Calculus based approach to statistics: Mean, Median, Variance, and Standard Deviation

Mean

In the last section we saw that if savings and loan institutions are continuously failing at a rate of 5% per year, then the associated probability density function is f (x) = 0.05e−0.05x with domain [0, +∞). An interesting and important question to ask is: What is the average length of time such an institution will last before failing? To answer this question, we use the following:

Mean or Expected Value

If X is a continuous random variable with probability density function f defined on an interval with (possibly infinite) endpoints a and b, then the mean or expected value of X is

$$E(X) = \int_a^b x f(x) dx.$$

E(X) is also called the average value of X. It is what we expect to get if we take the average of many values of X obtained in experiments.

Why is E(X) given by that integral formula?

Suppose for simplicity that the domain of *f* is a finite interval [*a*, *b*]. Break up the interval into *n* subintervals [*xk*−1, *xk*], each of length _*x*, as we did for Riemann sums. Now, the probability of seeing a value of *X* in [*xk*−1, *xk*] is approximately *f* (*xk*)_*x* (the approximate area under the graph of *f* over [*xk*−1, *xk*]). Think of this as the fraction of times we expect to see values of *X* in this range. These values, all close to *xk*, then contribute approximately

xk f (*xk*)_*x* to the average, if we average together many observations of *X*.

Adding together all of these contributions, we get

$$E(X) \approx \sum_{k=1}^{n} x_k f(x_k) \Delta x.$$

Now, these approximations get better as $n \longrightarrow \infty$, and the sum above is a Riemann sum converging to

$$E(X) = \int_a^b x f(x) dx,$$

which is the formula we have been using.

We can compute, one by one, the expected values of the distributions we discussed in the preceding section.

Mean of a Uniform Distribution

If X is uniformly distributed on $[a, b]$, then

$$E(X) = \frac{a+b}{2}.$$

Variance and Standard Deviation

Statisticians use the variance and standard deviation of a continuous random variable X as a way of measuring its dispersion, or the degree to which is it "scattered." The definitions are as follows:

Notes

1. In order to calculate the variance and standard deviation, we first need to calculate the mean.

2. $Var(X)$ is the expected value of the function $(x - \mu)2$, which measures the square of the distance of X from its mean. It is for this reason that $Var(X)$ is sometimes called the *mean square deviation*, and $\sigma(X)$ is called the *root mean square deviation*. $Var(X)$ will be larger if X tends to wander far away from its mean, and smaller if the values of

X tend to cluster near its mean.

3. The reason we take the square root in the definition of $\sigma(X)$ is that $Var(X)$ is the expected value of the *square* of the deviation from the mean, and thus is measured in square units. Its square root $\sigma(X)$, therefore, gives us a measure in ordinary units.

We now state the variances and standard deviations of the distributions we discussed in the previous section. We'll leave the actual computations (or special cases) for the exercises.

You can see the significance of the standard deviation quite clearly in the normal distribution. As we mentioned in the previous section, σ is the distance from the maximum at μ to the points of inflection at $\mu - \sigma$ and $\mu + \sigma$. The larger σ is, the wider the bell. Figure 15 shows three normal distributions with three different standard deviations (all with $\mu = 0.5$).

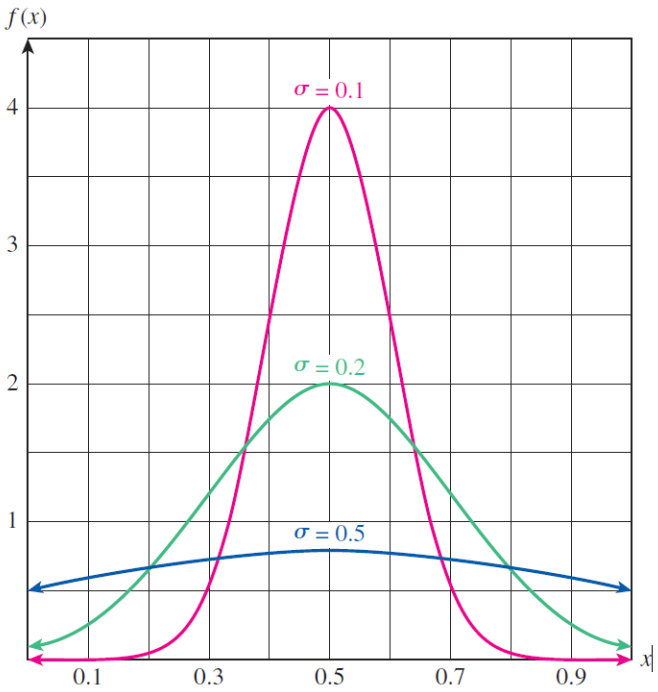

Figure 15

Again, a small standard deviation means that the values of X will be close to the mean with high probability, while a large standard deviation means that the values may wander far away with high probability.

Median

Let X be a continuous random variable. The median of X is the number M such that

$P(X \leq M) = ½$

Then, $P(M \leq X) = ½$ also.

What is the difference between the median and the mean?

Roughly speaking, the median divides the area under the distribution curve into two equal parts, while the mean is the value of X at which the graph would *balance*. If a probability curve has as much area to the left of the mean as to the right, then the mean is equal to the median. This is true of uniform and normal distributions, which are *symmetric* about their means. On the other hand, the medians and means are different for the exponential distributions and most of the beta distributions, because their areas are not distributed symmetrically.