3.3: Measures of dispersion

Introduction

The statistics of data exploration involve calculating estimates of the middle, central tendency, and the variability, or dispersion, about the middle. Statistics about the middle were presented in the previous section, Chapter 3.2. Statistics about **measures of dispersion**, and how to calculate them in R, are presented in this page. Use of Z score to standardize or normalize scores is introduced. Statistical bias is also introduced.

Describing the middle of the data gives your reader a sense of what was the typical observation for that variable. Next, your reader will want to know something about the variation about the middle value — what was the smallest value observed? What was the largest value observed? Were data widely scattered or clumped about the middle?

Measures of dispersion or **variability** are descriptive statistics used to answer these kinds of questions. Variability statistics are very important and we will use them throughout the course. A key descriptive statement of your data, how variable?

> Note:
>
> **Data** is plural = a bunch of points or values; **datum** is the singular and rarely used.

Examine the two figures below (Fig. 3.3.1 and Fig. 3.3.2): the two **sample frequency distributions** (put data into several groups, from low to high, and membership is counted for each group) have similar central tendencies (mean), but they have different degrees of variability (standard deviation).

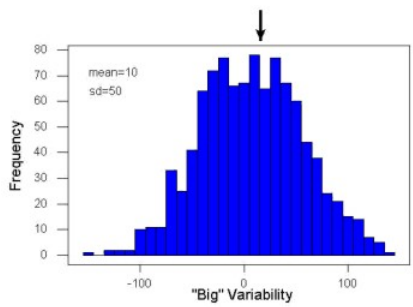

Figure 3.3.1: A histogram which displays a sampling of data with a mean of 10 (arrow marks the spot) and standard deviation (sd) of 50 units.

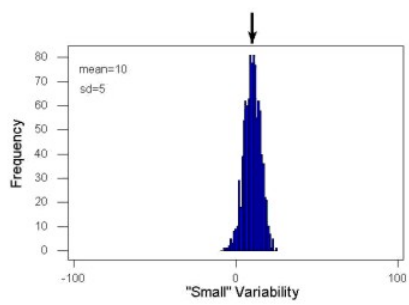

Figure 3.3.2: A histogram which displays a sampling of data with the same mean of 10 (arrow marks the spot) displayed in Fig. 3.3.1, but with a smaller standard deviation (sd) of 5 units.

Clearly, knowing something about the middle of a data set is only part of the required information we need when we explore a data set; we need measures of dispersion as well. Provided certain assumptions about the frequency of observations hold, estimates of the middle (e.g., mean, median) and the dispersion (e.g., standard deviation) are adequate to describe the properties of any observed data set.

For measures of dispersion or variability, the most common statistics are: Range, Mean Deviation, Variance, Standard Deviation, and Coefficient of Variation.

The range

The range is reported as one number, the difference between maximum and minimum values.

$$range = X_{max} - X_{min}$$

Arrange the data from low to high, subtract the minimum, X_{min}, from the maximum, X_{max}, value and that's the range.

For example, the minimum height of a collection of fern trees might be 1 meter whereas the maximum height might be 2.2 meters. Therefore, the range is 1.2 meters (= 2.2 – 1).

The range is useful, but may be misleading (as index of true variability in population) if there is one or more exceptional "outlier" (one individual that has an exceptionally large or small value). I often just report X_{min} and X_{max}. That's the way R does it, too.

If you recall, we used the following as an example data set.

```
x = c(4,4,4,4,5,5,6,6,6,7,7,8,8,8,15)
```

The command for range in R is simply `range()` .

```
range(x)
[1] 4 15
```

There is apparently no universally accepted symbol for "range," so we report the statistic as $range = 15 - 4 = 11$.

You may run into an **interquartile range**, which is abbreviated as **IQR**. It is a trimmed range, which means, like the trimmed mean, it will be robust to outlier observations. To calculate this statistic, divide the data into fourths, termed **quartiles**. For our example variable, x, we can get the quartiles in R:

```
quantile(x)
  0%  25%  50%  75% 100%
 4.0  4.5  6.0  7.5 15.0
```

Thus, we see that 25% of the observations are 4.5 (or less), the second quartile is the median, and 75% of the observations are less than 7.5. The IQR is the difference between the first quartile, Q1, and the third quartile, Q3.

The command for obtaining the IQR in R is simply `IQR()` . Yes, you have to capitalize "IQR" — R is case sensitive; that means IqR is not the same as IQR or iqr.

```
IQR(x)
[1] 3.5
```

and we report the statistic as *IQR* = 3.5. Thus, 75% of the observations are within 3.5 points of the median. The IQR is used in box plots (see Chapter 4). Quartiles are special cases of the more general **quantile**; quantiles divide up the range of values into groups with equal probabilities. For quartiles, four groups at 25% intervals. Other common quantiles include **deciles** (ten equal groups, 10% intervals) and percentiles (100 groups, 1% intervals).

The mean deviation

$$sample\ mean\ deviation = \frac{\sum_{i=1}^{n} \left| X_i - \bar{X} \right|}{n}$$

Subtract each observation from the sample mean; each $X_i - \bar{X}$ is called a ***deviate***: some observations will be positive (greater than $\bar{X}$) and some will be negative (less than $\bar{X}$).

Take the absolute value of the deviation and then add up the absolute values of the deviations from the mean. At the end, divide by the sample size, n. Large values for this statistic imply that much of the data is spread out, far from the mean. Small values in turn imply that each observation is close to the mean.

Note:

The mean deviation will always be positive, which is why we take the absolute. By taking the absolute value of each deviate, then the sum is greater than zero. Now, we rarely use this statistic by itself — but that difference, $X_i - \bar{X}$, is integral to much of the statistical tests we will use. Look for this difference in other equations!

In R, we can get the mean deviation with the `mad()` function. At the R prompt, then

```
mad(x)
[1] 2.9652
```

Population variance and population standard deviation

The population variance is appropriate for describing variability about the middle for a **census**. Again, a census implies every member of the population was measured. The equation of the population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{N}$$

The population standard deviation also describes the variability about the middle, but has the advantage of being in the same units as the quantity (i.e., no "squared" term). The equation of the population standard deviation is

$$\sigma = \sqrt{\sigma^2}$$

Sample variance and sample standard deviation

The above statements about the population variance and standard deviation hold for the sample statistics. The equation of the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)}{n-1}$$

and the equation of the sample standard deviation is

$$s = \sqrt{s^2}$$

Of course, instead of taking the square-root of the sample variance, the sample standard deviation could be calculated directly.

$$s = \sqrt{\frac{X_i - \bar{X}}{n-1}}$$

Note:

See the difference between calculation of the population parameter and the sample statistic estimates? The difference between the formulas for population and sample variances — We divide by `n - 1` instead of `N` . This is **Bessel's correction** and we will take a few moments here and in class to show you why this correction makes sense. Bessel's correction to the sample variance illustrates a basic issue in statistics: when estimating something, we want the estimator (i.e., the equation), to be an unbiased value for the population parameter it is intended to estimate.

Statistical bias is an important concept in its own right; bias is a problem because it refers to a situation in which an estimator consistently returns a value different from the population parameter for which it is intended to estimate. Thus, the sample mean is said to be an **unbiased estimator** of the population mean. Here, bias means that the formula will give you a good estimate of the value. This turns out not to be the case for the sample variance if you divide by `n` instead of `n - 1` . Now, for very large values of `n` , this is not much of an issue, but it shows up when `n` is small. In R you get what you ask for — if you ask for the sample standard deviation, the software will return the correct value; calculators, go to watch out for this — not all of them are good at communicating which standard deviation they are calculating, the population or the sample standard deviation.

Winsorized variances

Like the trimmed mean and winsorized mean (Chapter 3.2), we may need to construct a robust estimate of variability less sensitive to extreme **outliers**. Winsorized refers to procedures to replace extreme values in the sample with a smaller value. As noted in Chapter 3.2, we chose the level ahead of time, e.g., 90%. Winorized values then The **winsorized variance** is just the sample variance of the winsorized values. In R, we use `winvar()` from the WRS2 package.

Making use of the sample standard deviation

Given an estimate of the mean and an estimate of the standard deviation, one can quickly determine the kinds of observations made and how frequently they are expected to be found in a sample from a population (assuming a particular population distribution). For example, it is common in journal articles for authors to provide a table of summary statistics like the mean and standard deviation to describe characteristics of the study population (aka the reference population), or samples of subjects drawn from the study population (aka the sample population). The CDC provides reports of attributes of for a sample of adults (more than forty thousand) from the USA population (Fryar et al 2018). Table $3.3.1$ shows a sample of results for height, weight, and waist circumference for men aged 20 – 39 years.

Table $3.3.1$. Summary statistics mean ($\pm$ standard deviation) of height, weight, and waist circumference of 20-39 year old men, USA.

Years	Height, inches	Weight, pounds	Waist Circumference, inches
1999 – 2000	69.4 (0.1)	185.8 (2.0)	37.1 (0.3)
2007 – 2008	69.4 (0.2)	189.9 (2.1)	37.6 (0.3)
2015 – 2016	69.3 (0.1)	196.9 (3.1)	38.7 (0.4)

We introduced the normal deviate as a way to normalize scores, and which we we will use extensively in our discussion of the normal distribution in Chapter 6.7, as a way to standardize a sample distribution, assuming a normal curve.

$$Z = \frac{X_i - \mu}{\sigma}$$

For data that are normally distributed, the standard deviation can be used to tell us a lot about the variability of the data:

62.26% of the data will lie between $\pm 1 \cdot \sigma$ of $\bar{X}$

95.46% of the data will lie between $\pm 2 \cdot \sigma$ of $\bar{X}$

99.0% of the data will lie between $\pm 3 \cdot \sigma$ of $\bar{X}$

This is known as the **empirical rule**, where 68% of the observations will be within one standard deviation, 95% of observations will be within two standard deviations, and 99% of observations will be within three standard deviations.

For example, men aged 20 years in the USA are on average μ = 5 feet 11 inches tall, with a standard deviation of σ = 3 inches. Consider a sample of 1000 men from this population. Assuming a normal distribution, we predict that 623 (62.26%) will be between 5 ft. 8 in. and 6 ft. 2 in., or $\pm 1 \cdot \sigma$.

Where did the 5 ft. 8 in. and the 6 ft. 2 in. come from? We add or subtract multiples of standard deviations. Thus, 6 ft. 2 in. = 5 ft. 11 in. $+1 \cdot \sigma$ (replace σ with 3 in.) and 5 ft. 8 in. = 5 ft. 11 in. $-1 \cdot \sigma$ (again, replace σ with 3 in.).

We can generalize to any distribution. **Chebyshev's inequality** (or theorem) guarantees that no more than a particular fraction $1/k^2$ of observations can be a specified k standard deviations distance away from the mean (k needs to be greater than 1). Thus, for $k = 2$ standard deviations, we expect a minimum of 75% of values $\left(1 - 1/2^2\right)$ within two standard deviations away from the mean, or for $k = 3$, then 89% of values $\left(1 - 1/3^2\right)$ will be within three standard deviations from the mean.

Hopefully you are now getting a sense how this knowledge allows you to plan an experiment.

For example, for a sample of 1000 observations of height of men, how many do we expect to be greater than 6 ft. 7 in. tall? Apply the empirical rule and do a little math — 79 inches (6 ft. 7 in) minus 71 inches (the mean) is equal to 8. Divide 8 by 3 (our value of σ) and you'll get 2.66667; so, 6 ft. 7 in. tall is about 2.67 standard deviations greater than the mean. From Chebyshev's inequality we have $1/2.67^2 = 0.140$, or 14% of observations less than or greater than the mean $(\pm\mu)$. Our question asks only about expected number of observations greater than $+2.67 \cdot \sigma$; divide 14% in half — we therefore expect about 70 individuals out of 1000 men

sampled to be 6 ft. 7 in. or taller. Note that we made no assumption about the shape of the distribution. If we assume the sample of observations comes from the normal distribution, then we can apply the Z score: about 4 individuals out of 1000 are expected to be $+2.67 \cdot \sigma$ from the mean (see R code in Note box). We extend the Z-score work in Chapter 6.7.

Note:

R code:

```
round(1000*(pnorm(c(79), mean=71, sd=3,lower.tail=FALSE)),0)
```

or alternatively

```
round(1000*(pnorm(c(2.67), mean=0, sd=1,lower.tail=FALSE)),0)
```

R output

```
[1] 4
```

R Commander command

Rcmdr: **Distributions** → **Continuous distributions** → **Normal distribution** → **Normal probabilities …**, select Upper tail.

For another example, this time we will use a data set available in R: the diabetic retinopathy data set in the survival package. It contains ages of the 394 subjects ranged from 1 to 58. Mean was 20.8 ± 14.81 years. How many out of 100 subjects do we expect to be greater than 50 years old? With Chebyshev's inequality we have $1/1.97^2 = 0.257$, or 25.7% of observations less than or greater than the mean, so about 13. If we assume normality, then the Z score (Upper tail) is 28.5% and we expect 28 subjects older than 50. Checking the data set, only eight subjects were older than 50. Our estimate by Chebyshev's inequality was closer to the truth. Why? Take a look at the histogram of the ages.

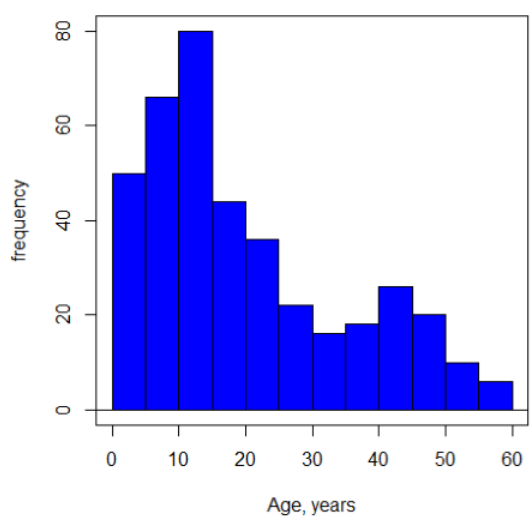

Figure 3.3.3: Histogram of subject ages from diabetic retinopathy data set in the R survival package.

Doesn't look like a normal distribution, does it?.

Z-score or Chebyshev's inequality, which to use? Chebyshev's inequality is more general — it can be used whether the variable is discrete or continuous, and without assumption about the distribution. In contrast, the Z score assumes more is known about the variable: random, continuous, drawn from a normally distributed population. Thus, as long as these assumptions hold, the Z score approach will give a better answer. This makes intuitive sense — if we know more, our predictions should be better.

In summary from the above points, and perhaps more formally for our purposes, the standard deviation is a good statistic to describe variability of observations on subjects, it's integral to the concept of precision of an estimate and is part of any equation

for calculating confidence intervals (CI). For any estimated statistic, a good rule of thumb is to always include a confidence interval calculation. We introduced these intervals in our discussion of risk analysis (approximate CI of NNT), and we will return to confidence intervals more formally when we introduce *t*-tests.

When you hear people talk about "**margin of error**" in a survey, typically they are referring to the standard deviation — to be more precise, they are referring to a calculation that includes the standard deviation, the standard error and an accounting for confidence in the estimate (see also Chapter 3.5 – Statistics of error).

Corrected Sums of Squares

Now, take a closer look at the sample variance formula. We see a *deviate*:

$$X_i - \bar{X}$$

The variance is the average of the squared deviations from the mean. Because of the "squared" part, this statistic will always be positive and greater (or typically, equal to) zero. The variance has squared units (e.g., if the units were grams, then the variance units are grams^2). The sample standard deviation has the same units as the quantity (i.e., no "squared" term). The numerator is called a **sums of squares**, and will be abbreviated as SS. Much like the deviate, it will show up frequently as we move forward in statistics (e.g., it's key to understanding ANOVA).

Other standard deviations of means

Just as we discussed for the arithmetic average, there will be corresponding standard deviations for the other kinds of means. With the geometric mean, one would calculate the **standard deviation of the geometric mean**, s_{gm}, as

$$s_{gm} = \pm exp\sqrt{\frac{\sum_{i=1}^{n}\left(\ln\frac{X_i}{\bar{X}_{gm}}\right)^2}{n-1}}$$

where exp is the exponential function, ln refers to natural logarithm, and $\bar{X}_{gm}$ refers to the sample geometric mean.

For the sample harmonic mean, it turns out there isn't a straightforward formula, only an approximation (which I will spare you — it involves use of expectations and moments).

Base R, and therefore Rcmdr, doesn't have built in functions for these, although you could download and install some R packages which do (e.g., package `NCStats`, and the function is `geosd()`).

If we run into data types appropriate for the geometric or harmonic means and standard deviations I will point these out; for now, I present these for completeness only.

Coefficient of variation (CV)

An unfortunate property of the standard deviation is that it is related (= "correlated") to the mean. Thus, if the mean of a sample of 100 individuals is 5 and variability is about 20%, then the standard deviation is about 1; compare this to a sample of 100 individuals with mean = 25 and 20% variability, where the standard deviation is about 5. For means ranging from 1 to 100, here's a plot to show you what I am talking about.

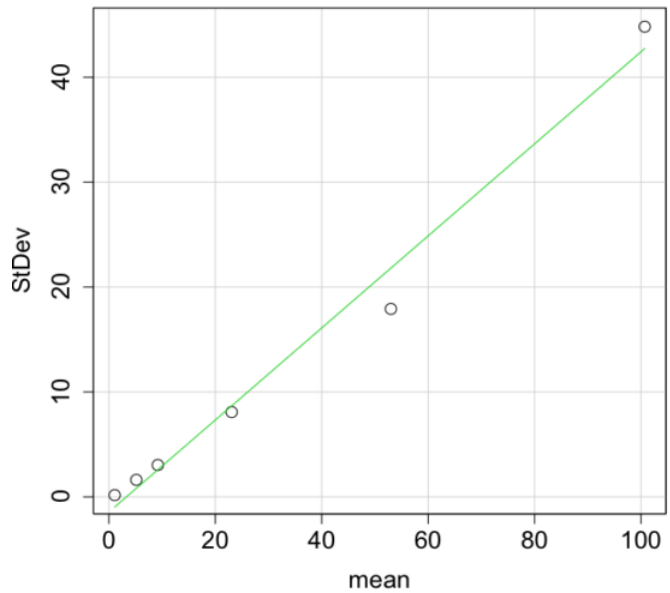

Figure 3.3.4: Scatter plot of the standard deviation (StDev) by the mean. Data sets were simulated.

The data presented in the Fig. 3.3.4 graph were simulated. Is this bias, a correlation between mean and standard deviation, something you'd see in "real" data? Here's a plot of height in inches at withers for dog breeds (Fig. 3.3.5). A line is drawn (ordinary linear regression, see Chapter 12) and as you can see, as the mean increases, the variability as indicated by the standard deviation also increases.

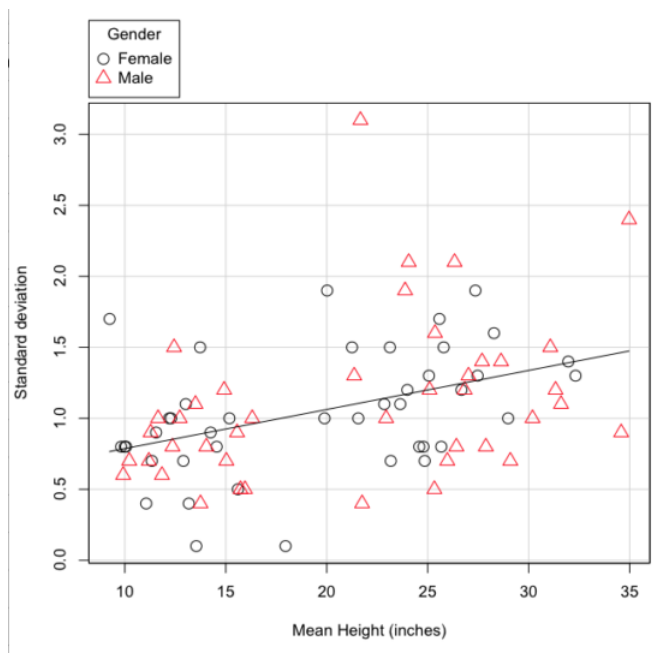

Figure 3.3.5: Plot of the standard deviation by the mean for heights of different breeds of dogs.

So, to compare variability estimates among groups when the means also vary, we need a new statistic, the coefficient of variation, which is abbreviated as CV. Many statistics textbooks not aimed at biologists do not provide this estimator (e.g., check your statistics book).

$$CV = \frac{s}{\bar{X}} \cdot 100$$

This is simply the standard deviation of the sample divided by the sample mean, then multiplied by 100 to give a percent. The CV is useful when comparing the distributions of different types of data with different means.Many times, the standard deviation of a distribution will be associated with the mean of the data.

Example. The standard deviation in height of elephants will be larger on the centimeter scale than the standard deviation of the height of mice. However, the amount of variability RELATIVE to the mean may be similar between elephants and mice. The CV might indicate that the relative variability of the two organisms is the same.

The standard deviation will also be influenced by the scale of measurement. If you measure on the millimeter scale versus the meter scale, the magnitude of the SD will change. However, the CV will be the same!

By dividing the standard deviation by the mean, you remove the measurement scale dependence of the standard deviation and generally, you also remove the relationship of the standard deviation with the mean. Therefore, CV is useful when comparing the variability of the data from distributions with different means.

One disadvantage is that the CV is only useful for ratio scale data (i.e., those with a true zero).

The coefficient of variation is also one of the statistics useful for describing the precision of a measurement. See Chapter 3.4: Estimating parameters.

Standard error of the mean (SEM)

All estimates should be accompanied by a statistic that describes the accuracy of the measure. Combined with the confidence interval (Chapter 3.4), one such statistic is called the standard error of the mean, or SEM for short. For the error associated with calculation of our sample mean, it is defined as the sample standard deviation divided by the square root of n, the sample size.

$$sem \rightarrow s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

The concept of error for estimates is a crucial concept. All estimates are made with error and for many statistics, a calculation is available to estimate the error (see Chapter 3.4).

Although related to each other, the concepts of sample standard deviation and sample standard error have distinct interpretations in statistics. The standard deviation quantifies the amount of variation of observations from the mean, while the standard error quantifies the difference between the sample mean and the population mean. The standard error will always be smaller than the standard deviation and is best left for reporting accuracy of a measure and statistical inference rather than description.

Questions

1. For a sample data set like `y = c(1,1,3,6)` , you should now be able to calculate, by hand, the
   - range
   - mean
   - median
   - mode
   - standard deviation
2. If the difference between Q1 and Q3 is called the interquartile range, what do we call Q2?
3. For our example data set, `x <- c(4,4,4,4,5,6,6,6,7,7,8,8,8,8,8)` calculate
   • IQR
   • sample standard deviation, *s*
   • coefficient of variation
4. Use the `sample()` command in R to draw samples of size 4, 8, and 12 from your example data set stored in `x` . Repeat the calculations from question 3. For example `x4 <- sample(x,4)` will randomly select four observations from *x*, and will store it in the object *x4*, like so (your numbers probably will differ!) `x4 <- sample(x,4); x4 [1] 8 6 8 6`
5. Repeat the exercise in question 4 again using different samples of 4, 8, and 12. For example, when I repeat sample(`x` ,4) a second time I get `sample(x,4) [1] 8 4 8 6`
6. For Table 1, determine how many multiples of the standard deviation for observations greater than 95-percentile (e.g., determine the observation value for a person who is in the 95-percentile for Height in the different decades, etc.
7. Calculate the sample range, IQR, sample standard deviation, and coefficient of variation for the following data sets
   - Basal 5 hour fasting plasma glucose-to-insulin ratio of four inbred strains of mice,
     `x <- c(44, 100, 105, 107) #(data from Berglund et al 2008)`

- Height in inches of mothers,
  ```
  mom <- c(67, 66.5, 64, 58.5, 68, 66.5) #(data from GaltonFamilies in R package HistData)
  ```
  and fathers,
  ```
  dad <- c(78.5, 75.5, 75, 75, 74, 74) #(data from GaltonFamilies in R package HistData)
  ```
- Carbon dioxide (CO_2) readings from Mauna Loa for the month of December for years <-c (1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010, 2015, 2020)
  ```
  years <-c (1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010, 2015, 2020) #obviously, do not calculate statistics on years; you can use to make a plot
  co2 <- c(316.19, 319.42, 325.13, 330.62, 338.29, 346.12, 354.41, 360.82, 396.83, 380.31, 389.99, 402.06, 414.26) #(data from Dr. Pieter Tans, NOAA/GML (gml.noaa.gov/ccgg/trends/) and Dr. Ralph Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/))
  ```
- Body mass of *Rhinella marina* (formerly *Bufo marinus*)
  ```
  bufo <- c(71.3, 71.4, 74.1, 85.4, 85.4, 86.6, 97.4, 99.6, 107, 115.7, 135.7, 156.2)
  ```

This page titled 3.3: Measures of dispersion is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michael R Dohm via source content that was edited to the style and standards of the LibreTexts platform.