BISC 305 Statistics for the Biological Sciences
Discussion #1
4^{th} September, 2019

Topics covered: Introduction (Chapter 1, 1-26); Description of Samples and Populations. (Chapter 2, 27-44)

1. In Fall 1973 UC Berkeley's (UCB) admission figures (Table 1) showed that men applying were more likely than women to be admitted. The difference was large enough for it to be arising purely out of chance. UCB faced a lawsuit. Assuming UCB has only two departments whose admission ratios are also shown below (Table 2), discuss if UCB's admission process was indeed biased.

	Female	Male
Applicants	550	550
Admitted	28.2%	41.8%

Table 1: UCB's acceptance percentage

	Female		Male	
	Applicants	Admitted	Applicants	Admitted
Department A	150	50%	400	50%
Department B	400	20%	150	20%

Table 2: UCB's acceptance percentage for each department

2. Submission process to research journals can happen through two channels. In the first channel (C_1), the submitter chooses the potential reviewers who are then invited to review the research article without discussing the identity of the submitter. While submissions through second channel (C_2) are assigned randomly from a pre-available list of reviewers via a computer program. Both the submitter and the reviewers remain anonymous to each other. C_1 and C_2 are examples ___________ study. (Choose one from options below)

   (i) blinded, blinded

   (ii) double-blinded, blinded

   (iii) double-blinded, double-blinded

   (iv) blinded, double-blinded

3. For each of the following cases indicate if the sampling was stratified random, random cluster or simple random:

   (a) Names of 5 students being chosen out of a hat containing names of all 18 students:

   (b) Calculating the proportion of female tigers in a dense forest by dividing into small zones, sampling a zone randomly and calculating the proportion for such samples:

   (c) Literacy rate of a area accounting for age and income strata:

4. USC is trying to decide if it can stop operating grass sprinkler system since there are claims that it might have aggravated California's drought situation since a lot of it anyways goes waste. USC's FMS wants to re-evaluate if watering the grasses has a positive effect on its growth rate. What sort of experiment (case-control/longitudinal) would you advise FMS to conduct? Also discuss confounding variables that you would want to control for.

5. The following graph indicate the number of cars parked in a USC parking lot over different days of a particular week.

   (a) Name the type of graph.

   (b) What percentage of cars were parked in the parking lot over the weekend (Sa/Su)?

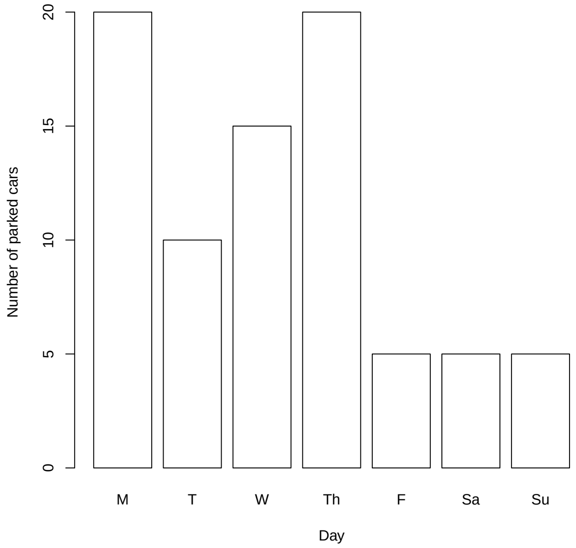


Figure 1: Number of cars parked in a USC parking lot over different days of a particular week

6. For each of the following variables indicate whether they are categorical or numeric. If numeric, indicate whether they are discrete or continuous:

   (a) Age in days:

   (b) Student run clubs at USC:

   (c) Number of student members per student club at USC:

   (d) Manufacture of cars:

   (e) Cost of a food bowl (in dollars) at Verde:

   (f) Cost of a coffee (in cents) at Dulce:

7. For the following two hypothetical samples, calculate the mean, median and mode. Verify your answers using R.

   (a) 10.1, 12.4, 13, 15.1, 20.3, 23.1

   (b) 10.1, 12.4, 13, 15.1, 20.3, 23.1, 100

1. **Observational study:** An experimental study where the independent variable is not under the control of the experimenter; the experimenter draws inferences from a sample to a population. The experimenter can only observe, but cannot manipulate the conditions.

   Case-control studies: Case-control studies involve studying the outcome of experiments involving two existing groups that differ on some supposed causal attribute. Example: Incidences of lung cancer in smokers versus non-smokers; blood glucose levels in diabetics who have been injected with insulin versus those who were injected with saline (placebo).

2. **Sampling bias:** A type of bias introduced by non-uniform sampling of the population such that some members of the population have higher probability of being sampled.

   Non-response bias: A bias caused by persons not responding to some (or all) of the questions in a survey or not returning the survey.

3. **Blind experiment:** An experiment where the treatment assignment is kept secret from the experimental subject. This us used to control for the fact that expectations can influence the results of the experiment. Example: I am getting a drug, so I should feel better.

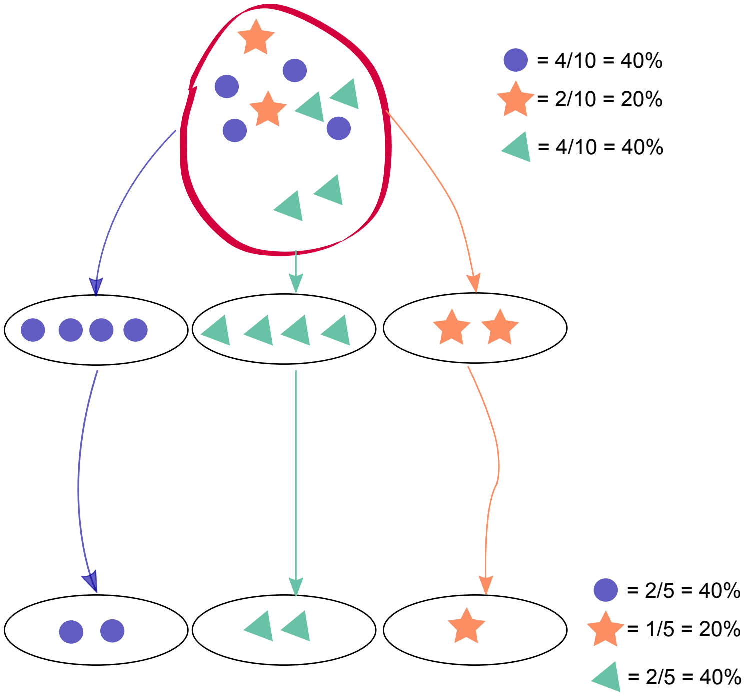
   Double-blind experiment: An experiment in which both subjects and the persons making the evaluation are *blinded* (Neither the doctor nor the patient know what treatment[drug/placebo] they are receiving).

4. **Simple Random Sample:** A simple random sample of a population has the following two properties:

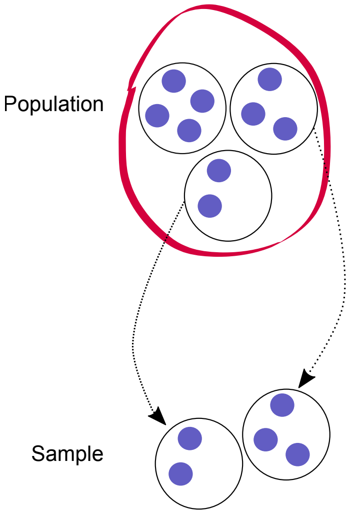
   1. every member of the population has the same chance of being *sampled*
   2. the members of the sample are chosen independently (the chance of a member being chosen does not depend on which other members were chosen)

   Non-Simple Random Sampling:

   1. Stratified Random Sampling: A sampling method for population that can be partitioned into subpopulation. The population is first divided into homogeneous groups and sampling is done over these groups proportional to their size. Example: Poll survey for presidential elections by stratifying age, sex, income etc.
   2. Random cluster sampling: A sampling method used when population contains mutually homogeneous yet internally heterogeneous groupings. Example: counting butterflies in a forest by defining small homogeneous geographical locations



Stratified Random Sampling



Cluster random sampling

5. **Variable:** A characteristic of a person or a thing that can be assigned a number or a category:

   Categorical: A type of variable that records which of the several categories a thing or a person is in. They are unordered. Examples:

   - Blood group: A, B, AB, O
   - Football teams: USC, Stanford, UCLA ...
   - Colors: Red, Blue, Purple ...

   Numeric: A numeric variable records the “amount” of something:

   - **Continuous:** A numeric variable that is measured on a continuous scale. Example: blood glucose levels, pH of a solution, rainfall (in centimeters), age (in years)
   - **Discrete:** A discrete variable is a numeric variable for which it is possible to list the possible values. In principle, you should be able to count them explicitly. Example: age (in days), number of cars in a parking lot.

6. **Frequency distribution**: A display of the frequency (the number of occurrences) of each value in the dataset. The information is generally presented with a graph:

   - Bar graph: A graph for categorical data showing the number of observations in each category. (Mnemonic: **cat** goes to the **bar**)
   - Dotplot: A graph used to display the distribution of a numeric variable when the sample size is small.
   - Histogram: A histogram displays the frequency distribution of a continuous numeric variable, similar to a dot plot. The area of each bar is proportional to the corresponding frequency.

   Shape features of a histogram Consider the following histogram of distribution of Creatine phosphokinase (CK) drawn from 36 male volunteers. The peak marks the **mode**, i.e. the most frequent CK values while on either side of this mode the frequency declines and defines the tails of distribution.

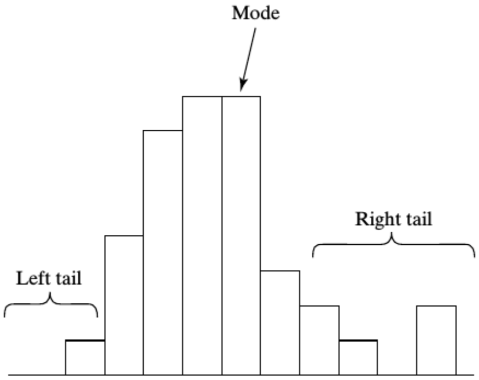


Figure 2: Shap features of a histogram

The area of each bar in histogram can be used to measure the corresponding relative frequency. For example in the following histogram, the blue shaded region comprises of $\frac{7+8}{1+4+7+8+8+3+2+1+0+2} = 15/36 = 42\%$.

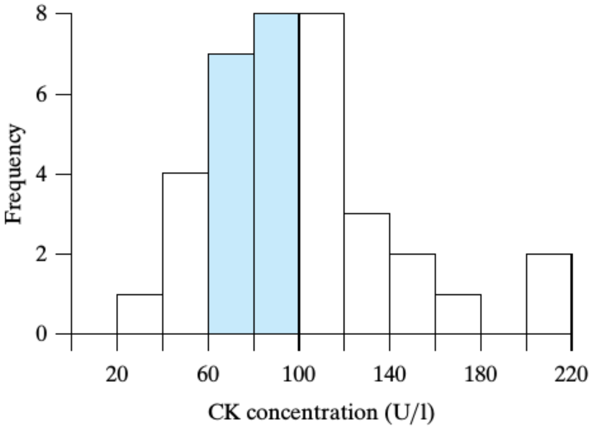


Figure 3: Shaded area is proportional to the relative frequency of CK values in the range of 60-100 U/l.

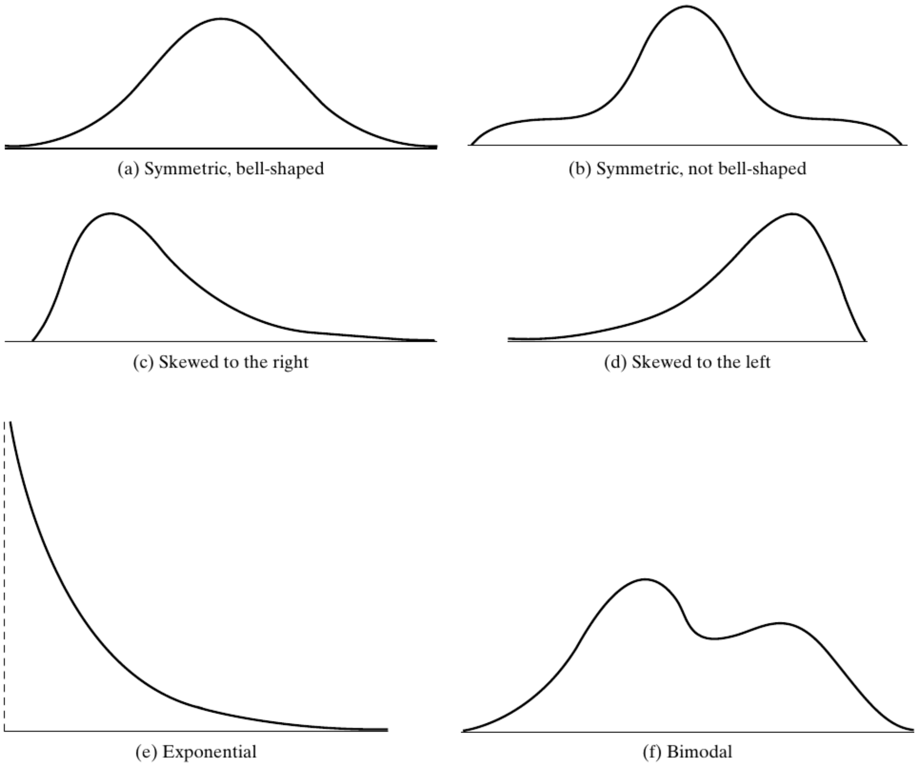


Figure 4: Possible shapes of a distribution

7. **Statistics**: A numerical ,measure calculated from sample data.
   Descriptive statistics describe a set of data.
   Median: The value that most nearly lies in the middle of the sample. If the number of observations are odd, it is the middle most value. If the number of observations is even, it is the mean of the middle most two values.
   Mean: sum of observation divided by number of observations

   Example: Consider the following lamb weight gain data from Example 2.3.1: 11, 13, 19, 2, 10, 1. The mean of the sample acts as the point of balance for the data while the median as fulcrum can lead to unbalanced *seesaw*.

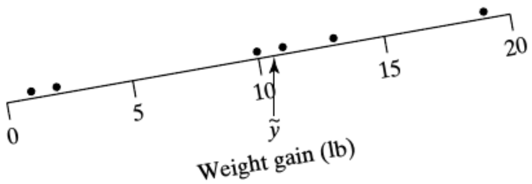


Figure 2.3.2 Plot of the lamb weight-gain data with the sample median as the fulcrum of a balance

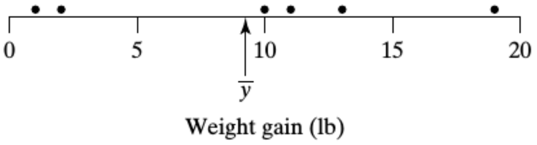


Figure 2.3.3 Plot of the lamb weight-gain data with the sample mean as the fulcrum of a balance