Review of Descriptive Statistics and Conceptualizations of Variance

EPSY 905: Multivariate Analysis

Online Lecture #1

Learning Objectives

- Univariate descriptive statistics
  - Central tendency: Mean, median, mode
  - Variation/spread: Standard deviation, variance, range
- Bivariate descriptive statistics
  - Correlation
  - Covariance
- Types of variable distributions:
  - Marginal
  - Joint
  - Conditional
- Bias in estimators

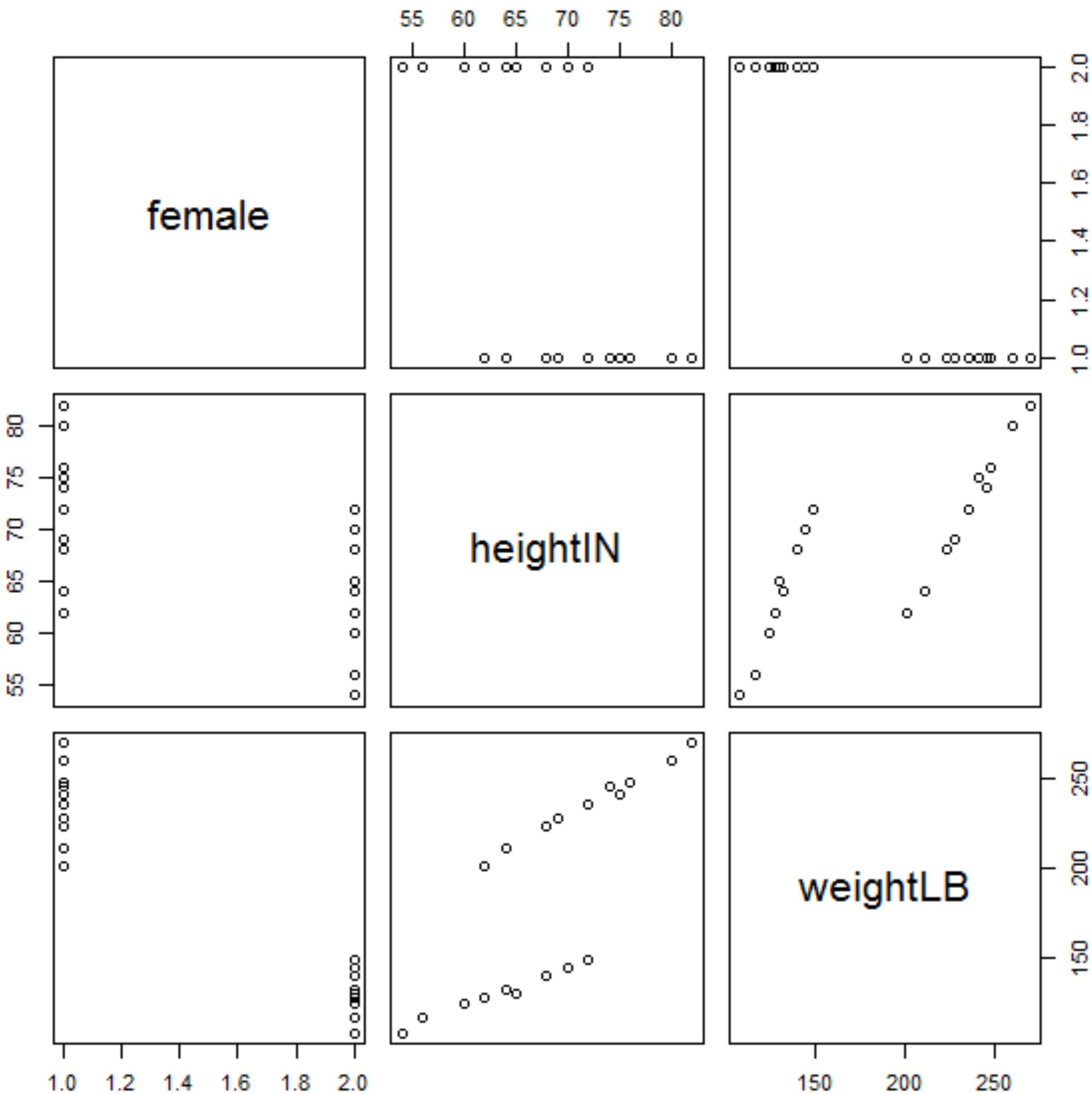
Data for Today's Lecture

- To help demonstrate the concepts of today's lecture, we will be using a data set with three variables
  - Female (Gender): Male (=0) or Female (=1)
  - Height in inches
  - Weight in pounds

- The end point of our second lecture will be to build a **linear model** that predicts a person's weight
  - **Linear model**: a statistical model for an outcome that uses a linear combination (a weighted sum) of one or more predictor variables to produce an estimate of an observation's predicted value

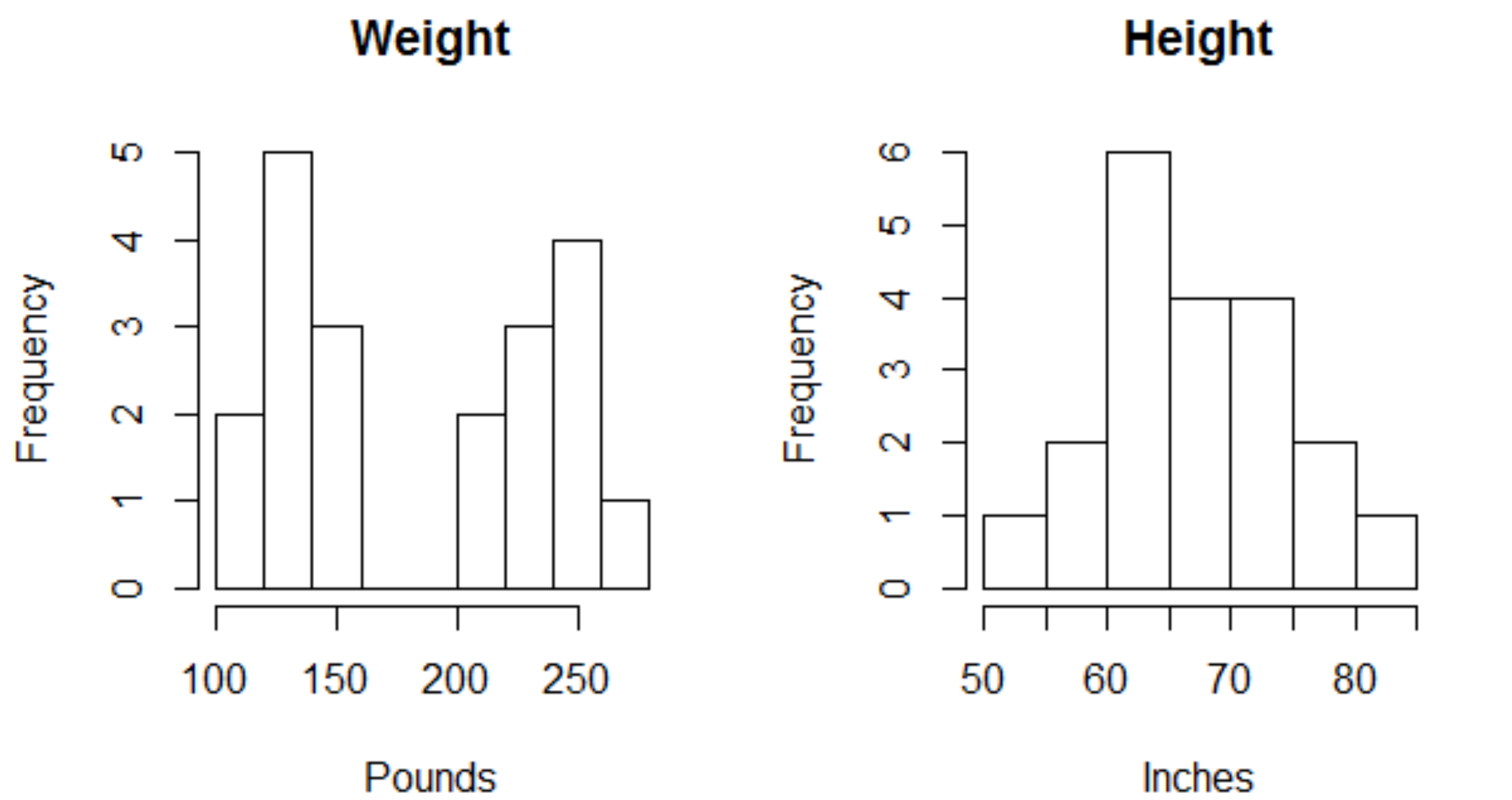
- **What you will learn is that models underlie all statistics**

Visualizing the Data



Histograms of Height and Weight

- The weight variable seems to be bimodal – should that bother you? (hint: it shouldn’t…yet)



Descriptive Statistics

- We can summarize each variable **marginally** through a set of descriptive statistics
  - **Marginal:** one variable by itself

- **Common marginal descriptive statistics:**
  - Central tendency: *Mean*, Median, Mode
  - Variability: *Standard deviation (variance)*, range

- We can also summarize the **joint** (bivariate) **distribution** of two variables through a set of descriptive statistics:
  - **Joint distribution:** more than one variable simultaneously

- **Common bivariate descriptive statistics:**
  - Correlation and covariance

THE UNIVERSITY OF KANSAS

Descriptive Statistics for Height/Weight Data

Variable	Mean	SD	Variance
Height	67.9	7.44	55.358
Weight	183.4	56.383	3,179.095
Female	0.5	0.513	0.263

Diagonal: Variance

Above Diagonal: Covariance

Correlation /Covariance	Height	Weight	Female
Height	55.358	334.832	-2.263
Weight	.798	3,179.095	-27.632
Female	-.593	-.955	.263

Below Diagonal: Correlation

Re-examining the Concept of Variance

- Variability is a central concept in advanced statistics
  - In multivariate statistics, covariance is also central

- Two formulas for the variance (about the same when N is large):

$$S^2_{Y_1} = \frac{1}{N-1}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)^2$$

Unbiased or "sample"

$$S^2_{Y_1} = \frac{1}{N}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)^2$$

Biased/ML or "population"

Here: p = person; 1 = variable number one

Interpretation of Variance

- The variance describes the spread of a variable in squared units (which come from the $\left(Y_{1p} - \bar{Y}_1\right)^2$ term in the equation)

- Variance: **the average *squared* distance of an observation from the mean**
  - Variance of Height: 55.358 inches squared
  - Variance of Weight: 3,179.095 pounds squared
  - Variance of Female – not applicable in the same way!

- Because squared units are difficult to work with, we typically use the standard deviation – which is reported in units

- Standard deviation: **the average distance of an observation from the mean**
  - SD of Height: 7.44 inches
  - SD of Weight: 56.383 pounds

Variance/SD as a More General Statistical Concept

- Variance (and the standard deviation) is a concept that is applied across statistics – not just for data
  - Statistical parameters have variance
    - e.g. The sample mean $\bar{Y}_1$ has a "standard error" (SE) of $S_{\bar{Y}} = \frac{s_Y}{\sqrt{N}}$
- The standard error is another name for standard deviation
  - So "standard error of the mean" is equivalent to "standard deviation of the mean"
  - Usually "error" refers to parameters; "deviation" refers to data
  - Variance of the mean would be $S_{\bar{Y}}^2 = \frac{s_Y^2}{N}$
- More generally, variance = error
  - You can think about the SE of the mean as telling you how far off the mean is for describing the data

Correlation of Variables

- Moving from marginal summaries of each variable to joint (bivariate) summaries, the Pearson correlation is often used to describe the association between a pair of variables:

$$r_{Y_1,Y_2} = \frac{\frac{1}{N-1}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)\left(Y_{2p} - \bar{Y}_2\right)}{S_{Y_1} S_{Y_2}}$$

- The correlation is **unitless** as it ranges from -1 to 1 for continuous variables, regardless of their variances
  - Pearson correlation of binary/categorical variables with continuous variables is called a point-biserial (same formula)
  - Pearson correlation of binary/categorical variables with other binary/categorical variables has bounds within -1 and 1

More on the Correlation Coefficient

- The Pearson correlation is a **biased** estimator
  - **Biased estimator:** the expected value differs from the true value for a statistic
    - Other biased estimators: Variance/SD when $\frac{1}{N}$ is used
- The unbiased correlation estimate would be:

$$r^{U}_{Y_1,Y_2} = r_{Y_1,Y_2}\left[1 + \frac{\left(1 - r^{2}_{Y_1,Y_2}\right)}{2N}\right]$$

  - As N gets large bias goes away; Bias is largest when $r_{Y_1,Y_2} = 0$
  - Pearson is an underestimate of true correlation
- If it is biased, then why does everyone use it anyway?
  - Answer: forthcoming when we talk about (ML) estimation

Covariance of Variables: Association with Units

- The numerator of the correlation coefficient is the covariance of a pair of variables:

$$S_{Y_1,Y_2} = \frac{1}{N-1}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)\left(Y_{2p} - \bar{Y}_2\right)$$

Unbiased or "sample"

$$S_{Y_1,Y_2} = \frac{1}{N}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)\left(Y_{2p} - \bar{Y}_2\right)$$

Biased/ML or "population"

- The covariance uses the units of the original variables (but now they are multiples):
  - Covariance of height and weight: 334.832 inch-pounds
- The covariance of a variable with itself is the variance
- The covariance is often used in multivariate analyses because it ties directly into multivariate distributions
  - But…covariance and correlation are easy to switch between

THE UNIVERSITY OF KANSAS

Going from Covariance to Correlation

- If you have the covariance matrix (variances and covariances):

$$r_{Y_1,Y_2} = \frac{S_{Y_1,Y_2}}{S_{Y_1} S_{Y_2}}$$

- If you have the correlation matrix and the standard deviations:

$$S_{Y_1,Y_2} = r_{Y_1,Y_2} S_{Y_1} S_{Y_2}$$