Quantitative Understanding in Biology
Module I: Statistics
Lecture I: Characterizing a Distribution

Mean and Standard Distribution

Biological investigation often involves taking measurements from a sample of a population.

The mean of these measurements is, of course, the most common way to characterize their distribution:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The concept is easy to understand and should be familiar to everyone. However, be careful when implementing it on a computer. In particular, make sure you know how the program you are using deals with missing values:

`> x <- rnorm(10)`
`> x`
` [1] -0.05102204  0.38152698  0.66149378`
` [4]  0.41893786 -1.01743583 -0.55409120`
` [7] -0.14993880 -0.31772140 -0.44995050`
`[10] -0.69896096`
	Generates 10 random samples from a normal distribution.

`> mean(x)` `[1] -0.1777162`	Computes the mean

`> x[3] <- NA`
`> x`
` [1] -0.05102204  0.38152698          NA`
` [4]  0.41893786 -1.01743583 -0.55409120`
` [7] -0.14993880 -0.31772140 -0.44995050`
`[10] -0.69896096`
	Indicate that one of the values is unknown

`> mean(x)`
`[1] NA`
`> mean(x, na.rm=TRUE)`
`[1] -0.2709618`
	The mean cannot be computed, unless you ask that missing values be ignored.

`> sum(x)`
`[1] NA`
`> sum(x, na.rm=TRUE)`
`[1] -2.438656`
`> length(x)`
`[1] 10`
`> length(na.omit(x))`
`[1] 9`
`> sum(x, na.rm=TRUE)/length(na.omit(x))`
`[1] -0.2709618`
	Computing the mean 'manually' requires careful attention to NAs.

Similar principles hold when using Microsoft Excel. Try using the AVERAGE( ) and SUM( ) functions. What is the difference in behavior when you leave a cell empty vs. when you use the NA( ) function.

In addition to the mean, the standard deviation and (to a lesser extent) the variance are also commonly used to describe a distribution of values:

$$\begin{pmatrix} Sample \\ Variance \end{pmatrix} = s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

$$\begin{pmatrix} Sample \\ Standard \\ Deviation \end{pmatrix} = s = \sqrt{Sample\ Variance} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Observe that the variance is an average of the square of the distance from the mean. All terms in the summation are positive because they are squared.

When computing the variance or standard deviation (SD) of a whole population, the denominator would be N instead of N-1. The variance of a sample from a population is always a little bit larger, because the denominator is a little bit smaller. There are theoretical reasons for this having to do with degrees of freedom; we will chalk it up to a "weird statistics thing".

Observe that the standard deviation has the same units of measure as the values in the sample and of the mean. It gives us a measure of how spread out our data is, in units that are natural to reason with.

In the physical sciences (physics, chemistry, etc.), the primary source of variation in collected data is often due to "measurement error": sample preparation, instrumentation, etc. This implies that if you are more careful in performing your experiments and you have better instrumentation, you can drive the variation in your data towards zero. Think about measuring the boiling point of pure water as an example. Some argue that if you need complex statistical analysis to interpret the results of such an experiment, you've performed the experiment badly, or you've done the wrong experiment.

Although one might imagine that an experimenter would always use the best possible measurement technology available (or affordable), this is not always the case. When developing protocols for CT scans, one must consider that the measurement process can have deleterious effects on the patient due to the radiation dose required to carry out the scan. While more precise imaging, and thus measurements (say of a tumor size), can often be achieved by increasing the radiation dose, scans are selected to provide just enough resolution to make the medical diagnosis in question. In this case, better statistics means less radiation, and improved patient care.

In biology, the primary source of variation is often "biological diversity". Cells, and in particular, patients, are rarely in identical states, and you expect a non-trivial variation, even under perfect experimental conditions. In biology, we must learn to cope with this naturally occurring variation.

Communicating a Distribution

$\bar{x}$ and SD have a particular meaning when the distribution is normal. For the moment, we'll not assume anything about normality, and consider how to represent a distribution of values.

Characterizing a Distribution

Histograms convey information about a distribution graphically. They are easy to understand, but can be problematic because binning is arbitrary. There are essentially two arbitrary parameters that you select when you prepare a histogram: the width of the bins, and the alignment, or starting location, of the bins. For non-large N, the perceptions suggested by a histogram can be misleading.

> set.seed(0) > x <- rnorm(50) > hist(x, breaks=seq(-3,3,length.out=6)) > hist(x, breaks=seq(-3,3,length.out=7)) > hist(x, breaks=seq(-3,3,length.out=12))	Three histograms are prepared; the same data are presented, but, depending on the binning, a different underlying distribution is suggested.

When preparing histograms, be sure that the labels on the x-axis are chosen so that the binning intervals can be easily inferred. The first plot would better be prepared including one additional option: `xaxp = c(-3,3,5)`. See the entry for `par` in the R help for this any many other plotting options; type `?par` at the R prompt.

R has a less arbitrary function, `density`, which can be useful for getting the feel for the shape of an underlying distribution. This function does have one somewhat arbitrary parameter (the bandwidth); it is fairly robust and default usually works reasonably well.

```
> hist(x, breaks=seq(-3,3,length.out=13), xaxp=c(-3,3,4),
probability=TRUE); lines(density(x))
```

Note that we add the `probability` option to the `hist` function; this plots a normalized histogram, which is convenient, as this is the scale needed by the overlayed density function.

You should be wary of using summary statistics such as $\bar{x}$ and SD for samples that don't have large N or that are not known to be normally distributed. For N=50, as above, other options include:

- A table of all the values: `sort(x)`
- A more condensed version of the above: `stem(x)`

For graphical presentations, do not underestimate the power of showing all of your data. With judicious plotting choices, you can often accomplish this for N in the thousands.

`stripchart(x)` shows all data points. For N = 50, `stripchart(x, pch="|")` might be more appropriate.

If you must prepare a histogram (it is often expected), overlaying the density curve and sneaking in a stripchart-like display can be a significant enhancement:

```
> hist(x, breaks=seq(-3,3,length.out=13), xaxp=c(-3,3,4),
probability=TRUE); lines(density(x))

> rug(x)
```

For larger N, a boxplot can be appropriate:


```
> x <- rnorm(1000); boxplot(x)
```

You can overlay (using the `add=TRUE` option) a stripchart to show all data points. With many data points, a smaller plotting symbol and the `jitter` option are helpful.

```
 > stripchart(x, vertical=TRUE, pch=".", method="jitter", add=TRUE)
```

Note that boxplots show quartiles. The heavy bar in the middle is the median, not the mean. The box above the median is the third quartile; 25% of the data falls in it. Similarly, the box below the median holds the second quartile. The whiskers are chosen such that, if the underlying distribution is normal, roughly 1 in 100 data points will fall outside their range. These are putative outliers that you may want to further inspect.

The concept of quartiles can be generalized to quantiles. Another way to characterize distributions is by reporting quantiles; quartiles and deciles are favorites:

```
> quantile(x, (0:4)/4)
         0%         25%         50%         75%        100%
-2.99767066 -0.69364940 -0.01546943  0.65434645  3.02193840

> quantile(x, (0:10)/10)
         0%         10%         20%         30%         40%
-2.99767066 -1.20812215 -0.87560155 -0.53779019 -0.26516716
        50%         60%         70%         80%         90%
-0.01546943  0.22308820  0.48496338  0.78565873  1.18193333
       100%
 3.02193840
```

SD is a representation of how spread out your data are. If the underlying distribution is normal and N is large, then 95% of the samples are expected to fall within the range: $\bar{x} \pm 1.96 \cdot SD$.

```
> x <- rnorm(100000)
> mean(x)
[1] 0.001076443
> sd(x)
[1] 1.000764
> quantile(x, (0:40)/40)
          0%         2.5%           5%         7.5%
-4.754242304 -1.964334170 -1.650846246 -1.442248402
         10%        12.5%          15%        17.5%
-1.280851350 -1.147004677 -1.035610266 -0.935114705
         20%        22.5%          25%        27.5%
-0.841499568 -0.754756972 -0.679036632 -0.600832587
         30%        32.5%          35%        37.5%
-0.526768394 -0.455779370 -0.385446675 -0.317184080
         40%        42.5%          45%        47.5%
-0.252345155 -0.187829088 -0.123271243 -0.058831369
         50%        52.5%          55%        57.5%
 0.003971025  0.066956941  0.129356108  0.192253012
         60%        62.5%          65%        67.5%
```



```
 0.257026661  0.321183502  0.388136537  0.458046456
         70%          72.5%          75%          77.5%
 0.528821444  0.601289931  0.677759750  0.758717314
         80%          82.5%          85%          87.5%
 0.842717888  0.933548945  1.035464420  1.145487565
         90%          92.5%          95%          97.5%
 1.277188560  1.435926218  1.637997636  1.964227885
        100%
 4.336132109
```

We expect the mean to be zero, the SD to be unity, the 2.5% quantile to be at -1.96, and the 97.5% quantile to be at +1.96.

Standard Deviation vs. Standard Error of the Mean

An important, but very different, question that statistics can help us with is how well we can estimate the mean. Two factors influence this: how spread out the data are, and how much data we have. A new quantity, the Standard Error of the Mean, is introduced:

$$SEM = \frac{SD}{\sqrt{n}}$$

For large N, we can be 95% sure that the true mean of the underlying population is in the range…

$$\bar{x} \pm 1.96 \cdot SEM$$

…where $\bar{x}$ is the sample mean. We will formalize and extend this result in another session.

Here is an experiment to demonstrate this. We generate a sample from a known normal distribution where the mean is zero and the standard deviation is unity, then compute a confidence interval (CI) for the mean. We expect that this CI will contain the true mean (which we know to be zero) roughly 19 out of 20 times.

```
> for (i in 1:100) {
+   x <- rnorm(10000)
+   print(mean(x) + c(-1.96, 0, 1.96) * sd(x) / sqrt(length(x)))
+ }
```

```
[1] -0.024301502 -0.004775178  0.014751146
[1] -0.026626053 -0.006999663  0.012626728
[1] -0.021006574 -0.001665145  0.017676283
[1] -0.023918612 -0.004202195  0.015514221
[1] -0.0389625436 -0.0193659724  0.0002305987
[1] -0.035444374 -0.015853783  0.003736808
[1] -0.02646289 -0.00695293  0.01255703
[1] -0.014265428  0.005169996  0.024605420
[1] -0.006374521  0.013183087  0.032740695
[1] -0.027726532 -0.008195687  0.011335157
[1] -0.028349554 -0.008883653  0.010582248
[1] -0.031477297 -0.011958221  0.007560854
[1] -0.024676765 -0.005038202  0.014600361
[1] 0.001781225 0.021552050 0.041322876
[1] -0.027024247 -0.007516368  0.011991510
[1] -0.011989703  0.007447232  0.026884167
[1] -0.015610630  0.004039557  0.023689743
[1] -0.013781684  0.005884203  0.025550091
```

```
[1] -0.026578015 -0.006881238  0.012815538
[1] -0.0205013516 -0.0007041864  0.0190929788
[1] -0.016042673  0.003399703  0.022842080
[1] -0.002855254  0.016817403  0.036490060
[1] -0.034024678 -0.014692288  0.004640103
[1] -0.007221995  0.012408992  0.032039980
[1] -0.009129527  0.010746901  0.030623328
[1] -0.013079007  0.006672627  0.026424261
[1] -0.022694849 -0.003307884  0.016079082
[1] -0.027884371 -0.008532565  0.010819241
[1] 0.00974305 0.02926519 0.04878734
[1] -0.027234717 -0.007591476  0.012051764
[1] -0.003150395  0.016241695  0.035633785
[1] -0.011179020  0.008338268  0.027855557
[1] -0.009063754  0.010312788  0.029689330
[1] -0.016275525  0.003311566  0.022898658
[1] -0.041542741 -0.021898218 -0.002253696
[1] -0.0004399937  0.0190984618  0.0386369172
```



```
[1] -0.0395574949 -0.0198794721 -0.0002014492
[1] -0.030405467 -0.010958771  0.008487925
[1] -0.026741095 -0.007219373  0.012302349
[1] -0.0195650260  0.0001406561  0.0198463383
[1] -0.010852923  0.008932587  0.028718097
[1] -0.021023184 -0.001362284  0.018298616
[1] -0.032128012 -0.012800292  0.006527427
[1] -0.014295601  0.005436455  0.025168511
[1] -0.010071603  0.009548244  0.029168092
[1] -0.023738067 -0.004131066  0.015475936
[1] -0.009742657  0.009975540  0.029693737
[1] 0.008632623 0.028290878 0.047949132
[1] -0.017761529  0.001915428  0.021592384
[1] -0.01060836  0.00924938  0.02910712
[1] -0.0201786752 -0.0005449171  0.0190888409
[1] -0.022384361 -0.002820761  0.016742839
[1] -0.021063830 -0.001345081  0.018373668
[1] -0.016937999  0.002667279  0.022272557
[1] -0.018183755  0.001535473  0.021254700
[1] -0.036550951 -0.016922209  0.002706532
[1] -0.022773824 -0.003336013  0.016101799
[1] -0.015164353  0.004381009  0.023926372
[1] -0.012301662  0.007435353  0.027172368
[1] -0.017349224  0.002106028  0.021561280
[1] -0.016890972  0.002804223  0.022499418
[1] -0.013442029  0.006092335  0.025626700
[1] -0.014835716  0.004868907  0.024573530
[1] -0.008241882  0.011418128  0.031078138
[1] -0.02932350 -0.00965467  0.01001416
[1] -0.024917504 -0.005348383  0.014220739
[1] -0.030790123 -0.011296159  0.008197805
[1] -0.036026598 -0.016414003  0.003198591
[1] -0.034872802 -0.015233129  0.004406544
[1] -0.018245860  0.001475052  0.021195964
[1] -0.023516307 -0.003742883  0.016030542
[1] -0.024021403 -0.004490213  0.015040977
[1] -0.010607586  0.009004412  0.028616410
[1] -0.009601862  0.009955746  0.029513354
[1] -0.020615433 -0.001143053  0.018329327
[1] -0.016035976  0.003725782  0.023487540
[1] -0.0203469118 -0.0006296587  0.0190875944
[1] -0.03135966 -0.01174369  0.00787228
[1] -0.016250733  0.003365887  0.022982508
[1] -0.041422524 -0.021735918 -0.002049312
[1] -0.027268162 -0.007798383  0.011671397
[1] -0.029318299 -0.010000627  0.009317046
[1] -0.024506008 -0.004866441  0.014773125
[1] -0.023804302 -0.004134864  0.015534575
[1] -0.017442125  0.002055394  0.021552913
[1] -0.022553148 -0.003004035  0.016545078
[1] -0.004416301  0.014960588  0.034337478
[1] -0.009665545  0.009813423  0.029292391
[1] -0.034476602 -0.014773927  0.004928748
[1] -0.025977733 -0.006251579  0.013474575
[1] -0.002699036  0.017051743  0.036802521
[1] -0.010480175  0.009363427  0.029207029
[1] -0.032002546 -0.012307894  0.007386758
[1] -0.029632377 -0.009711457  0.010209464
[1] -0.0203449679 -0.0006713485  0.0190022710
[1] -0.0200708700 -0.0004346696  0.0192015309
[1] -0.023238939 -0.003659484  0.015919971
[1] -0.024821591 -0.004881682  0.015058228
[1] -0.011626945  0.007985632  0.027598210
[1] -0.0194360411  0.0001868259  0.0198096930
```

This is pretty close to what was expected; in this particular case the true mean was not within the CI in six cases out of 100 (we expected about five).

To reiterate, understanding the difference between the SD and the SEM is critical. The SD gives us an indication of how spread out the data in the underlying population is. The SEM is an indication of how confident we are in our estimate of the true mean of the underlying population.

Many plots in publications show error bars. There is no standard as to what these represent; it could be ±SD, ±SEM, ±1.96SD, ±1.96SEM, or, as we will see later, something else. If the publication does not explicitly state what the error bars represent, they are of no use to you (and you might begin to question the underlying analysis).

