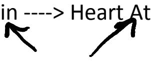
1/26/2015

Monday, January 26, 2015
9:31 AM

- Scribing- completion for making an effort with detailed notes- sign up sheet
- Help section- Thursday 3-5 pm in CBA 4.326
  - Just informative, available to help with software, homework, questions, etc.
  - Website shows R-walkthroughs and readings and class agenda
  - Second set of exercises up due next Monday

- Warm up into Statistical Modelling
  - SAT walkthrough - shows what you should type into R and what you should see in return
  - Review of last week:
    - Looked at questions
    - **Longitudinal**- over time
      - If they all implemented cell-phone bans, they could view the change over time
      - Control variable is the unit itself
      - Austin before vs. Austin after
    - **cross-sectional studies**- multiple units at the same time
      - Take arbitrary political boundaries- like with state lines - texas/arkansas border, some have an enforcement in law while others don't
    - **Natural experiments**- seemingly randomized event
      - Nature gave you the ability to do the experiment
      - Ex: Does becoming rich make you happy- Lottery winners compared to those who played and didn't win- experimental intervention
    - **Matching**- pair control group to variable group by factors such as age, background, etc.
    - **Endogenous vs. Exogenous variable**
      - Aspirin ---> Rate of Heart Attack
        - Experimental situation
        - Exogenous- outside the system
          - Aspirin is a force made upon them
          - No inward pointing link
      - Aspirin ----> Heart Attack



        - Consciousness
        - Ask randomly if someone is taking aspirin and if they have had a heart attack in the last year
        - Could also be for other reasons
          - Health Consciousness
        - Aspirin is endogenous
          - It is being controlled by other variables
      - Is there an inward pointing link to the statement of affair?
      - Also known as confounded and not confounded
  - Homework Exercise
    - Question 1
      - Blueberries
        - Good- study was randomized, double blind, and placebo controlled
        - Bad- study was very small group- women from 40-60
      - Vitamin D
        - Good- very careful
        - Bad- other confounding variables
        - They didn't make any experiment, the article wasn't clear, had to go back to original article
        - Picked 300 people, and looked at those with colon cancer and worked backward
        - Control group- they matched by age
        - **Case control study-** it is everything
          - Not at all the same as an experimental intervention
          - Matching could go horribly wrong
            - Ex: case control of pregnant women, negative vs. positive baby effects and matching backwards, need to look at more than just alcohol (such as control group using cocaine vs. those that don't) - results from matching could go horribly wrong
      - Mice with eating less hours

        - □ Good- long study, test out multiple variables
        - □ Bad- mice doesn't mean it will be the same on humans
    - ○ Question 2

$$f(\theta) = \sum_{i=1}^{N} (y_i - \theta)^2$$

SSE - sum of square areas

TV

$$f'(\theta) = \sum_{i=1}^{N} \cancel{2}(y_i - \theta) = 0$$

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \theta = 0$$

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \theta$$

$$= n\theta$$

$$\sum \frac{y_i}{n} = \theta$$

$$\bar{y} = \theta$$

- ○ Question 3- link on website to answer and explanation

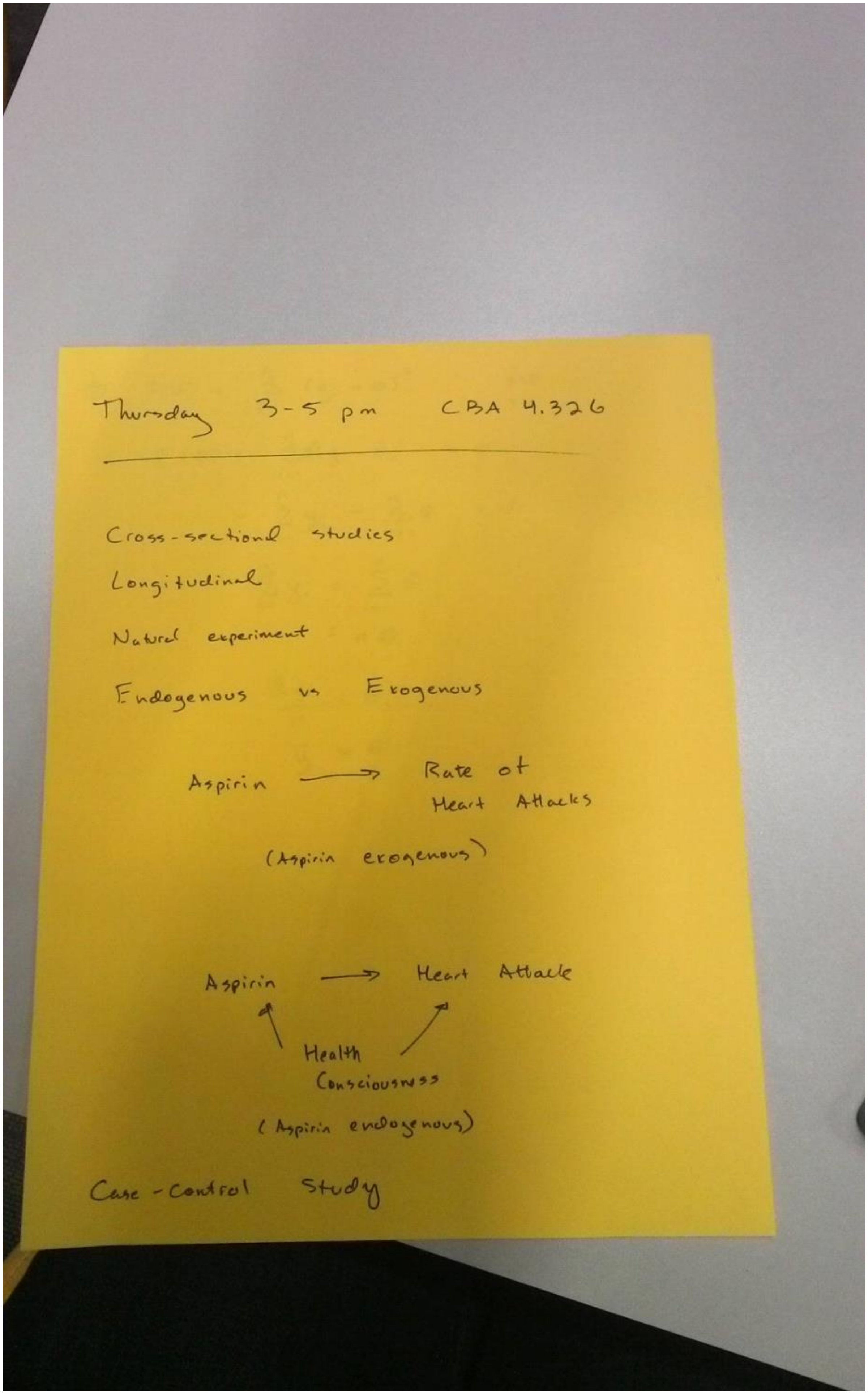
- Graphics
    - Gives force to a story and pictures you are telling
    - Good
        - ○ Vehicles for comparison - used for making comparisons

  - Multivariate
  - Truthful about magnitude
  - Usually not for small data sets- Size matters
- Bad- Scott does not like pie charts
- Clipped y-axis- gives incorrect impression that something can be skyrocketing when it's not
- Figures should be truthful about magnitude- lies about numbers
- 3 dimensional figures deceive the eye so it looks more important than it actually is
- Good charts are transparent to the reader
- Implicit causal story of a graph- what can you infer from it
- Very easy to mislead graphs with lurking variables (events in years for data overtime)

- Exploratory data analysis: the basics
  - **Categorical or grouping variables VS. numerical or quantitative variables**
    - **Contingency table** - presents the data with the info going into a chart and explains why you can create the graphs you make - used for CATEGORICAL VARIABLES
      - Tables that strategy by more than two variables
      - 2 X 2 table- two rows, two columns
      - 3 way table- three categorical variables compared
    - **Histogram** - presents the data with the info going into a chart and explains why you can create the graphs you make - used for NUMERICAL VARIABLES
      - Can compare histograms
        - Need the x and y axis to match
        - And bin sizes need to match as well
          - Creates an honest distribution and comparison
    - **Standard deviation**- standard way we measure dispersion of distribution
    - **Coverage interval**- specified at any level of coverage you want (25%, 95%, etc.)
      - Comes from the **quantiles** of distribution
        - Percentage of the distribution covered
  - **Numerical vs. Grouping**
    - When graphs includes aspects of both
    - Statistical model- you've taken the data and come up with a variable that can be tested with the info
    - As a statistical model- only gives you the overall data - doesn't give dispersion
    - A **box plot** gives you more information about the dispersion of the data
      - **Between group vs. within group variation**
        - Between- how variable are the centers of the groups
        - Within- what is the dispersion within a group
          - R^2 is it less or more than 50%
      - Graph on SAT scores and correlation with college
        - Yi: outcome for the case I
        - Decompose Yi = group mean + deviation
        - Yi= fitted value + residual

$$y_i = \hat{y}_i + e_i$$

    - R-skills practice- step by step for R-script with SAT Data on the website
      - Need to install library
        - Packages--> Install --> Mosaic
        - Do 4 walkthroughs on your own time this week
        - Paste into your own script and save it
        - Execute it and compare
        - Anything with # is output



Thursday 3-5 pm CBA 4.326

Cross-sectional studies

Longitudinal

Natural experiment

Endogenous vs Exogenous

Aspirin → Rate of Heart Attacks

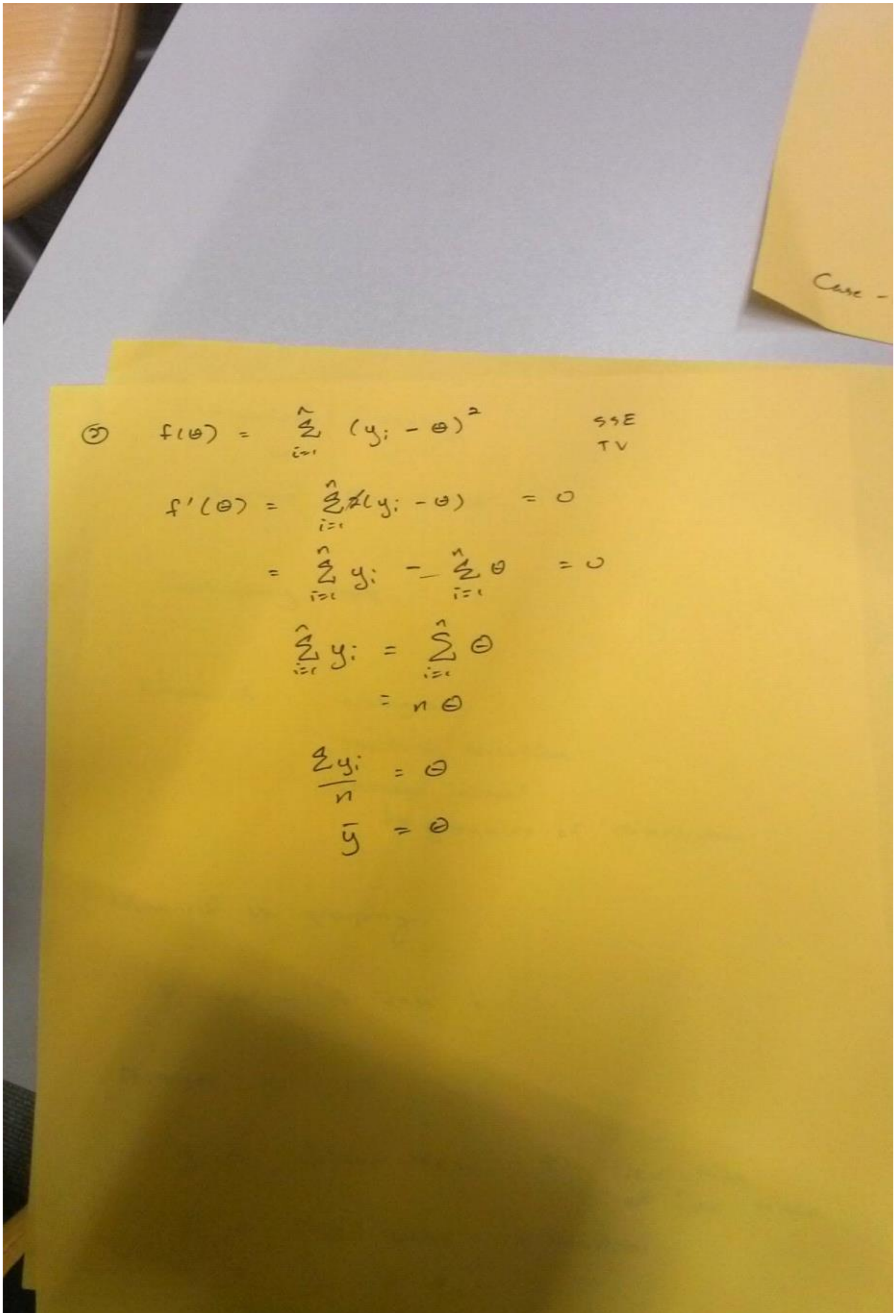
(Aspirin exogenous)

Aspirin → Heart Attack

Health Consciousness

(Aspirin endogenous)

Case-Control Study



Case -

⑦ $f(\theta) = \sum_{i=1}^{n} (y_i - \theta)^2$ SSE TV

$f'(\theta) = \sum_{i=1}^{n} \cancel{2}(y_i - \theta) = 0$

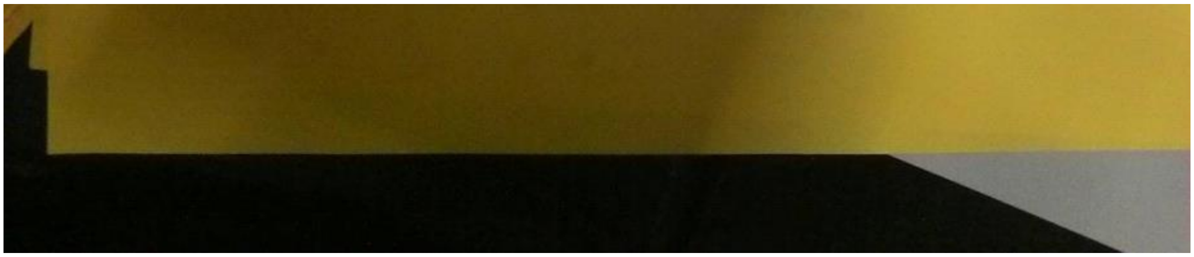
$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \theta = 0$

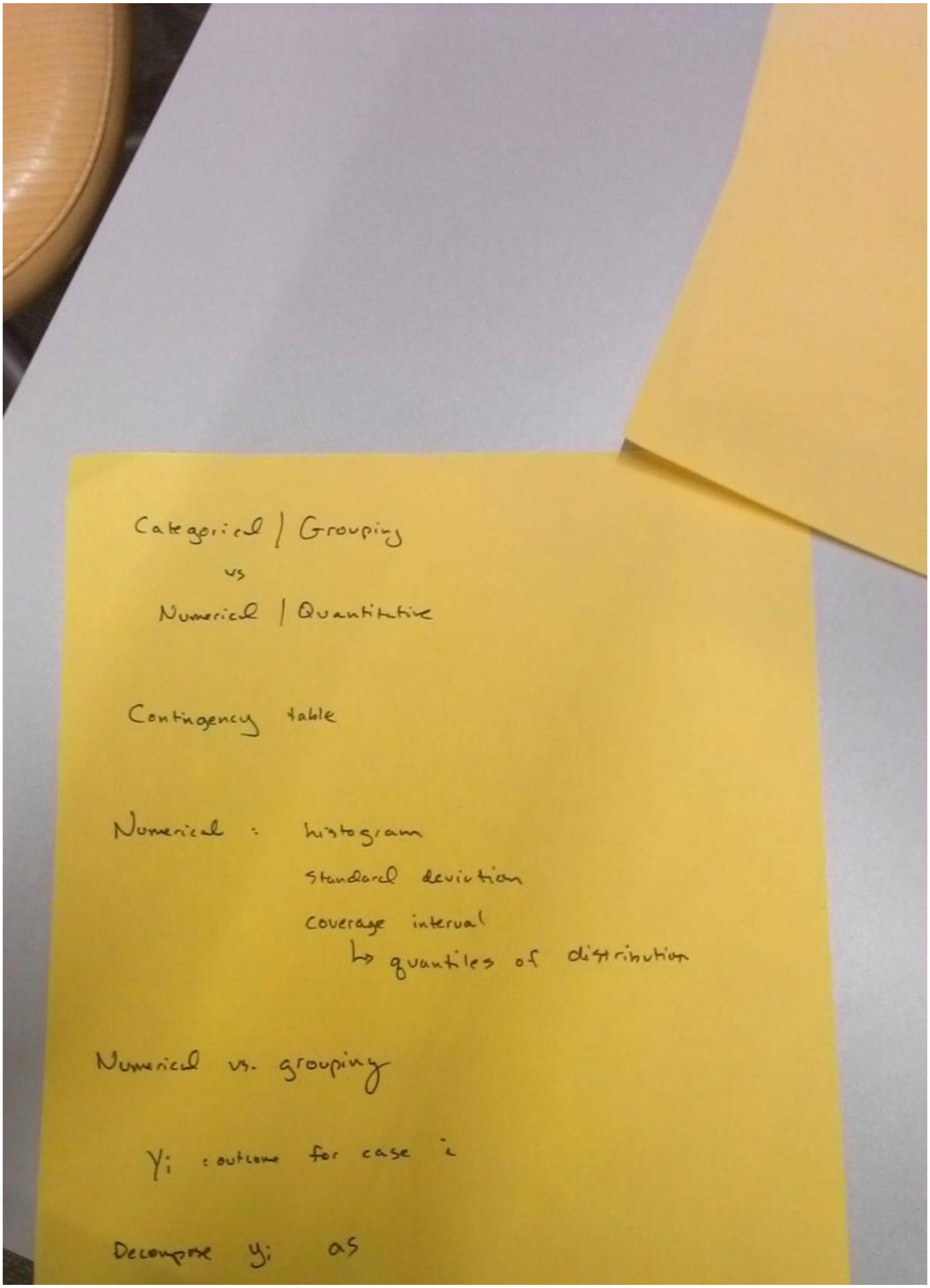
$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \theta$

$= n\theta$

$\frac{\sum y_i}{n} = \theta$

$\bar{y} = \theta$





Categorical | Grouping

vs

Numerical | Quantitative

Contingency table

Numerical : histogram

standard deviation

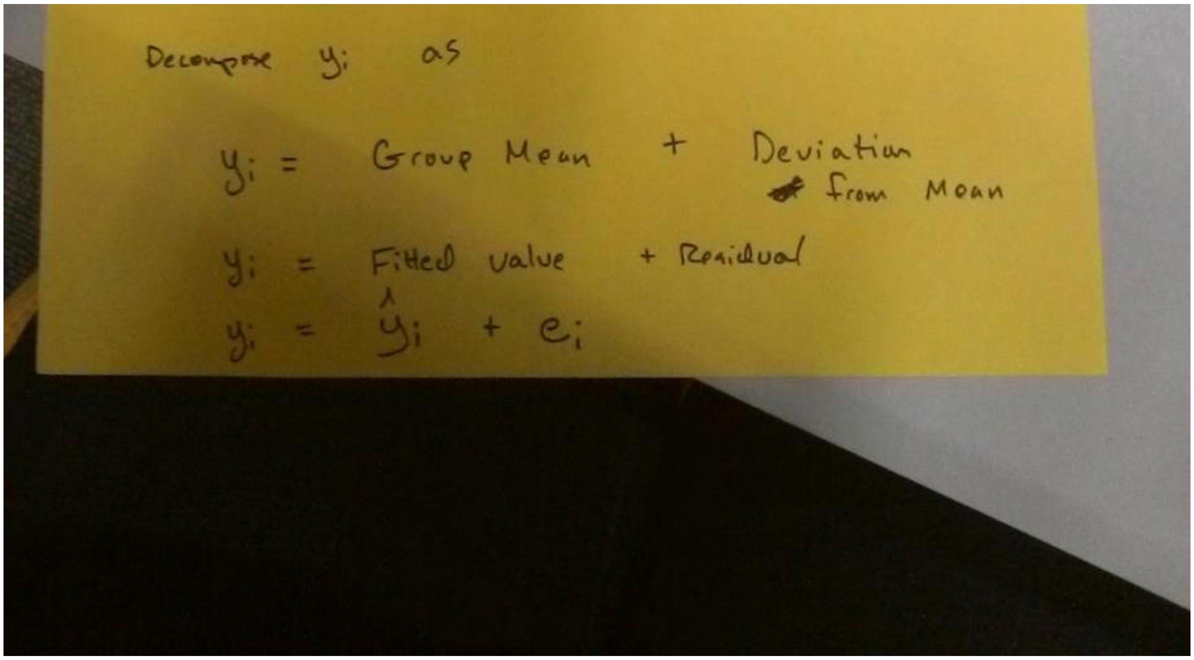
coverage interval

↳ quantiles of distribution

Numerical vs. grouping

y_i : outcome for case i

Decompose y_i as



Decompose y_i as

y_i = Group Mean + Deviation from Mean

y_i = Fitted value + Residual

$y_i = \hat{y}_i + e_i$