LOGNORMAL II

Suppose Y is a lognormal random variable – so X = logY is normal, say with mean and standard deviation μ and σ. Now suppose that $y_1, y_2, \ldots, y_n$ form a random sample from Y. Define $x_i = \log(y_i)$. Then $x_1, x_2, \ldots, x_n$ form a random sample from X. So their arithmetic mean $\bar{x}$ is a data-based estimate for μ. Hence $e^{\bar{x}}$ is a reasonable estimate for e^{μ}. But

$e^{\bar{x}}$ = ____,

which is the __________________________________ of $y_1, y_2, \ldots, y_n$. Hence we call e^{μ}

the __________________________ of Y.

Recall: e^{μ} is also the _____________ of Y.

We analogously define the *geometric standard deviation* of $y_1, y_2, \ldots, y_n$ to be

$\mathrm{GSD}(y_1, y_2, \ldots, y_n) = \exp(\text{standard deviation of } x_1, x_2, \ldots, x_n) =$

exp(__________________________________)

Caution: In defining the geometric standard deviation, some people use the definition of standard deviation with n in the denominator (the formula for the standard deviation of a population with n values), some the definition with n-1 in the denominator (the sample standard deviation. Both of these yield estimators of σ (one biased, the other unbiased), so $e^{\mathrm{GSD}(y1,\, y2,\, \ldots,\, yn)}$ estimates e^{σ}. We thus analogously call $\sigma^* = e^{\sigma}$ the *geometric standard deviation* of Y.

Exercise: Look again at Problem 2(e) from the handout Logs and Means:

e. The following quote is from p. 54 of the article "Wading in Waste", by Mallin, Michael A., *Scientific American*, Jun 2006, Vol. 294, Issue 6:

"*The U.S. Public Health Service has set a nationwide safety standard for shellfish beds using measurements of fecal coliform bacteria, a broad category of microorganisms found in the intestines of humans and animals. Shellfish cannot be harvested from the area if the geometric mean of the bacterial counts in 30 sets of samples is higher than 14 colony-forming units (CFU) per 100 milliliters of seawater. (A geometric mean is a type of average that minimizes the effects of outlying values.)* "

Outlying values are data values that are either noticeably smaller or noticeably larger than other data values occurring.

i. Explain why the geometric mean minimizes the effect of outlying values compared to the arithmetic mean.

ii. Why would it be appropriate (or not) to minimize the effect of outlying values in this situation?