Part 1: Descriptive Statistics

STAT 324

Midterm: Thursday, February 20th

Contents

Terminology…

Formulas…

R code…

Basic Terminology

Population: *the entire collection of well-defined objects*

Census: *information from every unit of the population*

Parameters: *numeric summaries of population characteristics*

Sample: *a subset of a population, containing the objects or outcomes that are actually observed*

Simple random sample (SRS): *each group of same size* n *is equally likely to be drawn as the sample*

Statistic: *numeric summaries of sample characteristics*

Randomization: *the process of intentional arbitrary selection to increase the validity of interference*

Experimental study: *the researcher actively manipulates certain variables*

Bias: *the degree to which a procedure systematically overestimates or underestimates population value*

Types of Data

Quantitative data: *values with unit of measure*

Discrete data: *numeric data where the scale is made up numbers with games (i.e. counting numbers)*

Continuous data: *numeric data where the values are taken off of any interval (i.e. time, length)*

Categorical data: *values vary in kind; different levels*

Nominal: *no natural order (i.e. gender, color)*

Ordinal: *an order exists but no numerical measurements (i.e. agree, disagree)*

Methods of Visualizing Data

Dot plot: *chart with a number line and a point for each datum above the line at its value*

Histogram: *used to display the frequency, percentage, or density of measurements falling to a range of values with rectangles with heights equal to frequency, percentage, or density*

Box plot: *displays 5 number summaries and outlying values in a box with whiskers*

First quartile: *the median of the lower half of a data set; 25th percentile*

Second quartile: *the median of the set; 50th percentile*

Third quartile: *median of the upper half of a data set; 75th percentile*

Interpreting Visuals

Symmetric data: *upper and lower halves of the data have very similar to identical shapes*

Right skewed: *data the graphs tail is extended to the right*

Left skewed data: *the graphs tail is extended to the left*

Uniform: *histogram where every interval has proportional number of observations*

Unimodal: hi*stogram with one major peak*

Bimodal: *histogram with two major peaks*

Interquartile range (IQR): *difference between the first and third quartiles; range of 50% of data*

$$IQR = Q3 - Q1$$

Range: *difference between the maximum and minimum values*

Formulas

Population

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

$$\mu = \frac{1}{n}\sum x_i$$

Sample

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Compliment of an event

$$P(A) = 1 - P(\neg A)$$

Basic probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Descrete random variables

$$\mu_X = E(X) = \sum x_i \cdot P(X = x_i)$$

$$\sigma_X^2 = Var(X) = \sum (x_i - \mu_x)^2 \cdot P(X = x_i)$$

Binomial random variable

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} = \frac{n!}{x!\,(n - x)!} \pi^x (1 - \pi)^{n-x}$$

$$\mu_X = \pi n$$

$$\sigma_X^2 = n\pi(1 - \pi)$$

Linear transformation of random variables

Let $Y = X + c$

$$E(Y) = E(X + c) = E(X) + c = \mu_X + c$$

$$Var(Y) = Var(X + c) = Var(X) = \sigma_X^2 \Rightarrow Sd(Y) = Sd(X) = \sigma_{\mathrm{X}}$$

Let $P = aX$

$$E(P) = E(aX) = a \cdot E(X) = a \cdot \mu_X$$

$$Var(P) = Var(aX) = a^2 \cdot Var(X) = a^2 \cdot \sigma_X^2 \Rightarrow Sd(P) = |a|Sd(X) = |a|\sigma_X$$

Let $L = aX + c$

$$E(L) = E(aX + c) = a \cdot E(X) + c = a \cdot \mu + c$$

$$Var(L) = a^2 \cdot Var(X) \Rightarrow Sd(L) = |a|\sigma_X$$

z-score

$$z = \frac{x - \mu}{\sigma}$$

Normal distribution

$$P = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$X \sim N(\mu_X, \sigma_X^2)$$

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

$$z^2 = -2\ln\left[\sigma\sqrt{2\pi}P\right]$$

Sampling distribution of the sample sum

$$Sum = X_1 + X_2 + \cdots + X_n$$

$$\mu_S = n \cdot \mu_X$$

$$\sigma_S^2 = n \cdot \sigma_X^2 \Longrightarrow \sigma_X = \sqrt{n} \cdot \sigma_X$$

Sampling distribution of the sample mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$\mu_{\bar{X}} = \mu_X$$

$$\sigma_{\bar{X}}^2 = \frac{1}{n}\sigma_X^2 \Longrightarrow \mathrm{\sigma}_{\bar{\mathrm{X}}} = \frac{1}{\sqrt{n}}\sigma_X$$

R Code

range(x)	
IQR(x)	
sd(x)	Sample standard deviation
par(mfrow=c(2,1))	Allows two charts, arranged vertically to be displayed
hist(x,breaks=””,xlim=range(),xlab=””,ylab=””,plot=TRUE)	
boxplot(x,xlab,ylab)	
mean(x)	
median(x)	
pnorm(x, mean, sd, lower.tail)	Single parameter → computes via z-score Gives distribution function
qnorm(x, mean, sd, lower.tail)	Single parameter → computes via z-score Gives quantile function
dbinom(x, size, prob)	Give density of binomial distribution