

TTh Final Review


Note Title 12/2/2010

Project


- * Effort based
- * Summarize data
- * Relationships
- * Inference - CI
- Novel/novel

Final Essays (Paragraph)
 * graded on correctness (thoroughness)
 * Add inference (CI) to exam 1.

Categorical: data that falls into categories
 Quantitative: measured data

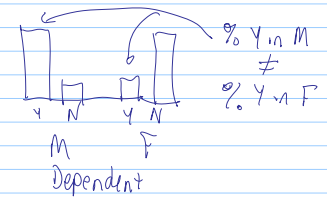
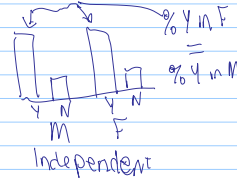
Categorical: Bar Chart
 cont →  gap: separate categories

- * No outlier, Shape
- * No skew, symm

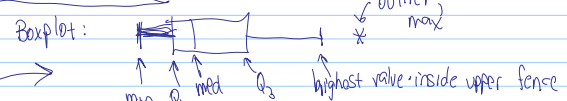
Quantitative: Histogram
 frequency →  gap: no count

- * Order of Bars matters
- * CAN Symm, Skew

Compare 2 Categorical Variables
 * Side-by-side bar chart ← use % in outermost category.

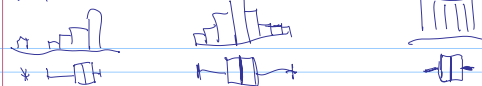


Quantitative Data



Q_1 = value that separates lower 25% of data (from upper 75%)
 Skew right

Hist/Box:



IQR = width of box = $Q_3 - Q_1$

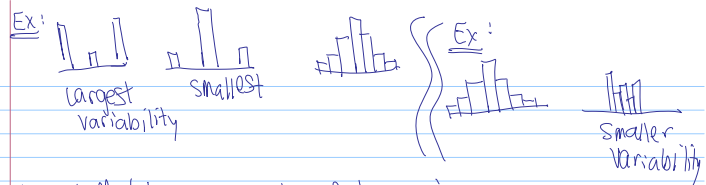
Skew data → center = median spread = IQR } resistant to outliers

Symm data → center = mean spread = SD.

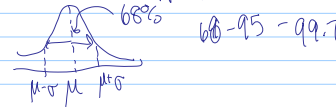
mean = center of balance on histogram
 SD = \approx distance of data to center

5# Summary: { min, Q_1 , med, Q_3 , Max }

variation (variability/spread = distance to center)

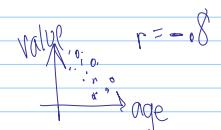


Normal Model: Appropriate if data is symm, unimodal.

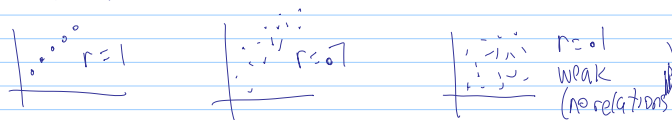


Relationships: 2 Quant. C78,9

Picture: Scatterplot
 Ex: age of BMW and value



If linear, find strength → r = linear corr. coefficient.
 $-1 \leq r \leq 1$



Describe:

Form: Linear / Not Linear / No Pattern
 Direction: + / -
 strength: Strong / weak

If linear, find linear model (regression):

$$\text{value} = 42000 - 4000(\text{age})$$

Prediction: Plug x-value in, calculate y-value.

BWARE: extrapolation ← model only useful for x-values w/in range of data.

Interpret:

* slope: (-4000) ← For each year older, the value goes down by \$4000.

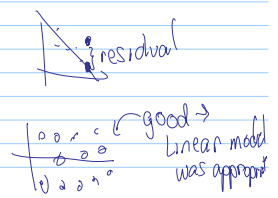
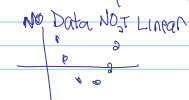
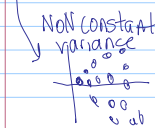
* y-int: 42000 ← value of a new car.

After regression → Diagnostics

R^2 = measures strength of adherence to model $R^2 > 50\% \rightarrow$ Good.

Residual Plot

Residual = Obs - expected = data - predicted x-value



Outliers *



← No linear pattern

If outlier → Do analysis 2x

Association ≠ Causation

Surveys / Bias

over- (or under-) representation part of pop.



Population μ, σ

sample \bar{y}, \hat{p}

} CI / Hyp tests

Eliminate Bias;

RANDOM

Obs Study

Experiment ← Experimenter RANDOMLY assigns treatment

↖ Cause/Effect conclusion

Principles

- ① Randomization
- ② Block
- ③ Control
- ④ Replicate

LN: Draw MANY (millions) of samples, the samples start to look like population.

CLT: Repeatedly draw a fixed number of samples from pop, and compute same statistic on each (\bar{y}, \hat{p})
↳ collection of all these stats ← sampling distribution
↳ "large" sampling distribution will be normal

$$E(\hat{p}) = p$$

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

$$E(\bar{y}) = \mu$$

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

$n \uparrow$ SD \downarrow

Data Analysis (Essay)

*Review Exam 1

*Consider add CI.