Stat 257.01 Fall 2004
Assignment #1 Solutions

**(2.1)** There are two definitions of "statistics" that are possible here. A *statistic* is simply a number computed from data; for example, the sample mean or sample median. The subject of *statistics* is that branch of science that deals with the analysis of data. There is much that can be said about the role of statistics in modern society: it is hard to imagine life without statistics. Politicians and public officials use statistics to make policy decisions in fields as diverse as: defense, health care, and transportation, etc. Companies use statistics to market their products, and television producers are obsessed with the Nielsen ratings as a measure of a TV show's success. It is hard to listen to a sporting event without the announcer commenting on the "statistics" of a particular team or player: penalty minutes, first downs, runs batted in, free throw percentage ....

**(2.3)** When given a data set, it is wise to begin by visually examining the data. Become familiar with the data: What measurement is being taken? and What is being observed? Is the data quantitative (numerical) or categorical (labels)? Are there any outliers? Often a graphical display reveals many patterns. There might be clusters or outliers. The data could be symmetric or skewed; unimodal or multi-modal. If the data are paired, then a scatterplot may reveal a relationship: linear or otherwise. Also, it is straightforward to compute the summary statistics (as appropriate): mean, median, mode, standard deviation, interquartile range, maximum, minimum, range, correlation and/or covariance. Most of this is easily accomplished with a computer and statistical software such as SAS.

**(2.4)** A statistic is a number computed from data, while a parameter is a number associated with a population. Most often, a parameter is unknown, and a statistic is used to estimate a particular parameter.

**(2.5)** By definition,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Squaring, we see that $(y_i - \overline{y})^2 = y_i^2 - 2y_i\overline{y} + \overline{y}^2$, and distributing the sum gives

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - 2\overline{y}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\overline{y}^2.\right]$$

Since $\overline{y}$ is constant, and since $\sum y_i = n\overline{y}$, we see that

$$\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - 2\overline{y}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\overline{y}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - 2\overline{y}\cdot n\overline{y} + n\overline{y}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right]$$

as required.

**(2.13)** Very briefly:

**(a)** The average is $M/R$, and each average is exact (having factored the billions and thousands into the computation). Ordered by state we have:

| State | average |
|---|---|
| Alaska | 8492.6 |
| New York | 11053.1 |
| Rhode Island | 11146.5 |
| Florida | 11422.8 |
| California | 11593.9 |

**(b)** The average across the 5 states is $491\,000\,000\,000/43\,103\,000 = 11\,391.3$.

**(c)** Yes, the 15 000 mile figure seems reasonable because the average in **(b)** is close (enough) to 10 000.

**(2.14)** If a statistic is viewed as a random variable (as opposed to deterministic) sampled from a population, then the distribution of that statistic is known as its sampling distribution.

**(2.17)** Very briefly:

**(a)** Compare the column totals for *exercise vigorously* and *cigarette smoking*. Since 10907 and 11092 are nearly the same, and 11017 and 11014 are nearly the same, the randomization scheme did a good job in controlling these variables.

**(b)** No, since $5431/11017 \approx 0.49$ and $5488/11014 \approx 0.50$ are nearly the same.

**(c)** No, since $2997/10907 \approx 0.27$ and $3060/10921 \approx 0.28$ are nearly the same.

**(2.21)** An estimator is a statistic that is computed in an attempt to estimate a particular parameter.

**(2.22)** In order to evaluate the goodness of an estimator $\hat{\theta}$, one should compute both $E(\hat{\theta})$ and $\mathrm{Var}(\hat{\theta})$.

**(2.23)** Suppose that $\hat{\theta}$ is used an estimator of $\theta$. Two desirable properties of $\hat{\theta}$ are that $E(\hat{\theta}) = \theta$ and that $\mathrm{Var}(\hat{\theta})$ is as small as possible.

**(2.24)** We call $\hat{\theta}$ an unbiased estimator of $\theta$ if $E(\hat{\theta}) = \theta$.

**(2.25)** The error of estimation made when using $\hat{\theta}$ to estimate $\theta$ is given by the (random) quantity $|\hat{\theta} - \theta|$.

**(2.26)** Since Chebychev's theorem tells us that for any distribution, 75% of the observations must fall within 2 standard deviations of the mean, it seems reasonable to use $2\sigma_{\hat{\theta}}$ as a bound on the error of estimation. In the particular case of the normal distribution, we see that just over 95% of the data lie within 2 standard deviations of the mean. (Actually, 95% lies within 1.96 SDs of the mean.) This is how the magic "95% confidence interval" is often justified.

**(2.27)** By definition,

$$\mathrm{Var}(y) = \sum_x (x-\mu)^2 P(y=x)$$

where $\mu = E(y)$ and $P(y=x) = p(x)$ is the *probability density function* of the random variable $y$. In the case given, the possible values of the random variable $y$ are $u_1, u_2, \ldots, u_N$, and each occurs with probability $P(y = u_i) = 1/N$. Hence, we conclude that

$$\mathrm{Var}(y) = \sum_x (x-\mu)^2 P(y=x) = \sum_{i=1}^{N} (u_i-\mu)^2 P(y=u_i) = \frac{1}{N}\sum_{i=1}^{N}(u_i-\mu)^2.$$