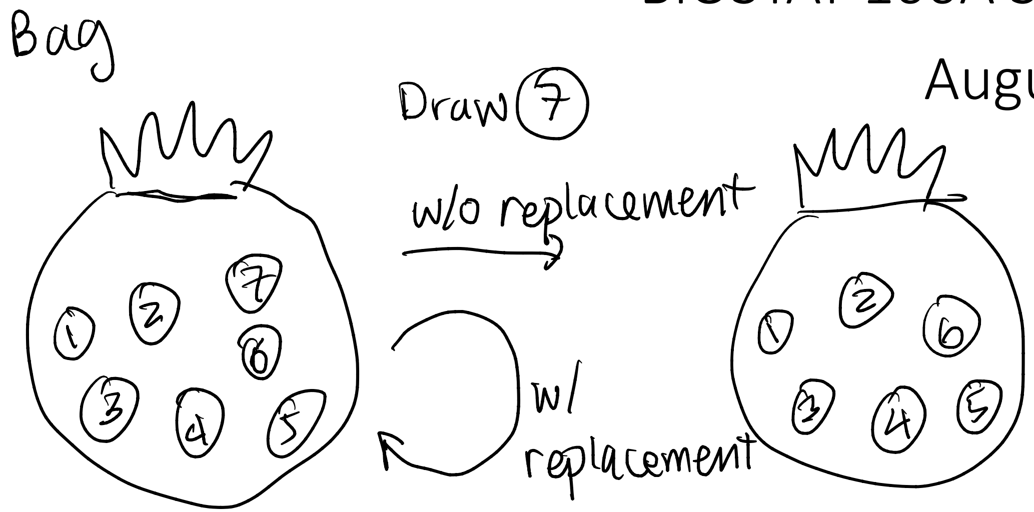
Week 2 Review

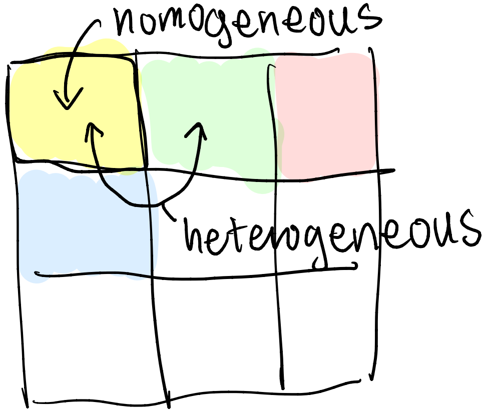
Cindy J. Pang

BIOSTAT 100A Summer Session C 2024

August 16, 2024



Probability (Random) Sampling

Type of Sampling	**When/Why** do this type of Sampling randomization	**Selection Mechanism,** or **How** to conduct this type of Sampling
① Simple Random Sampling (SRS) every sample has an equal chance of being selected (1) w/o replacement - we don't put the draw back (2) w/ replacement - put the element back.	convenience Assumption - assume respondents in the same are homogeneous. This is problematic b/c most pop'ns are not homogeneous	(1) Population Listing / Pop'n Frame (2) Assign unique ID to each person in the frame (3) Draw randomly.
② Stratified random Sampling  	When: there are homogeneous subpopulations (strata) / covariates within the pop'n you are interested in ⇒ homogeneous within; heterogeneous across. $n_1 = \left(\frac{N_1}{N}\right)n = \left(\frac{30}{100}\right)50 = 15$ N = pop'n size $n_2 = \left(\frac{N_2}{N}\right)n = \left(\frac{60}{100}\right)50 = 30$ n = sample size	÷ the pop'n into strata and take a SRS of each strata. Ex: Proportional Allocation - how to select the # of people to include in a strata Suppose you have pop'n w/ 3 strata $= \{G_1, G_2, G_3\}$ where $N = 100$ and $N_1 + N_2 + N_3 = N$, but we can only sample 50 people

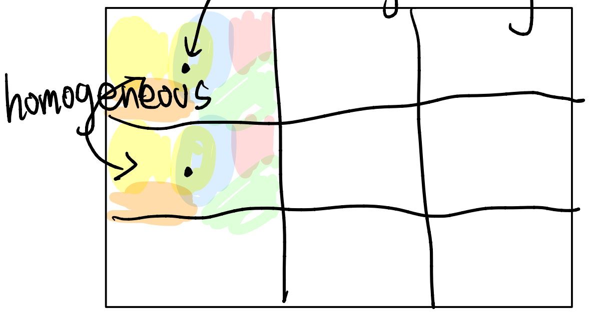
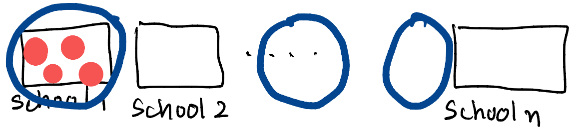
$n_3 = \left(\frac{N_3}{N}\right)n = \left(\frac{10}{100}\right)50 = 5$ $N = 100$

G_1	G_2	G_3
$N_1 = 30$	$N_2 = 60$	$N_3 = 10$
$n_1 = 15$	$n_2 = 30$	5

$n_i = \left(\frac{N_i}{N}\right)\cdot n$ $n = 50$

$i = 1, \dots, \#$ of strata.

Probability (Random) Sampling

Type of Sampling	**When/Why** do this type of Sampling	**Selection Mechanism,** or **How** to conduct this type of Sampling
③ Cluster Random Sampling "the opposite" of stratified heterogeneity  	when: 1) convenient, ↓ costs 2) Data has natural groups or "clusters" Ex: Test Performance across CA  1-Stage: Take SRS of schools and sample all students in the school 2-Stage: After selecting schools, take SRS of teachers within selected schools and sample all their students	$\boxed{\text{1-Stage}}$ Use SRS to select clusters and then sample all elements within the cluster $\boxed{\text{2-Stage}}$ 1-Stage & SRS to select subgroups within the cluster; take elements in subgroup.
④ Systematic Random Sampling (Line Sampling)	• Population is dynamic • Pop'n frame is in line format ex: list of addresses, phone #s, etc.	 randomly select starting point & select every k^{th} element

Data Display

- We can estimate the **Frequency Distribution** with:
  - (1) Tables and Graphs
    - Frequency Table
    - Histogram ("Bar Graph")
      - Information from a histogram
    - Cumulative Frequency Polygon
      - percentiles
    - Boxplot
      - rank statistics
      - parts of a boxplot
      - skewness, right vs left skew
  - (2) "Theoretical" Description
    - Normal (Gaussian) Distribution
    - Log-Normal Distribution
      - Why is this distribution useful?
    - Exponential Distribution
      - When is this distribution useful?
  - (3) Numerical (next lecture)
- Sensitivity vs Specificity
  - Trade-off between Sensitivity and Specificity → What happens when you move the line?
- Outliers
  - how to identify outliers
  - what do you do about outliers?

$$\text{Sensitivity} = \frac{\boxed{TP}}{FN + \boxed{TP}} \times 100 \rightarrow$$ "% of people who correctly diagnosed if they have the disease"

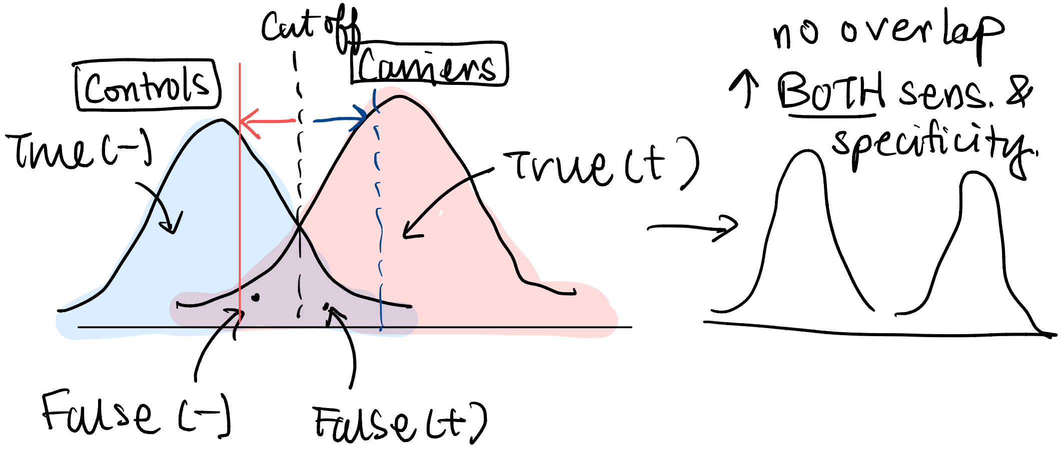
$$= \frac{P(\text{Have Disorder} \cap \text{Test} +)}{P(\text{Have Dis.})}$$

$$= P(+ \mid D)$$

$$\text{Specificity} = \frac{\boxed{TN}}{FP + \boxed{TN}} \rightarrow$$ "% of people correctly diagnosed if they don't have the disease"

$$= P(- \mid \bar{D})$$

	(D) Disorder	($\bar{D}$) No Dis.
(+) Test	True (+) (TP)	False (+) (FP)
(-) Test	False (-) (FN)	True (-) (TN)

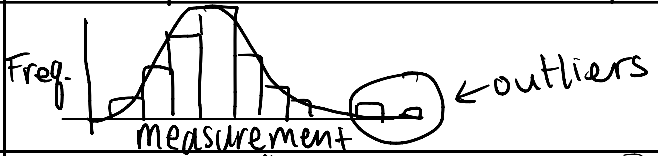
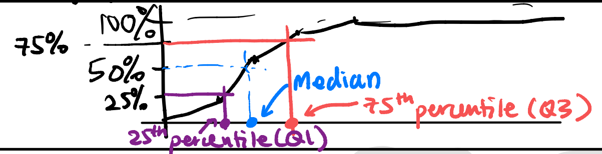
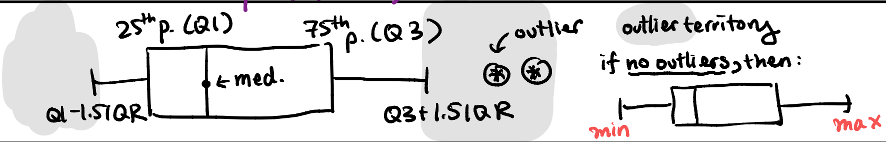
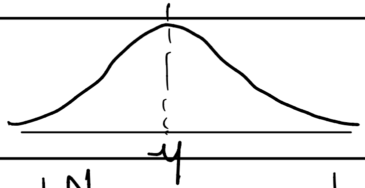
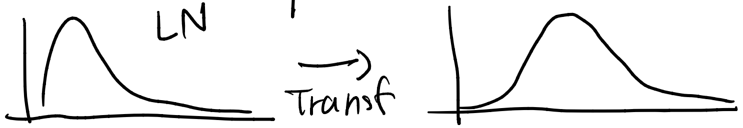
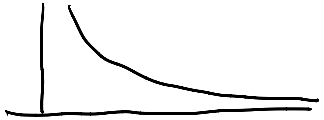


Trade-off between Sensitivity and Specificity ⇓ Inverse Relationship

← FP ↑, TN ↓
Specificity ↓, Sensitivity ↑

→ FN ↑, TP ↓, ↓ FP
Specificity ↑, Sensitivity ↓

Descriptive Statistics

	Tables and Graphs	
	What it looks like	What does it tell us / Utility?
Frequency Tables	counts — Percentage — Cumulative (Last Row = total # of obs. ; Last Row = 100%) Interval | Abs. Freq. | Rel. Freq (%) | Abs. Freq. | Rel. Freq. 1-10 | 5 | 1-2% | 5 | 1-2% 11-20 | 6 | 6-7% | 11 | 7-9% ⋮	Gives you the numbers
Histogram		· Shape of our frequency distr. · Outliers → observations that appear "extreme"
Cumulative Frequency Polygon		Estimating Percentiles
Boxplot		· shape · identify outliers (outside whiskers)
	Theoretical Descriptions	
Normal Distribution	 $N(\mu, \sigma^2)$ — mean, variance	· symmetric data · mean = median
Log-Normal Distribution		· skewed data (right skewed)
Exponential Distribution		survival data

Numerical Descriptions of Data

- Measures of Location
  - (1) Arithmetic Mean (average)
  - (2) Median – how to find the median when it is even vs odd
  - (3) Geometric Mean
  - (4) Mode
  - (5) Midrange

Measure of Location

	Formula/ How to calculate it	When to use it	Statistic → sample	Parameter → Population
Arithmetic Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{n}$	symmetric distr.	$\bar{x}$ = sample mean	μ = pop'n mean
Median	order all values → ½ < 50% percentile ½ > 50% percentile	skewed.		
Geometric Mean	1.) $\log(x_i)$, $i = 1, \ldots, n$ 2.) mean of the logs $\bar{x}_{\log} = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$	3.) $10^{\bar{x}_{\log}}$ exponential distr.		
Mode	most freq. value	skewed data, bimodal ← no mode		
Midrange	$\frac{\text{Max} + \text{Min}}{2}$	Quick & Dirty Statistic		