

7:35 pm Final Review

Note Title 12/1/2010

Project
 * Summarizing
 * Relationships
 * Inference ← CI

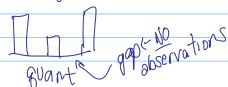
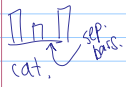
Final Exam (Essay)
 3-6 } Given 1-2 variables
 at a time to analyze.

Paragraph.
 Graded on correctness/ thoroughness

Categorical Data (3, 26, 18, 19, 20, 21)

Graph: * Bar Chart (1 var)
 * Side-by-Side bar charts (2 var)
 ↳ Use % within outermost category.

Bar vs Histogram

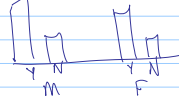


categorical: data in categories
 Quant: measures

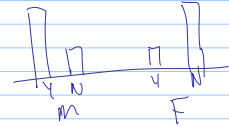
Bar order not important

Histogram: Bar order is important
 → CAN discuss skew/symm

Compare ← Independence



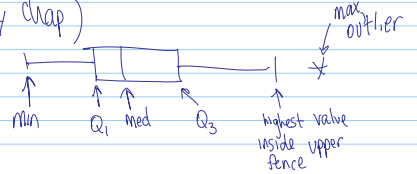
Gender and answer INDEP



Dependent.

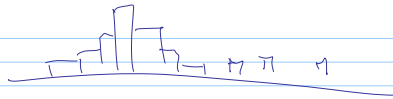
Quant. Data (Many Chap)

Boxplot



Q_1 = value that separates lower 25% from upper 75%.

Skew right.



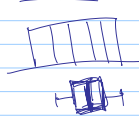
skew left



symm



uniform



$IQR = Q_3 - Q_1$ = width of box.

skew data → center = median

spread = IQR ← resistant to outliers

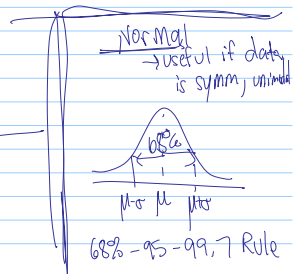
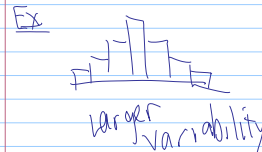
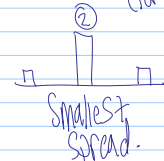
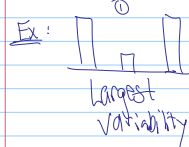
symm data → center = mean

spread = SD

mode = area of relatively high frequency.

5 # summary = { min, Q_1 , med, Q_3 , max }

variation / variability / spread = distance from the center (for bulk of data)

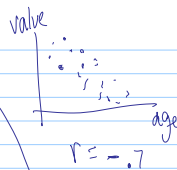


Relationships (2 Quant) - C 7, 8, 9

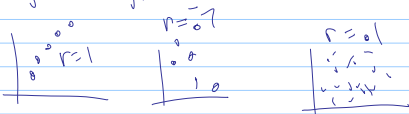
BMW new \$42000 drop \$4000/year.

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

strength of linearity = r = corr. coef.
 $-1 \leq r \leq 1$



Direction: pos/neg
 Form: Linear/non linear/no pattern
 Strength: strong/weak



Predict the value of a 1 yr old car: Plug 1 into
 → Stick to range of x values of data

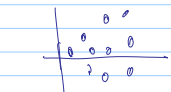
Interpret slope: -4000. For each year older, the value of the car decreases by \$4000.

y-int: \$42000. For a new car, the value is \$42000.

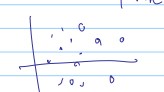
AFTER performing linear regression, examine R^2 , residual plot.

R^2 = measures strength
 Good: $R^2 > 50\%$

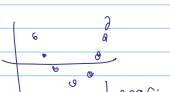
Residual Plot:
 residual = obs - expect
 = data - predicted by line



Non constant variance

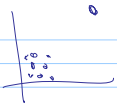


Linear model was good



Non linear pattern.

Outlier



Data not linear

} run analysis 2x

Association ~~≠~~ Causation.

Surveys/Bias

over-representation of part of population
(or under-)

Eliminate: RANDOMIZE.

Population μ, σ



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

} Generalizing sample to
Pop \rightarrow CI / Hyp test

Obs Study \leftarrow subject chooses treatment

Experiment \leftarrow experimenter Randomly assigns treatment
 \uparrow cause/effect poss.

Principles

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

LLN: Draw MANY samples from population, the samples start to look like population

CLT: Repeatedly draw samples of a fixed size and calculate a statistic on each.
"Distribution of all these statistics = Sampling Dist.
"n large", the sampling dist. is approximately Normal.