Chapter 2 Class Notes – Sample & Population Descriptions

Classifying variables

Random Variables (RVs) are

--- quantitative
- |--- discrete
- |
- |--- continuous

--- qualitative
- |--- nominal
- |
- |--- ordinal

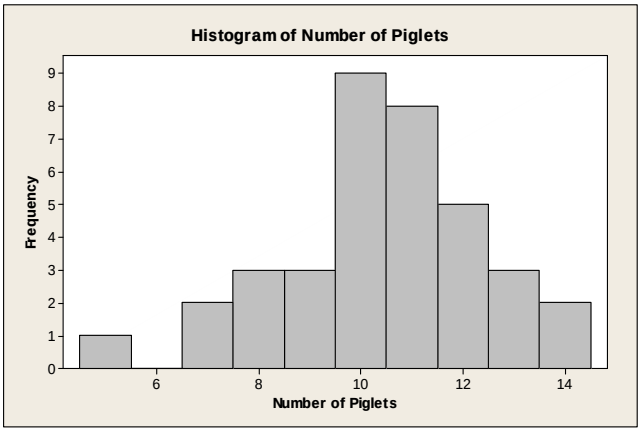
Notation and Definitions:

- a Sample is a collection of objects
- a Sample is of size 'n'; Population size is 'N'
- we measure a response variable (denoted Y)

Some graphical displays – Frequency Distributions (§2.2)

- Bar charts for qualitative variables (see pp.28-29)
- Frequency histograms for quantitative variables (see below)

Sow Litter Size (discrete; p. 31)



A frequency histogram for a continuous variable is the Serum CK example on p.33; note the *right skew* in this distribution. The sample mean for these data is 98.3 and the sample median is 94.5; indeed, the mean is to the right of the median. This will always be the case. And the reverse is true when the distribution is *skewed to the left*.

2.3 Measuring the Center (Mean, Median, Mode)

Lamb weight data (p.40):

$$y_1 = 11,\ y_2 = 13,\ y_3 = 19,\ y_4 = 2,\ y_5 = 10,\ y_6 = 1$$

The sample average or sample mean is

$$\bar{y} = \frac{y_1 + y_2 + \cdots y_n}{n} = \frac{\sum_{k=1}^{n} y_k}{n}$$

The sample median is the 50th percentile or center value, and the sample mode is the most common value. Sample measures such as the mean, median and mode are called sample statistics.

For the lamb weight data, the sample mean is

$$y = \frac{11 + 13 + 19 + 2 + 10 + 1}{6} = \frac{56}{6} = 9.333\ pounds$$

To find the median, we order the values:

1 2 10 11 13 19

and we average the center two values (since n = 6 is even); thus, the median is 10.5 pounds.

We will use the convention given in the main text on pp.45-46 to find the median as well as the first quartile (Q_1) and the third quartile (Q_3). Then, the interquartile range is IQR = Q_3 - Q_1. Also, Boxplots (§2.4 on pp.45-51)) graphically show the minimum and maximum values as well as Q_1, the median (Q_2), and Q_3. See text p.48 for an example. On pp. 48-49, the issue of Outliers is raised: note that this is a somewhat controversial subject.

Read §2.5 on relationships between variables (categorical-categorical, numeric-categorical, and numeric-numeric).

2.6 Measuring Variability or Dispersion

One such measure is the range – another is the IQR (which is more robust). But more common/useful is the sample standard deviation ("SD" or "s"), which is the positive square root of the sample variance (s^2). The sample variance is:

$$s^2 = \frac{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1} = \frac{SS}{n-1} = \frac{\sum_{k=1}^{n}(y_k - \bar{y})^2}{n-1}$$

We call the component $(y_k - \bar{y})$ the *deviation* associated with the observation y_k. To illustrate, for the lamb data ($\bar{y} = 9.333$):

11	13	19	2	10	1

the deviations are

1.667	3.667	9.667	-7.333	0.667	-8.333

(these always must sum to zero), so that the sample variance is

$$s^2 = \frac{(1.667)^2 + (3.667)^2 + (9.667)^2 + (-7.333)^2 + (0.667)^2 + (-8.333)^2}{6-1}$$
$$= \frac{233.33}{5} = 46.667 = 6.8313^2$$

and the sample SD is s = 6.8313 pounds. Further, the coefficient of variation (useful in ecology, agriculture, and medicine) is

$$CV = \frac{s}{\bar{y}} \times 100\% = \frac{6.8313}{9.333} \times 100\% = 73.2\%$$

In practice, to find the sample variance, we usually use the short-cut formula:

$$s^2 = \frac{\left(\sum_{k=1}^{n} y_k^2\right) - \frac{\left(\sum_{k=1}^{n} y_k\right)^2}{n}}{n-1}$$

To illustrate, for the Serum CK values on p.49 bottom, n = 36, Σy_k = 3,538, $\bar{y}$ = 98.2778, and Σy_k^2 = 404,784, so that

$$s^2 = \frac{404,784 - \frac{3538^2}{36}}{35} = 40.383^2$$

Old Quiz Exercise: Ten women were asked how many hours per week they exercise. Their answers were as follows:

5 13 3 6 14 3 1 3 8 4

Calculate the coefficient of variation for these data. Hint: the sum of these data is 60 and the sum of their squares is 534.

[Answer is (4.397/6)*100% = 73.3%]

An important application of the above is the Empirical Rule (p.46), which states that for a 'nicely shaped' distribution

- **about 68% of the sample values will lie in the interval from $(\bar{y} - s)$ to $(\bar{y} + s)$ – *within one SD of the mean***
- **about 95% of the sample values will lie in the interval from $(\bar{y} - 2s)$ to $(\bar{y} + 2s)$ – *within two SD's of the mean***
- **> 99% of the sample values will lie in the interval from $(\bar{y} - 3s)$ to $(\bar{y} + 3s)$ – *within three SD's of the mean***

To illustrate, for the Serum CK values $\bar{y}$ = 98.278 and s = 40.383, so the Empirical Rule says that

- **approximately 68% of the sample values will lie in the interval (57.90 , 138.66);**
- **approximately 95% of the sample values will lie in the interval (17.51 , 179.04); and**
- **over 99% of the sample values will lie in the interval (-22.87 , 219.43)**

By returning to the *actual data* on p.68 and counting, we see that the actual percentages in these respective intervals are

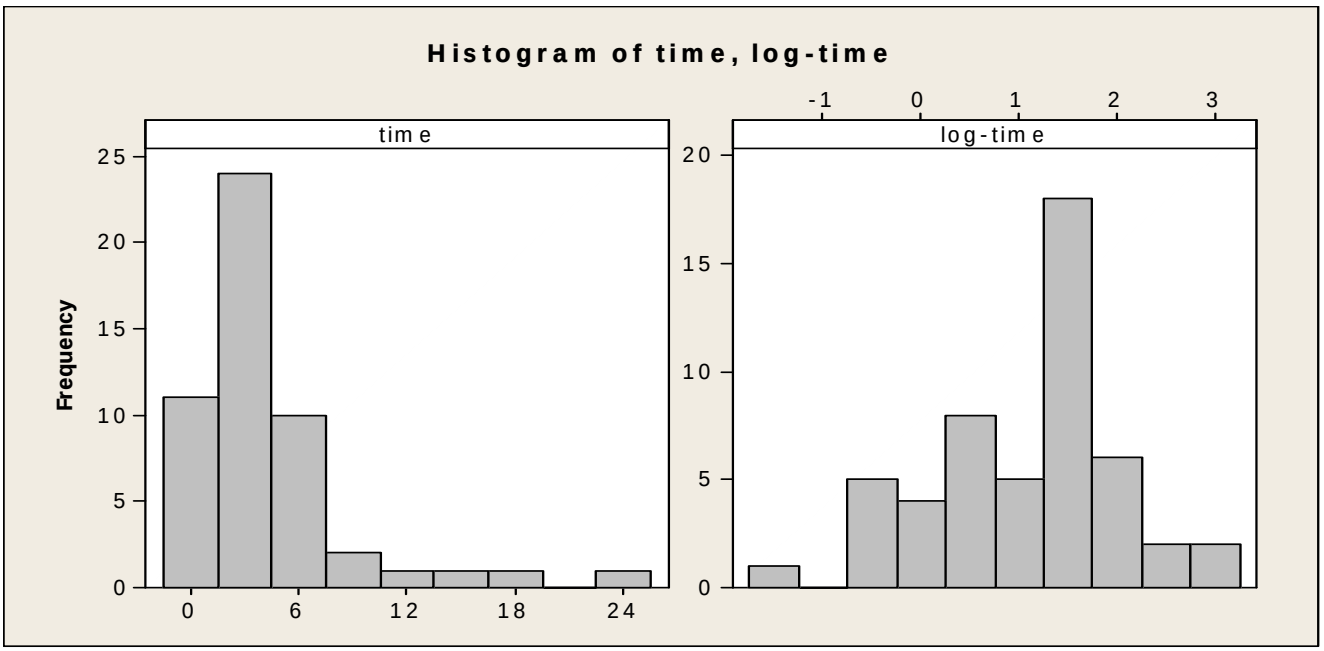
- **26 / 36 = 72.2%**
- **34 / 36 = 94.4%**
- **36 / 36 = 100%**

Note that these values are indeed close to those predicted by the Empirical Rule.

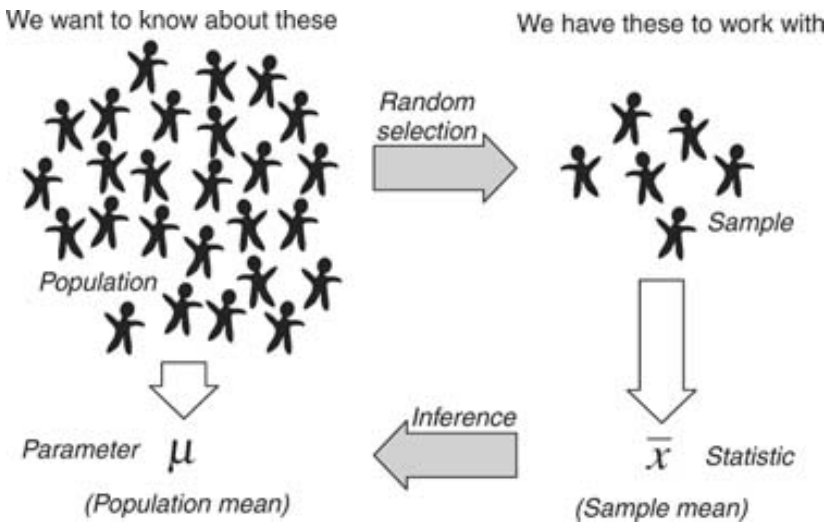
Read on p.65 how you can estimate the mean and SD from a histogram; note Example 2.6.10 on p.66.

§2.7: Sometimes, obtaining a *transformation* of a RV (Y) is helpful – for example to achieve a more symmetric distribution of the transformed variable (Y'). A linear transformation is used to convert degrees Celsius to degrees Fahrenheit; taking a square root or a logarithm involves a nonlinear transformation.

The Cricket Singing Time (Y) data of p.72 can be transformed using the natural logarithm transformation [$Y' = \log(Y)$]; see the following plots.



§2.8: As mentioned, the following diagram is helpful to visualize an important application of statistical science called Statistical Inference. A representative *Sample* is taken from the *Population* (so as to remove any *Bias*), and the sample *Statistic* (such as sample mean) is used to estimate the population *Parameter*.



For example, a selected group of n = 25 Loyola undergraduates had an average age of $\bar{y}$ = 19.4 and $\hat{p}$ = 64.0% were Female. Making statements about *all* Loyola undergraduates falls under the heading of 'statistical inference'. It's important to consider if the sample is representative and consider whether the sampling methodology may result in any bias. Sometimes defining the population can be challenging. For example, if the sample was chosen "at random" by knocking on doors in the student dormitories, these issues may be very important.

Note that a sample characteristic is called a statistic and a population characteristic is called a parameter. Some examples include the following (see p.79):

Characteristic	Sample Statistic	Population Parameter
Mean	$\bar{y}$	μ
SD (standard deviation)	s	σ
Proportion	$\hat{p}$	p