Statistical Science

Chapter 3.3 Descriptive Statistics and Rescaling

ReCap. Quantitative Reasoning(Ch 1)
Quantities (Ch2)
Re-Scaling (Ch3)
3.1 Logical Re-scaling
3.2 Operations on Ratio Scale Quantities
3.3 Descriptive Statistics and Rescaling
Normalization – General Definition
Normalization to the minimum – Scope
Normalization to the sum
Normalization to the mean
Normalization to measures of dispersion
3.4 Unit Conversion and Rigid Rescaling

Recap Chapter 1
Quantitative reasoning: Example of scallops, which combined stats and models
Recap Chapter 2
Quantities: Five part definition
Measurements made on four types of scale: nominal, ordinal, interval, ratio
Recap Chapter 3. Re-scaling
Logical rescaling (from one type of unit to another).
Re-scaling is a common technique in quantitative biology.
Operations on measured quantities differ from operations on numbers.
-the rules differ
-physically interpretable, not just abstract mathematical procedures

Today: Normalization in science and in statistics.

Wrap-up:
We can convert a scaled quantity to a ratio with no units by rescaling it to a quantity with the same units. This is called normalization. We can renormalize to the maximum value, resulting in a ratio between zero and one. We can renormalize to the minimum value, resulting in a scope. Statistical renormalization results from scaling to a statistic such as a sum, a mean, a range, or a standard deviation.

Normalization – General Definition

We can reduce a scaled quantity to a ratio with no units by normalizing each value. Normalization occurs when we divide a value by a reference value having the same units. The generic expression for normalization is

$$\left(\frac{Q}{Q_{ref}} \right)^{\beta}$$

The ratio has no units. The magnitude of a scaled quantity thus becomes independent of the units of measurement. In most applications $\beta = 1$. But we can use any value of β when rescaling quantities. Normalization where $\beta = 1$, allows us to substitute one measurement unit for another, as in Galileo's use of spearlengths to measure velocity. It is the basis of classical dimensional analysis.

A convenient reference quantity Q_{ref} is the largest observed or largest possible value, resulting in a reduced variable that can range from 0 to 1. An example is running speed measured relative to the maximum for that species. Yet another useful reference quantity is the minimum observed or possible value Q_{min}, which yields a scaled variable that ranges upward from one. An example is metabolic rate as a multiple of the standard metabolic rate SMR, which is measured at rest and in the absence of absorptive activity by the gut. Scaling relative to Q_{min} expresses the quantity Q in steps that are relevant to that variable. In physiology the scaled quantity Q_{max} / Q_{min} is called a scope. The definition of scope can be extended to any measured quantity

$$scope(Q) \equiv \frac{Q_{\max}}{Q_{\min}}$$

We can use scope to
compare the capacity of measurement instruments,
compare the information content of graphs,
compare variability of physical systems, or biological systems.

In addition to normalizing to a reference value, we can normalize the values of a quantity relative to a statistic with the same units such as a sum, such as the mean, the range, or the standard deviation. For the sake of clarity, we'll call this statistical normalization, to distinguish it from other forms of normalization. Normalization has several meanings, so we'll be specific about the normalization: normalizing to the mean, to the range, *etc.*

Normalization to the minimum --> scope (Schneider 2009 Chapter 11)
Physical quantities often have a large scope.

Quantity: mass of hydrogen atom H_2

Scope = mass(H_2)/mass(earth)

mass(H_2) = 1.0079 g·mol^{-1} · 1 mol·6.02^{-1}10^{-23} atoms · 2 atom· molecule^{-1}
= $3.3 \cdot 10^{-24}$ g

mass(earth) = $5.5 \cdot g \cdot cm^{-3} \cdot 4\pi\ 3^{-1}\ (12.756\ km/2)^3$
= $5.98 \cdot 10^{19}$ g = $5.98 \cdot 10^{7}$ Tg (teragrams)

Scope = $5.98 \cdot 3.3^{-1} \cdot 10^{31} = 1.8 \cdot 10^{31}$

Biological quantities often have a smaller scope than physical quantities.

Quantity: respiration rate

Scope is of the order of 10 (maximum is ca 10 times the minimum).

Quantity: body mass

Scope = 10^{21}

= ratio of mass of *Mycoplasma* (the smallest organism) to mass of Blue Whale.

Measurement instruments have a scope, defined as the maximum over the minimum reading.

Example: 1 kg / 1 microgram = 10^9 if we have a scale that will record masses to the nearest microgram, up to a maximum of 1 kg.

Scope of a metre-stick = 1m / 1cm = 100 (if marked in centimetres)
Scope of a metre-stick = 1m / 1mm = 1000 (if marked in millimetres)

A survey will also have a scope. Surveys are carried out by
-defining the sample unit,
-listing all possible units (the frame),
-then either sampling all possible units (complete census)
or sampling units at random.

The scope is the ratio of the frame size to the unit size.

For example a salmon survey might employ 100 km transects along river.

The unit is the 100 km transect, the frame is length of the river, and the scope is the number of possible transects along length of the river.

Extending the survey to all rivers in Labrador enlarges the scope. The unit is still the 100 km transect, the frame is now the sum of the length of each river. The scope increases to the number of possible transects along all rivers.

Normalization to minimum --> scope

Experiments also have a scope.

A physiological experiment carried out on 4 samples from the liver tissue from one mouse has a scope equal to the volume of the liver divided by the volume of each sample.

An experiment carried out on 4 tissue samples from each of 20 different mice has a greater scope: the ratio of the liver volume of 20 mice, divided by volume of each sample. This increase in volumetric scope tells us something about the generality of the result. The experiment with the greater scope is more convincing because carried out over a greater variety of tissue states, due to variation among mice.

If the experimental unit is a duration, then this is used in determining the scope.

For example an experiment on mortality of bacterial colonies in an agar plate, measured daily over 10 days, has a temporal scope of 10 days / 1 day = 10.

If the experiment is repeated 10 times, the temporal scope rises to

10 * 10 days / 1 day = 100.

This increase in scope again tells us something about the generality of the result, which applies at several times, not just at one point in time.

The scope of measured quantities is used in comparing survey designs and evaluating the limits on statistical inference from field and laboratory experiments.

See Chapter 11. The Scope of Quantities
Chapter 12. The Scope of Research Programs.

Schneider, D.C. 2009. *Quantitative Ecology: Measurement, Models, and Scaling.* San Diego: Academic Press.

Normalization relative to a statistic. In statistical analysis, we often renormalize relative to a descriptive statistic, such as the sum, the mean, the range, or the standard deviation.

Normalization to a sum

A familiar example of renormalizing is taking a percentage: adding up the parts to compute the whole, then taking each part as a ratio relative to the whole. For a percentage the reference quantity Q_{ref} is the sum of all the values of Q and the exponent is $\alpha = 1$, resulting in dimensionless values that can range from 0 to 1.

$$\% = \left(\frac{Q_i}{\sum_n^{i=1} Q_i} \right)^1$$

Normalization to a sum

Of particular interest in statistical analysis is the scaling of counts derived from nominal scale scoring of units. For example Mendel scored 929 pea plants as having either white flowers (224) or purple flowers (705). Thus 24% of the plants had white flowers.

$$\pi = \left(\frac{224}{929}\right) = 0.24$$

The same information is also expressed as an odds ratio. The odds of a flower being white are 0.318 to 1.

$$Odds = \frac{\pi}{1-\pi} = \left(\frac{224/929}{705/929}\right) = \left(\frac{224}{705}\right) = 0.318 \ : \ 1$$

Normalization to the mean of the quantity Q, where $mean(Q) = \frac{1}{n}\sum_{n}^{i=1} Q$

1. Wind speed vel = [……..]·m/sec
measured hourly at St. John's airport on ....(day with rise and fall over 8 hours.

2. Number of vascular plant species on 7 of the Canary Islands, in the eastern Atlantic. Data from K. Lems 1960 Floristic botany of the Canary Islands *Sarracenia* 5: 1-94.

Nplant = [366 348 763 1079 539 575 391] · sp/island

mean(*Nplant*) = n^{-1} ΣNplant
mean(*Nplant*) = 7^{-1} ·4061 · species/island = 580

dev(*Nplant*) = *Nplant* – mean(*Nplant*)

dev(*Nplant*) = [–214 –232 +182 +498 –41 –5 –189] · sp/island

kdev(*Nplant*) = [*Nplant* – mean(*Nplant*)] / mean(*Nplant*)

kdev(Nplant) = [–0.36 –0.56 +0.31 +0.89 –0.071–0.0086 –0.33] · sp/island

This particular normalization is often used in the physical sciences, notably meteorology and oceanography, where it is called the anomaly. An example is the annual temperature anomaly, the degree to which average temperature in the current year differs from a longer term average (warm year, cold year, *etc*).

Normalization to the mean

Normalization to the mean is central to genetics, population biology, and evolutionary biology, including the calculation of selection coefficients.

An example is incidence of pale-eyed (I = ge) or wild type (I = wild) flies in 8 progenies of the housefly *Musca domestica* (Sokal and Rohlf 2012 p 731). A new variable, the frequency of wild type flies f_{wild} is obtained by logical rescaling from nominal to ratio scale:

f_{wild} = [83 77 110 92 51 48 70 85] · flies. $\Sigma f_{wild} = 616$

This is normalized to the mean number of flies per progeny,

N_b = [130 120 206 150 82 109 112 151] · flies. $\Sigma N_b = 1060$

The normalized value is the proportion of pale eyed flies

$p_{wild} = f_{wild}/N_b$ = [0.64 0.64 0.53 0.61 0.62 0.44 0.63 0.56] ·wildtype/progeny

mean(p_{wild}) = $\Sigma f_{wild}/\Sigma N_b$ = 616/1060 = 0.58 wildtype/progeny

Proportions are usually normalized as a ratio, rather than as a difference.

p_{wild}/mean($p_{I=wild}$) = [1.10 1.10 0.92 1.06 1.07 0.76 1.08 0.97]

The proportion of wild type flies was highest in progenies 1 and 2, lowest in 6. Selection coefficients are calculated from these proportions.

Normalization to the mean – The Coefficient of Variation

The examples so far have been for each value of a variable. Normalization is also applied to measures of variability, resulting in a single ratio. The most familiar example is the coefficient of variation.

$$CV \equiv \frac{stdev(Q)}{mean(Q)}$$

Taking $\overline{Q}$ as the average value of Q he standard deviation is defined as:

$$stdev(Q) = \sqrt{\frac{1}{n}\sum_{n}^{i=1}\left(Q-\overline{Q}\right)^2}$$

The coefficient of variation is a unitless ratio that permits comparison of the variability of two scaled quantities, free of the effects of choice of measurement scale or the effects of size. For example, we can use the CV to compare morphological variability in mice and elephants.

Normalization to a range

The range is defined as the largest minus smallest value, usually in a sample. Normalization to the range (or scope) reduces the quantity to the range of 0 to 1.

$$Q' \equiv \frac{Q - Q_{\min}}{Q_{\max} - Q_{\min}}$$

This form of normalization is called ranging (Sneath and Sokal 1973) or minmax scaling. In data science is it called feature scaling, or more ambiguously "data normalization."

Normalization to the standard deviation

Normalization to a measure of dispersion, the standard deviation, is common in statistical applications. Taking $\overline{Q}$ again as the average value of Q, the standard deviation is defined as:

$$stdev(Q) = \sqrt{\frac{1}{n}\sum_{n}^{i=1}\left(Q - \overline{Q}\right)^2}$$

Returning to the number of plant species on 7 Canary Islands, we have:

dev(*Nplant*) = [–214 –232 +182 +498 –41 –5 –189] · sp/island

dev^2(*Nplant*) = [45857 53890 .....] · (sp/island)^2

mean squared deviation

msd(*Nplant*) = n^{-1} Σdev^2(*Nplant*) = 7^{-1} 419537 = 59934 (sp/island)^2

root mean squared deviation

rmsd(*Nplant*) = sqrt(msd) = sqrt(59934) = 245 sp/island

variance

var(*Nplant*) = (n–1)$^{-1}$ Σdev^2(*Nplant*) = 6^{-1} 419537 = 69923 (sp/island)^2

standard deviation

std(*Nplant*) = sqrt(var(*Nplant*)) = sqrt(69923) = 264 sp/island

standard deviates

nscore(*Nplant*) = (*Nplant* - mean(*Nplant*)) / std(*Nplant*) or nscores

nscore(Nplant) = [–214/264 –232/264 ...] = [–0.81 –0.88 ...]

$$Q' \equiv \frac{Q - mean(Q)}{stdev(Q)}$$

This form of normalization is called standardization. The results are called normal scores. These are dimensionless numbers because all of the terms in the formula have the same units, that of Q. Normal scores permit comparison of quantities that differ in magnitude and variability. Legendre and Legendre (1998) discuss applications and potential problems of this and other statistical reductions to dimensionless ratios.