The Lognormal Distribution

The lognormal distribution applies when the measured scores (assumed all positive) vary over several orders of magnitude and the probability of the log of a score is approximated by a normal distribution. Let $x$ stand for the score and let $a$ and $b$ stand for the mean and standard deviation respectively of the $\ln(x)$, i.e., the $\ln(x)$ is distributed like $N(a, b)$. From this we can deduce the distribution of the random variable $x$.

Under the assumption of the normality of $\ln(x)$, the probability of a measurement less than or equal to $x$ is given by

$$F_n(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\ln(x) - a} e^{-z^2/2} dz & \text{if } x < 0 \\ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-z^2/2} dz & \text{if } x > 0 \end{cases}.$$ 

It follows immediately that $F_n(e^a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-z^2/2} dz = \frac{1}{2}$, so the median of a lognormal distribution is $e^a = \exp(a)$. The probability density function (pdf) for the scores $x$ is the derivative of this distribution function and is given by the following.

$$f_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{\sqrt{2\pi b x}} \exp \left( -\frac{(\ln(x) - a)^2}{2b^2} \right) & \text{if } x > 0 \end{cases} = \frac{1}{\sqrt{2\pi b}} \exp \left( -\frac{2b^2 \ln(x) + (\ln(x) - a)^2}{2b^2} \right).$$

At the median, the pdf has the value $f_n(e^a) = \frac{1}{\sqrt{2\pi b}} \exp(-a) = \frac{1}{\sqrt{2\pi b e^a}}$.

The derivative of the pdf is given by

$$\frac{d}{dx} f_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ -\left( \frac{b^2 + \ln(x) - a}{\sqrt{2\pi b^3x}} \right) \exp \left( -\frac{2b^2 \ln(x) + (\ln(x) - a)^2}{2b^2} \right) & \text{if } x > 0 \end{cases},$$

which is only zero at $\ln(x) = a - b^2$, so the mode of a lognormal distribution is $e^{a-b^2} = \exp(a - b^2)$. At the mode, the pdf has the value

$$f_n(e^{a-b^2}) = \frac{1}{\sqrt{2\pi b}} \exp \left( -a + \frac{b^2}{2} \right) = \frac{e^{b^2/2}}{\sqrt{2\pi b e^a}}.$$ 

Graphs of the pdf of the lognormal distribution for various values of $a$ and $b$ are displayed on the next page.
The moment generating function is not useful for the lognormal distribution since
\[ M(t) = \left( e^{xt} \right) = \int_{0}^{\infty} \frac{1}{\sqrt{2\pi b x}} \exp \left( x t - \frac{\ln(x) - a}{2b^2} \right) dx \]
diverges for all positive values of \( t \).

However all of the moments of the lognormal distribution can be calculated by explicitly evaluating the integral, \( \langle x^p \rangle = \int_{0}^{\infty} x^p f(x) dx \). Under the substitution, \( x = e^y \),
\[ \langle x^p \rangle = \int_{-\infty}^{\infty} e^{py} f_{\ln}(e^y) e^y dy = \frac{1}{\sqrt{2\pi b}} \int_{-\infty}^{\infty} e^{-(y-a)^2/(2b^2)} e^{py} dy \]
\[ = \frac{1}{\sqrt{2\pi b}} \int_{-\infty}^{\infty} e^{-(y^2-2ay+2ab^2+a^2)/(2b^2)} dy = \frac{1}{\sqrt{2\pi b}} \int_{-\infty}^{\infty} e^{-(y^2-2(a+pb^2)y+(a+pb^2)^2)/(2b^2)} e^{y(a+pb^2)^2/(2b^2)} dy \]
\[ = e^{pa+pb^2/2} \frac{1}{\sqrt{2\pi b}} \int_{-\infty}^{\infty} e^{-\left(y-(a+pb^2)^2/(2b^2)\right)} dy \]

Now, making the substitution, \( z = \frac{y-(a+pb^2)}{b} \),
\[ \langle x^p \rangle = e^{pa+pb^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = e^{pa+pb^2/2} = \exp \left( pa + \frac{p^2 b^2}{2} \right) \]

Hence the mean of the lognormal distribution is given by
\[ \mu_x = \langle x \rangle = e^{a+b^2/2} = \exp \left( a + \frac{b^2}{2} \right) \].
The Lognormal Distribution

The second moment is $\langle x^2 \rangle = e^{2a+4b^2/2} = \exp \left( 2a + 2b^2 \right)$.

So the variance and standard deviation of a lognormal distribution are given by

$$
\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2 = e^{2(a+b^2)} - \left( e^{a+b^2/2} \right)^2 = e^{2a+b^2} \left( e^{b^2} - 1 \right)
$$

$$
\sigma_x = \sqrt{e^{2a+b^2} \left( e^{b^2} - 1 \right)} = e^{a+b^2/2} \sqrt{e^{b^2} - 1} = \mu_x \sqrt{e^{b^2} - 1}.
$$

For an independent sample of size $n$ taken from a lognormal distribution the expected value of the sample mean is the population mean, $\langle \bar{x} \rangle = \mu_x = \exp \left( a + b^2/2 \right)$ and the standard error of the mean is given by $\frac{\sigma_x}{\sqrt{n}} = \mu_x \sqrt{\frac{e^{b^2} - 1}{n}}$.

Note that $a = \langle \ln (x) \rangle < \ln \left( \langle x \rangle \right) = a + b^2/2$. For large samples, the spread about $\mu_x$ goes to zero like the reciprocal of the square root of the sample size as expected.

Instead of the arithmetic mean, $\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$, consider the geometric mean of a sample of $n$ scores, $\text{GM}(x) = \sqrt[n]{x_1 x_2 \cdots x_n} = \left( \prod_{j=1}^{n} x_j \right)^{1/n}$. As is shown in the lemma of the appendix, for any sample size larger than one the geometric mean is always smaller than the arithmetic mean.

Consider an independent sample of size $n$ taken from a lognormal distribution the expected value of the geometric mean of the sample is given by the $n$-fold integral.

$$
\langle \text{GM}(x) \rangle = \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{n} x_j f_\ln \left( x_1 \right) f_\ln \left( x_2 \right) \cdots f_\ln \left( x_n \right) dx_1 dx_2 \cdots dx_n
$$

$$
= \int_0^\infty \left( x_1 \right)^{1/n} f_\ln \left( x_1 \right) dx_1 \int_0^\infty \left( x_2 \right)^{1/n} f_\ln \left( x_2 \right) dx_2 \cdots \int_0^\infty \left( x_n \right)^{1/n} f_\ln \left( x_n \right) dx_n
$$

$$
= \left( \int_0^\infty \left( x \right)^{1/n} f_\ln \left( x \right) dx \right)^n = \left( \langle x^{1/n} \rangle \right)^n = \left[ \exp \left( \frac{a}{n} + \frac{b^2}{2n^2} \right) \right]^n
$$

$$
= \exp \left( a + \frac{b^2}{2n} \right)
$$

As anticipated, for samples of more than one score, $\langle \text{GM}(x) \rangle < \langle \bar{x} \rangle = \exp \left( a + b^2/2 \right)$.

By a similar argument the second moment of the sample geometric mean can be computed to yield an expression for the variance of the sample geometric mean.
The Lognormal Distribution

\[ \langle \text{GM}(x)^2 \rangle = \int_{0}^{\infty} \int_{0}^{\infty} \cdots \int_{0}^{\infty} (\sqrt[2]{x_1x_2\cdots x_n})^2 f_{\ln}(x_1)f_{\ln}(x_2)\cdots f_{\ln}(x_n)dx_1dx_2\cdots dx_n \]

\[ = \int_{0}^{\infty} (x_1)^{2/n} f_{\ln}(x_1)dx_1 \int_{0}^{\infty} (x_2)^{2/n} f_{\ln}(x_2)dx_2 \cdots \int_{0}^{\infty} (x_n)^{2/n} f_{\ln}(x_n)dx_n \]

\[ = \left( \int_{0}^{\infty} (x)^{2/n} f_{\ln}(x)dx \right)^n = (\langle x^{2/n} \rangle)^n = \left[ \exp\left( \frac{2a + 2b^2}{n} \right) \right]^n \]

\[ = \exp\left( 2a + \frac{2b^2}{n} \right) \]

\[ \text{var}(\text{GM}(x)) = \langle \text{GM}(x)^2 \rangle - (\langle \text{GM}(x) \rangle)^2 = \exp\left( 2a + \frac{2b^2}{n} \right) - \exp\left( a + \frac{b^2}{2n} \right)^2 \]

\[ = \exp\left( 2a + \frac{2b^2}{n} \right) - \exp\left( 2a + \frac{b^2}{n} \right) = \exp\left( 2a + \frac{b^2}{n} \right) \left( \exp\left( \frac{b^2}{n} \right) - 1 \right) \]

So the standard error of the geometric mean of a lognormal distribution is given by

\[ \sigma(\text{GM}(x)) = \sqrt{\exp\left( \frac{b^2}{2n} \right) - 1} = \frac{\text{GM}(x)}{\sqrt{n}} \cdot \sqrt{\exp\left( \frac{b^2}{2n} \right) - 1}. \]

For large samples \( \langle \text{GM}(x) \rangle \rightarrow e^a \) and \( \sigma(\text{GM}(x)) \rightarrow e^a \frac{b}{\sqrt{n}}. \)

\[ a = \langle \ln(x) \rangle < \ln(\langle \text{GM}(x) \rangle) = a + \frac{b^2}{2n}, \text{ however in contrast to the arithmetic mean the log of the geometric mean approaches the mean logarithm as the sample size increases.} \]

Defining the population geometric mean, \( \Gamma M_x \), as the infinite sample size limit of the expected value of the sample geometric mean gives \( \Gamma M_x = \lim_{n \to \infty} \langle \text{GM}(x) \rangle = e^a \) which is identical to the median score of a lognormal distribution.

Thus, for data that is spread out over several orders of magnitude (like aerosol particle size) the geometric mean gives a more representative measure of a typical score than the arithmetic mean which will be skewed to right because of large measurements.
Appendix: Lemma: The Arithmetic Mean is Greater than the Geometric Mean.

George Polya's Proof:
Consider for positive $x$, the continuous function $f(x) = x - 1 - \ln(x)$.

As $x \to 0^+$, $f(x) \to \infty$ and as $x \to \infty$, $f(x) = (x-1)\left(1 - \frac{\ln(x-1)}{x-1}\right) \to \infty$. Therefore, $f(x)$ must have an absolute minimum on $(0, \infty)$. The only zero of $f'(x) = 1 - \frac{1}{x}$ is at $x = 1$. Hence, $f(x) \geq f(1) = 0$, or equivalently for all positive $x$, $\ln(x) \leq x - 1$ with equality only at $x = 1$. Consider a set of $n$ positive numbers $\{u_j\}_{j=1}^n$. The arithmetic mean of the $u_j$ is given by $\bar{u} = \frac{\sum_{j=1}^n u_j}{n}$. Now for each $j$, $1 \leq j \leq n$, $\ln\left(\frac{u_j}{\bar{u}}\right) \leq \frac{u_j}{\bar{u}} - 1$.

Therefore, $\sum_{j=1}^n \ln\left(\frac{u_j}{\bar{u}}\right) \leq \sum_{j=1}^n \left(\frac{u_j}{\bar{u}} - 1\right) = \frac{1}{\bar{u}} \sum_{j=1}^n (u_j - \bar{u}) = 0$. Hence,

$$\ln\left(\prod_{j=1}^n \frac{u_j}{\bar{u}^n}\right) = \ln\left(\prod_{j=1}^n \frac{u_j}{\bar{u}}\right) \leq 0,$$

or $\prod_{j=1}^n \frac{u_j}{\bar{u}} \leq 1$. Note unless each $\frac{u_j}{\bar{u}} = 1$, which is only true for a set of positive constants, $\frac{\prod_{j=1}^n u_j}{\bar{u}^n}$ is strictly less than 1. Therefore for any set of positive scores $\{u_j\}_{j=1}^n$, $\prod_{j=1}^n u_j \leq \bar{u}^n$ or $\sqrt[n]{\prod_{j=1}^n u_j} \leq \bar{u}$ with equality if and only if all the scores are equal or the special case of only one score sampled.