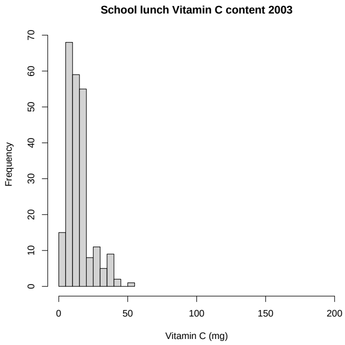
Taking logs - why and how?

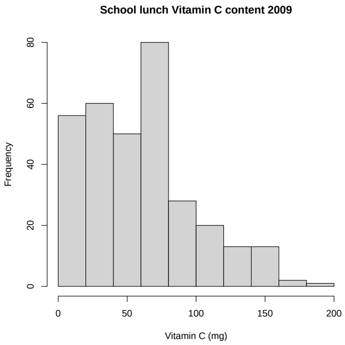
J N S Matthews

Biostatistics Research Group, Newcastle University

Some data

Consider the data on Vitamin C intake (mg) from school lunches [from Reception, Years 1 & 2]





Left hand from 2003 and right hand from 2009 (from Spence *et al.* 2013)

Comments

Year	n	Mean	SD	Min	Q1	Median	Q3	Max
2003	233	14.5	8.6	0.1	8.5	12.5	17.4	52.4
2009	323	60.0	38.4	5.8	26.2	59.1	77.8	184.7

Table: Summary statistics for the vitamin C intakes (mg)

- Aim is to compare intakes between 2003 and 2009
- Thoughts of using a *t*-test fade as data look skewed
- Also means are less than two SDs, so again unlikley to be Normal as Vitamin C is non-negative
- SDs very different

So what does the non-statistician do?

Tendency is to reach for non-parametric aka distribution-free aka rank-based methods. Is this OK?

1. Tests hypothesis $F_1(\cdot) = F_2(\cdot)$. I.e. samples are from same distribution - like a *t-test* only if equal variances assumed
2. Based on ranks - is this OK?
3. Estimation preferred over testing - now focuses on medians not means. Medians recommended as they have a high *breakdown point* of 50%. Is this a good thing?
4. Usually no SEs with medians - OK as confidence intervals are available. However, usually based on assumption $F_2(x) = F_1(x - \theta)$. So equal dispersion assumed - method not assumption free (or even non-parametric)
5. Distribution-free methods, at least the common ones, usually not rich enough for most purposes

Need we bother?

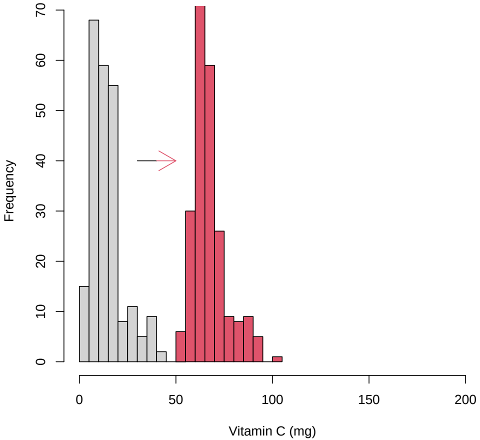
- Could just ignore skewness
- For samples of reasonable size, and inference based on median, distribution anxiety might be assuaged by Central Limit Theorem
- Might be being a bit cavalier with differences in SDs

So what does the statistician do?

- Of course, most statisticians would analyse the logs of the Vitamin C values. Why?
- Well, often explained in terms of distributional shape - log of Vit C will be closer to Normal.
- Yes, OK, but
- arguably because of inadequacy of approaches that are essentially additive when applied to positive, skewed data.

Are additive effects OK?

Difference in mean (or median for that matter) of Vit C between surveys is about 50 mg. Adding 50 to mean of 2003 data gives following:

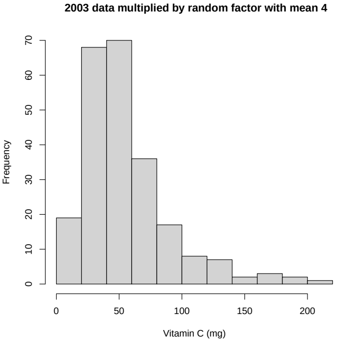


Red plot looks nothing like 2009 data:
wrong shape;
wrong spread.

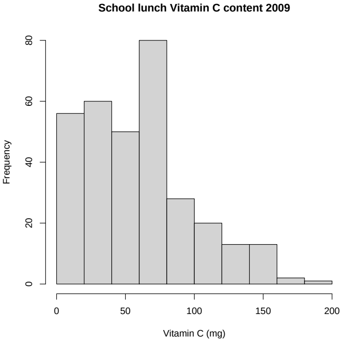
What about multiplication?

- Mean in 2009 about four times that in 2003
- Suppose $x_i, i = 1, \ldots, 233$ are the 2003 Vit C values
- Suppose f_i are 233 independent realisations of a gamma variate with mean 4 and variance 1.
- Form $x_i f_i, i = 1, \ldots, 233$ and plot these

Multiplicative effects



Scaled by f_i



2009 data

So multiplying 2003 data by a four-fold factor looks more convincing - and logs can mediate between additive and multiplicative effects

Transformations

General approach is:

1. Select transformation g such that the $g(x_i), i = 1, \ldots, n$ are Normal
2. Analyse the $g(x_i)$
3. Present results *on original scale*, with appropriate use of $g^{-1}(\cdot)$

Often point 1 receives most attention cf. Box & Cox (1964)

But unless point 3 is done convincingly and understandably, whole exercise is less compelling

Taking logs

Suppose Vit C values in 2003 are the xs and in 2009 the ys, then we calculate

$$m_3 = \frac{1}{n_3} \sum_{i=1}^{n_3} \log x_i : \qquad m_9 = \frac{1}{n_9} \sum_{i=1}^{n_9} \log y_i$$

- These means will not look like Vit C values
- So we report $\exp(m_3)$ and $\exp(m_9)$ as plausible and comprehensible measures of location

What about the difference between 2003 & 2009?

- Difference between Normal variables are appropriate, so $m_9 - m_3$ is a suitable measure of difference
- But it is on the log scale - does evaluating $\exp(m_9 - m_3)$ make sense?
- Yes it does

$$\exp(m_9 - m_3) = \frac{\exp(m_9)}{\exp(m_3)}$$

- So discrepancy between 2003 & 2009 is now in terms of a *ratio* of the individual year means
- This is the Heineken property - only logs can do this

Other transformations

- If we had, e.g., used $g(x) = \sqrt{x}$, with m_9, m_3 (appropriately redefined) now Normal on the square root scale, then $m_9 - m_3$ would still be a suitable measure of difference
- But $(m_9 - m_3)^2$ is no longer just a function of m_9^2 and m_3^2
- So no simple form for the discrepancy on original scale, based on some measure of discrepancy between m_3^2, m_9^2, arises naturally.
- While $g(\cdot) \neq \log(\cdot)$ may give more Normal data, this lack of a compelling way to back-transform makes non-log transformation much less attractive.

Geometric means

Now returning to the log transformation

- We can readily contrast 2003 and 2009 using the $\exp(m_3), \exp(m_3)$ but what are they? Are they means?
- They are, but not *arithmetic means*. They are *geometric means*, defined as, e.g.,

$$\exp(m_3) = \exp\left(\frac{1}{n_3}\sum_{i=1}^{n_3}\log x_i\right) = \sqrt[n_3]{\left(\prod_{i=1}^{n_3} x_i\right)},$$

Properties of geometric means (GMs)

- GMs defined for positive values only
- If A, G are the arithmetic and geometric means, respectively, of some data then $G \leq A$, with equality only if all values are equal.
- With positively skewed data, median is less than arithmetic mean, and often closer to geometric mean
- Large values perturb GM less than the AM - useful alternative to median
- GM can be sensitive to changes in small values.

Some theoretical considerations

Although logs work well with many skewed distributions, most insight comes from assuming Y is log-Normal - i.e. $Y = \exp(X)$ where X is Normal with mean μ and variance σ^2.

Worth recalling that the moment generating function of a Normal is:

$$M(t) = \mathsf{E}[\exp(tX)] = \exp(\mu t + \tfrac{1}{2}t^2\sigma^2)$$

Theoretical comments on log-Normal

1. $\mathsf{E}[Y] = M(1) = \exp(\mu + \frac{1}{2}\sigma^2)$, so AM larger than $\exp(\mu)$
2. As $\exp$ is monotone increasing,
   $\frac{1}{2} = \Pr(X < \mu) = \Pr(Y < e^{\mu})$, so e^{μ} is the median of Y
3. The variance of Y is

$$M(2) - M(1)^2 = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

so SD of Y is proportional to its mean. The CV of Y, i.e. SD divided by mean is $\sqrt{\exp(\sigma^2) - 1}$ and for small σ this is $\approx \sigma$.

Yet more theoretical comments

- For a (positive) random variable Y its geometric mean is defined as
$$\exp(\mathsf{E}[\log(Y)])$$
- For log-Normal Y this is e^{μ}, which coincides with the median
- Regardless of the distribution of Y, the GM of Y is less than $\mathsf{E}[Y]$, i.e. its AM, so the AM-GM inequality holds for random variables. To see this, apply Jensen's inequality and note that $\log$ is concave.

Practical arithmetic

This is very truncated - for details see Section 4 of the associated document.

Summary of data

Year	n	Mean	SD	Mean (logs)	SD (logs)	Geometric mean
2003	233	14.5	8.6	2.491	0.680	12.1
2009	323	60.0	38.4	3.853	0.752	47.2

Main comparison is 3.853-2.491=1.362. But this is on the log-scale, so antilog

$$\exp(3.853 - 2.491) = \exp(1.362) = 3.90 = \frac{\exp(3.853)}{\exp(2.491)},$$

So difference is described on original scale by a ratio - and of GMs not AMs - i.e. GM in 2009 is about four times that in 2003

Unlogging the CI

Apply standard methods to logged value to get 95% CI for difference in means on log scale

$$60.0 - 14.5 \pm 1.96 \times 0.723\sqrt{\left(\frac{1}{233} + \frac{1}{323}\right)} = (1.241, 1.485)$$

So, point estimate 1.362 is anti-logged to get 3.90 and interval estimate (1.241,1.485) is anti-logged to get interval estimate for 3.90, namely 3.46 to 4.41.

Should I anti-log the estimated SE?

No

Hypothesis test

- Hypothesis of equality of AMs on logged scale is $\mu_1 = \mu_2$
- This is the same as testing $\exp(\mu_1) = \exp(\mu_2)$, i.e. testing equality of GMs on original scale
- So P-value to be reported is unaffected by the transformation

Back to SE

- Why shouldn't you anti-log the SE?
- Presumably would want to get a measure of uncertainty
- Not needed as you have an interval estimate
- Also, $\exp(s)$ does not provide a measure of uncertainty, at least not analogous to an SE.
- For log-Normal, the sampling distribution of the sample GM is log-Normal, with expectation and SD, respectively

$$\exp(\mu + \tfrac{1}{2n}\sigma^2) \qquad \exp(\mu + \tfrac{1}{2n}\sigma^2)\sqrt{\exp(\sigma^2/n) - 1}$$

- Sampling variation depends on μ, but s is not dependent on μ, so $\exp(s)$ cannot provide the relevant information

Miscellaneous comments

Several issues are mentioned in the accompanying article, two of which are mentioned without expansion here.

- For most purposes, with a skew distribution, the GM is a highly suitable summary for location.
- For cost data, which are often skew, it is the AM that is pertinent. If the mean cost per patient is m, then the cost of treating N patients is Nm only if m is the AM *not* the GM.
- Zeroes in the data. Faced with skewed data that one would like to log, zeroes are a real pain. Sensible ways round this depend on the context and the provenance of the zeroes.