Stats.

ANOVA:

Analysis of variance.

Main idea:

figure out how much of the total variance comes from:

The variance <u>between</u> groups

———— <u>within</u> groups.

linear regression assumption:

- error 的 mean 为 0
- 对于所有X, ε的方差为 constant
  ↓观测
- 误差项ε是一个服从正态分布之 R.V.
  且相互独立, $\varepsilon \sim N(0, \sigma^2)$

线性: 因变量与自变量之间的线性关系.

正态: 因变量的正态性.

残差: 独立同分布.

正交: 残差与自变量不相关.
期望为0

0 同 残 独 M

灵童 残线长 Master.

Lufy　冰线呆头　麦哲伦　Ceasar.

Wikipedia 上答案:

1. Weak exogeneity:
   即 predictor variable x can be treated as fixed values, rather than r.v.
2. Linearity:
   response variable is a linear combination of the parameters (regression coefficients)
3. constant variance.
4. error 之间互相独立.
5. predictor 之间 No multicolinearity

AMS
161017,

The common tests arranged:

1 sample proportions	1 sample t	
2 sample proportions.	2 Sample t 2 Sample t paired t	Correlation/ Regression.
	One-way ANOVA.	

categorical data
分类即可.

Quantitative. data.

Q2 How many samples do you have?

→ or how many groups do you have.

→ historically，你通常有两个。

- test group：做了某种test的
- Control group：用于比较，没做任何test的。

One sample:

→ Usually ~~I~~ means I make a comparison against a historical or global value.

Two samples:

→ Traditional control group vs. test group in an experiment.

→ comparing one group to the other group.

Confidence intervals.

$\bar{X} \pm t \frac{s}{\sqrt{n}}$ - 总体方差未知，用 t - statistics.

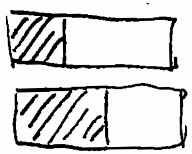
$\bar{X} \pm z \frac{\sigma}{\sqrt{n}}$ 总体方差已知，用 z - statistics.

Choosing which stat to use:

means, proportions, relationships
■ ● ▲

Test for proportion ●

Difference of two proportions ●



Chi-sq test for independence ▲

Difference of two means (independent samples) ■

Regression analysis ▲

Test for a mean ■

Difference of two means paired ■

Quantitative research,
most likely, you perform some hypothesis test in the end.
→ 哪种 test? 由你 collected 的 data 所决定.

Many possibilities:

- Estimate population propotion
- Estimate ——— mean
- One sample proportion
- Two ———
- One sample t (mean)
- Two sample t (mean)
- paired t
- correlation/regression analysis.
- One way ANOVA
- Two way ANOVA
- Chi square ~~test~~
- One sample variance
- Two sample ———

One sample: against hypothesized value.

- mean 250g
- 82% export quality.

Two samples:

A B

compared with each other:

- Heavier?
- No difference?
- More export quality?

One sample,
two measurements.

- circumference
- Weight.

or: color grade.

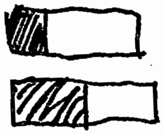
Test
for propotion

One Sample

Test for a mean

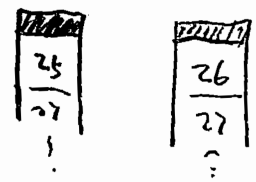
合适的比例…

Difference of
two propotions.



Two samples.

Differences of two



Two sets of information
on the sample

Chi-sq test
for independence

Regression analysis.

Difference of
two means,
paired.

to how many nuts chocolate sold v.s. weather

or male white chocolate
female milk chocolate

哪种性别更prefer哪种.

chi-sq & regression

都是用来look for ~~regression~~ relationships.

	D	M	W
men			
women			

数据是个table

regression: 数据是个散点.
scatter plot.

purpose:

- test against a hypothesis value
- comparing two statistics.
- looking for a Relationship.

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$$

但是

$\mathrm{cov}(X,Y)$ 不好用.

因为 X, Y 是有 units 的.

$$\underset{(X-\mu_X)}{1} \quad \underset{(Y-\mu_Y)}{-1\times10^6}$$

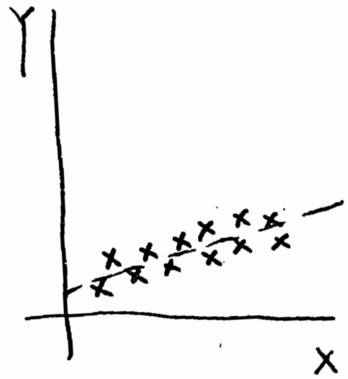
~~[illegible]~~ 90

$$\underset{(X-\mu_X)}{10^6} \quad \underset{(Y-\mu_Y)}{-1}$$

取值一样. 但实际数字可以差很多!!

所以人更常用是 correlation, normalize it, to make it unitless.

$$\mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot \mathrm{Var}(Y)}}$$

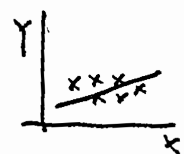


散点图说明 X, Y 正相关. $\mathrm{cov}(X,Y) > 0$

就是 if $(X > \mu_X)$, 欲保证 $\mathrm{cov}(X,Y)>0$.

须 $Y > \mu_Y$

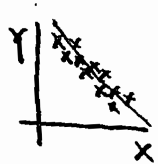
正相关 $\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$



			cov
$X > \mu_X$	+	+	+
$X < \mu_X$	−	−	+

X, 与 Y 同涨或同跌 (cov > 0)

负相关 $\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$



			cov
$X > \mu_X$	+	−	−
$X < \mu_X$	−	+	−

X 涨, 则 Y 跌 (cov < 0)

Confidence intervals are based on sample data, give a range of plausible value for a <u>parameter</u>.

We never construct confidence intervals for statistics (like $\bar{x}$)

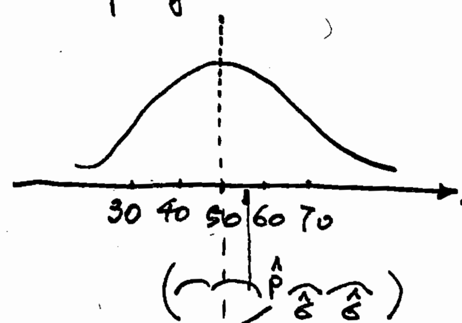
for parameters.

最common之choice.

We may be 95% confident that μ lies in the interval $(-0.2, 3.1)$

50% students like iphone.
you sample 25.

Sampling dist.



这条线是 a pile of $\hat{p}$.

true p

$\hat{p}$是statistics.
由sampling得到的
sampling取样的.

$\mu_{\hat{p}} = p$

mean of the sample distribution ($\hat{p}$).

$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.5(.5)}{25}} = .10$

std of the sample dist.

通过sample

你得到一个$\hat{p}$
往上,往下两个σ
得到一个置信区间.

真实的p落在此区间中的概率是95%

$\hat{p}$的分布满足 sampling distribution.
是个正态分布.

I don't know what my true p is, all I know is my $\hat{p}$

from sampling

cheby:

描述的是：~~[illegible]~~.

越远离均值，可能性越小.

这个可能性由方差衡量

衡量偏离均值的程度.

$$P(|X-\mu| \geq a) \leq \frac{Var(X)}{a^2}$$

什么叫 Inference?

有 10,000 个苹果, 求平均质量.

你选了 100 个 —— 推断 → 总体, 即称为 Inference.

Sampling

↓ 由样本推总体.

S_1 → 150g, S_2 → 130g, S_3 → 140g.

—— 总是有 sampling error 的!

- confidence interval depends on 什么?

- variance
- sample size.

Population with low variation

↓

samples w. low variation

↓

narrow confidence interval.

Small samples vary more from each other, and have less info.

↓ leads to

wider confidence intervals.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

→ σ 大

→ n 小.

由中心极限定理所推得的!!

Confidence Interval for a mean

$$\bar{X} \pm t \frac{s}{\sqrt{n}}$$

s → sample std. $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

n: sample size

$\bar{X}$ → 由sample 得到之平均值

t → 由置信度或显著性水平所得之值，查表可得

t	Sample size				
	10	15	20	30	100
90%	1.833	1.761	1.729		
95%	2.262	2.145			
99%					

$\frac{s}{\sqrt{n}}$ gives us standard error.

→ 为什么是 $\sqrt{n}$?

→ reflects the more ~~information~~ sample we have, the less information we get from new samples.

$t \frac{s}{\sqrt{n}}$: margin error.

We're 90% confident

95%

99%

要求越精确，Confident interval 越大.

总体是一个随机变量 $X \sim F(x)$

样本: $(X_1, \cdots, X_n)$

实际上是n维随机变量

→ n维之联合分布.

若为正态, 则分量必为正态.

→ 加减操作亦为正态.

- X, Y 各自正态 $\not\to$ (X, Y) 正态

可以保证.

- X, Y 独立正态 $\longrightarrow$ $X+Y$ 正态

独立很重要, 加了独立, 怎么折腾都行.

→ (X, Y) 正态 $N(\cdots, 0)$ → $\rho = 0$! 这是独立非常美之结果.

二维亦正态.

- 简单随机样本. → i.i.d.

X: 总体.

$X_1, X_2, \cdots, X_n$ ① X_i 独立 ② $X_i \sim F(x)$ 独立同分布 (i.i.d).

样本之分布函数.

$F(X_1, X_2, \cdots, X_n) = P(X_1 \le x_1, X_2 \le x_2, \cdots, X_n \le x_n)$ 独立

n维随机变量

$= P(X_1 \le x_1) \cdots P(X_n \le x_n)$

$= F(x_1) \cdots F(x_n) = \prod_{i=1}^{n} F(x_i)$

即把总体之分布函数拿来，x换成x_i，连乘起来.

样本之分布函数

类似的，

样本之概率密度

$X \sim f(x)$

$(X_1, X_2, \cdots, X_n)$ —— $\prod_{i=1}^{n} f(x_i)$

简单随机样本

样本之概率密度函数.

样本之数字特征.

$\bar{X} = \frac{1}{n} \sum_{1}^{n} X_i$

样本之均值是n个随机变量求平均

$S^2 = \frac{1}{n-1} \sum_{1}^{n} (X_i - \bar{X})^2$

定义样本之方差为$\frac{1}{n-1}$

$E\bar{X} = \frac{1}{n} \sum_{1}^{n} EX_i = \mu$ 就是总体之期望.

ES^2

DS^2

$D\bar{X} =$

$\bar{X}$, S^2, $E\bar{X}$, $D\bar{X}$, ES^2, DS^2 统计量.

e.g. $f(x) = \begin{cases} 2x & 0<x<1 \\ 0 & \text{其他} \end{cases}$

$X_{(4)} = \max(X_1, X_2, X_3, X_4)$ $\quad X_1, \dots, X_4$ 独立:

求 $f_{X_{(4)}}(x)$

Soln.

$P(X_1 \le x)P(X_2 \le x)P(X_3 \le x)P(X_4 \le x)$

$|||$

★. $F_{X_{(4)}}(x) = P(X_4 \le x) = P(\max(X_1, X_2, X_3, X_4) \le x)$

$= [F(x)]^4$

于是先写 $F(x)$.

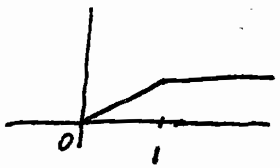
$$F(x) = \begin{cases} 0 & x<0 \\ x^2 & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$$

$$F_{X_{(4)}}(x) = F^4(x) = \begin{cases} 0 & x<0 \\ x^8 & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$$

$$f_{X_{(4)}}(x) = \begin{cases} 8x^7 & 0 \le x < 1 \\ 0 & \text{其他} \end{cases}$$

$X \sim U(0,1)$

分布函数:



$\bar{X}, S^2, E\bar{X}, D\bar{X}, ES^2, DS^2$

χ^2, t, F

1). Chi

$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$

$\sim \chi^2(n)$

$f(x) =$

不必记.

$E(\chi^2) = n,$

$D(\chi^2) = 2n.$

X_i 的服从.

① $X_i \sim N(0,1)$ 标准正态

② X_i 独立

当然可以 $\int x f(x)$

Expectation 线性.

$E\chi^2 = E(X_1^2 + \cdots + X_n^2) = EX_1^2 + \cdots + EX_n^2$

典型模式

$= \underset{1}{\underset{\shortparallel}{DX_1}} + \underset{0}{\underset{\shortparallel}{(EX_1)}}^2 + \cdots$

X_i 标准正态.

$= n$

分布:

$$T=\frac{X}{\sqrt{Y/n}} \sim t(n) \qquad \begin{cases} X:\text{标准正态}. \\ Y\sim \chi^2(n) \\ X,Y\text{独立}. \end{cases}$$

t分布最重要，密度函不必掌握

↓但是个偶函数. α分位点对称.

F分布:

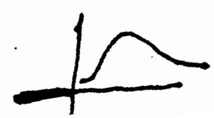
$$F=\frac{X/n_1}{Y/n_2} \sim F(n_1,n_2) \qquad \begin{cases} ① \ X\sim \chi^2(n_1) \\ ② \ Y\sim \chi^2(n_2) \\ ③ \ X,Y\text{独立}. \end{cases}$$

$\frac{1}{F}$ 亦为F分布

$\frac{1}{F} \sim F(n_2,n_1)$

χ^2	t	F
$X_1^2+X_2^2+\cdots+X_n^2$	$\frac{X}{\sqrt{Y/n}}$	$\frac{X/n_1}{Y/n_2}$
$X_i \sim N(0,1)$ 独立	$X\sim N(0,1)$ $Y\sim \chi^2(n)$ X,Y独立	$X\sim \chi^2(n_1)$ $Y\sim \chi^2(n_2)$ X,Y独立.
$EX^2=n$ $DX^2=2n$	$f(x)$ 偶函数	$\frac{1}{F}\sim F(n_2,n_1)$

平方和.
取值均为正.



正态总体. $X\sim N(\mu,\sigma^2)$ 总体正态

样本 $x_1, x_2, \cdots, x_n$, i.i.d $\sim N(\mu,\sigma^2)$

联合起来是n维正态.

1.

$$\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

sample mean 不是n维正态了!

2. $\bar{X}$ 与 S^2 独立.

3. $$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

两个正态总体. 暂略.

$\sum_1^n (X_i-\bar{X})^2$ 自由度只有n−1个!!

因为 $\sum_1^n (X_i-\bar{X})\equiv 0$

有一个约束! 看似n个, 实际是n−1个自由度

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

$$(n-1)\,S^2=\frac{\sum_1^n (X_i-\bar{X})^2}{n-1}=\sum_1^n (X_i-\bar{X})^2$$

$$\frac{(n-1)S^2}{\sigma^2}=\sum_1^n\frac{(X_i-\bar{X})^2}{\sigma^2}$$

X_i: 正态.

$\bar{X}=\frac{1}{n}(X_1+\cdots+X_n)$ 正态

σ: 常数

$$E\left(\frac{X_i-\bar{X}}{\sigma}\right)=\frac{EX_i-E\bar{X}}{\sigma}=0$$

$$D\left(\frac{(n-1)S^2}{\sigma^2}\right)=2(n-1).$$

常数提出, 要平方.

$$\frac{(n-1)^2}{\sigma^4}D(S^2)=2(n-1)\quad\Rightarrow\quad DS^2=\frac{2\sigma^4}{n-1}$$

$X \sim B(1, p)$

X服从 0,1 分布.

即一次 Bernoulli 试验

一个简单随机样本，$X_1, X_2, \cdots, X_n$.

求 $X_1, X_2, \cdots, X_n$ 之分布律.

X	0	1
	q	p

Soln: 难点在于如何把离散分布列写成分布函数

分布列不太好写.

$$P(X=k) = C_1^k \, p^k q^{1-k} \qquad k=0,1 \qquad C_1^k = C_1^0 = 1, \; C_1^1 = 1$$

$$= p^k q^{1-k}$$

分布函数.

$$P(x) = \begin{cases} p^x q^{1-x} & x=0,1 \\ 0 & 其他 \end{cases}$$

乘法定律（把 x 换成 x_i）

$$P(X_1, \cdots, X_n) = \begin{cases} p^{x_1+\cdots+x_n} q^{n-x_1-\cdots-x_n} & x_i=0,1 \\ 0 & 其他 \end{cases}$$

$$= \begin{cases} p^{n\bar{x}} q^{n(1-\bar{x})} & x_i=0,1 \\ 0 & 其他. \end{cases}$$

2. $X \sim P(\lambda)$ $\bar{X}$ 分布律.

~~$\bar{X}$~~

Soln:

$\bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$

X_i 样本! 必独立.

$X_1 + X_2 \sim P(2\lambda)$ 对于泊松分布, 有可加性.

$\sum_{1}^{n} X_i \sim P(n\lambda)$

即 $P(\sum_{1}^{n} X_i = k) = \frac{(n\lambda)^k}{k!} e^{-n\lambda}$ $k = 0, 1, 2, \cdots, \cdots$

于是

$P(\frac{1}{n}\sum_{1}^{n} X_i = \frac{k}{n}) = P(\bar{X} = \frac{k}{n}) = \frac{\lambda^k}{k!} e^{-\lambda}$

不再是泊松分布

$X_1, X_2, X_3, X_4 \sim N(0, 2^2)$ 样本
(简单随机)

$$\chi^2 = a(X_1-2X_2)^2 + b(3X_3-4X_4)^2 \sim \chi^2(n).$$

Soln:

- 随机变量函数之分布. 用什么? 定义法可以但巨麻烦
- χ^2 用典型模式!!

$X_1-2X_2 \sim N(0, 20)$

$D(X_1-2X_2) = D(X_1) + 4DX_2 = 5\sigma^2$

$\sigma^2 = 2^2$

非标准正态, 于是标准化.

$\frac{X_1-2X_2}{\sqrt{20}} \sim N(0,1)$

于是令 $a = \frac{1}{20}$

$\left(\frac{X_1-2X_2}{\sqrt{20}}\right)^2$ 是一个标准正态之平方.

于是 $n=2$

何能 a 或 $b=0$. 于是 $n=1$.

4. $T \sim t(n)$

则 T^2 服从何分布？

Soln:

$T \sim t(n)$

$T = \frac{X}{\sqrt{Y/n}}$, $T^2 = \frac{X^2}{Y/n} \Rightarrow F(1,n)$.

$\chi^2(1)$

$\chi^2(n)$?

看到 χ^2, t, F, 不要用定义做！

因为不记它们 $f(x)$.

Stat 11.

$X_i \sim N(0,1)$ 来自于标准正态总体的简单随机样本

问 1. $n\bar{X} \sim N(0,1)$?

2. $nS^2 \sim \chi^2(n)$

3. $\dfrac{(n-1)\bar{X}}{S} \sim t(n-1)$.

Xi相互独立

1. $n\bar{X} = \sum_{1}^{n} X_i \sim N(0,n)$

$\bar{X} \sim N(0, \frac{1}{n})$ 方差相加.

$n\bar{X} \sim N(0,n)$

n乘进去,要乘以n^2

2. $S^2 = \dfrac{1}{n-1}\sum (X_i - \bar{X})^2$

$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ ✓

$nS^2 \sim \chi^2(n)$ ✗

S^2中隐含n.

3. $\sigma^2 = 1$

$(n-1)S^2 \sim \chi^2(n-1)$

$\dfrac{\bar{X}}{\sqrt{Y/n}}$

$\dfrac{(n-1)^2\bar{X}}{(n-1)S^2} = \dfrac{\bar{X}}{(n-1)S^2/(n-1)^2}$

亦不对.

$X \sim N(0,\sigma^2)$, $X_1,\cdots,X_9$

9个样本.

求σ使

$P(1<\bar{X}<3)$最大.

Soln:

$$\bar{X} = \frac{1}{n}\sum X_i \sim N\left(0, \frac{\sigma^2}{9}\right)$$

$\Downarrow n=9$

于是标准化

$$\frac{\bar{X}-0}{\sigma/3} = \frac{3\bar{X}}{\sigma}$$

$$P\left(\frac{3}{\sigma}<\frac{3\bar{X}}{\sigma}\le\frac{9}{\sigma}\right) = \Phi\left(\frac{9}{\sigma}\right) - \Phi\left(\frac{3}{\sigma}\right) = P(\sigma)$$

$\downarrow$ 标准正态!

$$P'(\sigma) = \varphi\left(\frac{9}{\sigma}\right)\left(-\frac{9}{\sigma^2}\right) - \varphi\left(\frac{3}{\sigma}\right)\left(\frac{-3}{\sigma^2}\right) = 0$$

$$\varphi\left(\frac{9}{\sigma}\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{81}{2\sigma^2}}$$

$\vdots$

$$\sigma = \frac{6}{\sqrt{\ln 3}}.$$

X_1, X_2, X_3 独立 $N(0,\sigma^2)$

证明：$\sqrt{\frac{2}{3}}\,\frac{X_1+X_2+X_3}{|X_2-X_3|} \sim t(1)$

Soln: 典型模式 $T=\frac{X}{\sqrt{Y/1}}$ ① $X\sim N(0,1)$ ② $Y\sim \chi^2(1)$ ③ X, Y 独立.

$(X_1+X_2+X_3) \sim N(0, 3\sigma^2)$

$\frac{1}{\sqrt{3}}(X_1+X_2+X_3) \sim N(0,\sigma^2)$

标准化一下

$\frac{X_1+X_2+X_3}{\sqrt{3}\sigma} \sim N(0,1)$.

$Y = \frac{|X_2-X_3|}{\sqrt{2}\sigma}$ $\qquad X_2-X_3 \sim N(0, 2\sigma^2)$

$Y^2 = \frac{(X_2-X_3)^2}{2\sigma^2}$ $\qquad \frac{X_2-X_3}{\sqrt{2}\sigma} \sim N(0,1)$.

$= \left(\frac{X_2-X_3}{\sqrt{2}\sigma}\right)^2$

标准正态.

$Y^2 \sim \chi^2(1) \Rightarrow T=\frac{X}{\sqrt{Y}} \sim t(1)$

needs证.

$X_1+X_2+X_3$ 与 X_2-X_3 独立.

即证.

X_2+X_3 与 X_2-X_3 独立.

X_1, X_2, X_3 独立, 则

X_2+X_3 与 X_2-X_3 均正态.

~~联合~~ 正态情况下, 独立 ≡ 相关系数=0

$\mathrm{COV}(X_2+X_3, X_2-X_3)$

$= \mathrm{COV}(X_2,X_2) - \mathrm{COV}(X_3,X_2) + \mathrm{COV}(X_2,X_3) - \mathrm{COV}(X_3,X_3)$

$\sigma^2 - \sigma^2 = 0$ 故不相关 故独立

ES^2

$X \sim N(\mu, \sigma^2)$

正态总体. 样本: $X_1, X_2, \cdots, X_{2n}$. $\bar{X} = \frac{1}{2n}\sum_1^{2n} X_i$

$Y = \sum_1^n (X_i + X_{n+i} - 2\bar{X})^2$ 求 EY.

Soln.

$\cancel{EY = \sum_1^n (EX_i + EX_{n+i}}$

新样本 $(X_1 + X_{n+1}), (X_2 + X_{n+2}), \cdots (X_n + X_{2n})$

n个分量.

这个样本取自 $N(2\mu, 2\sigma^2)$

均值: $\frac{1}{n}\sum_{i=1}^{n}(X_i + X_{n+i})$

$= \frac{1}{n}\sum_{i=1}^{2n} X_i = 2\bar{X}$

新样本 方差: $\frac{1}{n-1}\sum_{i=1}^{n}(X_i + X_{n+i} - 2\bar{X})^2 = \frac{Y}{n-1}$

$E\left(\frac{Y}{n-1}\right) = 2\sigma^2$ → 新总体之方差.

$EY = 2(n-1)\sigma^2$.

参数估计. 关键是 L 如何写.

- 矩估计, 最大似然估计.

  估计方法:

    无偏估计, 一致性, 有效性.

    何为无偏?

    你得到一个 estimator, $\hat{\theta}$, 求期望

    $$E\hat{\theta} = \theta.$$

- 估计量之求法.

- 区间估计.

$$X_i \sim N(\mu, \sigma^2)$$

$$\hat{\sigma}^2 = c\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2$$

欲使 $E\hat{\sigma}^2 = \sigma^2$, $c = ?$

即 $\hat{\sigma}^2$ 是为 σ^2 之无偏估计.

$$E\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2 = E\sum_{i=1}^{n-1}(X_{i+1}^2 - 2X_iX_{i+1} + X_i^2)$$

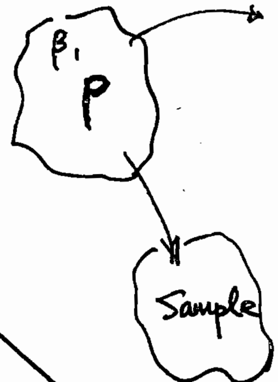
$$E\sum_{i=1}^{n-1}\left[(X_{i+1}-\mu) - (X_i-\mu)\right]^2 = E\sum_{i=1}^{n-1}\underbrace{(\quad)^2}_{\sigma^2} - \underbrace{2(\quad)(\quad)}_{0} + \underbrace{(\quad)^2}_{\sigma^2}$$

Stats

F-test.

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

We had some sort of population



within this population, we have some true value (β_1)

我们只有 population 的一部分 (Sample)
利用这有限的数据，我们要来做 test.
(Not 全集)

test if $\beta_1 = 0$.

- β_1是 population 中的 true mean.
- 这便是 so-called 假设检验.

H_0: null hypothesis.

$$\beta_1 = 0.$$

F和t假设检验之 ultimate power 来自于什么?

— 中心极限定理.

Sample distribution 趋于正态分布.

借助于中心极限定理的强大力量，构建 t-test

对于单变量 regression 而言，
该问题

若 test 多个变量, $\beta_0, \beta_1, \cdots$
Not feasible within current framework.

to Null hypothesis:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

Stats.

$\beta_1, \beta_2, \beta_3 = 0$

P

P代表population

即 $\beta_1 = \beta_2 \cdots = \beta_p = 0$

Ho:

Null hypothesis 是 $\beta_1, \beta_2, \cdots \beta_3$ 是 non-significant 的.

H1:

Alternative hypothesis 是 any of β_i 是 significant 的.

$$\beta_i \neq 0$$

任何一个 β_i 不为0, 会 lead to us rejecting the NULL hypothesis.

从 unrestricted regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \hat{u}$$

得到 $SSR = \sum_{i=1}^{N} \hat{u}_i^2 = (u_1^2 + u_2^2 + \cdots + u_N^2)$

estimation of population error.

We call residues.

"Statistically significant variables"

再 form restricted model.

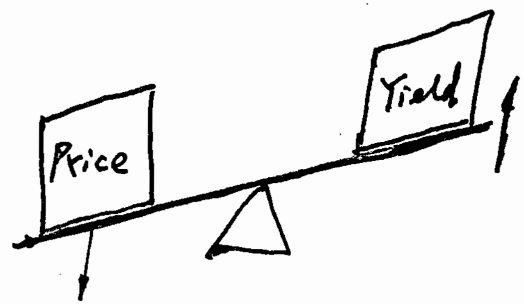
$$y = \alpha \qquad SSR_R > SSR_{UR}$$

Bond:

Price and yield are inverse,

Longer duration = more sensitive to rate moves

Bonds up
↳ 是 prices up (yield 其实 down 了)



Old bond

Price \$100

Coupon \$5

Yield 5%

所以 old bond
你还卖100块，
便没人买，于是
你只好降价.

If price falls to \$83.33, a coupon payment of \$5, will yield 6%
(5/83.33)

New bond

Price \$100

Coupon \$6

Yield 6%

obviously, every investor will choose the new bond, interest rate ~~goe~~ has gone up, so get more interest on their loans.

↳ interest rate goes higher, so lenders demanding higher yield.

F2N

Stats.

为什么只有 | F-test | ANOVA
| t-test |

? 因为人们往往只关心到二阶!

高阶的并不常用!

Volatility $\sim X^2$

Inference for Proportions.

Stats

What is the probability of developing a surgical site infection (SSI) after a certain type of surgery?

随机选400个 patients, who had this surgery, 12 develop an SSI.

$$\hat{p} = \frac{12}{400} = 0.03.$$

- $\hat{p}$: (sample proportion) is a point ~~proportion~~ estimation 点估计. of p (population proportion)
- $\hat{p}$ When we draw our sample, $\hat{p}$ will get a value.

  ↳ But this is not usually the case, in statistics. Just giving a point estimation, is not good enough.

  ↳ We want to use the data to say as much as we can about the parameter p.

- 兩种常見之做法:
  - Construct a confidence interval for p.
  - We may have a hypothesis value about p to test.

To construct 一个合适之 Inference procedures for p,

We need to know the sampling distribution of the sample $\hat{p}$.

样本 propotion

(由样本得到的比例)

$\hat{p}$: has a mean of p.

- 样本 propotion is an unbiased estimator of the population propotion.
  总体
- standard deviation of sampling distribution of $\hat{p}$

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

总体方差是 p(1-p)
但是 p 是总体的参数
→ 未知量!!

← We'll be able to walk around that.

- Sampling distribution of $\hat{p}$ is approximately normal if the sample size is large.

The assumptions of the one-sample inference procedures for a single propotion

一次抽样.

↘ We have a simple random sample from the population of interest.
总体.

↘ The sample size is large enough, for the normal approximation to be reasonable.

- It's reasonable 用正态近似, if:

$n\hat{p} \geq 15$ & $n(1-\hat{p}) \geq 15$.

即若 $\hat{p}$ 小, 则需要之 sample size 须更大.

A $(1-\alpha)100\%$ confidence interval for p is given by:

$$\hat{p} \pm \underbrace{Z_{\alpha/2} \times SE(\hat{p})}_{\text{margin of error}}$$

$\hat{p}$: best estimator of p.

$SE(\hat{p})$: standard error of $\hat{p}$.

↳ recall true standard deviation of sampling distribution of $\hat{p}$

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

↳ 我们不知 p. 只是 do the next best thing:

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

假设检验的原理.

Mathematics stat.

一个班学生之得分.

假设为 80分平均分, 5分之标准差.

开始检验

你上来抽一个50分. 距你的80分有6σ.

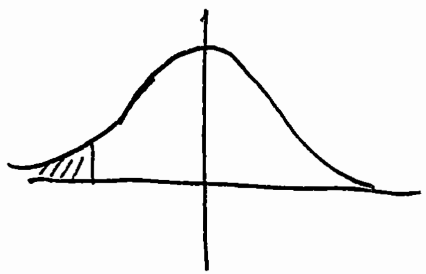
其概率是很低的.

于是你大概率上认为

你的假设有问题.

either 平均分,

抑或是标准差出了问题.



否定域, 值落于面积为5%的区域, 便否定之.

显著性水平 α.

通常取5%.

但有时也取10% (对结果之要求没那么严格, 判错了也没关系时)

stat.

linear regressi

- 从回归方程可知，计算过程中并不需要知道Y与X间是否有线性相关之关系。
  - 若不存在，则求得之回归方程毫无意义。
  - 故需对回归方程进行假设检验。

- 统计上讲 β_1 是 $E(Y)$ 随 X 线性变化之变化率。若 $\beta_1=0$，则 $E(Y)$ 并不随 X 线性变化。
  仅当 $\beta_1 \neq 0$ 时，$E(Y)$ 随 X 线性变化，此时回归方程才有意义。

$H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$

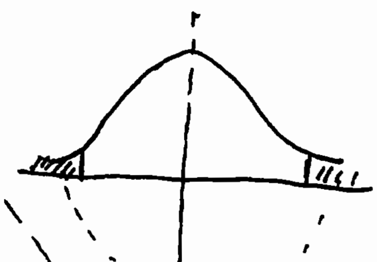
通常采用三种方法（构造统计量）

即 $\beta_1 = 0$ 时

1) t 检验法：H_0 成立时，统计量

$$T = \frac{\hat{\beta}_1}{sd(\hat{\beta}_1)} = \frac{\hat{\beta}_1\sqrt{S_{xx}}}{\hat{\sigma}} \sim t(n-2)$$

对于给定之显著性水平 α，检验拒绝域为：$|T| > t_{\alpha/2}(n-2)$



落于此二区域内，为拒绝域
拒绝 H_0，即 $\beta_1 \neq 0$。
即 $E(Y)$ 随 X 线性变化。

服从以 $(n-2)$ 为自由度之 t 分布。
即，有20个样本点，
则服从 $t(18)$

$\hat{\beta}_1$，$\sqrt{S_{xx}}$，$\hat{\sigma}$ 均可通过样本来算出，不包含未知参数。

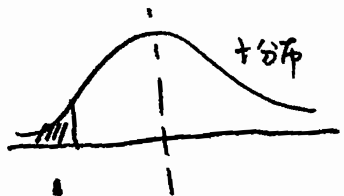
$t = \frac{\hat{\beta}_1\sqrt{S_{xx}}}{\hat{\sigma}}$ 即 R 中 lm 得之 t 值

linear regression.

t value　　　$Pr(>|t|)$

13.51　　　9.50 e-08

是建构造的统计量 $T=\frac{\hat{\beta}_1\sqrt{S_{xx}}}{\hat{\sigma}}$ 在当前样本下之取值.



t落于此区域，才是说是 X, Y 有线性关系这一假设有统计学意义.

$Pr(>|t|) = 9.50e-08$

之意义: 随机得到此t值之 prob 是如此之小.

于是有统计规律在里面.

→ "认为回归方程是显著的"

→ $\beta_1=0$. (即 X, Y 无线性关系)

2) F检验法，构造之统计量不同，当 H_0 成立时，

$$F=\frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} \sim F(1, n-2)$$

对于给定的显著性水平 α，检验拒绝域为 $F > F_\alpha(1, n-2)$

3). 相关系数检验法

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

，称为样本相关系数，对于给定之显著性水平 α. 查相关系数临界值表得 $r_\alpha(n-2)$

则检验拒绝域为: $|R| > r_\alpha(n-2)$

Regression.
1
Simple leanear regression.

- regression analysis explores the relationship between
  - a quantitative response variable
  and
  - one or more explanatory variables.

- 若只有一个 explanatory: simply linear regression.
  多个 —— : multiple ——.

- 有何用 (linear regression).
  - 比如一个区域有野生的鹿，也有狼.
    鹿 好 measure.
    狼 更危险些. → 于是若知二者有线性关系，则测鹿的数目即可得狼的数目.
  - 又比如某种树，想知其硬度.
    但硬度难测.
    → 于是可测 density.
    → density 与硬度成正比.

The first 4 observation (of 36)

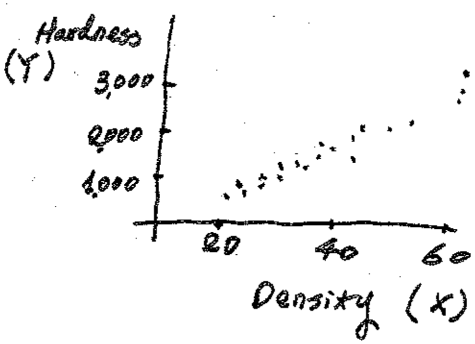
X	Density (lb/ft^3)	24.7	39.4	53.4	24.8
Y	Hardness (pound force)	484	1210	1880	427

- X: Explanatory var, independent var.
  Y: Response var, dependent var.

- The data set contains 36, (X, Y) pairs.

$$(X_1, Y_1): (24.7, 484)$$
$$\vdots \qquad \vdots$$
$$(X_4, Y_4): (24.8, 427)$$

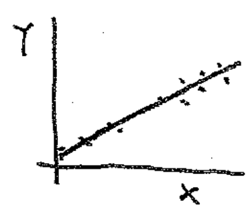
- First to do, plot our data to see what we're working with.



$\xrightarrow{so}$ Can we use a known value of density (X) to help predict the hardness (Y)?

Yes.

fit a line 來做預測



↘ 你可能会问：

- How to come up with a good fitting line?
- Is a line a good summary between those vars?
- Is the relationship strong enough, such we can use it for prediction?

↘ To measure the strength of the linear relationship, 用 correlation.

We assume a linear relationship between Y & X.

$$\underline{\mu_{Y|X}} = \beta_0 + \beta_1 X$$

The true mean of Y, given X.

亦可写成：

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

Expectation of Y given X, 等于 $\beta_0 + \beta_1 X$

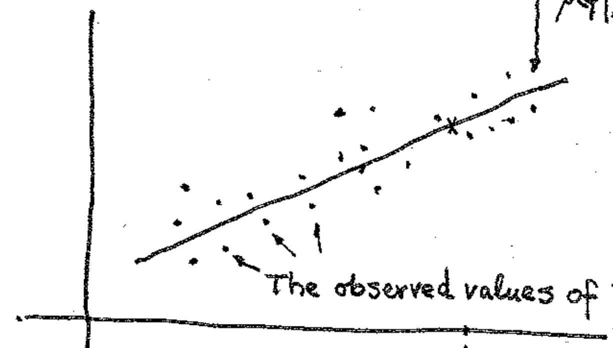
We assume, this is the true relationship.

β_0: y intercept

β_1: the slope

即为 true relationship between Y & X.

$\mu_{Y|X} = \beta_0 + \beta_1 X$



The observed values of Y vary about the line.

X

给一个X值，则 theoritical mean of Y, falls on that line.

但真实之 measurements 不会落在此理论预测值上. 因为有 variation.

We account for that variability around the line, using ε.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

random error component

represents the fact, Y will vary about the line.

- In the future, to do inference, have to ~~do~~ make some assumption on distributions of that random error components.

β_0, & β_1 are parameters, They are typically unknown values,

So we want to estimate them.

We will use sample data to obtain the estimated regression line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

这两个带hat的是 sample statistics,

↘ We usually use the method of least square to estimate $\underline{\beta_0}$ and $\underline{\beta_1}$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

↪ Assumed relationship between Y and X.

We use data to find the estimated regression line.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

predicted value of Y.

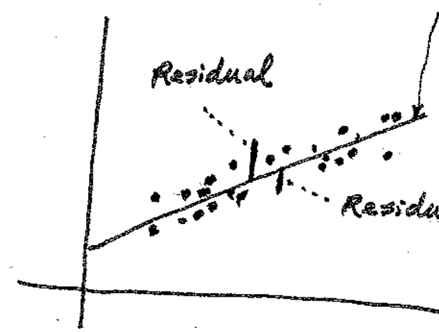
$\hat{\beta}_0$ is a statistic that estimates β_0

$\hat{\beta}_1$ is a ________________ β_1

↘ Residual = Observed − Predicted

$e_i = Y_i - \hat{Y}_i$ --- every observation has a residual associated w. it.

predicted value from the model.



Residual: $\underline{\text{signed}}$ distance
(从点到预测线的)
是的，是有正负号的

What's the best line?

↘ minimizes:

$$\Sigma |e_i| = \Sigma |Y_i - \hat{Y}_i|$$

↘ Sum of vertical distances between the points and the line.

↘ $\hat{\beta}_0$ & $\hat{\beta}_1$ are chosen to minimize

$$\Sigma \hat{e}_i^2 = \Sigma (Y_i - \hat{Y}_i)^2 = \Sigma (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

→ each observation (Y_i) 减去 prediction ($\hat{Y}_i$) 的差，再平方，相加。
－得到的和求最小值。

This is called the method of least squares.

↘ If the error terms are normally distributed, it's the best possible line to use.

↘ minimized summed least square residuals.

↘ $SS_{XX} = \Sigma (X_i - \bar{X})^2$ → Variance of X (samples), $S_X^2 = \frac{SS_{XX}}{n-1}$

$SS_{YY} = \Sigma (Y_i - \bar{Y})^2$ Variance of Y, $S_Y^2 = \frac{SS_{YY}}{n-1}$

$SP_{XY} = \Sigma (X_i - \bar{X})(Y_i - \bar{Y})$ $cov(X, Y) = \frac{SP_{XY}}{n-1}$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$ (mean of Y 减去 $\hat{\beta}_1$ 乘上 mean of X)

$$\hat{\beta}_1 = \frac{SP_{XY}}{SS_{XX}} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})(X_i - \bar{X})}$$

$$= \frac{COV(X, Y)}{Var(X)}$$

实例：R做 linear regression:

Coefficients:

	Estimate	Std Error	t	Pr(>|t|)
Intercepts	-1160.50	108.50	-10.69	2.07e-12
Density	57.507	2.279	25.24	< 2e-16

slope，labeled with explotory var X.

$\hat{\beta}_0 = -1160.50$

$\hat{\beta}_1 = 57.507$

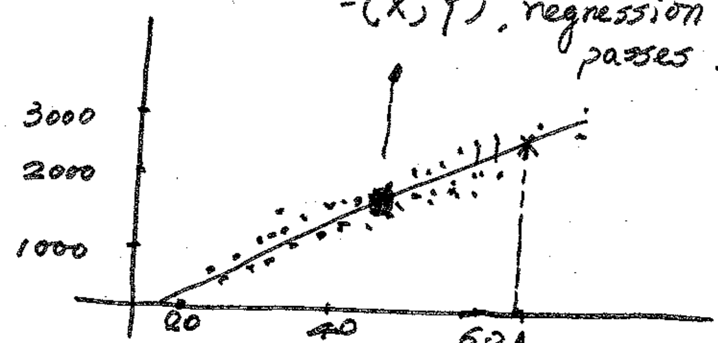
即 least square regression line: $\hat{Y} = -1160.50 + 57.507 X$

$\hat{\beta}_0$ $\hat{\beta}_1$

带符号距离，往下为负，往上为正.

- 所有之 sample 点到 regression line 之距离为0

- $(\bar{X}, \bar{Y})$, regression line always passes through that.



有了 prediction line，知道个 x，便可以预测其 predicted $\hat{Y}$

For least square regression:

→ The residual sum to 0 $\sum e_i = 0$

↳ mathematically, every time, least square sum to zero.

→ The line alway passes through $(\bar{X}, \bar{Y})$

The sample linear regression model:

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

→ The true mean of Y given X, 等于 $\beta_0 + \beta_1 X$

当 $X=0$ 时
$\mu_{Y|X} = \beta_0$
即 β_0 是 true mean of Y, when $X=0$.

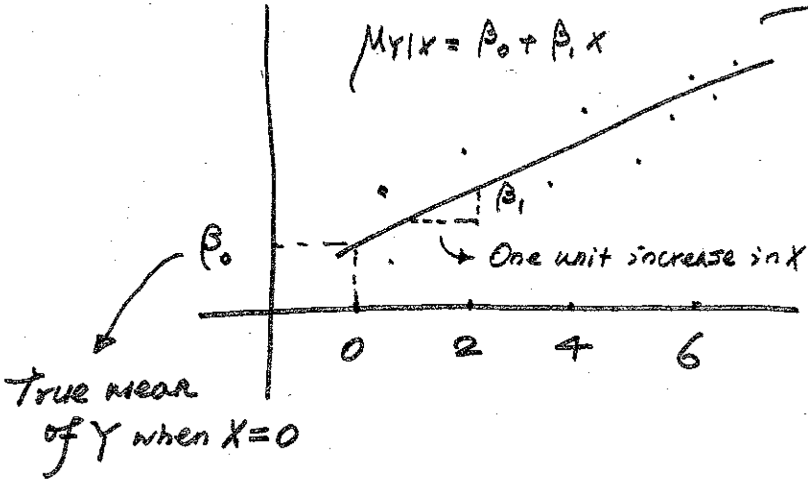
→ 但 practically 可能没太用,
因为 $X=0$ 可能比较 rare.

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

↳ Y value itself does not fall precisely on this line, 有一个 random variable components.

β_0, β_1 are parameters

β_1: 当 X 变化 1 时,
true mean of Y, 变化 β_1



通常我们不可能知道

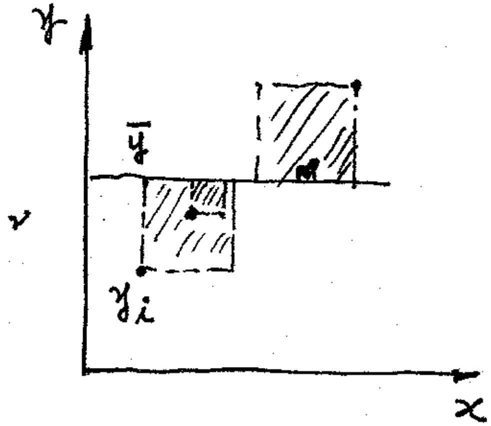
- these true parameters in any given situation.
- We're going to get sample data and estimate β_0 & β_1

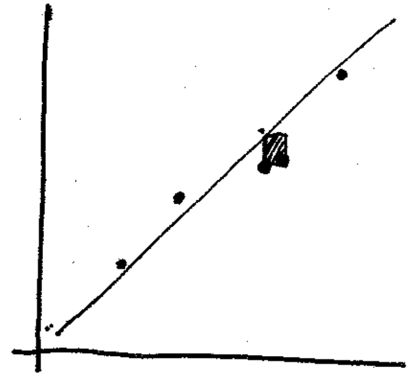
线性回归基本假设:

1. 随机误差项是一个期望值或均值为0的R.V.
2. 对于explotary var之所有之observation，随机误差有相同方差.
3. 随机误差彼此不相关.
4. Explotary var是确定性变量，不是R.V.，与随机误差之间彼此相互独立.
5. 随机误差服从正态分布.

0 同方独确正.

“灵童独雀呈”





total sum of squares.

正比于 variance of the data

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

最 no-brain 的，是用均值这条平行于 x 轴之线来预测。

$$SS_{reg} = \sum_i (\overset{*}{y}_i - \hat{y})^2$$

经过 linear regression 后，误差变小了。

本来之误差平方和

regression 后之误差平方和

$$R^2 \equiv \frac{SS_{tot} - SS_{reg}}{SS_{tot}}$$

$$= 1 - \frac{SS_{reg}}{SS_{tot}}.$$

→ 二者做差后，便是线性回归模型所做之贡献。