results in an ordering consistent with the inequality concerning the measures of central tendency:

$$x_{(1)} = 1 \qquad h = 1.61 \qquad g = 1.88 \qquad \bar{x} = 2.2 \qquad q = 2.53 \qquad x_{(10)} = 5.$$

1.4 Estimating dispersion

Central tendency is just one metric associated with a population probability distribution. Measuring how far a random variable might stray from a central value is a second key aspect of a population probability distribution. This section introduces statistics that are used by statisticians for measuring dispersion. We begin with the most important measure of dispersion: the sample variance.

Sample variance

One way to begin to develop a statistical analog of the population variance is to return to the empirical distribution. Recall from Definition 1.4 that the empirical distribution was developed by assigning a probability of $1/n$ to each data value. This allowed us to see that the "plug-in" estimator of the population mean was the mean of this empirical distribution. We will now determine the plug-in estimator of the population variance.

As in the previous two sections, let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. We will apply the formula for the population variance from probability theory

$$\sigma^2 = V[X] = E\left[(X-\mu)^2\right]$$

to the empirical probability distribution associated with the data values. Since the empirical probability distribution is a discrete probability distribution, the formula for the population variance becomes

$$\sigma^2 = V[X] = E\left[(X-\mu)^2\right] = \sum_{\mathcal{A}} (x-\mu)^2 f(x),$$

where $\mathcal{A}$ is the support of the random variable X. Since $f(x_i) = 1/n$, for $i = 1, 2, \ldots, n$ and μ can be replaced by its plug-in estimator $\bar{x}$, the plug-in estimator for the population variance from the empirical distribution becomes

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

At first glance, this estimator seems quite reasonable. It is the arithmetic average of the squared deviations of each data value from the sample mean. One flaw will become apparent through the following thought experiment.

Let's say you go to Mars and observe the height of $n = 1$ Martian, who is exactly 2 feet tall. So the estimate of the population mean is just the sample mean: $\bar{x} = 2$. Now you would like to estimate the population variance σ^2. The summation in the plug-in estimator for the population variance has only a single term, which is zero because $x_1 = \bar{x}$, so $\hat{\sigma}^2 = 0$. But is a population variance estimate of zero appropriate here? A variance of zero implies that *all* Martians are exactly 2 feet tall. All Martians *might* indeed be 2 feet tall or they might range from 1 foot tall to 10 feet tall. You just cannot tell by observing a single Martian. In fact, you cannot estimate the population variance at all when $n = 1$. So the definition of the sample variance given next is undefined when $n = 1$, and this is accomplished by dividing the sum of squares by $n - 1$ rather than n.

Definition 1.10 Let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. The *sample variance* is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 .$$

The positive square root of the sample variance is the *sample standard deviation* S.

The $n-1$ in the denominator is counter-intuitive to most people who encounter this version of the sample variance for the first time. Here are three comments on the unusual denominator.

- As mentioned earlier, a sample size of $n = 1$ means that S^2 is undefined.
- The summation $\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a statistic with $n-1$ "degrees of freedom." In many application areas, statisticians divide statistics by their degrees of freedom. So dividing by $n-1$ here is consistent with that practice. One degree of freedom is lost by using $\bar{X}$ rather than μ in this expression.
- Perhaps more important than the first two comments, the sample variance is an unbiased estimate of the population variance, that is $E\left[S^2\right] = \sigma^2$ for any population with finite population mean and finite population variance. This result will be formally stated and proved in the next chapter. The important take-away here is that the sample variance S^2 is "on target" for estimating σ^2. Using anything other than $n-1$ in the denominator of the formula for S^2 would result in a sample variance that is not aimed at the population variance σ^2.

Computing the sample variance is not a trivial matter for large data sets. The "defining formula" given in Definition 1.10 leads to a two-pass algorithm for computing s^2, in that the data must be passed over twice in order to compute s^2: once to compute $\bar{x}$, then a second time to compute $\sum_{i=1}^{n}(x_i - \bar{x})^2$. This procedure is inefficient computationally. Some simple algebra can lead to a one-pass algorithm:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right].$$

In the one-pass algorithm, the data is passed over just once, computing the running values of $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i$.

Even with the one-pass algorithm, problems still arise. For huge data sets that must be processed on a computer using floating-point arithmetic, there is the possibility of round-off and overflow errors. There is additional risk when the data values themselves are large numbers. A dozen or so algorithms have been designed to decrease the possibility of these issues having an impact on the computed value of s^2.

The sample variance is computed in R with the `var` function, which is illustrated next for a small sample.

Example 1.27 Consider again the data set of $n = 10$ observations from Example 1.15, where kindergarteners were polled concerning the number of children in their family. The data values $x_1, x_2, \ldots, x_{10}$ are

$$3, 1, 5, 1, 3, 2, 1, 1, 3, 2.$$