STAT 241: Statistics (Advanced Level)

Typeset by Kevin Matthews
Some diagrams by Anonymous
Updated March 30, 2016

Chapter 1 - Introduction

Population: the collection of potential study objects. 2016 01 04

process: a mechanism by which the data are generated.

EXAMPLE **1.** The underlying probability distribution from which the sample data are generated.

sample:

- Usually a small proportion of the population
- Needs to be representative

Variate (Variable): characteristics measured on the subjects.

- continuous
- discrete
- categorical $\begin{cases}\text{nominal}\\ \text{ordinal}\end{cases}$

Attributes

Data collection:

(1) sample surveys

(2) observational study

(3) Experimental design

Data:

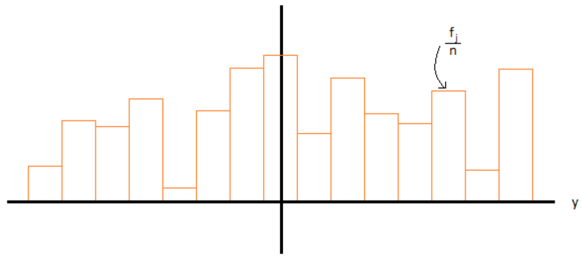
subject	gender	age	weight	$\cdots$	Y
1	M	x	x		y_1
2	F	x	$\vdots$		y_2
3	F	x	$\vdots$		y_3
$\vdots$	$\vdots$	$\vdots$			$\vdots$
n	M	x	x		y_n

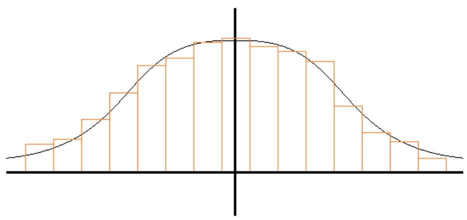
Histogram.

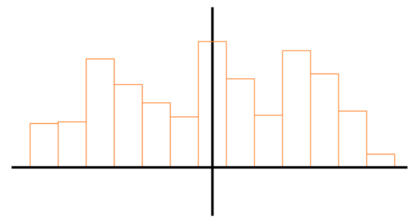
(1) Partition the range of y into k non-overlapping (equal-length) intervals $I_j = [a_{j-1}, a_j)$, $j = 1, \ldots, k$.

(2) Calculate $f_j = \#$ of y_j's there are in I_j, $j = 1, \dots, k$.

"relative frequency" histogram

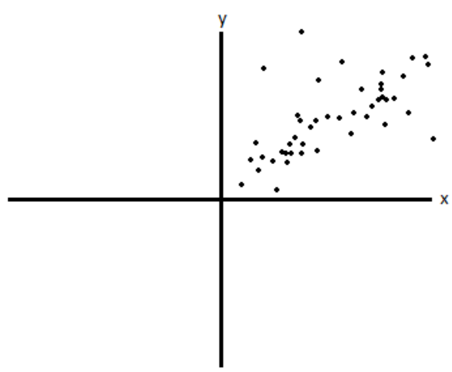






Scatterplot

$$\begin{array}{cc} X & Y \\ x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{array}$$



Numerical Summaries of the data:

(1) Measure of location $\qquad \{y_1, \dots, y_n\} \qquad y_{(1)} \le y_{(2)} \le \dots \le y_{(n)}$

(i) mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

(ii) median $\begin{cases} \text{if } n \text{ is odd, then median } = y_{\left(\frac{n+1}{2}\right)} \\ \text{if } n \text{ is even, then median } = \frac{1}{2}\left(y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}\right) \end{cases}$

median is more robust against outliers compared to mean.

(iii) mode

(2) measure of variability

(i) sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$s = \sqrt{s^2}$

(ii) <u>range</u>: $y_{(n)} - y_{(1)} = \max - \min$

(3) measure of skewness or shape

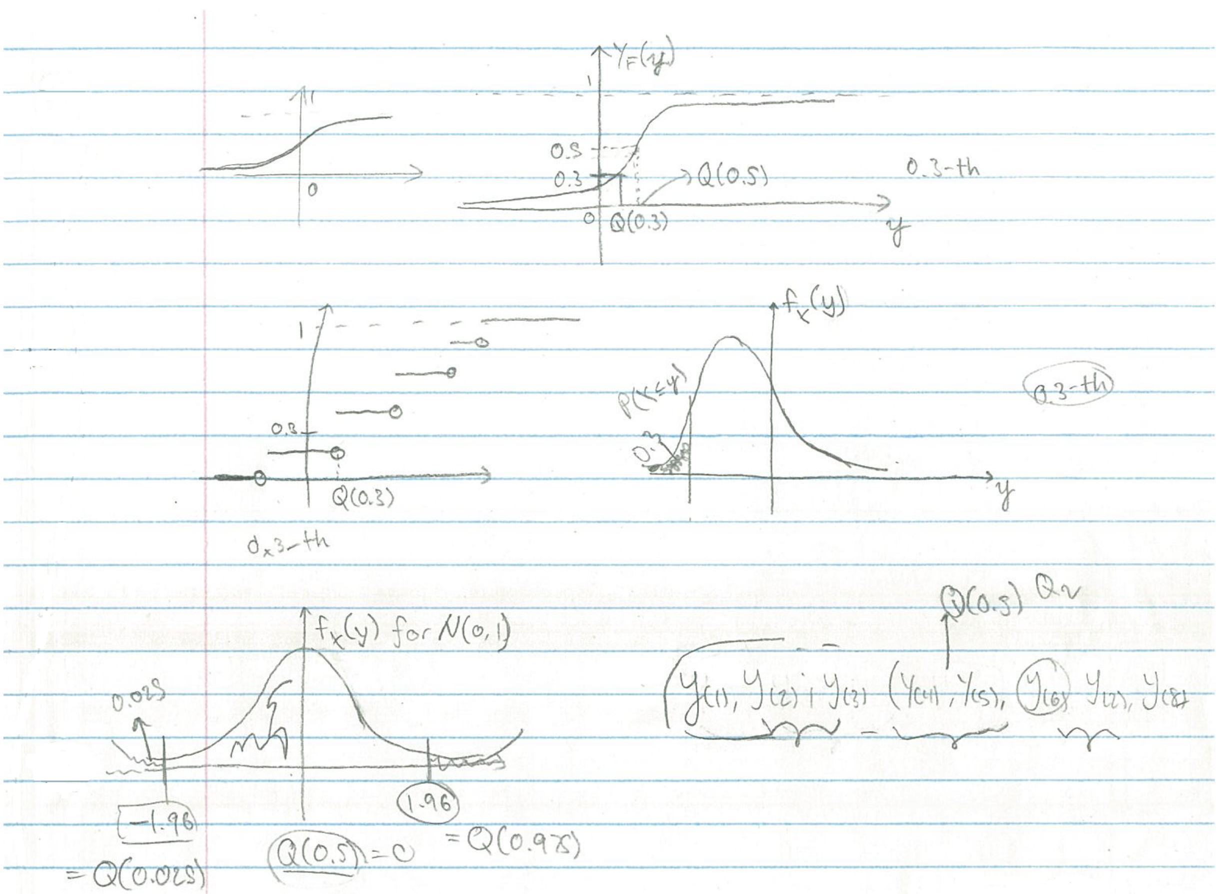
Numerical Summaries: 2016 01 06

1) measure of location: mean, median, mode

2) measure of variability: sample variance s^2, range, inter-quartile range (IQR)

3) measure of skewness and shape:

<u>IQR</u>: <u>Def of quantiles</u>: Let Y be a random variable with CDF $F_Y(y) = P(Y \le y)$. the p-th quantile of Y is $Q_Y(p) \equiv F_Y^{-1}(p) \equiv \inf \underbrace{\{y;\, F_Y(y) \ge p\}}_{Q_Y(p)=y \text{ s.t. } F_Y(y)=p}$ where $p \in [0,1]$.



some special quantiles:

lower quartile $Q(0.25)$

medain $Q(0.5)$

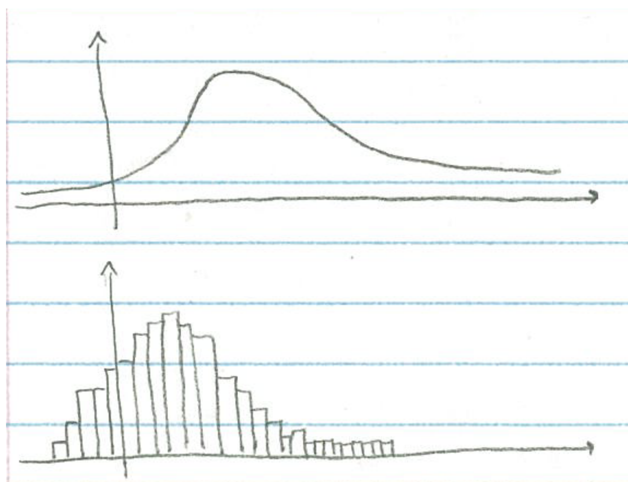
upper quartile $Q(0.75)$

IQR= $Q(0.75) - Q(0.25)$

3) measure of skewness and shape:

   this measure indicates how the distribution of the data differs from a Normal distribution

   i) skewness: measures the asymmetry of the data



sample skewness $= \dfrac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{3/2}}$

$\rightarrow$ skewed to the right $\implies$ sample skewness > 0

   ii) sample kurtosis $= \dfrac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{2}}$

   the sample kurtosis for Normal distribution ≈ 3

   Data that are very peaked have a sample kurtosis > 3

the five-number summary: $y_{(1)}, Q_1, Q_2, Q_3, y_{(n)}$

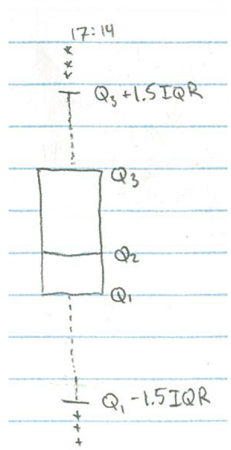
Sample Correlation: for $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, $\rho = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ where $S_{yy} = \sum_{i=1}^n (x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$

$-1 \le \rho \le 1$

ρ measures the linear relationship between x and y

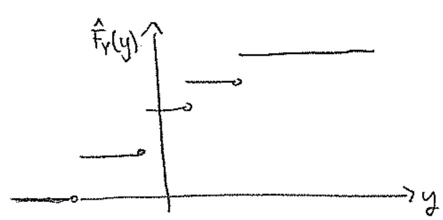
$$\begin{cases} \text{when } \rho \text{ is close to 1, } x \text{ and } y \text{ have a strong positive linear relationship} \\ \text{when } \rho \text{ is close to } -1\text{, } x \text{ and } y \text{ have a strong negative linear relationship} \end{cases}$$

Boxplot: $y_{(1)}, Q_1, Q_2, Q_3, y_{(n)}$



1) Draw a box with ends at the lower and upper quartiles
2) Add a line at the median
3) Draw two lines outside the box to $y_{(1)}$ and $y_{(n)}$. If $y_{(1)}$ or $y_{(n)}$ is more than 1.5IQR then add lines at the most extreme data points within $Q_1 - 1.5\text{IQR}$ and $Q_3 + 1.5\text{IQR}$
4) Plot any additional points beyond ± 1.5IQR using "+" or "$\star$"

ECDF $\hat{F}_Y(y) = \dfrac{\#\text{ of values in } \{y_1,\ldots,y_n\} \text{ that are} \le y}{n}$ 2016 01 11



Ch 2. Maximum Likelihood Estimation R demo

$Y \sim \text{Binomial}(n,p)$

k possible outcomes: $1,\ldots,k$

probability of the i-th outcome: $\theta_1, \ldots, \theta_k$

now we have n independent trials

let Y_i be the # of the i-th outcome for these n trials

$(Y_1, \ldots, Y_k) \sim \text{Multinomial}\left(n, \vec{\theta}\right)$

$P\left(Y_1 = y_1, \ldots, Y_k =_k; \vec{\theta}\right) = \dfrac{n!}{y_1! \cdots y_k!} \theta_1^{y_1} \cdots \theta_k^{y_k}$ where $\sum_{i=1}^{k} y_i = n$ $\sum_{i=1}^{k} \theta_i = 1$

$Y \sim \text{Binomial}\,(n, p)$

$P\,(Y = y;\, p) = \binom{n}{y} p^y \,(1-p)^{n-y}$

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

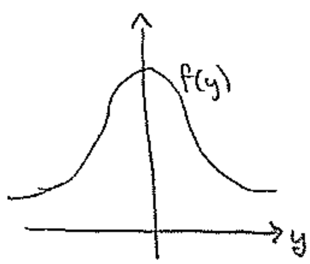
Def: (Estimator/Estimate)

Let $\vec{Y}$ be the data vector (random vector) and $\vec{y}$ the observed value of $\vec{Y}$

An estimator of a parameter θ is a function of $\vec{Y}$ and possibly other known quantities such as n

An estimate of θ is the value of an estimator evaluated at the data $\vec{y}$

Def: the likelihood function for θ is $L(\theta) = L\,(\theta;\, \vec{y}) = \underline{f\,(\vec{y};\, \theta)}, \theta \in \Omega$



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

$f(y; \theta)$

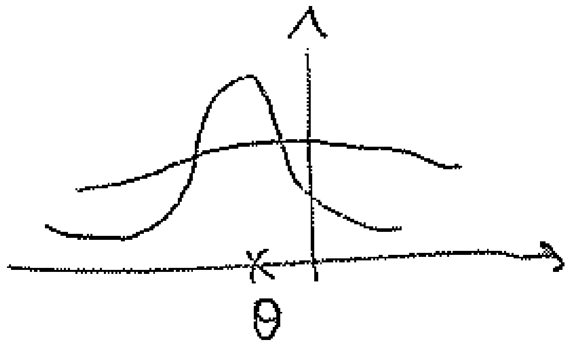
Def: the value of θ that maximizes $L\,(\theta)$ for given data $\vec{Y}$ is called the MLE of θ, denoted by $\hat{\theta}$

Back to $Y \sim \text{Binomial}\,(n, p)$, $\rightarrow \hat{p} = \frac{y}{n}$, $\hat{p} = \frac{Y}{n}$

Likelihood $L(\theta) \equiv L\,(\theta, \vec{y}) = f\,(\vec{y};\, \theta)$ MLE $\hat{\theta} \equiv \arg\max_{\theta} L\,(\theta)$ 2016 01 13

$\binom{n}{y} \theta^y \,(1-\theta)^{n-y}$ $e^{[]}$

Def: $\ell\,(\theta) = \log L\,(\theta)$: log-likelihood. $\hat{\theta}$ is usually derived by solving $\underbrace{\dfrac{d\ell\,(\theta)}{d\theta} = 0}_{\text{score equation}}$



often $\vec{Y} = (Y_1, \ldots, Y_n)$ where Y_i<u>'s are a random sample from</u> the population

Y_i's are independent

often we assume Y_i's have the same distribution

$\Rightarrow Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f(y;\theta) \Rightarrow L(\theta) = f(\vec{y};\theta) = \prod_{i=1}^n f(y_i;\theta) = \prod_{i=1}^n L(\theta; y_i)$

<u>Ex1</u> MLE for <u>Exponential</u> Let Y denote the lifetime of a randomly selected light bulb. $Y \sim \text{Exp}(\theta) \implies 4(y,\theta) = \frac{1}{\theta} e^{-\frac{y}{\theta}} \quad \theta > 0$

a random sample $Y_1, \ldots, Y_n \implies L(\theta) = \prod_{i=1}^n f(y_i;\theta) = \dfrac{1}{\theta^n} e^{-\frac{\sum_{i=1}^n Y_i}{\theta}} \quad \theta > 0$

$$\implies \ell(\theta) \log L(\theta) = -n\log\theta - \frac{1}{\theta}\sum_{i=1}^n Y_i$$

$$\implies \frac{d\ell(\theta)}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n Y_i = 0 \implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i = F$$

To check that $\hat{\theta}$ is the MLE $\qquad \left.\dfrac{d^2\ell(\theta)}{d\theta^2}\right|_{\hat{\theta}} < 0$

<u>Ex2:</u> $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \qquad f(y;\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \qquad -\infty < \mu < \infty \quad \sigma > 0$

$$L(\vec{\theta}) = \prod_{i=1}^n f(y_i, \vec{\theta}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} e^{-\frac{\sum_{i=1}^n (Y_i-\mu)^2}{2\sigma^2}}$$

$$\implies \ell(\vec{\theta}) = \log L(\hat{\theta}) = c - n\log\sigma - \frac{\sum_{i=1}^n (Y_i-\mu)^2}{2\sigma^2} \implies \frac{\partial \ell(\hat{\theta})}{\partial\vec{\theta}} = \begin{cases} \frac{2\sum_{i=1}^n (Y_i-\mu)}{2\sigma^2} = 0 \\ -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (Y_i-\mu)^2}{\sigma^3} = 0 \end{cases}$$

$$\implies \begin{pmatrix} \hat{\mu} = \bar{Y} \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2 \end{pmatrix} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2 \qquad E[s^2] = \sigma^2$$

<u>Ex3:</u> MLE for Multinomial

$\vec{Y} = (Y_1, \ldots, Y_k) \sim$ Multinomial $\implies L(\hat{\theta}) = \frac{n!}{Y_1! \cdots Y_k!} \theta_1^{Y_1} \cdots \theta_k^{Y_k}$ where $\sum_{i=1}^k \theta_i = 1 \quad \sum_{i=1}^k Y_i = n$

$\hat{\theta}_i = \frac{Y_i}{n} \quad i = 1, \ldots, k$

<u>Ex4:</u> $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F(x;\theta_1,\theta_2) = 1 - \left(\dfrac{\theta_1}{x}\right)^{\theta_2} \qquad x \geq \theta_1, \theta_1 > 0, \theta_2 > 0$

$f(x,\theta_1,\theta_2) = \theta_1^{\theta_2}\theta_2 x^{-\theta_2-1} \quad x \geq \theta_1, \theta_1 > 0, \theta_2 > 0$

$L(\theta_1,\theta_2) = \prod_{i=1}^n f(x_i;\theta_1,\theta_2) = \theta_1^{n\theta_2}\theta_2^n \left(\prod_{i=1}^n x_i\right)^{-\theta_2-1} \quad 0 < \theta_1 \leq X_{(1)}, \theta_2 > 0$

$\hat{\theta}_1 = X_{(1)}$

$L(\theta_2) = L\left(\hat{\theta}_1,\theta_2\right) = \hat{\theta}_1^{n\theta_2}\theta_2^n \left(\prod_{i=1}^n x_i\right)^{-\theta_2-1} \quad \theta_2 > 0$

$\implies \ell(\theta_2) = \log\ L(\theta_2) = n\theta_2 \log\hat{\theta}_1 + n\log\theta_2 - (\theta_2+1)\sum_{i=1}^n \log X_i$

$\implies \dfrac{d\ell(\theta)_2}{d\theta_2} = n\log\hat{\theta}_1 + \dfrac{y}{\theta_2} - \sum_{i=1}^n \log X_i = 0 \implies \hat{\theta}_2 = \dfrac{n}{\sum_{i=1}^n \log X_i - n\log\hat{\theta}_1}$

Check that $\hat{\theta}_2$ is the MLE

Ex 5: let $\theta \equiv$ # of coliform bacteria in one ml of water:

Then for a water sample of v ml the average # of bacteria is $v\theta$

Let $Y \equiv$ actual # of bacteria in a water sample of v ml

Suppose that $Y \sim \text{Poisson}(\theta v)$

1) Suppose that we can precisely count the # of bacteria. How to estimate θ using the MLE approach?

   Randomly select n water samples with volume $v_1,\ldots,v_n$

   let Y_i dentoe the # of bacteria in sample i. Then $Y_i \sim \text{Poisson}(v_i\theta)$

   $L(\theta) = \prod_{i=1}^n f(y_i;\theta) = \prod_{i=1}^n \left[\dfrac{(\theta v_i)^{Y_i}}{Y_i!}e^{-\theta v_i}\right] = \dfrac{\prod_{i=1}^n v_i^{Y_i}}{\prod_{i=1}^n Y_i!}\theta^{\sum_{i=1}^n Y_i}e^{-\theta\sum_{i=1}^n v_i}$

   $\ell(\theta) = \log L(\theta) = c + \sum_{i=1}^n Y_i \log\theta - \theta\sum_{i=1}^n v_i \implies \dfrac{d\ell(\theta)}{d\theta} = \dfrac{\sum_{i=1}^n Y_i}{\theta} - \sum_{i=1}^n v_i = 0 \implies \hat{\theta} = \dfrac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n v_i}$

   To check that $\hat{\theta}$ is the MLE

   $\left.\dfrac{d^2\ell\theta}{d\theta^2}\right|_{\hat{\theta}} = -\dfrac{\sum_{i=1}^n Y_i}{\hat{\theta}^2} < 0$

2016 01 18

2) Suppose now that we can only detect the presence/absence of bacteria

   how to estimate θ using the ML method?

   $(Y > 0)$, $(Y = 0)$ If we define $z = \begin{cases} 1 & Y > 0 \\ 0 & Y = 0 \end{cases}$, $\quad Y \sim \text{Poisson}(\theta V)$

   $Z \sim \text{Bernoulli}(P(Z=1)) = \text{Ber}(P(Y>0)) = \text{Ber}(1-P(Y=0)) = \text{Ber}(1-e^{-\theta v})$

   $f_Y(y) = P(Y=y) = \dfrac{\theta^y}{y!}e^{-\theta}$

Now randomly take water samples with volume $v_1, \ldots, v_n$

Z_i, $i = 1, \ldots, n$, $\quad Z_i \sim \text{Ber}\left(1 - e^{-\theta v_i}\right)$

$$L(\theta) = \prod_{i=1}^{n} f(Z_i, \theta) = \prod_{i=1}^{n} \left(1 - e^{-\theta v_i}\right)^{Z_i} \left(e^{-\theta v_i}\right)^{1-Z_i}$$

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \left[Z_i \log\left(1 - e^{-\theta v_i}\right) - (1 - Z_i)\,\theta v_i\right]$$

$$\frac{d\ell}{d\theta} = \sum_{i=1}^{n} \left[\frac{v_i Z_i e^{-\theta v_i}}{1 - e^{-\theta v_i}} - (1 - Z_i)\, v_i\right] = 0$$

Newton-Raphson

Thm: (invariance property of the MLE) If $\hat{\theta}$ is the MLE of θ, then the MLE of $g(\theta)$ is $g\left(\hat{\theta}\right)$, where $g \in \mathcal{C}^0$ or g is continuous

$$\text{e.x.} \quad N\left(\mu, \sigma^2\right), \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

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

Merits of MLE:

1. bias $E\left(\hat{\theta}\right) - \theta \to 0$ as $n \to \infty \qquad E\left(\hat{\theta}\right) = \theta$
2. efficiency

problems of MLE:
not robust to model misspecification

Asymptotic Statistics

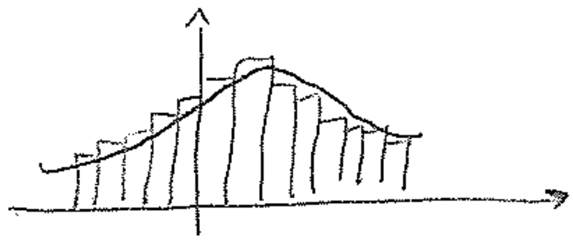
Model Checking:

- one way to check the adequacy of a model is to compare model-based probabilities to sample-based frequencies

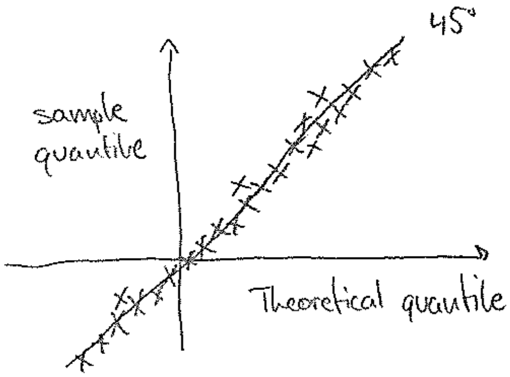
partition the range of Y into $[a_{j-1}, a_j]$, $j = 1, \ldots, J$

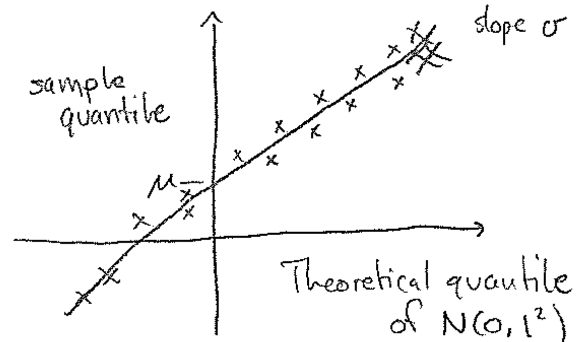
$Y \sim f(Y; \theta) \qquad \hat{\theta} \qquad P\left(a_{j-1} \le Y < a_j; \hat{\theta}\right)$

If the model is appropriate, then these probabilities should be close to the corresponding relative frequencies



<u>QQ-plots</u> $\{Y_1, \ldots, Y_n\}$ and $f\left(y, \hat{\theta}\right) = N\left(\mu, \sigma^2\right)$





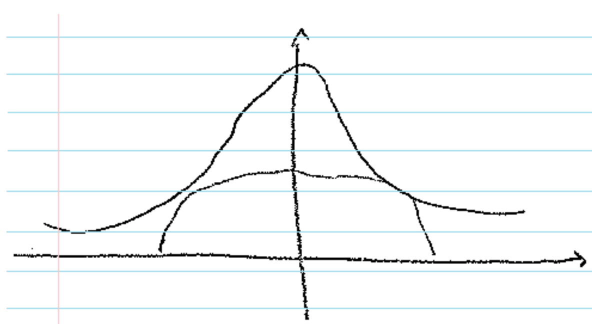
$Y \sim N\left(\mu, \sigma^2\right), \qquad Z \sim N(0,1)$

$y = Q_Y(\tau) \qquad P(Y \le y) = \tau = P\left(\frac{Y-\mu}{\sigma} \le \frac{y-\mu}{\sigma}\right) = P\left(Z \le \frac{y-\mu}{\sigma}\right)$

$\implies Q_Z(\tau) = \frac{y-\mu}{\sigma} = z$

$y = \sigma z + \mu$

R demo



Ch4 Interval Estimation

2016 01 20

- Sampling distribution of the MLE $\quad \hat{\theta} = \hat{\theta}\left(\vec{Y}; n\right)$
    - point estimator
    - uncertainty of $\hat{\theta}$
    - sampling distribution of $\hat{\theta}$
- Finding the sampling distribution of $\hat{\theta}$ is generally very difficult

- $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N\left(\mu, \sigma^2\right)$ $\quad \hat{\mu} = \bar{Y}_n \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ $\quad P\left(|\hat{\mu} - \mu| \le 0\right)$
- $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \operatorname{Exp}(\theta)$ $\quad \hat{\theta} = \bar{Y}_n \overset{\text{app}}{\sim} N\left(\mu_Y, \dfrac{\sigma_Y^2}{n}\right)$
- large sample approximation is commonly used to derive asymptotic distribution of $\hat{\theta}$

- Interval Estimator
  - a way of indicating the uncertainty of $\hat{\theta}$
  - In the form of $\left[L\left(\vec{Y}\right), U\left(\vec{Y}\right)\right]$
  - we would like $P\left(L\left(\vec{Y}\right) \le \theta \le U\left(\vec{Y}\right)\right)$ to be large (skip sec 4.3)

Def: $C(\theta) = P\left(L\left(\vec{Y}\right) \le \theta \le U\left(\vec{Y}\right)\right)$ is called the coverage prob. of the interval estimator

Note: we'd like $C(\theta)$ to be close to 1 $(0.95, 0.99)$ while keeping the length of the interval short

For a fixed $C(\theta)$ the interval estimators are called confidence intervals

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

Def: A $100p\%$ for θ is an interval estimate $\left[L(\vec{y}), Y(vecy)\right]$ such that $P\left(L\left(\vec{Y}\right) \le \theta \le U\left(\vec{Y}\right)\right) = p$
Here p is called the confidence coefficient

Ex 4.4.1 $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ μ unknown σ^2 known $\quad \bar{Y}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
Consider $\left[\bar{Y} - 1.96\dfrac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\dfrac{\sigma}{\sqrt{n}}\right]$

$$P\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = P(-1.96 \le \underbrace{\frac{\mu - \bar{Y}}{\frac{\sigma}{\sqrt{n}}}}_{\sim N(0,1)} \le 1.96) = 0.95$$

Note: 1 Suppose we observed $y_1, \ldots, y_n$ $\quad \left[\bar{Y} - 1.96\dfrac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\dfrac{\sigma}{\sqrt{n}}\right]$

$$P\left(Y - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right) \ne 0.95$$

We have 95% confidence that $\left[Y - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ covers μ

2 CI gets narrower as $n \uparrow$

Def: A function $Q_n = g\left(\vec{Y}, \theta\right)$ of the data $\vec{Y}$ and unknown θ is called a pivotal quantity if the distribution of Q_n is completely known

Suppose now we have a pivotal quantity Q_n

1) find a and b st $P\left(a \le Q_n\left(\vec{Y};\theta\right) \le b\right) = p$

2) solve for θ from $a \le Q_n\left(\vec{Y};\theta\right) \le b$ to get $L\left(\vec{Y}\right) \le \theta \le U\left(\vec{Y}\right)$

<u>Ex 4.4.2</u> $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N\left(\mu, \sigma^2\right)$ μ unknown σ^2 is known

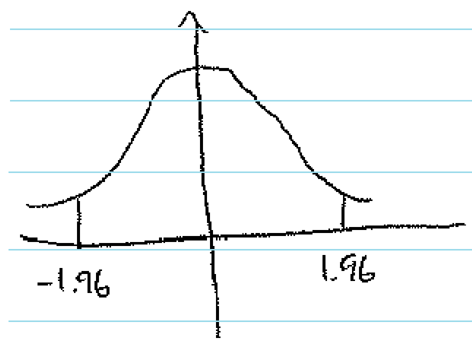
$Q_n\left(\vec{Y};\mu\right) = \dfrac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ so $Q_n\left(\vec{Y};\mu\right)$ is a pivotal quantity

to construct a 95% CI for μ

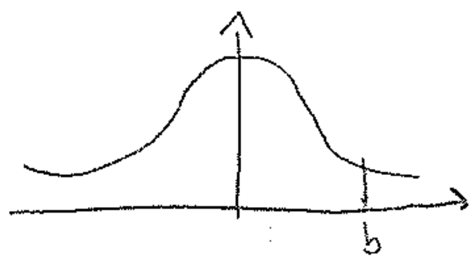
$P\left(a \le \dfrac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \le b\right) = 0.95$ → then solve for μ from $a \le \dfrac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \le b$ to get

$\implies \bar{Y} - b\dfrac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} - a\dfrac{\sigma}{\sqrt{n}}$

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



<u>Note:</u> ① $\left[\vec{Y} - 1.96\dfrac{\sigma}{\sqrt{n}}, \vec{Y} + 1.96\dfrac{\sigma}{\sqrt{n}}\right]$ takes the form point estimator ± c · sd (point estimator) which is known as "two-sided" CI



② If we choose $a = -\infty$, $b = 1.645$, then we have $\left[\vec{Y} - 1.645\dfrac{\sigma}{\sqrt{n}}, \infty\right)$

Similarly, we can take $a = -.1645$, $b = \infty$ to get another "one-sided" CI for μ

$(-\infty, \square]$ $\quad P(\mu \le \square] = 0.95$

How to obtain a pivotal quantity?

for most problems, it's not possible to get an "exact" pivotal quantity so we turn to "asymptotic" pivotal quantity, $Q_n\left(\vec{Y};\theta\right)$ st the distribution of Q_n is known as $n \to \infty$

<u>Ex 4.4.3</u> $Y \sim \text{Binomial}(n;\theta)$ we want a 95% CI for θ

Based on CLT, $\dfrac{Y-\theta}{\sqrt{n\theta(1-\theta)}} \sim N(0,1)$ as $n \to \infty$ $\quad Z_1,\dots,Z_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta) \quad Y = \sum_{i=1}^{n} Z_i$

Ex 4.4.3 $Y \sim \text{Binomial}(n,\theta)$ we want a 95% CI for θ 2016 01 25

$$Q_n(\theta) = \frac{Y-n\theta}{\sqrt{n\theta(1-\theta)}} \sim N(0,1) \text{ as } n \to \infty$$

$$P(a < Q_n(\theta) < b) \approx 0.95 \implies \text{ solve for } \theta \text{ from } -1.96 \le \frac{Y-n\theta}{\sqrt{n\theta(1-\theta)}} \le 1.96$$

$$a = -1.96, b = 1.96 \qquad \implies \tilde{Q}_n(\theta) = \frac{Y-n\theta}{\sqrt{n\hat{\theta}1-\hat{\theta}}} \text{ where } \hat{\theta} = \frac{\bar{Y}}{n}$$

$$\sim N(0,1) \text{ as } n \to \infty$$

$$\text{solve for } \theta \text{ from } -19.6 \le \frac{Y-n\theta}{\sqrt{n\hat{\theta}\left(1-\hat{\theta}\right)}} \le 1.96 \implies \bar{Y}-1.96\sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}} \le \theta \le \bar{Y}+1.96\sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}$$

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

Ex 4.4.5 $Y \sim \text{Binmoal}(n,\theta)$ a 95% CI for θ is [,]
Suppose now we would like the length of the 95% CI no longer than a pre-fixed Δ then what n should we take?

$$\text{The length of the 95\% CI is } 2\cdot 1.96\sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}} \le \Delta \implies n \ge \left(\frac{2\cdot 1.96}{\Delta}\right)^2 \underbrace{\hat{\theta}\left(1-\hat{\theta}\right)}_{\le 0.25}$$

$$\implies n \ge \left(\frac{2\cdot 1.96}{\Delta}\right)^2 \cdot 0.25 \quad \text{because } 0 < \hat{\theta} < 1$$

for example, if $\Delta = (0.03)\cdot 2 \implies n \ge 1067.1$ or $n \ge 1068$
Thus by taking $n \ge 1068$, we have $P\left(\left|\hat{\theta}-\theta\right| \le 0.03\right) \ge 95\%$

Def Let $Z_1,\dots,Z_k \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $X = \sum_{i=1}^{k} Z_i^2$ the we call the distribution of X the $\chi^2(k)$
•distribution with k df

The pdf of a $X \sim \chi^2(k)$ is $f(x;k) = \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad x > 0$

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

i) $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$

ii) $\Gamma(n) = (n-1)!$

iii) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

when $k = 1$ $\quad X \sim \chi^2(1)$ find the pdf $f_X(s)$ $\qquad X = Z^2$ where $Z \sim N(0,1)$

for $x \ge 0$ $P_X(X \le x) = P\left(Z^2 \le x\right) = P\left(-\sqrt{x} \le Z \le \sqrt{x}\right) = \Phi\left(\sqrt{x}\right) - \Phi\left(-\sqrt{x}\right)$ $\quad \Phi$ is the CDF of $Z \sim$

$N(0,1)$ ϕ is the pdf of $Z \sim N(0,1)$

$$f_X(x) = \frac{dP_x(X \le x)}{dx} = \phi\left(\sqrt{x}\right)\tfrac{1}{2}x^{-\frac{1}{2}} + \phi\left(-\sqrt{x}\right)\tfrac{1}{2}x^{-\frac{1}{2}}$$
$$= x^{-\frac{1}{2}}\phi\left(\sqrt{x}\right) = \frac{1}{\sqrt{2\pi}}x^{-\frac{1}{2}}e^{-\frac{x}{2}} \quad x \ge 0$$

<u>Thm:</u> Let $W_1, \ldots, W_n$ be independent random variables with $W_i \sim \chi^2(k_i)$, $i = 1, \ldots, n$
then $\sum_{i=1}^{n} W_i \sim \chi^2\left(\sum_{i=1}^{n} k_i\right)$

solve for θ from $-1.96 \le \dfrac{Y - n\theta}{\sqrt{n\hat{\theta}(1-\hat{\theta})}} \le 1.96 \implies \bar{Y} - 1.96\sqrt{\dfrac{\hat{\theta}(1-\hat{\theta})}{n}} \le \theta \le \bar{Y} + 1.96\sqrt{\dfrac{\hat{\theta}(1-\hat{\theta})}{n}}$

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

Moment Generating Function of a distribution
$X \sim \chi^2(k)$ MGF of X is $M_x(t) = E\left(e^{tX}\right)$

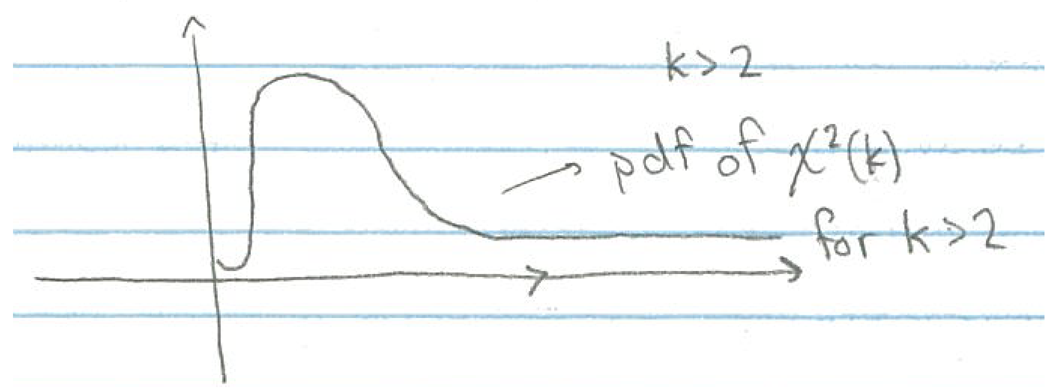
let $\quad y = (1-2t)x$
$dy = (1-2t)dx$
$dx = \dfrac{dy}{1-2t}$

$$\otimes M_x(t) = \int_0^\infty e^{tx}\underbrace{\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}}_{c}x^{\frac{k}{2}-1}e^{-\frac{k}{2}}\,dx = \int_0^\infty c \cdot x^{\frac{k}{2}-1}e^{-\frac{1-2t}{2}x}\,dx$$

$$= \int_0^\infty c\left(\frac{y}{1-2t}\right)^{\frac{k}{2}-1}e^{-\frac{y}{2}}\frac{dy}{1-2t} = \frac{1}{(1-2t)^{k/2}}\int_0^\infty y^{\frac{k}{2}-1}e^{-\frac{y}{2}}\,dy = \frac{1}{(1-2t)^{k/2}} \qquad t < \frac{1}{2}$$

$$EX = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = k(1-2t)^{-\frac{k}{2}-1}\Big|_{t=0} = k$$

$$EX^2 = \left.\frac{d^2M_X(t)}{dt^2}\right|_{t=0} = k(k+2)(1-2t)^{-\frac{k}{2}-1}\Big|_{t=0} = k(k+2)$$

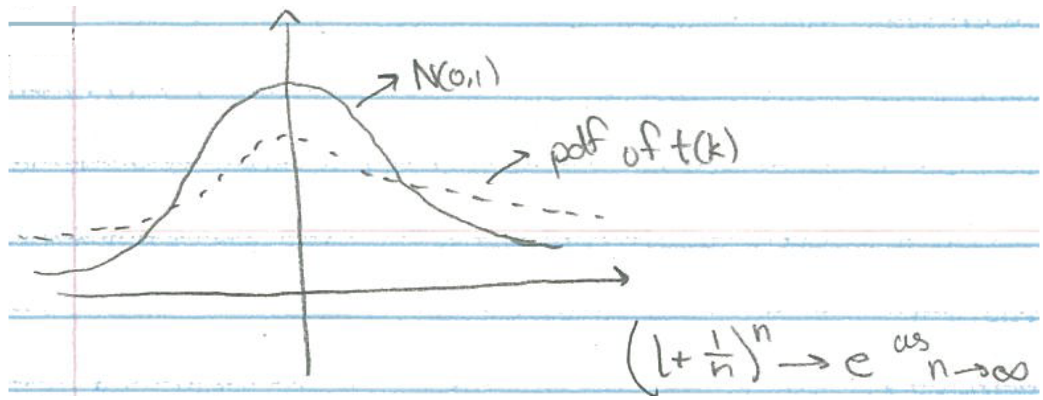


<u>Def:</u> Suppose $X \sim N(0,1)$, $Y \sim \chi^2(k)$ and $X \perp Y$ then we call the distribution of $T = \dfrac{X}{\sqrt{\frac{Y}{k}}}$

the $t(k)$ distribution with k df

$T \sim t(k) \quad f_T(t;k) - C_k\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}$

$\overset{k\to\infty}{\to} e^{-\frac{t^2}{2}}$



Def: $\Lambda \equiv \Lambda(\theta) \equiv -2\log\frac{L(\theta)}{L(\hat\theta)}$, where $\hat\theta$ is the MLE, is called the likelihood-ratio statistic 2016 01 27

$\Lambda = 2\log L(\hat\theta) - 2\log L(\theta) = 2\ell(\hat\theta) - 2\ell(\theta)$

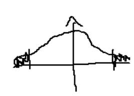
Thm: Suppose θ is the true value of the parameter, then $\Lambda(\theta) \sim \chi^2(1)$ as $n \to \infty$
Thus $\Lambda(\theta)$ is asymptotically pivotal
To construct a 100% CI
Step 1: find a value c s.t. $P(W \le c) = p$ where $W \sim \chi^2(1)$
Step 2: Solve for θ from $\Lambda(\theta) \le c \implies \{\theta : \Lambda(\theta) \le c\}$

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

Ex. $Y \sim \text{Binomial}(n,\theta) \quad P(Y=y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = L(\theta)$

$\hat\theta = \frac{Y}{n} \implies \frac{L(\theta)}{L(\hat\theta)} = \left(\frac{\theta}{\hat\theta}\right)^Y\left(\frac{1-\theta}{1-\hat\theta}\right)^n - Y \implies \Lambda(\theta) = -2Y\log\frac{\theta}{\hat\theta} - 2(n-Y)\log\frac{1-\theta}{1-\hat\theta}$

To get a 95% CI for θ



Step 1: find c s.t. $P(W \le c) = .95$ where $W \sim \chi^2(1) \quad W = Z^2 \quad Z \sim N(0,1)$

$\implies c = 1.96^2 = 3.841 \qquad P(-1.96 \le Z \le 1.96) = 0.95$

$P(W \le c) = 0.95$

Step 2: solve for θ from $\Lambda(\theta) \le 3.84$.

$Y = y = 40 \ n = 100 \implies \hat\theta = \frac{Y}{n} = 0.4 \implies \Lambda(\theta) = -80\log\frac{\theta}{0.4} - 120\log\frac{1-\theta}{0.6}$

Section 4.5 Inference for $N(\mu,\sigma^2)$

$Y_1,\cdots,Y_n \overset{\text{i.i.d}}{\sim} N(\mu,\sigma^2)$ our interest is to estimate both μ and σ^2

Point estimators: $\hat\mu = \frac{1}{n}\sum_{i=1}^n Y_i \quad \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n\left(Y_i - \bar Y\right)^2 \quad S^2 = \frac{1}{n-1}\sum_{i=1}^n\left(Y_i-\bar Y\right)^2 \quad ES^2 = \sigma^2$

Interval estimators:

For μ $\begin{cases} i) \text{ with know } \sigma^2 & \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \implies 95\% \text{ CI for } \mu \text{ is } \left[\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right] \\ ii) \text{ with unknow } \sigma^2 & \underline{\text{Thm:}}\ T = \frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t(n-1) \quad Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \end{cases}$

$$= \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

To ~~con~~ get a 95% CI for μ

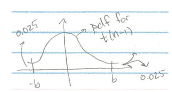
step 1: $P\left(a \le \frac{\bar{Y}-\mu}{S/\sqrt{n}} \le b\right) = 0.95$

step 2: solve for μ from $a \le \frac{Y-\mu}{S/\sqrt{n}} \le b$

$\implies$ a 95% CI for μ is $\left[\bar{Y} - b\frac{S}{\sqrt{n}}, \bar{Y} - a\frac{S}{\sqrt{n}}\right]$

if we take $a = -b$ then

$$\left[\bar{Y} - b\frac{S}{\sqrt{n}}, \bar{Y} + b\frac{S}{\sqrt{n}}\right]$$



For σ^2 $\underline{\text{Thm:}}$ $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ then $U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

Thus U is a pivotal quantity

To construct a 100% CI for σ^2

step 1: find a and b s.t $P\left(a \le \frac{(n-1)S^2}{\sigma^2} \le b\right) = p$

step 2: solve for σ^2 from $a \le \frac{(n-1)S^2}{\sigma^2} \le b$ gives $\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right]$ is a 100% CI for σ^2

"equal=tail" CI

If we are just interested in the upper bound of σ^2

we can take $b = \infty$ so that $P\left(\frac{(n-1)S^2}{\sigma^2} \ge a\right) = p$

then a "one-sided" 100% CI for σ^2 is $\left[0, \frac{(n-1)S^2}{a}\right]$

Prediction $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$ Y is a new random draw from the same $N(\mu, \sigma^2)$ projector 2016 02 01

we want to $\underline{\text{predict}}$ Y

point prediction: $\bar{Y_n}$ or μ

Interval prediction:

$$Y - \underline{\bar{Y} \sim N\left(\underline{0}, \left(1 + \tfrac{1}{n}\right)\sigma^2\right)} \quad \text{Var}\left(Y - \bar{Y_n}\right) = \text{Var}(Y) + \text{Var}\left(\bar{Y_n}\right) = \sigma^2 + \frac{\sigma^2}{n}$$

1 σ^2 is known $\frac{Y - \bar{Y_n}}{\sigma\sqrt{1+\frac{1}{n}}} \sim N(0,1)$ $\quad P\left(-1.96 \le \frac{Y - \bar{Y_n}}{\sigma\sqrt{1+1\frac{1}{n}}} \le 1.96\right) = 0.95$

$$\implies 95\% \text{ PI for } Y \text{ is } \left[\bar{Y_n} - 1.96\sigma\sqrt{1+\tfrac{1}{n}}, \bar{Y_n} + 1.96\sigma\sqrt{1+\tfrac{1}{n}}\right]$$

② σ^2 unknown $\quad \dfrac{y-\bar{Y_n}}{\sigma\sqrt{1+\frac{1}{n}}} \sim N(0,1) \quad \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \quad \bar{Y_n} \perp S^2$

$$\frac{Y-\bar{Y_n}}{S\sqrt{1+\frac{1}{n}}} = \frac{\frac{Y-\bar{Y_n}}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \sim t(n-1)$$

95% PI $P\left(-t_{0.975}(n-1) \le \dfrac{Y-\bar{Y_n}}{S\sqrt{1+\frac{1}{n}}} \le t_{0.975}(n-1)\right) = 0.95$

$\implies$ a 95% PI for Y is $\left[\bar{Y_n} - t_{0.975}(n-1)S\sqrt{1+\frac{1}{n}}, \bar{Y_n} + t_{0.975}(n-1)S\sqrt{1+\frac{1}{n}}\right]$

95% CI for μ is $\left[\bar{Y_n} - t_{0.975}(n-1)S\sqrt{\frac{1}{n}}, \bar{Y_n} + t_{0.975}(n-1)S\sqrt{\frac{1}{n}}\right]$

CI $\bar{Y_n}$ μ

PI $\bar{Y_n} \to \mu \to Y$

$\Lambda(\theta) \equiv -2\log\frac{L(\theta)}{L(\hat{\theta})} \to \chi^2(1)$ as $n\to\infty$

Ex. $Y_1,\ldots,Y_n \overset{\text{i.i.d.}}{\sim} N(\mu,\sigma^2)$ where μ is unknown, σ^2 is known, we want a 95% CI for μ

$f(y;\mu) = \dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

$\implies L(\mu) = \prod_{i=1}^{n} f(Y_i;\mu) \qquad \Lambda(\mu) = -2\log\dfrac{L(\mu)}{L(\hat{\mu})} = s\ell(\hat{\mu}) - 2\ell(\mu)$ where $\ell(\mu) = \log L(\mu)$

$$= 2\left[-\frac{\sum_{i=1}^{n}(Y_i-\hat{\mu})^2}{2\sigma^2} + \frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(Y_i-\mu)^2 - (Y_i-\hat{\mu})^2\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(Y_i-\hat{\mu}+\hat{\mu}-\mu)^2 - (Y_i-\hat{\mu})^2\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\left(\bar{Y_n}-\mu\right)^2\right] = \frac{n\left(\bar{Y_n}-\mu\right)^2}{\sigma^2} \sim \chi^2(1) \text{ as } n\to\infty$$

$$= \frac{\bar{Y_n}-\mu}{\sigma/\sqrt{n}}^2 \sim \chi^2(1)$$

to get a 95% CI for μ $\quad P\left(\left(\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}\right)^2 \leq \chi_0^2.95\,(1)\right) = 0.95$

$-\sqrt{\chi_0^2.95\,(1)} \leq \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \leq \sqrt{\chi_0^2.95\,(1)}$

95% CI for μ is $\left[\bar{Y}_n - \sqrt{\chi^2_{0.95}\,(1)}\frac{\sigma}{\sqrt{n}}, \bar{Y}_n + \sqrt{\chi^2_{0.95}\,(1)}\frac{\sigma}{\sqrt{n}}\right]$

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

exer 2.2

exer 4.30
2016 02 03

exer 4.30

exer 4.31
Midterm I

2016 02 08

Chapter 5 – Hypothesis Testing

Simple example: flipping a coin and counting number of tails/heads to see if it is fair

Hypothesize coin is fair and collect data

relation between lung cancer and smoking? hypothesis no. follow smokers and non-smokers for 5 or 10 years and see how many get cancer. collect data and see if they support hypothesis

null hypothesis is that drug has no effect at all. see if data is for or against hypothesis

Example **2.** [Lady Tasting Tea - R.A. Fisher] he gave her 8 cups of tea, 4 TM (tea added to milk) 4 MT (milk added to tea) in random order

the lady is told that there are 4 and 4, is asked to tell which 4 are TM and which 4 are MT.

Suppose that she correctly tells all 4 TM. Does this really mean that she has the ability to tell which is which?

H_0: she has no ability to tell which is which (null hypothesis; she is just guessing) under H_0 there are in total $\binom{8}{4} = 70$ different ways to randomly choose 4 and only one of these ways is correct. Since she correctly tells which 4 are TM, we have two possibilities:

1. H_0 is true, but an event with probability 1/70 occurred
2. H_0 is false

Since the probability 1/70 is small, the observed data is against (1), or against H_0, so we reject H_0

Now suppose that the lady gets 3 TM correct. The probability that, by purely guessing, the lady can tell at least 3 TM correctly is

$$P\left(\text{she can tell at least 3 TM} \mid H_0\right) = \frac{1 + \binom{4}{3}\binom{4}{1}}{70} = \frac{17}{70} = 0.243.$$

Why are we looking at *at least* 3, as opposed to exactly 3? One reason is that when we have continuous things, the probability of a single value is zero, so we have no choice but to consider some interval, and the interval of the event together with the less likely things seems natural and works well.

We still have two possibilities:

1. H_0 is true but an event with probability 0.243 occurred
2. H_0 is false

Since 0.243 is not very small, the observed data doesn't provide evidence against H_0 so we do not reject H_0

In practice, a level α needs to be pre-fixed so that when the calculated probability is less than α we consider that the data provides enough evidence against H_0

We look at a *null hypothesis*, rather than the alternative to the null hypothesis, because we do calculations with the null hypothesis. I think?

A summary of steps needed for a hypothesis testing problem:

1. Specify the "NULL Hypothesis" H_0
2. Find a test statistic D

   REMARK **3.** i) A test statistic is a function of the observed data, and is used to measure how well the observed data agrees with H_0 (In the previous example we had $D = 4$ and then $D = 3$.)

   ii) $P\left(D \geq d \mid H_0\right)$ is the probability that, under H_0, we observe the current event and events that are even less likely to occur; it is called the p-value

3. Under H_0, calculate $P\left(D \geq d \mid H_0\right)$ hwere d is the observed value of D
4. Draw a conclusion by comparing the p-value with a pre-fixed threshold α:

If the p-value is less than α then we reject the null hypothesis H_0.

If the p-value is greater than α then we fail to reject H_0.

Hi Wilson

It's not terribly important, for practical purposes, what to do if the p-value is equal to α. The value of α is artificial, and for experimental measurement, are you likely to get equality? People often use 0.05. Sometimes 0.1 or 0.01.

If we had an alternative hypothesis, apparently it would be notated as H_a (a for alternative)

Remark **4.** By setting α at a low level, such as 0.05 and 0.01, rejecting H_0 means we have strong evidence against H_0. However, not rejecting H_0 doesn't mean that H_0 is true; it only means that the observed data doesn't provide enough evidence to say H_0 is false.

Example **5.** To study if a coin is fair. The null hypothesis is $H_0 : \theta = 0.5$, or $\theta - 0.5 = 0$. Out of 100 tosses, we observed 52 heads: HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-HHHHHHHHHHHHHHHHHHHHHH.

Let Y denote the number of heads, a random variable I presume. We take the test statistic to be $D = |Y - 50|$. This is the discrepancy between the actual data and the expected data. Later we will discuss a more formal/systematic way to select test statistics.

The p-value is

$$P(D \geq d \mid H_0) = P(|Y - 50| \geq |52 - 50| \mid H_0) = P(Y \leq 48 \text{ or } Y \geq 52 \mid H_0)$$

Under the null hypothesis H_0, $Y \sim \text{Binomial}(100, \theta = 0.5)$. Hence the above is

$$1 - P(Y = 49 \text{ or } Y = 50 \text{ or } Y = 51 \mid H_0) \approx 0.76 > 0.05.$$

Hence we fail to reject the null hypothesis H_0.

People do not tend to work backwards, as in "what would I need to get as observed data in order to reject H_0 for this α", then just take the data and calculate the p-value and compare to α.

Example **6.** [6-sided die] Null hypothesis: $H_0 : \theta = 1/6$, where θ is $P(\#1 \text{ showing up})$. (An alternative hypothesis would be $H_a : \theta \neq 1/6$.) Suppose we observed 44 ones out of 180 rolls. Let Y be the number of ones rolled out of n rolls. Take the test statistic $D = \left|Y - \frac{n}{6}\right|$. The p-value is then

$$P(D \geq d \mid H_0) = P\left(\left|Y - \frac{n}{6}\right| \geq \left|44 - \frac{180}{6}\right| \mid H_0\right) = P(Y \leq 16 \text{ or } Y \geq 44 \mid H_0).$$

Under the null hypothesis, $Y \sim \text{Binomial}(180, \theta)$. Thus the p-value is, after calculation, approximately $0.007 < \alpha = 0.05$. Therefore we reject the null hypothesis.

2016 02 10

Midterm Return 2016 02 22

We reject H_0: we have strong evidence against H_0. Suppose that we suspect that the number one turns up more often than if the die were fair. Suppose again $n = 180$ and $y = 44$. Do we have enough evidence to say that $\theta > 1/6$?

$H_0 : \theta = 1/6$ vs $H_a = \theta > 1/6$

Consider for this purpose the test statistic $D = \max\{Y - n/6, 0\}$. Take $d = \max\{y - n/6, 0\} = 14$. We calculate the p-value to be

$$P(D \geq d \mid H_0) = P(\max\{Y - n/6, 0\} \geq 14 \mid P(Y \geq 44 \mid H_0)) \approx 0.005 < 0.05.$$

So we reject H_0: we have strong evidence that $\theta > 1/6$.

Suppose that instead of $y = 44$ we observed $y = 35$. Then the p-value is

$$P(D \geq d \mid H_0) = P(D \geq 5 \mid H_0) = P(Y \geq 35 \mid H_0) \approx 0.18 > 0.05.$$

We fail to reject H_0.

Consider $H_0 : \theta = 1/6$ vs $H_a : \theta > 1/6$. We have $D = Y - n/6$, $d = y - n/6$. The p-value is $P(D \geq d \mid H_0) = P(Y - n/6 \geq d \mid \theta = 1/6)$ where $Y \sim \text{Binomial}(n, 1/6)$.

Consider $H_0 : \theta = 1/6$ vs $H_a : \theta < 1/6$. Here $D = Y - n/6$ and $d = y - n/6$. The p-value is $P(D \leq d \mid H_0)$. Small values of D provide evidence against H_0 in the direction of H_a.

Remark **7.** Different test statistics may be used to solve the same hypothesis testing problem. In hypothesis testing, there are two types of errors:

Type I error: $P(\text{reject } H_0 \mid H_0 \text{ is true})$

Type II error: $P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$

Usually, we would like to control type I error to a small level (say 0.05) and then try to reduce Type II error, or increase

$$1 - \text{Type II error} \equiv P(\text{reject } H_0 \mid H_0 \text{ is false}) = \text{power}.$$

Testing hypothesis under a normal model

We have $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ where μ and σ^2 are unknown.

Hypothesis for μ: $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$. Consider $T = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t(n-1)$. Take $D = |T| = \left|\frac{\bar{Y} - \mu}{s/\sqrt{n}}\right|$.

Given $y_1, \ldots, y_n$ we have $d = \left|\frac{\bar{y} - \mu}{s/\sqrt{n}}\right|$. The p-value is

$$P(D \geq d \mid H_0) = P(|T| \geq d \mid H_0) = 1 - P(|T| \leq d) = 1 - P(-d \leq T \leq f).$$

Consider $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$. Use the test statistic $D = T$. Large values of D provide evidence against H_0 in the direction of H_a. The p-value is $P(D \geq d \,|\, H_0)$.

Consider $H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$. Use the test statistic $D = T$. Small values of D provide evidence against H_0 in the direction of H_a. The p-value is $P(D \leq d \,|\, H_0)$.

Example **8.** [5.1.2] Let $n = 10$, $\bar{y} = 0.9810$, $s = 0.0170$. Take $H_0 : \mu = 1$ vs $H_a : \mu \neq 1$. Then

$$D = |T| = \left| \frac{\bar{Y} - \mu}{s/\sqrt{n}} \right|$$

and

$$d = \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| = 3.534.$$

The p-value is

$$P(D \geq d \,|\, H_0) = P(|T| \geq 3.534) = 0.0064 < 0.05$$

where $T \sim t(9)$. We reject H_0.

Remark **9.** Although there is strong evidence against H_0, we can say nothing about the magnitude of the deviation between the true value μ and 1.

A 95% CI for μ using $T = \frac{\bar{Y}-\mu}{s/\sqrt{n}}$ is

$$\left[\bar{y} - 2.2622 s/\sqrt{10}, \bar{y} + 2.2622 s/\sqrt{10}\right] = [0.969, 0.993].$$

Although the 95% CI doesn't contain 1, the true value of μ may not be far away from 1.

Connection between hypothesis testing and CI using the same pivotal quantity. 2016 02 24

First consider the normal distribution case. We have $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. We have the null hypothesis $H_0 : \mu = \mu_0$ and the alternate hypothesis $H_a : \mu \neq \mu_0$. Take test statistic $D |T| = \left| \frac{\bar{Y}-\mu}{s/\sqrt{n}} \right|$. The p-value is

$$P(D \geq d \,|\, H_0) = P\left(\left| \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \right| \geq \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \right) = P\left(|T| \geq \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \right)$$

where $t \sim t(n-1)$. So the p-value is at least 0.05 if and only if

$$\begin{aligned} P\left(|T| \leq \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \right) \leq 0.95 \Leftrightarrow & \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \leq t_{0.975}(n-1) \\ \Leftrightarrow & \mu_0 \in \left[\bar{y} - t_{0.975}(n-1) s/\sqrt{n}, \bar{y} + t_{0.975}(n-1) s/\sqrt{n}\right]. \end{aligned}$$

Suppose $[L(\mathbf{Y}), U(\mathbf{Y})]$ is a 95% CI for θ. For $\theta^* \in [L(\mathbf{y}), R(\mathbf{y})]$, we can test $H_0 : \theta = \theta^*$ vs $H_a : \theta \neq \theta^*$. It turns out the p-value of this problem is at least 0.05. For any $\theta^* \in$

$[L(\mathbf{y}), U(\mathbf{y})]$ the p-value is less than 0.05. On the other hand if the p-value for $H_0 : \theta = \theta^*$ is

$$\begin{cases} \geq 0.05 \text{ then } \theta^* \in [L(\mathbf{y}), U(\mathbf{y})] \\ < 0.05 \text{ then } \theta^* \notin [L(\mathbf{y}), U(\mathbf{y})] \end{cases}$$

Section 5.3 Likelihood Ratio Test

The goal is to test $H_0 : \theta = \theta_0$. We discussed, but did not prove, that $\hat{\theta}_{MLE} \xrightarrow{P} \theta_0$ (and this uses the implicit assumption that the maximizer is unique). Therefore, if θ_0 is the true value, then $\hat{\theta}_{MLE}$ should "be close" to θ_0 and thus $L(\hat{\theta})$ should be close to $L(\theta_0)$. (Note L is the likelihood function.) That is,

$$R(\theta_0) = \frac{L(\theta_0)}{L(\hat{\theta})}$$

should be close to one (if n is large (where n is the number of sample data)).

So small values of $R(\theta_0)$ provides evidence against H_0. To calculate the p-value, we need the distribution of $R(\theta)$ under H_0.

Theorem **10.** *If $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} f(y; \theta)$ and θ is a scalar, then*

$$\Lambda(\theta) = -2\log(R(\theta)) = -2\log\left(\frac{L(\theta)}{L(\hat{\theta})}\right) \overset{app.}{\sim} \chi^2(1)$$

as $n \to \infty$. (This appears to be a definition of Λ.) Note $\Lambda(\theta)$ is called the likelihood ratio.

Under the null hypothesis $H_0 : \theta = \theta_0$, $\Lambda(\theta_0) \sim \chi^2(1)$ as $n \to \infty$. Since small values of $R(\theta)$ corresponds to large values of $\Lambda(\theta)$, large values of $\Lambda(\theta_0)$ provides evidence against H_0. The p-value is $P(\Lambda(\theta_0) \geq \lambda(\theta_0) \mid H_0)$. Note that λ is the observed value of Λ.

Example **11.** [5.3.2] We have some things and we want to measure their lifetime or something. Anyway, we end up with $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\theta)$. The observed data is $n = 20$ and $\sum_{i=1}^n y_i = 38524$. Take the null hypothesis $H_0 : \theta = 2000$, and the alternate hypothesis $H_a : \theta \neq 2000$. We have/known/claim(?)

$$f(y; \theta) = \frac{1}{\theta} e^{-y/\theta},$$

for $\theta > 0$, $y > 0$. The likelihood function is

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

Skipping some computation, we end up with $\hat{\theta} = \bar{Y}$. The likelihood ratio statistic therefore is

$$\begin{aligned}\Lambda(\theta) &= 2\log\left(\frac{L(\theta)}{L(\hat{\theta})}\right)\\ &= 2\log\left(L(\hat{\theta})\right) - 2\log\left(L(\theta)\right)\\ &= 2n\left(\log(\theta) + \frac{\bar{Y}}{\theta} - \log(\hat{\theta}) - \frac{\bar{Y}}{\hat{\theta}}\right)\\ &= 2n\left(\log\left(\frac{\theta}{\bar{Y}}\right) + \frac{\bar{Y}}{\theta} - 1\right).\end{aligned}$$

The p-value is

$$P\left(\Lambda(\theta) \geq \lambda(\theta) \mid H_0\right) = P(\underbrace{\Lambda(\theta_0)}_{\sim\chi^2(1)} \geq 0.028) \approx 0.87 > 0.05.$$

So we fail to reject H_0.

Example **12.** [5.2.3] Suppose we have $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, where μ is unknown but σ^2 is known. Consider $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$. Take the test statistic to be

$$D = \left|\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right| = |Z|$$

where $Z \sim N(0,1)$. We have

$$F(y;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

and hence

$$L(\mu) = \prod_{i=1}^{n} f(Y_i;\mu) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{2\sigma^2}}.$$

This means

$$l(\mu) = \log(L(\mu)) = c - \frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{2\sigma^2}.$$

Some computation not included here goes to show that $\hat{\mu} = \bar{Y}$. Note that

$$\sum_{i=1}^{n}(Y_i\mu)^2 = \sum_{i=1}^{n}\left(Y_i - \bar{Y} + \bar{Y} - \mu\right)^2 = \cdots = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 + n\left(\bar{Y} - \mu\right)^2.$$

Now we see

$$\begin{aligned}
\Lambda(\mu) &= -2l(\hat{\mu}) - 2l(\mu) \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(Y_i-\mu)^2 - \sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2\right) \\
&= \ldots ? \\
&= \frac{n}{\sigma^2}\left(\bar{Y}-\mu\right)^2 \\
&= \left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1).
\end{aligned}$$

The p-value is

$$P(D \geq d \,|\, H_0) = P\left(\left|\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}\right| \geq d\right) = P\left(\left(\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \geq d^2\right).$$

Now we generalize to having more than one parameter. Suppose we have $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f(y; \boldsymbol{\theta})$, where $\boldsymbol{\theta}_{k\times 1} \in \Omega$ and $\dim(\Omega) = k$. Let $\Omega_0 \subset \Omega$ with $\dim(\Omega_0) = r < k$. Now we want to test $H_0 : \boldsymbol{\theta} \in \Omega_0$.

Example **13.** For the normal distribution $N(\mu, \sigma^2)$ we have $\boldsymbol{\theta} = (\mu, \sigma^2)$. We have

$$\Omega = \left\{\left(\mu, \sigma^2\right); \mu \in \mathbb{R}, \sigma > 0\right\}.$$

We see $\dim(\Omega) = 2$. To have null hypothesis $H_0 : \mu = \mu_0$ we set $\Omega_0 = \{(\mu, \sigma^2); \mu = \mu_0, \sigma > 0\}$. To have null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ we set $\Omega_0 = \{(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma = \sigma_0\}$.

Example **14.** Suppose $S_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma_2^2)$. Let $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, so $\dim(\Omega) = 4$. To take null hypothesis $H_0 : \mu_1 = \mu_2$, we take $\Omega_0 = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2); \mu_1 = \mu_2 \in \mathbb{R}, \sigma_1 > 0, \sigma_2 > 0\}$, so $\dim(\Omega_0) = 3$.

The *likelihood-ratio statistic* is 2016 02 29

$$\Lambda = -2\log\left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right)$$

where

$$\begin{aligned}
\hat{\theta}_0 &= \arg\max_{\theta \in \Omega_0} L(\theta) \\
\hat{\theta} &= \arg\max_{\theta \in \Omega} L(\theta)
\end{aligned}$$

and recall the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y; \boldsymbol{\theta}).$$

Theorem 15. [Wilks' Theorem] *Under H_0, $\Lambda \overset{app}{\sim} \chi^2(k-r)$ as $n \to \infty$.*

The p-value for the above test is $P(\Lambda \geq \lambda \mid H_0) = P(W \geq \lambda)$ where $W \sim \chi^2(k-r)$.

Example 16. [5.4.4] Suppose $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ where μ and σ^2 are unknown. Take the null hypothesis to be $H_0 : \sigma^2 = \sigma_0^2$. The parameter space is $\Omega = \{(\mu, \sigma^2) ; \mu \in \mathbb{R}, \sigma > 0\}$, and $\Omega_0 = \{(\mu, \sigma^2); \mu \in \mathbb{R}\}$. The likelihood function is

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

So

$$l\left(\mu, \sigma^2\right) = \log\left(L\left(\mu, \sigma^2\right)\right) = -n \log \sigma - \frac{\sum_{i=1}^{n}\left(y_i - \mu\right)^2}{2\sigma^2} + c.$$

One calculates that $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}^2)$ where $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$.

Something about $\hat{\theta}_? = \hat{\mu}_0 = \bar{y}$???

Now calculate

$$\begin{aligned}
\Lambda &= -2\log\left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right) \\
&= 2l(\hat{\theta}) - 2l(\hat{\theta}_0) \\
&= -2n\log(\hat{\sigma}) - \frac{\sum_{i=1}^{n}(y_i - \hat{\mu})^2}{\hat{\sigma}^2} + 2n\log(\sigma_0) + \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_0)^2}{\sigma_0^2} \\
&= -n\log\left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right) + \sum_{i=1}^{n}(y_i - \bar{y})^2\left(\frac{1}{\sigma_0^2} - \frac{1}{\hat{\sigma}^2}\right) \\
&= -n\log\left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right) + n\left(\frac{\hat{\sigma}^2}{\sigma_0^2} - 1\right).
\end{aligned}$$

Definitely must be familiar with and able to do these types of calculations for test.

This finishes chapter 5, though we will revisit likelihood ratio test in chapter 7, looking at applications of Wilks' Theorem.

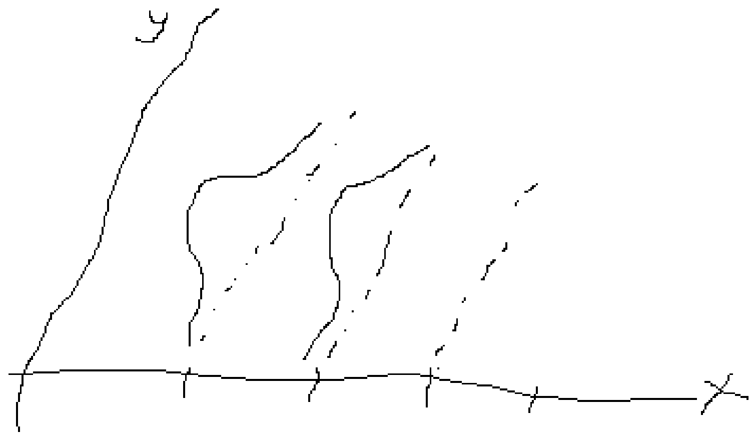
Chapter 6 – Gaussian Response Model

We always want to model exactly one random variable, denoted Y. We assume it follows some sort of distribution, and study the consequences. In practice, many variables can be explained with other variables.

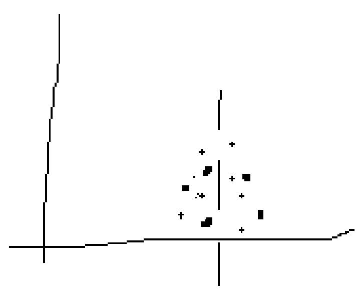
The goal is to use $\mathbf{X}$ to explain the distribution of Y. We assume $Y \sim N(\mu, \sigma^2)$ where μ and σ are unknown constants. In this chapter, we allow μ and σ to depend on other random variables or factors.

We will assume that, given $\mathbf{X}_i = \mathbf{x}_i$, $Y_i \sim N(\mu(x_i), \sigma^2(x_i))$.

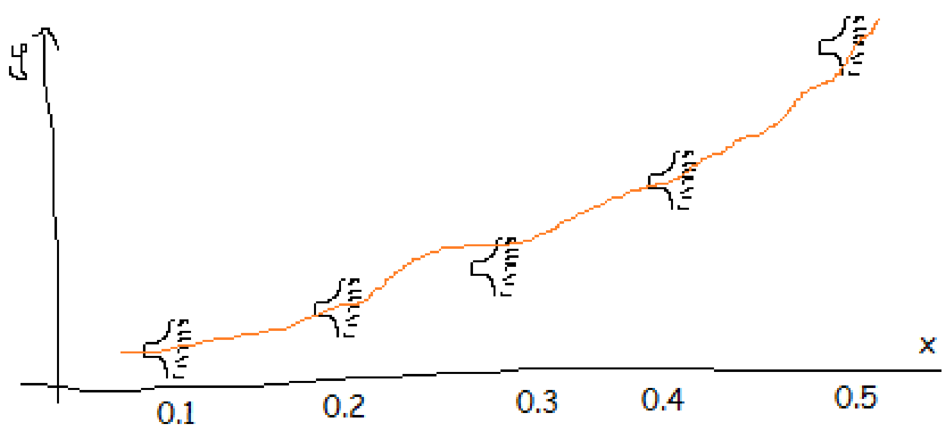
Y: "response variable" or "outcome"
$\mathbf{X}$: "explanatory variables" or "covariates"



If $\mathbf{X}$ is continuous, then we won't see these lines, we will see more like:



Example **17.** [6.1.3 with some exaggeration] Consider



We see that variance does not depend on x_i but mean does. So $Y_i \sim N(\mu(x_i), \sigma^2)$. Looking at the curve we fitted, it seems reasonable to assume it is quadratic, so $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. The unknown parameters are $\beta_0, \beta_1, \beta_2, \sigma^2$; these need to be estimated from the observed data.

It is common to assume (and we will, for this course) that $\mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ (here $\mathbf{x}$ is a k-dimensional vector of covariates) and $\sigma^2(\mathbf{x}) \equiv \sigma^2$.

Under these assumptions, the Gaussian response model is also called "linear regression model". This assumption is partly made by convention. It can also be made because most often the quantity of interest are the β's. The variance is less of interest.

Remark **18.** The term "linear regression" means linear in the regression coefficients $\beta_0, \dots, \beta_k$ but not necessarily in $x_1, \dots, x_k$.

Example **19.** If we have $Y_i \sim N(\mu(\mathbf{x}_i), \sigma^2)$ where $\mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 e^{x_2} + \beta_4 \log(x_3)$ then it is still linear.

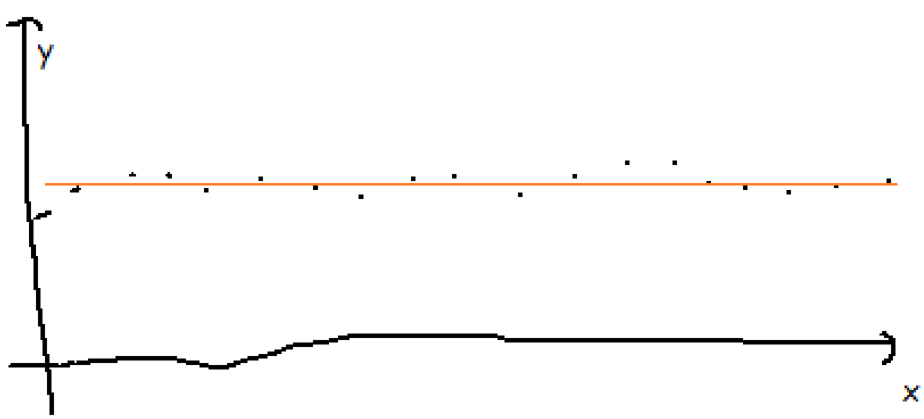
Another commonly used way to write the linear regression model is $Y_i = \mu(\mathbf{x}_i) + \varepsilon_i$ where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

Special Case: Linear regression with no covariates. We have $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, or equivalently, $Y_i = \mu + \varepsilon_i$, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. In this case, $\mu(x) \equiv \mu = \beta_0$ is just a constant.

The MLE of μ (or β_0) is $\hat{\beta}_0 = \bar{y}$. By taking the logarithm,

$$\begin{aligned}
\hat{\mu} &= \arg\max_{\mu} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\sum_{y_i-\mu}^{2}}{2\sigma^2} \right) \\
&= \arg\max_{\mu} \left[-\sum_{i=1}^{n} \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] \\
&= \arg\min_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2 .
\end{aligned}$$

So $\hat{\mu}$ is also the solution to a "least square" problem.



2016 03 02

Simple Linear Regression

We will look at the model $Y_i = \mu(x_i) + \varepsilon_i$, where $\mu(x_i) = \alpha + \beta x_i$ and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. We have three unknown quantities α, β, and σ.

The MLE of α, β, and σ is

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

The log likelihood function

$$l(\alpha, \beta, \sigma^2) = c - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i).$$

Taking derivatives we have

$$\begin{cases} \dfrac{\partial l}{\partial \alpha} & = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\ \dfrac{\partial l}{\partial \beta} & = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0 \\ \dfrac{\partial l}{\partial \sigma} & = -\dfrac{n}{\sigma} + \dfrac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = 0. \end{cases}$$

Solving yields

$$\begin{cases} \hat{\alpha} & = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} & = \dfrac{S_{x,y}}{S_{x,x}} \\ \hat{\sigma}^2 & = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}(S_{y,y} - \hat{\beta}S_{x,y}). \end{cases}$$

(Recall that

$$S_{x,y} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

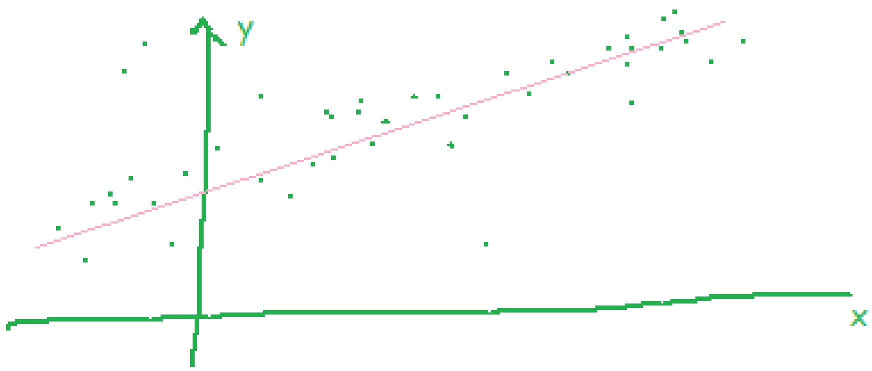
was defined long ago.) One can calculate that

$$S_{x,x} = \sum_{i=1}^{n}(x_i - \bar{x})x_i$$

by factoring out on copy of $(x_i - \bar{x})^2$ and getting a difference of two sums.

Remark **20.** Note $(\hat{\alpha}, \hat{\beta})$ actually minimizes

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



Also note that in the following, we will use

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

as an estimator of σ^2 instead of using the MLE $\hat{\sigma}^2$ because $E[S_e^2] = \sigma^2$.

Confidence Interval for β

Interpretation of β: since $E[Y|x] = \alpha + \beta x$, β can be interpreted as the average increase in y for one unit increase in x.

If $\beta = 0$, then x has no effect on y, assuming the linear regression model is correct.

Our goal is to find the distribution of $\hat{\beta}$. Now

$$\hat{\beta} = \frac{S_{x,y}}{S_{x,x}} = \sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{x,x}}y_i \equiv \sum_{i=1}^{n}a_i y_i.$$

Hence $\hat{\beta}$ is a linear combination of the y_i and thus $\hat{\beta} \sim N(\ ,\ )$. Well,

$$E(\hat{\beta}) = \sum_{i=1}^{n}a_i E(Y_i) = \sum_{i=1}^{n}a_i(\alpha + \beta x_i) = \alpha\underbrace{\sum_{i=1}^{n}a_i}_{=0} + \beta\sum_{i=1}^{n}a_i x_i = \beta\underbrace{\sum_{i=1}^{n}\frac{(x_i-\bar{x})x_i}{S_{x,x}}}_{=1} = \beta.$$

So $\hat{\beta} \sim N(\beta,\ )$. Now

$$Var(\hat{\beta}) = \sum_{i=1}^{n}a_i Var(Y_i) = \sum_{i=1}^{n}a_i^2\sigma^2 = \sigma^2\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{S_{x,x}^2} = \frac{\sigma^2}{S_{x,x}}.$$

Hence $\hat{\beta} \sim N(\beta, \sigma^2/S_{x,x})$.

The following fact will be provided on test/final if it is needed; we will not discuss proof:

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

$$\hat{\beta} \perp S_e^2 \implies \frac{\dfrac{\hat{\beta}-\beta}{\sqrt{\dfrac{\sigma^2}{S_{x,x}}}} \sim N(0,1)}{\sqrt{\dfrac{(n-2)S_e^2}{(n-2)\sigma^2}}} = \frac{\hat{\beta}-\beta}{S_e/\sqrt{S_{x,x}}} \sim t(n-2).$$

Therefore a 95% confidence interval for β is

$$\left[\hat{\beta} - t_{0.975}^{(n-2)} \frac{S_e}{\sqrt{S_{x,x}}}, \hat{\beta} + t_{0.975}^{(n-2)} \frac{S_e}{\sqrt{S_{x,x}}}\right].$$

Confidence interval for $\mu(x) = \alpha + \beta x$ at a given x

Note $\mu(x)$ is the population mean of Y for given x. The MLE of $\mu(x)$ is, by the invariance property,

$$\begin{aligned}\hat{\mu}(x) &= \hat{\alpha} + \hat{\beta}x = \bar{Y} - \hat{\beta}\bar{x} + \hat{\beta}x = \bar{Y} + \hat{\beta}(x-\bar{x}) = \bar{Y} + \frac{S_{x,y}}{S_{x,x}}(x-\bar{x}) = \frac{1}{n}\sum_{i=1}^n y_i + \sum_{i=1}^n \frac{x_i-\bar{x}}{S_{x,x}}(x-\bar{x})y_i \\ &= \sum_{i=1}^n \underbrace{\left(\frac{1}{n} + \frac{(x_i-\bar{x})(x_i-\bar{x})}{S_{x,x}}\right)}_{a_i} y_i = \sum_{i=1}^n a_i y_i.\end{aligned}$$

Now

$$E[\hat{\mu}(x)] = \sum_{i=1}^n a_i E(Y_i) = \sum_{i=1}^n a_i(\alpha + \beta x_i) = \alpha \underbrace{\sum_{i=1}^n a_i}_{=1} + \beta \sum_{i=1}^n a_i x_i = \alpha + \beta x$$

since

$$\sum_{i=1}^n a_i x_i = \sum_{i=1}^n \left(\frac{x_i}{n} + \frac{(x_i-\bar{x})x_i(x-\bar{x})}{S_{x,x}}\right) = \bar{x} + (x-\bar{x}) = x.$$

For variance, we have

$$\begin{aligned}Var(\hat{\mu}(x)) &= \sum_{i=1}^n a_i^2 Var(Y_i) = \sigma^2 \sum_{i=1}^n a_i^2 = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} + \frac{2}{n}\frac{(x_i-\bar{x})(x-\bar{x})}{S_{x,x}} + \frac{(x_i-\bar{x})^2(x-\bar{x})^2}{S_{x,x}^2}\right) \\ &= \sigma^2\left(\frac{1}{n} + 0 + \frac{(x-\bar{x})^2}{S_{x,x}}\right).\end{aligned}$$

So

$$\hat{\mu}(x) \sim N\left(\mu(x), \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{x,x}}\right)\right).$$

This, together with

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2) \qquad \text{and} \qquad \chi^2(n-2)$$

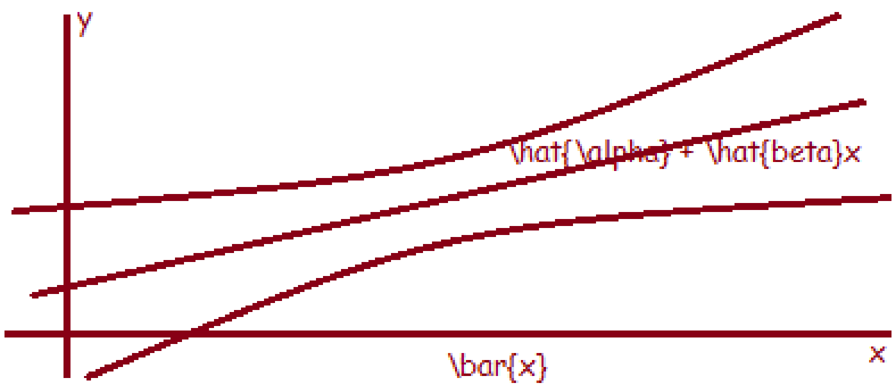
imply that

$$\frac{\dfrac{\hat{\mu}(x)-\mu(x)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{x,x}}\right)}}}{\sqrt{\dfrac{(n-2)S_e^2}{(n-2)\sigma^2}}} = \frac{\hat{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{x,x}}}} \sim t(n-2).$$

So a 95% confidence interval for $\mu(x)$ is

$$\left[\hat{\mu}(x) - t_{0.975}(n-2)S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{x,x}}}, \hat{\mu}(x) + t_{0.975}(n-2)S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{x,x}}}\right]$$

Remark **21.** The length of the 95% confidence interval for $\mu(x)$ is $2t_{0.975}(n-2)S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{x,x}}}$, and is the smallest when $x = \bar{x}$. Graphically,



By taking $x = 0$, we get a 95% confidence interval for α:

$$\left[\hat{\alpha} - t_{0.975}(n-2)S_e\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{x,x}}}, \dots\right].$$

Inference on the intercept of α is usually of less interest than inference on β.

2016 03 07

6.3.3 — Prediction Interval for an Individual Response

We have $Y_i = \mu(x_i) + \varepsilon_i$ where $\varepsilon_i \sim G(0,\sigma)$ and $\mu(x_i) = \alpha + \beta x_i$. substitute

Covariates x, responses $Y = \mu(x) + \varepsilon$, $\varepsilon \sim G(0, \sigma)$ and is independent of $Y_1, \ldots, Y_n$. Our best guess for a point prediction of Y is $\hat{\mu}(x)$.

The error is

$$Y - \hat{\mu}(x) = Y - \mu(x) + \mu(x) - \hat{\mu}(x) = \varepsilon + (\mu(x) - \hat{\mu}(x)).$$

So we understand the error as the sum of the two possible sources of error.

Since both ε and $\mu(x) - \hat{\mu}(x)$ follow Gaussian distribution, and they are independent, $Y - \hat{\mu}(x)$ follows a Gaussian distribution.

Now we calculate the expectation and variance, to find exactly which Gaussian distribution we have:

$$\begin{aligned} E(Y - \hat{\mu}(x)) &= E(\varepsilon) + E(\mu(x) - \hat{\mu}(x)) = 0, \\ Var(Y - \hat{\mu}(x)) &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{x,x}} \right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{x,x}} \right). \end{aligned}$$

Thus

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

To construct a prediction interval for Y, we use the following pivotal quantity:

$$\frac{Y - \hat{\mu}(x)}{S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{x,x}}}} = \frac{\dfrac{Y - \hat{\mu}(x)}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{x,x}}}}}{\sqrt{\dfrac{(n-2)S_e^2}{(n-2)\sigma^2}}} \sim t(n-2).$$

Therefore a $100p\%$ interval for Y is

$$\left[\hat{\mu}(x) - bS_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{x,x}}}, \cdots \right],$$

where $P(-b \le T \le b) = p$ for $T \sim t(n-2)$.

A comparison between CI for $\mu(x)$ and prediction interval for an individual response at x. Recall that the MLE of $\mu(x)$ is $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}(x)$ which has distribution

$$G\left(\mu(x), \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{x,x}}}\right).$$

Therefore a $100p\%$ CI for $\mu(x)$ is given by

$$\left[\hat{\mu}(x) - bS_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{x,x}}} \right],$$

where $P(-b \le T \le b) = p$ for $T \sim t(n-2)$.

Thus the prediction interval is always wider than the CI.
Reason: If our goal is to predict Y at x,
Since $Y \sim G(\mu(x), \sigma)$
If we know $\mu(x)$, then the variance of the error is $Var(Y) = \sigma^2$. However, we don't know $\mu(x)$, and we only have $\hat{\mu}(x)$ as its estimation. Using $\hat{\mu}(x)$ to substitute $\mu(x)$ as a prediction for Y introduces an extra error $\mu(x) - \hat{\mu}(x)$. The total error at predicting Y using $\hat{\mu}(x)$ is $Y - \hat{\mu}(x) = (Y - \mu(x)) + (\mu(x) - \hat{\mu}(x))$ and this error is quantified as

$$\begin{aligned} Var(Y - \hat{\mu}(x)) &= Var(Y - \mu(x)) + Var(\mu(x) - \hat{\mu}(x)) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{x,x}} \right) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{x,x}} \right). \end{aligned}$$

Therefore the prediction interval is wider than CI.

6.3.4 — Verifying the assumptions for the simple linear regression model

There are two main assumptions made for Gaussian Linear Regression model.

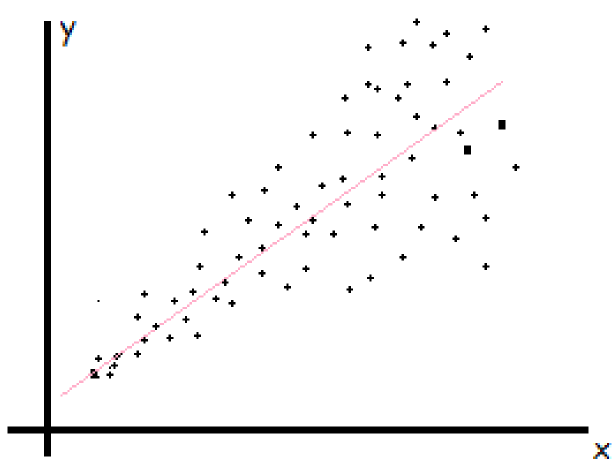
i) The error terms $\epsilon_i \overset{\text{i.i.d.}}{\sim} G(0, \sigma)$ with a constant standard deviation σ.

ii) $E(Y_i) = \mu(x_i)$ is a linear combination of the covariates with unknown coefficients.

In practice, it is important to check both of the two assumptions. We mainly focus on graphical ways of model checking.

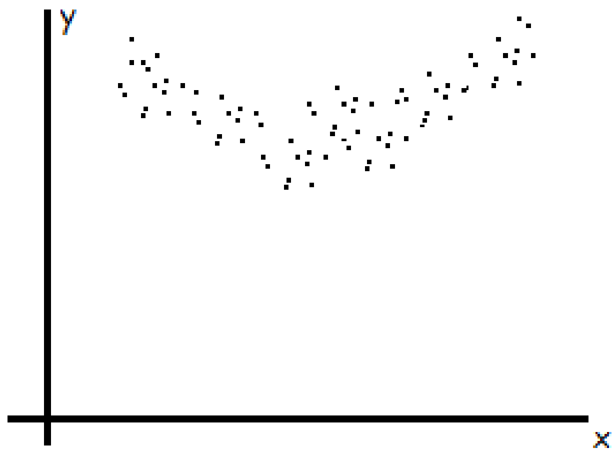
Scatter plot of (x, y) is a useful tool.

Example **22.** [1]



$E(Y)$ seems to be a linear function of x and it is reasonable to assume $\mu(x) = \alpha + \beta x$, but he assumption of constant variance may be violated, as the variance of Y increases as x increases.

EXAMPLE **23.** [2]



Then $E(Y)$ seems to be a quadratic function of x, and it is reasonable to assume $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$.

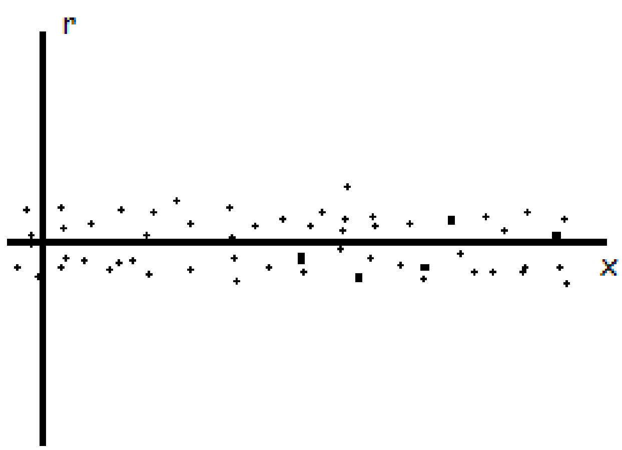
Another type of scatter plot is the "residual plot". The residual for subject i is $r_i = y_i - \hat{\mu}(x)$.

For simple linear regression, $r_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. Note that

$$\frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta} x_i) = \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0.$$

Plot the points (x_i, r_i). If our model is satisfactory, r_i should behave roughly like a random sample from $G(0, \sigma)$.

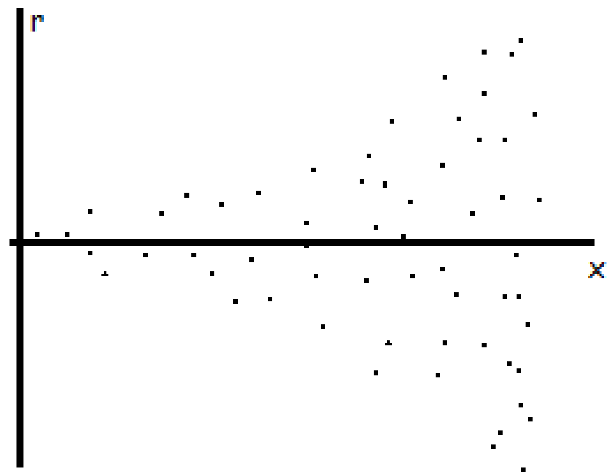
(x_i, r_i) should lie more or less horizontally within a band around the line $r = 0$, That is,



When we have multiple covariates, that is, $\mu(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, then we can plot $(\hat{\mu}(x_i), r_i)$.

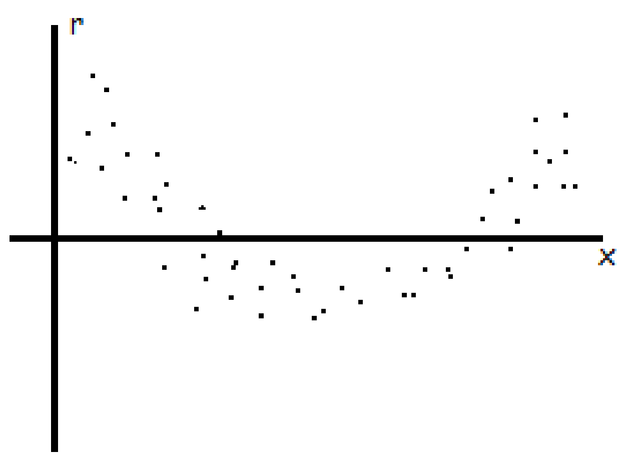
Departure from the above pattern suggests problems (with?) the model.

Example **24.** For example,



indicates that $Var(Y_i)$ may not be a constant, but may depend on x.

Example **25.**



indicates that $\mu(x)$ is not a linear function of x.

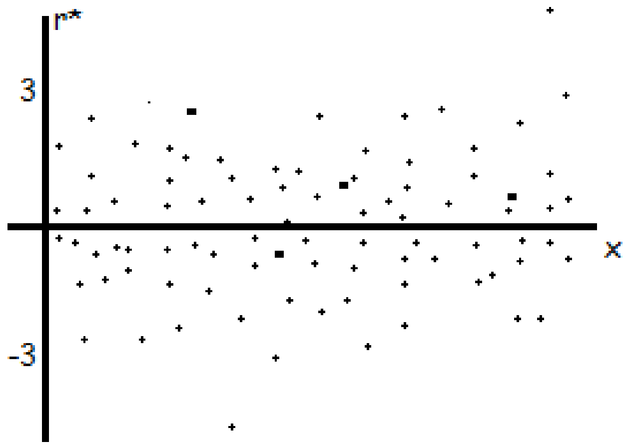
We can define standardized residual:

$$r_i^* = \frac{r_i}{S_e} = \frac{y_i - \hat{\alpha} - \hat{\beta} x_i}{S_e}$$

for $i \in \{1, \ldots, n\}$, and make plots using r_i^* instead of r_i.

The patterns of the plots are unchanged, but the r_i^* values tend to lie in the range $(-3, 3)$. Reason: if the model is satisfactory, then the r_i's are roughly a random sample from $G(0, \sigma)$

and S_e is an estimate of σ. So 95% of r_i^* values should be in $(-2, 2)$, 99.7% of 4_i^* values should be in $(-3, 3)$.



REMARK **26.** 1. QQ plot of r_i or r_i^* may be used to check the Gaussian distribution assumption

2. Most plots in practice do not have clear patterns as in the examples. Reading these plots is something of an art, and we should not over-read them.

2016 03 09
TA

Comparing two Poisson means

We have

$$Y_{11}, \ldots, Y_{1n_1} \sim \text{Poi}(\mu_1),$$
$$T_{21}, \ldots, Y_{2n_2} \sim \text{Poi}(\mu_2).$$

Our null hypothesis is $\mu_1 = \mu_2$. The likelihood function is

$$L = \prod_{i=1}^{n_1} \frac{\mu_1^{Y_{1i}} e^{-\mu_1}}{y_{1i}!} \prod_{j=1}^{n_2} \frac{\mu_2^{Y_{2j}} e^{-\mu_2}}{y_{2j}!}.$$

The log likelihood function is found by taking the logarithm (and ignoring constants if you want because they don't change were the maximum occurs).

Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu$, we take the derivative with respect to μ to maximize and find

$$\hat{\mu} = \frac{\sum_{i=1}^{n_1} Y_{1i} + \sum_{j=1}^{n_2} Y_{2j}}{n_1 + n_2}.$$

Under $H_1 : \mu_1 \neq \mu_2$, we must take the derivative with respect to both μ_1 and μ_2. Solving yields

$$\widehat{\mu}_1 = \overline{y}_1,$$
$$\widehat{\mu}_2 = \overline{y}_2.$$

And it went on for a bit...

Comparison using paired data 2016 03 14

Often times experimental studies comparing difference in population means are conducted using pairs of units. Say we have

$$Y_{1,1}, Y_{1,2}, \dots, Y_{1,n} \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2),$$
$$Y_{2,1}, Y_{2,2}, \dots, Y_{2,n} \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma_2^2).$$

Then $Y_{1,i}$ and $Y_{2,i}$ are not independent. So the previous method cannot be used. How do we construct a 95% CI for $\mu_1 - \mu_2$?

We may assume that the pairs $(Y_{1,i}, Y_{2,i}) \overset{\text{i.i.d.}}{\sim}$ Bivaraite Normal Distribution (which we have not covered). It can be shown that

$$Y_{1,i} - Y_{2,i} \overset{\text{i.i.d.}}{\sim} N(\mu_1 - \mu_2, \sigma^2)$$
$$\sigma^2 = Var(Y_{1,i} - Y_{2,i}) = Var\,(Y_{i,1}) + Var(Y_{2,i}) - 2Cov(Y_{1,i}, Y_{2,i}).$$

It looks like σ^2 depends on i, but it doesn't actually in the end.

Therefore, if our interest is inference about $\mu_1 - \mu_2$, we can use the data

$$X_1 \equiv Y_{1,1} - Y_{2,1}, \dots, X_n \equiv Y_{1,n} - Y_{2,n}.$$

We will have $X_i \overset{\text{i.i.d.}}{\sim} N(\mu_1 - \mu_2, \sigma^2)$. This reduces the problem to a one-sample problem.

When $Y_{1,i}$ and $Y_{2,i}$ are positively correlated, using the paired data increases the precision of estimating $\mu_1 - \mu_2$. This is because

$$Var(\overline{Y}_1 - \overline{Y}_2) = Var(\overline{Y}_1) + Var(\overline{Y}_2) - 2Cov(\overline{Y}_1, \overline{Y}_2) \leq Var(\overline{Y}_1) + Var(\overline{Y}_2).$$

6.4 — More general Gaussian response models

Our response Y_i now depends on more covariates

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p c_{i,p} + \varepsilon_i,$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. We can write this in vector/matrix notations as

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i.$$

Here,

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

With

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix},$$

we have $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \text{MVN}\left(\mathbf{0}, \sigma^2 I\right)$.

The MLE for $\boldsymbol{\beta}$ and σ^2 is

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2}{2\sigma^2}\right).$$

We can take the logarithm to get the log likelihood function:

$$l(\boldsymbol{\beta}, \sigma^2) = c - n\log\left(\sigma\right) - \frac{\left(\mathbf{Y} - X\boldsymbol{\beta}\right)^T \left(\mathbf{Y} - X\boldsymbol{\beta}\right)}{2\sigma^2}.$$

Taking derivatives and setting them equal to zero, we would find that

$$\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}$$
$$\widehat{\sigma}^2 = \frac{1}{n}\left(\mathbf{Y} - X\boldsymbol{\beta}\right)^T \left(\mathbf{Y} - X\boldsymbol{\beta}\right),$$

assuming $X^T X$ is invertible.

Define

$$S_e^2 = \frac{n\widehat{\sigma}^2}{n - (p+1)}.$$

Note that this is an unbiased estimator on σ^2, as $E(S_e^2) = \sigma^2$.

What is the distribution of $\widehat{\beta}$? Well,

$$E(\widehat{\beta}) = (X^T X)^{-1} X^T E(\mathbf{Y}) = (X^T X)^{-1} X^T X\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Note that it is a fact that

$$Var(A\mathbf{Y}) = AVar(\mathbf{Y})A^T.$$

In particular, $Var(c\mathbf{Y}) = c^2 Var(\mathbf{Y})$. Actually, maybe we should define variance and expectation of vectors. That might be helpful. They are defined as

$$Var(\mathbf{Y}) = E\left(\left(\mathbf{Y} - E\left(\mathbf{Y}\right)\right)\left(\mathbf{Y} - E\left(\mathbf{Y}\right)\right)^T\right),$$
$$E(\mathbf{Y}) = \begin{bmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{bmatrix}.$$

Therefore we can calculate

$$Var(\widehat{\beta}) = (X^T X)^{-1} X^T Var(\mathbf{Y}) X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Thus we have

$$\widehat{\beta} \sim \text{MVN}\left(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1}\right).$$

Remark **27.** [$p = 1$ case] This is just the case we already saw in previous subsections. We just have

$$\begin{bmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}, \sigma^2 \underbrace{\left(\begin{bmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{bmatrix}\begin{bmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}\right)^{-1}}_{\frac{1}{\sum_{i=1}^n x_i^2 - (n\overline{x})^2}\begin{bmatrix}\sum_{i=1}^n x_i^2 & -b\overline{x}\\ -n\overline{x} & n\end{bmatrix}}\right)$$

Go check that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

It can be shown that

$$W = \frac{(n-(p+1))S_e^2}{\sigma^2} \sim \chi^2(n-(p+1))$$

and $\boldsymbol{\beta} \perp W$. Based on these results we have

$$\frac{\widehat{\beta}_j - \beta_j}{S_e\sqrt{C_{(j+1),(j+1)}}} \sim t(n-(p+1)),$$

where $C_{(j+1),(j+1)}$ is the $(j+1),(j+1)$th entry of $(X^TX)^{-1}$.

Using this, we can give a 95% CI for β_j:

$$\left[\widehat{\beta}_j \pm t_{0.975}(n+(p-1))S_e\sqrt{C_{(j+1),(j+1)}}\right].$$

Midterm II

2016 03 16

Chapter 7 - Tests and Inference Problems Based on Multinomial Distribution

7.1 - General Theory

You will not be tested on matrix notation on the final exam. There is no new theory in this chapter, supposedly.

Suppose that our data is $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim \text{Multinomial}\,(n; \boldsymbol{\theta})$ with probability mass function (discrete version of pdf (probability density function))

$$P(Y_1 = y_1, \ldots, Y_k = y_k; \theta_1, \ldots, \theta_k) = f(\mathbf{y}; \boldsymbol{\theta}) = \frac{n!}{y_1! \cdots y_k!} \theta_1^{y_1} \cdots \theta_k^{y_k},$$

where $y_j = 0, \ldots, n$ and $\sum_{i=1}^{n} y_j = n$.

Suppose now we suspect that $\boldsymbol{\theta}$ depends on a lower dimensional parameter $\boldsymbol{\alpha}$ and wish to test $H_0 : \theta_j = \theta_j(\boldsymbol{\alpha})$ for $j = 1, \ldots, k$ where $\dim(\boldsymbol{\alpha}) = p < k - 1$. We use the likelihood ratio statistic to test H_0. The likelihood function is $L(\boldsymbol{\theta}) = c\theta_1^{y_1} \cdots \theta_k^{y_k}$. Maximizing $L(\boldsymbol{\theta})$ subject to $\sum_{k=1}^{n} \theta_j = 1$ yields the MLE of θ_j:

$$\widehat{\theta}_j = \frac{y_j}{n}, \quad j = 1, \ldots, k.$$

Now under H_0,

$$L(\boldsymbol{\alpha}) = c \prod_{i=1}^{k} \theta_j(\boldsymbol{\alpha})^{y_j}.$$

Maximizing $L(\boldsymbol{\alpha})$ leads to the MLE $\hat{\alpha}$, and therefore the MLE of θ_j under H_0 is $\theta_j(\widehat{\boldsymbol{\alpha}})$.

The likelihood ratio statistic is

$$\begin{aligned}
\Lambda &= -2 \log \left(\frac{L(\boldsymbol{\theta}(\widehat{\boldsymbol{\alpha}}))}{L(\widehat{\boldsymbol{\theta}})} \right) \\
&= 2 \log \left(L(\widehat{\theta}) \right) - 2 \log \left(L(\boldsymbol{\theta}(\widehat{\boldsymbol{\alpha}})) \right) \\
&= 2 \left(\sum_{j=1}^{k} Y_j \log(\widehat{\theta}_j) - \sum_{j=1}^{k} Y_j \log(\theta_j(\widehat{\boldsymbol{\alpha}})) \right) \\
&= 2 \sum_{j=1}^{k} Y_j \log \left(\frac{Y_j}{E_j} \right),
\end{aligned}$$

where $E_j = n\theta_j(\widehat{\boldsymbol{\alpha}})$. Note E_j can be viewed as the "expected frequency" of the jth outcome under H_0. Under H_0, $\Lambda \underbrace{}^{\text{approximately}} \chi^2(k + 1 - p)$. Then the p-value is

$$P(\Lambda \geq \lambda \,|\, H_0) \approx P(W \geq \lambda),$$

where $W \sim \chi^2(k - 1 - p)$.

Remark 28.

1. $\log\left(\frac{Y_j}{E_j}\right)$ quantifies the difference between the observed data and the "expected data" (if H_0 is true). So if Λ is very large, H_0 is unlikely to be true.
2. An alternative test statistic is the Pearson goodness-of-fit statistic:

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j}.$$

It can be shown that $D \overset{\text{app}}{\sim} \chi^2(k - 1 - p)$ under H_0 when n is large.

7.2 - Examples on Testing Goodness-of-fit

Suppose $\mathbf{Y} = (Y_1, Y_2, Y_3) \sim \text{Multinomial}(n; \theta_1, \theta_2, \theta_3)$. The observed data is $n = 100$, $y_1 = 17$, $y_2 = 46$, $y_3 = 37$. We want to test that $H_0 : \theta_1 = \alpha^2$, $\theta_2 = 2\alpha(1-\alpha)$, $\theta_3 = (1-\alpha)^2$. Under H_0,

$$L(\alpha) = C\theta_1(\alpha)^{y_1}\theta_2(\alpha)^{y_2}\theta_3(\alpha)^{y_3} = C\alpha^{80}(1-\alpha)^{120}.$$

Maximizing $L(\alpha)$ yields $\widehat{\alpha} = 0.40$. Therefore $e_1 = n\theta_1(\widehat{\alpha}) = n\widehat{\alpha}^2 = 16$, $e_2 = \cdots = 48$, $e_3 = \cdots = 36$. Therefore

$$\lambda = s\sum_{j=1}^{3} y_j \log\left(\frac{y_j}{e_j}\right) = 0.17.$$

The p-value is

$$P(\Lambda \geq 0.17 \,|\, H_0) \approx P(W \geq 0.17) = 0.68 > 0.05$$

where $W \sim \chi^2(1)$. So we fail to reject H_0.

Example **29.** [7.2.2 - Goodness-of-fit of an exponential model] Suppose that an Exponential distribution is assumed for a random variable T and a random sample $t_1, \ldots, t_n$ is collected. We wish to test $H_0 : f(t, \alpha) = \frac{1}{\alpha}e^{-\frac{t}{\alpha}}$.

We can check graphically whether or not the data follows the model we want to assume.

To test the null hypothesis H_0, we partition the support of T into k intervals:

$$[0, x_1), [x_1, x_2), \ldots, [x_{k-1}, \infty).$$

Let $Y_j \equiv \#$ of subjects which fall into the jth interval, and let $p_j \equiv P(T \in \text{the } j\text{th interval})$. Then $\mathbf{Y} = (Y_1, \ldots, Y_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k)$. Under H_0,

$$P_j = P_j(\alpha) = \int_{x_{j-1}}^{x_j} \frac{1}{\alpha}e^{-\frac{t}{\alpha}}\, dt.$$

Now suppose $n = 100$, and we partition $[0, \infty)$ into 7 intervals:

$$[0, 100), [100, 200), [200, 300), [300, 400), [400, 600), [600, 800), [800, \infty).$$

And $y_1 = 29$, $y_2 = 22$, $y_3 = 12$, $y_4 = 10$, $y_5 = 10$, $y_6 = 9$, $y_7 = 8$.

There were some calculations and then we got a p-value of 0.68.

2016 03 21 Midterm Return

We have $\mathbf{Y} = (Y_1, \ldots, Y_k) \sim \text{Multinomial}(n; \boldsymbol{\theta})$. Our null hypothesis is $H_0 : \theta_j = \theta_j(\boldsymbol{\alpha})$, $j \in \{1, \ldots, k\}$, $\dim(\boldsymbol{\alpha}) < k-1$. In previous lecture we saw the likelihood ratio is

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

where $E_j = n\theta_j(\hat{\alpha})$.

7.3 – Two-Way Tables

7.3.1 – Testing for Independence of Two Variables

We wish to test whether two categorical Y random variables A and B are independent.

EXAMPLE **30.** A =smoking, B =lung cancer

We will consider the case where A and B take on a fairly small number of possible values. Suppose that, for A, there are a mutually exclusive types $A_1, \ldots, A_a$. Suppose that, for B, there are b mutually exclusive types $B_1, \ldots, B_b$. Assume $a, b \geq 2$. Let $\theta_{i,j}$ be the probability that a randomly selected subject is of type (A_i, B_j), i.e. $\theta_{i,j} = P(A_i \cap B_j)$.

	B_1	B_2	$\cdots$	B_b
A_1	$\theta_{1,1}$	$\theta_{1,2}$	$\cdots$	$\theta_{1,b}$
A_2	$\theta_{2,1}$	$\theta_{2,2}$	$\cdots$	$\theta_{2,b}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_a	$\theta_{a,1}$	$\theta_{a,2}$	$\cdots$	$\theta_{a,b}$

For a random sample with size n, let $Y_{i,j}$ be the number of units that are of type (A_i, B_j). Then $\mathbf{Y} = (Y_{1,1}, Y_{1,2}, \ldots, Y_{i,j}, \ldots, Y_{a,b}) \sim \text{Multinomial}(n; \theta_{1,1}, \theta_{1,2}, \ldots, \theta_{i,j}, \ldots, \theta_{a,b})$, where

$$\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{i,j} = n, \qquad \sum_{i=1}^{a}\sum_{j=1}^{b} \theta_{i,j} = 1.$$

Let

$$\alpha_i = P(\text{a subject is of type } A_i) = \sum_{j=1}^{b} \theta_{i,j},$$

$$\beta_i = P(\text{a subject is of type } B_j) = \sum_{i=1}^{a} \theta_{i,j}.$$

	B_1	B_2	$\cdots$	B_b	total
A_1	$\theta_{1,1}$	$\theta_{1,2}$	$\cdots$	$\theta_{1,b}$	α_1
A_2	$\theta_{2,1}$	$\theta_{2,2}$	$\cdots$	$\theta_{2,b}$	α_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	$\theta_{a,1}$	$\theta_{a,2}$	$\cdots$	$\theta_{a,b}$	α_a
total	β_1	β_2	$\cdots$	β_b	1

The independence of A and B is equivalent to $\theta_{i,j} = \alpha_i\beta_j$ for all i, j. Thus $H_0 : \theta_{i,j} = \alpha_i\beta_j$, $i \in \{1, \ldots, a\}, j \in \{1, \ldots, b\}$.

The likelihood ratio is

$$\Lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{i,j} \log\left(\frac{Y_{i,j}}{E_{i,j}}\right),$$

where $E_{i,j} = n\theta_{i,j}(\hat{\alpha}, \hat{\beta})$. Under the null hypothesis H_0 we can compute

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = C \prod_{i=1}^{a} \prod_{j=1}^{b} \theta_{i,j}(\hat{\alpha}, \hat{\beta})^{Y_{i,j}} = C \left(\prod_{i=1}^{a} \alpha_i^{Y_{i,+}} \right) \left(\prod_{j=1}^{b} \beta_j^{Y_{+,j}} \right)$$

where

$$Y_{i,+} = \sum_{j=1}^{b} Y_{i,j}, \qquad Y_{+,j} = \sum_{i=1}^{a} Y_{i,j}.$$

So we want to maximize $L(\hat{\alpha}, \hat{\beta})$ subject to $\sum_{i=1}^{a} \alpha_i = 1$ and $\sum_{j=1}^{b} \beta_j = 1$. This will give

$$\hat{\alpha}_i = \frac{Y_{i,+}}{n}, \qquad \hat{\beta}_j = \frac{Y_{+,j}}{n}.$$

	B_1	B_2	$\cdots$	B_b	total
A_1	$Y_{1,1}$	$Y_{1,2}$	$\cdots$	$Y_{1,b}$	$Y_{1,+}$
A_2	$Y_{2,1}$	$Y_{2,2}$	$\cdots$	$Y_{2,b}$	$Y_{2,+}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	$Y_{a,1}$	$Y_{a,2}$	$\cdots$	$Y_{a,b}$	$Y_{a,+}$
total	$Y_{+,1}$	$Y_{+,2}$	$\cdots$	$Y_{+,b}$	n

So

$$E_{i,j} = n\theta_{i,j}(\hat{\alpha}, \hat{\beta}) = n\hat{\alpha}_i\hat{\beta}_j = n\frac{Y_{i,+}}{n}\frac{Y_{+,j}}{n} = \frac{Y_{i,+}Y_{+,j}}{n}.$$

Under H_0, $\Lambda \sim \chi^2((a-1)(b-1))$. The p-value is $P(\Lambda \geq \lambda \mid H_0)$.

2016 03 23

7.3.2 Testing for Homogeneity of Multiple Groups

Suppose the whole population is divided into a sub-populations $A_1, \ldots, A_a$ and each unit in the population is one of the b types $B_1, \ldots, B_b$.

Remark **31.** Independence and Homogeneity are mathematically the same procedure, but the two problems and their interpretations are different. The course notes don't really distinguish between the two.

Let $\theta_{i,j} \equiv P(\text{a unit from sub-population } i \text{ is of type } j) = P(B_j | A_i)$. Let $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,2}, \ldots, \theta_{i,b})$. We wish to test the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_a \equiv \boldsymbol{\theta}$. That is, the proportions of units of types $B_1, B_2, \ldots, B_b$ are the same for each sub-population.

Example **32.** We wish to test whether the proportions of different age groups are the same across different countries.

For each group i, suppose we collect n_i units. Among them there are $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,b}$ units that are of types $B_1, B_2, \ldots, B_b$ respectively. Let $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,b})$. Therefore $\mathbf{Y}_i \sim \text{Multinomial}(n_i; \theta_{i,1}, \ldots, \theta_{i,b})$ where

$$\sum_{j=1}^{b} Y_{i,j} = n_i, \qquad \sum_{j=1}^{b} \theta_{i,j} = 1.$$

	B_1	B_2	$\cdots$	B_b	total
A_1	$\theta_{1,1}$	$\theta_{1,2}$	$\cdots$	$\theta_{1,b}$	1
A_2	$\theta_{2,1}$	$\theta_{2,2}$	$\cdots$	$\theta_{2,b}$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	$\theta_{a,1}$	$\theta_{a,2}$	$\cdots$	$\theta_{a,b}$	1

	B_1	B_2	$\cdots$	B_b	total
A_1	$Y_{1,1}$	$Y_{1,2}$	$\cdots$	$Y_{1,b}$	n_1
A_2	$Y_{2,1}$	$Y_{2,2}$	$\cdots$	$Y_{2,b}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	$Y_{a,1}$	$Y_{a,2}$	$\cdots$	$Y_{a,b}$	n_b

(Note that in the following, when we maximize to get MLEs, we have the constraints that some things sum to one and such.) We now have a Multinomial distributions, one for each sub-population. The joint likelihood function is

$$L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_a) = \prod_{i=1}^{a} \left(\frac{n_i!}{Y_{i,1}! \cdots Y_{i,b}!} \prod_{j=1}^{b} \theta_{i,j}^{Y_{i,j}} \right) = c \cdot \prod_{i=1}^{a} \prod_{j=1}^{b} \theta_{i,j}^{Y_{i,j}}.$$

One can calculate that the MLE is

$$\hat{\theta}_{i,j} = \frac{Y_{i,j}}{n_i}.$$

Under the null hypothesis H_0

$$L(\boldsymbol{\theta}) = c \cdot \prod_{i=1}^{a} \prod_{j=1}^{b} \theta_j^{Y_{i,j}} = \prod_{j=1}^{b} \theta_j^{Y_{+,j}},$$

and hence we can calculate the MLE to be

$$\hat{\theta}_j = \frac{Y_{+,j}}{\sum_{j=1}^{b} Y_{+,j}} = \frac{Y_{+,j}}{n}.$$

It can be shown that (exercise)

$$\Lambda = 2 \sum_{i=1}^{a} \sum_{j=1}^{b} Y_{i,j} \log\left(\frac{Y_{i,j}}{E_{i,j}} \right)$$

where

$$E_{i,j} = n_i \frac{Y_{+,j}}{n}.$$

So under H_0, $\Lambda \overset{\text{app}}{\sim} \chi^2((a-1)(b-1))$. The p-value is $P(\Lambda \geq \lambda \mid H_0) = P(W \geq \lambda)$ where $W \sim \chi^2((a-1)(b-1))$.

The final result is the same as the previous thing we did, but the steps to get there were different.

(The previous table can actually be:

	B_1	B_2	$\cdots$	B_b	total
A_1	$Y_{1,1}$	$Y_{1,2}$	$\cdots$	$Y_{1,b}$	$n_1 = Y_{1,+}$
A_2	$Y_{2,1}$	$Y_{2,2}$	$\cdots$	$Y_{2,b}$	$n_2 = Y_{2,+}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	$Y_{a,1}$	$Y_{a,2}$	$\cdots$	$Y_{a,b}$	$n_b = Y_{a,+}$
total	$Y_{+,1}$	$Y_{+,2}$	$\cdots$	$Y_{+,b}$	n

)

REMARK **33.** 1. $H_0 : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_a$ means that $\boldsymbol{\theta}_i$ doesn't depend on i; that is, $P(B_j \mid A_i) = P(B_j)$, which essentially means independence.

2. For both testing problems, we can follow the same procedure to calculate the p-value:

   i) lay out data in the two-way table;

   ii) "expected frequencies" $e_{i,j}$ under H_0:

   $$e_{i,j} = \frac{Y_{i,+}Y_{+,j}}{n}$$

   (I don't think this is a definition of H_0.);

   iii)

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

   iv) the p-value is approximately $P(W \geq \lambda)$ where $W \sim \chi^2((a-1)(b-1))$.

EXAMPLE **34.** [EXAMPLE 7.3.1] $n = 300$

	O	A	B	AB	total
Rh+	82	89	54	19	244
Rh-	13	27	7	9	56
total	95	116	61	28	300

We can calculate

$$e_{i,j} = \frac{y_{i,+}y_{+,j}}{n}, \qquad \text{e.g. } e_{1,1} \approx 77.3$$

$$\lambda = 2\sum_{i=1}^{2}\sum_{j=1}^{4} y_{i,j} \log\left(\frac{y_{i,j}}{e_{i,j}}\right) = 8.52$$

$$p\text{-value } = P(\Lambda \geq 8.52 \mid H_0) \approx P(W \geq 8.52) = 0.036 < 0.05$$

where $W \sim \chi^2(3)$.

2016 03 28

Review of Course

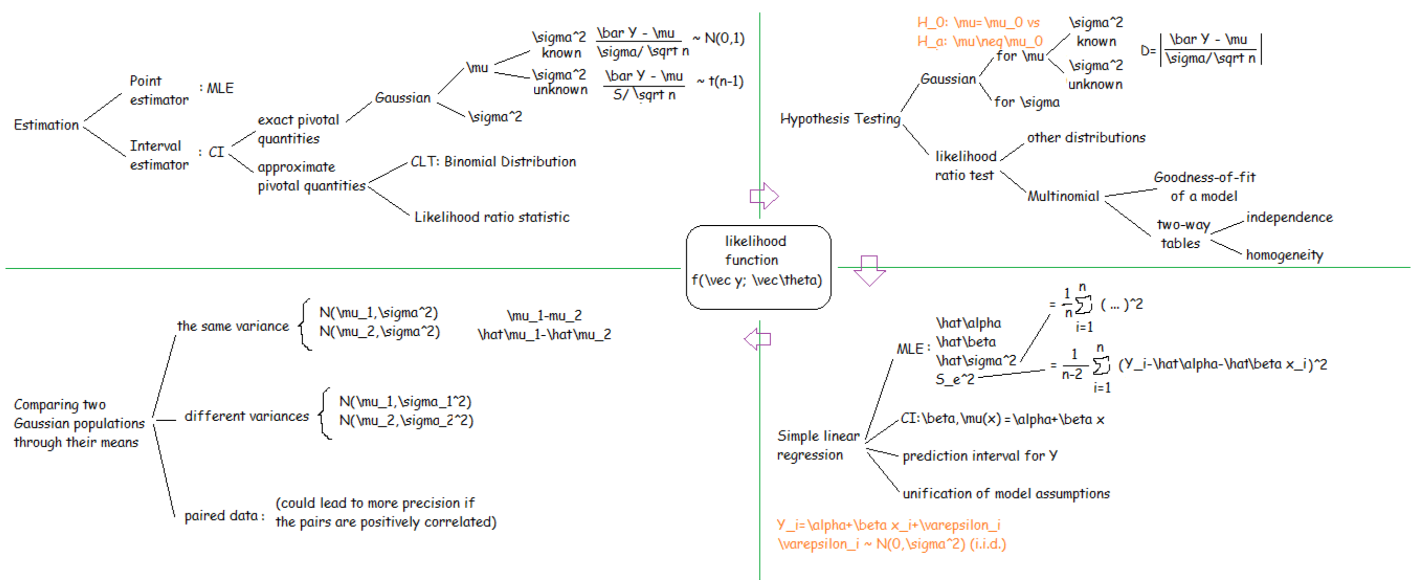
Almost everything is build on likelihood.

Problem with point estimator is that two different data sets give different estimates. Naturally leads to interval estimator; deals with uncertainty.

There are different estimators other than *maximum likelihood* estimators, but MLE is the most widely used due to it's nice properties (it is consistent when the data set is large, it has the least variation). We didn't take about properties, just how to find/derive/check.

Only for nice/special distribution can we find exact pivotal quantities. For example, we can actually do it for Gaussian.

More comments were said after this point, but I was unable to write them down and copy the digram.



Then we did some problems/exercises from the course notes.

Some discussion/problems/something about chapter 7.

2016 03 30

Chapter 8: Causal Relationships

It is difficult to formally define what a causal relationship is. One idealized definition is as follows: If all other factors affecting Y are held constant and if the distribution of Y changes with the change of factor X, then we say X has causal effect on Y. Problem: in a study, we don't even know what all the factors are, so how can we hold them all constant?

REMARK **35.** It is relatively easy to study causal effect in experimental studies, but difficult to study causal effect in experimental studies.

EXAMPLE **36.** [8.3.1] The data of applications and admissions to graduate studies in Engineering and Arts faculties in a university over the past five years are available.

	# Applied	# Admitted	% Admitted	
Engineering	1000	600	60%	Men
	200	150	75%	Women
Arts	1000	400	40%	Men
	1000	800	44%	Women
Total	2000	1000	50%	Men
	2000	950	47.5%	Women

The above feature (men appearing to have better chances than women overall, even though when you break it down this is clearly not the case) is called Simpson's paradox.

Mathematically, we may have $P(A \mid B_1 C_i) > P(A \mid B_2 C_i)$ for $i \in \{1, \dots, k\}$ but $P(A \mid B_1) < P(A \mid B_2)$, because

$$P(A \mid B_1) = \sum_{i=1}^{k} P(A \mid B_1 C_i) P(C_i \mid B_1),$$
$$P(A \mid B_2) = \sum_{i=1}^{k} P(A \mid B_2 C_i) P(C_i \mid B_2).$$