MATH 10

INTRODUCTORY STATISTICS

Ramesh Yapalparvi

Good News About Final Exam

- It is “cumulative” but ***tentatively***, only starting from sampling distributions and confidence intervals.

- Do note that many concepts like mean, standard deviation, Pearson’s r (in regression), and probability ***are used*** in later chapters.

- Will have a meeting to finalize this soon. Let you know on Tuesday.

Week 7

- **Chapter 12 – Test of Means**

More hypothesis testing!

- **Chapter 13 – Interestingly titled "Power"** ← **today's lecture**

The idea of the "power" of a test.

- **Chapter 14 – (brief introduction to) Regression** ← **today's lecture**

Aside : The Law of Large Numbers (not on the exam!)

- We have actually indirectly learned “The Law of Large Numbers”.

- Sampling distribution has true population mean.

- Standard errors have n in the denominator.

Aside : The Law of Large Numbers (not on the exam!)

- Sampling distribution has true population mean.
- Standard errors have n in the denominator.

- So, as sample size n become larger…
- …the sample mean/proportion becomes closer to true population mean/proportion.

This is the law of large numbers!

Hypothesis Testing Example : Difference Between Means

- Textbook, Chapter 12, Exercise 8
- One group threw darts at a target with their preferred hand (sample 1), another group threw darts with their non-preferred hand (sample 2).

Hypothesis Testing Example : Difference Between Means

- Textbook, Chapter 12, Exercise 8
- One group threw darts at a target with their preferred hand (sample 1), another group threw darts with their non-preferred hand (sample 2).

- Assume you can treat both samples as independent simple random samples from the respective hypothetical population of the score of a single dart throw.

- Additional assumptions: populations have the same unknown variance, populations normally distributed, both groups consists of $n = 5$ participants.

Hypothesis Testing Example : Difference Between Means

- Additional assumptions: populations have the same unknown variance, populations normally distributed, both groups consists of $n = 5$ participants.

- Sample 1 (preferred) : $\bar{X}_1 = 10.6$, $S_1^2 = 5.3$
- Sample 2 (non-pref) : $\bar{X}_2 = 8.6$, $S_2^2 = 1.3$

- $H_0 : \mu_1 - \mu_2 = 0$ *i.e. no difference*
- $H_A : \mu_1 - \mu_2 > 0$ *i.e. better to throw with preferred hand*

Hypothesis Testing Example : Difference Between Means

- Populations normal, unknown variances => sampling dist. is t-dist.
- Mean of sampling dist. = $\mu_1 - \mu_2$

- Standard error = $\sqrt{\frac{S_1^2+S_2^2}{n}} = \sqrt{\frac{5.3+1.3}{10}} = 0.81240$-ish.
- Degrees of freedom = (n – 1) + (n – 1) = 4 + 4 = 8.

Hypothesis Testing Example : Difference Between Means

- Standard error = $\sqrt{\frac{S_1^2+S_2^2}{n}} = \sqrt{\frac{5.3+1.3}{10}} = 0.81240$-ish.
- Degrees of freedom = (n – 1) + (n – 1) = 4 + 4 = 8.
- $\bar{X}_1$ - $\bar{X}_2 = 10.6 \; - 8.6 = 2.$

- $P(\, sample\ difference \;\geq 2\,) = P(\,T \geq \frac{10.6\ -8.6}{0.81240}\,)$
- $P\left(\,T \geq \frac{10.6\ -8.6}{0.81240}\right) = P(\,T \geq 2.461\,) < P(\,T \geq 2.31\,) = 0.025$

Hypothesis Testing Example : Difference Between Means

- $P(\, sample\ difference\ \geq 2\,) = P(\, T \geq \frac{10.6 - 8.6}{0.81240}\,)$
- $P\left(\, T \geq \frac{10.6 - 8.6}{0.81240}\,\right) = P(\, T \geq 2.461\,) < P(\, T \geq 2.31\,) = 0.025$

- The question does not specify a significance level α.
- But we know that the condition for rejecting H_0 is :

$$P(\, sample\ difference\ \geq 2\,) < \frac{\alpha}{2}$$

Hypothesis Testing Example : Difference Between Means

- But we know that the condition for rejecting H_0 is :

$$P(\,sample\ difference\ \geq 2\,) < \frac{\alpha}{2}$$

- Depends on what α is given, this condition may or may not be met.

- Condition met : reject H_0 at α level of significance. Throwing with the preferred hand will *probably* give a higher score.

- Condition not met : we do not reject H_0 at α level of significance. Inconclusive.

Chapter 13 - Power

- Recall : Type I and II errors. → exact definition / lingo not required

- Probability of rejecting a true $H_0 = \alpha$ *(yes, the significance level)*

Chapter 13 - Power

- Recall : Type I and II errors. → exact definition / lingo not required
- Probability of rejecting a true $H_0 = \alpha$ *(yes, the significance level)*
- Probability of failing to reject a false $H_0 = \beta$.
- Remember these by realizing that H_0 is either true or false.

Chapter 13 - Power

- Probability of failing to reject a false null hypothesis $= \beta$.
- Power $= 1 - \beta$.
- Cannot be calculated unless we specify a particular value for the alternative hypothesis.

Chapter 13 - Power

- Probability of failing to reject a false null hypothesis $= \beta$.
- Power $= 1 - \beta$.

Example of power calculation

(in this course we will only do this for normal distributions)

$H_0 : \mu = 50$, $H_A : \mu > 50$, let's say the true mean is 60.

Chapter 13 - Power

Example of power calculation

(in this course we will only do this for normal distributions)

$H_0 : \mu = 50$, $H_A : \mu > 50$, let's say the true mean is 60.

Population variance given : $\sigma^2 = 25$.

Significance level $\alpha = 0.1587$. @__@

Chapter 13, Section 6 – Factors Affecting Power

- Sample size - *larger sample size, higher power*.
- Standard deviation - *lower variance, higher power*.

- Difference between hypothesized and true mean.
- Significance level. → interesting trade-off
- One vs. Two-tailed tests.

Aside: Neyman-Pearson Lemma (not in exam!!!)

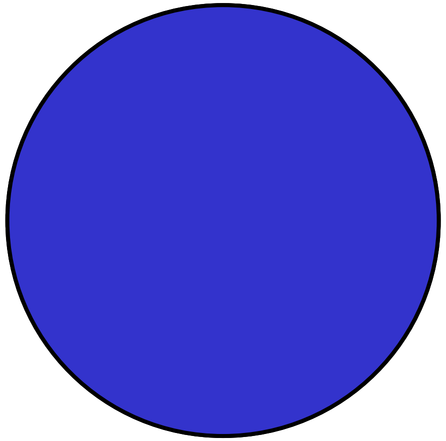
- One of the key ideas in statistical hypothesis testing.
- Given any significance level α, what is a hypothesis test that maximizes power $1 - \beta$?
- Neyman-Pearson proved that it is the Likelihood Ratio Test.
- Not the popular framework that we are learning now.

Break time!! \o/

- No exercise today! Go enjoy your break. ^__^
- There are only so many ways I can write hypothesis testing questions. ☹
- Circle is a timer that becomes blue. O_o →

(please ignore if it glitches)

12 minutes



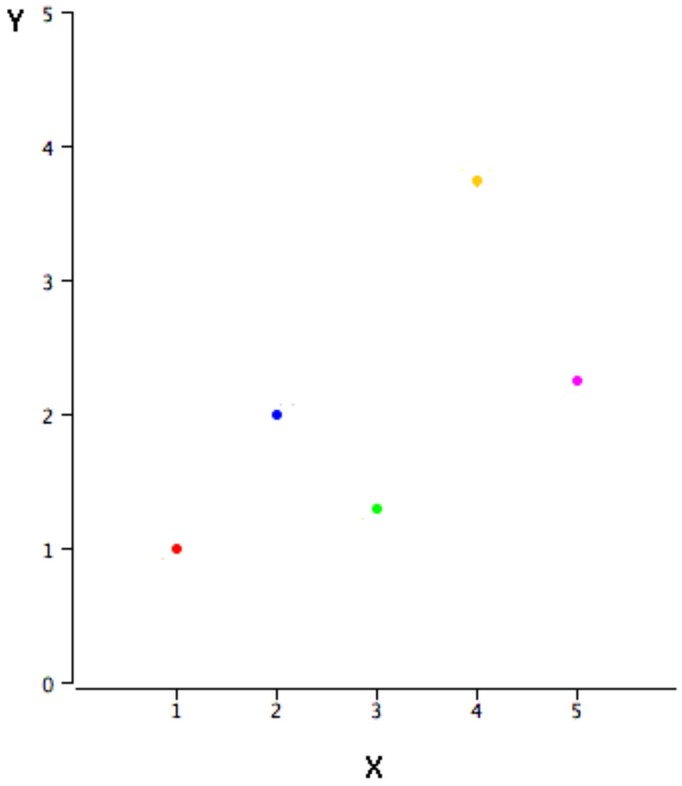
Chapter 14 - Regression

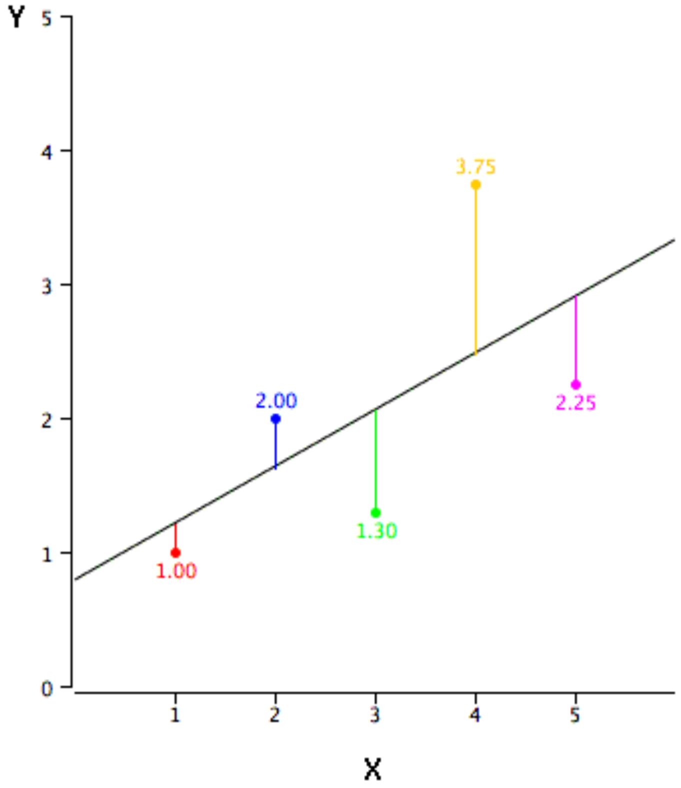
- Bivariate data (remember Pearson's r?).

- Pick one to be the independent variable, X.
- Pick one to be the dependent variable, Y.

- One independent/predictor variable => simple linear regression.
- Want to plot predictions of Y as a function of X using a straight line.

Want to find the “best fit” line.





Errors of prediction (or residuals)

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

- $\hat{Y}_i = bX_i + a$ → *recall the slope-intercept definition of a line.*
- $Y_i = i$th actual value.

Errors of prediction (or residuals)

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

- $\hat{Y}_i = bX_i + a$ → *recall the slope-intercept definition of a line.*
- $Y_i = i$th actual value.

- We want to minimize the sum of squared errors $\sum_{i=1}^{n} e_i^2$.
- Remember how the mean is defined as the quantity that minimizes the sum of squares deviations?

Computing Regression Line

- Slope coefficient :

$$b = r\ s_Y / s_X$$

- r = Pearson correlation coefficient.

- Intercept :

$$a = M_Y - bM_X$$

Standardized Variables

- To standardize a variable, you subtract its mean from it and divide the result with the standard deviation.

$$\frac{X \ - mean}{standard\ deviation}$$

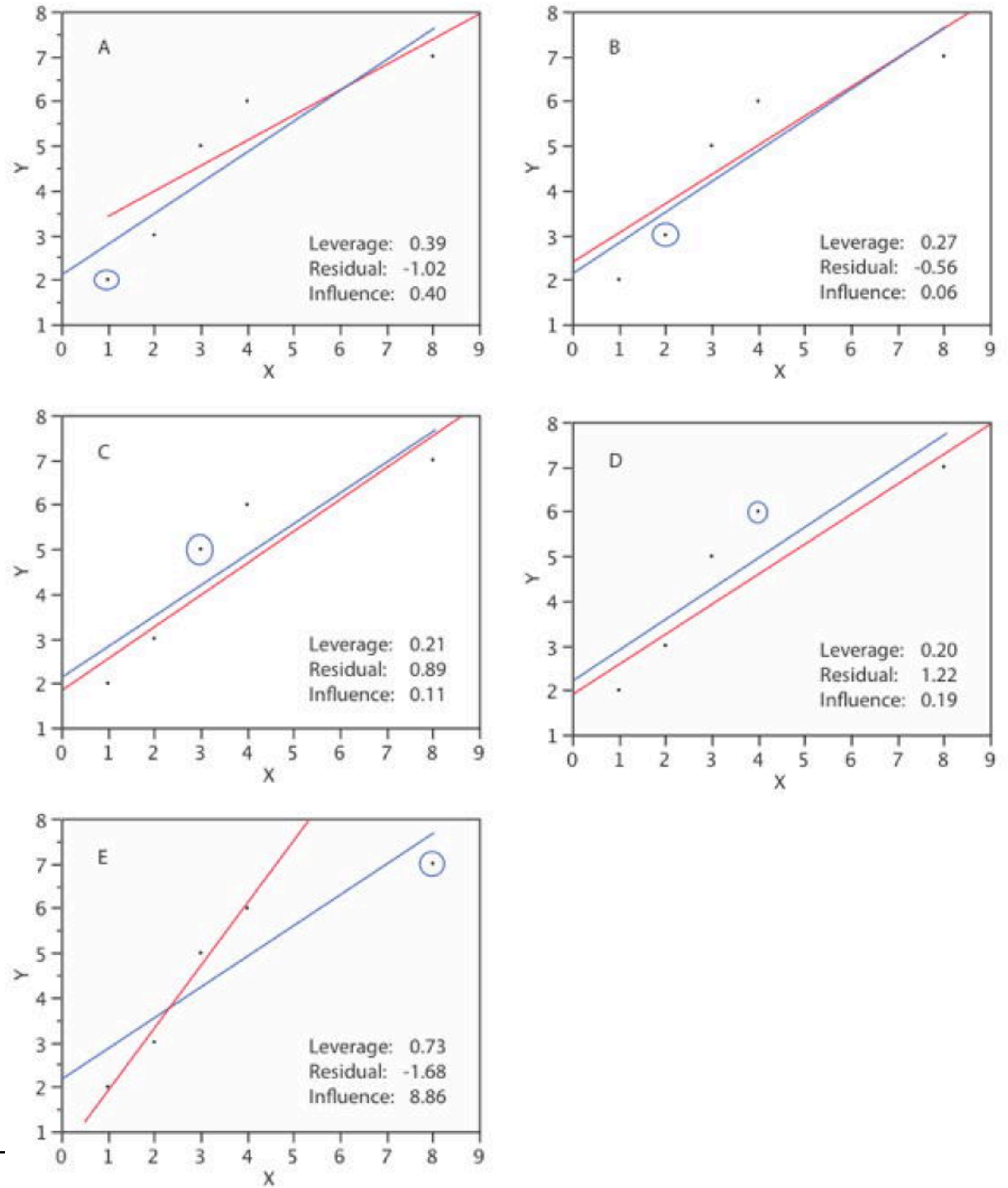
- We have been doing this to turn variables into standard normal and standard t-dist.
- Then, regression line becomes :

$$\hat{Z}_Y = r\, Z_X$$

Chapter 14, Section 7 – Influential Observations

- Textbook : leverage and influence.
- Both not required.

- Just an intuitive understanding of what outliers do to the regression line.



Chapter 14, Section 8 – Regression towards the Mean

- Slope coefficient :

$$b = r\, s_Y / s_X$$

- A change in one standard deviation in X is predicted by the regression model to result in a change in r standard deviations in Y.

Chapter 14, Section 8 – Regression towards the Mean

- Slope coefficient :

$$b = r\, s_Y / s_X$$

- A change in one standard deviation in X is predicted by the regression model to result in a change in r standard deviations in Y.

- So, if X and Y are similar measurements, e.g. heights of children and parents, then higher than average X would appear to be associated with a Y that is less over the average.

Chapter 14, Section 4 – Partitioning Sums of Squares

- Sum of squared deviations of Y from its mean :

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

- 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

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

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

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

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).

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

- Another way of writing this :

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

- Population versions.

Chapter 14, Section 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 6 – Hypothesis Testing with Regression

- Assumptions:

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

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

Chapter 14, Section 6 – Hypothesis Testing with Regression

- 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 $s_b = \frac{s_{est}}{\sqrt{SSX}}$.
- $SSX = \sum (X \; - \bar{X})^2$

Public Service Announcement

- Chapter 14, Section 9, Introduction to Multiple Regression
- Small part of Chapter 14 : Significance Test for the Correlation.

- Not required