***COURSERA: STATS W/ R SPECIALIZATION***

***COURSE 1 - Introduction to Probability and Data***

**WEEK 1- Introduction to Data**

* **Numerical variables** = **quantitative** = sensible to add/subtract, take averages, etc.
* **Discrete** --> one of a *specific set* of numeric values, so we can account for/enumerate/count ALL numeric possibilities (ex: # of cars in lot) = *COUNTS*
* **Continuous** --> infinite # of values w/in given range (height, weight)
* \* *Rounding of continuous can make a variable LOOK discrete*
* **Categorical variables** = **qualitative** = *limited #* of distinct categories, which can be IDed w/ numbers but are not sensible to do mathematical operations on (1 = male, 2 = female)
* **Ordinal** - levels w/ *inherent ordering* (not satisfied, satisfied, very satisfied)
* **Regular Categorical** - levels w/ *NO* inherent ordering (morning, afternoon, or night person)
* **Observational**: study in which we collect data in a way that does not directly interfere w/ how data arises ("observes")
* can ONLY establish **ASSOCIATION/CORRELATION** between **explanatory/dependent variables/predictors** and **independent/response variables/outcomes**
* RETROSPECTIVE --> data from past
* PROSPECTIVE --> data collected as study goes on to predict future
* **Experiment** --> random assignment of subjects to Tx.s in order to assign causal connections between variables
* **Confounding Variables** = extraneous variables affecting both predictors and response to make it seem like there is some relationship between them
* What determines if we can infer **causation** is the *type of study*
* Observation = correlation Experiment = causation
* There’re *some* more advanced methods (**causal inference**) that allow us to make CAUSATION statements *from observational studies*
* Conducting a **census** (entire population) takes lots of resources, some individuals may be hard to locate/measure and may be different from the rest of the population
* Ex: US Census --> illegal immigrants often not recorded properly: tend to be reluctant to fill out census forms, w/ the concern this info could be shared with immigration
* If they possess characteristics different than the rest of the population, not getting info from them might result in very unreliable data from geographical regions w/ high concentrations of illegal immigrants.
* Also, populations rarely stand still --> never really possible to get a perfect measure.
* **Sampling** is actually quite natural.
* **Convenience sample bias =** easily accessible individuals are more likely to be included in the sample.
* **Non-response** happens if only a *non-random fraction* of the *randomly sampled people* respond to a survey, such that the sample is no longer representative of the population
* Ex: lower socioeconomic status = less likely to respond to a city survey
* **Voluntary response bias =** sample consists of only people who volunteer to respond b/c they have strong opinions on an issue.
* Ex: online polls --> people who responded definitely do not make up a representative sample of a world population = people who happen to have visited a website the day the poll was posted and felt strongly enough to vote.
* Difference between *voluntary* response bias and *non*-response bias
* Non-response = *has a random sample*, but people who respond are not representative of the sample
* Voluntary response = *no initial random sample.*
* Ex: Literary Digest polled 10M + got responses from ~2.4M (reliable polls in US routinely poll about 1.5K people, so this = a huge sample)
* Showed Landon = overwhelming winner + FDR = 43% of the votes.
* FDR won w/ 62% b/c magazine surveyed OWN readers, registered automobile owners, + registered telephone users.
* These groups had incomes well above national average of the day (great depression era) + therefore were far more likely to support Republicans than a truly "typical" voter of the time
* While poll was based on a huge sample, it was biased + did not yield an accurate prediction
* Sampling methods
* **Simple Random Sampling (SRS)** = randomly select cases from population such that each case is *equally likely to be selected* (randomly draw names from a hat)
* **Stratified Sampling** = 1st divide population into homogenous groups (**strata**) + randomly sample from *within each stratum*.
* Ex: Both genders are equally represented in a study -> divide population into males + females + *then* randomly sample from within each group.
* **Cluster Sampling** = divide population into **clusters,** *then*randomly sample a few *clusters*, and *then* sample ALL observations w/in these random clusters.
* **Clusters**, unlike strata, are *heterogeneous* w/in themselves, + each cluster is similar to another, such that we can get away w/ just sampling from a few of the clusters.
* **Multistage sampling** = adds another step to *cluster* sampling
* Divide population into clusters, randomly sample a few clusters, + then *randomly sample observations from w/in these clusters.*
* Usually use cluster + multistage sampling for economical reasons.
* Might divide a city into geographic regions that are, on average, similar to each other + then sample randomly a few of these regions,
* Then go to these randomly picked regions + sample a few people from w/in these regions to avoid need to travel to ALL regions in the city.
* **QUIZ**
* ARetail Store randomly samples 1K of their credit card holders to survey via phone during work hours, so a lower rate of response is recorded from members who work in these hours
* **NON-RESPONSE BIAS** = initially random, but not everyone in random sample is reached
* 4 Principles Of Experimental Design
* **Control** = compare *Tx.* of interest *to a control* group
* **Randomize** = randomly assigning subjects to Tx.s
* **Replicate** = collect sufficiently large sample w/in the study or in order to replicate entire study
* **Block** = If there’re variables known/suspected to affect the response variable, 1st group subjects into blocks *based on these variables* + then randomly place cases from *w/in each block* to Tx. groups.
* Ex: Design experiment to investigate if energy gels make you run faster.
* Tx. group gets energy gel while control group = does not
* Its suspected energy gels might affect pro + amateur athletes differently, so block for pro status
* *Divide sample* into pro + amateur athletes (2 groups) + then randomly assign both pro + amateur athletes to both Tx. + control groups
* Therefore pro + amateur athletes are equally represented in both resulting Tx. + control groups
* Now if we DO find a difference in running speed between Tx. + control groups, we can attribute it to the Tx. + be assured the difference is NOT due to pro status (since both athletes were equally represented in both the Tx. + control groups)
* Difference between a **blocking variable** and an **explanatory variable**:
* Explanatory variables (**factors/predictors**) = conditions *we can impose* on our experimental units
* Blocking variables = characteristics experimental units *come with* that we’d like to control for
* Blocking is like stratifying, except used in *experimental* settings *when randomly assigning* as *opposed to when sampling.*
* Ex: Study designed to test light + noise level on exam scores, while assuming both might have different effects on males + females
* Want both genders equally represented
* We have 2 explanatory variables (light, noise), 1 blocking variable (gender), 1 response variable (exam score)
* **Placebo** = fake Tx. often used as the control group for medical studies.
* **Placebo Effect** = when experimental units show improvement simply b/c they believe they’re receiving a special Tx.
* **Blinding** = experimental units do not know whether they are in the control or Tx. group
* **Double-Blind** = BOTH experimental units + researchers do not know who is in the control or Tx.
* **Random sampling** and **random assignment** sound similar, but serve quite different purposes in study design.
* **Random sampling** *occurs* when subjects are *being selected for a study.*
* If subjects are selected randomly from a population, each subject is equally likely to be selected + a resulting sample is likely representative of the population
* So the study results are generalizable to the population at large.
* **Random assignment** occurs *only in experimental* settings + at the time when *subjects are assigned to various Tx’s*
* Usually subjects in a sample exhibit slightly different characteristics from one another.
* Through random assignment, we ensure these different characteristics are represented equally in both Tx. + control groups
* This allow us to attribute any observed difference between Tx. + control to the Tx. being observed, since otherwise these groups are essentially the same.
* In other words, random ASSIGNMENT allows us to make *CAUSAL* conclusions based on a study.
* Ex: Want to conduct a study evaluating whether people read serif fonts or sans serif faster.
* Ideally, 1st randomly sample subjects from a population, *then* randomly assign subjects in the sample to 2 Tx. Groups
* 1 = read some text in serif font
* 2 = read same text in sans serif font.
* Through random *assignment*, we ensure other factors (confounder) that may be contributing to reading speed (ex: fluency, how often a subject reads for leisure, etc.) are represented equally in the 2 groups
* In this setting, if we observe any difference between average reading speeds of the 2 groups, we can attribute it to the Tx. (font type) + know it is likely not due to a confounder
* So to recap, **sampling happens first + assignment happens second.**
* *Ideal experiment employs both random sampling AND random assignment to get causal conclusions that can be generalized to a whole population*
* But such studies are usually difficult to carry out, especially if experimental units are humans, since it may be difficult to randomly sample people from a population + then impose Tx’s on them.
* Typical experimental study = recruit *volunteer* subjects which is NOT *random* sampling, but they apply random *assignment*
* Gives *causal* conclusions, but those that *only apply to the sample* + cannot be generalized
* Typical observational study = DOES use random sampling but NO random assignment
* Gives *correlation* statements but results CAN be generalized to the population at large.
* UN-ideal observational study = does not use random assignment OR random sampling + can only be used to make correlational statements that are NOT generalizable