random variables)

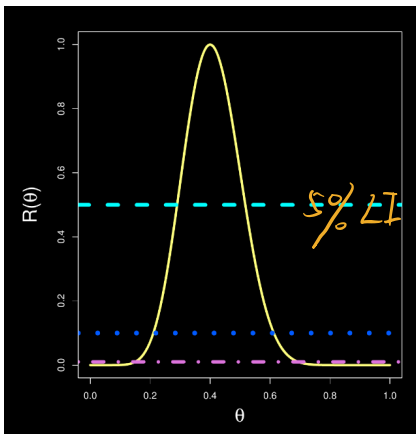
Q. Toss a coin 200 times.

Count # heads. Calculate 10% likelihood interval of each.

Reality : $P(\text{head}) = \theta = 0.5$

Experiment :

Sample	Heads	$\hat{\theta}$
1	80	0.400
2	102	0.510
...	...	...
99	86	0.430
100	109	0.545



$\theta = 0.5$ 50% LI

Sample	Heads	$\hat{\theta}$	50% LI	Covers θ?
1	80	0.400	[0.360, 0.441]	No
2	102	0.510	[0.468, 0.552]	Yes
...	...	...	...	...
99	86	0.430	[0.389, 0.472]	No
100	109	0.545	[0.503, 0.586]	No

→ 检查是否包含 θ

74 intervals covers $\theta = 0.5$. coverage of 10% LI is 74%

$\theta = 0.5$ 10% LI

Sample	Heads	$\hat{\theta}$	10% LI	Covers θ?
1	80	0.400	[0.325, 0.475]	No
2	102	0.510	[0.434, 0.585]	Yes
...	...	...	...	...
99	86	0.430	[0.356, 0.506]	→ Yes
100	109	0.545	[0.469, 0.620]	→ Yes

→ $\theta = 0.5$

97 intervals covers $\theta = 0.5$. coverage of 10% LI is 97%

likelihood level ↑ → LI (likelihood interval) narrower → coverage ↓

* higher coverage is prefered : the higher coverage, the more likely it is that our sample will give us an interval that covers the true value

4.4 Confidence intervals & pivotal quantities ! 相互转化

- confidence interval

A confidence interval (C.I.) for a population parameter θ is a range of values defined so that there is a specific probability that the true value of the parameter lies within that range.

100p% confidence interval for θ is an interval estimator $[L(Y), U(Y)]$:

$$P(\theta \in [L(Y), U(Y)]) = P(L(Y) \le \theta \le U(Y)) = p$$

coverage probability for the interval $[L(y), U(y)]$: $P(\theta \in [L(Y), U(Y)])$

confidence coefficient : p

* θ 是个未知 is constant. 改 random variable. 设 distribution.

Q. Gaussian distribution (未知 $\bar{x}$, 已知 σ^2) (米)

$Y \sim G(\mu, \nu)$

interval estimate of μ : $[\bar{y} - 1.96 \cdot \frac{2}{\sqrt{n}}, \bar{y} + 1.96 \cdot \frac{2}{\sqrt{n}}]$

↳ 将 $\mu = \bar{y}$ 对称

$\because \bar{Y} \sim G(\mu, \frac{2}{\sqrt{n}})$

$\therefore P(\mu \in [\bar{Y} - 1.96 \cdot \frac{2}{\sqrt{n}}, \bar{Y} + 1.96 \cdot \frac{2}{\sqrt{n}}])$

$= P(\bar{Y} - 1.96 \cdot \frac{2}{\sqrt{n}} \le \mu \le \bar{Y} + 1.96 \cdot \frac{2}{\sqrt{n}})$

$= P(-1.96 \le \frac{\bar{Y} - \mu}{\frac{2}{\sqrt{n}}} \le 1.96)$

$= 0.95$

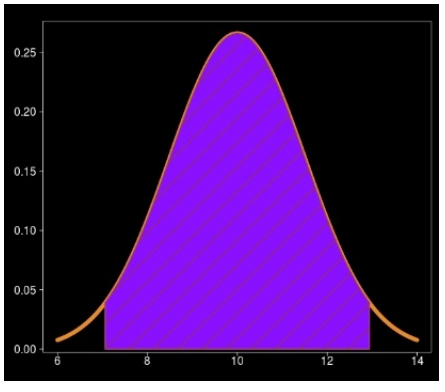
$\frac{\bar{Y} - \mu}{\frac{2}{\sqrt{n}}} = Z$ → 查表 $Z \sim G(0, 1)$

```
> pnorm(1.96) - pnorm(-1.96)
[1] 0.9500042
```

$\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ is 95% confidence interval for μ.

* interval doesn't depend on μ. depend on σ.

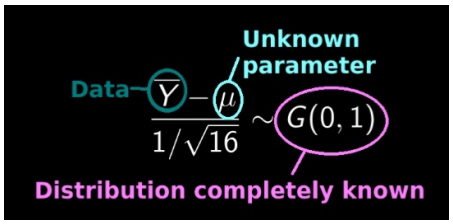
* $n \uparrow$. → interval gets narrower.



- Pivotal quantity $Q = Q(Y; \theta)$

一个数据 Y 与未知参数 θ 的 function. 使 Q 的 distribution 完全已知.

$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim G(0, 1)$$

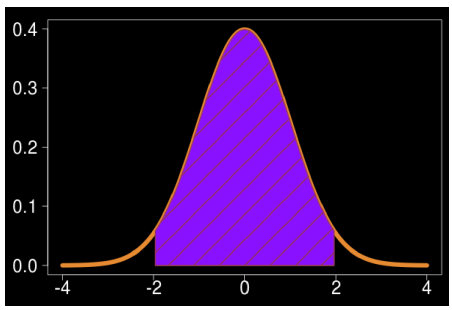


$P(Q \le b)$ & $P(Q \ge a)$ depend on a & b but not on θ or other information.

Use pivotal quantity to constnct CI :

pivotal quantity : $Q(Y; \mu) = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim G(0, 1)$ σ 已知

constnct CI for $p = 0.95$



1. Determine a. b. s.t $P(a \le Q(Y; \theta) \le b) = p$.

$$P\left(a \le \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \le b\right) = 0.95$$

$a = -1.96$ $b = 1.96$

有无数 in (a, b) 可选, we choose $(-1.96, 1.96)$ which gives narrowest CI

2. Re-express $a \le Q(Y; \theta) \le b$. in the form $L(Y) \le \theta \le U(Y)$

$$p = P(a \le Q(Y; \theta) \le b) = P(L(Y) \le \theta \le U(Y)) \quad \leftarrow p: \text{coverage prob.}$$

$$0.95 = P\left(-1.96 \le \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right)$$

$$= P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{Y} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\underbrace{\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}}_{L(Y)} \le \mu \le \underbrace{\bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}}_{U(Y)}\right)$$

3. $100p\%$ CI $= [L(y), U(y)]$

95% CI for μ is $\left[\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

ep. 测光波测 64 measurement ($n = 64$). sample mean $= 276 \times 10^6$ m/s sample s.d $= 55 \times 10^6$ m/s

95% CI for μ : $\left[\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right] = \left[276 \pm 1.96 \cdot \frac{55}{\sqrt{64}}\right]$

- Two-sided, equal-tailed CIs.

A $100p\%$ CI for μ is of the form (PP TS-ET):

point estimate ± distribution quatile × σ(estimator)

ep. Construct a $100p\%$ confidence interval for μ in Gaussian Data where σ known.

1. 找 a s.t $P(-a \le Z \le a) = p$. $\quad Z \sim G(0,1)$

(相当于 $P(Z \le a) = \frac{1+p}{2}$)

2. $100p\%$ confidence interval for μ is: $\bar{y} \pm a\frac{\sigma}{\sqrt{n}}$

- Changing confidence (CI)

CI: $\bar{y} \pm a\frac{\sigma}{\sqrt{n}}$ $\quad$ width $= 2a\frac{\sigma}{\sqrt{n}}$

- 若 confidence level change, 则只有 a change
- sample size ↑ → CI narrower.
- σ ↑ → CI wider

confidence level	CI
90%	$\bar{x} \pm 1.6449\cdot\frac{\sigma}{\sqrt{n}}$
95%	$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$
99%	$\bar{x} \pm 2.5758\frac{\sigma}{\sqrt{n}}$

可根据 c_1, c_2 求 σ

- Approximate pivotal quantities (Q_n)

We can always find $Q_n = Q_n(Y_1, \cdots, Y_n; \theta)$

s.t $n \to \infty$, the distribution of Q_n unlikely to depend on θ or other unknown information.

n越大，Q_n 越不容易被影响

Gaussian	Binomial
Pivotal quantity: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	Approx. pivotal quantity: $\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim G(0,1)$ approx.
1. Quantiles: $P\left(-a \le \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \le a\right) = p$	1. Quantiles: $P\left(-a \le \frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \le a\right) \approx p$
2. Rearrange: $P\left(\bar{Y} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + a\frac{\sigma}{\sqrt{n}}\right) = p$	2. Rearrange: $P\left(\tilde{\theta} - 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \le \theta \le \tilde{\theta} + 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}\right) \approx p$
3. CI: $\left[\bar{y} - a\frac{\sigma}{\sqrt{n}}, \bar{y} + a\frac{\sigma}{\sqrt{n}}\right]$	3. Approx. CI: $\left[\frac{y}{n} - 1.96\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}, \frac{y}{n} + 1.96\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right]$

- For Gaussian data:

$\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim G(0,1)$ is a pivotal quantity, which allows us to find a value s.t:

$$P\left(-a \le \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \le a\right) = p$$

- For Non-Gaussian data:

Using approximate pivotal quantity to construct an approximate 100p% CI:

Approximate 95% CI for Binomial

1. Determine a & b s.t. $P(a \le Q_n(Y;\theta) \le b) \approx p$

$$P\left(-1.96 \le \frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \le 1.96\right) \approx 0.95$$

2. Express $a \le Q(Y;\theta) \le b$ in the form $L(Y) \le \theta \le U(Y)$

then coverage probability: $p = P[a \le Q_n(Y;\theta) \le b] = P(L(Y) \le \theta \le U(Y))$

$$0.95 \approx P\left(\tilde{\theta}-1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \le \theta \le \tilde{\theta}+1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}\right)$$

3. For observed data y, approximate 100p% CI for θ is $[L(y), U(y)]$

$\because \tilde{\theta} = \frac{Y}{n}$ $\therefore$ an approximate 95% CI for θ is:

$$\left[\frac{y}{n}-1.96\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}},\ \frac{y}{n}+1.96\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right] = \hat{\theta} \pm 1.96\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}$$

- Sample size calculation 计算 n.

$\hat{\theta}=0.758$ $n=240$ 95% CI: [70.4%, 81.2%] narrower ✓

$n=120$ 95% CI: [68.2%, 83.5%]

Suppose we want a 95% CI of width $\le 2l$. Then we require n s.t

$$2 \times \left(1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right) \le 2l$$

$$n \ge \left(\frac{1.96}{l}\right)^2 \hat{\theta}(1-\hat{\theta})$$

$\frac{2l}{0.06}$ → $l=0.03$ $\hat{\theta}=0.5$.

$$n \ge \left(\frac{1.96}{0.03}\right)^2 \times 0.5 \times 0.5$$

↳ $n \ge 1067.1$

If $n=1068$, then the approximate 95% CI for θ have width <0.03 for all values of $\hat{\theta}$

4.5.1 Chi-square

- Gamma function (用于描述持续时间)

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy \qquad \alpha > 0$$

properties: 1) $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
2) $\Gamma(\alpha) = (\alpha-1)!$ $\alpha = 1, 2, \cdots$
3) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

ep. Exponential distribution

$$E(X) = \int_{-\infty}^{\infty} x f(x)$$
$$= \int_{-\infty}^{\infty} x \frac{1}{\theta} e^{-\frac{x}{\theta}}\, dx \qquad \text{let } y = \frac{x}{\theta} \quad dx = \theta\, dy$$
$$= \int_0^\infty y e^{-y}\, \theta\, dy$$
$$= \theta \int_0^\infty y^1 e^{-y}\, dy$$
$$= \theta\, \Gamma(2)$$
$$= \theta(1!)$$
$$= \theta$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)$$
$$= \int_{-\infty}^{\infty} x^2 \frac{1}{\theta} e^{-\frac{x}{\theta}}\, dx \qquad \text{let } y = \frac{x}{\theta} \quad dx = \theta\, dy$$
$$= \int_0^\infty \theta y^2 e^{-y}\, \theta\, dy$$
$$= \theta^2 \int_0^\infty y^2 e^{-y}\, dy$$
$$= \theta^2 \Gamma(3)$$
$$= \theta^2 (2!)$$
$$= 2\theta^2$$

If $X \sim Exp(\theta)$, then $E(X) = \theta$ $Var(X) = \theta^2$

- chi-squared distribution $\chi^2(k)$ gamma distribution 的一种特殊情况

连续分布. 随机变量非负.

a continuous family of distributions on $(0, \infty)$ with p.d.f:

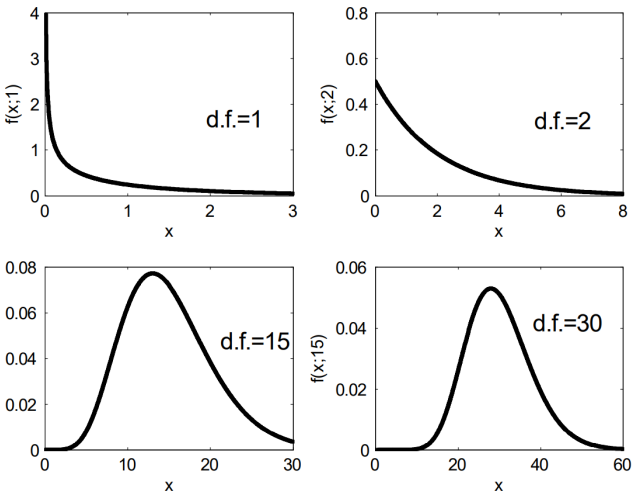
$$f(x; k) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \qquad x > 0$$

↳ degree of freedom

$$E(X^j) = 2^j \frac{\Gamma(\frac{k}{2}+j)}{\Gamma(\frac{k}{2})}$$

ep. $X \sim \chi^2(k)$ $E(X) = k$ $Var(X) = 2k$

图像形状取决于 degrees of freedom (d.f.)：　　d.f. = (列数-1)·(行数-1)



$j=1$

$E(X)=2\cdot\frac{\gamma(\frac{k}{2}+1)}{\gamma(\frac{k}{2})}$

$j=2$

$E(X^2)=2^2\cdot\frac{\gamma(\frac{k}{2}+2)}{\gamma(\frac{k}{2})}$

$=k(k+2)$

$j=15$

$j=30$

- Properties

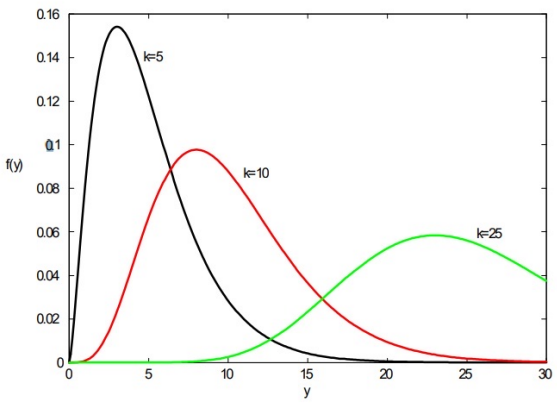
1. If $W\sim\chi_k^2$, then $E(W)=k$　$Var(W)=2k$

2. Sums of chi-squared distributions.

If $W_i\sim\chi_{k_i}^2$, then $S=\sum_{i=1}^{n}W_i\sim\chi^2_{\sum k_i}$

also follows chi-squared distribution

eg. $W_1\sim\chi_1^2$　$W_2\sim\chi_3^2$　$\Rightarrow$　$W_1+W_2\sim\chi_4^2$



k越大，p.d.f越symmetric along $y=k$

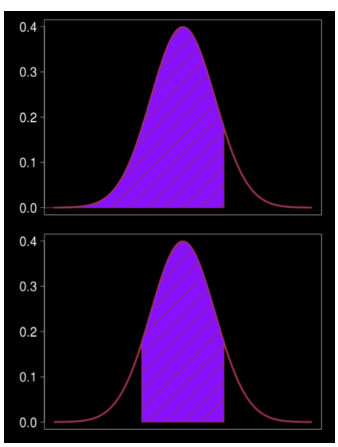
3. $p.d.f=\frac{1}{2^{\frac{k}{2}}\gamma(\frac{k}{2})}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$

4. • If $Z\sim G(0,1)$, then $Z^2=W\sim\chi_1^2$

$P(W\le w)=P(Z^2\le w)$

$=P(-\sqrt{w}\le Z\le\sqrt{w})$

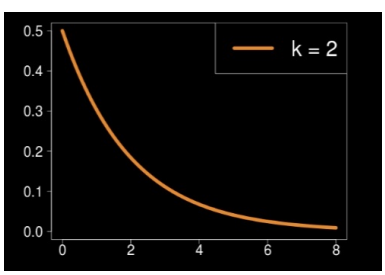
$=2P(Z\le\sqrt{w})-1$



• If $Z_1,Z_2,\cdots,Z_n\sim G(0,1)$, then $S=\sum_{i=1}^{n}Z_i^2\sim\chi_n^2$

5. If $W\sim\chi_2^2$, then $W\sim Exp(2)$

$P(W\le w)=1-e^{-\frac{w}{2}}$　$P(W\ge w)=e^{-\frac{w}{2}}$



a chi-squared distribution with 2 d.f. $\equiv$ Exp(2)

4.5.2 t Distribution

- Student's t distribution

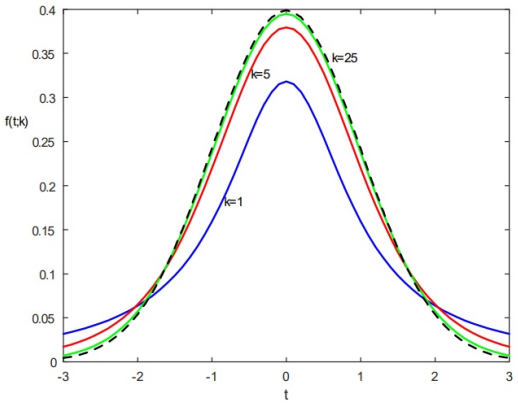
用于小样本情况下的推断

$T \sim t(k)$: T has a t distribution with $d.f = k$.

$$f(t;k) = c_k\left(1+\frac{t}{k}\right)^{-\frac{k+1}{2}}$$

$$c_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}$$

k越大，t(k)图像越接近 G(0,1)



适用条件： $Z \sim G(0,1)$. $U \sim \chi^2(k)$ independently

$$T = \frac{Z}{\sqrt{\frac{U}{k}}}$$

$Z \sim G(0,1)$

$U \sim \chi^2_k$

↳ Then T has a Student's t distribution with k degree of freedom

- t distribution

Suppose $Y_1, Y_2, \cdots, Y_n$ is a random sample from $G(\mu, \sigma)$

- σ known $\quad \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim G(0,1)$
- σ unknown $\quad \frac{\bar{Y}-\mu}{\frac{S}{\sqrt{n}}}$ $\quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$ (4.7 也有)

Q: Is this a pivotal quantity? What is its distribution?

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \Rightarrow V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\frac{\bar{Y}-\mu}{\frac{S}{\sqrt{n}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}} \sim t_{n-1}$$

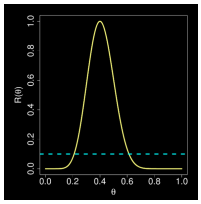
4.6 Likelihood intervals and Confidence intervals

- Likelihood interval

100p%　　$\{\theta : R(\theta) \geq p\}$

Values of θ s.t $R(\theta) \geq p$.

通过 uniroot function 得到：　$R(\theta) \geq 0.15$



$$R(\theta) = \left[\left(\frac{\theta}{0.5}\right)\left(\frac{1-\theta}{0.5}\right)\right]^{200} = [4\theta(1-\theta)]^{200} \quad \text{for} \quad 0 \leq \theta \leq 1$$

```
uniroot(function(x)((4*x*(1-x))^200-0.15),lower=0.42,upper=0.46)$root
uniroot(function(x)((4*x*(1-x))^200-0.15),lower=0.52,upper=0.56)$root
```

Given confidence level q the likelihood interval level p is found by:

- Find c such that $q = P(W \leq c) = P(|Z| \leq \sqrt{c})$ where q is the coverage probability (confidence level) and $W \sim \chi^2_1$.
- The likelihood interval is then given by $\{\theta : R(\theta) \geq e^{-c/2}\}$.
- And the corresponding likelihood interval level is $p = e^{-c/2}$.

Given likelihood level p the confidence interval level q is found by:

- $q = P(\Lambda(\theta) \leq -2 \log p)$ where $\Lambda(\theta) \sim \chi^2_1$

- Confidence interval

100q%　　$\{\theta : R(\theta) \geq e^{-\frac{c}{2}}\}$

Values of θ s.t $P(L(Y) \leq \theta \leq U(Y)) = q$　　$P(W \leq c) = q$　$W \sim \chi_1^2$

通过 gamma function 得到

特征	Likelihood Interval	Confidence Interval
定义	表示参数值的估计范围，使得在给定数据下，该参数值具有较高的可能性。	表示参数值的估计范围，使得在多次重复抽样的情况下，真实参数值落在该区间的概率达到指定的置信水平。
目的	强调在给定数据下对参数值的可能性进行推断。	强调在多次抽样下对参数值的估计精度进行推断。
对称性	不一定对称，取决于估计方法和分布形状。	对称，置信区间的上限和下限相对于估计值对称。
Confidence Level	不涉及置信水平的概念。	置信水平表示在多次抽样中，估计的置信区间包含真实参数值的概率。常见的置信水平包括95%和99%。
计算方法	基于似然函数构建，考虑参数值的可能性。	基于抽样分布构建，考虑在多次抽样中参数估计的变异。
适用情况	适用于小样本或非正态分布情况，强调在给定数据下的推断。	适用于大样本且满足正态性假设的情况，强调对参数估计的精度。

- Thm **100p% LI ≈ 100q% CI**

A 100p% likelihood interval is an approximate 100q% confidence interval where $q = 2P(Z \le \sqrt{-2\log p}) - 1$. $Z \sim N(0,1)$

proof: 100p% likelihood interval of $\{\theta; R(\theta) \ge p\}$

$$\{\theta; R(\theta) \ge p\} = \{\theta : -2\log\left(\frac{L(\theta)}{L(\hat\theta)}\right) \le -2\log p\}$$

↳ Can be approximated by

$$\begin{aligned} P(\Lambda(\theta) \le -2\log p) &= P\left(-2\log\left(\frac{L(\theta)}{L(\hat\theta)}\right) \le -2\log p\right) \\ &\approx P(W \le -2\log p) \qquad W \sim \chi^2(1) \\ &= P(|Z| \le \sqrt{-2\log p}) \qquad Z \sim N(0,1) \\ &= 2P(Z \le \sqrt{-2\log p}) - 1 \end{aligned}$$

ex. 10% LI ≈ 96.8% CI

$p = 0.1 \qquad c = -2\log p = 4.605$

$q = P(W \le c) = P(W \le -2\log p) = P(W \le 4.605) = 0.968$

- Thm **LI ≈ 100p% CI**

If a is a value s.t. $p = 2P(Z \le a) - 1 \qquad Z \sim N(0,1)$

Then the likelihood interval $\{\theta : R(\theta) \ge e^{-\frac{a^2}{2}}\}$ is an approximate 100p% confidence interval.

proof: $P\left(\frac{L(\theta)}{L(\hat\theta)} \ge e^{-\frac{a^2}{2}}\right) = P\left(-2\log\left(\frac{L(\theta)}{L(\hat\theta)}\right) \le a^2\right)$

$$\begin{aligned} &\approx P(W \le a^2) \qquad W \sim \chi^2(1) \\ &= 2P(Z \le a) - 1 \qquad Z \sim N(0,1) \\ &= p \end{aligned}$$

ex. ∵ $0.95 = 2P(Z \le 1.96) - 1$ where $Z \sim N(0,1)$

$e^{-\frac{1.96^2}{2}} = e^{-1.9208} \approx 0.15$

∴ a 15% likelihood interval for θ = an approximate 95% CI for θ.

- likelihood ratio statistic $\Lambda(\theta)$

$$\Lambda(\theta) = -2\log\left(\frac{L(\theta)}{L(\hat{\theta})}\right)$$

relative likelihood $R(\theta)=\frac{L(\theta)}{L(\hat{\theta})}$	$\Lambda(\theta) = -2\log(R(\theta))$
• 用于估计参数. 给定 model. 找参数值.	• 用于选择模型 判断哪个模型更符合观测数据

- Thm

If $Y = (Y_1, Y_2, \cdots, Y_n)$. n: size θ: true value

n large $\Rightarrow \Lambda(\theta)$ 可用作 pivotal quantity.

$n \to \infty \Rightarrow \Lambda(\theta) = -2\log\left(\frac{L(\theta)}{L(\hat{\theta})}\right) \sim \chi^2_1$

Using approximate pivotal quantity to construct an approximate 100p% CI:

$p = 0.95$.

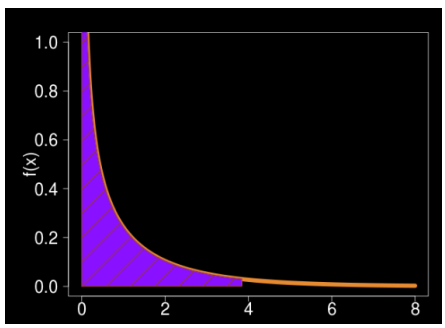
1. Determine c s.t. $P(-2\log\left(\frac{L(\theta;Y)}{L(\hat{\theta};Y)}\right) \le c) \approx p$

$P(W \le c) = 0.95$ $W \sim \chi^2_1$

$\because W = Z^2$ $Z \sim G(0,1)$

$\therefore P(Z^2 \le c) = P(-\sqrt{c} \le z \le \sqrt{c}) = 0.95$.

$c = 1.96$




```
> qchisq(0.95, 1)
[1] 3.841459
```

2. Express the inequality $-2\log\left(\frac{L(\theta;Y)}{L(\hat{\theta};Y)}\right) \le c^2$ in the form $l(Y) \le \theta \le u(Y)$

$\{\theta : l(y) \le \theta \le u(y)\} = \{\theta : -2\log\left(\frac{L(\theta;Y)}{L(\hat{\theta};Y)}\right) \le 1.96^2\}$ 不用 rearrange

3. An approximate p% CI for θ: $[l(y), u(y)]$.

- Approximate CI for $Bi \sim (n, \theta)$

2 methods for obtaining approximate 95% CI:

① 15% LI

② $\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$ $\quad \hat{\theta} = \frac{Y}{n}$

pros & cons: Bi model involves fewer modal assumptions but gives less precise interval

19. The lifetime T (in days) of a particular type of light bulb is assumed to have a distribution with probability density function

$$f(t;\theta) = \frac{1}{2}\theta^3 t^2 e^{-\theta t} \quad \text{for } t > 0 \text{ and } \theta > 0$$

(a) Suppose $t_1, t_2, \ldots, t_n$ is a random sample from this distribution. Find the maximum likelihood estimate $\hat{\theta}$ and the relative likelihood function $R(\theta)$.

(b) If $n = 20$ and $\sum_{i=1}^{20} t_i = 996$, graph $R(\theta)$ and determine the 15% likelihood interval for θ which is also an approximate 95% confidence interval for θ. The interval can be obtained from the graph of $R(\theta)$ or by using the function `uniroot` in R.

(c) Suppose we wish to estimate the mean lifetime of a light bulb. Show $E(T) = 3/\theta$. **Hint:** Use the Gamma function. Find an approximate 95% confidence interval for the mean.

(d) Show that the probability p that a light bulb lasts less than 50 days is

$$\begin{aligned} p &= p(\theta) \\ &= P(T \le 50;\theta) \\ &= 1 - e^{-50\theta}[1250\theta^2 + 50\theta + 1] \end{aligned}$$

Determine the maximum likelihood estimate of p. Find an approximate 95% confidence interval for p from the approximate 95% confidence interval for θ. For the data referred to in (b), the number of light bulbs which lasted less than 50 days was 11 (out of 20). Using a Binomial model, obtain an approximate 95% confidence interval for p. What are the pros and cons of the second interval over the first one?

a) $L(\theta) = \prod_{i=1}^{n} f(t_i;\theta)$

$= \prod_{i=1}^{n} \frac{1}{2}\theta^3 t_i^2 e^{-\theta t_i}$

$= (\frac{1}{2}\theta^3)^n \prod_{i=1}^{n} t_i^2 \cdot e^{-\theta \sum_{i=1}^{n} t_i}$

$L(\theta) = \theta^{3n} \cdot e^{-\theta \sum_{i=1}^{n} t_i}$

$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{3n} e^{-\theta \sum t_i}}{\hat{\theta}^{3n} e^{-\hat{\theta} \sum t_i}} = \frac{\theta^{3n} e^{-\theta \frac{3n}{\hat{\theta}}}}{\hat{\theta}^{3n} e^{-3n}} = \left(\frac{\theta}{\hat{\theta}}\right)^{3n} e^{3n(1-\frac{\theta}{\hat{\theta}})}$

for $\hat{\theta}$: $l(\theta) = 3n \log\theta - \theta \sum_{i=1}^{n} t_i$

$\frac{dl}{d\theta} = \frac{3n}{\theta} - \sum_{i=1}^{n} t_i$

$\hat{\theta} = \frac{3n}{\sum_{i=1}^{n} t_i} \rightarrow \sum_{i=1}^{n} t_i = \frac{3n}{\hat{\theta}}$

b) $\hat{\theta} = \frac{3n}{\sum t_i} = \frac{2\times 20}{996} = \frac{5}{83}$

15% LI: $R(\theta) \ge 15\%$.

c) $E[T] = \int_0^\infty \frac{1}{2}\theta^3 t^2 e^{-\theta t}\, dt$

$= \frac{1}{2}\int_0^\infty (\theta t)^3 e^{-\theta t}\, dt$

$= \frac{1}{2\theta}\int_0^\infty x^3 e^{-x}\, dx \quad x = \theta t$

$= \frac{1}{2\theta}\Gamma(4)$

$= \frac{1}{2\theta} 3! = \frac{3}{\theta}$

95% CI: $\left[\frac{3}{0.0462}, \frac{3}{0.0768}\right]$

d) $P(T \le 50) = \int_0^{50} \frac{\theta^3}{2} t^2 e^{-\theta t}\, dt$

$= \frac{\theta^3}{2}\left[\frac{-2\times 50}{\theta} e^{-50\theta} - \frac{100}{\theta^2} e^{-50\theta} + \frac{2}{\theta^2}\left(-\frac{1}{\theta} e^{-50\theta} + \frac{1}{\theta}\right)\right]$

$= 1 - (1250\theta^2 + 50\theta + 1)e^{-50\theta}$

$[p(0.0461), p(0.0768)] = [0.608, 0.738]$

② Binomial $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{11}{20} \pm 1.96\sqrt{\frac{\frac{11}{20}\cdot\frac{9}{20}}{20}}$

$= [0.332, 0.768]$

Bi model involves fewer model assumptions but gives less precise interval

4.1 CI for $G(\mu, \sigma)$ — μ, σ 未知

- σ^2 的两种算法.

1) m.l.e for σ^2 : $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

2) sample variance estimator : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ ← prefer.

- 将 t distribution 运用到 $G(\mu, \sigma)$

Suppose $Y_1, Y_2, \cdots, Y_n \sim G(\mu, \sigma)$

① σ known

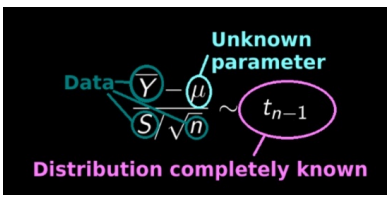
$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim G(0, 1)$$

CI for μ: $\bar{y} \pm a\frac{\sigma}{\sqrt{n}}$ $\quad P(Z \le a) = \frac{1+p}{2}$ $\quad Z \sim G(0,1)$

② σ unknown

completely known → is pivotal quantity.

$$\frac{\bar{Y} - \mu}{\frac{S}{\sqrt{n}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}} \sim t_{n-1}$$



$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

CI for μ : $\bar{y} \pm a\frac{s}{\sqrt{n}}$ $\quad P(T \le a) = \frac{1+p}{2}$ $\quad T \sim t_{n-1}$

∴ 即使 σ 未知, 也可 construct CI for μ.

σ known	σ unknown
Pivotal quantity: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	Pivotal quantity: $\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$
1. Quantiles: $P\left(-a \le \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \le a\right) = p$	1. Quantiles: $P\left(-a \le \frac{\bar{Y}-\mu}{S/\sqrt{n}} \le a\right) = p$
2. Rearrange: $P\left(\bar{Y} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + a\frac{\sigma}{\sqrt{n}}\right) = p$	2. Rearrange: $P\left(\bar{Y} - a\frac{S}{\sqrt{n}} \le \mu \le \bar{Y} + a\frac{S}{\sqrt{n}}\right) = p$
3. CI: $\left[\bar{y} - a\frac{\sigma}{\sqrt{n}}, \bar{y} + a\frac{\sigma}{\sqrt{n}}\right]$ $P(-a \le Z \le a) = p, Z \sim G(0,1)$	3. CI: $\left[\bar{y} - a\frac{s}{\sqrt{n}}, \bar{y} + a\frac{s}{\sqrt{n}}\right]$ $P(-a \le T \le a) = p, T \sim t_{n-1}$

- Sample size calculation : Gaussian data

→ σ 已知

Let Y denote the rent for a randomly selected apartment, and assume Y ~ G (μ, 250).
I'm trying to estimate μ by taking a random sample of apartments currently on the market.
How many apartments should I sample to obtain a 95% confidence interval of width $200?
What about of width $100?

Width of 95% CI : $2a\frac{\sigma}{\sqrt{n}} = 2\times 1.96\times\frac{250}{\sqrt{n}}$

For width \$200 : $2\times 1.96\times\frac{250}{\sqrt{n}} \le 200$ $n = 24.01$ ⇒ choose $n = 25$

For width \$100 : $2\times 1.96\times\frac{250}{\sqrt{n}} \le 100$ $n = 96.04$ ⇒ choose $n = 97$

→ σ 未知 needs estimate

How to construct a CI for σ ?

从 point estimator 入手. form a pivotal quantity.

$Y_1, Y_2, \cdots, Y_n \sim G(\mu, \sigma)$

$E(Y_i) = \mu$ $sd(Y_i) = \sigma$ both unknown. estimate σ.

estimator for σ^2 : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

100p% CI:

CI for μ, σ known	CI for σ^2
Pivotal quantity: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	Pivotal quantity: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
1. Quantiles: $P\left(-a \le \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \le a\right) = p$	1. Quantiles: $P\left(a \le \frac{(n-1)S^2}{\sigma^2} \le b\right) = p$
2. Rearrange: $P\left(\bar{Y} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + a\frac{\sigma}{\sqrt{n}}\right) = p$	2. Rearrange: $P\left(\frac{(n-1)S^2}{b} \le \sigma^2 \le \frac{(n-1)S^2}{a}\right) = p$
3. CI: $\left[\bar{y} - a\frac{\sigma}{\sqrt{n}}, \bar{y} + a\frac{\sigma}{\sqrt{n}}\right]$ $P(-a \le Z \le a) = p, Z \sim G(0,1)$	3. CI: $\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right)$ $P(a \le W) = P(b \ge W) = (1-p)/2, W \sim \chi^2_{n-1}$

不关于 s^2 对称

The professor weighs their sample of 30 TimBites and finds their sample variance is s^2 = 0.311.
Let Y = the weight of a randomly selected TimBite. Y∼G (μ, σ), with μ, σ both unknown.
Construct a 95% confidence interval for σ^2 based on these data.

n=

95% CI : $\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right)$

→ choose a, b s.t $P(W \le a) = 0.025$

$P(W > b) = 0.025$ $\quad W \sim \chi^2_{n-1}$

→
```
> qchisq(0.025, 29)
[1] 16.04707 ← a
> qchisq(0.975, 29)
[1] 45.72229 ← b
```

$P(W \le 16.05) = \frac{1-0.95}{2} = 0.025$

$P(W \le 45.72) = \frac{1+0.95}{2} = 0.975$

→ 95% CI for σ^2 :

$\left(\frac{29 \times 0.311}{45.72}, \frac{29 \times 0.311}{16.05}\right) = (0.190, 0.543)$

sample size small → quantile from t distribution >> quantile from G (0, 1)

sample size ↑ → quantile from t distribution 逐渐接近 G (0, 1)

For the good movies, the sample size was 58, the sample mean was 97.707 minutes,
and the sample standard deviation was 18.229.
Our 95% confidence interval was [92.914, 102.500]. Consider the following changes:

1 Confidence level increases to 99%
2 Sample size increases to 100
3 Sample standard deviation decreases to 10
4 Sample mean decreases to 95

CI ↑ → wider
n ↑ → narrower
σ ↓ → narrower
$\bar{y}$ ↓ → 不变

width ↑

4.8 Chapter 4 Summary

Approximate Confidence Intervals based on Likelihood Intervals

A $100p\%$ likelihood interval is defined as $\{\theta : R(\theta) \geq p\}$ where $R(\theta) = R(\theta; \mathbf{y})$ is the relative likelihood function for θ based on observed data $\mathbf{y}$ (possible a vector). Likelihood intervals must usually be found using a numerical method such as the `uniroot` function in R.

A$100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where $q = P(W \leq -2\log p)$ and $W \sim \chi^2(1)$. (Note: $q =$ `pchisq`$(-2 * \log p, 1)$ in R.)

An approximate $100p\%$ confidence interval is given by a $100\left(e^{-b/2}\right)\%$ likelihood interval where b is the value such that $p = P(W \leq b)$ and $W \sim \chi^2(1)$. (Note: $b =$ `qchisq`$(p, 1)$ in R.)

These results are derived from the fact that $-2\log R(\theta; \mathbf{Y})$ is an asymptotic pivotal quantity with approximately a $\chi^2(1)$ distribution.

CI

Table 4.3
Approximate Confidence Intervals for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Observed Data	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Asymptotic Gaussian Pivotal Quantity	Approximate $100p\%$ Confidence Interval
Binomial(n, θ)	y	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}}{n}}$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\frac{\tilde{\theta}}{\sqrt{n}}}$	$\hat{\theta} \pm a\frac{\hat{\theta}}{\sqrt{n}}$

Note: The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0, 1)$. In R, $a =$ `qnorm`$\left(\frac{1+p}{2}\right)$

Table 4.4
Confidence/Prediction Intervals for Gaussian and Exponential Models

Model	Unknown Quantity	Pivotal Quantity	$100p\%$ Confidence/Prediction Interval
$G(\mu,\sigma)$ σ known	μ	$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	$\bar{y} \pm a\sigma/\sqrt{n}$
$G(\mu,\sigma)$ σ unknown	μ	$\frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\bar{y} \pm bs/\sqrt{n}$
$G(\mu,\sigma)$ μ unknown σ unknown	Y	$\frac{Y-\overline{Y}}{S\sqrt{1+\frac{1}{n}}} \sim t(n-1)$	$100p\%$ Prediction Interval $\bar{y} \pm bs\sqrt{1+\frac{1}{n}}$
$G(\mu,\sigma)$ μ unknown	σ^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\frac{(n-1)s^2}{d}, \frac{(n-1)s^2}{c}\right]$
$G(\mu,\sigma)$ μ unknown	σ	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\sqrt{\frac{(n-1)s^2}{d}}, \sqrt{\frac{(n-1)s^2}{c}}\right]$
Exponential(θ)	θ	$\frac{2n\overline{Y}}{\theta} \sim \chi^2(2n)$	$\left[\frac{2n\bar{y}}{d_1}, \frac{2n\bar{y}}{c_1}\right]$

Notes: (1) The value a is given by $P(Z \le a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
In R, $a =$ `qnorm`$\left(\frac{1+p}{2}\right)$

(2) The value b is given by $P(T \le b) = \frac{1+p}{2}$ where $T \sim t(n-1)$. In R, $b =$ `qt`$\left(\frac{1+p}{2}, n-1\right)$

(3) The values c and d are given by $P(W \le c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n-1)$.
In R, $c =$ `qchisq`$\left(\frac{1-p}{2}, n-1\right)$ and $d =$ `qchisq`$\left(\frac{1+p}{2}, n-1\right)$

(4) The values c_1 and d_1 are given by $P(W \le c_1) = \frac{1-p}{2} = P(W > d_1)$ where $W \sim \chi^2(2n)$.
In R, $c_1 =$ `qchisq`$\left(\frac{1-p}{2}, 2n\right)$ and $d_1 =$ `qchisq`$\left(\frac{1+p}{2}, 2n\right)$

5. Hypothesis Testing

5.1 Introduction.

Test a hypothesis in the light of observed data or information

- Test Statistic / discrepancy measure 差异测量 D

A function of data Y that constructed to measure degree of agreement between data Y and null hypothesis H 数据Y与假设Y no一致程度

$\begin{cases} D = 0 & \text{best possible agreement} \\ D >> 0 & \text{poor agreement} \end{cases}$

- def. p-value

p-value of the test hypothesis H_0 using test statistic D is $P(D \geq d; H_0)$

* probability of observing data 至少和 actual data 一样 no 概率

p-value 小表示：If null hypothesis were true, then unlikely to have observed data at least as surprising as the data actually observed.

Q. Approximate p-values

$Y \sim Bi(100, 0.5)$ $D(Y) = |Y - 50|$

observed value for Y: $y = 52$ observed value for D: $|52 - 50| = 2$.

When H_0 true. What is the probability that the discrepancy measure $\geq d$?

$$\begin{aligned} p\text{-value} &= P(D \geq d; H_0) \\ &= P(|Y - 50| \geq |52 - 50|; H_0) \\ &= P(|Y - 50| \geq 2) \quad Y \sim Bi(100, 0.5) \\ &= 1 - P(49 \leq Y \leq 51) \\ &= 1 - \binom{100}{49} 0.5^{100} - \binom{100}{50} 0.5^{100} - \binom{100}{51} 0.5^{100} \\ &\approx 0.76 \end{aligned}$$

- Steps of hypothesis's test.

1) 列出 null hypothesis H_0 to be tested using Y.

$H_0: \theta = 0.5$. Y = # 'A' answers out of 25 $Y \sim Bi(25, \theta)$

2) Define a test statistic or discrepancy measure $D(Y)$

D越大. H_0越 less consistent.

Let $d = D(y)$ be the corresponding observed value of D.

$D = |Y - E(Y)|$ $d = |15 - 12.5| = 2.5$

3) Calculate p-value $= P(D \geq d)$ assume H_0 is True

p-value $= P(D \geq 2.5)$

which is same as $P(|Y - 12.5| \geq 2.5)$ if $Y \sim Bi(25, 0.5)$

4) Draw a conclusion based on p-value.

p-value	Interpretation
$p > 0.1$	There is no evidence against H_0 based on the data.
$0.05 < p \leq 0.1$	There is weak evidence against H_0 based on the data.
$0.01 < p \leq 0.05$	There is evidence against H_0 based on the data.
$0.001 < p \leq 0.01$	There is strong evidence against H_0 based on the data.
$p \leq 0.001$	There is very strong evidence against H_0 based on the data.

* p值大说明原理解释：

p-value small → H_0 False

p-value large → H_0 true

- Central Limit Thm (CLT) 用于 approximate p-value.

$Y \sim (n, \theta)$ n large $\Rightarrow \frac{Y - n\theta}{\sqrt{n\theta(1-\theta)}} \sim G(0, 1)$

- p-hacking

具体的例子要具体的去分析，不能按一个固定的标准来，比如说设定p小于等于某个数才算重要的话，实验的时候还是他们计算出的p value恰好卡在那个线那里，会有人刻意选一些合适的数据让它们小于那个数来证明它们是重要的，不是真实的数据

5.2 Hypothesis testing for $G(\mu, \sigma)$

- Test H_0 (null hypothesis)

For testing null hypothesis $H_0: \mu = \mu_0$. Use $D = |\bar{Y} - \mu_0|$

$p\text{-value} = P(|\bar{Y} - \mu_0| \geq |\bar{y} - \mu_0|)$ $\quad \bar{Y} \sim G(\mu_0, \frac{\sigma}{\sqrt{n}})$

eg. $Y \sim G(\mu, \sigma)$ → volume of tea in a randomly chosen cup.

test $H_0: \mu = 5$.

We observe a sample of n cups with sample mean $\bar{y} = 5.5$.

discrepancy measure: $D = |\bar{Y} - 5|$

计算 $P(D \geq d; H_0)$:

$d = |\bar{y} - 5| = 5.5 - 5 = 0.5$

$p\text{-value} = P(D \geq 0.5; H_0)$

$= P(|\bar{Y} - 5| \geq 0.5)$

$= P(\bar{Y} \geq 5.5) + P(\bar{Y} \leq 4.5)$

- Test $H_0: \mu = \mu_0$; σ 未知

pivotal quantity

To test $H_0: \mu = \mu_0$; σ 未知. Use $D = \dfrac{|\bar{Y} - \mu_0|}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

$P(D \geq d) = P\left(\dfrac{|\bar{Y} - \mu_0|}{\frac{S}{\sqrt{n}}} \geq \dfrac{|\bar{y} - \mu_0|}{\frac{S}{\sqrt{n}}}\right)$ ← d

test statistic

$= P(|T| \geq d)$ $\quad T \sim t_{n-1}$

$= 2[1 - P(T \leq d)]$

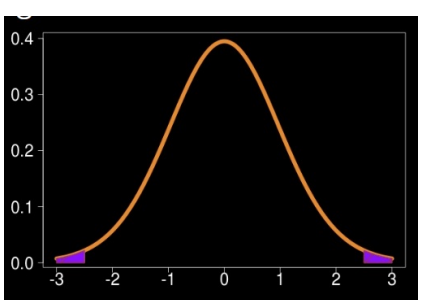
eg. 接上面. $\bar{y} = 5.5$. $S = 1.0$. $\mu_0 = 5.0$ $\mu = 2.5$.

$d = \dfrac{|\bar{y} - \mu_0|}{\frac{S}{\sqrt{n}}} = 2.5$

$p\text{-value} = P(D \geq d) = P(D \geq 2.5)$

$= P(|T| \geq 2.5)$ $\quad T \sim t_{24}$

$= 2(1 - P(T < 2.5))$




```
> pt(2.5, 24)
[1] 0.9901729
> 2*(1 - pt(2.5, 24))
[1] 0.01965418
```

- Test H_0 : $\sigma^2 = \sigma_0^2$, μ 未知

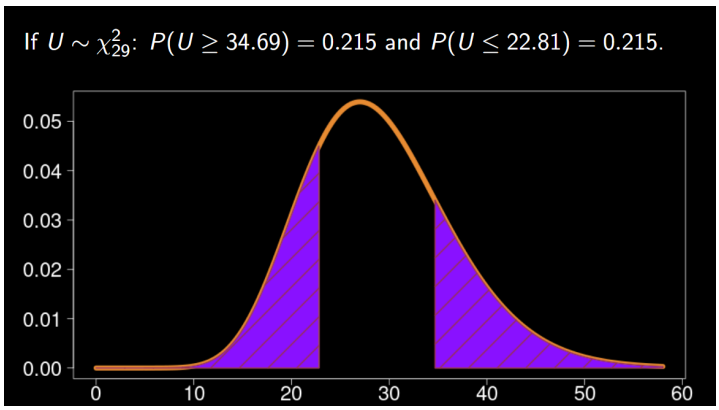
To test H_0 : $\sigma^2 = \sigma_0^2$, μ 未知 Use $D = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

test statistics : $U = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$

$p\text{-value} = P(D' \le -|d'|) + P(D' \ge |d'|)$

eg.

In our TumBites example we're testing $H_0 : \sigma^2 = 0.26$ and found $u = \frac{(n-1)s^2}{\sigma_0^2} = 34.69$ with $n = 30$. We consider χ^2_{29}:



$H_0 : \sigma^2 = \sigma_0^2$ $G(\mu, \sigma^2)$

Step 1. Draw a sample of size n. with variance s^2

Step 2. Compute Discrepancy measure $u = \frac{(n-1)s^2}{\sigma_0^2}$

Step 3. Compute $P(U \le u)$ $U \sim \chi^2_{n-1}$

Step 4. If s^2 large s.t $P(U \ge u) < 0.5$. Then $p = 2P(U \ge u)$

If s^2 small s.t $P(U \le u) \le 0.5$. Then $p = 2P(U \ge u)$

p-value : 描述若 H_0 True. 得到 in data 有多离谱
出现 sample 0 和出现比 Sample 0 还离谱 in 概率
(未知 H_0 与 true value in 偏差)

CI : H_0 与 true value 偏差多少

Statistical significant 用于描述: when a hypothesis's test returns a small p-value

Practical significant 用于描述: when our results have 现实中 in 影响

- Hypothesis & CI

Suppose we test $H_0: \mu = \mu_0$ for $G(\mu, \sigma)$.

$$p\text{-value} \geq 0.05.$$

$$P\left(\frac{|\bar{Y}-\mu_0|}{\frac{S}{\sqrt{n}}} \geq \frac{|\bar{y}-\mu_0|}{\frac{S}{\sqrt{n}}}\right) \geq 0.05$$

$$P\left(|T| \geq \frac{|\bar{y}-\mu_0|}{\frac{S}{\sqrt{n}}}\right) \geq 0.05 \qquad T \sim t_{n-1}$$

$$P\left(|T| \leq \frac{|\bar{y}-\mu_0|}{\frac{S}{\sqrt{n}}}\right) \geq 0.95$$

找 $a \geq \frac{|\bar{y}-\mu_0|}{\frac{S}{\sqrt{n}}}$ s.t $P(|T| \leq a) = 0.95$

$-a \leq \frac{\bar{y}-\mu_0}{\frac{S}{\sqrt{n}}} \leq a$ s.t $P(|T| \leq a) = 0.95$

if $p \geq 0.05$, then $\mu_0 \in [\bar{y} - a\frac{S}{\sqrt{n}}, \bar{y} + a\frac{S}{\sqrt{n}}]$

- If our p-value ≥ 0.05 then μ_0 is inside a 95% CI for μ
- If μ_0 is inside a 95% CI for μ then our p-value ≥ 0.05

A certain company sells fruit juice in cartons. The amount of juice in a carton has a normal distribution with a standard deviation of 3 ml.

The company claims that the mean amount of juice per carton μ is 60 ml. A trading inspector has received complaints that the company is overstating the mean amount of juice per carton and he wishes to investigate this complaint. The trading inspector took a random sample of 16 cartons which gave a mean of 59.1 ml.

Using a 5% level of significance, and stating your hypotheses clearly, test whether or not there is evidence to justify this complaint.

The hypotheses are:

H_0: $\mu = 60$ H_1: $\mu < 60$

This is like the 'evidence' presented at a trial.

The sample gives $n = 16$ and $\bar{x} = 59.1$

$$P(\bar{X} \leqslant 59.1 | \mu = 60) = P\left(Z \leqslant \frac{59.1 - 60}{\frac{3}{4}}\right)$$

$$= P(Z \leqslant -1.2)$$

$$= 0.1151$$

$0.1151 > 0.05$ so the result is not significant and there is insufficient evidence to reject H_0, that $\mu = 60$.

The conclusion should incorporate two statements:

1 State whether or not the test is significant.

2 Interpret this in the context of the question.

There is insufficient evidence to support the complaint.

Remember, H_0 must specify a particular value of μ. The inspector therefore will assume that the company is innocent and wants to formulate a null hypothesis to express this idea in terms of the parameter μ.

If the company is guilty then μ must be less than 60 (there would be few complaints if the cartons contained on average more than 60 ml) and so the alternative hypothesis is H_1: $\mu < 60$. This means the test is one-tailed.

The inspector (like the jury in a trial) then has to calculate the probability of obtaining evidence 'as bad or worse' than this, assuming that the null hypothesis is true.

The alternative hypothesis is that the company is deceiving customers and that $\mu < 60$; the inspector's sample gave a mean of 59.1 and so any value of the sample mean less than or equal to 59.1 will be 'as bad or worse'.

You know that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ so standardise to use the tables.

This probability is greater than the 5% significance level so there is no reason to suspect the validity of H_0.

test statistic:

$$D = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

The following four steps summarise the stages in answering questions about hypothesis tests for the mean μ.

1 Identify the sample mean $\bar{x}$ and value for the population mean given by the null hypothesis.

2 Write down the null (H_0) and alternative (H_1) hypotheses. The alternative hypothesis will determine whether you want a one-tailed or a two-tailed test.

3 Calculate the value of the test statistic $z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

4 Either using the critical region for Z, or by calculating a probability, complete the test and state your conclusions. The following points should be addressed.

a Is the result significant or not?

b What are the implications in terms of the context of the original problem?

5.3 Likelihood ratio test statistic

- likelihood ratio statistic (4.6)

$$\Lambda = -2\log\left(\frac{L(\theta)}{L(\hat{\theta})}\right) = -2\log\left(\frac{L(\theta;Y)}{L(\tilde{\theta};Y)}\right)$$

Likelihood ratio statistic.

$H_0: \theta = \theta_0$.

1. 建立模型. Form $L(\theta)$. Derive expression for $\hat{\theta}$
2. Gather data and calculate $\hat{\theta}$ for observed data.
3. Compute the observed value of test statistic

$$\lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right) = -2\log(R(\theta_0))$$

4. $p\text{-value} \simeq P(W \geq \lambda(\theta_0))$. $W \sim \chi_1^2$

- Likelihood ratio test for Bi model

若 $Y \sim Bi(n,\theta)$. $H_0: \theta = \theta_0$.

我们可通过 $\frac{Y - n\theta}{\sqrt{n\theta(1-\theta)}} \sim G(0,1)$ 找到 CI of θ

→ Approximate pivotal quantity

$$\Lambda = -2\log\left(\frac{L(\theta)}{L(\tilde{\theta})}\right) \sim \chi_1^2$$

1. H_0 True.

$$\Lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\tilde{\theta})}\right) \sim \chi_1^2$$

2. Data inconsistent with H_0

$$\lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right) = -2\log(R(\theta_0))$$

$R(\theta_0)$ is large (close to 1) $\Leftrightarrow$ $-2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right)$ small

$R(\theta_0)$ is small (close to 0) $\Leftrightarrow$ $-2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right)$ large

ex. Suppose I bought 30 pizza and won 6 dipping sauces.

$n=30 \quad y=6 \quad \hat{\theta}=\frac{6}{30}=0.2$

$$R(\theta_0)=\frac{L(\theta_0)}{L(\hat{\theta})}=\frac{\theta_0^y(1-\theta_0)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}}=0.27$$

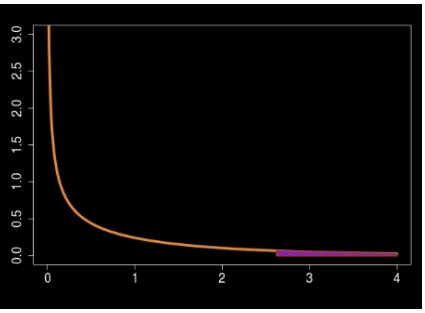
$$\lambda(\theta_0)=-2\log(R(\theta_0))=2.622$$

$p\text{-value} \approx P(W \geq 2.622) \qquad W \sim \chi_1^2$

Using R:

```
> 1 - pchisq(2.621527, 1)
[1] 0.1054229
```

There is no evidence against $H_0 : \theta = \frac{1}{3}$ based on the data!



Q. 3000个人中抽6000人 vote, 4440人 vote "yes".

How valid do you think this estimate is? (5.6 7d)

Since this estimat is not based on random sample.

It is not possible to say how accurate it is.

自愿-投票 → not random

5.5 Chapter 5 Summary

Test of Hypothesis based on Likelihood Ratio Statistic

Suppose $R(\theta) = R(\theta; \mathbf{y})$ is the relative likelihood function for θ based on observed data $\mathbf{y}$ (possible a vector). To test the hypothesis $H_0 : \theta = \theta_0$ we can use the likelihood ratio statistic $-2logR(\theta_0; \mathbf{Y})$ as the test statistic. Let $\lambda = -2logR(\theta_0; \mathbf{y})$ be the observed value of the likelihood ratio statistic for the data $\mathbf{y}$. The corresponding $p-value$ is approximately equal to $P(W \geq \lambda)$ where $W \sim \chi^2(1)$. In R this can be calculated as `1-pchisq(lambda,1)`.

This result is based on the fact that $-2\log R(\theta_0; \mathbf{Y})$ has approximately a $\chi^2(1)$ distribution assuming $H_0 : \theta = \theta_0$ is true.

Table 5.2
Hypothesis Tests for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Test Statistic for $H_0 : \theta = \theta_0$	Approximate $p-value$ based on Gaussian approximation
Binomial(n, θ)	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right)$ $Z \sim G(0,1)$
Poisson(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}\right)$ $Z \sim G(0,1)$
Exponential(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}\right)$ $Z \sim G(0,1)$

Note: To find $2P(Z \geq d)$ where $Z \sim G(0,1)$ in R, use $2 * (1 -$ `pnorm`$(d))$

Table 5.3
Hypothesis Tests for Gaussian and Exponential Models

Model	Hypothesis	Test Statistic	Exact $p-value$
$G(\mu,\sigma)$ σ known	$H_0:\mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{\sigma/\sqrt{n}}$	$2P\left(Z\geq\frac{\lvert\bar{y}-\mu_0\rvert}{\sigma/\sqrt{n}}\right)$ $Z\sim G(0,1)$
$G(\mu,\sigma)$ σ unknown	$H_0:\mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{S/\sqrt{n}}$	$2P\left(T\geq\frac{\lvert\bar{y}-\mu_0\rvert}{s/\sqrt{n}}\right)$ $T\sim t(n-1)$
$G(\mu,\sigma)$ μ unknown	$H_0:\sigma=\sigma_0$	$\frac{(n-1)S^2}{\sigma_0^2}$	$\min(2P\left(W\leq\frac{(n-1)s^2}{\sigma_0^2}\right),$ $2P\left(W\geq\frac{(n-1)s^2}{\sigma_0^2}\right))$ $W\sim\chi^2(n-1)$
Exponential(θ)	$H_0:\theta=\theta_0$	$\frac{2n\bar{Y}}{\theta_0}$	$\min(2P\left(W\leq\frac{2n\bar{y}}{\theta_0}\right),$ $2P\left(W\geq\frac{2n\bar{y}}{\theta_0}\right))$ $W\sim\chi^2(2n)$

Notes:

(1) To find $P(Z\geq d)$ where $Z\sim G(0,1)$ in R, use 1− `pnorm`(d)

(2) To find $P(T\geq d)$ where $T\sim t(k)$ in R, use 1− `pt`(d,k)

(3) To find $P(W\leq d)$ where $W\sim\chi^2(k)$ in R, use `pchisq`(d,k)

6. Gaussian Response Models

6.1 Introduction

- Gaussian response model

One for which the distribution of the response variate Y, given the associated vector of covariates $x = (x_1, x_2, \cdots, x_k)$ for an individual unit.

$$Y \sim G(\mu(x), \sigma(x))$$

↳ 针对 response variate Y in 解释变量.

$$Y_i \sim G(\mu(x_i), \sigma(x_i)) \qquad i = 1, 2, \cdots, n$$

- Gaussian linear model

Assume $\sigma(x_i) = \sigma$ → constant. $\mu(x_i)$ be linear function of the covariates.

↑Model is Gaussian linear model.

$$Y_i \sim G(\mu(x_i), \sigma) \qquad i = 1, 2, \cdots, n \qquad \mu(x_i) = \beta_0 + \sum_{j=0}^{k} \beta_j x_{ij}$$

β_j: regression coefficient

6.2 Linear Regression

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \qquad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)$$

$$\text{corr}(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E[(X - \mu_x)(Y - \mu_x)]}{\sigma_x \sigma_Y}$$

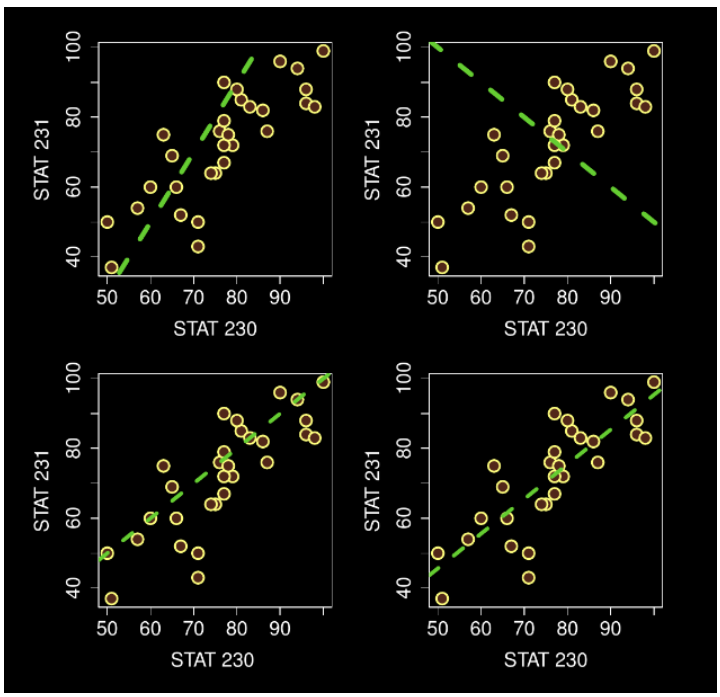
- least square estimate

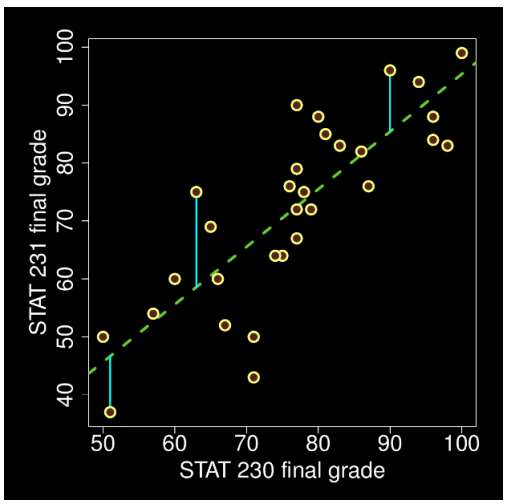
用于找 best fit line : $y = \hat{\alpha} + \hat{\beta} x_i$

How to quantify which line is better than others?

↓

minimizes the distance between itself & data pts





minimize $g(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

通过对 α 和 β 求导，得 $\hat{\alpha}$. $\hat{\beta}$

$$\frac{\partial g}{\partial \alpha} = \frac{\partial}{\partial \alpha} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i)(-1) = 0$$

$$\frac{\partial g}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i)(-x_i) = 0$$

↳ 化简得 $\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \\ \hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = r\sqrt{\dfrac{S_{yy}}{S_{xx}}} \end{cases}$

← least square of α

← least square of β

* $\hat{\beta}$ & r have the same sign
* $S_{yy} >> S_{xx} \Leftrightarrow \hat{\beta}$ is larger than r.

sum of squared residuals : $SS = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$

- Likelihood function for α & β

$Y_i \sim G(\alpha+\beta x_i, \sigma)$ minimize $\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$

σ : the variability in response variate Y in study population

- likelihood function for θ based on y

$$\mathcal{L}(\alpha,\beta)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y_i-\alpha-\beta x_i)^2}$$
$$=e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2} \quad (\text{Since } \sigma \text{ is known})$$

* maximize $\mathcal{L}(\alpha,\beta)$ 相当于 minimize $\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$

↳ $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$ $\hat{\beta}=\frac{S_{xy}}{S_{xx}}$

- Distribution of $\tilde{\beta}$

$\tilde{\beta} \sim G(\mu_\beta, \sigma_\beta)$ $E(\tilde{\beta})=\mu_\beta$ $Var(\tilde{\beta})=\sigma_\beta^2$

$Y_i \sim G(\alpha+\beta x_i, \sigma)$ $\tilde{\beta} \sim G(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$

- Interval estimation

对于固定 x, 可通过 $y=\hat{\alpha}+\hat{\beta}x$ 得到一个 point estimate

$Y_i \sim G(\alpha+\beta x_i, \sigma)$

$$\hat{\beta}=\frac{S_{xy}}{S_{xx}}=\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \qquad \hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}.$$
$$=\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})\,y_i$$

$\tilde{\beta}=\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})Y_i$ is a linear combination of Gaussian random variables Y_i

- Pivotal quantity for β

→ σ known

$\tilde{\beta} \sim G(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$ pivotal quantity: $\frac{\tilde{\beta}-\beta}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim G(0,1)$

→ σ unknown.

$Y \sim G(\mu, \sigma)$ pivotal quantity: $\frac{\tilde{\beta}-\beta}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

σ known	σ unknown
Pivotal quantity: $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	Pivotal quantity: $\frac{\overline{Y}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
1. Quantiles: $P\left(-a \le \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \le a\right) = p$	1. Quantiles: $P\left(-a \le \frac{\overline{Y}-\mu}{S/\sqrt{n}} \le a\right) = p$
2. Rearrange: $P\left(\overline{Y} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{Y} + a\frac{\sigma}{\sqrt{n}}\right) = p$	2. Rearrange: $P\left(\overline{Y} - a\frac{S}{\sqrt{n}} \le \mu \le \overline{Y} + a\frac{S}{\sqrt{n}}\right) = p$
3. CI: $\left[\overline{y} - a\frac{\sigma}{\sqrt{n}}, \overline{y} + a\frac{\sigma}{\sqrt{n}}\right]$ $P(-a \le Z \le a) = p, Z \sim G(0,1)$	3. CI: $\left[\overline{y} - a\frac{s}{\sqrt{n}}, \overline{y} + a\frac{s}{\sqrt{n}}\right]$ $P(-a \le T \le a) = p, T \sim t_{n-1}$

- Estimate σ^2 in simple linear regression

sum of squared errors

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})$$

s_e^2: mean squared error — used to estimate σ^2

s_e^2 is not maximum likelihood estimate of σ^2, but we use it to estimate σ^2, since $E(S_e^2) = \sigma^2$

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2$$

$$\tilde{\beta} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})Y_i \qquad \tilde{\alpha} = \overline{Y} - \tilde{\beta}\overline{x}$$

If S_e^2 is the estimator of the mean squared error, then $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$

degree of freedom: $n-2$

- Interval estimation

$Z \sim G(0,1)$ $\quad U \sim \chi^2_k$ $\quad T = \frac{Z}{\sqrt{\frac{U}{k}}} \sim t_k$

1. Quantities. 找 a:

$P(-a \le T \le a) = p$ $\quad T \sim t_{n-2}$.

$p = P(-a \le T \le a) = P\left(-a \le \frac{\tilde{\beta}-\beta}{\frac{S_e}{\sqrt{S_{xx}}}} \le a\right)$ $\quad$ (Since $\frac{\tilde{\beta}-\beta}{\frac{S_e}{\sqrt{S_{xx}}}} \sim t_{n-2}$)

2. Rearrange. 分出 β.

$p = P\left(\tilde{\beta} - a\frac{S_e}{\sqrt{S_{xx}}} \le \beta \le \tilde{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right)$

3. CI. $100p\%$ CI for β:

$\left[\hat{\beta} - a\frac{S_e}{\sqrt{S_{xx}}}, \hat{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right]$ $\quad P(T \le a) = \frac{1+p}{2}$ $\quad T \sim t_{n-2}$

$G(\mu,\sigma)$, σ unknown	Linear Model
Pivotal quantity: $\frac{\overline{Y}-\mu}{s/\sqrt{n}} \sim t_{n-1}$	Pivotal quantity: $\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}} \sim t_{n-2}$
1. Quantiles: $P\left(-a \le \frac{\overline{Y}-\mu}{s/\sqrt{n}} \le a\right) = p$	1. Quantiles: $P\left(-a \le \frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}} \le a\right) = p$
2. Rearrange: $P\left(\overline{Y} - a\frac{s}{\sqrt{n}} \le \mu \le \overline{Y} + a\frac{s}{\sqrt{n}}\right) = p$	2. Rearrange: $P\left(\tilde{\beta} - a\frac{S_e}{\sqrt{S_{xx}}} \le \beta \le \tilde{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right)$
3. CI: $\left[\overline{y} - a\frac{s}{\sqrt{n}}, \overline{y} + a\frac{s}{\sqrt{n}}\right]$ $P(-a \le T \le a) = p, T \sim t_{n-1}$	3. CI: $\left[\hat{\beta} - a\frac{S_e}{\sqrt{S_{xx}}}, \hat{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right]$ $P(-a \le T \le a) = p, T \sim t_{n-2}$

$100p\%$ CI: $\hat{\beta} \pm a\frac{S_e}{\sqrt{S_{xx}}}$ $\quad$ width $= 2a\frac{S_e}{\sqrt{S_x}}$

Set ↑ ($Y_i \sim G(\alpha+\beta x_i, \sigma)$ 中 ↑ in σ) → width ↑.

variability ↑ → uncertainty ↑

Confidence interval for : $\mu_x = \alpha + \beta x$.

Based on my population of n observations, what is a plausible range of values for the average STAT 231 grade of all students who score x in STAT 230?

point estimate : $\hat{\mu}_x = \hat{\alpha} + \hat{\beta} x = \bar{y} + \hat{\beta}(x - \bar{x})$

estimator : $\tilde{\mu}_x = \tilde{\alpha} + \tilde{\beta} x = \bar{Y} + \tilde{\beta}(x - \bar{x})$

$\therefore \tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$

$\therefore \tilde{\mu}_x = \bar{Y} + \tilde{\beta}(x - \bar{x})$

$= \frac{1}{n} \sum_{i=1}^{n} Y_i + (x - \bar{x}) \cdot \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$

$= \sum_{i=1}^{n} \left(\frac{1}{n} + (x - \bar{x}) \frac{(x_i - \bar{x})}{S_{xx}} \right) Y_i$ $\qquad Y_i \sim G(\alpha + \beta x_i, \sigma)$

↳ $\tilde{\mu}_x \sim G\left(\mu_x, \sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$

→ CI for β

pivotal quantity : $\frac{\tilde{\beta} - \beta}{\frac{S_e}{\sqrt{S_{xx}}}} \sim t_{n-2}$

$100p\%$ CI : $\hat{\beta} \pm \frac{a S_e}{\sqrt{S_{xx}}}$

→ CI for μ_x

pivotal quantity : $\frac{\tilde{\mu}_x - \mu_x}{S_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

$100p\%$ CI : $\hat{\alpha} + \hat{\beta} x \pm a S_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$

↳ $T \sim t_{n-2}$ $\quad P(T \in a) = \frac{1+p}{2}$

Confidence interval for Y at x

individual response

Based on my population of n observations, what is a plausible range of values for the STAT 231 grade of a new student who scored x in STAT 230?

Y = potential observation for given value of x.

$Y = \mu_x + R$ where $R \sim G(0, \sigma) \Rightarrow Y \sim G(\alpha + \beta x, \sigma)$

$\tilde{\mu}_x \sim G\left(\mu_x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$

Goal: $\#$ distribution of $Y - \mu_x$ (error in point estimate of Y)

$$E[Y - \tilde{\mu}_x] = E(R + [\mu_x + \tilde{\mu}_x])$$
$$= E(R) + E[\mu_x] - E[\tilde{\mu}_x]$$
$$= 0$$

$$Var(Y - \tilde{\mu}_x) = Var(Y) + Var(\tilde{\mu}_x)$$
$$= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)$$
$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)$$

$$Y - \tilde{\mu}_x \sim G\left(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

$$\frac{Y - \tilde{\mu}_x}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0, 1)$$

→ $100p\%$ CI for μ_x

$$\hat{\alpha} + \hat{\beta}x \pm a\,se\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

→ $100p\%$ prediction interval for future observation Y:

$$\hat{\alpha} + \hat{\beta}x \pm a\,se\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

6.3 Checking the model

- Linear regression models

Wish to fit the model : $E[Y_i] = \mu(x_i) = \beta_0 + \sum_{j=0}^{k} \beta_j x_{ij}$

find $\beta_0, \beta_1, \cdots, \beta_k$ that minimize $\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2$

- Hypothesis test

For model $y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$.

test $H_0 : \beta_j = 0$ use test statistic $t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} = \frac{\text{estimate}}{\text{standard error}}$

→ H_0 true $\Rightarrow$ $T_j \sim t_{n-k-1}$

ep. model $y = \alpha + \beta x$ test $H_0 : \beta = 0$ use test statistic $\frac{\hat{\beta}}{se/\sqrt{S_{xx}}}$

- Model checking (check model assumption for Gaussian response models)

Two main assumptions:

1) Y_i has a Gaussian distribution with standard deviation σ which doesn't depend on covariates.

a) Y_i has a Gaussian distribution

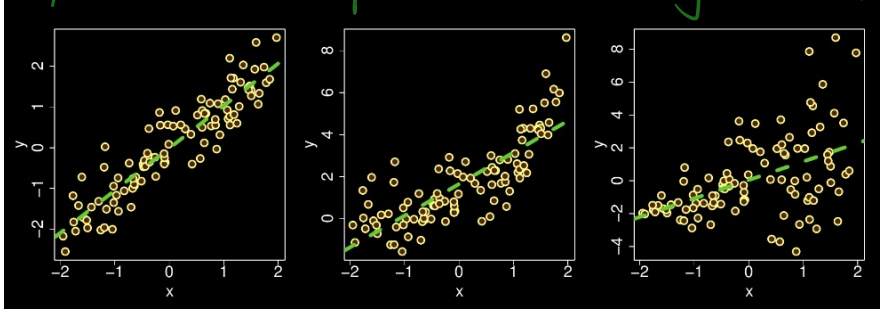
b) The distribution has standard deviation σ which does not depend on the covariates.

2) $E(Y_i) = \mu(x_i)$ is a linear combination of known covariates $x_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$ and unknown regression coefficients $\beta_0, \beta_1, \cdots, \beta_k$

Methods: (for 1b & 2)

1. Scatter plot

判斷 point whether fit reasonably along a straight line.



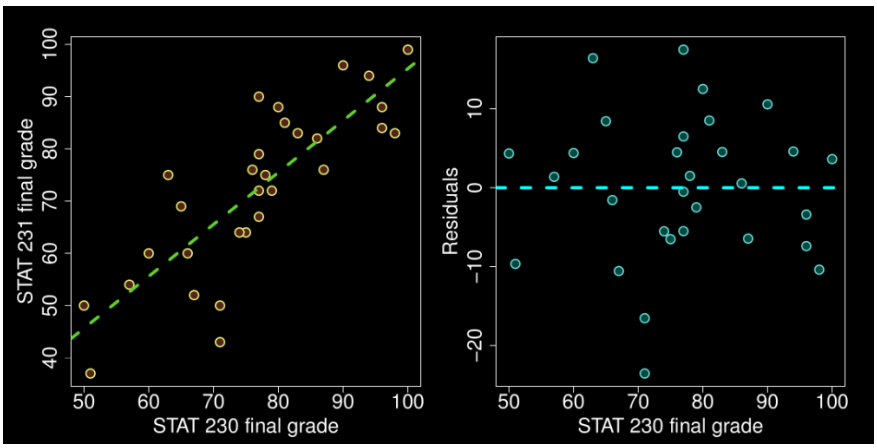
2. Residual plot (针对 Scatterplot hard to read in 情况)

'fitted' response: $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$

residual: $\hat{r}_i = y_i - \hat{\mu}_i$ $\quad Y_i = \mu_i + R_i \quad R_i \sim G(0, \sigma)$

↳ 残差 $\hat{r}_i$ represents what is 'left over' after model has been fitted to the data

standard residual: $\hat{r}_i^* = \frac{\hat{r}_i}{Se} = \frac{y_i - \hat{\mu}_i}{Se}$ $\quad i = 1, 2, \cdots, n$



Standardized residual plot 可用于检测 Scatter plot 相同 in 假设:

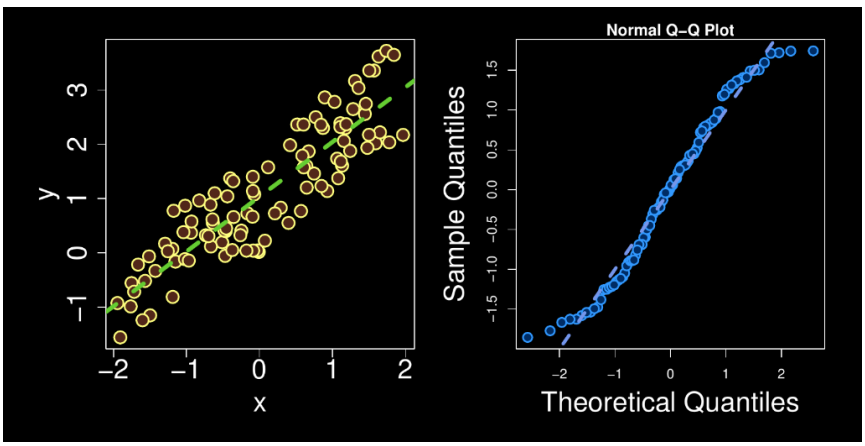
residual plot in 优点:

1) make visualization easier: 只需判断点是否处于一条"平" in 线上

2) more general: 适用于多个 covariate in 情况

Method: (for 1a) Q-Q plot of $\hat{r}_i^* = \frac{\hat{r}_i}{Se} = \frac{y_i - \hat{\mu}_i}{Se}$

Assumed model: $\frac{R_i}{\sigma} = \frac{Y_i - \mu_i}{\sigma} \sim G(0, 1)$



若 assumption model holds, Q-Q plot 应为 a straight line

6.4 Compare means of 2 populations

- 2 Gaussian population with common variance

$Y_{11}, Y_{12}, \cdots, Y_{1n_1} \sim G(\mu_1, \sigma)$

$Y_{21}, Y_{22}, \cdots, Y_{2n_2} \sim G(\mu_2, \sigma)$

The likelihood function for μ_1, μ_2, σ is

$$L(\mu_1, \mu_2, \sigma) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_{ji} - \mu_j)^2\right] \quad \text{for } \mu_1 \in \Re,\ \mu_2 \in \Re,\ \sigma > 0$$

Maximization of the likelihood function gives the maximum likelihood estimates

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = \bar{y}_1$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = \bar{y}_2$$

$$\text{and } \hat{\sigma}^2 = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2\right]$$

An estimate of the variance σ^2 called the *pooled estimate of variance* is

$$\begin{aligned} s_p^2 &= \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2\right] \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{n_1 + n_2}{n_1 + n_2 - 2}\hat{\sigma}^2 \end{aligned}$$

where

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2$$

In particular a $100p\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

A $100p\%$ confidence interval for σ is

$$\left[\sqrt{\frac{(n_1 + n_2 - 2)\, s_p^2}{b}}, \sqrt{\frac{(n_1 + n_2 - 2)\, s_p^2}{a}}\right]$$

where

$$P(U \le a) = \frac{1 - p}{2}, \quad P(U \le b) = \frac{1 + p}{2}, \quad \text{and} \quad U \sim \chi^2(n_1 + n_2 - 2)$$

point estimator of σ^2: $(E(S_p^2) = \sigma^2)$

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

- 2 Gaussian population with unequal variance (unknown)

$Y_{11}, Y_{12}, \cdots, Y_{1n_1} \sim G(\mu_1, \sigma_1)$ $\quad Y_{21}, Y_{22}, \cdots, Y_{2n_2} \sim G(\mu_2, \sigma_2)$

$G(\mu_2, \sigma_2)$ but $\sigma_1 \neq \sigma_2$. If σ_1 and σ_2 are known then we could use the pivotal quantity

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim G(0,1) \tag{6.15}$$

A $100p\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

eg. 95% CI $a = 1.96$

where $P(Z \leq a) = \frac{1+p}{2}$ and $Z \sim G(0,1)$. To test $H_0 : \mu_1 - \mu_2 = 0$ we use the test statistic

$$D = \frac{|\overline{Y}_1 - \overline{Y}_2 - 0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{|\overline{Y}_1 - \overline{Y}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

with

$$p - value = P\left(|Z| \geq \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = 2\left[1 - P\left(Z \leq \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)\right]$$

where $Z \sim G(0,1)$.

Remark The R command `t.test(y1,y2,var.equal=T,conf.level=p)`, where y_1 and y_2 are the data vectors,.will carry out the test above and give a $100p\%$ confidence interval for $\mu_1 - \mu_2$.

- Unpaired: (比较整体)

$Var(Y_{1i}) = \sigma_1^2$ $\quad Var(Y_{2i}) = \sigma_2^2$

$Var(\overline{Y}_1 - \overline{Y}_2) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2Cov(\overline{Y}_1, \overline{Y}_2)$

$d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\quad p = 2[1 - P(T \leq d)]$ $\quad T \sim t_{n_1+n_2-2}$

- Paired. (把两组数放在一起, 1↔1 . 算差)

define $Y_i = Y_{1i} - Y_{2i} \sim G(\mu, \sigma)$ $\quad H_0 : \mu_0 = 0$

$d = \frac{|\bar{y} - 0|}{s/\sqrt{n}}$ $\quad p = 2[1 - P(T \leq d)]$ $\quad T \sim t_{n-1}$

$\tilde{\mu}_1 - \tilde{\mu}_2 = \overline{Y}_1 - \overline{Y}_2$

$Var(\overline{Y}_1 - \overline{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}$

R code

(e) What is a 95% confidence interval for β?

(f) What is a 90% confidence interval for the mean body density of males with a skinfold measurement of 2?

(g) What is a 99% prediction interval for the body density of a male with skinfold measurement of 1.8?

(h) What is a 95% confidence interval for σ?

(e) From the R output

```
> # 95% Confidence interval for slope
> confint(RegModel,level=0.95)
2.5 %       97.5 %
(Intercept) 1.15035436   1.17192390
x      -0.06872823  -0.05540425
```

the 95% confidence interval for β is $[-0.06872823, -0.05540425]$.

(f) From the R output

```
> # 90% confidence interval for mean response at x=2
> predict(RegModel,data.frame("x"=2),interval="confidence",level=0.90)
fit      lwr      upr
1 1.037007 1.034394 1.03962
```

the 90% confidence interval for the mean body density for a skinfold measurement of 2 is $[1.034394, 1.03962]$

(g) From the R output

```
> # 99% prediction interval for response at x=1.8
> predict(RegModel,data.frame("x"=1.8),interval="prediction",level=0.99)
fit     lwr      upr
1 1.04942 1.028503 1.070336
```

a 99% prediction interval for the body density of a male with skinfold measurement of $x = 1.8$ is $[1.028503, 1.070336]$.

(h) From the R output

```
> a<-qchisq(0.025,df)
> b<-qchisq(0.975,df)
> int<-c(se*sqrt(df/b),se*sqrt(df/a))
> cat("95% confidence interval for sigma:  ",int)
95% confidence interval for sigma:  0.006875574 0.009223456
```

the 95% confidence interval for σ is $[0.006875574,\ 0.009223456]$

CI – Linear Regression $y = \alpha + \beta x$

Table 6.1
Confidence/Prediction Intervals for
Simple Linear Regression Model

Unknown Quantity	Estimate	Estimator	Pivotal Quantity	$100p\%$ Confidence/ Prediction Interval
β	$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$	$\tilde{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}$	$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}} \sim t(n-2)$	$\hat{\beta} \pm a s_e/\sqrt{S_{xx}}$
α	$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$	$\tilde{\alpha} = \overline{Y} - \tilde{\beta}\bar{x}$	$\frac{\tilde{\alpha}-\alpha}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	$\hat{\alpha} \pm a s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$
$\mu(x) = \alpha + \beta x$	$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$	$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x$	$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	$\hat{\mu}(x) \pm a s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$
σ^2	$s_e^2 = \frac{S_{yy}-\hat{\beta}S_{xy}}{n-2}$	$S_e^2 = \frac{\sum_{i=1}^{n}(Y_i-\tilde{\alpha}-\tilde{\beta}x_i)^2}{n-2}$	$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$	$\left[\frac{(n-2)s_e^2}{c}, \frac{(n-2)s_e^2}{b}\right]$
Y	$\hat{Y} = \hat{\alpha} + \hat{\beta}x$		$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	Prediction Interval $\hat{\mu}(x) \pm a s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$

Notes: The value a is given by $P(T \leq a) = \frac{1+p}{2}$ where $T \sim t(n-2)$.
The values b and c are given by $P(W \leq b) = \frac{1-p}{2} = P(W > c)$ where $W \sim \chi^2(n-2)$.

H_0 – Linear Regression

Table 6.2
Hypothesis Tests for
Simple Linear Regression Model

Hypothesis	Test Statistic	$p-value$
$H_0 : \beta = \beta_0$	$\frac{\lvert\tilde{\beta}-\beta_0\rvert}{S_e/\sqrt{S_{xx}}}$	$2P\left(T \geq \frac{\lvert\hat{\beta}-\beta_0\rvert}{s_e/\sqrt{S_{xx}}}\right)$ where $T \sim t\,(n-2)$
$H_0 : \alpha = \alpha_0$	$\frac{\lvert\tilde{\alpha}-\alpha_0\rvert}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$	$2P\left(T \geq \frac{\lvert\hat{\alpha}-\alpha_0\rvert}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right)$ where $T \sim t\,(n-2)$
$H_0 : \sigma = \sigma_0$	$\frac{(n-2)S_e^2}{\sigma_0^2}$	$\min\left(2P\left(W \leq \frac{(n-2)s_e^2}{\sigma_0^2}\right), 2P\left(W \geq \frac{(n-2)s_e^2}{\sigma_0^2}\right)\right)$ $W \sim \chi^2\,(n-2)$

Figure 6.4 shows a the scatterplot of the data together with the fitted line, $y = \hat{\alpha}+\hat{\beta}x = 34.05413 + 0.6352523x$. The fitted line passes through the points but we notice that there is a quite a bit of variability about the fitted line.

The $p-value$ for testing $H_0 : \beta = 0$ is

$$\begin{aligned} & 2P\left(T \geq \frac{\left|\hat{\beta}-0\right|}{s_e/\sqrt{S_{xx}}}\right) \\ = \; & 2P\left(T \geq \frac{|0.6352523-0|}{(8.263079)/\sqrt{10813.75}}\right) \\ = \; & 2P\left(T \geq 7.994522\right) \approx 0 \end{aligned}$$

where $T \sim t\,(63)$. Therefore there is very strong evidence against the hypothesis $H_0 : \beta = 0$. This is also consistent with what we see in Figure 6.4. The data suggest there is linear relationship between exam mark and midterm mark.

CI 2 sample

Table 6.3
Confidence Intervals for Two Sample Gaussian Model

Model	Parameter	Pivotal Quantity	$100p\%$ Confidence Interval
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ σ_1, σ_2 known	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\sim G(0,1)$	$\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 = \sigma_2 = \sigma$ σ unknown	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\sim t(n_1 + n_2 - 2)$	$\bar{y}_1 - \bar{y}_2 \pm b s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ μ_1, μ_2 unknown	σ^2	$\frac{(n_1+n_2-2)S_p^2}{\sigma^2}$ $\sim \chi^2(n_1 + n_2 - 2)$	$\left[\frac{(n_1+n_2-2)s_p^2}{d}, \frac{(n_1+n_2-2)s_p^2}{c}\right]$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 \neq \sigma_2$ σ_1, σ_2 unknown	$\mu_1 - \mu_2$	asymptotic Gaussian pivotal quantity $\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ for large n_1, n_2	approximate $100p\%$ confidence interval $\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Notes:

The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.

The value b is given by $P(T \leq b) = \frac{1+p}{2}$ where $T \sim t(n_1 + n_2 - 2)$.

The values c and d are given by $P(W \leq c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n_1 + n_2 - 2)$.

Table 6.4
Hypothesis Tests for
Two Sample Gaussian Model

Model	Hypothesis	Test Statistic	$p-value$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ σ_1, σ_2 known	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ σ unknown	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$2P\left(T\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$ $T\sim t(n_1+n_2-2)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ μ_1, μ_2 unknown	$H_0:\sigma=\sigma_0$	$\frac{(n_1+n_2-2)S_p^2}{\sigma_0^2}$	$\min(2P\left(W\leq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right),$ $2P\left(W\geq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right))$ $W\sim\chi^2(n_1+n_2-2)$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ $\sigma_1\neq\sigma_2$ σ_1, σ_2 unknown	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	approximate $p-value$ $2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$

7. Multinomial models & Goodness of fit tests

7.1 Likelihood Ratio Test

- If n large. H_0 true. Then $\Lambda(\theta_0) = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right) \sim \chi^2_{k-1-p}$

Case: 打台球，每个洞进球概率不一样，distribution 不同

multinomial distribution $f(y_1, y_2, \cdots, y_k; \theta_1, \theta_2, \cdots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$

likelihood function $\mathcal{L}(\theta_1, \theta_2, \cdots, \theta_k) = \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k} = \prod_{j=1}^{k} \theta_j^{y_j}$

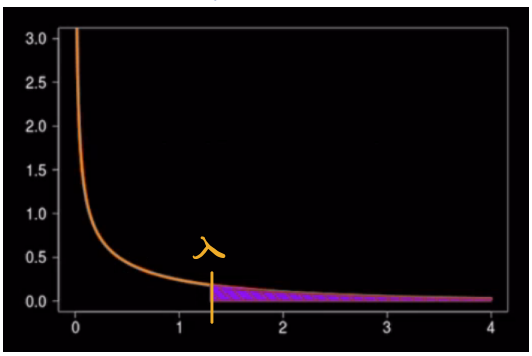
$$\Lambda(\theta_0) = -2\log\left(\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\tilde{\theta})}\right) \qquad \mathcal{L}(\theta_0) = \prod_{j=1}^{k}\left(\frac{1}{k}\right)^{y_j}, \quad \mathcal{L}(\tilde{\theta}) = \prod_{j=1}^{k}\left(\frac{y_j}{n}\right)^{y_j} \rightarrow (\text{since } \mathcal{L}(\theta) = \prod_{j=1}^{k}(\theta_j)^{y_j})$$

$$= -2\log \prod \left(\frac{\left(\frac{1}{n}\right)^{y_j}}{\left(\frac{y_j}{n}\right)^{y_j}}\right)$$

$$= -2\log \prod_{j=1}^{k}\left(\frac{n/k}{Y_j}\right)^{Y_j}$$

$$= -2\log \prod_{j=1}^{k}\left(\frac{E_j}{Y_j}\right)^{Y_j}$$

$$= 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right)$$



observed value: $\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{e_j}\right)$ y_j = observed, e_j = expected

$\begin{cases} y_j = e_j \Rightarrow j \text{ 不会影响 statistic} \\ y_j > e_j \Rightarrow j \text{ 使 statistic 增加} \\ y_j < e_j \Rightarrow j \text{ 使 statistic 下降} \end{cases}$

* category 之间是 dependent 的，有 $y_j > e_j$，必有其它 category: $y_j < e_j$

p-value: $p\text{-value} = P(W \geq \lambda(\theta_0))$ $W \sim \chi^2(k-1-p)$

k = # categories p = # parameters estimated in forming H_0

degree of freedom = # values which "free to move" = # variables − # constraints

- Pearson goodness of fit test ☆

For large n, $D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j} \sim \chi^2_{k-1-p}$ observed value $d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}$

p-value: $p\text{-value} = P(D \geq d)$ $D \sim \chi^2_{k-1-p}$

↑ 当有 observed 与 expected value 时直接套公式，求 p-value.

7.2 Goodness of Fit Tests

- Test of fit of Poisson model

Example 7.2.2 Goodness of fit and Poisson model

The number of service interruptions in a communications system over 200 separate days is summarized in the following frequency table:

Number of interruptions:	0	1	2	3	4	5	>5	Total
Frequency observed y_j:	64	71	42	18	4	1	0	200

Let Y_j = number of times j interruptions are observed. The joint model for the Y_j's is Multinomial.

Y_j is distribution ①

We wish to test whether a Poisson model for Y = the number of interruptions on a single day is consistent with these data. The null hypothesis is

$$H_0 : \theta_j = \frac{\theta^j e^{-\theta}}{j!} \quad \text{for } j = 0, 1, \ldots$$

H_0 : θ_j $\begin{cases} \theta_j = \dfrac{\theta^j e^{-\theta}}{j!} \\ \theta_j = \sum_{y=7}^{\infty} \dfrac{\theta^j e^{-\theta}}{j!} \end{cases}$

(Note that we are using θ rather than α as the parameter of interest.) The maximum likelihood estimate of θ based on the observed data in the table is ②

$$\hat{\theta} = \frac{1}{200}\left[0\,(64) + 1\,(71) + 2\,(42) + 3\,(18) + 4\,(4) + 5\,(1)\right] = \frac{230}{200} = 1.15$$

$\hat{\theta}$

The observed and expected frequencies assuming a Poisson(1.15) distribution are given in the table below

小于5 → using chi-squared approximation to obtain a p-value

No. of interruptions	0	1	2	3	4	≥ 5	Total
y_i	64	71	42	18	4	1	200
e_i	63.33	72.83	41.88	16.05	4.61	1.30	

拒绝 H_0

求出每个 e_j

where

$$e_j = 200\frac{(1.15)^j e^{-1.15}}{j!} \quad \text{for} \quad j = 0, 1, \ldots, 4$$

θ

and the last category is obtained by subtraction. Since the expected frequency in the last category is less than 5 we combine the last two categories to obtain

No. of interruptions	0	1	2	3	≥ 4	Total
y_i (e_i)	64(63.33)	71(72.83)	42(41.88)	18(16.05)	5(5.91)	200

likelihood ratio statistics ③

$$\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{e_j}\right)$$

The observed value of the likelihood ratio statistic is

$$\begin{aligned} & 2\left[64\log\left(\frac{64}{63.33}\right) + 71\log\left(\frac{71}{72.83}\right) + 42\log\left(\frac{42}{41.88}\right) + 18\log\left(\frac{18}{16.05}\right) + 5\log\left(\frac{5}{5.91}\right)\right] \\ = \;& 0.43 \end{aligned}$$

The collapsed table has five categories so $k = 5$ and only one parameter θ has been estimated under H_0 so $p = 1$. The degrees of freedom for the Chi-squared approximation equal $k - 1 - p = 5 - 1 - 1 = 3$. Since

deg of freedom = k-1-p

$$\begin{aligned} p - value &\approx P(W > 0.43) \quad \text{where} \quad W \sim \chi^2(3) \\ &= 0.93 > 0.1 \end{aligned}$$

p-value ④

there is no evidence against the Poisson model based on the observed data.

so 大于 p-value

- Test of fit of Exp model

Example 7.2.3 Goodness of fit and Exponential model

Continuous distributions can also be tested by grouping the data into intervals and then using the Multinomial model. Example 2.6.2 previously did this in an informal way for an Exponential distribution and the lifetimes of brake pads data.

Suppose a random sample $t_1, t_2, \ldots, t_{100}$ is collected and we wish to test the hypothesis that the data come from an Exponential(θ) distribution. We partition the range of T into intervals $j = 1, 2, \ldots, k$, and count the number of observations y_j that fall into each interval. Assuming an Exponential(θ) model, the probability that an observation lies in the j'th interval $I_j = (a_{j-1}, a_j)$ is

$$p_j(\theta) = \int_{a_{j-1}}^{a_j} f(t;\theta)dt = e^{-a_{j-1}/\theta} - e^{-a_j/\theta} \quad \text{for } j = 1, 2, ..., k \qquad (7.8)$$

and if y_j is the number of observations (t's) that lie in I_j, then $Y_1, Y_2, \ldots, Y_k$ follow a Multinomial$(100; p_1(\theta), p_2(\theta), \ldots, p_k(\theta))$ distribution.

Suppose the observed data are

Interval	$0-100$	$100-200$	$200-300$	$300-400$	$400-600$	$600-800$	>800
y_j	29	22	12	10	10	9	8
e_j	27.6	20.0	14.4	10.5	13.1	6.9	7.6

so $k = 7$. To calculate the expected frequencies under the null hypothesis (7.8) we need an estimate of θ which is obtained by maximizing the likelihood function

$$L(\theta) = \prod_{j=1}^{7} [p_j(\theta)]^{y_j}$$

Since there is only one unknown parameter θ under (7.8), $p = 1$. It is possible to maximize $L(\theta)$ to obtain $\hat{\theta} = 310.0$. The expected frequencies, $e_j = 100p_j(\hat{\theta})$, $j = 1, 2, \ldots, 7$, are given in the table.

The observed value of the likelihood ratio statistic (7.5) is

$$2\sum_{j=1}^{7} y_j \log\left(\frac{y_j}{e_j}\right) = 2\left[29\log\left(\frac{29}{27.6}\right) + 22\log\left(\frac{22}{20}\right) + \cdots + 8\log\left(\frac{8}{7.6}\right)\right] = 1.91$$

The degrees of freedom for the Chi-squared approximation equal $k - 1 - p = 7 - 1 - 1 = 5$. Since

$$\begin{aligned} p-value &\approx P(W \geq 1.91) \quad \text{where } W \sim \chi^2(5) \\ &= 0.86 > 0.1 \end{aligned}$$

there is no evidence against the model (7.8) based on the observed data.

A goodness of fit test has some arbitrary elements, since we could have used different intervals and a different number of intervals. Theory has been developed on how best to choose the intervals. For this course we only give rough guidelines which are: chose $4 - 10$ intervals, so that the observed expected frequencies under H_0 are at least 5.

7.3 Two-way Contingency Table

- Cross-Classification of a Random Sample of individuals

$A\backslash B$	B_1	B_2	$\cdots$	B_b	Total
A_1	y_{11}	y_{12}	$\cdots$	y_{1b}	r_1
A_2	y_{21}	y_{22}	$\cdots$	y_{2b}	r_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_a	y_{a1}	$\cdots$	$\cdots$	y_{ab}	r_a
Total	c_1	c_2	$\cdots$	c_b	n

← two-way table

y_{ij} : A type: A_i, B type: B_j

→ test hypothesis

$H_0 : \theta_{ij} = \alpha_i \beta_j$ $\quad \sum_{i=1}^{a} \alpha_i = 1 \quad \sum_{j=1}^{b} \beta_j = 1$

→ expected

likelihood function for y_{ij} : $L(\alpha, \beta) = \prod_{i=1}^{a} \prod_{j=1}^{b} (\alpha_i \beta_j)^{y_{ij}}$

m.l.e : $\hat{\alpha}_i = \frac{r_i}{n}$ $\quad \hat{\beta}_j = \frac{c_j}{n}$

$e_{ij} = n \hat{\alpha}_i \hat{\beta}_j = \frac{r_i c_j}{n}$

→ likelihood ratio statistic for H_0

$$\lambda = 2 \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij} \log \left(\frac{y_{ij}}{e_{ij}} \right)$$

→ deg of freedom

$k - 1 - p = (ab - 1) - (a - 1 + b - 1) = (a - 1)(b - 1)$

→ p-value

$p - value = P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda)$ where $W \sim \chi^2((a-1)(b-1))$

Table 5.1: Guidelines for interpreting $p - values$

$p - value$	Interpretation
$p - value > 0.10$	No evidence against H_0 based on the observed data.
$0.05 < p - value \leq 0.10$	Weak evidence against H_0 based on the observed data.
$0.01 < p - value \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p - value \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p - value \leq 0.001$	Very strong evidence against H_0 based on the observed data.

- Example:

Example: I asked students in a previous class whether they liked hockey and whether their hometown was in Canada. We can present the results in what is known as a 2×2 *contingency table*:

	Canadian hometown	Non-Canadian hometown	Total
Hockey :)	33	9	42
Hockey :(	22	43	65
Total	55	52	107

Is there a relationship between hometown and hockey love?

relative risk of Hockey :) among people with Canadian hometown:

$$\frac{(33/55)}{(9/52)} = 3.467$$

If hometown & hockey love are independent, the table expected to be:

	Canadian hometown	Non-Canadian hometown	Total
Hockey :)	(θ_{11}) 22	(θ_{12}) 20	42
Hockey :(	(θ_{21}) 33	(θ_{22}) 32	65
Total	55	52	107

:)	$\alpha\beta$	$\alpha(1-\beta)$
:(	$(1-\alpha)\beta$	$(1-\alpha)(1-\beta)$

$$\Lambda = 2\left[y_{11}\log\left(\frac{y_{11}}{e_{11}}\right) + y_{12}\log\left(\frac{y_{12}}{e_{12}}\right) + y_{21}\log\left(\frac{y_{21}}{e_{21}}\right) + y_{22}\log\left(\frac{y_{22}}{e_{22}}\right)\right]$$

degree $= k - 1 - p$ → # parameters estimated in forming H_0.

↓ # categories

$k - 1 - p = 2 \times 2 - 1 - 2 = 1$

- `1 - pchisq(21.4034, 1)`

  `[1] 3.72e-6` → p-value $P(W \geq 21.4034)$

8. Causal Relationships

- causation: (x是否导致y)

If all other factors that affect y are held constant,

Let us change x (or observe different value of x) and see if y changes.

Let us change x and see if some specified attribute of y changes.

⇒ If (specified attribute) y changes then x has causal effect on y.

若当所有影响Y in因素不变，x变会导致Y in distribution变化.

则 x has a casual effect of Y.

- Reasons 2 variates can be related

1. explanatory variate (x) 直接导致 response variate (y)
2. response variate (y) 导致 explanatory variate (x) 变化.
3. explanatory variate (x) 导致 response variate (y) 变化. (not only sole)
4. both variates are changing with time.
5. 巧合 (coincidence)
6. both variates results common cause

- 判断: x是否导致y

(whether x is a causal effect on a response variate y)

排除法: 排除可能由x导致 ≠ y (方法: 控制变量x)

1) Hold other possible explanatory variates fixed.
2) Use randomization to control other variates.

- Randomization (establish causality from observational data)

1) The association between the two variates must be observed in many studies of different types among different groups. This reduces the chance that an observed association is due to a defect in one type of study or a peculiarity in one group of subjects.

2) The association must continue to hold when the effects of plausible confounding variates are taken into account.
Many possible sources of confounding variates have been examined in these studies and have not been found to explain the association.

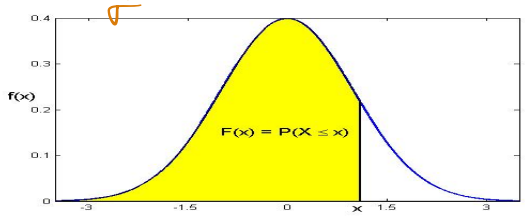
3) There must be a plausible scientific explanation for the direct influence of one variate on the other variate, so that a causal link does not depend on the observed association alone.

4) There must be a consistent response, that is, one variate always increases (decreases) as the other variate increases.

The evidence for causation here is about as strong as non-experimental evidence can be.

N(0,1) Cumulative Distribution Function

$z = \frac{x-\mu}{\sigma}$



This table gives values of $F(x) = P(X \le x)$ for $X \sim N(0,1)$ and $x \ge 0$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

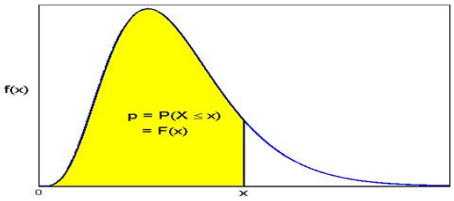
N(0,1) Quantiles: This table gives values of $F^{-1}(p)$ for $p \ge 0.5$

p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.075	0.08	0.09	0.095
0.5	0.0000	0.0251	0.0502	0.0753	0.1004	0.1257	0.1510	0.1764	0.1891	0.2019	0.2275	0.2404
0.6	0.2533	0.2793	0.3055	0.3319	0.3585	0.3853	0.4125	0.4399	0.4538	0.4677	0.4959	0.5101
0.7	0.5244	0.5534	0.5828	0.6128	0.6433	0.6745	0.7063	0.7388	0.7554	0.7722	0.8064	0.8239
0.8	0.8416	0.8779	0.9154	0.9542	0.9945	1.0364	1.0803	1.1264	1.1503	1.1750	1.2265	1.2536
0.9	1.2816	1.3408	1.4051	1.4758	1.5548	1.6449	1.7507	1.8808	1.9600	2.0537	2.3263	2.5758

$X \sim \chi^2(10)$ $P(X \le 2.6) \approx 0.01$ pchisq (2.6, 10)

$X \sim \chi^2(10)$ $P(X \le a) = 0.01$ qchisq (0.01, 10)

Chi-Squared Quantiles



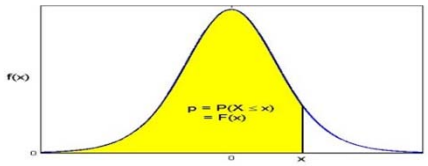
This table gives values of x for p = P(X ≤ x) = F(x)

df\p	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	2.706	3.842	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.146	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.054	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.391	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.430	104.210
80	51.172	53.540	57.153	60.391	64.278	96.578	101.880	106.630	112.330	116.320
90	59.196	61.754	65.647	69.126	73.291	107.570	113.150	118.140	124.120	128.300
100	67.328	70.065	74.222	77.929	82.358	118.500	124.340	129.560	135.810	140.170

$T \sim t(10)$ $P(T \le 0.88) = 0.8$ pt(0.88, 10)

$T \sim t(10)$ $P(T \le a) = 0.8$ qt(0.8, 10)

Student t Quantiles



This table gives values of x for $p = P(X \le x) = F(x)$, for $p \ge 0.6$

df \ p	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.3249	0.7265	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	0.2887	0.6172	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	0.2767	0.5844	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	0.2707	0.5686	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.2672	0.5594	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.2648	0.5534	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.2632	0.5491	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.2619	0.5459	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.2610	0.5435	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.2602	0.5415	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.2596	0.5399	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.2590	0.5386	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.2586	0.5375	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.2582	0.5366	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.2579	0.5357	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
16	0.2576	0.5350	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
17	0.2573	0.5344	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
18	0.2571	0.5338	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216
19	0.2569	0.5333	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
20	0.2567	0.5329	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
21	0.2566	0.5325	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193
22	0.2564	0.5321	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921
23	0.2563	0.5317	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676
24	0.2562	0.5314	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454
25	0.2561	0.5312	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251
26	0.2560	0.5309	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066
27	0.2559	0.5306	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896
28	0.2558	0.5304	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739
29	0.2557	0.5302	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594
30	0.2556	0.5300	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
40	0.2550	0.5286	0.8507	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069	3.5510
50	0.2547	0.5278	0.8489	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614	3.4960
60	0.2545	0.5272	0.8477	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317	3.4602
70	0.2543	0.5268	0.8468	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108	3.4350
80	0.2542	0.5265	0.8461	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953	3.4163
90	0.2541	0.5263	0.8456	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019
100	0.2540	0.5261	0.8452	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
>100	0.2535	0.5247	0.8423	1.2832	1.6479	1.9647	2.3338	2.5857	3.1066	3.3101