STATS 105 10/01

Review of Probability

Probability is common sense of uncertainty

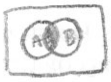
Sample Space: the set of all possible outcomes (Ω, S)

Event: statement about the outcome. Subset of the Sample Space ($A, B, C, \ldots$)

ϕ : NULL/Empty Event. Probability is defined on the event

Relations:

AND : $A \cap B$



OR : $A \cup B$

NOT : $\bar{A}$

Equally likely Setting

- Randomly sample a single person in a population. All people are equally likely to be sampled

$P(A) = \frac{|A|}{|\Omega|}$, A = sub-population $|A|$: # of elements in A
Ω = entire population

Axioms:

(1) $P(\Omega) = 1$

(2) $P(A) \geqslant 0$

(3) Additivity. If $A \cap B = \phi$ (disjoint) then $P(A \cup B) = P(A) + P(B)$

Probability is a measure

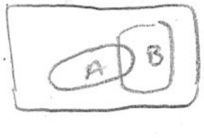
σ algebra - collection of all meaningful statements
start from simple statements

Probability can be interpreted as long run frequency.
If we repeat the experiment a large # of times,
$P(A)$ can be interpreted as how often A happens.

Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Geometry



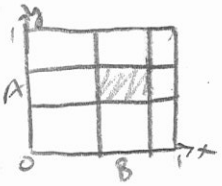
$P(A|B)$
↳ falls into B (new Ω)

Long-run frequency:
How often A happens when B happens

Independence

A and B are independent. $P(A|B) = P(A)$. $P(B|A) = P(B)$

disjoint ≠ independent



$|A \cap B| = |A| \cdot |B|$ $P(A \cap B) = P(A)P(B)$

Random Variables:

Population $\Omega = \{w_1 \quad w_2 \cdots w_M\}$

Randomly sample a person

$\mathbb{X}$= height of this person. ⇒ height = $\mathbb{X}(w)$

Randomly sample $w \in \Omega$

Experiment $\longrightarrow$ outcome $\xrightarrow{\text{label}}$ number

$A: \mathbb{X} > 6ft \Rightarrow A = \{w : \mathbb{X}(w) > 6\} \subset \Omega \Rightarrow P(A) = P(\mathbb{X} > 6) = P(\{w : \mathbb{X}(w) > 6\})$

Event: Statement about outcome

w/ numbers: inequalities, equation

Basic Events:

Discrete: $\mathbb{X} = x$ $\{w : \mathbb{X}(w) = x\}$

Continuous: $\mathbb{X} \in (x, x+\Delta x)$ $\{w : \mathbb{X}(w) \in (x, x+\Delta x)\}$

STATS 105 10/03

In general, a special event: $X \le x$
$\hookrightarrow$ RV $\hookrightarrow$ Specific value

$F(x) = P(X \le x)$ CDF goes from 0 to 1.

Basic Events

Discrete : $X = x$ $P(X = x) = p(x)$ PMF

Continuous: $X \in (x, x+\Delta x)$ $P(X \in (x+\Delta x)) \approx f(x)\Delta x$ for small Δx

$$f(x) = \lim_{\Delta x \to 0} \frac{P(X \in (x, x+\Delta x))}{\Delta x}$$ PDF

PDF: density $= \frac{\text{\# of people in } (x; x+\Delta x)}{\Delta x}$

Summarizing $p(x)$ or $f(x)$:

Expectation:

$$E(X) = \begin{cases} \sum_x x \cdot p(x) \\ \int x f(x)\,dx \end{cases}$$
- Average
- Long run average

Variance/Volatility:

$$Var(x) = E[(X - E(x))^2] = E[X^2] - E[X]^2$$

$$aX+b \Rightarrow E[aX+b] = aE[X]+b$$
$$var(aX+b) = a^2 Var[X]$$

Two Random Variables

$$COV(X,Y) = E((X-\mu_x)(Y-\mu_Y))$$


cov>0


cov<0

$$X+Y: E[X+Y] = E[X] + E[Y]$$

$$VAR[X+Y] = E[[(X+Y)-(\mu_X+\mu_Y)]^2] = E[((X-\mu_X)+(Y-\mu_Y))^2]$$
$$= E[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y))$$
$$= Var(X) + Var(Y) + 2\,Cov(X,Y)$$

If X & Y are independent, $E(XY) = E[X]E[Y]$, $cov(X,Y) = 0$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $E[\bar{X}] = \mu$, $var(\bar{X}) = \frac{\sigma^2}{n}$

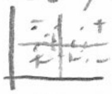
STATS 105 10/08

$(X, Y) \sim f(x,y)$ $\quad f(x,y) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \dfrac{P(X \in (x, x+\Delta x) \;\&\; Y \in (y, y+\Delta y))}{\Delta x \Delta y}$

$\mathrm{COV}(X,Y) = E[(x-\mu_x)(y-\mu_y)]$

cov>0

cov<0



cov=0

$$\mathrm{CORR}(X,Y) = \frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \mathrm{COV}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right)$$

Vector Representation

Ω	X	Y
1	$X(1)$	$Y(1)$
2	$X(2)$	$Y(2)$
$\vdots$	$\vdots$	$\vdots$
ω	$X(\omega)$	$Y(\omega)$
$\vdots$	$\vdots$	$\vdots$
m	$X(m)$	$Y(m)$

$\rightarrow$

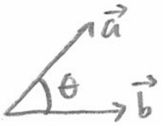
Ω	$\tilde{X}$	$\tilde{Y}$
	$X(1)-\mu_x$	$Y(1)-\mu_y$
	$X(2)-\mu_x$	$Y(2)-\mu_y$
	$\vdots$ $\tilde{X}(\omega)$	$\vdots$
	$X(\omega)-\mu_x$	$Y(\omega)-\mu_y$
	$\vdots$	$\vdots$
	$X(\omega)-\mu_x$	$Y(\omega)-\mu_y$

Recall:

$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$ $\quad \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$

$\langle \vec{a}, \vec{b} \rangle = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^{n} a_i b_i$

$|\vec{a}|^2 = \langle \vec{a}, \vec{a} \rangle = \sum_{i=1}^{n} a_i^2$



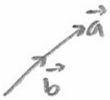
$$\cos\theta = \frac{\langle \vec{a}, \vec{b} \rangle}{|a| \cdot |b|} = \frac{\frac{1}{M}\langle \vec{a}, \vec{b} \rangle}{\sqrt{\frac{1}{M}|\vec{a}|^2}\sqrt{\frac{1}{M}|\vec{b}|^2}} = \frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \in (-1,1)$$

$\mathrm{Corr}(X,Y) = \cos\theta$

$$\frac{1}{M}\langle \vec{a}, \vec{b} \rangle = \frac{1}{M}\sum_{\omega=1}^{M} (X(\omega)-\mu_x)(Y(\omega)-\mu_y) = E[(X-\mu_x)(Y-\mu_y)] = \mathrm{COV}(X,Y)$$

$$\frac{1}{M}|\vec{a}|^2 = \frac{1}{M}\sum_{\omega=1}^{M} (X(\omega)-\mu_x)^2 = E[(X-\mu_x)^2] = \mathrm{Var}(X)$$

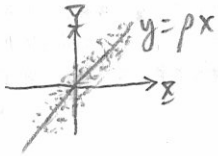
Corr = 1:



$\vec{b} = \lambda \vec{a}$, $\lambda > 0$

For Simplicity Assume

$E[X] = E[Y] = 0$ \$ $Var(X) = Var(Y) = 1$



Find ρ, min $E[(Y - \rho X)^2]$ (Least Square Estimation)

$$= \frac{1}{M}\sum_{\omega=1}^{M}(Y(\omega) - \rho X(\omega))^2 = \frac{1}{M}|\vec{b} - \rho\vec{a}|^2$$

$$\frac{|\rho\vec{a}|}{|\vec{b}|} = \cos\theta$$

$|\vec{b}| = |\vec{a}|$

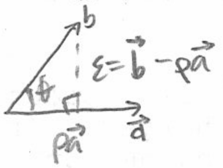
$$1 - \rho^2 = \frac{|\vec{b} - \rho\vec{a}|^2}{|\vec{b}|^2} = \frac{E[(Y - \rho X)^2]}{Var(Y)}$$

$\rho = \cos\theta = corr(X, Y)$

$R^2 = \cos^2\theta = \frac{|\rho\vec{a}|^2}{|\vec{b}|^2}$ = % variation of Y explained by linear relationship

↳ strength of linear relationship

$Var(Y) = Var(\rho X) + Var(\varepsilon)$



$Y = \rho X + \varepsilon$

STATS 105 10/10

Independently & Identically Distributed (iid) Random Vorriables

$X_1, X_2, \ldots, X_n \underset{iid}{\sim} f(x)$

If $X \sim f(x)$

$\mu = E[X]$
$\sigma^2 = Var[X]$ $\Rightarrow S = \sum_{i=1}^{n} X_i \;,\; \bar{X} = \frac{S}{n}$

$$E[S] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = n\mu$$

$$Var[S] = Var(\sum_{i=1}^{n} X_i) = \sum Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j) = n\sigma^2$$

$$E[\bar{X}] = \frac{E[S]}{n} = \frac{n\mu}{n} = \mu$$

$$Var[\bar{X}] = Var\left(\frac{S}{n}\right) = \frac{1}{n^2} Var(S) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Law of Large Numbers

$$P(|\bar{X} - \mu| > \varepsilon) \xrightarrow[\infty]{n} 0$$

Central Limit Theorem

$$\sqrt{n}(\bar{X} - \mu) \rightarrow E[\sqrt{n}(\bar{X} - \mu)] = 0$$
$$Var[\sqrt{n}(\bar{X} - \mu)] = \sigma^2$$

Follows Normal distribution

$N(0, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

If $Y \sim N(0, \sigma^2)$

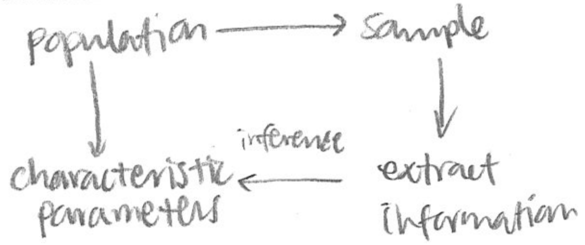
$$P(\sqrt{n}(\bar{X} - \mu) \leq y) \rightarrow P(Y \leq y), \forall y$$

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \rightarrow N(0,1).$$

$Z \sim N(0,1)$

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \quad P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z\right) \rightarrow P(Z \leq z)$$

STATISTICS



Two Purposes
- understanding
- predictions

Estimate Population proportion

Population $\longrightarrow$ n people

$X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p)

$X_i = 1$ if vote for

$\quad\ \ 0$ otherwise

$\downarrow$ (from Population) population proportion "p" $\longleftarrow$ $\hat{p} = \bar{X} = \frac{\sum_i x_i}{n}$

Bernoulli RV

$P(X_i = 1) = p \qquad E[X_i] = 1 * p + 0 * (1-p) = p$

$P(X_i = 0) = 1-p \qquad Var[X_i] = (1-p^2)*p + (0-p)^2*(1-p) = p(1-p)$

$\hat{p}$ is a random variable.

Hypothetic Repetition: Sample $X_1, \ldots, X_n$ again, get a different $\hat{p}$.

$E[\hat{p}] = E[\bar{X}] = \mu = p$ (unbiased)

$Var[\hat{p}] = Var[\bar{X}] = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$. $\quad s.d.(\hat{p}) = \sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$ if you don't know p, plug in $\hat{p}$.

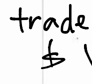
standard error, $s.e.(\hat{p}) = \widehat{s.d.(\hat{p})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\longleftarrow$

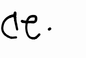
$\hat{p}_1 = \frac{S}{n} = \bar{X} \qquad E[\hat{p}_2] = \frac{E[S + \frac{\sqrt{n}}{2}]}{n+\sqrt{n}} = \frac{E[S] + \frac{\sqrt{n}}{2}}{n+\sqrt{n}} = \frac{np + \frac{\sqrt{n}}{2}}{n+\sqrt{n}} = \frac{\sqrt{n}p + \frac{1}{2}}{\sqrt{n}+1}$

$\hat{p}_2 = \frac{S + \frac{\sqrt{n}}{2}}{n+\sqrt{n}}$

Bias: $E[\hat{p}_2] - p = \frac{\frac{1}{2} - p}{\sqrt{n}+1}$

"unbiased" "biased"

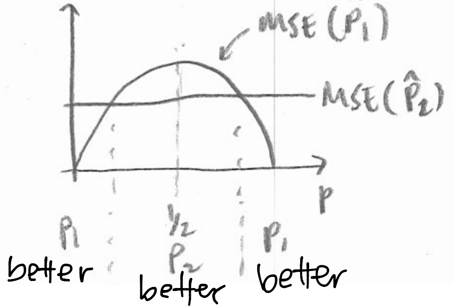




$Var(\hat{p}_2) = \frac{Var(S)}{(n+\sqrt{n})^2} = \frac{np(1-p)}{(n+\sqrt{n})^2} = \frac{p(1-p)}{(\sqrt{n}+1)^2} < Var(\hat{p}_1)$ trade off between bias & variance.

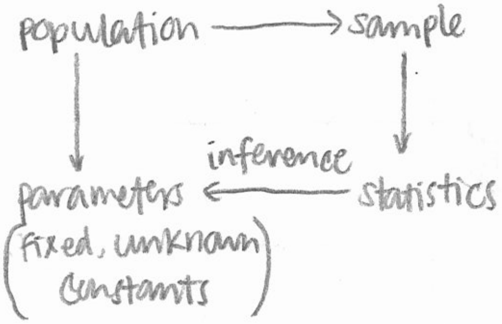
$MSE(\hat{p}_1) = \frac{p(1-p)}{n}$

$MSE(\hat{p}_2) = \frac{(\frac{1}{2}-p)^2}{(\sqrt{n}+1)^2} + \frac{p(1-p)}{(\sqrt{n}+1)^2} = \frac{\frac{1}{4}}{(\sqrt{n}+1)^2}$



STATS 105 10/15

Point Esimation



$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x,\theta)$

sample, data. model of population

$\hat{\theta} = h(x_1, x_2, \ldots, x_n)$

Estimate is a function of the data

$h(\,)$: estimator

Value of $\hat{\theta}$: estimate

$X_1, X_2, \ldots, X_n$ are random $\Rightarrow$ different under hypothetical repetitions and as a result, $\hat{\theta}$ is also random.

bias: $E[\hat{\theta}] - \theta$. unbiased if $E[\hat{\theta}] = \theta$. ✓

variance: $Var(\hat{\theta})$ ✓

Mean Square Error (MSE): $E[(\hat{\theta}-\theta)^2] = bias(\hat{\theta})^2 + Var(\hat{\theta})$ ✓

Proof: $\mu = E[\hat{\theta}] \Rightarrow E[(\hat{\theta}-\theta)^2] = E[((\hat{\theta}-\mu)-(\theta-\mu))^2]$

$$= E[(\hat{\theta}-\mu)^2 + (\theta-\mu)^2 - 2(\hat{\theta}-\mu)(\theta-\mu)]$$

$$= E[\hat{\theta}-\mu]^2 + E[(\theta-\mu)^2] - 2E[(\hat{\theta}-\mu)(\theta-\mu)]$$

$$= Var(\hat{\theta}) + \underbrace{[E(\hat{\theta})-\theta]^2}_{\text{bias}} - 2\cancel{(\theta-\mu)E[\hat{\theta}-\mu]} \hookrightarrow = 0$$

Example:

$X_1, X_2, \ldots, X_n \sim Bernoulli(p)$

$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$

$X_1, X_2, \ldots, X_n \sim Exponential(\lambda)$ $f(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$ (waiting time)

Method of Moments

Find θ. $E_\theta(x) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ → estimating equation → solution: $\hat{\theta}$

Another notation $\langle X \rangle_{f_\theta} = \langle X \rangle_{data}$

(model)

$X_1, X_2, \ldots, X_n \sim N(\mu, \underset{\text{known}}{\sigma^2})$ $E_\mu[X] = \mu$

$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \underset{\text{solution}}{\hat{\mu}} = \bar{X}$

Get this by trying to match model to data.

$X_1, X_2, \dots, X_n \sim Exp(\lambda)$. $E[X] = \frac{1}{\lambda}$

estimating equation: $\frac{1}{\lambda} = \bar{X}$

solution: $\hat{\lambda} = \frac{1}{\bar{X}}$

$X_1, X_2, \dots, X_n \sim Bernoulli(p)$. $E[X] = p$

estimating equation: $p = \bar{X}$

solution : $\hat{p} = \bar{X}$

Two Unknown Parameters

$X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$

↳ unknown

$E[X] = \mu$

① $\mu = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Solution: $\hat{\mu} = \bar{X}$

$E[X^2] = E(X)^2 + Var(\bar{x}) = \mu^2 + \sigma^2$

② $E[X^2] = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

$\mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

Solution: $\hat{\sigma}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

In general: $X_1, X_2, \dots, X_n \sim f(x, \theta)$

$\theta = (\theta_1, \theta_2, \dots, \theta_m)$ needs "m" estimating equations

Estimating Equations: $E_\theta(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

Solutions: $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$

More Generally: $E_\theta(h_k(x)) = \frac{1}{n}\sum_{i=1}^{n} h_k(x_i)$, $k = 1, \dots, m$

Best Estimation Equation: maximum likelihood estimator

$X_1, X_2, \dots, X_n \overset{iid}{\sim} f(x, \theta_{TRUE})$

Likelihood : $L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$ $\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta)$

↳ independent $(X_1, X_2, \dots, X_n)$

$L(\theta)$ measures plausibility of θ in explaining $X_1, X_2, \dots, X_n$

MLE: most plausible explanation

log-likelihood

$$\ell(\theta) = \text{Log}(\theta) = \sum_{i=1}^{n} \log f(x_i ; \theta)$$

$$\ell'(\theta) = 0$$

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(x_i ; \theta) = 0 \quad \text{(MLE Equation)}$$

STATS 105

10/17

Maximum Likelihood Estimator (MLE)

$X_1, X_2, \dots, X_n \overset{iid}{\sim} p(x, \theta)$ Likelihood: $L(\theta) = \prod_{i=1}^{n} p(X_i; \theta)$

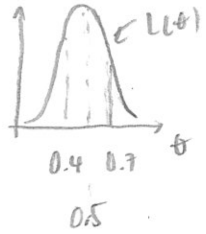
in general: $L(\theta) = \text{prob}(\text{data}; \theta)$

plausibility of θ for explaining the data

$\text{Prob}(\text{data}; \theta_1) > \text{Prob}(\text{data}; \theta_2) \Rightarrow \theta_1$ is more plausible than θ_2.

we want the most plausible value for θ.

$\hat{\theta}_{MLE} = \arg\max L(\theta)$



fix data in each repitition

θ changes as variable

If data $= (X_1, X_2, \dots, X_n)$ are independent, recall A & B independent

$P(A \cap B) = P(A)\,P(B)$

$P(X_1, X_2, \dots, X_n; \theta) = \prod_{i=1}^{n} P(X_i, \theta)$

log-Like: $l(\theta) = \text{Log } L(\theta) = \sum \log p(X_i; \theta)$

derivative: $l'(\theta) = \sum \frac{\partial}{\partial \theta} \log p(X_i; \theta)$

Estimating eqn: $l'(\theta) = 0$

Solve: $\hat{\theta}_{MLE}$ $l''(\hat{\theta}_{MLE}) < 0$

Continuous

$X_1, X_2, \dots, X_n \overset{iid.}{\sim} f(x; \theta)$ density

If $X \sim f(x; \theta)$, $\text{Prob}(X \in (x, x + \underset{\overset{\downarrow}{\text{precision}}}{\Delta x})) = f(x; \theta) \cdot \Delta x$

$L(\theta) = \prod_{i=1}^{n} \left(f(x; \theta) \overset{\text{constant}}{\Delta x} \right)$

Example 1:

$X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p)$

$X \sim \text{Ber}(p)$

	0	1
Prob	(1-p)	p

$P(X = x) = p^{x} (1-p)^{1-x}$

$$L(p)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$\ell(p)=\sum_{i=1}^{n}\left(X_i \log p + (1-X_i)\log(1-p)\right)$$

$$\ell'(p)=\sum_{i=1}^{n}\left(X_i\cdot\frac{1}{p} + (1-X_i)\cdot\frac{1}{1-p}\cdot(-1)\right)$$

Estimating equation: $\ell'(p)=0$

freq of 1s.

Let $S=\sum X_i$ = # of 1's. $\frac{S}{p}-\frac{n-S}{1-p}=0 \Rightarrow S(1-p)=(n-S)p \Rightarrow \hat{p}=\frac{S}{n}$

$$\ell''(p)=-\frac{S}{p^2}-\frac{n-S}{(1-p)^2}<0$$

Example 2

$$X_1, X_2, \dots, X_n \sim N(\overset{?}{\mu}, \overset{\text{known}}{\sigma^2})$$

$$f(x,\mu)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

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

$$\min R(\mu)=\sum_{i=1}^{n}(X_i-\mu)^2 \Rightarrow R'(\mu)=\sum_{i=1}^{n}2(X_i-\mu)(-1)=0$$

$$\sum(X_i)-n\mu=0 \Rightarrow \hat{\mu}=\frac{\sum X_i}{n}$$

Mixture Model

$$f(x,\theta)=\lambda\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)+(1-\lambda)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

$\theta=(\lambda,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$ (Method of moments could be used)

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

Regression

Observations	Predictor	Response
1	X_1	Y_1
2	X_2	Y_2
⋮	⋮	⋮
n	X_n	Y_n

error $\sim N(0,\sigma^2)$

$$Y_i \sim \alpha+\beta X_i+\varepsilon_i$$

fixed

$$Y_i \sim N(\alpha+\beta X_i, \sigma^2)$$

$$L(\theta)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(Y_i-(\alpha+\beta X_i))^2}{2\sigma^2}\right)$$

$$L(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - (\alpha + \beta X_i)\right)^2\right)$$

$$\min R(\theta) = \sum_{i=1}^{n}\left(Y_i - (\alpha + \beta X_i)\right)^2$$

$\hat{\alpha}, \hat{\beta}$ Least Square Estimators

Logistic Regression

$y_i \in \{0, 1\}$
$\{-, +\}$

$y_i \sim Ber(p)$ $\qquad \log \frac{p_i}{1-p_i} = \alpha + \beta x_i$

STATS 105 10/22

MLE

Example:

$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) \Rightarrow L(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right), \quad \theta = (\mu, \sigma^2)$$

$$\ell(\theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

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

$$= \text{constant} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = \frac{-1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)(-1) \Rightarrow \frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \quad (1)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad (2)$$

$$(1)\ \sum_{i=1}^{n}(x_i-\mu) = 0 \Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$(2)\ -n(\sigma^2) + \sum(x_i-\mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\ell'(\theta) = \begin{pmatrix} \frac{\partial \ell}{\partial \mu} \\ \frac{\partial \ell}{\partial \sigma^2} \end{pmatrix} \qquad \ell''(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} & \frac{\partial^2 \ell}{\partial (\sigma^2)^2} \end{pmatrix} \qquad \ell''(\hat{\theta}) < 0$$

negative definite

Example: Normal Linear Regression

observations	predictor	response
1	X_1	Y_1
2	X_2	Y_2
⋮	fixed	random
n	X_n	Y_n

$\{Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu_i, \sigma^2)$

$\mu_i = x_i\beta$

$y_i = x_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

$$f(y|x,\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left(-\frac{(y-x_i\beta)^2}{2\sigma^2}\right)$$

$$L(\theta) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right), \quad \mu_i = x_i\beta$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(y_i-x_i\beta)^2}{2\sigma^2}\right)$$

$E(\varepsilon_i) = 0$

$Var(\varepsilon_i) = \sigma^2$

$E[\mu_i + \varepsilon_i] = \mu_i$

$Var[\mu_i + \varepsilon_i] = \sigma^2$

$$l(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum (y_i - x_i\beta)^2$$

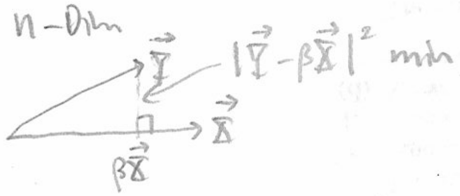
$$\underset{\substack{\downarrow \\ \text{max likelihood}}}{\max l(\theta)} \Longrightarrow \underset{\substack{\downarrow \\ \text{least squares}}}{\min \sum (y_i - x_i\beta)^2} = R(\beta)$$

$$\frac{\partial l}{\partial \beta} = \frac{-1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - x_i\beta)(-x_i) \quad (1)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum (y_i - x_i\beta)^2 \quad (2)$$

(1) $\sum (y_i - x_i\beta)x_i = 0 \Rightarrow \sum y_i x_i - \sum x_i^2 \beta = 0 \Rightarrow \hat{\beta} = (\sum X_i^2)^{-1} \sum x_i y_i$

(2) $\hat{\sigma}_2 = \dfrac{\sum (y_i - x_i\beta)^2}{n}$



$\langle \vec{X}, \vec{Y} - \beta\vec{X} \rangle = 0$ (inner product of orthogonal)

$\sum x_i(y_i - x_i\beta) = 0$

Multiple Linear Regression:

$y_i = \beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_p x_{i_p} + \varepsilon_i = y_i = \vec{X}_i^{T}\vec{\beta} + \varepsilon_i$

$$\underset{p\times 1}{\hat{\beta}} = \underset{p\times p}{\left(\sum \vec{X}_i X_i^{T}\right)^{-1}} \underset{p\times 1}{\left(\sum \vec{X}_i y_i\right)} \qquad \vec{x}_i = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad \vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Example:

$y_i \sim Ber(p_i)$ (x_i – height, y_i – gender)

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i\beta \Longleftrightarrow p_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} = \frac{1}{1+e^{-x_i\beta}}$$

$$L(\beta) = \prod_{i=1}^{n} p(y_i | x_i, \beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n}\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)^{y_i}\left(\frac{1}{1+e^{x_i\beta}}\right)^{1-y_i}$$

$$= \prod_{i=1}^{n}\left(\frac{e^{y_i x_i\beta}}{1+e^{x_i\beta}}\right) \Rightarrow l(\beta) = \sum_{i=1}^{n}\left(y_i x_i\beta - \ln(1+e^{x_i\beta})\right)$$

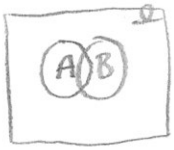
$l'(\beta) = 0 \Rightarrow \hat{\beta}_{MLE}$; $l'(\beta) = \sum_{i=1}^{n}\left(y_i x_i - \overbrace{\frac{e^{x_i\beta}}{1+e^{x_i\beta}}}^{\rightarrow p_i} \cdot x_i\right)$ want p_i close to y_i.

STATS 105 10/24

Bayes Rule

Conditional probability: $P(A|B) = \frac{P(A\cap B)}{P(B)}$

Geometric:



$P(A) = \frac{|A|}{|\Omega|}$

$P(A|B)$ (falls into B)

$P(A|B) = \frac{|A\cap B|}{|B|}$ how much A occupies B.

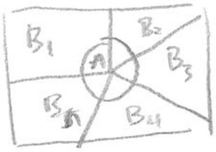
slogan: conditional probability is just a regular probability except B is a regular probability in a conditional probability, $P(A) = P(A|\Omega)$.

Frequency: $P(A|B)$: how often A happens when B happens.

Chain Rule: $P(A\cap B) = P(B)\,P(A|B)$
$= P(A)\,P(B|A)$

$$\begin{tikzcd} A \arrow[d, "\text{generative/causal}"] \\ B \arrow[u, bend left, dashed, "\text{inferential}"] \end{tikzcd}$$

Rule of Total Probability



Given: $P(B_i)$, $i=1,\dots,n$

$P(A) = \sum_{i=1}^{n} P(A\cap B_i)$ → Additivity

$= \sum_{i=1}^{n} P(B_i)\,P(A|B_i)$

$P(B_i|A) = \frac{P(B_i\cap A)}{P(A)}$

$= \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)}$

Example:

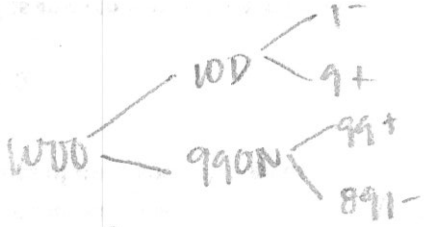
1% of population has disease.

Medical Test: $P(+|\text{Disease}) = 90\%$

$P(-|\text{Disease}) = 10\%$

$P(-|\text{Not Disease}) = 90\%$

$P(+|\text{Not Disease}) = 10\%$



$\frac{9}{99+9} = \frac{1}{12}$

$$P(D|+) = \frac{P(D\cap +)}{P(+)} = \frac{P(D)P(+|D)}{P(D\cap +) + P(N\cap +)} = \frac{P(D)P(+|D)}{P(D)P(+|D) + P(N)P(+|N)}$$

$$\begin{tikzcd} \text{Experiment} \arrow[r] & \text{outcome} \arrow[r] & \# \\ & \text{statement about} \arrow[u] \arrow[r] & \text{equation or inequality about} \arrow[u] \end{tikzcd}$$

$(X, Y) \sim P(x,y) = P(X=x, Y=y)$

$P(x) = P(X=x) = \sum_{y} P(x,y)$ $\quad P(x|y) = P(X=x|Y=y) = \frac{P(x,y)}{P(y)}$

$P(Y=y) = \sum_{x} P(x,y) = \sum_{x} P(x)\, P(y|x)$

$P(x|y) = \frac{P(x)P(y|x)}{\sum_{x} p(x)p(y|x)}$ $\quad$ $p(x)$ prior probability

$p(x|y)$ posterior probability

$(X, Y) \sim f(x,y)$

$P(X \in (x, x+\Delta x) \cap Y \in (y, y+\Delta y)) = f(x,y)\, \Delta x \Delta y$

$f(x) = \int f(x,y)\,dy$ $\quad f(y|x) = \frac{f(x,y)}{f(x)}$

$f(y) = \int f(x,y)\,dx$

$f(x|y) = \frac{f(x,y)}{f(y)}$

$f(y) = \int f(x,y)\,dx = \int f(x) f(y|x)\,dx$

$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(x,y)}{\int f(x) f(y|x)\,dx}$

Mixture Model:

$X \in \{0, 1\}$ gender $\quad f(y|0) = f_0(y)$ $\quad X \sim Ber(\lambda)$ $\quad P(1) = (\lambda)$

Y continuous height $\quad f(y|1) = f_1(y)$ $\quad P(0) = (1-\lambda)$

$f(y) = \sum_{x} p(x) f(y|x) = P(1) f(y|1) + P(0) f(y|0) = \lambda f_1(y) + (1-\lambda) f_0(y)$

$p(1|y) = \frac{f(1,y)}{f(y)} = \frac{p(1)\, f(y|1)}{p(1)\, f(y|1) + p(0)\, f(y|0)} = \frac{\lambda f_1(y)}{\lambda f_1(y) + (1-\lambda) f_0(y)}$

Normal Inference

$X \sim N(\mu, \sigma^2)$ $\quad f(x|y) = \frac{f(x)f(y|x)}{\int f(x)f(y|x)dx}$

$Y = X + \varepsilon$, $\varepsilon \sim N(0, \tau^2)$

$(Y | X=x) \sim N(x, \tau^2)$

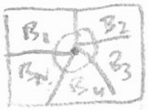
STATS 105

10/29

Probability Calculations

Events:

Chain Rule: $P(A \cap B) = P(B)\,P(A|B)$

Total: 

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)\,P(A|B_i)$$

Bayes Rule: $$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(B_i)\,P(A|B_i)}{\sum_{j=1}^{n} P(B_j)\,P(A|B_j)}$$

Random Variable:

Discrete: Chain Rule: $P(X,Y) = P(x)\,P(y|x)$

Marginality: $p(y) = \sum_x P(x,y) = \sum_x P(x)\,p(y|x)$

Bayes Rule: $$P(x|y) = \frac{P(x,y)}{P(y)} = \frac{P(x)\,P(y|x)}{\sum_x P(x)\,P(y|x)}$$

Continuous:

$$f(x,y) = f(x)\,f(y|x)$$

$$f(y) = \int f(x,y)\,dx = \int f(x)\,f(y|x)\,dx$$

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(x)\,f(y|x)}{\int f(x)\,f(y|x)\,dx}$$

Mixture Model Example:

$X \sim Ber(\lambda)$

X	0	1
Pro	$1-\lambda$	λ

$P(1) = \lambda$
$P(0) = 1-\lambda$

$[Y|X=1] \sim f_1(y) = f(y|1)$
$[Y|X=0] \sim f_0(y) = f(y|0)$

$f(y,x) = f(y|x)\,p(x)$
$f(y,1) = f(y|1)\,p(1) = \lambda f_1(y)$
$f(y,0) = f(y|0)\,P(0) = (1-\lambda)\,f_0(y)$

$$f(y) = \sum_x f(y,x) = \sum_x f(y|x)\,p(x) = \lambda f_1(y) + (1-\lambda) f_0(y)$$

$$P(x|y) = \frac{f(y,x)}{f(y)} = \frac{f(y|x)\,P(x)}{f(y)}$$

$$P(1|y) = \frac{f_1(y)(\lambda)}{\lambda f_1(y) + (1-\lambda) f_0(y)} \qquad P(0|y) = \frac{(1-\lambda) f_0(y)}{\lambda f_1(y) + (1-\lambda) f_0(y)}$$

$$f(y,x) = (\lambda f_1(y))^x\,((1-\lambda) f_0(y))^{1-x}$$

Current Population: proportion

$M = 300$ mil, male $= \lambda$, female: $(1-\lambda)$

\# of males: $M \cdot \lambda \rightarrow$ \# of males in $(y, y+\Delta y) = M\lambda f_1(y)\Delta y$

females: $M \cdot (1-\lambda) \rightarrow$ \# of females in $(y, y+\Delta y) = M(1-\lambda) f_0(y)\Delta y$

$$\# \text{ of ppl in } (y, y+\Delta y) = M\underbrace{(\lambda f_1(y) + (1-\lambda) f_0(y))}_{f(y)}\Delta y$$

Among all ppl in $(y, y+\Delta y)$:

$$\text{proportion male: } \frac{\cancel{M}\lambda f_1(y)\cancel{\Delta y}}{\cancel{M}(\lambda f_1(y) + (1-\lambda) f_0(y))\cancel{\Delta y}} = P(1|y) = P(X=1 \mid Y \in (y, y+\Delta y))$$

Example:

$(X_1, Y_1), (X_2, Y_2) \cdots (X_n, Y_n)$, $X_i \sim Ber(\lambda)$

$[Y_i \mid X_i = 1] = f_1(y) \sim N(\mu_1, \sigma_1^2)$ $\theta = (\lambda, \mu_1, \sigma_1^2, \mu_0, \sigma_0^2)$

$[Y_i \mid X_i = 0] = f_0(y) \sim N(\mu_0, \sigma_0^2)$

$$L(\theta) = \prod_{i=1}^{n} f(y_i, x_i) = \prod_{i=1}^{n} [\lambda f_1(y_i)]^{x_i} [(1-\lambda) f_0(y_i)]^{1-x_i}$$

$$l'(\theta) = 0 \Rightarrow \text{MLE} \quad \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \hat{\mu}_1 = \frac{\sum y_i x_i}{\sum x_i}, \quad \hat{\mu}_0 = \frac{\sum y_i (1-x_i)}{\sum (1-x_i)}$$

What if we do not observe X_i, $i = 1, \ldots, n$?

$$L(\theta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} (\lambda f_1(y) + (1-\lambda) f_0(y))$$

Expectation Maximization (EM) Algorithm

E step: Given current parameter values $\lambda^{(t)}, \mu_0^{(t)}, \sigma_0^{2(t)}, \mu_1^{(t)}, \sigma_1^{2(t)}$ Guess $\hat{x}_i = \dfrac{\lambda^{(t)} f_1(y)}{\lambda^{(t)} f_1(y) + (1-\lambda^{(t)}) f_0(y)}$

M step: $\lambda^{(t)} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

STATS 105   Homework Due Thurs after Midterm   10/31

All Homework Problems: All calculations in notes. Don't skip any steps

Bayes Rule

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

$Y \mid X = x \sim N(x, \tau^2)$

$$X \sim f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$[Y \mid X] \sim f(y \mid x) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(y-x)^2}{2\tau^2}\right)$$

$$[X, Y] \sim f(x,y) = f(y \mid x) f(x)$$

$$[X \mid Y] \sim f(x \mid y) = \frac{f(x,y)}{f(y)}$$ density of x w/ y fixed.

$$f(x \mid y) \propto f(x,y) \Rightarrow f(x \mid y) \propto \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma^2} - \frac{2\mu}{\sigma^2}x + \frac{x^2}{\tau^2} - \frac{2y}{\tau^2}x\right)\right)$$

$$f(x \mid y) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)x^2 - 2\left(\frac{\mu}{\sigma^2} + \frac{y}{\tau^2}\right)x\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)\left(x - \frac{\frac{\mu}{\sigma^2} + \frac{y}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)^2\right)$$

$$[X \mid Y] \sim f(x \mid y) \sim N\left(\frac{\frac{\mu}{\sigma^2} + \frac{y}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

Fix σ^2, $\tau^2 \to 0 \Rightarrow$ posterior mean $= y$

fix τ^2, $\sigma^2 \to 0 \Rightarrow$ posterior mean $= \mu$

↳ compromise between prior & evidence weighted by 1/var = precision

Inference:

$$X_1, X_2, \dots, X_n \sim N(\overset{\nearrow\text{unknown}}{\mu}, \overset{\text{given}}{\sigma^2}) \Rightarrow \hat{\mu}_{MLE} = \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Bayesian:

$$\mu \sim N(\overset{\nearrow\text{given}}{\alpha}, \tau^2) \Rightarrow [\mu \mid \bar{X}] \sim N\left(\frac{\frac{\alpha}{\gamma^2} + \frac{\bar{x}n}{\sigma^2}}{\frac{1}{\gamma^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\gamma^2} + \frac{n}{\sigma^2}}\right)$$

$$n \to \infty \qquad \begin{matrix}\downarrow \\ \bar{x}\end{matrix}$$

Bayesian vs. Frequentist:

μ: speed of light

Frequentist: μ unknown constant

Bayesian: since unknown, random variable

$$L(\theta) = f(x_1, x_2, \ldots, x_n | \mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

MLE: $\min \sum_{i=1}^{n} (x_i - \mu)^2 \longrightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$

regularization by penalty (penalize large μ)

$$\min \left[\sum_{i=1}^{n} (x_i - \mu)^2 + \underbrace{\lambda \mu^2}_{\text{prefer small } \mu}\right]$$

As if prior $\mu \sim f(\mu) \propto \exp(-\lambda\mu^2) \sim N(0, \frac{1}{2\lambda})$

Posterior $[\mu | x] \propto f(x_1, x_2, \ldots, x_n | \mu) f(\mu) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\lambda\mu^2}{2\sigma^2}\right)$

Recall Regression

obs	predictor	response
1	x_1	y_1
2	x_2	y_2
3	$\vdots$	$\vdots$
$\vdots$		
n	x_n	y_n

$y_i = \beta x_i \Rightarrow \min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2$

if x_i is p dim & so is β

$$\min_{\beta} \left[\sum_{i=1}^{n} (y_i - \vec{x}_i^T \vec{\beta})^2 + \underbrace{\lambda |\vec{\beta}|^2}_{\text{curve smooth.}}\right]$$

Election:

State wise probability/proportion biased polls

p_i: probability in state i.

$x_i | n_i, p_i \sim Bin(n_i, p_i)$

$p_i \sim N(\alpha, \gamma^2)$

Borrow strength from other similar states.

Batting Averages

player i $\overset{\text{hits}}{X_i} | \overset{\text{attempts}}{n_i}, p_i \sim Bin(n_i, p_i)$

$p_i \sim N(\alpha, \gamma^2)$

MLE $\hat{p}_i = \frac{X_i}{n_i}$

Baysian: pull $\frac{X_i}{n_i}$ towards α

STATS 105

10/31

Problem 3:

(1) $f(x,y) = f_x(x) f(y|x) \quad \propto \exp\left(-\frac{(y-\mu)^2}{2(\sigma^2+\tau^2)}\right) \Rightarrow N(\mu, \sigma^2+\tau^2)$

(2) $f(x|y) = \frac{f(x,y)}{f(y)}$ ← part(a) $\quad f(y) = \int f(x,y)\,dx$

Problem 4:

(1) MLE of μ. $L(\mu) \Rightarrow \ell(\mu) \Rightarrow \hat{\mu}_{MLE}$

(2) $p(\mu)$, $p(x_i|\mu)$ known, $p(\mu|x)$?

$$p(\mu|x) \propto p(\mu) \underbrace{p(x|\mu)}_{\prod_{i=1}^{n} p(x_i|\mu)} \quad \sim \text{Normal}$$

Midterm Review

$$\mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i) + \overset{\frac{n(n-1)}{2} \text{ terms}}{\sum_{i \neq j} \mathrm{Cov}(X_i, X_j)}$$

$E[aX+b] = aE[X] + b$

$\mathrm{Var}[aX+b] = a^2 \mathrm{Var}[X]$

MLE for θ

$L(\theta) \Rightarrow \ell(\theta) \Rightarrow \hat{\theta}_{MLE}$

(+/-) $\mathrm{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

$\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta})$

$\mathrm{Var}[X] = E[X^2] - E^2[X]$

$\mathrm{COV}[X,Y] = E[XY] - E[X]E[Y]$

$\mathrm{CORR}[X,Y] = \frac{\mathrm{COV}[X,Y]}{\sqrt{\mathrm{VAR}[X]\,\mathrm{VAR}[Y]}}$

$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{COV}(X,Y)$

$\mathrm{Ber}(p)$: $p(x=1) = p$, $p(x=0)$ $1-p$

$\mathrm{Bin}(n,p) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

Normal:

Exponential(λ) $f(x) = \lambda e^{-\lambda x}$

STATS 405 11/07

Confidence Interval

Probability Background

$X_1, X_2, \dots, X_n \sim f(x)$

for $X \sim f(x)$ $\mu = E[X]$, $\sigma^2 = Var[X]$

$\bar{X} = \frac{\sum x_i}{n} \Rightarrow E[\bar{X}] = \mu$, $Var[\bar{X}] = \frac{\sigma^2}{n}$

Law of Large #

$\bar{X} \xrightarrow[\infty]{n} \mu$ $P(|\bar{X} - \mu| > \overset{\text{fixed}}{\varepsilon}) \xrightarrow[\infty]{n} 0$

Central Limit Theorem

$\sqrt{n}(\bar{X} - \mu) \longrightarrow N(0, \sigma^2)$

$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0,1)$

Inference of population proportion

$$\begin{tikzcd}
\text{population} \arrow[r] \arrow[d] & \text{sample } (X_1, X_2, \dots, X_n \sim Ber(p)) \arrow[d] \\
\text{population proportion}(p) & \text{sample proportion}\left(\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}\right) \arrow[l]
\end{tikzcd}$$

L.L.N. $\hat{p} \to p$ consistent estimator

C.L.T. $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

$s^2 = \frac{\hat{p}(1-\hat{p})}{n}$; $s(\hat{p})$: standard error

Approximately $\hat{p} \sim N\left(p, \sigma^2 = \frac{p(1-p)}{n}\right)$, $s^2 = \hat{\sigma}^2$

Confidence interval of 95%

$\hat{p} \pm 1.96\, s(\hat{p})$ $[\hat{p} - 1.96\, s(\hat{p}), \hat{p} + 1.96\, s(\hat{p})]$

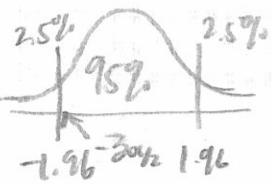
Margin of error $= 1.96\, s(\hat{p})$

Example : Call 1000 ppl $\hat{p} = 0.53$

CI : $0.53 \pm 1.96\left(\sqrt{\frac{0.53(1-0.53)}{1000}}\right)$

$$P\left(\underset{\underset{\text{random}}{\uparrow}}{\left[\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right]} \ni \underset{\uparrow\text{fixed}}{p}\right) = 95\%$$

$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$



$P(Z\in(-1.96, 1.96)) = 95\%$

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}} \Leftrightarrow -1.96 \le \frac{p-\hat{p}}{\sqrt{\frac{p(1-p)}{n}}} \le 1.96$$

1-α Confidence Interval

$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ 99% $\hat{p} \pm z_{.5\%}\sqrt{\frac{p(1-p)}{n}}$

Normal: $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$

$\hat{\mu} = \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ $95\% = \hat{\mu} \pm 1.96\, s(\hat{p})$

$\hat{\mu} \pm 1.96\frac{\sigma}{\sqrt{n}}$

2.2 Baysien:

$\mu \sim N(\alpha, \gamma^2)$

$$[\mu|\bar{X}] \sim N\left(\frac{\mu\frac{1}{\gamma^2} + \bar{x}\frac{n}{\sigma^2}}{\frac{1}{\gamma^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\gamma^2}+\frac{n}{\sigma^2}}\right)$$

$[\mu|\bar{X}]$ (noninformative prior)
as $\gamma^2 \to \infty$ $N(\bar{X}, \frac{\sigma^2}{n})$

$$P(\underset{\underset{\text{random}}{\downarrow}}{\mu} \in \underset{\underset{\text{fixed interval}}{\downarrow}}{[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]}) = 95\%$$

σ^2 unknown

$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ ← already used one degree of freedom for estimating μ at $\bar{X}$. $E[S^2] = \sigma^2$

$\sum_{i=1}^{n}(x_i - \bar{X}) = \sum x_i - n\bar{X} = \sum x_i - \frac{n\sum x_i}{n} = 0$

For large n:

95% C.I. $= \bar{X} \pm 1.96\frac{s}{\sqrt{n}}$, replace σ^2 by s^2

For small n:

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ (distribution spreads out)   1-α CI: $\bar{X} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

added variability   (T distribution)

STATS 105 11/19

Hypothesis Testing

$X_1, X_2, \dots, X_n \sim Ber(p)$

$H_0: p=p_0$ simple "="

$H_1: p>p_0$ complex/composite "$\neq$" "$<$" "$>$"

Attitude: H_0 dependent

Reject H_0 only when there is "strong" evidence.

Test statistic (calculated from the data)

$\hat{p} = (\sum_{i=1}^{n} X_i)/n \overset{H_0}{\sim} N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$ ← Null distribution

$z(H_0)$ ← normalization under null hypothesis $\Rightarrow z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim} N(0,1)$

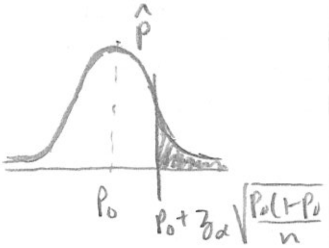
Compare $\hat{p}_{obs}$ or z_{obs} to their null distributions

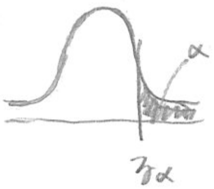
whether $\hat{p}_{obs}$ or z_{obs} is too extreme to be explained by H_0

Decision Rule:

Reject H_0 if $\hat{p} > p_0 + z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$

$z > z_\alpha$





Prob (Type 1 error) = α

Decision Rule ↔ Type I error

p-value (percentile)

p-value = P(observing something more extreme than the observed vale | H_0)

How extreme the observed value is.

$\hat{p}_{obs}=0.55$, $z=1$ p-value =  $= \int_1^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})\, dz$

Decision Rule with Type I = α

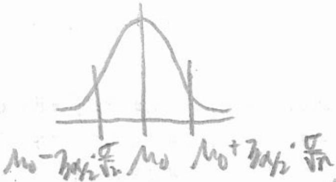
Reject H_0 if p-value $< \alpha$

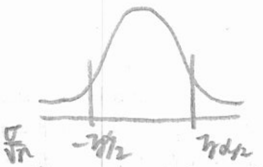
Example:

Admission

1. Admit if SAT score > 2300
2. Admit if SAT percentile > 98%

p-value < 2%





Two sided

known

$X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \overset{H_0}{\sim} N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \overset{H_0}{\sim} N(0,1)$$

Reject Ho if

$$|\bar{X} - \mu_0| > z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$|Z| > z_{\alpha/2}$$

$$P(|Z| \geq |z_{obs}|)$$

Example:

$H_0: X_1, X_2, \cdots, X_{300} \sim \text{Unif}[0,1]$

$H_1:$ not from Unif $[0,1]$

$\bar{X}\ obs = 0.52$

$\mu = E(X) = 0.5$

$$\sigma^2 = Var(X) = E(X^2) - E[X]^2$$
$$\frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

$$E(X^2) = \int_0^1 x^2 dx = \frac{X^3}{3}\Big|_0^1 = \frac{1}{3}$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1) \Rightarrow z_{obs} = \frac{\bar{X}_{obs} - 0.5}{\frac{1}{\sqrt{12}}/\sqrt{300}} = \frac{0.52 - 0.5}{1/60} = 1.2 < \overset{5\%}{1.96}$$

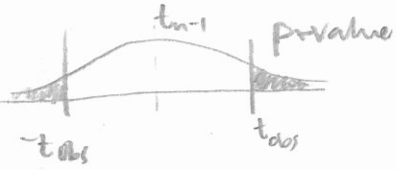
Two Sided

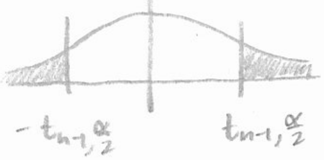
unknown

$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$

originally $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$

now $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$





Two Sample Test:

$X_1, X_2, \ldots, X_n \sim N(\mu_X, \sigma^2)$

$Y_1, Y_2, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$

$H_0: \mu_X = \mu_Y$

$H_1: \mu_X \neq \mu_Y$

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} \quad \left[\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)\right]$$

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\left(\sqrt{\frac{1}{n} + \frac{1}{m}}\right)}$$

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{m} (Y_i - \bar{Y})^2}{n+m-2}$$

Regression

obs	predictor X	response Y
1	X_1	Y_1
2	X_2	Y_2
3	$\vdots$	$\vdots$
·	X_i	Y_i
·	$\vdots$	$\vdots$
n		

$Y_i = \beta X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$

$Y_i \sim N(\beta X_i, \sigma^2)$

$H_0: \beta = 0$
$H_1: \beta > 0$

$Y_i = \beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i2} + \varepsilon_i$

$H_0: \beta_2 = 0 \longrightarrow$ simple model $Y_i = \beta_0 + \beta_1 X_{i_1} + \varepsilon_i$

$H_1: \beta_2 \neq 0. \longrightarrow$ complex model $Y_i = \beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i2} + \varepsilon_i$

Hypothesis testing helps avoid overfitting

STATS105 11/21/13

Likelihood Ratio Test

H_0: simple

H_1: simple

$X_1, X_2, \ldots, X_n \sim Ber(p)$

$H_0: p = \frac{1}{2} = 0.5$

$H_1: p = 0.7$

Recall likelihood & maximum likelihood

$$L(p) = \prod_{i=1} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

MLE = max $\log(L(p))$ over $p \in [0,1]$

Reject H_0 if $\frac{L(0.7)}{L(0.5)} > c$ (cutoff)

$$\Rightarrow \frac{0.7^{\sum x_i}\, 0.3^{n-\sum x_i}}{0.5^{\sum x_i}\, 0.5^{n-\sum x_i}}$$

if $\frac{L(H_1)}{L(H_0)} > c$

$$\Rightarrow \left(\frac{0.7}{0.5}\right)^{\sum x_i}\left(\frac{0.5}{0.3}\right)^{\sum x_i}\left(\frac{0.3}{0.5}\right)^{n}$$

$$\Rightarrow \left(\frac{0.7}{0.3}\right)^{\sum x_i}\left(\frac{0.3}{0.5}\right)^{n} > c$$

If H_0 value < H_1 value

$$\hat{p} = \frac{\sum x_i}{n} > c_3$$

if H_0 value > H_1 value

$$\hat{p} = \frac{\sum x_i}{n} < c_4$$

After log → $\sum x_i > c_1$ or $\frac{\sum x_i}{n} > c_2$

Form of Decision Rule

$H_0: p = p_0$

$H_1 = p = p_1$ $(p_1 > p_0)$

Reject H_0 if $\hat{p} > c$

Type I error $= \alpha$

$$\hat{p} \overset{H_0}{\sim} N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

$$c = p_0 + z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$$

In General:

$H_0: X \sim p_0(x)$

$H_1: X \sim p_1(x)$

Rejection Rule:

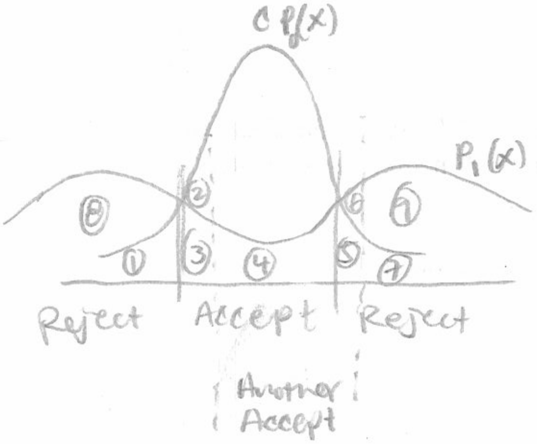
$$\frac{L(H_1)}{L(H_0)} = \frac{p_1(x)}{p_0(x)}$$

H_1 is more plausible then H_0

Two Decision Rules:

keep type I error fixed

compare type II error



Type I error

(① + ⑤ + ⑦)/c

Type II error:

③ + ④

Another Test:

I: (③ + ① + ② + ⑦)/c

II: ④ + ⑥ + ⑤

Same Type I

① + ⑤ + ⑦ = ③ + ① + ② + ⑦ ⇒ ⑤ = ③ + ②

Campare Type II:

③ + ④ Vs ④ + ⑥ + ⑤

= ④ + ⑥ + ③ + ②

bigger

Decision $\frac{p_1(x)}{p_0(x)} > 0 \Rightarrow$ reject H_0

Bayes Test:

prior: $p(H_1) = \lambda$

$P(H_0) = 1-\lambda$

posterior: $$p(H_1|x) = \frac{\lambda p_1(x)}{\lambda p_1(x) + (1-\lambda) p_0(x)}$$

$$p(H_0|x) = \frac{(1-\lambda) p_0(x)}{\lambda p_1(x) + (1-\lambda) p_0(x)}$$

STATS 105 11/12

<u>t-distribution</u>

unknown constant large "n"

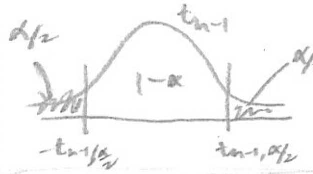
$X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$

$(1-\alpha)$CI: $[\bar{X} - t_{n-1,\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}]$

Z transformation:

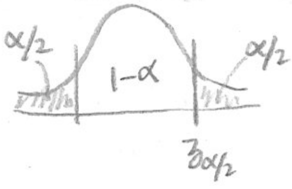
$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \xrightarrow[\infty]{n} N(0,1)$$



1-α confidence interval:

$[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$



claim: $P(\mu \in CI) = 1-\alpha$

μ ↓ fixed, CI ↓ random

$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$

$P(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}) = 1-\alpha \Rightarrow P(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

$P(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

What if σ^2 is unknown?

$S^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ $E(S^2) = \sigma^2 \Rightarrow S^2$ is unbiased $C^{\bar{Y}}$

$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum((X_i - \mu) - (\bar{X} - \mu))^2$
$\hookrightarrow Y_i$ $\hookrightarrow \bar{Y}$

$\bar{Y} = \frac{\sum Y_i}{n} = \frac{\sum(X_i - \mu)}{n} = \frac{\sum X_i - n\mu}{n} = \frac{\sum X_i}{n} - \mu$

$E[Y_i^2] = E[(X_i - \mu)^2] = \sigma^2$

$E[\bar{Y}^2] = E[(\bar{X}-\mu)^2] = Var[\bar{X}] = \frac{\sigma^2}{n}$

$\otimes\otimes$ $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i^2 - 2\bar{Y}Y_i + \bar{Y}^2)$

$= \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n}\bar{Y}Y_i + \sum_{i=1}^{n}\bar{Y}^2 = \sum Y_i^2 - 2\bar{Y}\sum Y_i + n\bar{Y}^2 = \sum Y_i^2 - 2\bar{Y} \cdot n\bar{Y} + n\bar{Y}^2 = \sum Y_i^2 - n\bar{Y}^2$

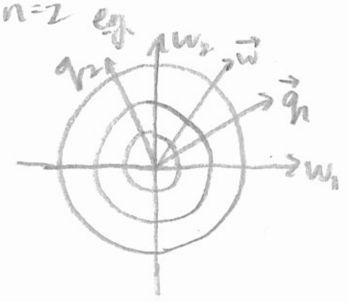
$E[\otimes\otimes] = \sum E[Y_i]^2 - nE[\bar{Y}]^2 = \sum \sigma^2 - \frac{n\sigma^2}{n} = (n-1)\sigma^2$

$\sum(X_i - \bar{X})^2 \rightarrow \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^{n} Y_i^2 - n\left(\frac{\sum Y_i}{n}\right)^2 = \sum Y_i^2 - \left(\frac{\sum Y_i}{\sqrt{n}}\right)^2$

$\frac{\sum(X_i - \bar{X})^2}{\sigma^2} = \sum(\frac{Y_i}{\sigma})^2 - \left(\sum\frac{Y_i/\sigma}{\sqrt{n}}\right)^2$, $w_i = \frac{Y_i}{\sigma} = \frac{X_i - \mu}{\sigma} \sim N(0,1)$ $= \sum w_i^2 - (\sum \frac{w_i}{\sqrt{n}})^2$

$f(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}}\exp(-\frac{w_i^2}{2})\right) = \frac{1}{(2\pi)^{n/2}}\exp(-\frac{1}{2}\sum_{i=1}^{n} w_i^2)$

$\vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \Rightarrow f(\vec{w}) = \frac{1}{(2\pi)^{n/2}}\exp(-\frac{1}{2}\|\vec{w}\|^2)$



new orthonormal system
$(q_1, q_2, \dots, q_n)$

$$\vec{w} \longrightarrow \begin{pmatrix} V_1 = \langle \vec{w}, \vec{q}_1 \rangle \\ V_2 = \langle \vec{w}, \vec{q}_2 \rangle \\ \vdots \\ V_n = \langle \vec{w}, \vec{q}_n \rangle \end{pmatrix} \Rightarrow f(\vec{V}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|\vec{V}\|^2\right)$$

$$\Downarrow$$
$$\vec{V}$$

$$V_1 = \frac{\sum_i w_i}{\sqrt{n}} = \frac{\sum_i \left(\frac{X_i - \mu}{\sigma}\right)}{\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma} \cdot n}{\sqrt{n}}$$

$$\sum w_i^2 = \sum V_i^2 \Rightarrow \sum w_i^2 - \left(\sum \frac{w_i}{\sqrt{n}}\right)^2 \quad \text{Let } \vec{q}_1 = \begin{pmatrix} \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{pmatrix} \quad = \sum_{i=1}^{n} V_i^2 - \langle \vec{w}, \vec{q}_1 \rangle^2$$

$$= \sum_{i=1}^{n} V_i^2 - \vec{V}_1^2 = \sum_{i=2}^{n} V_i^2 \sim \chi^2_{n-1}$$

Definition: $z_1, z_2, \dots, z_n \sim N(0,1)$

$\sum_{i=1}^{n} z_i^2 \sim \chi_n^2$ (chi squared)

(numerator & denominator are independent)

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/\sigma}{\frac{S/\sigma}{\sqrt{n}}} = \frac{\frac{\bar{X}-\mu}{\sigma}}{\left(\frac{\sum(X_i - \bar{X})^2}{(n-1)(\sigma^2)}\right)^{1/2} / \sqrt{n}} = \frac{V_1/\sqrt{n}}{\sqrt{\sum_{i=2}^{n} V_i^2 / (n-1) \cdot n}} = \frac{V_1}{\sqrt{\sum_{i=2}^{n} V_i^2/(n-1)}} \sim \frac{N(0,1)}{\sqrt{\boxed{\chi^2_{n-1}/(n-1)}}}$$

$\nsim 0$

Two Sample problem:

$X_1, X_2, \dots, X_n \sim N(\mu_X, \sigma^2)$

$Y_1, Y_2, \dots, Y_n \sim N(\mu_Y, \sigma^2)$

$\Delta = \mu_X - \mu_Y$

$\bar{X} \sim N(\mu_X, \frac{\sigma^2}{n})$, $\bar{Y} \sim N(\mu_Y, \frac{\sigma^2}{m})$

$$\hat{\Delta} = \bar{X} - \bar{Y} \sim N\Big(\underset{\Delta}{\underset{\parallel}{\mu_X - \mu_Y}}, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\Big)$$

$$\sim N\left(\Delta, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$$

$1-\alpha$ CI for Δ : $\hat{\Delta} \pm z_{\alpha/2} \cdot \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

$1-\alpha$ CI for Δ : $\hat{\Delta} \pm t_{n+m-2, \frac{\alpha}{2}} S\sqrt{\frac{1}{n} + \frac{1}{m}}$

$$S^2 = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{n+m-2}$$

STATS 105 11/14

Hypothesis Testing

Want to test whether a coin is fair

Flip n times Observed 55 heads

Null hypothesis: $H_0 : p = \frac{1}{2}$ (p is probability of head)

Alternative Hypothesis: $H_1 : p \neq \frac{1}{2}$

Collect Data: $X_1, X_2, \dots, X_n \sim Ber(p)$ $\hat{p} = \frac{\sum X_i}{n} \Rightarrow \hat{p}_{observed} = \frac{55}{100} = 0.55$

$\sim N\left(p, \frac{p(1-p)}{n}\right)$

Null Distribution, reference distribution

$\hat{p} \overset{H_0}{\sim} N\left(p = \frac{1}{2}, \frac{\frac{1}{2}(1-\frac{1}{2})}{100}\right) \sim N\left(\frac{1}{2}, \frac{1}{400}\right) \sim N\left(\frac{1}{2}, (0.05)^2\right)$

$z(H_0): z = \frac{\hat{p} - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{100}}} = \frac{\hat{p} - 0.5}{0.05} \overset{H_0}{\sim} N(0,1)$

$z_{obs} = \frac{0.55 - 0.5}{0.05} = 1$ compare z_{obs} w/ $N(0,1)$

Type 1 Error: H_0 is true, but reject it.

Prob(type 1 error) $= \alpha$ (level of significance)

Type 2 Error: accept H_0 when H_0 is false
H_1 is true

Prob(type 2 error) power of the test

Election Example:

survey n people

$x_i = \begin{cases} 1 & \text{vote for} \\ 0 & \text{not} \end{cases}$

$X_1, X_2, \dots, X_i \sim Ber(p)$

$H_0 : p = \frac{1}{2} = p_0$

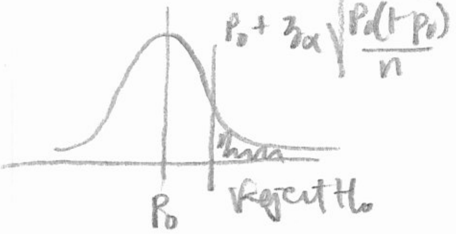
$H_1 : p > \frac{1}{2} = p_0$ (one sided H_1)

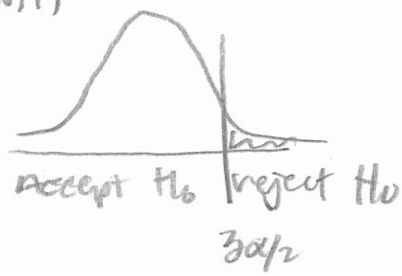
($p \neq \frac{1}{2}$ two sided)

$\hat{p} \sim N\left(\frac{1}{2}, \frac{\frac{1}{2}(1-\frac{1}{2})}{n}\right)$

$\sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$

$z(H_0): z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim} N(0,1)$



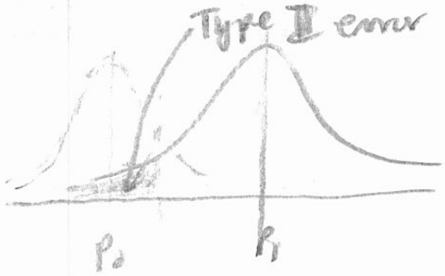


Type II error Accept H_0 when H_1 is true.

$P(\text{type II error}) = ?$

$H_0: p = p_0$ (0.5) $\hat{p} \overset{H_0}{\sim} N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$

$H_1: p = p_1$ (0.7) $\hat{p} \overset{H_1}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n}\right)$



STATS 105 11/26

1 sided, hand written, half sheet (No Project)

Recall Bayes Rule

population: $p(\underset{z=1}{male}) = \lambda$ $[X|Z=1] \sim f_1(x)$

$p(\underset{z=0}{female}) = 1-\lambda$ $[X|Z=0] \sim f_0(x)$

$P(1|x)$

$$p(z=1|X=x) = \frac{\lambda f_1(x)}{\lambda f_1(x)+(1-\lambda)f_0(x)} = \frac{p(z=1)f(x|z=1)}{p(z=1)f(X|z=1)+p(z=0)f(X|z=0)} = \frac{p(1,x)}{p(1,X)+p(0,x)} = \frac{p(1,x)}{p(x)} =$$

Count the Population:

M = total # of people

$\#$ of males $= M\lambda$ $\longrightarrow$ $\#$ of males in $(x, x+\Delta x) = M\lambda \overbrace{f_1(x)}^{density} \Delta x$

females $= M(1-\lambda)$ $\#$ of females in $(x, x+\Delta x) = M(1-\lambda) f_0(x) \Delta x$

$\#$ of people in $(x, x+\Delta x) = M(\lambda f_1(x) + (1-\lambda) f_0(x))\Delta x$

$= M f(x) \Delta x$

↳ density of entire population.

Among the slice of population in $(x, x+\Delta x)$, what is the proportion of males?

$$\frac{\cancel{M}\lambda f_1(x)\cancel{\Delta x}}{\cancel{M}(\lambda f_1(x)+(1-\lambda)f_0(x))\cancel{\Delta x}} = P(male|X \in (x, x+\underset{\substack{\downarrow\\0}}{\Delta x}))$$

$X = x$

Bayes Test

$H_0: X \sim P_0(x)$

$H_1: X \sim P_1(x)$

Prior prob: $P(H_1) = \lambda$

evidence: $[X|H_0] \sim P_0(x)$

$[X|H_1] \sim P_1(x)$

Posterior: $P(H_1|X=x) = \dfrac{\lambda P_1(x)}{\lambda P_1(x) + (1-\lambda)P_0(x)}$

Bayes Decision Rule:

Decision \ Truth	H_0	H_1
Accept H_0	0	Loss II
Reject H_0	Loss I	0

Risk: expected loss

① Risk of accepting $H_0 = P(H_0|X)\cdot 0 + P(H_1|X)\cdot$ Loss II

② Risk of rejecting $H_1 = P(H_0|X)\cdot$ Loss I $+ P(H_1|X)\cdot 0$

Reject H_0 if ② < ① or ① > ② or $\frac{①}{②} > 1$

$$\frac{\dfrac{\lambda P_1(x)}{\lambda P_1(x)+(1-\lambda)P_0(x)}\cdot \text{Loss II}}{\dfrac{(1-\lambda)P_0(x)}{\lambda(P_1(x))+(1-\lambda)P_0(x)}\cdot \text{Loss I}} > 1 \Rightarrow \overset{\substack{\text{likelihood}\\ \text{ratio}\\ \downarrow}}{\boxed{\frac{P_1(x)}{P_0(x)}}} > \frac{(1-\lambda)\,\text{Loss I}}{\lambda\,\text{Loss II}}$$

Reject H_0

$$\frac{P_1(x)}{P_0(x)} = \frac{\text{likelihood }(H_1)}{\text{likelihood }(H_0)} > \frac{\text{Prior }Pr(H_0)}{\text{Prior }Pr(H_1)} \cdot \frac{\text{loss of Type I}}{\text{loss of Type II}}$$

Generalize Likelihood Ratio

test whether die is fair.

$H_0: p_1 = \frac{1}{6}, p_2 = \frac{1}{6}, \ldots, p_6 = \frac{1}{6}$

$H_1: (p_1 \ p_2 \cdots p_6) \neq (\frac{1}{6} \ \frac{1}{6} \cdots \frac{1}{6})$

$\text{Likelihood}(p_1 \ p_2 \cdots p_6) = p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6}$

Suppose: $H_0: (p_1 \ p_2 \ p_3 \cdots p_6) = (\frac{1}{6} \ \frac{1}{6} \cdots \frac{1}{6})$

$H_1: (p_1 \ p_2 \cdots p_6) = (0.1, 0.1, 0.2, 0.2, 0.3, 0.1)$

$n_1 = \#$ of 1s
$n_2 = \#$ of 2s
$\vdots$
$+ n_6 = \#$ of 6s
$n =$ total $\#$ of repetitions

$$\text{likelihood}: \frac{L(H_1)}{L(H_0)} = \frac{0.1^{n_1+n_2+n_6} \, 0.2^{n_3+n_4} \, 0.3^{n_5}}{(\frac{1}{6})^{n_1+n_2+\cdots+n_6}}$$

What if $H_1: (p_1 \ p_2 \cdots p_6) \neq (\frac{1}{6} \ \frac{1}{6} \cdots \frac{1}{6})$

$$\frac{L(H_1)}{L(H_0)} = \frac{\hat{p}_1^{n_1} \hat{p}_2^{n_2} \cdots \hat{p}_6^{n_6}}{(\frac{1}{6})^{n_1+n_2+\cdots+n_6}} \qquad \hat{p}_1 = \frac{n_1}{n}, \ \hat{p}_2 = \frac{n_2}{n}, \ldots, \hat{p}_6 = \frac{n_6}{n}$$

$$2\log \frac{L(H_1)}{L(H_2)} \sim \chi^2_5 \qquad (\hat{p}_1 \ \hat{p}_2 \cdots \hat{p}_6) = (\frac{n_1}{n} \ \frac{n_2}{n} \cdots \frac{n_6}{n}) \quad \text{MLE}$$

$L(p_1 \ p_2 \cdots p_6) = p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6}$

$\ell(p_1 \cdots p_6) = n_1 \log p_1 + \cdots + n_6 \log p_6 = n_1 \log p_1 + \cdots + n_5 \log p_5 + n_6 \log(1-(p_1+\cdots+p_5))$

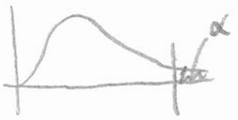
$$\left.\begin{aligned} \frac{\partial \ell}{\partial p_1} &= \frac{n_1}{p_1} - \frac{n_6}{\underbrace{1-(p_1+\cdots+p_5)}_{} \in p_6} = 0 \\ \frac{\partial \ell}{\partial p_2} &= \frac{n_2}{p_2} - \frac{n_6}{p_6} = 0 \end{aligned}\right\} \quad \frac{n_1}{p_1} = \frac{n_2}{p_2} = \cdots = \frac{n_6}{p_6} = C$$

$$\begin{cases} \hat{p}_1 = \frac{n_1}{C} \\ \hat{p}_2 = \frac{n_2}{C} \\ \vdots \\ \hat{p}_6 = \frac{n_6}{C} \end{cases} \qquad \begin{aligned} &\hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_6 = 1 \\ &\frac{n_1}{C} + \frac{n_2}{C} + \cdots + \frac{n_6}{C} = 1 \\ &C = n \end{aligned}$$

$$\frac{\hat{p}_1^{n_1} \hat{p}_2^{n_2} \cdots \hat{p}_6^{n_6}}{(\frac{1}{6})^{n_1+\cdots+n_6}} = \frac{\max_{H_1} p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6}}{\max_{H_0} p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6}}$$

$$\frac{\text{max out unknown in } H_1}{\text{max out unknown in } H_0}$$

2. log like ratio $\overset{H_0}{\sim} \chi^2$ # of free parameters $\underset{5}{\text{in } H_1}$ − # of free $\underset{0}{\text{parameters}}$ in H_0



overfitting

high cuttoff value to avoid overfitting

STATS 105

12/03

Final: Similar to homework/lecture (6 Questions)

Regression

obs	Predictor	Response
1	Y_1	Y_1
2	X_2	Y_2
	$\vdots$	$\vdots$
n	X_n	Y_n
	training	

Simplest Regression

$Y_i \approx \beta X_i$

more precisely

$Y_i \overset{iid}{\sim} N(\beta X_i, \sigma^2)$ — ①

$\updownarrow$

$Y_i = \beta X_i + \varepsilon_i \quad , \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ — ②

Maximum Likelihood: assume x_i are fixed (given) & y_i random

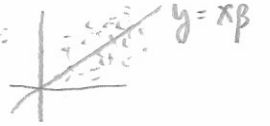
$$L(\beta) = \prod_{i=1}^{n} f(y_i | x_i, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right)$$

maximum likelihood $\Longrightarrow$ Least squares

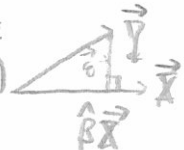
Find $\hat{\beta}$ by minimizing $\sum_{i=1}^{n} \underbrace{(y_i - x_i\beta)}_{\text{Residual}}^2 = R(\beta)$

$$R'(\beta) = \sum_{i=1}^{n} 2(y_i - x_i\beta)(-x_i) = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Scatter plot:
(2-Dim)



Vector:
(N-Dim)



$R(\beta) = |\vec{Y} - \beta\vec{X}|^2$

$\langle \vec{X}, \vec{Y} - \beta\vec{X} \rangle = 0 \Rightarrow \sum_{i=1}^{n} x_i(y_i - \beta x_i) = 0$

Under hypothetical Repetition, y_i random, β random (different from repetition to repetition)

$$\beta = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta_{TRUE} x_i + \varepsilon_i)}{\sum x_i^2} = \frac{\sum x_i^2 \beta_{TRUE} + \sum x_i \varepsilon_i}{\sum x_i^2} = \underset{\substack{\downarrow \\ \text{signal}}}{\beta_{TRUE}} + \underset{\substack{\downarrow \\ \text{noise}}}{\frac{\sum x_i \varepsilon_i}{\sum x_i^2}}$$

$$\left.\begin{aligned} E[\hat{\beta}] &= \beta_{TRUE} + \frac{\sum x_i E[\varepsilon_i]}{\sum x_i^2} = \beta_{TRUE} \text{ (unbiased)} \\ Var[\hat{\beta}] &= \frac{\sum x_i^2 Var(\varepsilon_i)}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2} \end{aligned}\right\} \hat{\beta} \sim N\left(\beta_{TRUE}, \frac{\sigma^2}{\sum x_i^2}\right)$$

Sampling Distribution

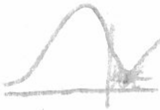
95% C.I.: $\beta \pm \underset{(=1.96)}{z_{\alpha/2}} \left(\frac{\sigma}{\sqrt{\sum x_i^2}}\right)$ (σ^2 known)

T-distribution (σ^2 unknown) by s^2.

double sided

Testing: $H_0: \beta_{TRUE} = 0$, $H_1: \beta_{TRUE} \neq 0$.

Under H_0: $\hat{\beta} \sim N\left(0, \frac{\sigma^2}{\sum x_i^2}\right) \Rightarrow Z = \frac{\hat{\beta}}{\left(\frac{\sigma^2}{\sqrt{\sum x_i^2}}\right)} \sim N(0,1)$



$(H_1: \beta > 0)$

decision rule: reject $Z > Z_\alpha$

Type I: α

Only if $\hat{\beta}$ (or Z) large enough we believe $\beta_{TRUE} \neq 0$.

Avoid overfitting.

Simple Regression:

$y_i \approx \alpha + \beta x_i$ $\qquad y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$y_i \overset{iid}{\sim} N(\alpha_{TRUE} + \beta_{TRUE} x_i, \sigma^2)$

Least Square: $R(\alpha, \beta) = \sum (y_i - (\alpha + x_i \beta))^2$

pretend β is known: $\hat{\alpha} \leftarrow \min \sum (\underset{\text{new } y_i}{\underset{\downarrow}{(y_i - \beta x_i)}} - \alpha_i \cdot \underset{\text{new } x_i}{\underset{\downarrow}{1}})^2 \Rightarrow \hat{\alpha} = \frac{\sum (y_i - \beta x_i) \times 1}{\sum 1^2 \ (= n)}$

$\hat{\alpha} = \bar{Y} - \bar{X}\beta$

$R(\alpha, \beta) = R(\beta) = \sum (y_i - ((\bar{Y} - \bar{X}\beta) + \beta x_i))$

$= \sum ((y_i - \bar{Y}) - \beta(x_i - \bar{X}))^2 = \sum (\tilde{y}_i - \beta \tilde{x}_i)^2$

$\hat{\beta} = \frac{\sum \tilde{x}_i \tilde{y}_i}{\sum \tilde{x}_i^2} = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sum (x_i - \bar{X})^2} = \frac{\frac{1}{n}\sum (x_i - \bar{X})(y_i - \bar{Y})}{\frac{1}{n}\sum (x_i - \bar{X})^2} = \frac{\text{Sample Covariance}}{\text{Sample Var of } X}$

$\hat{\alpha} = \bar{Y} - \bar{X}\hat{\beta} \Rightarrow \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$ $(\bar{X}, \bar{Y})$ on regression line

$\sum (x_i - \bar{X}) = \sum x_i - n\bar{X} = 0$

$\hat{\beta} = \frac{\sum (x_i - \bar{X}) y_i}{\sum (x_i - \bar{X}) x_i} = \frac{\sum \tilde{x}_i (\alpha_T + \beta_T x_i + \varepsilon_i)}{\sum \tilde{x}_i \cdot x_i} = \frac{\cancelto{0}{\alpha_T \sum \tilde{x}_i} + \beta_T \sum \tilde{x}_i x_i + \sum \tilde{x}_i \varepsilon_i}{\sum \tilde{x}_i x_i} = \beta_T + \frac{\sum \tilde{x}_i \varepsilon_i}{\sum \tilde{x}_i^2}$

$E[\hat{\beta}] = \beta_{TRUE}$ & $Var[\hat{\beta}] = \frac{\sigma^2}{\sum \tilde{x}_i^2} \Rightarrow \hat{\beta} \sim \left(\beta_{TRUE}, \frac{\sigma^2}{\sum \tilde{x}_i^2}\right)$ Sample distribution

Multiple Regression:

$$y_i = \beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\text{weight} = \underset{\beta_1 \quad 1}{\text{intercept}} + \beta_2 \underset{x_{i_2}}{\text{height}} + \beta_3 \underset{x_{i_3}}{\text{height}^2} + \beta_4 \underset{x_{i_4}}{\text{gender}} + \cdots + \text{error}$$

useful for predictions

linear in β's. can be nonlinear in variables

$$R(\beta) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 = \sum_{i=1}^{n} \left(y_i - \vec{x}_i^T \vec{\beta}\right)^2$$

$$\vec{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad \vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\frac{\partial R(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} 2\left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)(-x_{ij}) \Rightarrow \frac{\partial R}{\partial \beta} = \begin{pmatrix} \frac{\partial R}{\partial \beta_1} \\ \frac{\partial R}{\partial \beta_1} \\ \vdots \\ \frac{\partial R}{\partial \beta_p} \end{pmatrix} = -2\sum^{n}\left(y_i - \sum^{p} \beta x_{ij}\right)\begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$\xleftarrow{\quad} = -2\sum\left(y_i - \vec{x}_i^T \vec{\beta}\right)\vec{x}_i = 0$$

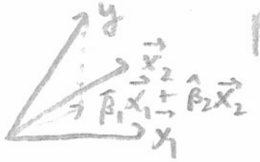
$$\sum^{n} \vec{x}_i\left(y_i - \vec{x}_i^T \vec{\beta}\right) = 0$$

$$= \sum \vec{x}_i y_i - \sum \vec{x}_i \vec{x}_i^T \beta = 0$$

$$\sum \vec{x}_i \vec{x}_i^T \beta = \sum \vec{x}_i y_i \Rightarrow \underset{p\times 1}{\hat{\beta}} = \underset{p \times p}{\left(\sum \vec{x}_i \vec{x}_i^T\right)^{-1}} \underset{p \times 1}{\sum \vec{x}_i y_i}$$

Simplest Regression:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$



$p = 2$

$$\langle x_j, y_i - \sum^{p} x_i \beta_j \rangle = 0$$

STATS 105 12/05

Regression

obs	predictors 1	2	3 ⋯	p	Response
1	X_{11}	X_{12}	⋯	X_{1p}	y_1
2	⋮	⋮	⋱	⋮	y_2
i	X_{i1}	X_{i2}	⋯	X_{ip}	⋮
n	X_{n1}	X_{n2}	⋯	X_{np}	y_n

Linear Regression

$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$

linear in β's.

May not be linear in original variables

Least Square Method:

find $(\beta_1, \beta_2, \ldots, \beta_p)$ by minimizing $\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \Rightarrow (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)$

Vector Notation

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad X_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$y_i = x_i^T \beta + \varepsilon_i$$

$$R(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

$$\hat{\beta} = (\sum_{i=1}^{n} x_i x_i^T)^{-1} \sum_{i=1}^{n} (x_i y_i)$$

Statisticians: understanding observed data

$H_0: \beta_j = 0$ & $H_1: \beta_j > 0$ (to avoid overfitting)

Machine Learners: predicting testing data

Regularization (to avoid overfitting)

$$\min (\sum_{i=1}^{n} y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \text{penalty}(\beta_1 \cdots \beta_p)$$

$$\text{penalty}(\beta_1 \cdots \beta_p) = \sum_{j=1}^{p} \beta_j^2$$

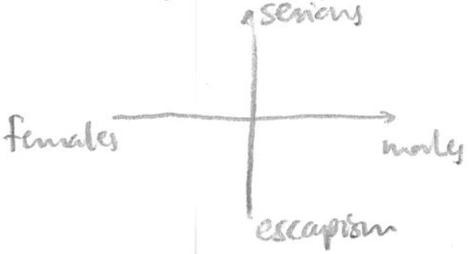
$$= \sum_{j=1}^{p} |\beta_j|$$ Lasso (Least Absolute shrinkage & selection sparsity operation)

Recommender System

Netflix

users \ items	1	2 ⋯	i	⋯ p
1				
2				
u			r_{ui}	
⋮				
n				

Aspects/factors/Dimensions:



item i: $a_i = \begin{pmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ik} \end{pmatrix}$ a_{ik} = appeal of item i in the k^{th} aspect

user u: $p_u = \begin{pmatrix} p_{u1} \\ p_{u2} \\ \vdots \\ p_{uk} \end{pmatrix}$ p_{uk}: preference of user u in k^{th} aspect

$$r_{ui} = \sum_{k=1}^{k} p_{uk} \cdot a_{ik} + \varepsilon_{ui}$$

$$r_{ui} = \mu + b_u + b_i + \sum_{k=1}^{k} p_{uk} a_{ik} + \varepsilon_{ui}$$

$$\min \sum_{\substack{\text{obs} \\ \text{ratings}}} \left(r_{ui} - (\mu + b_u + b_i + \vec{p}_u^{\,T} \vec{a}_i)\right)^2 + \lambda_1 |\vec{p}_u|^2 + \lambda_2 |\vec{a}_i|^2$$

$$\Longrightarrow \hat{\mu}, \hat{b}_u, \hat{b}_i, \hat{p}_u, \hat{a}_i$$

Simple Version:

$\sum \left(r_{ui} - \vec{p}_u^{\,T} \vec{a}_i\right)^2$ Alternating: Given $\vec{p}_u$, find $\vec{a}_i$

Given $\vec{a}_i$, find $\vec{p}_u$

Logistic Regression

$y_i \in \{0, +1\}$ - statistics

$\{-1, +1\}$ - machine Learning

Assumption: $y_i \sim Ber(p_i)$; $\log \frac{p_i}{1-p_i} = \sum_{j=1}^{p} \beta_j x_j = x^T \beta \Longleftrightarrow p_i = \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}$

Maximum likelihood:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}\right)^{y_i} \left(\frac{1}{1+\exp(x_i^T \beta)}\right)^{1-y_i}$$

$$= \prod_{i=1}^{n} \frac{\exp(y_i x_i^T \beta)}{1+\exp(x_i^T \beta)} \Rightarrow \ell(\beta) = \sum_{i=1}^{n} y_i x_i^T \beta - \log\left(1+\exp(x_i^T \beta)\right)$$

$$\ell'(\beta) = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{pmatrix} = \sum y_i x_i - \overbrace{\frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}}^{\rightarrow p_i} \cdot x_i = \sum_{i=1}^{n} x_i (y_i - p_i)$$

Learning Algorithm:

gradient ascent

start from $\beta^{(t)}$

iterate: $$\beta^{(t+1)} = \beta^{(t)} + \overset{\text{learning rate}}{\gamma} \sum_{i=1}^{n} x_i \left(y_i - \overset{\text{previous page}}{P_i^{(t)}}\right)$$

- adaboost (form of logistic regression)

  X_i = weak classifiers

  $\Downarrow$ boost

  Strong classifiers

- support vector machine (maximum margin principle)

  want a large Margin (M)