



Pathobiology for Investigators, Students and Academicians

# Lunch & Learn: Science, Statistics and Getting it Right



*A Workshop Sponsored by  
ASIP Committee for Career Development & Diversity  
and ASIP Education Committee*

**Dan A. Milner, Jr.**

American Society for Clinical Pathology, Chicago, IL

---

September 26, 2017 • 12:30 PM – 1:25 PM

**Vignette 1**

**Vignette 2**

*Vignettes Edited by Mark E. Sobel, MD, PhD, ASIP, Rockville, MD*



American Society for Investigative Pathology

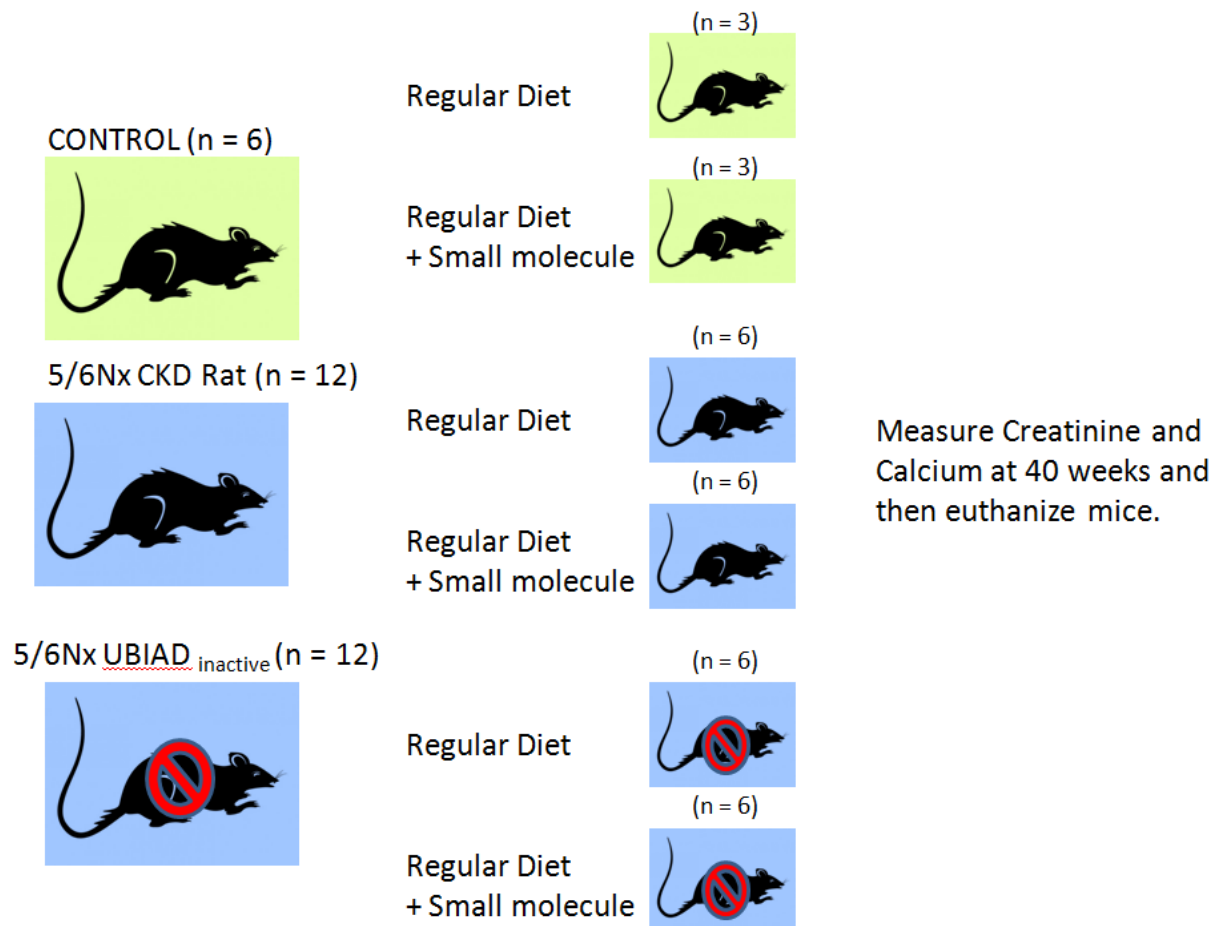


**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

**Vignette 1**

A small molecule screening program using human endothelial cell lines that develop calcification in the presence of calcification-inducing media identifies a molecule that drastically reduces the process. Interestingly, this molecule has extremely high oral bioavailability and is eliminated by direct excretion in the urine. The working biological hypothesis is that the molecule enhances the activity of UBIAD1 (an intracellular cholesterol regulator). Your laboratory has a working model of the 5/6Nx rat chronic kidney disease system and a collaborator happens to have a CRISPR/CAS9 tool to replace the UBIAD1 gene with an inactive form of the protein. In your system, the 5/6Nx rats develop chronic kidney disease including vascular calcifications and you monitor the disease using a peripheral blood measure of creatinine prior to euthanasia. In your collaborator's system, serum calcium is elevated in the UBIAD1 inactive form in normal rats.

You design an experiment to test the small molecule in your system as follows: At the end of 40 weeks, you measure the creatinine and calcium (see graph on the next page) of all of the rats and then sacrifice them. Using histology (H&E and Von Kassa stain) along with ImageJ (a free software program that allows you to do image based analyses, such as count cells or parse out a specific feature (nuclei, cytoplasm, etc) – download from <http://imagej.nih.gov/ij/>). You quantify the amount of calcification in the kidneys and the heart (see graph on the next page).

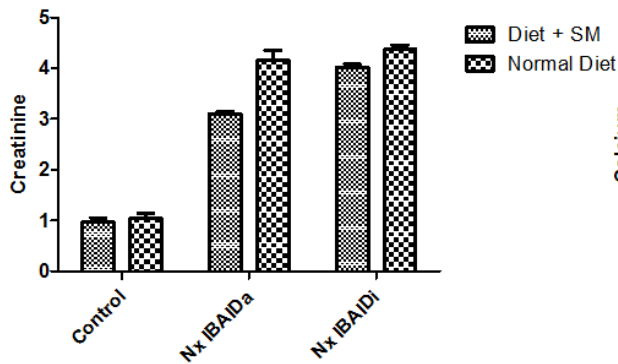


All rats are fed a 2% Ca, 1% P diet and followed for 40 weeks

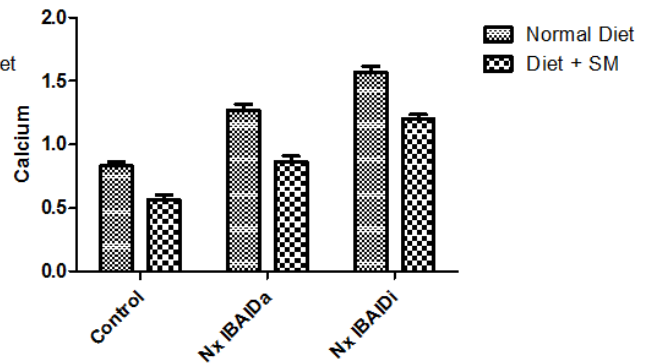
LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT

PISA 2017 WORKSHOP

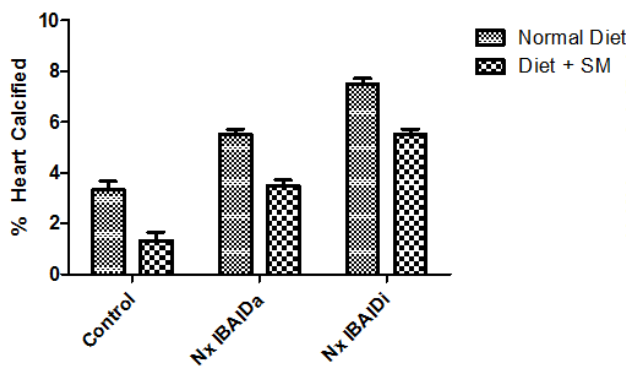
Creatinine Measures at 40 weeks



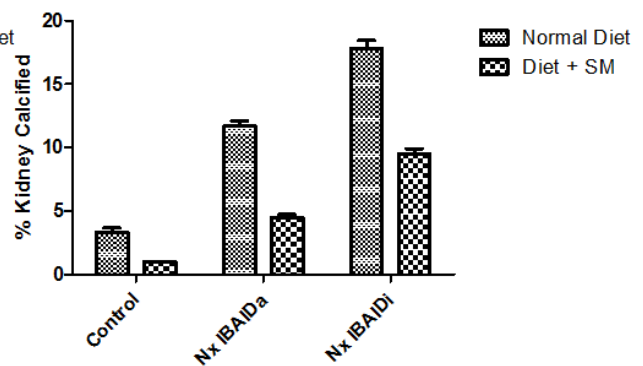
Calcium Measured at 40 weeks



Heart Area % Calcified at 40 weeks



Kidney Area % Calcified at 40 weeks



Questions and Discussion:

1. Are the differences in measurements of creatinine, calcium, and tissue calcification different between the groups? Which ones?
2. When approaching data such as this, a few questions need to be answered prior to beginning any analysis (and should best be thought of before designing the experiment!):

A. What kind of variables do I have?

In this case, we have two categorical variables (mouse and diet) and we have four different continuous variables (creatinine, calcium, heart % area calcification, kidney % area calcification). If we only had one categorical variable with two groups, then doing a t-test with any one of these continuous variables may be appropriate (see below). But, if we have a categorical variable with more than two categories (mouse), we have to use an ANOVA (analysis of variance) procedure to determine first if any of the groups are different from the others. If we only look by one category (mouse), this is a one-way ANOVA. But, we are also comparing by diet and thus one can visualize a 3 x 2 table where the cells contain the mean value for the given continuous variable.



**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

**B. What kind of test can I perform?**

This depends on the distribution of the data. For normally (or Gaussian) distributed data, parametric tests (t-test, ANOVA, etc) are the correct test to use. If the data are not normally distributed OR there are very few data points in the data set (very common with mouse experiments), it is best to use non-parametric tests (Wilcoxon Rank Sum, Kruskal-Wallis Test). A Wilcoxon Rank Sum tests looks at the order of all of the data in the data set to determine if one group's members fall towards the top or the bottom of the total ordered list (for two groups). Kruskal-Wallis does a similar method but for more than two groups. Unfortunately, there is not an easy way to do a non-parametric equivalent of a two-way ANOVA; thus, for this example, a two-way ANOVA is the best approach.

**C. What kind of result am I looking for?**

By "eyeballing" the graphs, it is pretty clear that for all four measures, there may be a difference between the controls and the Nx mice. It is pretty clear that the small molecule values are lower for calcium and calcification in the heart and kidney. But, the question is, "Is this result significant?" In the case of this design, we are specifically asking, "Is the difference in the two columns for each mouse more than I would expect by chance?" Another way of stating it is, "Given the power of my study (based on sample size), are the differences in the columns more than I would expect to see by chance?" In any case, it is clear that the two categories (mouse and diet) could be interacting to some degree so whatever test we perform, the result has to tell us something in that context.

**D. How do I perform the test?**

At this point, you will have to decide what your comfort level is with software such as GraphPad and/or Stata (or any other high level statistics package). Excel can even do many of the basic comparative statistics (t-test, ranksum, etc). GraphPad is a great tool because it not only produces a publication ready image but also performs the statistics without too much trouble. The images below illustrate what this data would look like in GraphPad and how to analyze it.

**1. Input the data:**

We use the rows to label our mice, the columns to label our diets, and then put in each value for each mouse. Note that for the control (where we only have three mice) we leave those blank. We create a sheet for EACH variable (calcium, creatinine, etc).

Table format: Grouped		A						B					
		Normal Diet						Diet + SM					
		A:Y1	A:Y2	A:Y3	A:Y4	A:Y5	A:Y6	B:Y1	B:Y2	B:Y3	B:Y4	B:Y5	B:Y6
1	Control	1.1	0.8	1.2				0.8	1.0	1.1			
2	Nx IBAIDa	4.0	3.8	4.2	3.7	4.1	5.1	3.1	3.2	2.9	3.3	3.1	3.0
3	Nx IBAIDi	4.3	4.4	4.2	4.5	4.1	4.7	4.1	3.9	3.9	4.0	3.9	4.3

**2. Graph the data:**

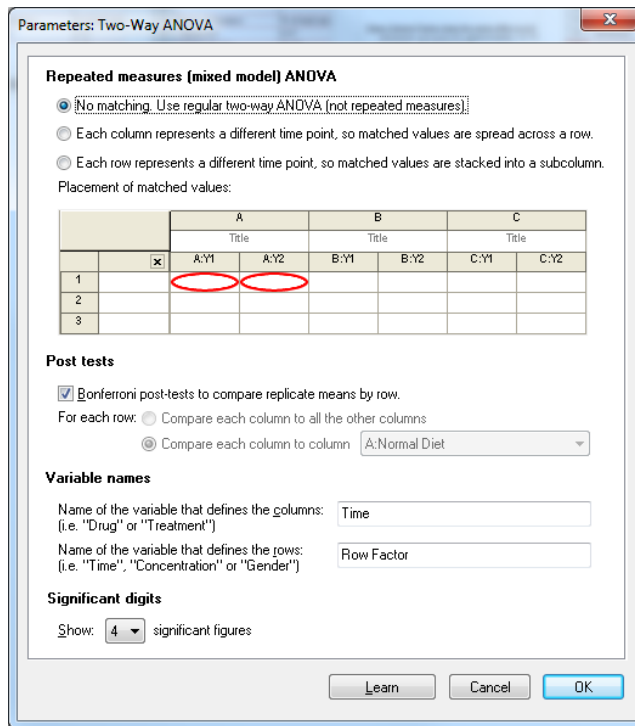
In GraphPad this is generated automatically (see above) but we can change the labels, axes, color, etc. using the options within the graph. If you are using Stata, for example, the commands begin with "graph" and

**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

searching for "graph" in the Stata library will show you the options and how to get the result you want (although it's much easier in GraphPad!). Other statistical packages have similar commands and indices.

3. Analyze the data:

Graphpad is button driven and when you click on "Analyze", you get a screen with these options:



**Parameters: Two-Way ANOVA**

**Repeated measures (mixed model) ANOVA**

No matching. Use regular two-way ANOVA (not repeated measures).  
 Each column represents a different time point, so matched values are spread across a row.  
 Each row represents a different time point, so matched values are stacked into a subcolumn.  
 Placement of matched values:

	A		B		C	
	Title		Title		Title	
	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1						
2						
3						

**Post tests**

Bonferroni post-tests to compare replicate means by row.  
 For each row:  Compare each column to all the other columns  
 Compare each column to column: A:Normal Diet

**Variable names**

Name of the variable that defines the columns: (i.e. "Drug" or "Treatment") Time  
 Name of the variable that defines the rows: (i.e. "Time", "Concentration" or "Gender") Row Factor

**Significant digits**

Show: 4 significant figures

Learn Cancel OK

For this analysis, we do not want to match because these experiments are not related. You would use matching if you followed, say, the same mouse over a 4-week period and had repeated measures of your variables from the same mouse. Here, they are independent so we do not match. We also want to see the "Bonferroni post-tests" so that we know WHICH of our subgroups are different. If we don't check this box, the software will only report the overall p-value which is not helpful. Once we click OK, we get the result below.

E. What is the result of the two-way ANOVA?

GraphPad analysis of the data for calcium (all of the analyses are similar except for creatinine) is shown on the next page. What does this tell us? First, there is little if any interaction between the mouse and the diet variables. Second, the row variable (mouse) in this case explains ~60% of the variation in the data while the column variable (diet) explains ~25%. Furthermore, the mouse and diet variables are significant in the model while the interaction term is not. Towards the bottom, we can see that the differences in the columns are reported and that they are significant for all three mouse groups. This suggests that the small molecule, regardless of the mouse, causes a reduction in calcium and calcification. Interestingly, the differences in creatinine are only significant for the Nx IBAIDa and are the reverse of the rest of the data. What this means biologically is unclear but statistically it suggests that creatinine is not related to calcium or calcification in this system.



**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

Table Analyzed	Calcium			
Two-way ANOVA				
Source of Variation	% of total variation	P value		
Interaction	0.57	0.4388		
Column Factor	25.28	< 0.0001		
Row Factor	60.75	< 0.0001		
Source of Variation	P value summary	Significant?		
Interaction	ns	No		
Column Factor	***	Yes		
Row Factor	***	Yes		
Source of Variation	Df	Sum-of-squares	Mean square	F
Interaction	2	0.01800	0.009000	0.8526
Column Factor	1	0.8008	0.8008	75.87
Row Factor	2	1.925	0.9623	91.17
Residual	24	0.2533	0.01056	
Number of missing values	6			
Bonferroni posttests				
Normal Diet vs Diet + SM				
Row Factor	Normal Diet	Diet + SM	Difference	95% CI of diff.
Control	0.8333	0.5667	-0.2667	-0.4826 to -0.05077
Nx IBAIDa	1.267	0.8667	-0.4000	-0.5527 to -0.2473
Nx IBAIDi	1.567	1.200	-0.3667	-0.5193 to -0.2140
Row Factor	Difference	t	P value	Summary
Control	-0.2667	3.179	P < 0.05	*
Nx IBAIDa	-0.4000	6.743	P<0.001	***
Nx IBAIDi	-0.3667	6.181	P<0.001	***







**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

## Vignette 2

A group of 1175 healthy subjects (43% Caucasian, 33% African or African American, 24% Hispanic/Latino) were recruited from college campuses in the Boston area (from among 26 different colleges) and were asked to provide a buccal swab for DNA sequencing along with a detailed questionnaire regarding their family history and medical health as well as a tube of blood for laboratory testing. They also agreed to complete a follow up survey every 5 years for the next 25 years in order to look at new diagnoses and diseases. For each patient, an aliquot of blood as well as the buccal swab were both used to sequence each patient to 40X coverage as well as perform comparative genomic hybridization to a sequenced and assembled reference genome. All genomes were cataloged for mutations included insertions/deletions, single nucleotide polymorphisms, and gene duplication. The survey included questions about all of the following: diabetes, hypertension, malignancy (specifically of breast, lung, colon, prostate, kidney and/or brain), infections (including frequency and specifically for mononucleosis, ear infections, head colds, urinary tract infections, toenail infections, persistent/excessive acne), diet, and exercise habits. All of the subjects were counseled to use a free pedometer (provided by the study team) which was connected to the internet and report their daily activity, which was monitored by the study.

After 10 years (3 total surveys), a manuscript was published by a non-competing group in a mouse model showing that a specific mutation of pyruvate dehydrogenase kinase 4 (PDK4) caused a massive decrease in mouse activity as well as obesity in mice. You propose to look at the pedometer data of the study's subjects' activity to see if there is an association with fewer steps and mutations in PDK4. Your PI, however, thinks that such an association may be polygenetic (or even spurious in the mouse) and the entire genome should be examined in the context of all of the data.

### Questions:

1. How would you go about investigating any potential associations in your data set?
2. What statistical considerations are important in thinking about this question?
3. How should the pedometer data be parsed for the analysis?

### Discussion:

Your first approach (just look at the pedometer data and the mutations in PDK4) may be an okay place to start to generate hypothesis. The easiest approach would be to look at the mutations in PDK4 and parse the data into those groups. For example, if there is only one single nucleotide polymorphism in PDK4, then you can divide the patients into two groups (those with and those without the major allele) and perform a t-test or rank sum test (depending on the normality of the data). You can go see everything that's known (pretty much) at this link: <http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=5166>

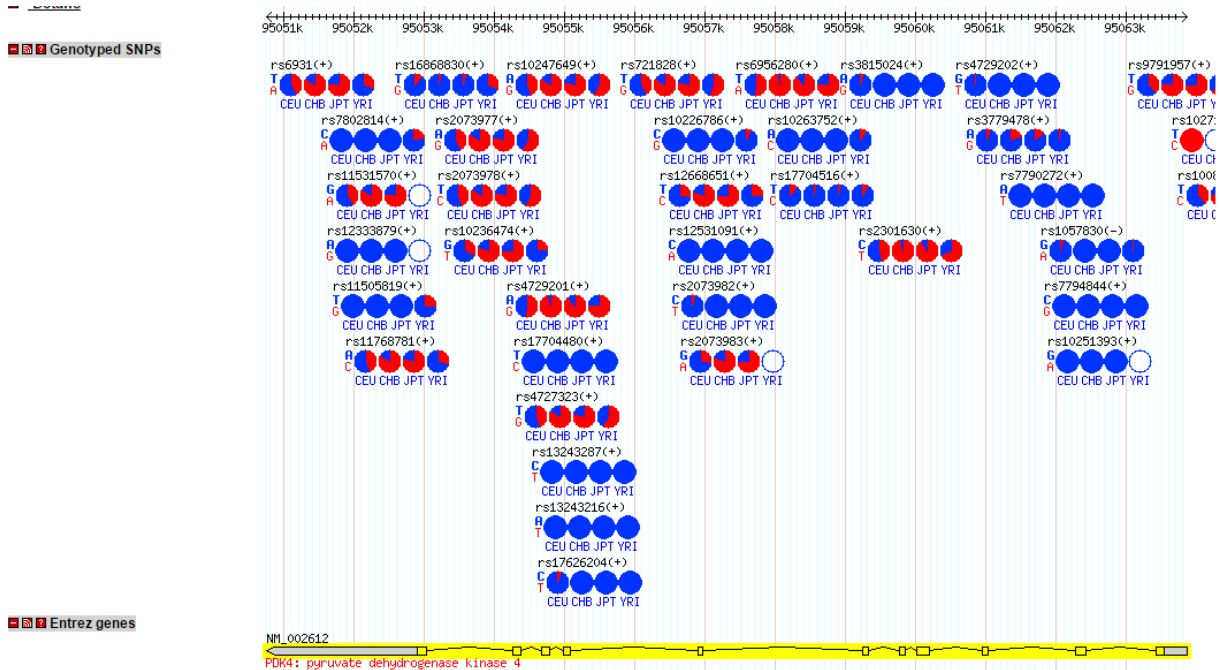
You can view the gene in the HapMap project (to look at variation) here:

[http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap24\\_B36/#search](http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap24_B36/#search)



**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

Here are some snapshots for discussion:



Uh oh! It looks like there are 35 SNPs in PDK4! Note the SNPs where all the circles are BLUE, which indicates that, in the original genotyping project, these SNPs were not seen (they have a low minor allele frequency). In the SNPs where there is quite a bit of RED, the minor allele frequency is higher/variable. Which SNP are you going to study? At this point, with so much variation, you would have to sequence across the gene for each person to confirm the haplotypes. These haplotypes (groups of associated SNPs that travel together) might THEN give you only two groups (or three) to compare, but possibly not. If there is a single important SNP in this group, it is likely that a long range haplotype analysis would point it out easily, but we can't know that from eyeballing. So, in this case, your boss is probably correct and the best approach is to assume nothing about genes of importance and look at the entire genome.

What statistical considerations are important in thinking about this question?

Number of samples: We have 1175 samples in this study, which seems like a lot but remember that we are going to be comparing thousands of SNPs so we have to be conscious of our power. This is why, in general, analyses like these should be considered hypothesis generating (and not definitive) because even with such a large sample size, we may not have captured all of the variation in the population (see SNP data above).

Distribution of SNPs: In our data we have nice coverage of the American population with Caucasians, African/African-Americans, and Hispanic/Latino. But in the original HapMap data, the SNPs were assessed in Africans, Caucasians, and Chinese. Since we haven't really explored the Hispanic/Latino population, we may find new SNPs or a single allele in that population. One approach to this could be to analyze the ethnic groups separately as a first pass to see if there are hits in the SNP data that are population specific.



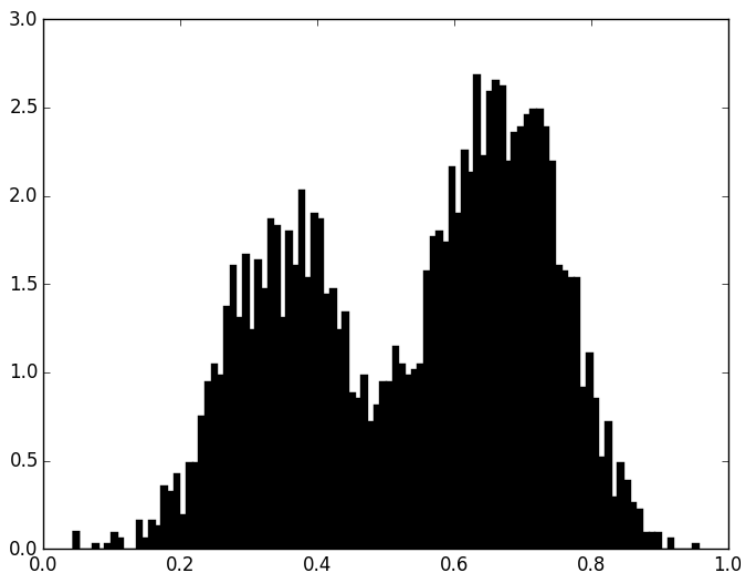
**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

Other variables: We took our data from college students and collected lots of information including exercise habits. If 25% of the students (regardless of ethnic background) were involved in a sports activity during college, this may have greatly skewed the pedometer count. To correct for this, we may need to select pedometer data only from after college and then classify individuals as active or sedentary based on their post-college level of activity. Looking back at the list of variables, it is clear that there could be many interactions or biases and we have to carefully choose how to analyze this cohort.

Censored data: When a patient or subject in a study leaves the study before the end (drops out, moves out of the area, dies of process unrelated to study), those data are said to be censored. What if 100 individuals who were highly active early on died of accidental causes 3 years into the study? They may have been at the top of the histogram if they had lived but would fall in the bottom of the histogram because they are missing years of data. Thus, removal of these individuals (or extrapolation of their projected data) is necessary to insure we are not getting an incorrect estimation.

How should the pedometer data be parsed at the individual variable level for the analysis?

Having no background in biostats, you consult with a local biostats expert who has worked with this dataset before and explain the variables of interest (pedometer measures). Before you do anything further, the biostats expert sends you a graph (the histogram below) and says that the data have been normalized relative to the highest number of steps attained by any person so that all study subjects have a value between 0 and 1 (essentially a % of the total reflective of their steps). All subjects included in this diagram have data for all the years of the study (censored patients removed). You are delighted because the data appear bimodal (there are two distinct peaks) so you can pick a cut off value (50% will probably work) to divide your pedometer data into a simple categorical variable (0 = < 50%, 1 = >50%).



value between 0 and 1 (essentially a % of the total reflective of their steps). All subjects included in this diagram have data for all the years of the study (censored patients removed). You are delighted because the data appear bimodal (there are two distinct peaks) so you can pick a cut off value (50% will probably work) to divide your pedometer data into a simple categorical variable (0 = < 50%, 1 = >50%).

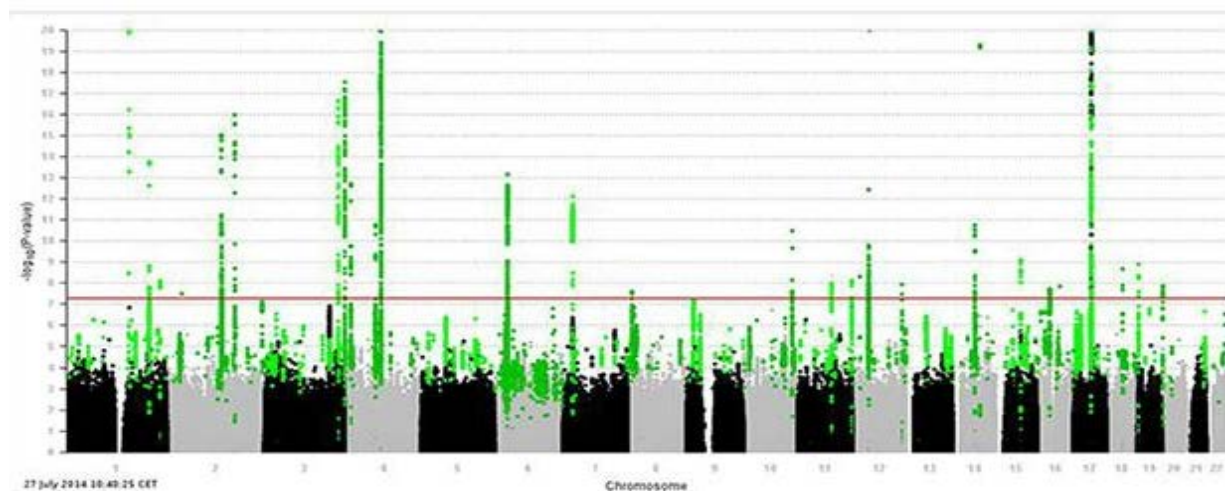
The biostats expert is happy and says all you need to do now is take each SNP in the dataset and create a two by two table as follows:

	Major Allele	Minor Allele
<50%	A	B
>50%	C	D

**LUNCH & LEARN: SCIENCE, STATISTICS, AND GETTING IT RIGHT**  
**PISA 2017 WORKSHOP**

A Fisher's exact test can now be performed, which will produce a p-value for EACH SNP based on the distribution of the major and minor alleles across the two different pedometer groups. So, you just have to do that about 100,000 times (and, hence, the reason you've asked the biostats person for help!). The result of such an analysis produces a "Manhattan Plot" (see below) which essentially graphs the p-value obtained for each SNP for the comparison between them. There are other tests that can be performed in a similar manner depending on the structure of the independent variable (in this case, pedometer category). For example, let's assume you don't want to categorize the patients as 0 or 1 but rather want to take the mean pedometer measure (or median) for each allele for a given SNP and compare them. This would be doing a t-test (or rank sum test) for each SNP and repeating for all SNPs. The result is still a Manhattan plot of p-values but the underlying test is different.

What do we learn from this Manhattan plot? The important concept to remember is that a Manhattan plot gives you potentially interesting locations in the genome from which you can generate hypotheses that need further testing. As we described above, if we don't control for anything and just run either our categorical or continuous approach with a single variable, our plot may look like this:



How do we read this? Each black or gray area across the bottom (in this example) represents a chromosome. The green dots represent p-values (signal) that are above expected noise (gray/black). The red line usually denotes the cut off for significance after corrections for multiple testing (such as Bonferroni where we divide 0.05 by the total number of tests performed). Note that we have hits on several chromosomes including chromosome 7 (where PDK4 is located). The plot is usually arranged long arm to short arm so this particular hit on chromosome 7 is on the telomere end of the long arm of chromosome 7 (which we could gut check is in the correct region for PDK4). But we have LOTS of hits on this plot, some much stronger than chromosome 7. Do we need to control for ethnicity or any of the other dozens of variables? Yes, we do. Assume we do and only a single peak remains. The challenge now is looking at that peak, the genes in the area, and making biological sense of what the results mean, which often takes confirmation and more carefully designed experiments to achieve a publishable result.