Biostat 514
Emerson, Fall 09
Homework #1 Key
October 30, 2009

In the following presentation of methods I suggest that the sample mean of a binary variable can be used to compute the proportion. If you use the commands given below, it will also compute the sample standard deviation of the binary variables. Why is this boring?

**Answer:**

Given binary random variables $X_1, X_2, X_3, \ldots, X_n$ with $X_i \sim \mathcal{B}(1, p)$, the sample mean is given by

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

and the sample variance is given by

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\ &= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right] \end{aligned}$$

.

Now because $X_i \in \{0, 1\}$ for all $i$, we also know that $X_i^2 = X_i$ for all $i$. Hence, we have

$$s^2 = \frac{n}{n-1}\left[\overline{X} - \overline{X}^2\right] = \frac{n}{n-1}\overline{X}(1 - \overline{X}).$$

Thus, the sample variance and the sample standard deviation are merely a transformation of the sample mean. No new information is provided by the sample standard deviation.

It is true that the Bernoulli distribution has a mean-variance relationship such that $Var(X) = E(X)(1-E(X))$. But it is not the mean-variance relationship of the distribution that makes the sample standard deviation completely boring. If we consider the Poisson distribution, we have that $Var(X) = E(X)$, but there is no simple relationship between the sample mean and the sample variance in the setting of the Poisson distribution. In fact, when doing Poisson based analyses, we often consider the difference between the model based estimates of the variance $(\overline{X})$ and the nonparametric estimator of the variance $(s^2)$ to detect violations of our assumptions. Such "model diagnostics" are not possible with Bernoulli data.

(Of course, the sample mean and sample variance are not totally independent statistics unless the data have a normal distribution.)