Announcements

Wednesday, February 29, 2012
1:25 PM

Feb 29

- Please submit the Self-evaluation form by Friday (Mar 2).
- The form is posted on the course website and will be distributed in class (today).

- You can take a look at your midterm later today.

Mar 2

- Please submit the self-evaluation form by the end of today.
  (5:20 PM)
- Note Part I.7 is @ the copy center and website.

Mar 9

- HW 5 is posted : Due Mar 16 (Fri)

Mar 14

-
- Note part II.1 is now available.

Mar 16

- Submit HW 5 by 5PM today
- Note Part II.2 is now available.

Mar 23

Wed.

Mar 23

- HW 6 (Mar 28) wed.
- Note Part II.3 and Part II.4 are available.
- Start preparing the $\underbrace{\text{A4 sheet}}_{\text{2-sided handwritten}}$ for the final exam.

Mar 28

- HW 7 and its solution will be posted later today.
- Note Part II.5 is available.
- There will be eval. form posted on the web site.

Mar 30

- Check the course website regularly (or at least before the final exam day)

9 Multiple RVs

Wednesday, February 29, 2012
1:25 PM

Previously, we talked about mean (average)
variance
standard deviation

of one set of data.

Now, we want to extend the calculation to > 1 sets of data.

Correlation
Covariance
correlation coefficient

One big example...

Suppose we have 10 students in a class...

student id	Midterm score x	Final score y
1	10	40
2	10	40
3	10	50
4	20	40
5	20	40
6	20	50
7	20	50
8	20	50
9	20	50
10	20	50

To summarize these data, use a table

$x \backslash y$	40	50
10	2	1
20	2	5

Randomly select one of the students

X = his/her midterm score

Y = his/her final score

What is the probability that $X = 10$ and $Y = 40$?

$$P[X = 10 \text{ and } Y = 40] = \frac{2}{10}$$

$$\equiv$$

$$p_{X,Y}(10, 40)$$

Defn $p_{X,Y}(x,y) = P[X = x \text{ and } Y = y]$

joint pmf

$p_{X,Y}(10, 20) = 0$

$p_{X,Y}(x,y)$

$x \backslash y$	40	50

$p_{X,Y}(10,20) = 0$

$x \backslash y$	40	50
10	1/5	1/10
20	1/5	1/2

What if my interest is only on the midterm score?

For example, randomly select a student

what is the probability that his/her midterm score = 10 ?

$$p_X(10) = P[X=10] = \frac{2+1}{10} = \frac{3}{10}.$$

same

$$= \frac{2}{10} + \frac{1}{10} = \frac{3}{10}$$

$$= p_{X,Y}(10,40) + p_{X,Y}(10,50)$$

$$= \sum_y p_{X,Y}(10,y)$$

$$p_X(20) = P[X=20] = \sum_y p_{X,Y}(20,y) = \frac{2}{10} + \frac{5}{10} = \frac{7}{10}$$

Formula :

$$\left.\begin{aligned} p_X(x) &= \sum_y p_{X,Y}(x,y) \\ p_Y(y) &= \sum_x p_{X,Y}(x,y) \end{aligned}\right\} \text{marginal pmf's.}$$

$$p_X(x) = \begin{cases} 3/10, & x = 10 \\ 7/10, & x = 20 \\ 0, & \text{otherwise} \end{cases} \qquad p_Y(y) = \begin{cases} 2/5, & y = 40 \\ 3/5, & y = 50 \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbb{E}X = \sum_x x\, p_X(x) = \frac{3}{1\cancel{0}} \times 1\cancel{0} + \frac{7}{1\cancel{0}} \times 2\cancel{0} = 17 \qquad \mathbb{E}Y = \sum_y y\, p_Y(y) = \frac{2}{5}\times 40 + \frac{3}{5}\times 50 = 46$$

Given that $X = 10$, what is the probability that $Y=50$?

$$= \frac{1}{3}$$

$$P[\overbrace{Y = 50}^{A} \mid \overbrace{X=10}^{B}] = \frac{P(A\cap B)}{P(B)} = \frac{P[X=10 \text{ and } Y=50]}{P[X=10]}$$

$$\parallel\!\!\!/$$

$$p_{Y|X}(50|10) = \frac{p_{X,Y}(10,50)}{p_X(10)} = \frac{1/10}{3/10} = \frac{1}{3}$$

Formula:

$$p_{Y|X}(y|x) \equiv P[Y=y \mid X=x] = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

conditional pmf

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

To calculate the total score, we use

midterm = 40% of total score

final = 40% of total score

$$t = \frac{x}{20} \times 40 + \frac{y}{50} \times 40 + 20$$

20: max midterm score

50: max final score

20: HW/Quiz/participation

t

$x \backslash y$	40	50
10	72	80
20	92	100

Let T be the total score of a randomly selected student.

$$p_T(t) = \begin{cases} 1/5 & t = 72, \\ 1/10 & t = 80, \\ 1/5 & t = 92, \\ 1/2 & t = 100, \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{E}T = \sum_t t\, p_T(t) = 90.8.$$

Alternatively,

$$\mathbb{E}T = \mathbb{E}\left[\frac{X}{20}40 + \frac{Y}{50}40 + 20\right] = \mathbb{E}\left[2X + \frac{4}{5}Y + 20\right]$$

$$= 2\,\mathbb{E}X + \frac{4}{5}\mathbb{E}Y + 20$$

$$= 2(17) + \frac{4}{5}(46) + 20 = 90.8.$$

In general, if $T = g(X,Y)$

$$\mathbb{E}T = \mathbb{E}[g(X,Y)] = \sum_x \sum_y g(x,y)\, p_{X,Y}(x,y)$$

9 Least Square

Friday, March 02, 2012
11:33 AM

Earlier in our class

Basic statistics ...

Suppose we have n observations : $x_1, x_2, x_3, x_4, \ldots, x_n$

Find "b" to represent this data set.

To determine how good "b" is,
we will calculate

$$\sum_{i=1}^{n} (x_i - b)^2$$

and minimize it.

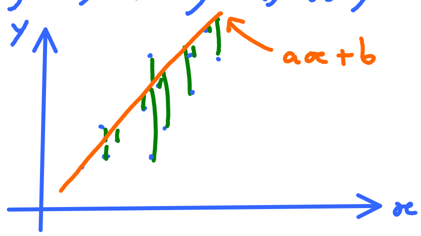
$\xrightarrow[\text{calculus}]{}$ the best "b" $= \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$

calculus

$\frac{d}{db} = 0$

Suppose that we have n pairs of observations

$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$



Find the "best" straight line that "fits" the observations.

Again, want to minimize square error

$$\sum_{i=1}^{n} \left(y_i - (a x_i + b) \right)^2$$

IEX $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

calculus

$\frac{\partial}{\partial a} = 0$

$\frac{\partial}{\partial b} = 0$

$$a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$\mathbb{E}Y \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$b = \bar{y} - a\bar{x}$$

$\mathbb{E}[X^2] \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

$\underbrace{\mathbb{E}[XY]}_{\text{correlation}} \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$

$$a = \frac{\mathbb{E}[XY] - \mathbb{E}X\,\mathbb{E}Y}{\mathbb{E}[X^2] - (\mathbb{E}X)^2} = \frac{\operatorname{cov}(X,Y)}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$$

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} \Rightarrow \operatorname{cov}(X,Y) = \rho\,\sigma_X\sigma_Y$$

$$y \approx ax + b$$

$$\begin{aligned} Y \approx aX + b &= aX + (\mathbb{E}Y - a\,\mathbb{E}X) \\ &= \mathbb{E}Y + a(X - \mathbb{E}X) \\ &= \mathbb{E}Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}X) \end{aligned}$$

Here is how things get crazy ... in statistics ...

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots, (X_n, Y_n)$$

Post-midterm Review

Friday, March 09, 2012
10:37 AM

Two RVs X, Y

Correlation coefficient

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho_{X,Y} \leq 1$$

$\rho_{X,Y}$ captures linear (affine) relationship btw X and Y.

Continuous RV X, Y, Z

~~pmf~~ $P[X=5] = P[X=3] = P[X=1.3] = P[X=x] = 0$

↳ use pdf $f_X(x)$

$$P[3 \leq X < 5] = \int_3^5 f_X(x)\,dx$$

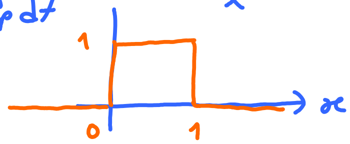
$$P[X^2 > 1] = \int_{-\infty}^{-1} f_X(x)\,dx + \int_1^{\infty} f_X(x)\,dx$$

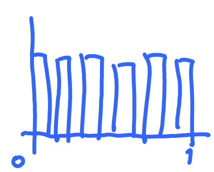
Important example of continuous random variables:

$X = \mathrm{rand}()$ → pdf $f_X(x)$

↓

histogram





Important properties of pdf: ≥ 0, $\int = 1$

Another important example: Gaussian / Normal RV

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

calculus

$$\mathbb{E}X = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx = m$$

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx = m^2 + \sigma^2$$

$$\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2 = m^2 + \sigma^2 - m^2 = \sigma^2$$

$$\sigma_X = \sqrt{\operatorname{Var} X} = \sigma$$

General Gaussian/Normal RV : $X \sim \mathcal{N}(m, \sigma^2)$

CDF : $F_X(x) = P[X \le x]$

$\downarrow$ $m = 0$, $\sigma = 1$

Standard Gaussian/Normal RV : $Z \sim \mathcal{N}(0,1)$

CDF : $F_Z(z) = P[Z \le z] = \Phi(z)$

can evaluate this using Table

$$P[a \le Z \le b] = \Phi(b) - \Phi(a)$$

This can be extended to find the probability involving general Gaussian RV $X \sim \mathcal{N}(m, \sigma^2)$.

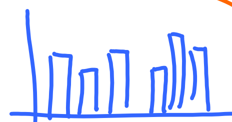
Link : $\dfrac{X - \mathbb{E}X}{\sigma_X} = \dfrac{X-m}{\sigma} \sim \mathcal{N}(0,1)$

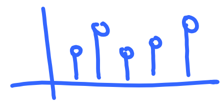
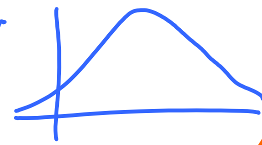
$$\Rightarrow P[a \le X \le b] = P\left[\frac{a-m}{\sigma} \le \frac{X-m}{\sigma} \le \frac{b-m}{\sigma}\right]$$

$Z \sim \mathcal{N}(0,1)$

$$= \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)$$

Distribution

data / observations
- histogram 
- relative freq. histogram

random variable
- discrete $\rightarrow$ pmf 
- continuous $\rightarrow$ pdf 

Sampling Distribution

For us, we focus on sample mean ($\bar{X}$)

$\downarrow$

estimate population mean (μ)

Assumption : For this part of the class, we will assume that the population is large. (infinite)

We will model the members of the population by random variables all with the same distribution (pmf, pdf).

Now, under the assumption above, we consider the distribution of $\bar{X}$.

Theorem 1: If the population is governed by

Normal/Gaussian distribution, $N(\mu, \sigma^2)$

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

Theorem 2 (CLT) : For any population, if n is large ($n \geq 30$), then

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

Goal : Estimate parameter of the population

↓

we focus on estimating the mean μ

- → point estimate : $\bar{X}$
- → interval estimate

↓

Find confidence interval (CI).

Section 13 Confidence interval (CI)

↳ interval estimate + confidence level (CL)

$[l, u]$ $\qquad$ $100(1-\alpha)\%$

Quality of CI

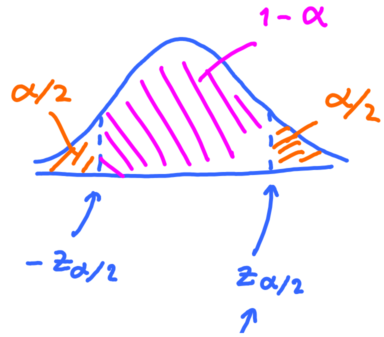
1) CL : $100(1-\alpha)\%$
2) Precision (inversely related to the width of CI)

Construction of CI

↳ 13.1) σ is known

$$l = \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$u = \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



get this value by finding the value of z in the Φ table such that $\Phi(z) = 1 - \frac{\alpha}{2}$.

13.2) σ is unknown

$$l = \bar{x} - t_{\alpha/2, n-1} \frac{\hat{\sigma}}{\sqrt{n}}$$

$$u = \bar{x} + t_{\alpha/2, n-1} \frac{\hat{\sigma}}{\sqrt{n}}$$

14. Tests of Hypotheses

Inferential Statistics : reaching conclusions based on sample information.

To do this, we use $\underbrace{\text{hypothesis}}_{\text{a statement about population parameter(s)}}$ testing

for us, we focus on the population mean (μ)

we wish to test

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

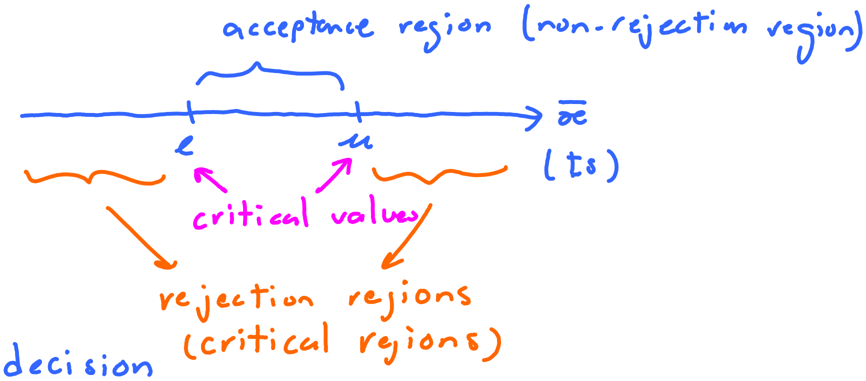
First step: form H_0, H_1

Two types of error

- False positive : Type I $\Rightarrow$ probability $= \alpha$ significant level
- False negative : Type II $\Rightarrow$ probability $= \beta$

Take a sample

Step 4
- Compute the test statistic (TS)
- compare with the dicision rule



- make a decision

If $\bar{x} \in [l, u]$, then conclude that ~~H_0 is true~~. we fail to reject H_0

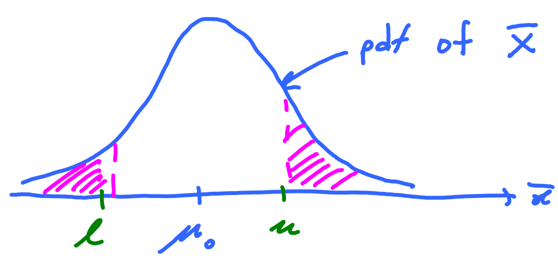
If $\bar{x} \notin [l, u]$, then conclude that ~~H_1 is true~~. we reject H_0

What is α?

↳ probability of type I error.

= probability of rejecting H_0 when $\underbrace{H_0 \text{ is true}}_{\mu = \mu_0}$.

$\bar{X} \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$



suppose the critical values are l and u

Then α = the shaded area.

Observe: small $\alpha \to$ larger $u - l \to$ more demanding test

Second step: choose the value of α

↳ usually,

$\alpha = 0.05$

Third step: Determine ℓ, u

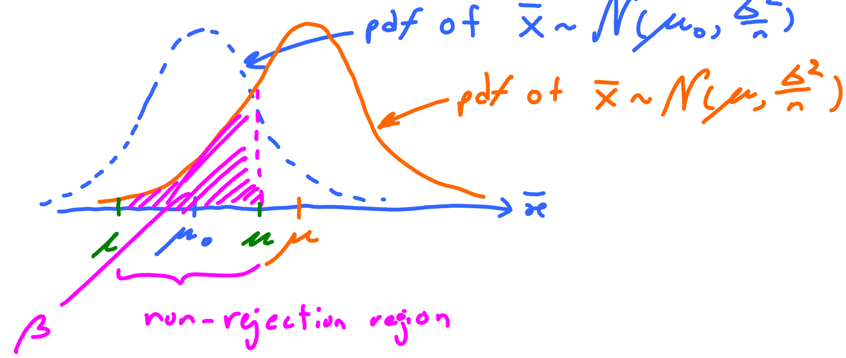
What about β? (more difficult)

probability of the false negative

= probability of failing to reject H_0 when H_1 is true.

$\mu \neq \mu_0$

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



Note: smaller α → non-rejection region will be expanded → larger β