It is easy to lie with statistics. It is hard to tell the truth without statistics.

– Andrejs Dunkels

(quoted in Maindonald & Braun 2003. Data Analysis and Graphics Using R.)

1 Definitions

Data = observations, measurements (data are plural; many data, one datum)

Statistics = analysis and interpretation of data

Biostatistics = statistics applied to biological problems

Descriptive statistics = ways of organizing & summarizing data

Inferential statistics = conclusions about whole (entire population) from data on parts

2 Types of Data

2.1 Ratio • constant interval size • true zero

2.2 Interval • constant interval size • not true zero

2.3 Ordinal • order or rank; not know values

2.4 Nominal • categories; trait or attribute

2.5 Continuous vs. Discrete

2.6 Accuracy vs. Precision

2.7 Significant figures

3 Populations & Samples

3.1 Populations – entire universe of objects; group of possible measurements

3.2 Samples – subset of population

3.3 Types of Sampling

– this course: random sampling:

– other kinds of sampling

(based on prior knowledge of population or practical considerations)

• stratified • clustered • systematic • multistage • double sampling • bootstrap

3.4 Parameters & statistics

parameter = characteristic of population (e.g., mean, variance, magnitude of effect)

statistic = estimate of parameter

4 Frequency Distributions

5 Descriptive Statistics

5.1 Measures of Location:

5.1.1 Arithmetic mean – most widely used (≅ average)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \sum_{1}^{4} X_i = X_1 + X_2 + X_3 + X_4$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

5.1.2 Median = middle measurement in ordered set of data (central data point)

$M = X_{(n+1)/2}$

if N even, average of 2

5.1.3 Quartiles: Inner quartile range (IQR) = range containing 50% of data, with 25% on either side of median

5.1.4 Mode = most frequent value

5.1.5 Geometric mean (GM)

$$\overline{X}_G = \sqrt[n]{X_1 X_2 X_3 \cdots X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

5.2 Measures of Variability:

5.2.1 Variance $s^2 = \sum \frac{(X_i - \overline{X})^2}{n-1}$

5.2.2 Standard Deviation $s = \sqrt{s^2}$

5.2.3 Standard Error $\mathrm{SE} = \sqrt{\frac{s^2}{n}}$

5.2.4 Confidence Interval $\mathrm{CI} = \overline{X} \pm t_{\alpha,n-1} \times \mathrm{SE}$