Review of Basic Probability Concepts (Chapters 1-4)

The goal of this course is to learn about *statistical inference*. Why do we need probability?

Statistical inference: learn about a population based on the information provided by a sample.

Population:

- Characterized by the probability distribution of a *random variable* X. Think of X as the value associated with a randomly selected individual from the population.
- The distribution of X depends on some unknown parameter θ.
- Goal: learn about θ.

Ex: What proportion of American voters think President Trump will do a good job? Describe the population random variable and the parameter of interest.

Population: American ~~[illegible]~~ voters' opinion regarding President Trump.

Associate an indicator variable (X) with each member of the population.

$$X = \begin{cases} 1, & \text{individual thinks that the president will do a good job} \\ 0, & \text{otherwise} \end{cases}$$

Population: $\{1, 0, 1, \ldots\}$, proportion of 1's in the population = prop. of American voters who think that the president will do a good job. $= p$ (this is what we wanted) (parameter of interest)

↓

Randomly select one individual from this population – X is associated with this.

Possible values: 0, 1. $P[X=1] = p$, $P[X=0] = 1-p$.

$P[X=x] = p^x (1-p)^{1-x}$, $x = 0, 1$

"follows"

$X \sim \text{Bernoulli}(p)$ parameter

"probability mass function"

A general framework for inference

Population: $X \sim f_\theta(x)$ ← probability dist. of X

θ = unknown parameter

↓ random sample

$\underbrace{X_1, X_2, \ldots, X_n}_{\text{random variables}}$ (n = sample size)

Def: $X_1, \ldots, X_n$ form a random sample if

(i) Each X_i has the same distr. as X. — $X_1, X_2, \ldots, X_n$ are identically distributed

(ii) $X_1, X_2, \ldots, X_n$ are independent

Random sample: $X_1, \ldots, X_n$ are IID.

NOTE: with replacement sampling: By def gives random sample

without replacement sampling: By def does not give identically distributed $X_1, \ldots, X_n$

If size of the population (N) is very large compared to the size of the sample (n),

without replacement $\approx$ with replacement sampling

"practically the same"

~~OK~~ Rule of thumb: OK if $\frac{n}{N} \leq 0.05$

(Otherwise some 'finite population adjustment' is needed)

NOT of interest in this course

<u>Statistical inference</u>: Learn from the data $X_1, \ldots, X_n$ about the parameter θ.

Two basic types of random variables

Discrete random variable: A rv X is *discrete* if the set of all its possible values is countable.

Ex:

- X = # hits on the UD website in a day; possible values: $\{0, 1, \ldots\}$ infinite but countable set
- X = # days it rains in the month of July
  possible values: $\{0, 1, 2, \ldots, 31\}$ countable set
- X = binary outcome
  possible values = $\{0, 1\}$.

Continuous random variable: A rv X is *continuous* if

Ex:

-
-
-