

# Mathematical Biostatistics Bootcamp: Lecture 14, Logs

Brian Caffo

Department of Biostatistics  
Johns Hopkins Bloomberg School of Public Health  
Johns Hopkins University

October 16, 2012

# Table of contents

- 1 Table of contents
- 2 Logs
- 3 The geometric mean
- 4 GM and the CLT
- 5 Comparisons
- 6 The log-normal distribution

# Logs

- Recall that  $\log_B(x)$  is the number  $y$  so that  $B^y = x$
- Note that you can not take the log of a negative number;  $\log_B(1)$  is always 0 and  $\log_B(0)$  is  $-\infty$
- When the base is  $B = e$  we write  $\log_e$  as just  $\log$  or  $\ln$
- Other useful bases are 10 (orders of magnitude) or 2
- Recall that  $\log(ab) = \log(a) + \log(b)$ ,  $\log(a^b) = b \log(a)$ ,  $\log(a/b) = \log(a) - \log(b)$  (log turns multiplication into addition, division into subtraction, powers into multiplication)

## Some reasons for “logging” data

- To correct for right skewness
- When considering ratios
- In settings where errors are feasibly multiplicative, such as when dealing with concentrations or rates
- To consider orders of magnitude (using log base 10); for example when considering astronomical distances
- Counts are often logged (though note the problem with zero counts)

## The geometric mean

- The (sample) **geometric mean** of a data set  $X_1, \dots, X_n$  is

$$\left( \prod_{i=1}^n X_i \right)^{1/n}$$

- Note that (provided that the  $X_i$  are positive) the log of the geometric mean is

$$\frac{1}{n} \sum_{i=1}^n \log(X_i)$$

- As the log of the geometric mean is an average, the LLN and clt apply (under what assumptions?)
- The geometric mean is always less than or equal to the sample (arithmetic) mean

## The geometric mean

- The geometric mean is often used when the  $X_i$  are all multiplicative
- Suppose that in a population of interest, the prevalence of a disease rose 2% one year, then fell 1% the next, then rose 2%, then rose 1%; since these factors act multiplicatively it makes sense to consider the geometric mean

$$(1.02 \times .99 \times 1.02 \times 1.01)^{1/4} = 1.01$$

for a 1% geometric mean increase in disease prevalence

- Notice that multiplying the initial prevalence by  $1.01^4$  is the same as multiplying by the original four numbers in sequence
- Hence 1.01 is constant factor by which you would need to multiply the initial prevalence each year to achieve the same overall increase in prevalence over a four year period
- The arithmetic mean, in contrast, is the constant factor by which your would need to *add* each year to achieve the same *total* increase  $(1.02 + .99 + 1.02 + 1.01)$
- In this case the product and hence the geometric mean make more sense than the arithmetic mean

## Nifty fact

- The *question corner* (google) at the University of Toronto's web site (where I got much of this) has a fun interpretation of the geometric mean
- If  $a$  and  $b$  are the lengths of the sides of a rectangle then
  - The arithmetic mean  $(a + b)/2$  is the length of the sides of the square that has the same perimeter
  - The geometric mean  $(ab)^{1/2}$  is the length of the sides of the square that has the same area
- So if you're interested in perimeters (adding) use the arithmetic mean; if you're interested in areas (multiplying) use the geometric mean



# Asymptotics

- Note, by the LLN the log of the geometric mean converges to  $\mu = E[\log(X)]$
- Therefore the geometric mean converges to  $\exp\{E[\log(X)]\} = e^\mu$ , which is *not* the population mean on the natural scale; we call this the population geometric mean (but no one else seems to)

- To reiterate

$$\exp\{E[\log(x)]\} \neq E[\exp\{\log(X)\}] = E[X]$$

- Note if the distribution of  $\log(X)$  is symmetric then

$$.5 = P(\log X \leq \mu) = P(X \leq e^\mu)$$

- Therefore, for log-symmetric distributions the geometric mean is estimating the median

## Using the CLT

- If you use the CLT to create a confidence interval for the log measurements, your interval is estimating  $\mu$ , the expected value of the log measurements
- If you exponentiate the endpoints of the interval, you are estimating  $e^{\mu}$ , the population geometric mean
- Recall,  $e^{\mu}$  is the population median when the distribution of the logged data is symmetric
- This is especially useful for paired data when their ratio, rather than their difference, is of interest

# Example

# Comparisons

- Consider when you have two independent groups, logging the individual data points and creating a confidence interval for the difference in the log means
- Prove to yourself that exponentiating the endpoints of this interval is then an interval for the *ratio* of the population geometric means,  $\frac{e^{\mu_1}}{e^{\mu_2}}$

## The log-normal distribution

- A random variable is **log-normally** distributed *if its log is a normally distributed random variable*
- “I am log-normal” means “take logs of me and then I’ll then be normal”
- Note log-normal random variables are not logs of normal random variables!!!!!! (You can’t even take the log of a normal random variable)
- Formally,  $X$  is  $\text{lognormal}(\mu, \sigma^2)$  if  $\log(X) \sim N(\mu, \sigma^2)$
- If  $Y \sim N(\mu, \sigma^2)$  then  $X = e^Y$  is log-normal

# The log-normal distribution

- The log-normal density is

$$\frac{1}{\sqrt{2\pi}} \times \frac{\exp[-\{\log(x) - \mu\}^2 / (2\sigma^2)]}{x} \quad \text{for } 0 \leq x \leq \infty$$

- Its mean is  $e^{\mu + (\sigma^2/2)}$  and variance is  $e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$
- Its median is  $e^{\mu}$

## The log-normal distribution

- Notice that if we assume that  $X_1, \dots, X_n$  are log-normal( $\mu, \sigma^2$ ) then  $Y_1 = \log X_1, \dots, Y_n = \log X_n$  are normally distributed with mean  $\mu$  and variance  $\sigma^2$
- Creating a Gosset's  $t$  confidence interval on using the  $Y_i$  is a confidence interval for  $\mu$  the log of the median of the  $X_i$
- Exponentiate the endpoints of the interval to obtain a confidence interval for  $e^\mu$ , the median on the original scale
- Assuming log-normality, exponentiating  $t$  confidence intervals for the difference in two log means again estimates ratios of geometric means

# Example