





Describing Data: Visual Summaries

Categorical Data

Gerow–Mathew (2003) undertook a study of feather color with a sample of artificial feathers. The teacher for this project, Mrs. Cowper, colored 17 identical feather images, choosing three different colors, then gave the images to the researcher. He counted 2 yellow feathers, 9 red, and 6 brown, and produced a bar chart, reproduced in . Bar charts are usually used for counts of observations in categories. The values you use for categorical data can be alphabetic (green, red blue, and so on) or numerical (1, 2, 3, ...); in either event, when you call for a bar chart the different “things” in the column are taken as identifying the category a given observation falls into. The bar chart reports the number of observations in each category. Often, you will be able to choose the scale on the vertical axis to be absolute frequencies (actual counts) or relative frequencies (as proportions or percents).

Quantitative Data

For numerical data, the equivalent (to a barchart) figure is a histogram. Emily Morrison (OTS-SA-05) studied limpet morphology at a seaside site (De Hoop) in South Africa. One of the measurements she made on each of 393 limpets was length . A histogram and a barchart may look very similar, but are constructed using different rules. For a histogram, the range of values and number of values will determine how many “bins” to make. Attempting a bar chart for numerical data usually yields a mess, simply (Emily’s limpets, revisited) .

There is an exception to this. If you have numerical data, with contiguous values (a range of integers, for example) a barchart tool will treat these values as category labels and the resulting figure will be a perfectly useful histogram. If the range of values is small, the difference in the appearance of the two tools is slight. Erik Fyfe and Vernon Visser (OTS-SA-05) recorded the number of hooks (thorns protruding from knobs) on five randomly selected knobs on sections of acacia branches . Note that this trick won’t work if there is a gap in your list. For example, if you have values like 1, 1.5, 2, 2.5, and 4, the “4” category will get plotted right beside the “2.5”.

Quantitative data in groups

For comparing distributions of numerical data among samples, side-by-side box-plots are absolutely wonderful. They illustrate, in a considered glance, much information: minimum, maximum, 1st and 3rd quartiles, median, possible outliers, symmetry (or lack thereof). Dorit Hockman (OTS-SA-05) counted the number of grass species near (10m or less) and far (greater than 10m) from termite mounds, illustrated here



Describing Data: Numerical Summaries

The Sample Mean

The formula for the sample mean is one of the simplest in statistics. Let Y be the label for the random variable, and Y_i be the i th sample value, with i running from 1 to the sample size n . The formula for the sample mean of Y is

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

The sample mean estimates the mean of some population, but in this fisheries case, of what population? Ponder that a bit for this dataset,

then .

Standard Deviation

What is standard deviation (SD)? I'll give two answers; one technical, one utilitarian. It's possible to make use of SD (and use it well) without use of the technical definition, so if formulae aren't your cup of tea, don't worry about it.

Technical Definition

The sample SD is computed as the square root of the sample variance, the variance being (almost) the average of squared deviations from the mean of the sample. Let Y be the label for the random variable, and Y_i be the i th sample value, with i running from 1 to the sample size n . The formula for the sample SD of Y is

$$SD(Y) = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$



In words: subtract the mean (\bar{Y}) from each observation; square the result and add them all up. Divide by $n-1$. Notice that we divide by $n-1$, not n . That's why I said the variance was *almost* the average of squared

deviations. Technically, $n-1$ are the degrees of freedom associated with the variance (and hence SD) estimate from a single sample as we have here. Why do we divide by $n-1$ instead of n ? If we knew the population mean of Y , the formula would use the population mean μ_Y instead of the sample mean. It turns out that individual sample values are closer to their own sample mean than to the population mean, with the result that the sample SD tends to under-estimate the true population SD. Dividing by the (slightly smaller) $n-1$ perfectly corrects for this (thanks to P. Velleman and R. Deveaux: this explanation appears in their intro stats text). Well, not quite. It turns out that there is a bit more correcting needed; this gets accomplished by using the t distribution with $n-1$ degrees of freedom. The formula for that distribution (think of it as a fudged version of the Normal distribution) explicitly has $n-1$ built in; that *and* dividing by $n-1$ does the trick. If you'd like some (easy!) practice

calculating SD:  .

Utilitarian Explanation of SD

I'll start by asking *you* a question: "What is a pound (as a unit of weight)?" If I ask that question in a group, the resulting discussion usually leads to three observations:

1. Nobody in the group can provide a precise definition;

2. In the rare event that someone can do so, the definition doesn't really enhance our intuitive understanding of "pound"; and
3. Despite that we don't have any real understanding of the definition it, we all happily and correctly use pounds in everyday life, quite content in our ignorance.

Of course, when I say, "we," I refer only to one little corner of the world. The other 95% of the planet uses kilograms as the basic unit of weight, but their discussion of kilograms would likely parallel ours of pounds.

Taking that same approach, we can say that SD is a unit used to measure variation. More variation, bigger SD. No variation, zero SD. It happens that there is one feature of SD that lends itself to intuitive use. The typical range of data values in a sample is approximately four or five SDs. So if you knew the SD of some data was just under 5, you might expect range of data to be somewhere around 20 to 25. The converse is also true. If you have an idea of the range of typical data (that is, the range for most (like 95%) of the values), then you can divide that number by four or five to estimate the SD. These translations from SD to range are not precise, but they are reasonably close in many real-life instances. For larger samples, the range tends to cover a larger number of SDs (5

might work better); smaller samples, a smaller number (4 might be better).

Degrees of Freedom

Degrees of freedom are a number associated with estimation of a variance (or, upon taking square roots, the SD). The actual formula varies depending on the statistical context (mean for one sample, difference in means between two samples, and so on), but it always has the same construction: sample size minus the number of parameters that must be estimated to compute the variance formula. In the current setting, we need only the sample mean (as an estimate of the population mean, an unknown parameter), so the formula is $n-1$.

Data-free Estimation of the Standard Deviation

In order to determine sample size requirements (based on power considerations or on precision requirements) *before starting a study*, you need to have an estimate of the relevant variability (usually embodied by the standard deviation) in the data you have not yet collected. This sounds like an impassable barrier to continuing. Clearly, all you can do is use an educated guess. Even so, you would have to be very experienced

with your particular type of data to be able to divine the size of the standard deviation without having first collected the data. Fortunately, there is a way around that problem.

It is quite possible, even with only a moderate amount of experience, to estimate the *range* (largest minus smallest) of observations one is likely to encounter in a given situation. Assuming the estimated range covers most of the observations you are likely to encounter, a simple estimate of the standard deviation can be formed by dividing the range by three (more conservative) or four. We'll illustrate this in class.

References

Gerow-Mathew (2003). My son Eugene's first statistical summary, done in kindergarden.

OTS-SA-05 The named person was a student on the OTS South Africa Course (semester-long undergraduate ecology and cultural experience).

OTS is the Organization for Tropical Studies.