

BIM 105

Probability and Statistics for

Biomedical Engineers

David M. Rocke

Department of Biomedical Engineering

Course Information

- <http://dmrocke.ucdavis.edu/courses.html>
 - Assignments, announcements
- <https://canvas.ucdavis.edu>
 - Access to text, online homework, grades
 - Bookshelf link for inclusive access
 - M-H Connect link for text and online homework

Class Meetings:	Tu/Th 6:10pm–7:30pm, Olson 206
Discussion:	A02: Wed 12:10pm–1:00pm, 1116 Academic Surge A03: Wed 1:10pm–2:00pm, 1116 Academic Surge A04: Tue 5:10pm–6:00pm, 1116 Academic Surge
Office Hours:	Tue & Wed 3:00pm–4:00pm. Other times by appointment.
Office:	140B Med Sci 1C (752-6999) E-mail: dmrocke@ucdavis.edu Web site: http://dmrocke.ucdavis.edu/ (assignments, schedule, announcements, etc.) Canvas site: BIM 105 FQ 2019. This will be used to access the electronic textbook, on-line homework, and test and homework scores.

Text and Software:	Statistics for Engineers and Scientists, Fifth Edition, William Navidi, McGraw Hill. The book and required web site are available online under the UCD Bookstore inclusive access program (see below). DO NOT BUY A PHYSICAL BOOK. MATLAB 2019a (available in the Engineering Computing Lab as well as virtually)
TA:	Hua Li (hihli@ucdavis.edu).
Course Grading:	<div>Midterm Examination 30%</div> <div>Final Examination 40%</div> <div>Homework (see below) 30%</div>

Homework:	<ul style="list-style-type: none"> • Each class will have a textbook reading assignment and associated LearnSmart assignment which is graded and which must be done on the Connect website before class. • Each homework assignment will have an online component which is graded and which must be done on the Connect website before class on the due date. • There will also be a component of each assignment that must be handed in on paper in class on the due date or e-mailed to the TA. • Late assignments will not be accepted.
Prerequisites	Math 21D (grade C– or better) ENG 6 (may be concurrent)
Final Exam	December 9, 2019, 8:30pm–10:30pm

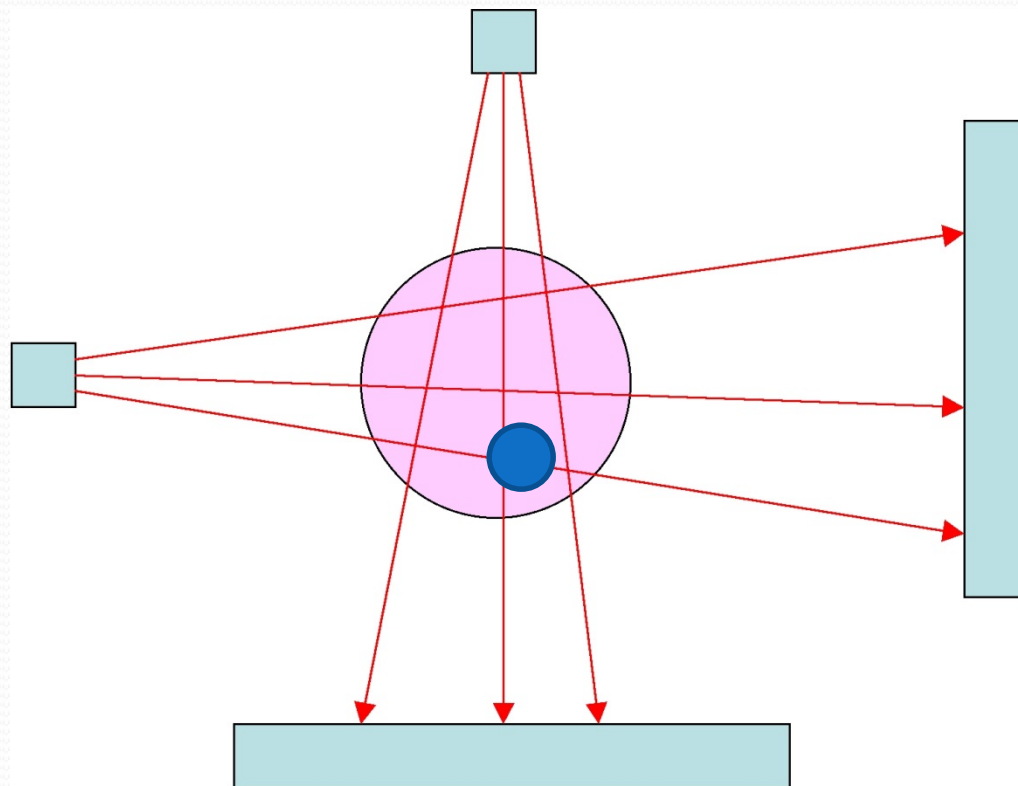
- This course is an introduction to probability and statistics as it is used in biomedical engineering.
- The emphasis is on understanding the methods available and how they are used.
- Homework assignments will be due weekly and will involve problem solving, conceptual understanding, computations, and some MATLAB work.
- Some parts of these MUST be done online via McGraw-Hill Connect.
- Reading assignments should generally be completed before class.
- Some material in the reading will not be repeated in class and some material presented in class supplements the reading.
- If you miss a class, you are responsible for obtaining the notes from another student.
- Homework assignments will be given weekly—late assignments will not be accepted.

- The midterm examination will be given in class (tentative date: Thursday, October 31)
- The final examination will be Monday, December 9, 8:30pm–10:30pm.
- Exams will require a scientific calculator.
- Please consult the course outline for details about McGraw-Hill Connect.
- It is now required by the Campus that all course outlines contain notice of the recently revised Code of Academic Conduct, which may be found at <http://sja.ucdavis.edu/files/cac.pdf>
Please read this important document.

Why This Course?

- Everything we know in biomedical engineering comes from one of three sources.
- Theory tells us what should happen based on mathematics and physics.
- Experiments tells us what happens when we affect a process.
- Computation can simulate a complex experiment when we cannot model it with simple mathematics.

Computerized Axial Tomography (CAT) Scan—*Theory*



The signal is weakened when the path contains denser tissue such as bone.

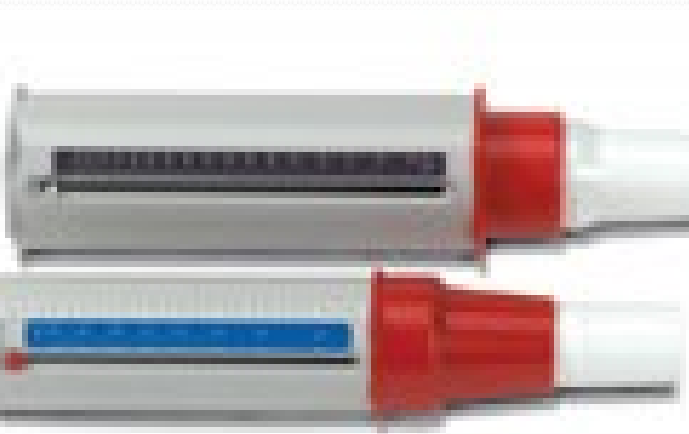
The signal is essentially the integral of the density along the path.

If we know the density everywhere, we can compute the signal at each point. This is the Radon transformation.

CAT scans invert the Radon Transformation to obtain the density from the collection of signals.

Measuring Peak Expiratory Flow Rate—Experiment

- This is a measure of lung function, important in assessing asthma, emphysema, and other types of COPD. The units are L/min.
- The original meter was designed by Martin Wright, a noted British biomedical engineer.
- Smaller, mini-Wright meters are often used now since they are easier for patients to carry, particularly children.
- The data on a following slide show a comparison between the two meters for a group of individuals.



subject	std.wright	mini.wright
1	494	512
2	395	430
3	516	520
4	434	428
5	476	500
6	557	600
7	413	364
8	442	380
9	650	658
10	433	445
11	417	432
12	656	626
13	267	260
14	478	477
15	178	259
16	423	350
17	427	451

How do the two measurements compare?

If we treat the standard Wright meter as accurate, what is the accuracy of the mini-Wright?

If we treat both measurements as variable, how does the variability compare?

Is the mini-Wright meter accurate enough to use?


And what does 'accurate' mean in this context?

Learning Objectives for BIM 105

- This class could be called Data Science for Biomedical Engineers.
 - Biomedical engineering is empirical—if you want to know if something works, you try it.
 - But things don't come out the same every time. If you want to build a stronger engineered cartilage, and you try something new, and you measure the strength, it will vary from try to try.
 - So if you want to know if your new method is better, you have to allow for variability, and that is what we learn how to do.

- We will learn how to take measurements—data—and convert them into knowledge. We first have to make measurements that reflect properties that are important to the engineering objective.
- “...when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.” (Sir William Thompson, Lord Kelvin, physics star).

- We will learn that everything varies, that you can't step into the same river twice (Heraclitus).
- What does it mean that you can't step into the same river twice?
- We will learn how to see through the variability to make conclusions:
 - How to compare a new method with an old one;
 - How to compare several methods with each other;
 - How to see if changing the temperature, the pressure, or the concentration of a reagent affects the results, and by how much.

- 
- We will learn how to take data, summarize it, graph it, and compute summaries from which we can answer the above kinds of questions, by hand in small problems, and by computer in MATLAB.
 - You will need to use methods like these to understand your experiments throughout your career, in the lab and in the field.

How to Succeed in this Class

- **The textbook is a resource.** You should look over the sections to be covered before class—this is encouraged by the LearnSmart component of the homework, which involves answering questions as you go through the text.
- **The lecture notes will always be posted.** These are not duplicates of the textbook, though obviously the material overlaps. When you are taking notes, you don't need to write down formulas etc. that are in the lectures, but you should take down the viewpoints presented during the lectures.
- **The posted lecture notes are not the lectures.** They are the skeleton of the lectures, which are fleshed out in class. If you don't come to class, you will often miss important points of emphasis, and that will later cause you trouble.

- **Do the homework. Yourself.** This consists of two on-line components in Connect and a written component. You may work together, but if you just copy someone else's homework, you won't learn the material, and you will probably do poorly on the exams. **If you can do the homework, you can do everything on the exams.**
- **Ask questions.** If there is something in class that you don't understand, ask. If you are working on the homework and don't know quite what to do, then ask in class, come to office hours, or email me.
- If you can't make office hours, just email me and we will schedule a time to talk.

- **If you don't do well on the midterm**, then talk to me, and we will try to figure out how to do a midcourse correction to improve your chances on the final.
- **If you do well on the midterm**, don't slack off. The material in the second half of the course is in some ways more complicated than the foundational material in the first half. It is also the most important for your future use in the lab and in the field.

- The following data are selected from the paper “Chondroitinase ABC Treatment Results in Greater Tensile Properties of Self-Assembled Tissue-Engineered Articular Cartilage,” by Roman M. Natoli, Christopher M. Revell, and Kyriacos A. Athanasiou, Tissue Engineering, A15, 2009. These consist of tensile-strength measurements in kPa for 6 self-assembled pieces of engineered cartilage.
- The values are 160, 250, 232, 309, 263, 288.
- We would like to beat a previous average level of 200 kPa. We can see that five of the six readings are higher than that and one is lower.
- How would we decide if the new procedure produces stronger engineered cartilage (at least on the average)?

Probability and Statistics

- Probability is used in *theory*. It tells us what outcomes to expect given some assumptions about the experiment.
- Statistics is used in *experiments*. It tells us what happened.
- We can also use probability and statistics to see if experiments came out the way we expected or not.
- Both are vital tools in practical biomedical engineering.

Types of Data

- A *variable* is a repeated quantity all of the same kind from different subjects, units, areas, etc.
- A *quantitative variable* is a measurement. For example the 17 measurements from the standard Wright meter in L/min are (494, 395, 516, 434, 476, 557, 413, 442, 650, 433, 417, 656, 267, 478, 178, 423, 427).
- We can think of this as a row vector or a column vector as is convenient.

Types of Data

- A *count variable* is one where the numerical values are counts (whole numbers), such as the number of defects on a silicon wafer. (0, 1, 1, 0, 0, 3, 1, 0).
- A *factor* or *categorical variable* is one where each observation is one of several possibilities, but these are not numbers as such. For example white blood cells can be categorized into T-cells, B-cells, etc., so a sample might look like this: (T-Cell, T-Cell, B-Cell, T-Cell, ...)

Summarizing a Quantitative Variable

- The 17 measurements from the standard Wright meter in L/min are (494, 395, 516, 434, 476, 557, 413, 442, 650, 433, 417, 656, 267, 478, 178, 423, 427).
- We can sort them in order.
- The largest value is 656. The smallest is 178.
- The range is $656 - 178 = 478$. This is a measure of “spread”.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
178	267	395	413	417	423	427	433	434	442	476	478	494	516	557	650	656

- There are 17 observations. The middle one, called the *median*, is observation number $(17 + 1)/2 = 9$. which is 434. This is a measure of location, a kind of average.
- If we had only 16 observations (omit the last one), then the median is observation number $(16 + 1)/2 = 8.5$, meaning we average the 8th and 9th observations for the median, obtaining $(433 + 434) = 433.5$.
- The median is relatively robust, meaning that it does not change much if there is an outlier or very large value. If the 17th observation was 1656 instead of 656, the median is not changed.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
178	267	395	413	417	423	427	433	434	442	476	478	494	516	557	650	656

- The **quartiles are the points that divide the data into four equal parts**. The median is the second quartile.
- The first quartile is the point with index $(n + 1)/4$, in this case $(17 + 1)/4 = 4.5$. We average the 4th and 5th observations to get $(413 + 417)/2 = 415$.
- The third quartile has index $3(n + 1)/4 = 13.5$, so we average the 13th and 14th observations to get $(494 + 516)/2 = 505$.
- These definitions differ slightly in different implementations.
- **We follow the book in always taking an unweighted average of the two adjacent values if the index is not a whole number.**
- *Remember the data need to be sorted first.*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
178	267	395	413	417	423	427	433	434	442	476	478	494	516	557	650	656

Values for Different Sample Sizes

Sample Size	Median		First Quartile		Third Quartile	
17	9	x_9	4.5	$(x_4 + x_5)/2$	13.5	$(x_{13} + x_{14})/2$
18	9.5	$(x_9 + x_{10})/2$	4.75	$(x_4 + x_5)/2$	14.25	$(x_{14} + x_{15})/2$
19	10	x_{10}	5	x_5	15	x_{15}
20	10.5	$(x_{10} + x_{11})/2$	5.25	$(x_5 + x_6)/2$	15.75	$(x_{15} + x_{16})/2$

The first quartile is the point with a quarter of the data below it. If there are 18 data points, then a quarter of the data consists of 4.5 data points. But any cutpoint will have either 4 points below it or 5 points below it, so really any number between the 4th and 5th data point is a reasonable value for Q1. This is why there are different definitions.

Five Number Summary

- **Min, Q₁, Median, Q₃, Max**
- 178, 415, 434, 505, 656
- These numbers form the basis of the *boxplot*, to be described later.
- We have a measure of location, the median = 434.
- We have a measure of variation or spread, the interquartile range (IQR) which is the difference between Q₃ and Q₁, or $505 - 415 = 90$.
- This is better than the range, which depends too much on the single largest and smallest values.

Average

- The word “average” is ambiguous. The median is a kind of average, but so is the mean.
- The *mean*, specifically the arithmetic mean, is obtained by adding up all the numeric values and dividing by the sample size.
- Like the median, this is a measure of location.

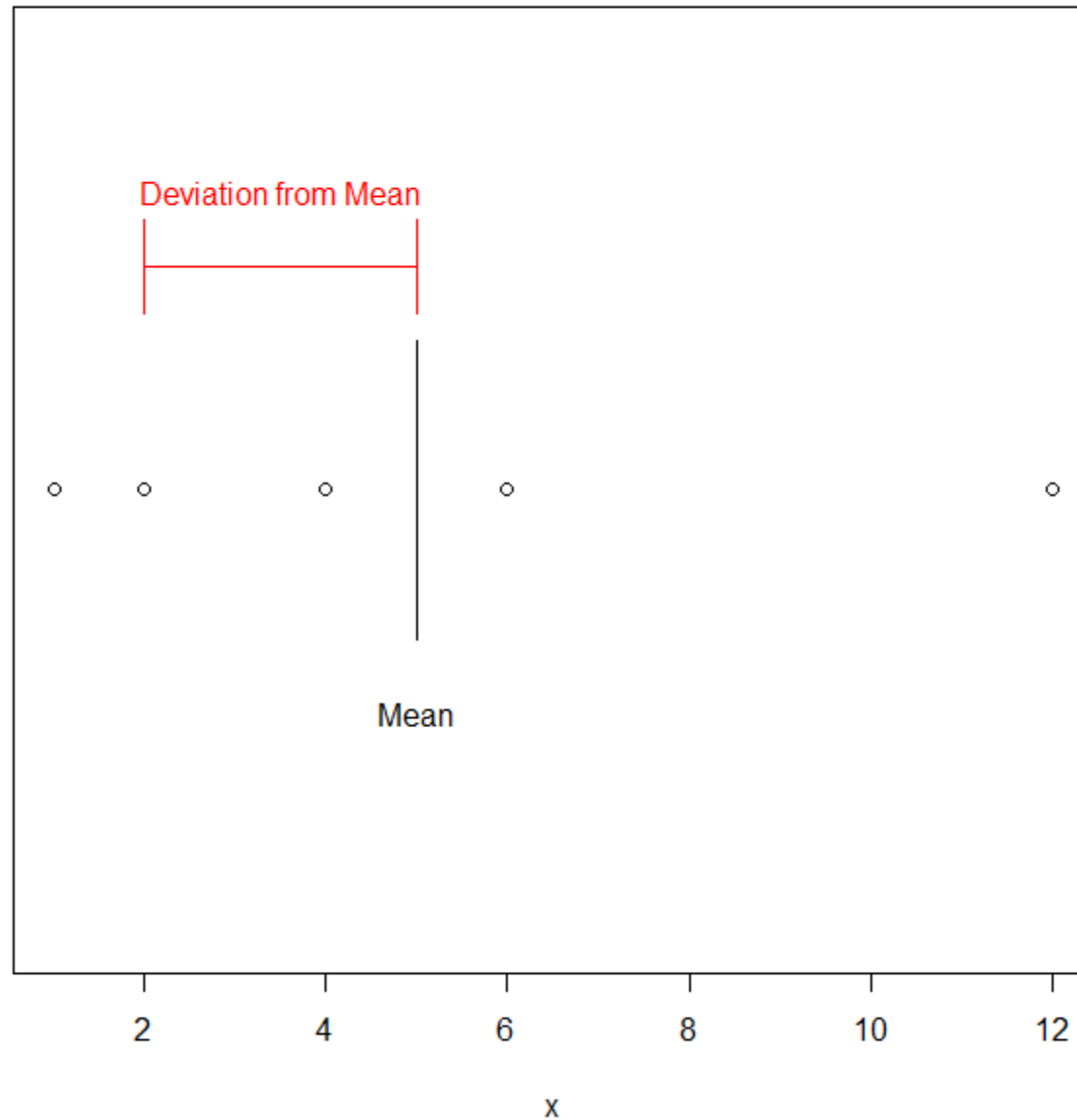
subject	std.wright	mini.wright
1	494	512
2	395	430
3	516	520
4	434	428
5	476	500
6	557	600
7	413	364
8	442	380
9	650	658
10	433	445
11	417	432
12	656	626
13	267	260
14	478	477
15	178	259
16	423	350
17	427	451
Sum	7656	7692
Average	450.3529412	452.4705882
	450.35	452.47

The mean for the standard Wright meter is 450.35 compared to the median of 434. This is not very different.

Normally, we don't report the mean to a high number of decimal places. One or two more than the data is usually enough.

Measures of Variation

- The IQR is a measure of variation.
- Another way to approach this is to find a *measure of location*, the center of the data, and find the *average* amount of deviation of the observations from the center. There are many ways to do this.
- A common way in statistics is to compute the *variance* and the *standard deviation*.
- Here we use the mean as the center and use the RMS average to compute the variance.
- While this may not seem the most natural choices, there are good reasons in statistical theory for these definitions.



The Variance

Index	Data	Deviation = Error	Square Error
1	4	-1	1
2	6	1	1
3	2	-3	9
4	8	3	9
5	5	0	0
Sum	25	0	20
Mean [$/n$]	5	0	
Sample Variance [$/(n-1)$]			5

We divide by the degrees of freedom (df), which is $n - 1 = 4$ in this case, because there are only four independent deviations. They must add to 0, so the fifth one is always known.

The Standard Deviation

- The sample standard deviation (s) is the square root of the variance (s^2).

x_1, x_2, \dots, x_n are the data of a sample of size n

The mean

$$\bar{x} = n^{-1} \sum_{i=1}^n x_i$$

The sample variance

$$s^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The sample standard deviation

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

subject	std.wright	Errors	Square Errors
1	494	43.65	1905.07
2	395	-55.35	3063.95
3	516	65.65	4309.54
4	434	-16.35	267.42
5	476	25.65	657.77
6	557	106.65	11373.60
7	413	-37.35	1395.24
8	442	-8.35	69.77
9	650	199.65	39858.95
10	433	-17.35	301.12
11	417	-33.35	1112.42
12	656	205.65	42290.71
13	267	-183.35	33618.30
14	478	27.65	764.36
15	178	-272.35	74176.12
16	423	-27.35	748.18
17	427	-23.35	545.36
Sum	7656		216457.88
Average	450.3529412	Var	13528.62
		SD	116.31

Round values for reporting, but do not round intermediate values in calculations. This can lead to large errors.

Summarizing Count and Categorical Data

- We can still compute the mean, median, standard deviation, etc. for count data, but not for categorical data.
- There may be many ties.
- If there are only a small number of values for count data, then recording the frequency of each makes sense. This is the only approach for categorical data.

Bladder Cancer Study

- Follow-up on 118 bladder cancer patients.
- The number of cancer recurrence events was recorded.
- The status is given as 0 = alive, 1 = dead from bladder cancer, 2 = dead from other causes.
- Status looks like a number, but it is not numerical, so has no average, only frequencies.
- Recurrences is numerical and can be averaged etc.
- The number of occurrences is grouped numerical data, such as there were 11 patients each of whom had 2 recurrences. We have to count each recurrence number by the number of patients who had that many.

Recurrences

Number of Recurrences	Number of Patients	Total Recurrences	Error	Square Error	Square Error times Number of Patients
0	56	0	-1.60	2.56	143.66
1	23	23	-0.60	0.36	8.33
2	11	22	0.40	0.16	1.75
3	8	24	1.40	1.96	15.64
4	4	16	2.40	5.75	23.01
5	8	40	3.40	11.55	92.38
6	1	6	4.40	19.35	19.35
7	1	7	5.40	29.14	29.14
8	3	24	6.40	40.94	122.81
9	3	27	7.40	54.73	164.20
	118	189/118 = 1.60			620.28/117 = 5.30 SD = 2.30

Status

Status	Number of Patients
Alive	88
Dead From Bladder Cancer	2
Dead from Other Causes	28

Changes in Scale

- For all of the measures of location, all of the quantiles, and the minimum and maximum, if we change units by multiplying by a conversion factor (e.g., 2.54 cm/in), then the summary also changes by multiplying by the same factor.
- A sample of measurements in inches is 3, 7, 2.5, 4, 5. The mean is 4.3. In cm, the measurements are 7.62, 17.78, 6.35, 10.16, 12.7 and the mean is $10.922 = 4.3 \times 2.54$.
- If we add 10 in to each measurement, the mean increases by 10.

Changes in Scale

- The standard deviation in inches is 1.79 and the standard deviation in cm is $1.79 \times 2.54 = 4.54$
- If we add 10 to all the measurements, the standard deviation does not change.
- The IQR also operates this way.
- For the variance, if we multiply the data by 2.54, the variance is multiplied by 2.54^2 .
- This is because the variance is on the scale of **square** deviations from the mean.

Outliers

- An outlier is a data value that seems too large or too small to be correct or at least much larger or smaller than the other values.
- If the data value itself is unlikely to be correct from subject matter knowledge, then it can be omitted.
- If we plate 10 cells each into wells of a microplate and return in 24 hours and count the live cells, then numbers as small as 0 are possible (they all died). If most of the values are in the range of 20-40, then a value of 153 is likely to be an error.

- If a value seems out of line and if an investigation of the generating mechanism shows an identifiable problem, then that data point can be omitted.
- A study of a plant converting ammonia to nitric acid shows one day's operations out of 21 at a lower efficiency. Further investigation shows that this occurred right after restarting the plant. This observation can be omitted.
- If the data point just looks different from the others, then it is not valid to omit it.
- **If the data point does not agree with your theory, it is especially invalid to omit it.**

Data Analysis in MATLAB

- MATLAB is not primarily a statistical package, so for serious work most people would use packages like R, SAS, Stata, or SPSS.
- It does have enough capabilities to do everything we need in this course.
- Today, we will talk about getting data into MATLAB and computing the numerical summaries we have discussed.
- Often, there are many ways to accomplish the same goal.
- The MATLAB online documentation is good, so Googling the topic works well “MATLAB import data”.

Importing Data

- Download `wright.csv` from the website to a directory of your choice (right click and choose save link as).
- Navigate to that directory using the file and folder tools on the left of the MATLAB window.
- Double click the file.
- You can import as separate column vectors or as a “table”.
- Or you can use the Import Data button on the toolbar.
- The table structure is generally preferred.

<http://www.mathworks.com/help/matlab/data-import-and-export.html>

```
>> wright
```

```
wright =
```

subject	stdwright	miniwright
1	494	512
2	395	430
3	516	520
4	434	428
5	476	500
6	557	600
7	413	364
8	442	380
9	650	658
10	433	445
11	417	432
12	656	626
13	267	260
14	478	477
15	178	259
16	423	350
17	427	451

```
>> mean(wright.stdwright)
>> median(wright.stdwright)
>> min(wright.stdwright)
>> max(wright.stdwright)
>> quantile(wright.stdwright,.25)
>> quantile(wright.stdwright,.75)
>> fns = [min(wright.stdwright) quantile(wright.stdwright,.25)
          median(wright.stdwright)
          quantile(wright.stdwright,.75) max(wright.stdwright)]
```

```
fns =
```

```
178.0000  416.0000  434.0000  499.5000  656.0000
```

The quartiles are a little different from the by-hand versions.

$Q1 = (413 + 417)/2 = 415$ and

$Q3 = (494 + 516)/2 = 505$

But both estimates of $Q1$ are between 413 and 417 and both estimates of $Q3$ are between 494 and 516.

```
>> summary1 = [mean(wright.stdwright) var(wright.stdwright)
                std(wright.stdwright)]
```

```
summary1 =
```

```
1.0e+04 *
```

```
0.0450    1.3529    0.0116
```

```
>> sprintf('%10.2f',summary1)
```

```
ans =
```

```
450.35 13528.62 116.31
```


Homework

- There are three types of homework in this class: online LearnSmart, online homework assignments, and written homework assignments.
- LearnSmart is a method of guided reading. First quickly review the assigned sections, then complete the LearnSmart questions, which are accessed from the assignment list on the Connect web site. This is due before class on each class day.
- The online assignments are homework problems. The assignment is due before class on the due date.
- The written assignment is due by the end of class on the due date.
- Ideally, don't drop the homework on the desk and then walk out! If you have better things to do than come to class, ask someone else to hand it in or e-mail to the TA.

Homework

- Usually the material discussed in class will be the focus of the LearnSmart assignment due before class, but today's material will be on the LearnSmart assignment due on Tuesday.
- There is also a LearnSmart assignment due Thursday before class, and most other class days as well.
- There is an online assignment due Thursday, and a written assignment due Thursday, which is on my website, not the Connect web site.