MATH 10

INTRODUCTORY STATISTICS

Tommy Khoo

Your friendly neighbourhood graduated student.

Homework 5

- You know what to do.

More Good News About The Final Exam

- "Cumulative" only starting from sampling distributions and confidence intervals!

- Be careful: previous concepts like Normal distribution, Pearson's r (in regression), probability etc ***are used*** in later chapters.

More Good News About The Final Exam

- "Cumulative" only starting from sampling distributions and confidence intervals!

- Be careful: previous concepts like Normal distribution, Pearson's r (in regression), probability etc ***are used*** in later chapters.

- 6 questions, 15 points each, total 90 points.
- 20 minutes per question, 2 hour exam but you have 3 hours. ^__^

Week 8 - 9

Finals : 1st June, Fri, 11:30 am

- Chapter 14 – Regression ← today's lecture
- Chapter 15 – Analysis of Variance ← today's lecture
- Chapter 16 – Chi Square

Tentative plan :

- Week 9 = finish up, review for final exam.
- Last lecture on Week 10, Tuesday = work through (new) sample qns.

Chapter 14 - Regression

- Bivariate data : $(X_i\ , Y_i)$

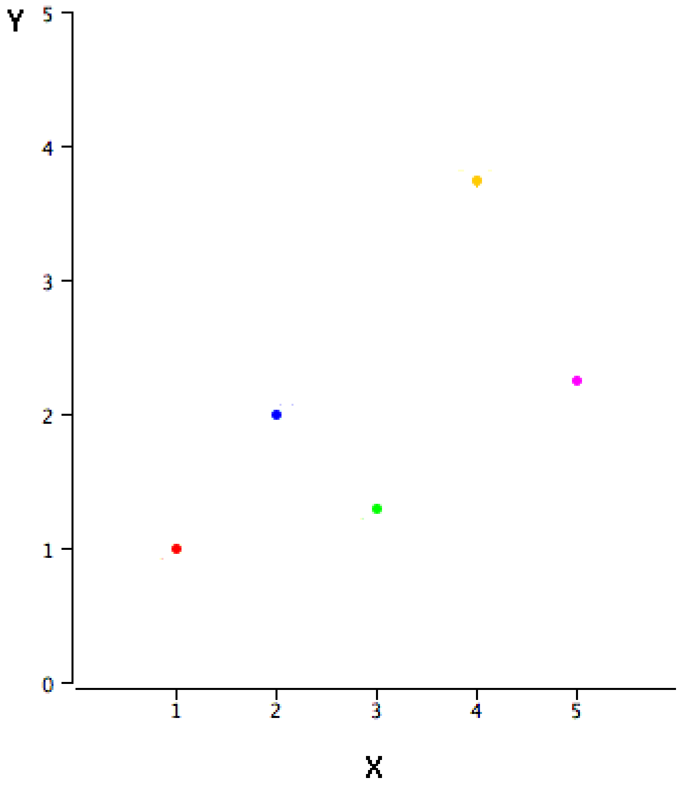
- For this course, we always take Y_i to be the *dependent* or *criterion* variable.
- X_i is the *independent* or *predictor* variable.

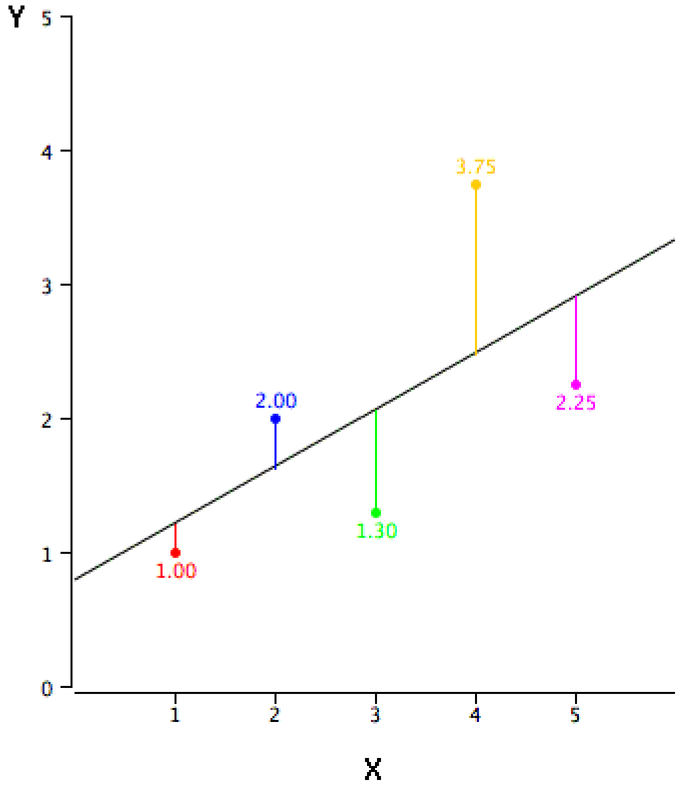
Chapter 14 - Regression

- Bivariate data : (X_i, Y_i)
- For this course, we always take Y_i to be the *dependent* variable.
- X_i is the *independent* or *predictor* variable.

- Believe there is a linear relationship, want to find "best fit" line.
- Best fit means to start with a line: $\hat{Y}_i = bX_i + a$
- Then, find a, b to minimize the sum of square errors.

Want to find the “best fit” line.





Errors of prediction (or residuals)

- $\hat{Y}_i = bX_i + a$
- $Y_i = i$th actual value.

- Difference between observed and predicted : $e_i = Y_i - \hat{Y}_i$

- We want to minimize the sum of squared errors $\sum_{i=1}^{n} e_i^2$.

Computing Regression Line

Important for Exam!!

- Slope coefficient :

$$\boldsymbol{b = r\, s_Y / s_X}$$

- **r** = Pearson correlation coefficient.

- Intercept *($\bar{Y}$ is the sample mean of all the Y values etc)* :

$$\boldsymbol{a = \bar{Y} - b\bar{X}}$$

Usage of Regression Line

Important for Exam!!

- Slope coefficient $\boldsymbol{b} = \boldsymbol{r}\ \boldsymbol{s_Y}/\boldsymbol{s_X}$ tells us the predicted change in Y, per unit change in X.

- Regression line is used to predict the value of Y, given a value of X.

- Formula for intercept coefficient tells us that regression line passes through the means $(\bar{X}, \bar{Y})$.

Chapter 14, Section 4 – Partitioning Sums of Squares

- Sum of squared deviations of Y from its mean :

$$\boldsymbol{SSY} = \sum (\boldsymbol{Y} - \overline{\boldsymbol{Y}})^{\mathbf{2}}$$

- SSY can be partitioned : SSY = SSY' + SSE

- SSY’ = sum of squares predicted
- SSE = sum of squares error

Chapter 14, Section 4 – Partitioning Sums of Squares

$$SSY = \sum (Y - \bar{Y})^2$$

- SSY can be partitioned : SSY = SSY' + SSE
- SSY’ = sum of squares predicted
- SSE = sum of squares error

$$\boldsymbol{SSE = \sum (Y - \hat{Y})^2} \quad , \qquad \boldsymbol{SSY' = \sum (\hat{Y} - \bar{Y})^2}$$

Chapter 14, Section 4 – Partitioning Sums of Squares

- Proportion explained = SSY'/ SSY = explained / total sum of squares.
- Proportion (of the variation) explained = r^2.

- Proportion not explained = SSE / SSY = residual errors / total sum of squares.

- The usual convention is to label these TSS, ESS, RSS.
- I am following the textbook's labels.

Chapter 14, Section 5 – Standard Error of the Estimate

- We can get a standard error of the estimate (sum of squares error).

$$s_{est} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N - 2}}$$

- Another way of writing this :

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

- Sample versions.

Chapter 14, Section 5 – Standard Error of the Estimate

- We can get a standard error of the estimate (sum of squares error).

$$\sigma_{est} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N}} = \sqrt{\frac{\sum e^2}{N}} \; (std\ dev\ of\ errors)$$

- Another way of writing this :

$$\sigma_{est} = \sqrt{\frac{(1 - \rho^2)SSY}{N}}$$

- Population versions.

Chapter 14, Section 6 – Hypothesis Testing with Regression → Important for exam!!

- Assumptions:

1. Linearity - true relationship exists and is linear.
2. Homoscedasticity - variance around regression line same for all values of X.
3. Errors are normally distributed.

- Significance test on whether slope b is zero.
- t-distribution, df $= N - 2$. → *confusing, probably be given.*

Chapter 14, Section 6 – Hypothesis Testing with Regression

→ Important for exam!!

- Significance test for the slope b.
- t-distribution, df $= N - 2$.

- General formula for t-test : $\frac{variable - hypothesized\ value}{estimated\ standard\ error}$

- Standard error for the slope is $s_b = \frac{s_{est}}{\sqrt{SSX}}$. → *number will be given.*
- $SSX = \sum(X - \bar{X})^2$

Chapter 14, Section 6 – Hypothesis Testing with Regression
→ Important for exam!!

- $H_0 : \beta = 0$, $H_A : \beta \neq 0$.

- E.g. $P(\, sample\ slope\ \geq b\,) = P\left(T \geq \frac{b - \beta}{s_{est}} \right) < \frac{\alpha}{2}$.

- Business as usual.

Chapter 14, Section 6 – Hypothesis Testing with Regression → Important for exam!!

- You can also do confidence intervals.

$$[\, b \;-\; t\,SE, b \;+\; t\,SE \,]$$

- I hope you are starting to see the pattern in both

- Do not confuse hypothesis testing with confidence intervals! Don't over think. :p

Public Service Announcement

- Chapter 14, Section 9, Introduction to Multiple Regression.
- Parts of Chapter 14 : Significance Test for the Correlation, Leverage, Influence.
- Not required.

Break time!! \o/

- Break starts after I hand out the exercise.

12 minutes

- Circle is a timer that becomes blue. O_o →

(please ignore if it glitches)

Chapter 15, Section 2 - Introduction

- Analysis of Variance (ANOVA)

- Test differences between two or more means, by analyzing their variance.

- For this course, we will only do One-Way or One-Factor ANOVA.

Chapter 15, Section 2 - Introduction

- **Between-subjects factors**

Different subjects are used to test different values of the factor.

- **Within-subjects factors** (not doing!!)

Same subject is exposed to the different values of the factor.

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

- Example: 4 samples, from 4 different populations.
- Want to know if all 4 populations have the same population mean.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

- Example: 4 samples, from 4 different populations.
- Want to know if all 4 populations have the same population mean.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A : at\ least\ two\ means\ are\ not\ equal$$

- You can write the alternative hypothesis in English. (writing in math not required)
- Don't confuse "at least two" with "exactly two"!

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

Assumptions for the hypothesis test :

- Populations have the same variance.
- Samples are independent.
- Populations are normally distributed.

Remember these for the t-distribution tests in difference between means?

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

- The idea is that we are computing the Mean Square Error (MSE) for each population.
- MSE = how the data points in each group varies internally.
- We then compute the Mean Square error Between-Groups (MSB).
- MSB = how the groups themselves varies relative to each other.

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

- **MSE** = how the data points in each group varies internally.
- **MSE** = sum each of the group's "sample variances" (estimators of the variance) and divided by number of group.

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

- **MSE** = how the data points in each group varies internally.
- **MSE** = sum each of the group's "sample variances" (estimators of the variance) and divided by number of group.

- **MSB** = how the groups themselves varies relative to each other.
- **MSB** = n times variance of the sample means.
- Variance of the sample means come from the sampling distribution.
- n = data points in a group/sample

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

The idea behind this test is that MSB estimates the same thing as MSE if populations do have same mean. Else, MSB much larger.

So, we compare MSB to MSE using the ratio:

$$F = \frac{MSB}{MSE}$$

Two different degrees of freedom here: df for MSB, df for MSE.

Chapter 15, Section 4 – One-Factor ANOVA (exam!!)

$$F = \frac{MSB}{MSE}$$

Two different degrees of freedom here: df for MSB, df for MSE.

- df MSB = number of groups – 1
- df MSE = N – number of groups → will give you these formulas

Public Service Announcement

Chapter 15

- Section 6, Multi-Factor Between Subjects
- Section 7, Unequal n
- Section 8, Tests Supplementing
- Section 9, Within-Subjects
- Section 10, Power of Within-Subjects Designs

- Not required.