Week 6 Statistic

- E(X) of population 不可得, Statistical Inference:
① assume RV follows some known distribution with some unknown parameters
② use a sample from distribution
③ estimate unknown quantity of the population

Sample is random vector $(X_1, X_2, \cdots, X_n)$ n is sample size, each component is sample point

- Random sampling: each sample point has exact the same distribution (population of X) and independent with other sample points. ⇒ iid: independent & identically distributed

representitive

Random sample 是 iid

- statistic: real-value function T of a random sample $T(X_1, X_2, \cdots, X_n)$
function of RVs → RV

distribution of statistic T is called 'sampling distribution'

{central location: mean/median (central tendency, sample)
{dispersion: sample variance, range, interquartile range

sample: central location & level of dispersion. ⇒ descriptive statistics (summary statistics) of a sample

left-skewed (extreme value) mean < median right-skewed mean > median

{skew 时 medium 更好
{mean 对outlier敏感
tractability

- 25^{th} percentile = 0.25 quantile = 1^{st} quartile (Q_1), 2^{nd} Q_2 50%, 3^{rd} Q_3 75%
IQR Inter-quartile range: $Q_3 - Q_1$. cannot negative Q3 一定比 Q1 大 upper limit $Q_3 + 1.5 \times IQR$

Week 7 Central Limit Theorem

Central limit theorem (CLT)

Suppose the random sample $X_1, X_2, \ldots, X_n$ follows an iid distribution $f(x)$ with mean μ and variance σ^2. (population)

CLT states that, for any $f(x)$, the distribution of $\overline{X_n}$ will become close to a normal distribution with mean μ and variance σ^2/n when the sample size n is very large. $n \geq 30$

n↑ approx↑ similar! 近似 population variance

Suppose X is a RV with mean μ_0 and variance σ^2. Thus, $\overline{X_n} \sim N(\mu, \sigma^2/n)$, by CLT for a large enough n.
By the W4 Tut, we can transform $\overline{X_n}$ to be a standard normal Z, via:

$$Z = \frac{\overline{X_n} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

dist $\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$ follows standard normal distribution N(0,1) when n large

CLT
{more normal X is, the smaller required sample size
{If X already normal, sample mean should autoly normal

Week 8 Confidence Interval

assume population follows certain distribution (statistical model)

Interval estimator for $\mu = E(X)$ is $[\overline{X_n} - a, \overline{X_n} + a]$ ∴ sample 变, interval 也变

find the shortest interval that meets a given performance criterion – confidence level

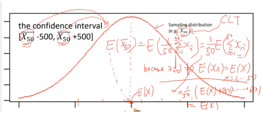
When the interval estimator $[\overline{X_n} - a, \overline{X_n} + a]$ contains the unknown parameter (μ) at least p% of the time (many random samples), we call this interval estimator

A p% Confidence Interval for μ

If I claim $[\overline{X_{50}}$ -500, $\overline{X_{50}}$ +500] is a 95% CI, then it means the true average μ will be contained in 95% of *the (many many) CI's based on this CI formula*

算 CI：重要的是找到 $[\overline{X_{50}} - a, \overline{X_{50}} + a]$ 的 "a"

CI 包含了在α significance 内所有不被 reject 的



How E(X) relate to $\overline{X_{50}}$

⇒Don't interpret them as "There is a 90% chance that the population mean is in [4767.383,5232.617] interval." The population mean is either in it or not in it (not probability). We just don't know it. We only know if you get many CI's, 90% of them will catch the (invisible) big fish!

- CI for σ known eg: "a" for 90% CI
find "a" to make $P\{\mu - a \leq \overline{X_{50}} \leq \mu + a\} = 0.9$
$F(\overline{X_{50}} = \mu + a) = 0.95 \Rightarrow F(Z = \frac{\mu + a - \mu}{\frac{\sigma}{\sqrt{n}}}) = 0.95$
$qnorm(0.95, 0, 1) = \frac{a}{\frac{\sigma}{\sqrt{n}}} \Rightarrow a = qnorm(0.95, 0, 1)\frac{\sigma}{\sqrt{n}}$

p% CI is $[\overline{X_n} - Z_{p'} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X_n} + Z_{p'} \cdot \frac{\sigma}{\sqrt{n}}]$ $p' = p + (100-p)/2$ %

same sample size, 99% CI contain 90% CI. To get shorter CI, larger sample size

- When σ is unknown, estimate σ with sample counterpart, the sample standard deviation S
replace Z score to t distribution
{t-distribution, degree of freedom (df) = n-1 n: sample score
{t score: cumulative probability of "q" and n-1 df as $t_q(n-1)$ qt(P, n-1)

* When $n \geq 30$ (large), close to N(0,1) → sample standard deviation of size n sample

p% CI is $[\overline{X_n} - t_q(n-1) * \frac{S_n}{\sqrt{n}}, \overline{X_n} + t_q(n-1) * \frac{S_n}{\sqrt{n}}]$ $q = p + (100-p)/2$ (%)

* 用Z.t score 基于 X~N(,)
- X is not normal but sample size ≥30 ⇒(CLT) Estimate E(X) based on t and S
- X is not normal but sample size small ⇒ {collect more sample n≥30 → CLT → CI
{computer simulation

Week 9 Hypothesis Testing

2. Hypothesis testing (HT)
Use the random sample to test a statement (H₀) about the value of "?"

Random sample → statistic model X~N(?,?)

Let me test whether ***"the average # of steps that a NTU student takes is equal to 5000" is plausible or not***

⇒Mathematically, we want to verify ***"μ = 5000," using a sample as evidence***
⇒ HT describes a method to converting a random sample to a yes-no conclusion about the statement

null hypothesis H_0: status quo / current knowledge of parameter
alternative hypothesis H_1: encompass all other possible value not in H_0
HT is a rule to decide support H_0 or not based on sample

reject or not? set a buffer zone where consider normal before rejecting H_0.

- Key concept #1:

We use the sample, either to reject H0, or NOT reject H0

- Key concept #2:

The sampling process induces uncertainty -> Rationally, we create a "buffer" in making the rejection decision

Type 1 error: reject H_0 when H_0 true
Type 2 error: not reject H_0 when H_1 true

- rejection region: wilder RR has higher Type 1 error, lower Type 2 error

基本不能同时减. ① set a maximum of Type 1. (1%, 5%, 10%) ← α value
② Among qualified RR, find one minimize Type 2.
⇒ RR unique when given α significance level

Steps for HT:

① Formulate H_0 and H_1

3. **RR:** Choose an acceptable type-I probability (α) ⇒ Rejection region (RR) in N(0,1) for Z qnorm($\frac{\alpha}{2}$, 0, 1) qnorm($1-\frac{\alpha}{2}$, 0, 1)
4. **Test statistics:** calculate $\overline{X_n}$, and then convert it to Z
5. **Decision:** If the Z score is in RR ⇒ Reject H_0!

p-value < 1%: strong
p-value < 5%: moderate
p-value < 10%: marginal evidence.

- the minimum α needed to reject H_0: p-value ← statistic

p value ×2 小于 α, reject H_0. (two side) p-value = 2× pnorm(Z, 0, 1)

p value = prob of observing value more extreme than observed test statistic if H_0 is true

One-tailed test: $H_0: \mu \geq \mu_0$ p-value = {pnorm(Z, 0, 1)
$H_0: \mu \leq \mu_0$ {1 - pnorm(Z, 0, 1)

p-value {σ^2 known: pnorm($\overline{X_n}$, μ_0, $\sigma/\sqrt{n}$), 或 pnorm(Z, 0, 1), where $Z = (\overline{X_n} - \mu_0)/(\sigma/\sqrt{n})$
{σ^2 unknown: pt(t, n-1), $t = (\overline{X_n} - \mu_0)/(S/\sqrt{n})$

Week 10 Linear Regression

model the conditional expectation 2 assumptions.

1. Linearity: assume conditional mean of Y is linear function of all X's
$E(Y | X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
x_1, x_2 are conditions, their value will be given a priori (先验), not random
2. Normality: assume Y follows a normal distribution, given the X's
$(Y|X_1, X_2) \sim N(E(Y|X_1, X_2), \sigma^2) \Rightarrow (Y|X_1, X_2) - E(Y|X_1, X_2) \sim N(0, \sigma^2)$. from (1), $E(Y|X_1, X_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
Thus, $(Y|X_1, X_2) - (\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ is $N(0, \sigma^2)$ random all
⇒ $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ the population model of regression
conditional expectation

- LR base on estimating the conditional expectation of "Y" β_i show effect size of variables X_i

ε: error term, RV, represent all the other independent variables related to Y

- Ordinary least Square (OLS)

find parameter $\beta_0, \beta_1 \cdots$ minimize sum of squared errors $\Sigma(Y_i - \hat{Y_i})^2$

Bibliographical notes

- The categorical variable estimates the differences in group means in DV, "ESBE" everything else being equal

You may see hypothesis tests related to differences in group means:

A two-level categorical variable: *the two-sample t-tests*

A categorical variable with more than two levels: *one-way ANOVA tests*

These tests in their primitive form does not "ESBE" and thus the result can be biased (refer to the W5 video and tutorial on *Correlation)*

⇒When you want to compare differences in group averages, use regression whenever possible to tease out the real, ELBE, differences

avg → true value.
coefficient are statistics, have sampling distribution
coefficient interpret: 基于 everything else equal.

H_0: $\beta_i = 0$ 看变量显著性
加 dummy variable: n类, n-1个 dummy V. 都是0的是 base variable

CI、HT都可

Week 11 采样和simulation都为statistical inferences. 区别只有sample collected/generated.

log 变换 研究 % change 而不是 unit change. → 不管变量单位

- $log(Y) = \beta_0 + \beta_1 X$ (2)

⇒The interpretation of β_1: "***per unit increase** in the IV is associated with a **(100* β_1)% change in Y***"

E.g., if $\beta_1 = 0.07, X$ *being education,* **then it means 1 more year of education increases the average salary by 7%.** (100*.07)%

- $log(Y) = \beta_0 + \beta_1 log(X)$ (3)

⇒The interpretation of β_1: "***per 1% increase*** in X is associated with a ***β_1% change in Y***"

E.g., if $\beta_1 = 1.7, X$ *being education,* **then it means 1% more education increases the average salary by 1.7%.**

coefficient interpretation
直接 lm(log(Y) ~ ··)

Week 6. Bootstrap

- Random sample (iid) collected by simple random sampling (everyone in population has equal probability to be selected into sample)
- stratified sampling: divide the entire population into several strata, perform simple random sampling in each strata and combine the data to get full sample. {within-strata homogeneity {between-strata heterogeneity strata 的 var 小 → true value

决定各类样本量: proportional allocation 根据各类在总体中的比例决定样本取多少

Bootstrap 原因: 1. classical methods 基于 assumption of {population distribution {large sample size (CLT) to establish sample distribution $\overline{X_n} \sim N(\mu, \sigma^2/n)$
2. 有些 statistics' distribution not analytically tractable eg: sd (correlation coefficient/sample median)
要很多 sample 才能 construct sampling distribution

idea: replace hypothetical population F with its estimate $\hat{F}$ constructed based on data of random sample

distribution used to generate bootstrap samples: $\hat{f}(x_i) = \frac{\text{\# of points in the sample} = x_i}{\text{Sample size}}$ is called empirical distribution (based on sample)

Sample comes from the population → provide approximation of population

- Estimate bias: estimator is a statistic to estimate a parameter of population
Bias of estimator $\hat{\theta}$ for parameter θ defined as:

Bias $= E(\hat{\theta}) - \theta$
($E(\hat{\theta})$: Expected value of the estimator (based on the true sampling distribution); θ: The true parameter value based on the population)

Bootstrap estimate of the bias $= E_B(\hat{\theta}^*) - \hat{\theta}$
($E_B(\hat{\theta}^*)$: Average of the B bootstrap samples (based on the simulated sampling distribution); $\hat{\theta}$: Sample mean based on the data)

- best for estimating 'shape-related' statistics
eg: sd, skewness

Week 7

- random sample $X_1, \cdots, X_n$, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$. Let $\overline{X_n} = \frac{\Sigma_{i=1}^n X_i}{n}$
Law of large numbers: $\lim Pr(|\overline{X_n} - \mu| < \varepsilon) = 1$ $\overline{X_n}$ will converge in probability to population mean μ
- Central limit theorem
$F_n(X)$ be CDF of $\sqrt{n}(\overline{X_n} - \mu)/\sigma$. By CLT $\sqrt{n}(\overline{X_n} - \mu)/\sigma$ will converge in distribution to N(0,1)

1. X_n converge in probability to X iff $\lim_{n\to\infty} Pr(|X_n - X| > \varepsilon) = 0$ $\forall \varepsilon > 0$ require direct probabilistic comparison between sequence and target RV
denote $plim(X_n) = X$, $X_n \xrightarrow{P} X$, X can be constant
verify $Pr(|X_n - X| > \varepsilon)$ by counting $\frac{\text{\#times that } |X_n - X| > \varepsilon}{\text{\# of trials}}$ n↑, ratio → 0. for any ε>0.
2. X_n converge in distribution to X iff $\lim_{n\to\infty} F_{X_n}(X) = F_X(x)$ $\forall x$ where $F_X(x)$ continuous
* convergence only requires the CDFs of sequence of RV's to be equal everywhere to another RV → weaker than converge in P.

Unbiasedness and consistency are two highly desirable properties of a good estimator:
- $\widehat{\theta_n}$ is an unbiased estimator for θ if Bias= $E(\widehat{\theta_n}) - \theta = 0$.
- $\widehat{\theta_n}$ is a consistent estimator for θ if $\hat{\theta}$ converges in p. to θ.

"Unbiasedness" is regarding one specific estimator, whereas "consistency" is regarding a sequence of estimators, whether the sampling distribution of $\widehat{\theta_n}$ will diminish under the theoretical infinite sample size.
Consistency is usually regarded as more important.

$X_1, \cdots, X_n$ iid U(0,1). $Y_n = max(X_1, \cdots, X_n)$ $plim(Y_n) = 1$
for any $0 < \varepsilon < 1$
$P(|Y_n - 1| > \varepsilon) = P(Y_n < 1 - \varepsilon)$
$= (1-\varepsilon)^n \to 0$
$Y_n = max(X_1, \cdots, X_n)$ {consistent of 1 {× unbias of 1
$K_n = 0.5 + X_n$ ⇒ K_n unbias of "1" not consistent

Week 8

multiple difference : if use traditional two-sample 95% CI 3 times

family-wise error rate (FWER) = 1 - P(all test correct) more test → higher FWER → lower confidence

1. Bonferroni correlation

$FWER \le \sum_{i=1}^{m} Pr(CI_i \text{ is incorrect})$ ← keep at a desirable level α.

eg: m=3. desired CI is 95%.

$FWER \le P(CI_1 \text{ wrong}) + P(CI_2 \text{ wrong}) + P(CI_3 \text{ wrong}) = 5\%$

use 1.67% as α to find CI for m = 1,2,3

t.test(---, ---, conf.level = ..).

$\Rightarrow$ let $P(CI_1 \text{ wrong}) = P(CI_2 \text{ wrong}) = P(CI_3 \text{ wrong}) = \frac{5\%}{3}$

2. Tukey's method : Assumption for each sample : normality, equal variances, independence

total sample size N. statistic $\max \left| \frac{(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{MSE}{N}}} \right|$ follows studentized range distribution

$\bar{x}_i - \bar{x}_j \pm q(\alpha)\sqrt{\frac{MSE}{N}}$ has coverage probability of α, where $q(\alpha)$ is quantile function of above distribution.

W9 Maximum likelihood estimate

Likelihood function of θ (unknown parameter) $\mathcal{L}(\theta) = Pr(X = x|\theta)$

MLE : value of p that maximize the value of $\mathcal{L}(p)$ $ best val 要取值

- The likelihood function is

$L(p) = \binom{10}{8} p^8(1-p)^2$=dbinom(8,10,p) The probability of observing X=8

The MLE is then the solution to $\text{Max}_{p\in\theta} L(p)$

```
# define the likelihood function
> L=function(p){-choose(10,8)*p^8*(1-p)^2}
# or, equivalently,
> L=function(p){-dbinom(8,10,p)}
# find the MLE using the Deoptim solver
>DEoptim::DEoptim(L, lower = 0,upper = 1,control = DEoptim.control(trace = FALSE))$optim
```

If the "-" sign is necessary because R's solver by d solves a minimization problem. This "-" is used to c the problem back to a maximization problem. E.g., Max X for 5<=X<=10 vs Min -X for 5<=X<=10. # choose(10,8) is a constant and thus can be ignor will not change the optimal solution.

Specify the parameter space θ

Suppress showing the intermediate computational outputs from the solver

(1) $L(\lambda) = \prod_{i=1}^{162} dpois(x_i, \lambda)$

As the sample size increases, the likelihood function becomes a high-dimensional problem of the parameter(s), due to the multiplication

(2) $\text{Log } L(\lambda) = \sum_{i=1}^{162} Log(dpois(x_i, \lambda))$

use log transformation to make the high-d problem -> lower-d problem. E.g., Log (X^10) = 10 Log X

(3) $\text{MLE} = \text{argmax}_{\lambda\in\theta}(\text{Log}(L(\lambda)))$

as log function is a monotonic increasing function, its optimal solution is the MLE

Bernoulli likelihood function 把乘求成 log 转化变为加和

$P(X=x|p) = p^x \cdot (1-p)^{1-x}$ $x = 0, 1$ 3次硬币

$\mathcal{L}(p) = (p^{x[1]} \cdot (1-p)^{1-x[1]}) \cdot (p^{x[2]} \cdot (1-p)^{1-x[2]}) \cdot (p^{x[3]} \cdot (1-p)^{1-x[3]})$

- continuous model

linear regression model $Y_i \sim N(b_0 + b_1 X_i, \sigma^2)$, where b_0, b_1, σ^2 are three unknown parameters of the model

$\mathcal{L}(\sigma, b_0, b_1) = \prod dnorm(Y_i, b_0 + b_1 X_i, \sigma^2)$ $\rightarrow \log(\mathcal{L}(\sigma, b_0, b_1)) = \sum \log(dnorm(Y_i, b_0 + b_1 X_i, \sigma^2))$

W10. OLS & Model selection

$y_i = \beta_0 + \sum \beta_p x_{pi} + \varepsilon_i$

minimize the sum of squared residuals

OLS find the β_p^* s.t. $\beta_0^* + \sum \beta_p \cdot x_p$ gives the 'best' predictions about y_i in the sample

$\sum_{i=1}^{n} (y_i - (\beta_0^* + \sum \beta_p^* \cdot x_{pi}))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (\text{residual})^2$

不知 true model : include more independent variables , include polynomial or more complicated function

future data don't exist during analysis ⇒ cross validation to measure model predictive performance

expected test MSE : $E(MSE) = E(\hat{Y}_0 - Y_0)^2$. use avg MSE values from cross-validation as an estimate for E(MSE)

$E(MSE) = Var\left(\hat{f}(X_0)\right) + E\left(\hat{f}(X_0) - y_0\right)^2 + Var(\epsilon)$

Variances: Variability of the predictions over many training datasets

Squared bias (SB): Squared difference between the avg predictions (over many training datasets) and the actual values y_0.

Irreducible errors: unpredictable, random errors bc we use $\hat{f}$ to approximate the unknown f.

	Bias	Variance
简单模型 :	↑	↓
复杂	↓	↑

λ↑ simpler model V↓ B↑

λ↓ flexible model V↑ B↓ (变量多)

Ridge will shrink the coefficients of correlated variables to a small values, whereas LASSO may pick one and discard others

In terms of model interpretability, LASSO has the upper hand ←可解释性

Panalty form

$$\text{Min } [\textstyle\sum_{i=1}^{n}\left(y_i - (\beta_0 + \sum\beta_p x_{pi})\right)^2] + \lambda\sum(\beta_p)^2 \text{ (Ridge)}$$

or

$$\text{Min } [\textstyle\sum_{i=1}^{n}\left(y_i - (\beta_0 + \sum\beta_p x_{pi})\right)^2] + \lambda\sum|\beta_p| \text{ (LASSO)}$$

"least absolute shrinkage and selection operator"

The $\lambda \ge 0$ is a user-determined ***tuning parameter***

Useful when have a lot of IV's and not as many sample points in the training data, want to keep fewer IV's in the model to improve interpretability

Code

W6 · df 筛选 · Bootstrap

```
mean(Seatbelts[Seatbelts[,'law']==0,'drivers'])
dim(ChickWeight[ChickWeight$weight<100,])
median( ChickWeight[ChickWeight$Diet==2,'weight'])
```

df 只选一列: W6$roa[W6$ind=="mfg"]

ifelse (条件, 1, 0)

sample( ): 默认 simple sample space, without replacement → replace = T 代表可放回

对于非 SSS 的 sample, sample (c(" rain"," no rain"), 7, prob = c(.4, .6), replace= T)

```
# Copy the flight delay data
W6=read.delim("clipboard")

# We draw 10k bootstrap samples from the bootstrap
population
B=10000
x_bar=numeric(B)
n=length(W6$Delay[delay$Carrier =="UA"])

for(i in 1:B) {
  #calculate the sample mean and store the result in x_bar
  x_bar[i]=mean(sample(W6$Delay,n,replace=TRUE))
}
hist(x_bar)

quantile(x_bar, c(0.025, 0.975))
```

```
e=0.01
n=500
prob=c(n)
set.seed(123)
#  sample size from 1 to n
for(i in 1:n){
  err_count=0

#  count the error probability under n
  for(j in 1:10000){
    if(abs(max(runif(i))-1)>e){
      err_count=err_count+1
    }
  }
  prob[i]=err_count/10000
}
plot(prob,ylab="Pr(|Yn-1|>0.01)",xlab="n")
```

W7 $Y = \max(X_1, \cdots, X_n)$. plim $(Y_n) = 1$ 的 simulation

W8 Bonferroni correlation

```
ope_cost = read.delim('clipboard', header = T)
A = t.test(ope_cost$Your.company,ope_cost$Firm.A,conf.level = (1-0.1/4))
B = t.test(ope_cost$Your.company,ope_cost$Firm.B,conf.level = (1-0.1/4))
C = t.test(ope_cost$Your.company,ope_cost$Firm.C,conf.level = (1-0.1/4))
D = t.test(ope_cost$Your.company,ope_cost$Firm.D,conf.level = (1-0.1/4))
A$conf.int
B$conf.int
C$conf.int
D$conf.int
```

做CI时 qt, qnorm 代 1-α/2. 若是sample的可直接 t.test( )

Turkey's Method

```
df2=df[df$CountryOfCtzName %in% c('Austria','Belgium','Netherlands'),]
```

df 类很多时先筛选 ↑

```
library(tidyr)
#w8.recipe=read.delim("clipboard")
w8.recipe=read.csv('/Users/apple/Desktop/NTU_Tri1/AN6005/Class\ Materi
data_long =  gather(w8.recipe, product, result, P1:P7)
# Tukey's command works as a post-hoc test of ANOVA in R
Model=aov(data_long$result~data_long$product)
TukeyHSD(Model)
T_result=TukeyHSD(Model)
plot(T_result, col="red")
```

要转 long form

, conf.level= 0.9/0.95

W9. t.test 可直接算 p-value : t.test(--, mu= ?) C_n^a = choose(n, a) 一定要注意 L function 的负号

Binomial

```
set.seed(520)
X=rbinom(3,10,0.8)
# 9 8 10 means we find 1 in 10 times for 9,8 and 10 times
L = function(p){-(dbinom(X[1],10,p) * dbinom(X[2],10,p) * dbinom(X[3],10,p)) }
DEoptim(L, lower = 0,upper = 1,control = DEoptim.control(trace = FALSE))$optim
```

泊松

```
LL=function(lambda){-sum(log(dpois(x$vihicles,lambda)))}
```

>x=sample(0:1,3,replace=T) Three random samples
L=function(p){- (p^x[1]*(1-p)^(1-x[1]))* (p^x[2]*(1-p)^(1-x[2]))* (p^x[3]*(1-p)^(1-x[3])) }

Bernoulli

线性

```
Y <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
X <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
LL= function(parameters){
  -sum(log(dnorm(Y,parameters[1]+parameters[2]*X,parameters[3])))
}
```

$Y_i \sim N(b_0 + b_1 X_i, \sigma^2)$

para 1 2 3

Logistic

$$L(b_0, b_1) = \prod_{i=1}^{n}\left(F(b_0 + b_1 Age_i)\right)^{y_i}\left(1 - F(b_0 + b_1 Age_i)\right)^{1-y_i}.$$

```
L1=function(parameters){
  -(sum(log(plogis(parameters[1]+parameters[2]*data$Age[data$ICU==1])))
      +sum(log(1-plogis(parameters[1]+parameters[2]*data$Age[data$ICU==0]))))
}
```

W10.

```
reg_age3=lm(OfficialTime_min~AgeOnRaceDay + Age2 + Age3,data=sample1)
summary(reg_age3)
MSE3=mean(reg_age3$residuals^2)
library(glmnet)
sample1=read.csv("d:/week10.Runners100.csv")
y=sample1[,"OfficialTime_min"  ]
x=sample1[,c("AgeOnRaceDay","Gender","Age2","Age3")]
# cv.glmnet can fit ridge and LASSO regression, by sele
lasso=cv.glmnet(as.matrix(x),y,alpha=1)
ridge=cv.glmnet(as.matrix(x),y,alpha=0)
plot(lasso)
plot(ridge)
# Extract the coefficients
coef(lasso,s="lambda.min")
coef(ridge,s="lambda.min")
```

```
# Extract coefficients from lasso$glmnet.fit$beta, based on the lambda value that gives the lowest
lasso$glmnet.fit$beta[,lasso$glmnet.fit$lambda==lasso$lambda.min]
ridge$glmnet.fit$beta[,ridge$glmnet.fit$lambda==ridge$lambda.min]
y_predicted <-
predict(lasso,s=lasso$lambda.min,newx=as.matrix(x))
rmse_lasso <- mean((y_predicted - y) ^ 2)
```

题: 当 upper/lower bound 大于max, min 则用 max, min

注意 sd(A+B), 不可分开

normal distribution 只是用 z, t 条件, 一定要看题有没有说 population sd!!

sd大, 要用更大 sample 来达到相同 abs error

3. A professor wants to test whether on average, his students are willing to pay less than $100 for the statistics textbook. The professor collects a sample of 35 students.

$H_0: \mu \ge 100$

$H_1: \mu < 100$

(1) Describe the Type 1 and Type 2 errors in this problem.

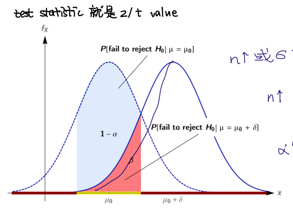
(2) Suppose $\sigma^2 = 25$, α=1%, and the unknown μ = $97. What are Pr(Type 1 errors) and Pr(Type 2 errors) = ?

RR= qnorm (0.01, 100, $\frac{25}{\sqrt{35}}$)

P(X>RR| M=97) = pnorm (98.033, 97, $\frac{25}{\sqrt{35}}$) = 0.889

P = 1 - 0.889 = 0.111

test statistic 就是 z/t value



n↑ 或 σ↑

n↑ $\frac{\sigma}{\sqrt{n}}$↓ 图更瘦

β ↓

α↑ 拒决域变大

变量解释

Year as a categorical variable: "Y2 students' avg spending is $44.31 higher than Y1"

Year as a continuous variable: "1 more year student's avg spending will increase $25.34 (ELBE)."

$H_0: \mu = 0$ p-value = p_i ⇒ $H_0: \mu \ge 0$ p-value = $p_i/2$

abs error: CI 的 α 的长度

cor (ChickWeight[, c('weight', 'Time')])

有非 numerical data 不能 cor(整个data) (factors类)

↳ class (...$..) 看类型

根据 diet 分类

boxplot (chick$ weight ~ chick$ Diet) (, data= xx)

quantile (xx , p=0.25/0.5/0.75)

IQR(xx).

Effectively, we look to obtain a vector of $\overline{X_3}$ that is 1000 in length. Thus,

```
> x_bar_3 = rep(0,1000)
for (idx in 1:1000) {
x_bar_3 [idx] = mean( sample(c(10,30,100,150), 3, prob = c( .3, .2,
.4,.1), replace = T) )
}
hist(x_bar_3)
#you can modify the codes to generate
Distributions for other sample sizes
```

$Pr(X = 10) = 0.3$
$Pr(X = 30) = 0.2$
$Pr(X = 100) = 0.4$
$Pr(X = 150) = 0.1$

pnorm( ), pt( ) 填值得CDF !

qnorm( ), qt( ) 填Pr 得值

```
# load data
> x<-read.csv(file.choose(),header=T)
# get sample size
>n<-length(x$weight)
25
# calculate sample mean
> X_bar<-mean(x$weight)
# calculate sample standard deviation
> S<-sd(x$weight)
12.67307
# calculate lower and upper limits of CI
>LL<-X_bar-S/sqrt(n)*qt(.975,n-1)
> UL<-X_bar+S/sqrt(n)*qt(.975, n-1)
c(LL,UL)
95.00881 105.47119
```

The p% CI is $[\overline{X_n} - t_q(n-1) * \frac{S_n}{\sqrt{n}}, \overline{X_n} + t_q(n-1) * \frac{S_n}{\sqrt{n}}]$

Find the p-value: when σ^2 is known

When σ^2 is known, p-value can be calculated by pnorm($\overline{X_n}$, μ_0, $\sigma/\sqrt{n}$) command; or via the standard norm. dist.: pnorm(z, 0, 1), where z= $(\overline{X_n} - \mu_0)/(\sigma/\sqrt{n})$

Consider the ChickenWeight data in R. Suppose the population standard deviation of chicken weight is 70 grams. H_0 is $\mu = 115$. Based on the sample, what is the conclusion of the test at the 5% level?

```
> library(datasets)
>length(chickWeight$weight)
578
>z=(mean(ChickWeight$weight)-115)/(70/sqrt(578))        #z=2.34>0; so right tail
>p_value = 2*(1-pnorm(z,0,1))
0.01919247
```

Conclusion: **We find (moderate) statistical evidence to reject H_0 (1%<p-value < 5%)**

Find the p-value: when σ^2 is unknown

When σ^2 is **unknown**, p-value can be found in R by pt(t,n-1), where t= $(\overline{X_n} - \mu_0)/(s/\sqrt{n})$

Consider the ChickenWeight data in R. Suppose the population standard deviation of chicken weight is **unknown**. H_0 is $\mu = 125$. Based on the sample, what is the conclusion of the test at the 5% level?

```
> library(datasets)
>length(chickWeight$weight)
578
>t=(mean(ChickWeight$weight)-125)/(sd(ChickWeight$weight )/sqrt(578))
> p_value=2*pt(t,577)        #t=-1.076266 is negative so is left-tail
0.2822582
```

Conclusion: **We do not reject H_0 (p-value >10%)**