1.1 Intro & Basic Term

QRMDAC

Scientific Method

1) Question
   - define population of interest

2) Research
   - what other research has been done?

3) Hypothesis / Prediction
   - what will happen if your right? wrong?
   ~~what variables do you need to measure~~

4) Collect Data
   - what variables do you need to measure

5) Analyze
   - Stats! { hypo test / Conf Intervals

6) Conclusions
   -

Population: All units of interest
- not a #
- 'target population'

Sample: subgroup of pop. from which we collect info
notation: 'n' = sample size

typically these two are the same but if they differ and unintentional it means bias

sampling unit: smallest unit selected from pop.

obs. unit: unit actually measured

variable: characteristic of units we want to learn

Parameter: summary of a variable for entire pop.
e.g. average of student height if we can measure every student

~~Sample~~

Statistic: summary of a variable for a sample

Stat. Inference: using sample info to make conclusions about the population

1.2 Types of Samples

Randomization: units selected for the sample randomly
- avoid systematic bias

Representativeness: Sample is similar to the population
- easier to say results in sample hold true to the pop

Precision: Stats from the sample are close to the true pop values

Feasibility: Sampling plan can be reasonably implemented

Biased Sample: more likely to produce some outcomes over others
- Convenience Samples: too easy
- Volunteer Response: tend to be very happy or very unhappy

- Avoid Biased Samples by using Probability / Random Sample

Probability / Random Sample: each member of pop. has a non-zero chance to be in the sample

Identically distr: all values follow same pattern, including mean & variance

Simple Random Sample:

1) every set of 'n' units has an equal chance of being picked
2) Compile list of all units: Sampling Frame
3) Select units from list using random process e.g. random # generator

Gold Std in avoiding bias
Not easy

Stratified Sample:

Pop. divided into strata
units inside a Strata are similar
Random Samples taken of each Strata

1) Compile a Sampling Frame for each Strata
2) Take an SRS from each sampling frame

Cluster Sample:

- Units in a Cluster are diverse
- Cheaper to collect; especially when units far apart
- Sampling frame needed only from clusters
- if cluster units not variable, sample may not represent pop well

1.3 Bias in Surveys & Samples

Selection Bias: only a subset of people selected to be in sample

Under-Coverage: when sampling frame does not include all of the population

- use a randomized/probability sampling method
- use a complete sampling frame

- Non-Sampling Errors & Bias can occur during entry errors or formatting

Non-Response Bias: some part of pop. refuses to participate
aka not every sampled unit do not become an observational unit
- small response rate may indicate problem

Limit By:
- multiple contact attempts
- monetary reward
- ensure anonymity

Response Bias: responses given differ from the truth

- Results from questions
  - ordering
  - illegal
  - wording
  - face to face

Limit:
- randomize order
- studies on wording
- anonymity

4 Study Design

Observational Study: no deliberate assignment; only witness current behavior

Lurking Variables: variables that influence response but not explicitly studied

* evidence of relationship does not imply causation

Response Variable: measures outcome variable
- "y"
- ~~out~~dependent variable

Explanatory Variable: explains changes in response variable
- "x"
- independent variable

Treatment: different levels of explanatory variable

Experiments: impose a difference in expl. variable to determine difference in outcome variable

Exp Units: units on which treatment assigned [e.g. 2 methods assigned to 2 different classes

Obs. Units: units on which treatments applied [but ~~the~~ its individual students we measure the response

1.5 Designing Experiments

Randomization: ~~any variable may be randomly assigned~~ experimental units are randomly assigned to different levels by a random mechanism

- this allows for all variables equal except for explanatory variable or lurking variables

Replication: when two or more exp. units are assigned to the same unit

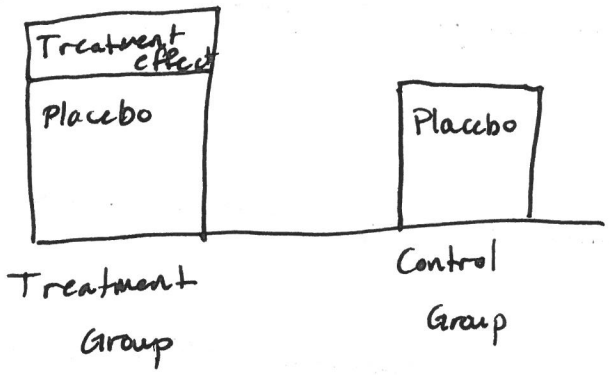
- this allows you to determine if any treatment effect is due to random chance or not; this is done by allowing one to calculate the variance of the experimental error to be compared to the size of the effect

Local Error Control: design techniques used to minimize experimental error; these errors can range from random error from natural variability to ~~S~~systematic bias

- "Control" can also mean to reference the 'control' group that allows one to determine if a treatment is more effective than if nothing had happened
- "Blocking": when exp. units are grouped based on some characteristic and ~~treatment~~ is different levels of treatment are assigned within each block; this allows for precise measurements by controlling for variation in the response due to blocking variables

- **Placebo Effect**: likelihood of a response regardless of an 'active' agent

* the placebo effect is in place for all units that receive any treatment, even if its the active agent because they expect a response... *



Obs. Studies vs. Experiments

- Obs. Studies can still have treatment & control groups
- In Obs. Studies, the subject chooses for themselves which group they want to be in
  - as opposed - to the researcher selecting the group

Random Selection vs. Random Assignment

RS: selects a random sample to represent population

RA: balances lurking variables between groups to attribute the differences to the treatment

* Both avoid bias *

1.7 Types of Exp. Designs

Good design is based in reducing variation & bi

to detect a treatment effect
- Local Err. Control
- matching
- blocking

} All help reduce variation

Completely Randomized Design:

each unit is randomly assigned to exactly one level of treatment

- simple
- only 2 variables
  - Explanatory
  - Response

- if a lot of variability it can be hard to measure a treatment effect

Matched Pair v. Block

Pair: on an individual level or often 2 units

Block: group level; multiple units per level of exp variable

Matched Pairs Design:

each unit receives each level of the explanatory variable

- a "unit" in this case could be either
  1) a single unit that ~~exists~~ is measured multiple times (order of treatments should be random!)
  2) multiple individual units that have been matched by sex, age, or other
     - usually in these pairs 1 will receive treatment and 1 will receive control and will always be compared to each other

- Reduce variability between Exp Units due to similar characteristics
- Matching can be difficult
- measurements are not independent between pairs and impacts analysis

Block Design:

- units divided into similar groups called blocks
- each level of explan. variable is applied in each block
- similar to Matched Pairs, just bigger scale
- Reduces variabi between Eu's since unit in same block are simil

Assignment of blocks is not random and is by characteristic

Assignment of treatment inside blocks is random

~~Reduces variabilit~~

- More representative samp since Eu's differ betwn blocks

- Need to account for variability with & w/o blocks

1.6 Bias in Experiments

Hawthorne Effect: people work harder when they're being monitored

Subject Bias:
- Subject might want a specific outcome
- might want to 'please researcher'

Researcher Bias: financial interest
assign subjects or report in biased manner

Limit by

Single Blinding: 1 party (subject or researcher) is not aware of which group a subject is assigned to

Double Blind: Neither party knows which treatment group a subject is assigned

Other issues

Non-response or dropout
- subjects rarely dropout in a random fashion...
the control group is more likely to drop out than a treatment group if they're very ill and the drug actually helps (thus keeping treatment group and a bigger gap between the two)

Non-adherence
- take other drugs
- not take all drugs prescribed

Generalization
- tests all college students? Can't apply to adults

2.1 Types of Data

- **Categorical:** places subject into a group aka 'qualitative'
  e.g. type of car

  Subtypes of Categorical:
  - **Nominal:** no ordering (e.g. blue, red..)
  - **Ordinal:** ordered (small, med, large)

- **Quantitative:** numeric (e.g. # of years)
  - **Discrete:** only specific values are possible (1, 2, 3, 4, 5..)
  - **Continuous:** any value possible (.001, .0001, 10.7)..
    - infinite # of possible values

- **Quant vs. Categ.:**
  - Quant. data that are categorized after can be summarized with both methods... ~~e.g. Age group 10-20, 20-30, etc.~~
  - If data was collected **in** categories to begin with, it must be treated as categorical
    e.g. Age = >18, 18-44, 45-64, 65>
    - this would be an **ordinal** variable
    - * if n/a or unknown is included as a value it then becomes a **nominal** variable, unless that category is treated as missing

Likert Scale Data: Categorical but often assigned numeric values

it can be invalid for: shorter scales (eg 5 vs 7 categories) small sam
individual items vs. combined items
(Likert type data) (Likert scale Data)

* We treat Likert as Categorical in this class

Graphical Displays:

Categorical: Bar/Mosaic; Pie charts (Counts/%'s of category)

Quantitative: Histogram, Box Plot, Scatterplot, Time Series, Heatmap

*Categorical Numeric summaries allow comparisons of individuals to the group or of a group to other groups

* Quantitative numeric summaries describe the main chunk of data w/ mean or median and variability/consistency w/ Variation, Std. Dev, range, and IQR's

2.2 Summarizing Categorical Data

- Count in Category
- $\hat{p} = \frac{\text{count}}{\text{sample size}} = \frac{y}{n}$ proportion
- tables displaying counts or %

Bar Charts: distorts vol/scale; need y-axis baseline of 0

- Display +1 or more category
- Grouped & Stacked & Mosaic represent a 2nd Categorical variable
- Mosaic's scale the width of each bar by the sample size by each group
- Flexibility in bar arrangement
  e.g. Ordinal/or in order
- Due to flexibility you can't talk about shape of a distr!

. use diverging stacked bar chart for Likert Data - or - a stacked (100%) Bar Chart

Pie Charts:

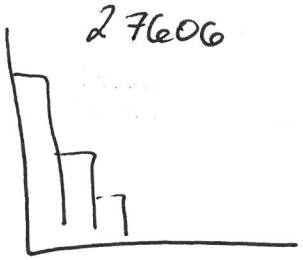
- only used if a unit can only be in 1 category
- 2nd Cat. variable can be explored w/ side by side pie charts

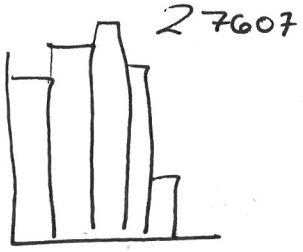
2.3 Histograms

- 1 numeric variable
- Horizontal axis: values of variable
- Vertical axis: frequency or count of variable
- Big difference between Histograms & Bar charts...
- Can compare across levels of categorical
  e.g. Sales Prices between Zip Codes

- Bars: range of values

distribution can be viewed in a Histogram





Distributions:

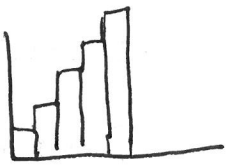
- 3 major elements to distributions

1) Shape 2) Center (Location) 3) Variability (Spread)

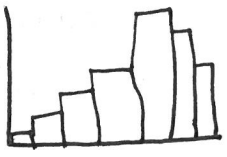
o Shape

Right Skewed: long tail to the right
(units stacked to lower limit)
but unlimited higher values

Left Skewed: long tail to the left
(units stacked to higher limit)
but unlimited lower values



Symmetric: tails are approx. even
major cluster far from limits



- modes: peaks

uniform = 0 peaks
unimodal = 1 peak
bimodal = 2 peaks

- Outliers: unusual values

Data Entry: correct if poss
remove if not

Invalid Point: remove

Actual Unusual Values: what info do
they give context to

* Careful about changing bin width

too wide: may hide important features

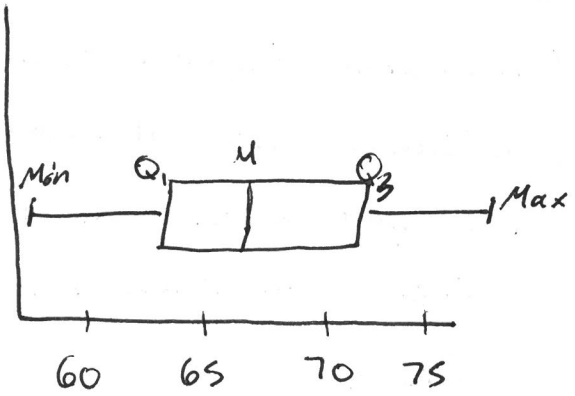
too narrow: too many features, not enough summary

2.4 Boxplots

Visual Display of the 5 number summary

1) Minimum
2) 1st Quartile Q1: 25% of values below it
3) Median Q2 : 50% of values below it
4) 3rd Quartile : 75% of values below it
5) Max

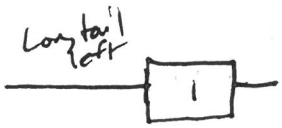
- 50% of data is inside the box



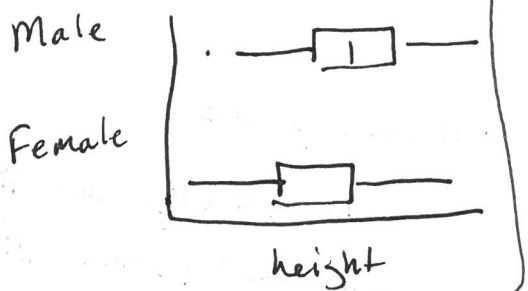
*Can't determine if multimodal

Skewed Right: Long tail to right

Skewed Left: Long tail left



- Side by Side box plots are a good way to summarize a quantitative variable within levels of a categorical variable



Timeplots:

- Scatterplot w/ Horizontal of time
- Time Series Data: Measurements taken at regular intervals over time
- Watchout for truncated y-axis!

Heat Maps:

- Represent w/ Spatial variable
- Spatial/Geospatial Data

* Watch for colors that go against convention!

Two Y-Axes Graphs:

Putting two variables measured on two different scales

2.5 Numeric Summaries of Quantitative Data

Measures of Central Tendency:

Mean, Median

- focused on main chunk of data

Measures of Variability:

Variation, Std. Dev, Range, IQR

- focused on consistency

Mean – Average: good for symmetric data / or w/ median if skewed

- Pop. Mean: μ "mu"
- Sample Mean: $\bar{y}$ "y-bar" $y = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1}{n}\sum y_i$

- sensitive to outliers
- sensitive to skewed data; will pull towards whichever side the data is skewed to

"the expected value"

- commonly used
- "typical value" – but not always the most common.

Median: good for skewed or symmetric

- middle value in a data set
- 2 middle values? take their average
- resistant to skew & outliers

"50% of ___ are above the median and 50% of ___ are below the median"

- Variance/Std Dev sensitive to skew
- Std. Dev more often reported

Measures of Variability:

Range: max – min (spread of entire data set) sensitive to outliers/skew quick looks

Interquartile Range (IQR): Q3 – Q1: spread of middle 50% (length of box of IQR)

- Q1: median of lower half
- Q3: median of upper half
- good for skewed or symmetric
- quick look

Variance: summarizes distance between all points & the mean

Pop. Variance: σ^2

Sample Varian: $s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$

Sensitive to skew/out better for symmetr

Std. Dev: sq root of variance report w/ mean & σ

"Avg distance between points and their mean or how far, on avg, values are from the mean"

2.6 Data Transformations

Adding or Subtracting: changes only measures of center

- measures of center move by same amounts (mean, median)
- measures of variability do not change (σ, σ^2)

Multiplying or Dividing: changes both

- mean, median, σ are multiplied & divided by the same amount

3.1 Random Variables & Probability

Random Variable: any characteristic whose value can change or vary between individuals

- are always quantitative
- for categorical, we assign numbers to categories (e.g. 1 for tails, 0 for heads on a coin flip)
- denoted by capital letters (X, Y, Z)
- specific values by lowercase (x, y, z)
- $P(X = x)$: "probability that random variable X equals the value x"

Probability: $P(A) = \frac{\#\text{ of ways } A \text{ can occur}}{\text{total } \#\text{ of possible outcomes}}$

- must always be between 0 and 1

Rules of Prob:

- Disjoint Addition Rule: if two events are disjoint/mutually exclusive or do not occur at the same time

$$P(A \text{ or } B) = P(A) + P(B)$$

- Multiplication for Independent: if two events do not impact the other's occurence

$$P(A \text{ and } B) = P(A)P(B)$$

Compliment of A: A and compliment of A are disjoint and means....

$$P(A) = 1 - P(A^c)$$

Types of Variables:

Discrete: outcomes are finite (0, 1, 2)

Continuous: any value in an interval (height, weight, etc.)

3.2 Distributions

Major Elements:

- Skew: Skew v. Symm, Modes, Outliers
- Center: Mean, Median
- Variability: Measured by Variance, σ, IQR

Probability Mass function: how likely a value of a discrete RV is to occur

- plug in specific values to calculate probabilities

1) Probabilities can never be negative
2) Sum of $P(X=x)$ must $= 1$

- Some PMF's are indexed by parameters
- each possible value of a parameter defines a different pmf

$X \sim$ Bernoulli(α)

"the rv X follows the Bernoulli Dst w/ parameter α"

$$P(X=x) = \begin{cases} \alpha & \text{if } x=1 \\ 1-\alpha & \text{if } x=0 \end{cases} \quad \text{for } 0 \leq \alpha \leq 1$$

e.g. $\alpha = .5$ $P(X=x) = \begin{cases} .5 & \text{if } x=1 \\ .5 & \text{if } x=0 \end{cases}$

Distributions for Continuous RV's

Probability Density Functions (pdf):

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

- $f(x)$ is just a function that defines shape
- needs to be integrated to find probabilities

Properties: $f(x)$ is non-negative

$\int f(x)\,dx = 1$ = valid pdf

$$P(.2 \le x \le .5) = f(x) \begin{cases} 3x^2 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\int_{.2}^{.5} 3x^2\,dx = x^3 \Big|_{.2}^{.5} = .117 = 11.7\%$$

4.1 Famous Discrete Distr. - ~~Bernoulli~~ Binomial

Ex Multiple choice test has 20 questions

a. probability of answering all correctly?

Multiplicative/Independent Rule from 3.1

p^{20} * where p is the probability a student answers a question correctly

b. probability of answering all incorrectly?

Compliment Rule of 3.1

$(1-p)^{20}$

c. 18 out of 20?

To capture all the different combinations of answers we need the following:

$\binom{n}{x}$: reads "n choose x" $= \frac{n!}{x!(n-x)!}$

$\binom{20}{18}$: $\frac{20!}{18!2!} = \boxed{190}$

So now we know there are 190 different ways of completing the test w 18 out of 20 correct answers... but what's the probability of getting 18/20 correct for any one way?

Probability of getting 18 correct: p^{18}

Probability of getting 2 incorrect: $(1-p)^2$

these events are independent, thus we multiply $p^{18}(1-p)^2$

Now we multiply the probability of answering the test to get a combination of 18/20 corre by the number of different ways you can do it: $(190)(p^{18})(1-p)^2$

thus, if p represents the probability of getting 18/20 correct ...

$p = .80$ and we apply that to our formula $(190)(.8^{18})(.2^{2}) = \boxed{.137}$

hus, we have the Binomial Distr

PMF: $P(X = x) = \binom{n}{x} p^{x}(1-p)^{n-x}$ Mean: np $\Sigma x\binom{n}{x}p^{x}(1-p)^{n-x}$

Variance: $np(1-p)$ $\Sigma (x-np)^{2}\binom{n}{x}p^{x}(1-p)^{n-x}$

- x can take integer values from n to x
- n = # of trials
- p = probability of success (must be same for each trial or PMF fails)

Conditions

1) Fixed # of trials (n)

2) Only 2 outcomes

3) Same prob. of success for each trial

4) Outcomes of trials must be independent

Ex) What is probability 70 heads out of 100 flips of a coin?

$p = .5$ of getting heads

p^{70} of getting heads 70 times

$(1-p)^{30}$ of getting tails 30 times

$n = 100$ thus, $P(X = 70) = \binom{100}{70}(.5)$

$\binom{n}{x} : \frac{n!}{x!(n-x)!}$

$\binom{100}{70} : \frac{100!}{70!(30)!} = ?$

$\binom{n}{x}(p^{x})(1-p)^{n-x}$

$\binom{100}{70}(.5^{70})(.5^{30})$

$= .000023$

7 heads out of 10 $\binom{n}{x}(p^{x})(1-p)^{n-x}$

$p = .5$ $n = 10$ $\binom{10}{7}(.5^{7})(.5^{3}) = .0009765625$

$x = 7$

ex| Which species of dinosaur does the skull belong to given the following info?

Species 1: . normal distr · $\mu = 619$ mm $\sigma = 3.4$mm

Species 2: normal distr · μ 641 mm $\sigma = 10$mm

*Remember 68% of all values take place w/in 1 σ

Species 2 since the distribution of species 2 skulls has a higher frequency than species 1 due to its larger std. dev & std score is smaller

$$Sp_1: z_1 = \frac{629 - 619}{3.4} = +2.94$$

$$Sp_2: z_2 = \frac{629 - 641}{10} = \boxed{-1.2}$$

smaller absolute value

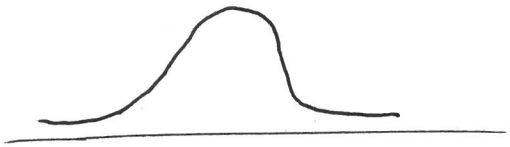
4.2 Normal Distr.

aka Gaussian

PDF: $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

Notation: $X \sim N(\mu, \sigma)$

- Bell shaped, unimodal, symmetric



- X can take any values $-\infty < x < \infty$
- μ must be $-\infty < \mu < \infty$
- σ must be ≥ 0

- 68% of possible values within 1σ of the μ
- 95% of possible values w/in 2σ of the μ
- 99% of possible values w/in 3σ of the μ

Distr has about 6σ's overall

Good guess for min value: $\mu - 3\sigma$

Good guess for max value: $\mu + 3\sigma$

Standard Scores:

To compare values from different scales, a uniform measurement is used in Standard Scores: $z = \frac{X - \mu}{\sigma}$

It is also found that if:

X is a normal distr. w/ μ and σ

then the Std. Score follows a normal distr. w/ $\mu = 0$ & $\sigma = 1$

called Std. Normal distr.

4.3 Probabilities & Percentiles for a Normal Distr.

Recall: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ represents a normal distr. curve

Also, remember we can integrate this to get probabilities/percentiles

Recall: People used to standardize each normal distr. using the Std. Sc

$z = \frac{X - \mu}{\sigma}$ since we didn't have computers

- Probability represents area under a curve

Methods to calc:

SAS: DATA temp; prob_x = cdf ('normal', UB, mean, std_dev);

PROC PRINT; var prob_x; run;

Online Calc: stattrek.com/online-calculator/normal.aspx

# of interest

Std. Score(x) 480

Cum. Prob. ____

Mean 430 μ

Std. Dev. 40 σ

Calculate

Cum. Prob = .894

E.g. Math test follows a normal distr

$\mu = 430$

$\sigma = 40$

Janice scored 480, what % of students scored below?

Finding Percentiles:

value of a variable such that a specified % is below that value

e.g. 75th percentile: X is a value such that 75% of area under the curve is less than X

SAS: * returns a z-score * will have to plug value into $z = \frac{X - \mu}{\sigma}$ for X to find value

```
DATA temp; z = probit(proportion); x = (z * std_dev) + mean;
PROC PRINT; var x; run;
```

Online Calc: Statkrek

Eg. what score is needed to be placed in the 90th percentile?

Std. Score (z) ____

Cum. Prob. .9

Mean 430

Std. Dev 40

Calc → P(X < 481.262) = .90

Cum. Prob = 481.262

4.3 Examples:

Schools Scores

$\mu = 75$

$\sigma = 5$

a) Prob of a school scoring below 70?

Online Calc

Std. Score	70
Cum. Prob	____ = .159
Mean	75
Std. Dev	5

$P(x < 70) = 15.9\%$

b) Prob of school scoring above 83?

Std. Score	83
Cum. Prob	____ = .945
μ	75
σ	5

$P(x > 83) = (1 - .945)$

$= .055$

or

5.5%

c) What score is required to be in top 20%?

Std. Score	____ = 79.208
Cum. Prob	.80
μ	75
σ	5

* set to 80% since to be in top 20% we need to be ahead of 80% of area under curve

4.4 Distributions for Sample Statistics

Sampling Variability: variation in sample statistics that results from selecting different random samples

Sampling Distr. of that statistic: consists of every value a sample stat could take based on every different possible sample taken

* As sample size increases, std. dev of sample means decreases

* As pop. std dev increases, variability in sample means increases

Std. Dev of Sample Mean: $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ ← pop std. dev.

"sigma of y-bar"

$$\sigma_{\bar{Y}} = \frac{2.8}{\sqrt{100}} = .28$$

$\mu = 65$ (pop. mean) $\quad \sigma = 2.8$ in (pop. std dev) $\quad n = 100$ (sample size)

Interpretation: the possible values of the sample mean are about .28in, on average, from the true value of the pop. mean

Right Skewed Pop:
Means are normal when n is large

* **Bi-Modal Pop:**
Means normal when $n > 30$

Bell Shaped Pop:
Means are normal even when n is small

4.5 Sampling Distr. of the Mean

Central Limit Theorem:

a distribution of sample means can be expected to be well modeled by a normal distr. if the sample size is large ($n > 30$) and

have: . mean: μ . Std Dev: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ ⇐

(Std Dev of Sample Mean)

* note that a distribution of samples will still look like the pop dist and the CLT does not apply

Examples: height follows normal distr. $\mu = 68$ in $\sigma = 3$ in

a) What is prob they're over 72 in tall?

X: 72

CumProb: .909

Mean: 68

StdDev: 3

$1 - .909 = .091$

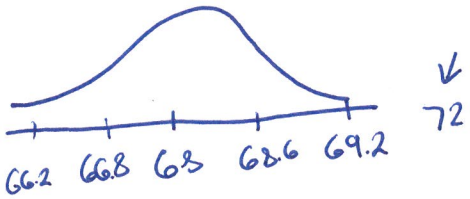
9.1% Chance that an individual from this pop. is taller than 72 in

b) Select a random sample of 25 from pop. What is the prob sample mean is > than 72 in?

Almost no chance the average from a sample of 25 people from this pop. would

$\sigma_{\bar{y}} = \frac{3}{\sqrt{25}} = .6$

$N(68, .6)$



1.6 Sampling Dist. of the Sample Prop

$$\hat{p} = \frac{\#\text{ of successes}}{\text{total }\#} = \frac{y}{n}$$

↑ Sample prop

- as sample size ↑, variability ↓
- Std. Dev of the Proportion depends on the population proportion
  - largest at $p = .5$ because outcome is most uncertain
  - smallest when the probabilities get closer to 0 or 1, e.g. @ .3 or .7

Std. Dev. of Sample Prop. :

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

↑ "sigma of p-hat"

ex 20% of city water customers are commercial

$n = 100$; What is the std error of the prop. in this situation?

$$\sigma_{\hat{p}} = \sqrt{\frac{(.2)(.8)}{100}} = .04$$

Interpretation: Possible values of sample proportion are about 4% away from the true value of population proportion, on average

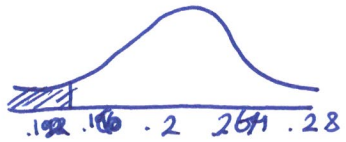
- Normal is good model if $np \geq 10$ and $n(1-p) \geq 10$
- Sample must be random and large
- Shape: approximately normal
- Center: on p (probability) ← $\mu_{\hat{p}} = .2$ since probability is 20%
- $\sigma_{\hat{p}} = .04$ due to calculations in the example above

Ex What is the probability the sample proportion is < 14%?

6.7% of possible sample proportions will be less than .14

$X = .14$
← Cum Prob: .067
Mean = .2
Std. Dev = .04

$N(.2, .04)$



5| Confidence Intervals

5.1 Point Estimation

We have gone over several pieces of the Scientific Process ... Step 4: Gathering Data & Step 5: Evaluating Data (Sample Sta

now we need to start applying the evaluations and drawing conclus with statistical inference between the sample and the populati our first tool to help us is... (while respecting sampling variabilit

Confidence Intervals: a range of reasonable estimates of the true population parameter based on the Sampling Distr.

Hypothesis Tests: provides evidence of a statistically significant effect in a population

Statistical Modeling: (regression analysis) explores the relationshi between variables, used for estimation & prediction

the Sampling Dist. & C.I. work together to connect the sample statistic to the best guess on the population parameter which we refer to as the point estimate

recall sample mean:

- random sample
- large $n \geq 30$
- Shape: normal
- Center: pop. mean μ
- Std. Dev: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

recall sample prop:

- random
- $np \geq 10$ and $n(1-p) \geq 10$

Shape: Normal

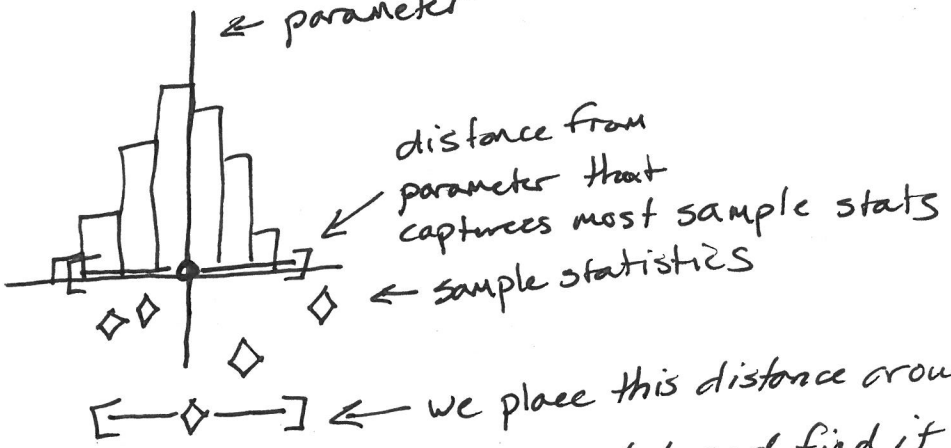
Center: Pop. prop p

Std. Dev: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

5.2 MoE

Margin of Error: numeric indicator of how far a statistic could be from a true parameter

we determine a distance that would capture most values in a sample's distribution, then we place that distance on all the sample distributions that still capture most of the potential values of a true parameter



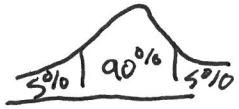
* the size of the MoE depends on the ~~means~~ std deviations of the statistic

for means: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

for prop: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

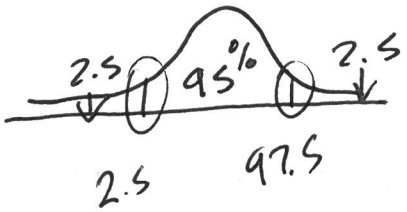
□ **How large should MoE be?** Depends on the % of the sampling distribution we would like to capture... called the **multiplier** or **confidence coefficient**

Ex $N(0,1)$
Divide N so that 90% of area is in the middle



	5th percentile	95th percentile
Std. Score:	-1.645	1.645
C.P. :	.05	.95
μ :	0	0
σ :	1	1

Ex Divide $N(0,1)$ so that 95% is in the middle



2.5^{th} percentile — Std. Score: -1.960

C.P.	2.5
μ	0
σ	1

97.5^{th} percentile — Std. Score: 1.960

C.P.	97.5
μ	0
σ	1

Ex 99%?

$.5^{th}$ percentile — Std. Score: -2.576

C.P.	.005
μ	0
σ	1

99.5^{th} percentile — Std Score: 2.576

C.P.	.995
μ	0
σ	1

95% in Middle — 2.5^{th} / 97.5^{th} percentile

$N(0,1) = 1.96$

99% in middle — $.5^{th}$ / 99.5^{th} percentile

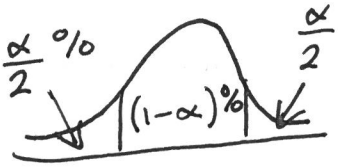
$N(0,1) = 2.576$

MoE: Multiplier x Std. Deviation $\left(1.96 \times \sigma_{\bar{y}}\right)$ $\left(2.576 \times \sigma_{\bar{y}}\right)$

* Std. Dev. of the statistic aka std. error

Proportions:

Multiplier = $z^*_{\alpha/2}$ (Mult) $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (Std. Dev.)



- found using
- $\alpha/2$ found by calculating tails

Means:

s = sample std. deviation

σ is known: $z^*_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

not known: $t^*_{\alpha/2, n-1} \times \frac{s}{\sqrt{n}}$

5.3 The t-distribution

the t. dist. is used because we rarely know the true pop. Std dev. as is used in the MoE formula for means

$$z^*_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$ ← pop std dev

our solution was to use 's', the sample std dev... but that is a 2nd estimation that isn't accounted for in the first MoE formul

to accomadate this 2nd estimation, let's see how big the difference is in the Std. Scores using Pop Std Dev & Sample Std. Dev

$\frac{y - \mu}{\sigma}$

Pop Std Dev w/ Z-score

Most values are between -3 and +3 std. deviations

$\frac{y - \mu}{s}$

Sample Std Dev w/ Z-score

Most values are between -4 and +4 std deviations

Thus we can see there is more variability in the distribu of Std. Scores of the Sample Std. Dev.... which means the normal distribution doesn't capture it well.. we need a new distr

... the t-distr

t-distr:

- commonly used for conducting inference for a pop mean
- models additional uncertainty
- bell shaped
- unimodal
- symmetrical

PDF: $$f(x;v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{\frac{-v+1}{2}}$$

- has "gamma" coefficient: Γ
- we use tech to help us find the percentiles since Γ is hard to solve/integrate
- depends on degrees of freedom (df)

v = df $(n-1)$

means the # of values in calculation of a statistic that are free to vary

logic:

Ex) $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

if you know the first n-1 data values and the sample mean, then the last data value can be calculated by...

$$y_n = n\bar{y} - \sum_{i=1}^{n-1} y_i$$

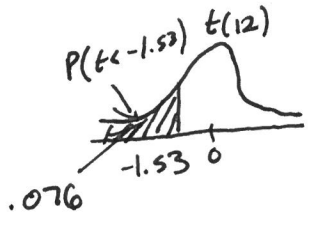
... thus n-1 values allowed to vary

* as df grows larger, the t-distr becomes more like the normal distr $[N(0,1)]$ you know everything about t-dist, if you know v aka df

- as df $\to \infty$, t-distr converges to $N(0,1)$

Ex) For t-dist w/ df = 12, find $P(t < -1.53)$

- mean of t-distr: 0
- variance: $\frac{v}{v+2}$



df: 12

t-score: -1.53

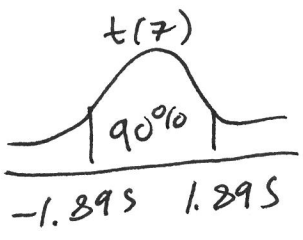
Probability: .0760

• Finding percentiles w/ t-distr

Ex| t(7), multiplier for 90% Confidence

rv t-score

dF: 7



5th $P(t < ?) = .05$ t-score: -1.895

95th

$P(T \leq t) = .05$

* report positive value so 1.895