In A

SHELL

nut-Shell

By Annie♡

Unit 1 – Different distributions and how to describe them

Populations and Samples

- Parameters are facts about population. σ and M_x
- Statistics are drawn from samples; do tests: $\bar{x}$ and s_x from samples to infer information about population.

numerical data

- numbers

Discrete - whole #'s

continuous - #'s with fractions

Organize: dot plots, box + whisker plot, stem + leaf, Histograms, cumulative.

Categorical Data

- nouns or proportions
  ex) 60% brown hair

Open - unlimited options

closed - limited set of options

Organize: bar graphs, frequency tables.

Describing Distributions

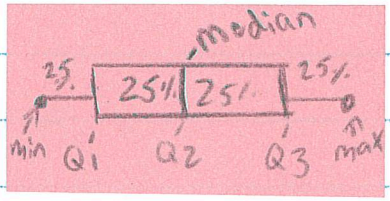
- shape: Symmetric, normal, skew left, skew right
- Center: median - more robust
  mean, IQR: $Q_3 - Q_1$
- Variability: range, Standard deviation
- Unusual features: outliers, clusters, gaps.
- Context: title, x and y axis, answer with it.

Calculator

For mean, s_x, Q_3, Q_1, etc:

Stat, edit, fill list, Calc, Vars

Box Plots



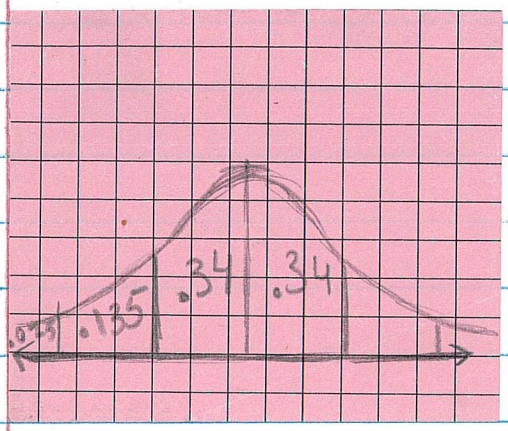
- 25% each section
- Median Center

Outliers:

IQR · 1.5 = fence ($Q_3 - Q_1$)

Q_1 - fence and Q_3 + fence

If # is outside that, it's an outlier.



Normal Graphs

- Bell Curve
- Mean in Middle
- Symmetric
- unimodle
- Empirical rule: .34, .135, .025

Z-Scores and Percentiles

Z-Score: how many standard deviations a value is away from the mean.

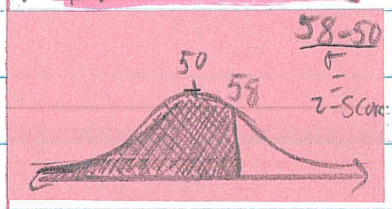
$$z = \frac{obs - exp}{\sigma}$$

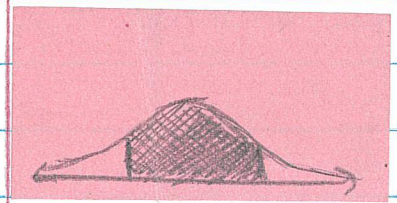
- Use to find the percent of the data below or above the value. (or inbetween values.)

Convert to % to find how much data is below/above it.

Calc: 2nd vars, normCdf, enter -10000 as lower if trying to find everything below point, 10000 as upper if trying to find all data above.

★ Draw Pictures





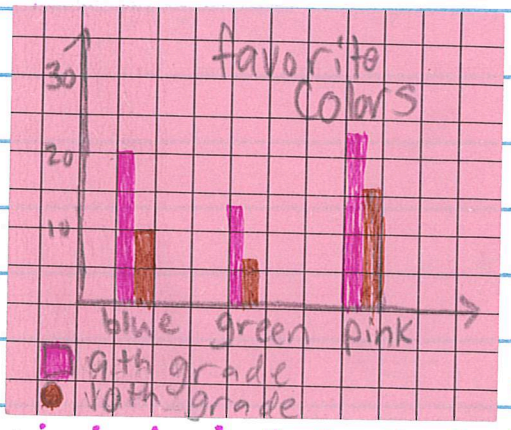
or... When finding % data between two values do the percent for both through z-scores, then subtract to get what you need.

or... Do 1 - combined value percents to find everything less than one value but greater than other value.

Unit 2 - Scatterplots, LSRL, Relative frequency, data graphs

Side by Side Graphs

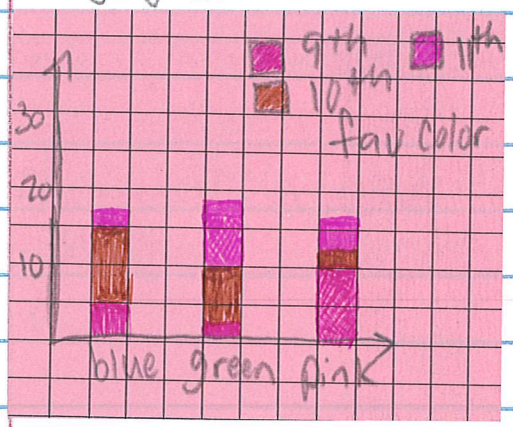
- Categorical, # values
- labels on sides
- next to each other
- swappable variables



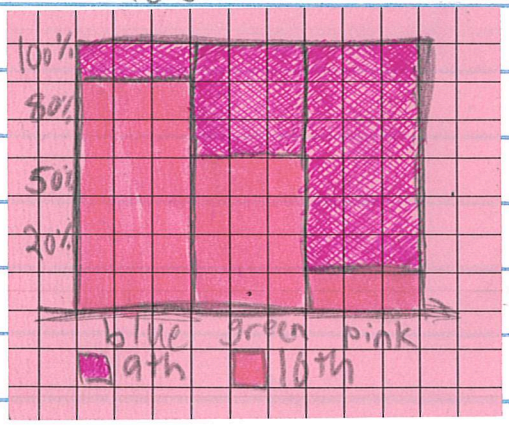
Segmented

- Swappable Variables
- 1st var on bottom, second as color on side

Segmented:



Mosaic:



Mosaic

- Uses relativ
frequencies to better compar
Variables.
- totals are not the same

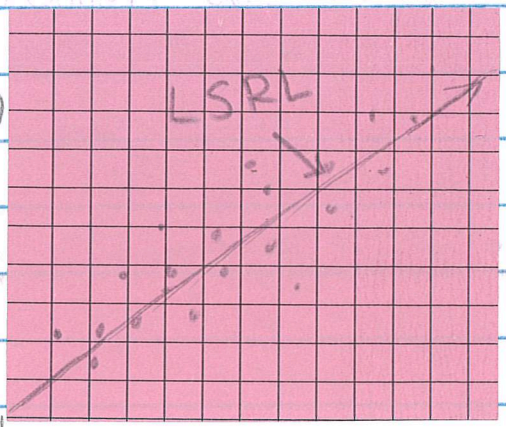
beware! The proportion of values may be higher, but the real value can be differen due to different totals.

Relative Frequency

- Also called Conditional frequency
- Probability of one event given another event must occur
- Value / column or row total
- In rel fre table, values must add to the totals for rows and columns.

Scatterplots

- Describe:
  - Direction (pos or neg)
  - unusual features: outliers, clusters, gaps
  - linear, non linear
  - strength: weak, med, strong. + Context



- Outliers
  - High leverage point
    - far from domain
    - doesn't affect residuals
    - close to LSRL
    - increases variance
  - Influential point
    - In domain
    - outside range
    - very influential on residuals
    - far from LSRL
    - A point can be both influential and high leverage

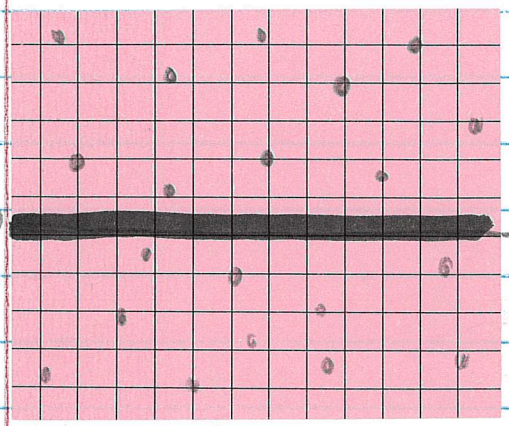
Correlation Coefficient (r) - How closely the scatterplot fits the least² Regline. From -1 to 1. -1 means perfect negative slope, 0 is no correlation, and 1 is perfect positive slope.

Coefficient of Determination (r^2) - What % of the respons variable is controlled by explainatory variable. Higher means x-var has more control over y-var. Smaller than r.

Residuals - How far observed value is from expected. Expected = plug in x value to LSRL. Ex) $y = 5.4(5) - 2$. Positive means underestimated, negative is overestimated. 0 = perfect.

Residual Graph

← linear model



- On x-axis is the observed value
- On y-axis is the residual value.
- A random pattern means linear is a good fit.
- Sres means the residuals are spread out by standard deviation of observed and residual values.

Calculator functions:

How to get residuals, expected, r^2, r, LSRL, Sres:

1. Enter x-values into L_1, y-values into L_2
2. To get **LSRL**, r^2, r: Stat, Calc, #4 LinReg
3. For **expected values**: go to L_3, enter $y=m(L_1)+b$ or the LSRL.
4. For **residual values**: go to L_4, enter L_2-L_3, or do obs-exp
5. For **Sres**: Calc, 1-var stats on L_4, Sx = Standard deviation of Res.

♡ Use this data to find out if the expected values are accurate or not, and if linear is a good fit.

Unit 3

- Essentially what type of study do you use in scenario.

Population vs. Sample

- A sample or survey can measure part of the population to conclude something about the population. When looking for answer about what a study proves ... taking a sample of population infers something about that specific population. When a survey covers whole population it is fact and not something we are inferring. Samples !Only! generalize to population they came from.

Causation vs. Correlation

- Causation: is when one event IS the result of another event. Proved that one var controls another by expiriament.

- Correlation: there is an observed correlation between variables through samples of observational study.

- Expiraments Design

Expiramental units - Subjects given the treatments. Manipulated variable

factors - What expiriament manipulates

Explainatory Var - Var manipulated

Response var - outcome of treatments

Confounding Vars - Vars related to exp var that can affect resp var. (Sometimes falsely)

- Inferences through obs. study

- Inferences are made on data collected, and deemed to be Signifigant or unSignifigant.

Signifigant - able to infer some fact about data.

IDEA: Random + bias free samples = inference for larger population.

types of studies

- Expiramental - assign treatment different ways:
  - Randomly - randomly assign treatments to subjects. Good = repeatable, a control group, limited Confounding Vars.
  - Block Designs: matches homogenous features to eliminate Conf Vars
    - Matched Pairs - matches close characteristics in pairs to eliminate possible confounding vars. Each subject in pair randomly assigned to either the expiriament or control group.
      ex) two arms, one with sunscreen, one without.
    - Randomized Block Design - Researcher divides subjects into homogeneous blocks, treatments randomly assigned to subjects in blocks
      ex) Divide subjects into blocks with same skin color, randomly assign sunscreen vs. none.
- Use expiraments to infer causation
- Observational study
  - Random - HAT METHOD! Put names in hat, shake, draw, collect samples.
  - Use the data to do different tests:
    - Z-test uses proportions of categorical data, t-test uses #'s of numerical data. One Sample for 1 var, 2 sample for different in variables.
- Use observational studies to infer correlation. Based on if data is significant from expected.

Unit 4

Different types of probability

- **Unions** - Either event happening, add probs.
  $P(A \cup B)$ → probability A or B happens
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ← to make sure you don't double count
  prob of A and B occuring simultaneously.

- **Intersections (and)** - Both events happening at same time. Multiply. Only works if Independent.
  $P(A \cap B)$ → prob A and B both happen
  $P(A \cap B) = P(A) \cdot P(B)$

- **With Replacement** -
  When drawing from group keep denominator same, and keep prob of certain event same.
  $\frac{7}{12} \cdot \frac{7}{12}$
  blue ball drawn, 2nd blue ball drawn

- **Without Replacement** -
  Change denominator and prob of certain event if drawn.
  ex) $\frac{4}{12} \cdot \frac{3}{11} \cdot \frac{2}{10}$

- **Conditional probability** - probability one event occurs given another event occurred. $P(A|B)$ → prob A given B
  divide prob A∩B by prob of B. $\frac{P(A \cap B)}{P(B)}$

- **Independent vs. Dependent**
  Independent $P(A) = P(A|B)$
  dependent $P(A) \neq P(A|B)$

- **Disjoint Events:** $P(A \cap B) = 0$ cause Mutually exclusive.

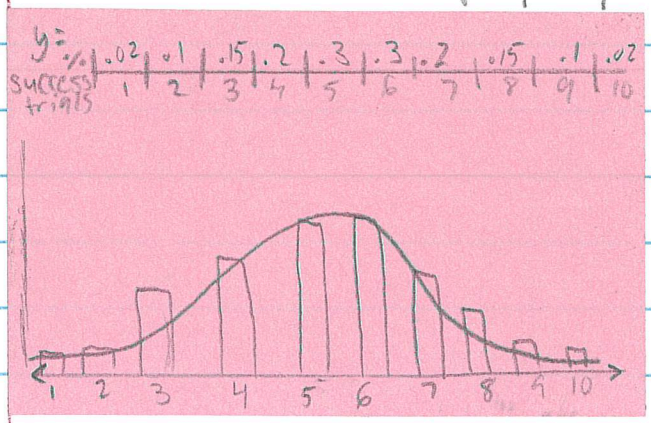
- **Mutually exclusive** - 2 events can't occur at same time
  inclusive means can happen together.

Geometric Distribution

- Solves for # of trials needed to obtain Sucess.
- $(1-ps)^{x-1}(ps)^1$ ps = probability of success.
  Expected Values = $\frac{1}{ps}$ $\sigma = \frac{\sqrt{(1-ps)}}{ps}$
- to easily see all values do $y = (1-ps)^{x-1}(ps)^1$

Binomial Distribution

- To find how many sucesses in a set # of trials.
- $_nC_r\,(p)^r(1-p)^{n-r}$ (n = # trials, r = # success, p = prob success)
- Also set equal to y to view all trials
- $M_x = n \cdot p$
- variance: $\sigma_x = \sqrt{n \cdot p(1-p)}$



After finding Mean and Sx you can set up graph and table to answer questions. If finding values greater/less than certain value, just find which values apply in the chart and add them up.

M_x = how we solve expected value

$\sigma_x = \sqrt{n \cdot p(1-p)}$ so $\sqrt{10 \cdot ps(1-ps)}$ (n = 10 trials)

Circles Combinations

- total # of circle arrangements without restrictions = $\frac{n!}{n}$
  ex) 7 people, how many ways to arrange: $\frac{7!}{7} = 720$
- If people must sit together, take the # that must sit together! times the normal equation but take away one from n for every new person that must be together. Ex) 3 people together: $3! \times \frac{5!}{5}$

Expected Value -

doesn't add to 1 →

p(x)	.2	.05	.01	.01
#	10	20	50	100

- Multiply probability by value of events. Often this involves game and how much won from game; Expected value = mean of data

ex) $.2(10) + .05(20) + .01(50) + .01(100) + .73(0)$

- Make sure the probability adds to 1; if not, see why and adjust. Possibly another factor like % 0, or divide prob by total probability there. ex) $\frac{.2}{.27} = .74$

Fundamental Counting Principal - how many possible outcomes

flip coin, roll dice → $2 \cdot 6 = 12$ outcomes

nothing repeats: SWAIN → $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ • Decreases each time

repeats: POOP: $\frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1}$ ← cause o's are same 'cause spots fill

• Decreasing #'s can be written as 5! → 5, 4, 3, 2, 1

Permutations - Order matters, line things up. Calc: Math → prob → nPr

nPr n = Starting # to choose from r = # you need

ex) line up 6 people from class of 20 → $_{20}P_6$

Combinations - When order doesn't matter calc: Math → prob → nCr

nCr n = total to choose from r = # needed

Combining Variables

- Expected Value / Combined means: $(P_x \cdot M_x) + (P_y \cdot M_y)$ probability x · mean x, + prob y · mean y
- for Combined standard deviation: $\sqrt{S_x^2 + S_y^2}$ only add variation

Scaling Variables -

- Mean: Just multiply Standard deviation: $\sqrt{x\sigma^2 + x\sigma^2}$ (scale factor)
- for adjusting means and standard deviation

for Sample, mean stays same, standard dev = $\frac{\sigma}{\sqrt{n}}$

Unit 5 ♡♡

Normal Distribution

- Normal: symmetric, mounded in middle, bell curve, from a normal population or Sample size > 30.

★ Keep discrete graphs and continuous graphs separated. If discrete, add up probability values to find what's needed. If continuous with fractions, use z-score. Real world usually continuous.

Central limiting theorem

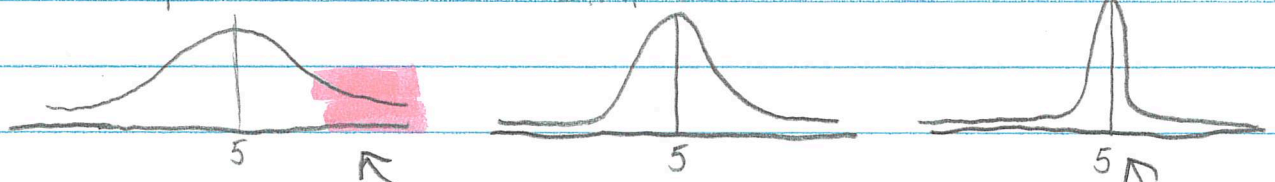
- Law of large #'s - More Samples run = closer to actual value. This means that eventually $\mu_{\bar{x}} = \mu$ (mean samples = true mean)

Essentially: $\bar{x} = \mu$ over repeated Samples (center samp. dist = pop center)

numerical → $\sigma_x = \frac{\sigma}{\sqrt{n}}$ over repeated samples because as you increase the sample size, the variance will decrease due to Central limiting theorem. (n = sample size)

Therefore... repeated samples = Same center, less variation so higher mounded data. (less spread out)

ex) 20 samples: 100 samples: 1000 samples:



Also note... less samples = larger tail ends
more samples = mounded in middle
more samples = closer to true mean

Categorical

$M\hat{p} = Mp$ Center population = Center samp dist.

$\sigma_s = \sqrt{\frac{p(1-p)}{n}}$ ← probability M, ← sample size

Conditions

numerical		Categorical	
random	Says	random	Says
Independent - small enough?	$n < .1N$	Independent	$n < .1N$ (sample size ↑, population ↑)
normal - big enough?	$n \geq 30$	normal $n \cdot p \geq 10$	and $n(1-p) \geq 10$

Bias check

- $M\hat{p} \pm Mp$. If more than 2 standard deviations away from parameter, the samples are likely biased. (If z-score of $M\hat{p}$ is greater than ± 2, it is likely biased.)

POINT ESTIMATOR – Statistic gathered from sample thats used to estimate parameter of population. Check for bias and conditions to see if statistic is good point estimator.

Sampling Distribution Differences

Numerical	Categorical
$M\hat{p}_1 - M\hat{p}_2 = p_1 - p_2$	$M\bar{x}_1 - M\bar{x}_2 = M_1 - M_2$
$\sigma_{\hat{p}_1} - \sigma_{\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$\sigma_{x1} - \sigma_{x2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
• Use to find differences in Means or SD from different groups.	• Use to find differences in proportions from different groups