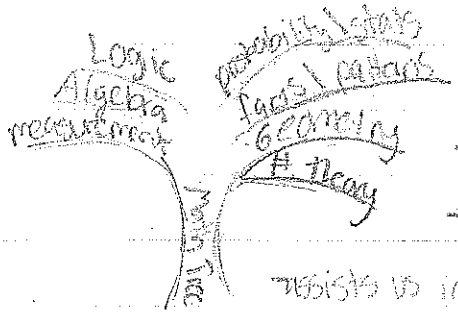
1.1 Notes: Def, Pop, Samples



- Statistical Data ...
  - is found almost everywhere
  - exists in a variety of forms

assists us in making informed decisions when faced w/ uncertainty

I. Definition

A. the collection, organization, Analysis, interpretation of numerical info.

Collect	Organize	Analyze	Interpret
surveys/polls - Sampling - experimenting accessibility - Census - Simulations	- frequency tables - charts - graphs (bar/pie, histogram, dot plot, scatter plot, boxplots, stem&leaf, trend lines)	- Center/shape/spread outliers/gaps - trends - measures of central tendency - measures of spread - confidence intervals - hypotheses tests	- assumptions - P-value re: Alpha level - confidence level - Z score - T score - written summary

- Correlation between events ≠ causation

II. Caution

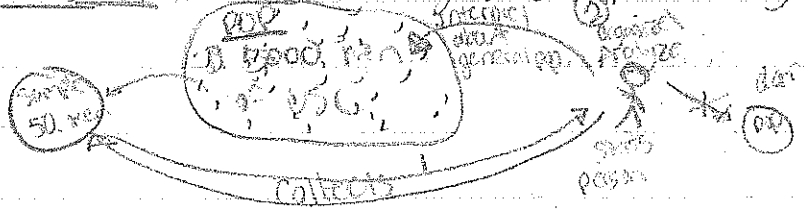
A. Can lie w/ any interpretation of stats

B. Work diligently to use stats to show entire picture/truth by good intentions & skill

III. Population

A. Is all measurements of interest, entire group of those being measured

B. Data from all



IV. Sample

A. Part of population

B. Smaller section of those being measured

C. Data from...

D. Random representative sample (to increase confidence level, ↑ size of random representative sample)

- list w/ random #s are reliable for getting a SRS

- categorical (qualitative) or quantative data

1.2 Notes: Random Sample

- Time of test(how long) - ratio
- Time of 1st class - interval
- Score on exam - ratio
- major - nominal
- course evaluation - ordinal

- Random # Table: 1) Start w/ a list 2) # all of sample 3) Prep something to land on # 4) Take first #s of sample (take first 3 digits if use 001-126) 5) Discard #s too big & continue on 6) Continue until you have amount you want *decide if going down or right, etc
- Calc: 1) Go to "math" 2) Go to "PRB" 3) Go to "randInt()" 4) Enter 5) Enter randInt(1, 126, 3) 1 = lowest # of sample, 126 = highest # of sample, 3 = how many you want to select
- Stratified is good, systematic can be bad, convenience is terrible, cluster and multistage are good,

I. Sampling Approaches

A. Convenience - bad, doesn't represent pop, (= worthless) it's easy, (= good starting place) Statistician feel good, lots o' bias
B. Stratified - ÷ pop into groups from a list & randomly sample
C. Systematic - # everybody, choose a # and then every multiple of #, so all that ends in #
D. Cluster - ÷ pop into different geographic areas & randomly select from them
E. Multistage - variety of methods

- A representative sample looks like the population
- Pop. consists of all subjects of interest (parameter)
- Sample is part of pop. & produces stats.
- Bias: • selection bias: sample doesn't represent pop.
  - undercoverage: low representation in sample
  - convenience sample: easy to get sample that doesn't represent pop
  - nonresponse bias: person chosen for sample don't participate
  - response Bias - occurs w/ research
    - social desirability: lies about info
    - leading questions: guides students to a certain answer

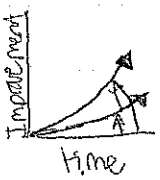
Blocking \ Binding Study Guide

2) A placebo treatment is a dummy treatment that can have no physical effect. It is a sham/fake pill. It is NO treatment.

3) The placebo effect is that many patients respond favorably to any treatment, even a placebo, because of trust in the doctor and expectations of a cure. The mind! nothing becomes something.

4) Two lurking variables in the original gastric freezing experiment was that the experimental data was misleading because of the placebo effect and the data reflects any special features of that particular study, such as a physician with a soothing manner. No control group.

5) A control group in the receives the placebo treatment in order to control the effects of lurking variables on the outcome. The experimental group receives the actual treatment and is the data that was being questioned by researcher.

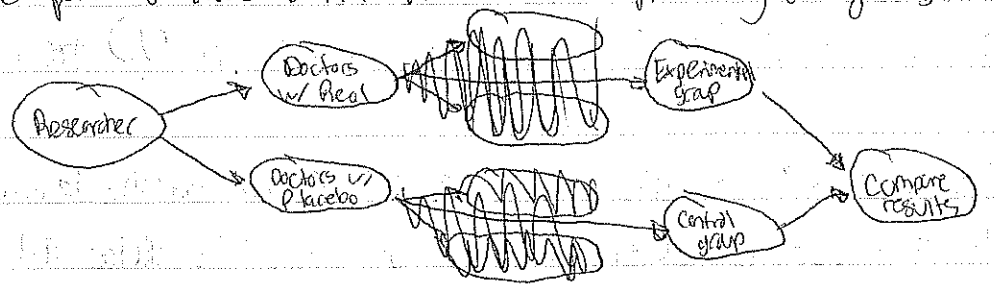


= placebo or other variable

6) Uncontrolled studies give new therapies inflated success because w/o comparison of treatments, experimental results can be dominated by the details of the experimental arrangement, the selection of subjects, and the placebo effect. The result is often bias.

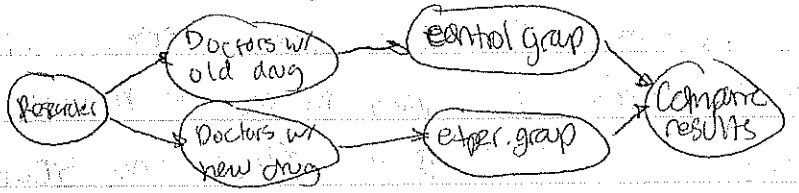
7) A double-blind experiment is that a researcher wants to see how effective a pill is. A researcher gives doctors a pill, not getting to know if the pill is the real pill or a sugar, placebo pill.

The patients also don't know what pill they are getting..



8) a) There is no control grap and the data would be inflated.

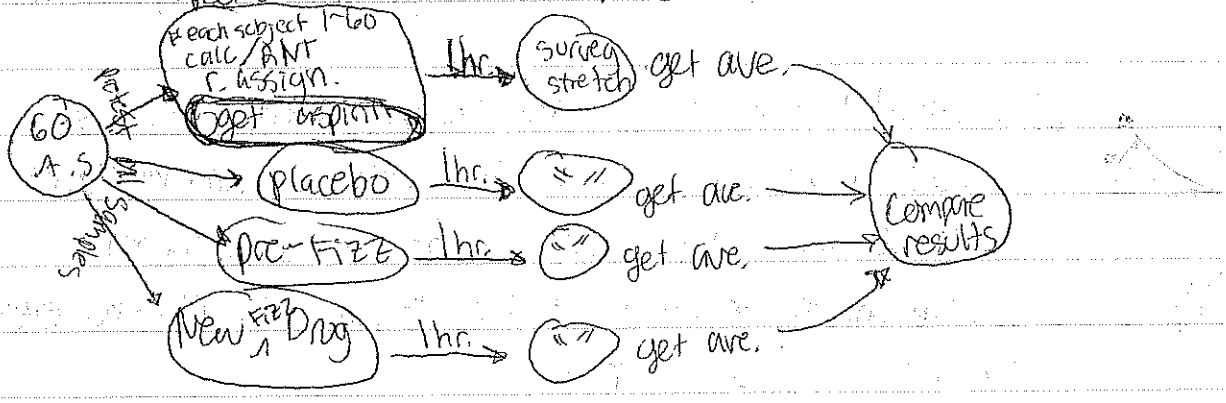
b)



c) No, this would create bias, & affect results.

d) Yes, less chance of bias. Should be double-blind.

9) Block design is the random assignment of units to treatments is carried out separately w/in each block. A block is a grap of experimental units or subjects that are similar in ways that are expected to affect the response to the treatments.



Maybe have a combo of drugs.

Simulation Notes

I. Steps to making a simulation

A. Describe Experiment -ex. BH NEL how many field goals are made; how might

B. State Assumptions -ex. distance, wind, could affect it, but assume the field goal is independent & field goal makes 90%.

C. Assign Digits on a RNT to simulate -ex. simulate next to bench

D. Repeat A LOT

E. State Conclusion that corresponds to real life

- can have #s repeated for a simulation
- written questions are better than verbal questions
- have questions w/ even #s (-ex. low satisfaction 1 2 can't put middle 3 high satisfaction 4
- space #s evenly; words centered over #s
- try to have ratio level questions
- have clear, concise questions; make respondent feel comfortable; assist them in giving truthful, accurate answers (eliminate fears, response bias, telescoping)
- assure them it's confidential and anonymous

2.2 Basic Graphs - Visual Representations of Data

I. Overview.

A. Info. can be represented quickly & efficiently

B. Info. can be presented more attractively

C. Keep it simple & understandable

D. Viewers will look for appx 7 seconds

E. It starts w/ a TITLE

F. Ends w/ Analysis; look @ highs/lows/trends & then ask why?

PUT ANALYSIS RIGHT AFTER GRAPH on page

II. Bar graphs

A. Allow viewer to see differences in data easily

B. Good tool for comparing several subjects

C. Categorical Data on horizontal axis D. Label axes
E. ~~KEEP~~ increments the same (can have a squiggle\jump once on axis)
↳ Does it zoom in for sharper comparisons or exaggerate
F. Keep the same width
G. Can have double bar graphs... USE a Key!
H. Put # inside the bar, NOT on top (adds height if put on top)
I. Pareto charts - vertical bar graphs arranged tall to small

III. Circle\Pie Graphs

A. Good for comparing parts of a whole; %s of 100
B. ÷'d into slices C. Put # or % inside sector or a key
D. Pieces can't overlap; total can't > 100%; can use "other" & extra

IV. Line Graphs

A. Shows change\change in time
B. When subject is tracked repeatedly
C. Put a dot as high as data should go
D. Connects the dots from L to R w/ a straight edge
E. Can have multiple lines... Make a key (..., ---)
F. Don't use squiggle in middle of axis
G. Time plots (time series) = special line graphs (time = x axis); notes trends

V. Picto Graphs

A. Rare, use repeated symbols; Use a key; full symbol = amount in key
B. Keep it simple; symbols same size & same spacing

VI. Dot Plots - newer graphs

A. Similar to vertical bar graph, but uses dots to build upward
B. Good for Analysis (center\shape\spread\outliers)

VII. Histograms

A. Bar graph in which bars touch
B. Width of bar has meaning
C. Data is grouped into classes

D. Useful w/ large amounts of data (big data)

E. Overall shape assists us in recognizing patterns

F. Miss out in weird data cuz everything clumped together

G. Exact/specific data is hidden

H. How choose # of classes (bars)? Ask "Would info. be hidden/lost?" "Will nothing stand out?"

I. Usually 5-10; 7 is perf J. Width = $\frac{\text{Big} - \text{small}}{\#\text{ of classes}}$ (ALWAYS round up)

K. Begin w/ smallest data value, then step by width

VIII. Frequency Tables

A. Colums to help organize raw data before graphically displaying them

B. Has class/groups; tallymarks; frequency #; class midpoints / class marks; relative frequency

C. First class begins w/ the smallest #

D. Put weirdo out into last column

IX. Types of Histograms

A. Normal (Symmetric) 

B. Rectangular 

C. Bimodal  (peaks & valleys)

D. Skewed Right 

E. Skewed Left 

X. Analysis (need a # & a word/note (words))

A. Center? B. Shape? C. Spread? (standard deviation) D. outliers?

II. Stem and Leaf Plots

A. Good for quantitative B. Easily made C. Groups data

D. Shape, distribution, still see raw data E. Not good w/ large amount of data F. Back to back G. Can split stems

H. max/min & gaps, & outliers

Ex. Period 1 in secs

0	8
0	2 6 8 9
1	0 2 5
2	1 5
3	4 5

314 = 34 secs after 60 secs

018 = 8 secs before 60 secs

Skewed right/ Skewed Right

Analysis: Center = 11 median Shape: Right

Spread: 6-15 secs Outliers: 8, 31; = median

MPG for Mack's car lot

-Ex. Jack's Dodge 16 mpg, Grand Prix 21 mpg, Ford Explorer 23 mpg; Nissans 29 mpg, Honda 31 mpg, Lexus 27 mpg

Domestic		Foreign
6	1	
4 3	2	7 9
	3	4

3/2/17 = 27 mpg for foreign & 23 mpg for domestic cars

Measures of Central Tendency

I. The Averages

A. Mean B. Median C. Mode D. Trimmed Mean

- 1D the Beast → Wha?, When?, Predictions, 10 Data Values → 10
- Mean, Median, Mode, IQR, Trimmed Mean, Trend?, Graphs

II. The Arithmetic Mean

A. $\frac{\text{sum}}{\#\text{ of terms}}$ B. $\bar{x}$ for sample data (C. "mu" (μ) for population

D. Based on numeric #s E. Easily affected by outliers

III. The Median

A. Middle value when data ordered small to large

B. Emphasizing position, not numerical value C. Resistant to outliers

D. Half data above median, half below median

IV. The Mode

A. Most often, most occurring value B. Useful for categorical data

C. Can be bimodal, can have no mode D. Outliers don't affect it

V. Trimmed Mean

A. Mean that resists the extremes B. Cut off % values from both ends

Measures of Variation Ch. 3.2

- average
- 1 # may not represent an entire set of #'s well. So a cross reference is the range, (measure of exact measure of variation / dispersion) the difference between the largest & smallest values of data
- 3 Measures of variation are range, variance, & standard deviation (σ pop or s sample)
- Pros of range: tells us difference between largest & smallest values in a distribution; total span; easy to get
  Cons of range: doesn't tell us how much other values vary from 1 another or from the mean; not important
- The measurement that helps us see how data is different from the mean is the standard deviation
- Divide by n-1 sometimes we say N other times because n-1 is for sample and N is for population ask
- Ex 8 (-5.5)$^{2n-1}$ - for ex. 5, His new formula is bad ask he is using a median and not the mean.

- The pop. formulas for mean & standard deviation when compared to the sample formulas are: the difference is by N or n-1 for standard deviation. Pop: $\mu = \Sigma x / n$ Sample: $\bar{x} = \Sigma x / n$ Means are same.

$$\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}} \qquad s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

- Ex. 2, 4, 6, 8, 10 of sample.
  $\bar{x} = 6$ median = 6 mode = none range = 10 - 2 = 8 (called total span)

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	$2-6=\|-4\|$	16
4	$4-6=\|-2\|$	4
6	$6-6=\|0\|$	0
8	$8-6=\|2\|$	4
10	$10-6=\|4\|$	16
		$\Sigma = 40$

$\frac{40}{n-1} = \frac{40}{5-1} = 10 = s^2$ (variance) $\quad s = \sqrt{10}$

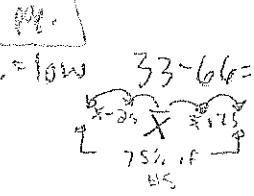
- Ex. 19.8, 43.8, 36.1, 52.4, 63.1, 20.7, 46.3 $\quad \bar{x} = 40.31$
  range = 63.1 - 19.8 = 43.3
  CV = $\frac{s}{\bar{x}} = \frac{14.82}{40.31} = 36.77\%$

x	$x-\bar{x}$	$(x-\bar{x})^2$	x	$x-\bar{x}$	$(x-\bar{x})^2$	$\Sigma = 1536.64$
19.8	19.8 - 40.31 = -20.51	420.66	63.1	22.79	519.39	7
43.8	43.8 - 40.31 = 3.49	12.18	20.7	-19.61	384.55	$\sqrt{319.50}$
36.1	36.1 - 40.31 = -4.21	17.72	46.3	5.99	35.88	$s = 14.82$
52.4	52.4 - 40.31 = 12.09	146.16				

- CV (coefficient of variation) = $\frac{s}{\bar{x}}$ (sample) or $\frac{\sigma}{\mu}$ (pop.) - can compare between populations.

for spread: - CV: 0-10% = tight · 13-30% = low 33-66 = medium 95 = high 100% or more = data all over the place

- Chebyshev's Theorem = $1 - \frac{1}{k^2}$ if $k = 2$, $1 - \frac{1}{2^2} = 75\%$ $k = 2, 3, 4$ usually



- need a one sentence conclusion: At least (75%) of the (students) would fit into the group that (volunteered) from (25.7) to (32.5) (hours each semester).

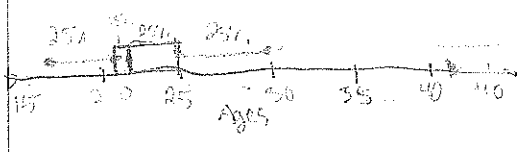
3.3:

- Median = 50th Percentile. Percentile: roadmark, placement in data set, compare ranking
- Quartiles = Special percentiles, splits data in 4ths
- 1st quartile = Q_1 = 25th percentile · 2nd quartile = Q_2 = 50th percentile = median · 3rd quartile = Q_3 = 75th percentile

To find Q_s - Arrange #s small to large · median · look @ data to left of median & find median of subgroup
- look at data to right of median & find median of this subgroup

- IQR = inner quartile range: $Q_3 - Q_1$ = 75th% - 25th% = spread of inner data
- Box and Whisker Plots

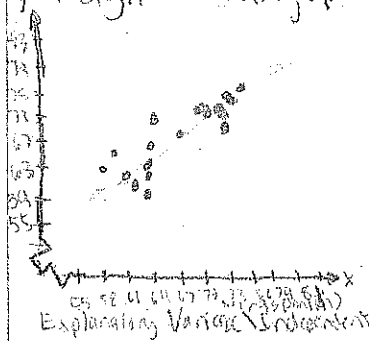
1) # line 2) median 3) LQ, UQ 4) small, large 5) draw box & line segments, (shows the length more spread data)



- min = 17 Q_1 = 21 Q_2 = 22 Q_3 = 25 max = 30 = 5 # summary
- IQR = 25 - 21 = 4 Upper Outer = $Q_3 + 1.5(4) = 25 + 6 = 31$ = no outliers
- Lower Outer = $Q_1 - 1.5(4) = 21 - 6 = 15$ = no outliers

4.1 Scatter Plots

Height v Wingspan of Students (closer, proportional)



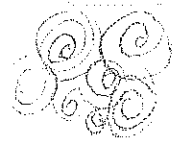
Analysis: + association

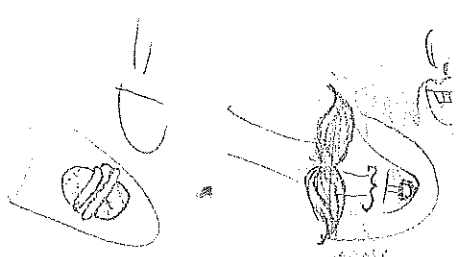
Trend: as x ↑, y ↑

Longer arms tend to be attached to longer/taller bodies

- go w/ 2 quantities = can color code, have different circle/dot marks, etc

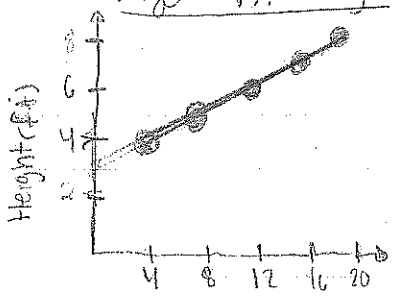
- Can have outliers that are far from line
- dot that follows line but is apart from rest of data (extreme) is called an influential observation





age (x)	4	8	12	16	20	6
ht. (y)	4	5	6	7	8	3

- $y = \frac{1}{4}x + 3$

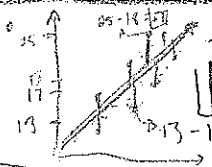


This is most likely false @ birth as the child would be 3 ft. long when it is born. The slope is $\frac{1}{4}$ & the y-int is 3. At some point, this person will stop growing. As age ↑, height ↑.

- Measures of variation: S^2 (variance) - s (standard deviation) - CV (coefficient of variation) - IQR (interquartile range) - r (range) - Chebychev's
- extrapolation: Things in future might not be same in future, so don't extend line
- Beyond parameters of data, safe to assume true if only slightly off from line (all yrs) but be sure to say nothing is certain

4.2

- Linear Regression & The Coefficient of Determination
- $\hat{y} = a + bx$ (ŷ = pred y-val, a = y-int, b = slope $\frac{\Delta y}{\Delta x}$)



- residuals: vertical distance from line to point: $y - \hat{y}$
- Line = Least squares Regression Line / Line of best fit

$$b = \frac{n\sum(xy) - (\sum x)(\sum y)}{n\sum(x^2) - (\sum x)^2}$$

n = # of lines, x = x values, Y = y values

$a = \bar{y} - b\bar{x}$ ($\bar{x}$ = mean of x values, $\bar{y}$ = mean of y values)

- put $(\bar{x}, \bar{y})$ on graph always! w/ a different mark & it above largest x-value & maybe largest y-value
- ins = influential observations & outs = outliers

4.1

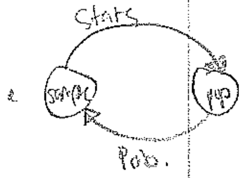
- Analysis for Scatterplot: +/- association, trend, outliers, C (# & word), r^2 = ____.
- r = how close points are to line = correlation coefficient; # between −1 & +1
- perfect correlation = −1 or +1 (sign depends on + or − slope) association
- no linear correlation = 0; • r^2 (coefficient of determination (always between 0 & 1)

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

- Since r is [close] to 1, we have an indication of a strong + linear (need) correlation between [x variable] & [y variable]
- As (x increases/decreases), (y increases/decreases). = Interpretation of r
- r^2 = (r^2(100)) % of variation in (y) is explained by the Least-Squares Regression line & variation in (x). $1 - r^2$ (%) remains unexplained. (give conjectures)

5.1

- CORRELATION DOES NOT MEAN CAUSATION!
- Probability - a measure of chance likelihood of an event happening, between $0 \le p \le 1$ (or %)
- Prob Notation: P(A) = Prob of event A "P of A" • Prob formula: # of desired results / # of total possible results
- Odds = (# of favorable / # of not favorable (f')) • P(6') = Probability of not getting 6 on die = 5/6 • S = Sample Space
- P(A') = complement = 1 − P(A); P(A) + P(A') = 1 • sample space = listing of all possible outcomes



−1⊃

Stats - sample is known • Draw conclusions about pop. based on results of an sample
Prob: pop. is known; ask "What are chances... what is likelihood on this try?" we are asking about sample
Stat reasons from sample toward pop. • Prob reasons from pop. toward sample
1- Pepsi • 2- RC • 3- Coke • S = CPR, CRP, PRC, PCR, RCP, RPC • expected # = who guesses correctly = probability × # of people in class = $\frac{1}{6} \cdot 22 = \frac{11}{3} = 3\frac{2}{3} = 4$ (whole #)

5.2

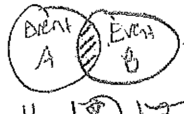
COMPOUND EVENTS

- two or more events that happen together / are dependent on each other

dependency = when things change to 1 affects the other.

Independent: $P(A \& B) = P(A) \times P(B)$ • Dependent: $P(A \& B) = P(A) \cdot P(B|A)$ (B given A) $P(B|A) = P(B)$

Or = 1 or other = sample space



And = Both events = sample space

- $P(A \cup B)$ = or = +
- $P(A \cap B) = \frac{1}{3} = \times$

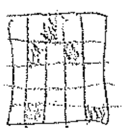
$P(\text{King} \cup \text{Ace}) = P(K) + P(A) = \frac{4}{52} + \frac{4}{52} = \boxed{\frac{8}{52}} = \boxed{\frac{2}{13}}$ ← disjoint / mutually exclusive = no overlap

$P(\heartsuit \cup K) = P(\heartsuit) + P(K) - P(\heartsuit \cap K) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \boxed{\frac{4}{13}}$ ← nondisjoint / non mutually exclusive = overlap

12 black sox — $P(\text{matching sox}) = P(B \cap B) + P(B \& B) + P(b \cap b) + P(G \cap G) + P(R \cap R) + P(B \cap B)$ (+ = or)

6 blue sox — $= P(BL) \cdot P(BL/BL) + \ldots$

8 Brown sox — $= \left(\frac{12}{36} \cdot \frac{11}{35}\right) + \left(\frac{6}{36} \cdot \frac{5}{35}\right) + \left(\frac{8}{36} \cdot \frac{7}{35}\right) + \left(\frac{5}{36} \cdot \frac{4}{35}\right) + \left(\frac{4}{36} \cdot \frac{3}{35}\right) + \left(\frac{1}{36} \cdot \frac{0}{35}\right)$



5 Grey Sox
4 Red Sox
1 Striped Sock

Fortune & Wanted

$P(1 \text{ win}) = P(\$ \cap S \cap S \cap S \cap S) = P(\$ \cdot \$/S \cdot \$/SS \cdot \$/SSS \ldots)$

5.3

conditional probability - probability that an event B will occur given the knowledge that an event A has already occurred. P(B Given A)

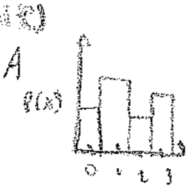
tree diagram = more attractive sample space (flows)

$P(B|A) = P(A \cap B) / P(A)$

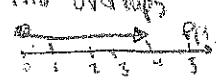
3 Red △ 2 Red ○ 4 Blue △ 4 Blue ○ $P(\triangle \text{ or Blue}) = \frac{7}{13} + \frac{8}{13} - \frac{4}{13} = \boxed{\frac{11}{13}}$

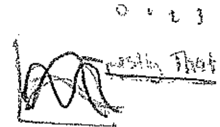
6.1

Discrete Random Variable - quantitative observations that are countable (finite)
-2, -1, 0, 1, 2 ... fractions & decimals - ex: # of students who earned an A
- histogram & bar's area of all bars added together = 1 (100%)
entire sample space present \ no gaps & mutually exclusive \ no overlaps



Continuous random Variable - countless, infinite observations





- ex: temp in room (has decimals & fractions) & time

Probability distribution - all probabilities + to 1. Has a mean & standard deviation
- μ = mean = $\Sigma x \cdot P(x)$ = expected value = balance \ central point = $\Sigma a\,p(a)$
- σ = St. dev. = $\sqrt{\Sigma (x-\mu)^2 \cdot P(x)}$

232-Ch.6.1: 10) a) P(60+ years-old) = P(60-69) + P(70-79) + P(80+) = .25 + .091 + .018 = .3590

x	24.5	34.5	44.5	54.5	64.5	74.5	84.5
P(x)	.057	.097	.145	.242	.25	.091	.018
x·P(x)	1.396	3.347	6.47	15.99	16.13	6.78	1.52

$\mu = \sum x \cdot p(x)$ → added together = $\mu = \Sigma$ = 53.72

x	x − μ	(x − μ)²		(x − μ)² p(x)
24.5	−29.2	852.64		48.6
34.5	−19.22	369.408		35.83
44.5	−9.2	84.64		16.50
54.5	.8	.64		.1869
64.5	10.78	116.21		29.05
74.5	20.8	432.64		39.37
84.5	30.78	947.41		17.053

$\Sigma(x-\mu)^2 \cdot p(x)$ = 48.6 + 35.83 + 16.5 + .1869 + 29.05 + 39.37 + 17.053

$\Sigma (x-\mu)^2 \cdot p(x) = \sqrt{186.6}$

σ = 13.66 years old

$CV = \frac{\sigma}{\mu} \cdot 100 = \frac{13.66}{53.72} \cdot 100 =$ 25.43 = low

p. 2

Binomial Experiments - Bernoulli Experiments

Jacob Bernoulli was a Swiss mathematician who studied binomial experiments in the late 1600s extensively.

The sort of problems which have exactly 2 possible outcomes is now called binomial / Bernoulli Experiments.

The central problem of a binomial experiment is finding the probability of r successes out of n trials.

Binomial experiments don't work with dependent situations. Only independent!

If asked what car is recommended to buy, n (# of trials) = # of teachers asked and # of outcomes possible = as many models of cars there are

This is NOT a binomial experiment, out 72 possibilities

$P(r) = C_{n,r} p^r q^{n-r}$ n = # of trials, p = prob. of success, q = 1 − p = prob. of failure, r = # of "successful" successes

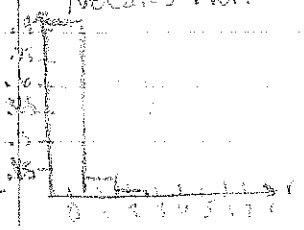
Ex. $P(r=6) = C_{10,6} (.75)^6 (.25)^{10-6} = .1460$

There exists a 14.60% chance that we will get 6 out of 10 people would be concerned about the security of the email.

If $r \geq 6$, then add probabilities of r=6, r=7, r=8, r=9, r=10.

Needles Hurt! test: n=8

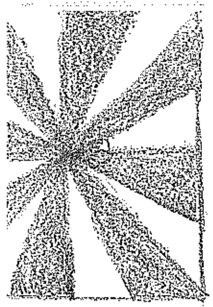
r	0	1	2	3	4	5	6	7	8
p(r)=.9	.923	.075	.003	0	0	0	0	0	0



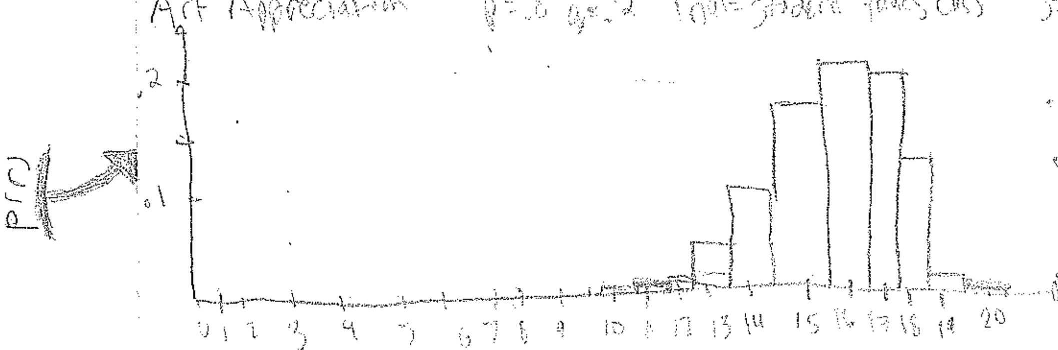
Shape:

Center: .08 = mean

Spread: .2618 = 6 = moderately high

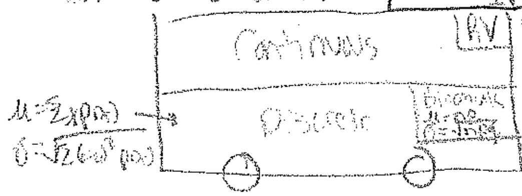


Art Appreciation Adult people learn independently. p=.8 q=.2 70%=student passes class S=Pass F=Fail n=20 r=0-20



- Center = μ = (20)(.8) = 16
- Shape = Skewed left
- Spread $\sqrt{16(.2)}$ = 1.79 = σ; low
- Outliers: 0-8

- For mean of binomials: $\mu = np$ $\sigma = \sqrt{npq} = \sqrt{\mu q}$

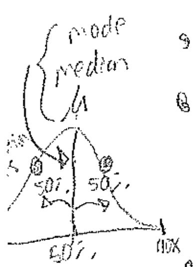


- Calculator: 2nd VARS → binompdf (just 1 r) or binomcdf (range of r's) → binompdf(n, p, r) or binomcdf(n, p, stopping r)

s not … n a bad … after all.

- Law of Large #s: Mean result of a large # of independent trials comes close to the true mean of the distribution
- myth of small #s = true mean will occur with low # of trials
- $\sigma^2_{(x+y)} = \sigma^2_x + \sigma^2_y$ variance of x & y = variance of x + variance of y

- high/low? trend & why? Analysis = bar, pie, picto, pareto, time plot, line
- CSSO Analysis = histogram, stemplot, box & whisker, dotplot
- For Scatter Analysis = +/- association, trends, influential observations/outliers, correlation (r = # & word) & r^2 = coefficient of determination (×100) = 49.73% of the variation in (y) is explained by the LSRL & the variation in "x".
- stratified = group people based on characteristics & cluster = group people based on location



Ch. 7. Continuous Random Variables - 7.1

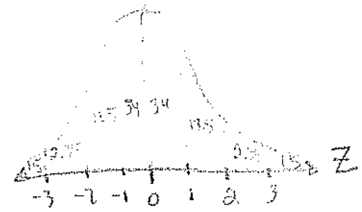
- Normal curve - area under curve = 1, curve is bell-shaped w/ highest point over mean (μ), curve is symmetrical about vertical line through μ, curve approaches but doesn't cross horizontal axis, inflection points occur @ $\mu \pm \sigma$
- empirical rule = 68% of data will fall w/in $\mu \pm \sigma$, 95% w/in $\mu \pm 2\sigma$, 99.7% w/in $\mu \pm 3\sigma$
- Need $p(19 \le x < 35)$ = 47.5%. There exists a 47.5% chance that...

.2

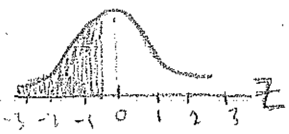
- z score = standard deviation distance from $\mu \rightarrow z = (x-\mu)/\sigma$
- raw score $x = z\sigma + \mu$ (derive from z score equation)

— ex. μ = 4.8 σ = .3 b) x < 4.2 $z = \frac{x-\mu}{\sigma} \rightarrow \frac{4.2-4.8}{.3} \rightarrow$ z < -2.00   e) 1.28 < z $1.28 = \frac{x-4.8}{.3} \rightarrow$ 5.184 <
f) x = 5.9 $\frac{5.9-4.8}{.3}$ | z = 3 2/3 ~~unusually high~~

Standard Normal Curve →



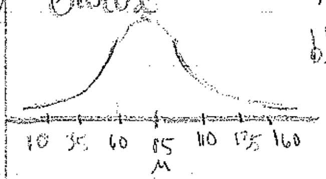
#(6) Draw curve left of $z = -.47$ & find area



-.4 → .7 → .3192 Use table

7.3

#25) Glucose



a) $P(x>60) \Rightarrow z = \frac{60-85}{25} = -1.00 \rightarrow 1 - .1587 =$ **.8413**

b) $P(x<110) \Rightarrow z = \frac{110-85}{25} = 1.00 \rightarrow$ **.8413**

c) $P(60<x<110) \Rightarrow .8413 - .1587 =$ **.6826**

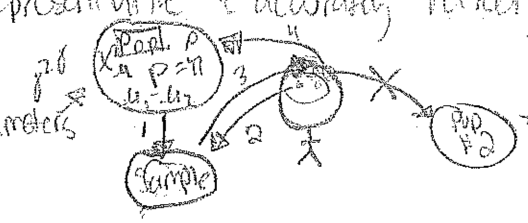
d) $P(x>140) \Rightarrow z = \frac{140-85}{25} = 2.20 = .9861 \rightarrow 1 - .9861 =$ **.0139**

There exists a 1.39% chance that an adult under 50 years old will have ≥140 milligrams of glucose in their blood.

7.4

Sampling Distributions - review of basic methods & terms

- Population - all measurements of interest
- Sample - subset of measurements from population
- Random Sample - representative & accurately reflects population
- Inferential Stats

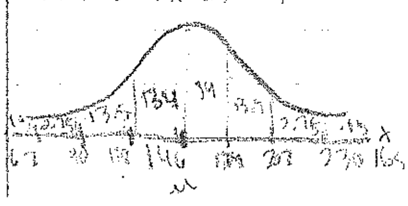


1) Get a random representative sample
2) Observe & collect data
3) Analyze results
4) Apply back to same pop
- Inferential stats can not reasonably be applied to a different population than the one in which sample was ...

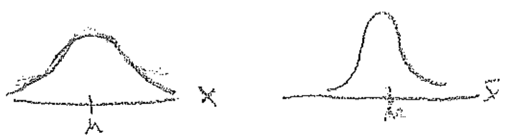
Sample Statistics
- Statistic - numerical descriptive measure of a sample. Sample proportion: $\hat{p}$ = phat = r/n & $\bar{x}, s, s^2$, etc
- Use sample stats to make inferences (conclusions / decisions) about population parameters
- Ch. 1-4: Descriptive Stats - Organize, summarize #s
- Ch. 5-7: Probability Theory & Distributions
- Ch. 8-11: Inferential Stats - Methods of using sample to obtain reliable info. about pop. (not 100% certain it will reflect popul.)

Geometric Distribution
- $P(x=n) = (1-p)^{n-1} p$ n = # we want   p = prob of success $P(x>n) = (1-p)^n$ $\mu = 1/p$
- $P(x=17) = ?$ p = .25   q = .75 $P(x=17) = (1-.25)^{16} .25$ **$P(x=17) = .002506$**
- $P(x>17) = ?$ $P(x>17) = (1-.25)^{17} =$ **.007517**
- $P(x \le 17) = 1 - P(x>17) = 1 - .007517 =$ **.9925**
- 2nd VARS → geompdf(p, n) if only 1 n   • 2nd VARS → geomcdf(p, stopping n)

- Ex. μ female ≈ 146 lbs   σ ≈ 28 lbs   ≥ 1100 lbs chair broke → caused ... for ...
  LBS of Americans



- normalcdf (lower value, upper value, μ, σ)
  - Never use normalpdf!
- invNorm (Area%, μ, σ) if given % or Area



7.5

- Standard error $= \sigma_{\bar{x}} = \sigma_x / \sqrt{n}$ $\quad z = \bar{x} - \mu_{\bar{x}} / \sigma_{\bar{x}}$
- Theorem 7.1: For a normal distribution: 1) $\bar{x}$ is a normal distribution 2) $\mu_{\bar{x}} = \mu_x$ 3) $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
- Central Limit Theorem: If $n \geq 30$, then Theorem 7.1 applies for any type of distribution

8.1

- E = error $= z_c \cdot \sigma/\sqrt{n}$ (z score × standard error) ; $\bar{x} - E < \mu < \bar{x} + E$ = point estimate
- sentence: If we took 100 samples of $n = \#$, we expect to capture population mean μ (x)(99 times) out of 100.
- $s \approx \sigma$ when n is big (≥ 30)

7.6

- If $np > 5$ and $nq > 5$, then r has a binomial distribution that is approx. normal
- If $np > 5$ and $nq > 5$, then the random variable $\hat{p} = r/n$ can be approximated by a normal random variable x w/ $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{pq/n}$ (proportions and %s) not ave.

8.2

- $E \approx t_c\, s/\sqrt{n}$ • degrees of freedom $= n-1$ *when σ is unknown*
- Ex. $n=4$ d.f. $=3$ $c = 90\%$ $\bar{x} = 33$ $s = 5$
  $E = t_c(s/\sqrt{n}) = 2.353(\frac{5}{\sqrt{4}}) = 5.8825$ $\boxed{27.12 < \mu < 38.88}$
  If we took 100 samples of size $n=4$, we expect to capture the population mean μ of brain cells of dead science teachers 90 times out of 100.
- W.S. Gossett - 1908 chemist @ Guinness brewery - found t_c.
- As n\d.f. increases, t_c decreases and goes toward z_c
- Ex. $n=14$ d.f. $=13$ $\bar{x} = 1.43$ $s = .38$ $c = 90\%$
  $E = t_c\, s/\sqrt{n} = 1.771(.38/\sqrt{14}) = .1799$ $\boxed{1.25 < \mu < 1.61}$
  If we took 100 samples of size $n=14$, we expect to capture the population mean μ of mL of depth perception 90 times out of 100.
- $n = (z_c\sigma / E)^2$ (round up to nearest whole #)
- t curve is less taller and wider than z score

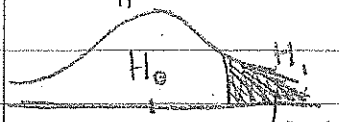
8.3

- To estimate proportion p: $\hat{p} - E < p < \hat{p} + E$ • $E = z_c\sqrt{pq/n}$ • $n = \frac{1}{4}(\frac{z_c}{E})^2$ ← this one if p estimate isn't known (if p estimate is known, rearrange this, so get n)
- If $np > 10$ and $nq > 10$ → then distribution is normal
- A normal approximation to the binomial is justified ("Anatbig") — say this!
- Sentence: If we took 1000 samples of size $n = \#$, we expect to capture population proportion p of x (950) times out of 1000.

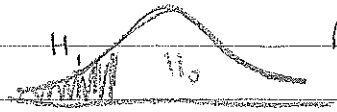
9.1

- hypothesis: an assumption or belief about a parameter
- hypothesis testing: procedure based on sample info, by which I accept/reject hypothesis
- null hypothesis: H_0 ; the hypothesis we are testing. "No change. No difference."
- Alternate hypoth: H_1 or H_A ; hypothesis to be accepted if null is rejected. "Ours is a better ..."

- Alt. hypothesis will always have 1 of 3 looks:

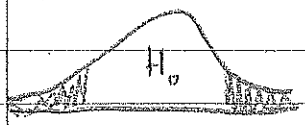
$H_A: \mu > k$ 

(one tailed test to right)



(one tailed test to left)

$H_A: \mu < k$



(two tailed test)

$H_A: \mu \neq k$

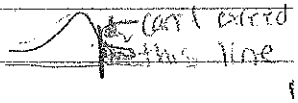
* If not given α, use .05

- A type I error is worst - when one rejects null when null is still the
- A type II error is also bad - when are accepts null when it/the larger should be

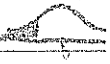
Error table:

	"Accept"	"Reject"
H_0 T	$1-\alpha$ (no error)	Type I error
H_0 F	Type II error	did right thing

- Level of significance (alpha) = α = P(willing to make a type I error)
  - $\alpha = .01$ = willing to risk Type I error 1/100 times
- Beta = β = P(making a Type II error)
- Power of a test $(1-\beta)$ = (bottom Rt. of table)
  - As $\uparrow n$, we $\uparrow$ power



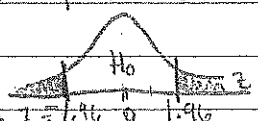
: 4 Ingredients to a Great Statistic:
- Null hypothesis - Alt. hypothesis - sample data $(\hat{p} \to z, t)$ - critical value

Ex. #12 a) $H_0: \mu = 30$ b) $H_A: \mu \neq 30$  c) $H_A: \mu > 30$  d) $H_A: \mu < 30$ 

- Conclusion: Accept \ Reject H_0, Reject \ Accept H_A, at what level of sig.
  : There exists sufficient statistical evidence to suggest that ... (H_0 \ H_A)

Ex: $\mu = 130°F$ $n = 81$ $\bar{x} = 131.08°F$ $\sigma = 1.5°F$ $\neq 130$

$H_0: \mu = 130°$ The company's claim that the fire sprinklers [remain] activated at a temp. of 130°

$H_A: \mu \neq 130$ 

$z = \frac{131.08 - 130}{1.5/\sqrt{81}} = 6.48$

conc: $\not H_0$ (H_A) $\alpha = .05$

: There exists sufficient statistical evidence to suggest that the sprinklers do indeed activate @ a temp $\neq 130°F$.

- P-value: assuming H_0 is true, probability that test statistic will take on values as extreme as \ more extreme than observed test statistic
  - Smaller the p-value computed from sample data, the stronger the evidence against H_0

Stats

- P-value = probability of chance; low P-value = results not due to random chance alone
- p-value for 2-tailed test = $2P(z > |z_1|)$
- If P-value $< \alpha$, then reject H_0. If P-value $> \alpha$, then don't reject H_0.
- There exists insufficient statistical evidence to suggest that...
  - if keep on accepting H_0 say this

9.3 Testing a population p

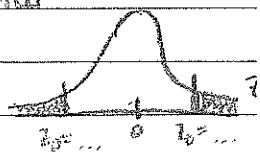
- $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$ · double ✓ that $np \geq 10$ & $nq \geq 10$ & $n \leq \frac{1}{20}N$ then write ANALYSIS and assure there is more than 10x sample size
- looking for true population percentage p (could also be μ)
- State "assuming this is a good, representative random sample" (for Ch. 8 & 9)

10.1 Hypothesis Testing w/ Dependent Graps

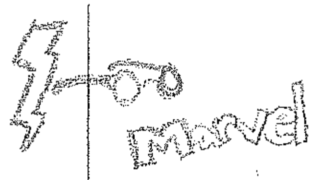
- 2 samples are dependent if each data value in 1 sample can be paired in meaningful way w/ corresponding data value in other sample
- involve paired data samples that help us draw conclusions about the difference of 2 graps
- frequently happens in before/after sits in which same item is measured prior to and after treatment (test before learning @ beginning of year; after test)
- sometimes there is a matching link, a natural match for collecting pairs even w/out a before/after feel.
- advantages: reduce danger of extra factors/variables except for 1 characteristics we want; reduces variance
- $\bar{d}$ bar ($\bar{d}$) is mean difference between paired data. s_d is sample standard deviation. (after − before)
- test is same as a t distribution. $H_0: \mu_d = 0$ (always 0; no difference before and after)
  $H_A: \mu_d \neq 0$. $\bar{d} \rightarrow t_{score} = \bar{d} - \mu_d / (s_d / \sqrt{n})$
- state why dependent

10.2 Two Dif. Populations

- 2 samples are independent if each sample is completely unrelated
- $H_0: \mu_1 - \mu_2 = 0$ or $H_0: \mu_1 = \mu_2$ (identify both μ) $H_A: \mu_1 \neq \mu_2$



- Large samples - difference of averages
  - if both $n \geq 30 \rightarrow$ CLT $\rightarrow$ Normal
- $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}} \overset{0}{=} \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$
- confidence interval: $(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$ where $E = z_{\frac{\alpha}{2}} \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$



CIIMHSMBWTHE

- $t = (\bar{x}_1 - \bar{x}_2) / s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ $d.f = n_1 + n_2 - 2$
- $E = t_c \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- Pool standard deviation only if standard deviation near =

10.3
- $n\hat{p} > 10$ & $n\hat{q} > 10$ for both populations → ANATBIJ
- $n < .1N$ for both pops. • sample = random, independent, & representative
- $H_0: p_1 = p_2$ or $\pi_1 = \pi_2$ • $z = \hat{p}_1 - \hat{p}_2 / \sqrt{(\bar{p}_1\bar{q}_2)/n_1 + (\bar{p}_2\bar{q}_2)/(n_2)}$
- $\bar{p} = r_1 + r_2 / (n_1 + n_2)$ • $\bar{q} = 1 - \bar{p}$

11.2
- asks if population fits given pattern
- $H_0: \chi^2 = 0$ The pop. fits the pattern/distribution, $H_1: \chi^2 > 0$ Pop. doesn't fit different pattern
- $\chi^2 = (O-E)^2/E$ • $E = \%$ of n • d.f. = # of categories − 1 $n = 137$

Ex.

	2005	2015 (O)	E	(O−E)	$(O-E)^2$	$(O-E)^2/E$
V	4%	3	4%×137=5.48	−2.48	6.15	1.122
B	65%	77	89.05	−12.05	145.2025	1.6306
Safety	13%	9	17.81	−8.81	77.6161	4.3580
Meds	12%	41	16.44	24.56	603.1936	36.6906
Ot	6%	7	8.22	−1.22	1.4884	.18107

~~p-value =~~ • calc: 2nd VARS → χ^2cdf(0, upper lim = crit value, degrees of freedom) → $\Sigma = 43.982$

$H_0: \chi^2 = 0$ Distribution of workers' wants of 2015 fits the distribution of workers' wants of 2005

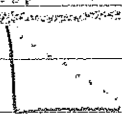
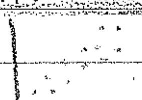
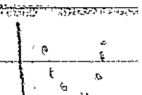
$H_A: \chi^2 > 0$ d.f. = 4

There exists sufficient statistical evidence to suggest that 2015 feelings are not fitting/following 2005 feelings.

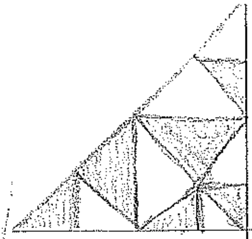
11.1 \ 2&3 − Assumptions: Each cell has expected value ≥ 5; sample is random, representative, & independent. $n < .1N$

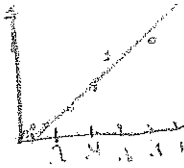
11.3
- $\chi^2_{variance} = (n-1)s^2/\sigma^2$ (d.f. = n−1) • $H_0: \sigma^2 = k(\#)$ • $H_A: \sigma^2 \lessgtr k(\#)$ Only χ^2 that isn't always > !
- (2-tailed) • To find $\chi_0^2 = \square$, go to table & look at d.f. (& d.f. / use right tail area → for left tail its $1-\alpha$)

- $H_0: \sigma^2 = k$ The variance remains the same
- There exists (in)sufficient statistical evidence to suggest that the variation in x is (indeed) (more than, less than, ≠) than claimed by (...).
- Assumptions − random, rep. independent sample; $n < .1N$

Warm up • perfect − linear correlation:  • hi + corr.  • no linear corr. 

Pg. 164 #15) • r^2 = coefficient of determination → between 0 & 1 → r^2% of variation in y is explained by variation in x & LSRL. $1-r^2$% remains unexplained.



pg. 171 a) 

- + association
- r = .9542, super high
- .9105 = ins
- trends as x↑, y↑

- b) good fit c) yes, sig.
- $\hat{y} = 17.18x - 2.48$
- residual = $y - \hat{y}_p$

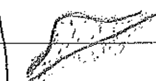
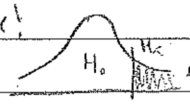
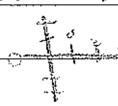
IDK

- r = linear coefficient → between -1 and 1; how lined are dots = coefficient of correlation

Analysis for graph
- ins\out, trends, + association, correlation ($r = 1 \pm$ & word)
  - ↳ influential obs. & outliers
  - ↳ as x(↑↓), y(↑↓)

Pop	Sample
μ μ_d χ^2	s $\bar{x}$ $\bar{d}$ $\hat{p}_1 - \hat{p}_2$
$p = \pi$ σ σ^2	$\hat{p}$ s^2 χ^2 $\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$ $\mu_1 - \mu_2$	r b $\hat{y} = a + bx$
$y = \rho + \beta x$ (ρ = rho, β = beta)	

11.4

- Assumptions: - set (x, y) of ordered pairs is random & normal distributions (majority in middle)  representat
- To test ρ: $H_0: \rho = 0$ x & y have no linear relationship\correlation. $H_a: \rho \gtrless 0$ [Fame]
- $t = r\sqrt{n-2}/\sqrt{1-r^2}$ d.f. = n-2
- To test β: $H_0: \beta = 0$ Pop. slope of ___ = 0 & LSRL slope is flat (⊢—) $H_a: \beta \gtrless 0$
- $t = b - B / S_e / \sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$ • t-value for ρ & β will be the same! 
- Calc: LinRegTTest & input values in Stats → Test
- Standard error = on ave. residual  = residual plot • No pattern = LSRL is fit\representative of (x,y) data
- $S_e = \sqrt{\Sigma(y - \hat{y})^2}/n-2$
- Pattern = LSRL isn't good fit & x,y aren't linear or perhaps x & y have Linear
- $S_e = \sqrt{(\Sigma(y^2) - a\Sigma y - b\Sigma xy)/n-2}$ d.f. = n-2
- Confidence interval: $y_p - E \le y \le y_p + E$ & $E = t_c S_e \sqrt{1 + \frac{1}{n} + (n(x-\bar{x})^2/n\Sigma(x^2) - (\Sigma x)^2)}$
- if α = p-value, still reject null (Ho (FTA))

Review
- Graph for discrete = histogram; $\mu = \Sigma(x \cdot p(x))$; $\sigma = \sqrt{\Sigma(x-\mu)^2 \cdot p(x)}$
- binomial = 2 outcomes w/ prob. being = , n = , ind. ; $\mu = np$ $\sigma = \sqrt{npq}$
- geometric = [illegible] $\mu = 1/p$ → until have success
- $\mu_{x \pm y} = E(x) + E(y)$ • $\sigma^2_{x \pm y} = \sigma(x) + \sigma(y) \rightarrow \sigma_{x \pm y} = \sqrt{(\sigma(x))^2 + (\sigma(y))^2}$