Starting with Modeling

Daniel T. Kaplan

March 6, 2014

Nick's comments:

In terms of this paper for CHANCE, I had several suggestions:

1. rather than “guide important lifestyle” would “shed light on important lifestyle” be better?
2. defining the usual categorization of BMI when it is introduced (< 20, 20-25, 25-30, 30+) would be helpful.
3. Interpreting the 30.94 and 24.51 after the code chunk mean(bmi ~ diab) would be helpful
4. can you please make the rounding coarser (“options(digits=3)” early on?)
5. Interpretation is also needed for the tally: need to be explicit that only 3.5% of those with BMI <= 30 are diabetic, vs. 12.2% for those with BMI > 30. Similarly for non diabetics (19% have BMI > 30) vs. diabetics (48% have BMI > 30.
6. Can you please number the figures, and refer to them in the text?
7. I’d suggest fitting just the diag ~ poly(age, 2)\*bmi model, along with the predicted proportion contour plot.
8. I’d suggest stopping there, with a note about whether this is what we are teaching in our intro stats courses, along with a note that most students stop after one course. A paragraph that the concepts are important, and the computation can be facilitated, etc. etc. I suspect that this is about the length of what Dalene is after.

==================

Does this sound familiar? "A parameter is to a population as a statistic is to a sample." [source](http://en.wikipedia.org/wiki/Statistical_parameter#Examples) In introductory statistics, students encounter descriptive statistics such as the proportion, mean, variance, and correlation coefficient. These descriptions then form the object of inferential statistics, what one can validly say about the population parameter based on sample statistics.

There's something important missing here: the *purpose* behind the statistical analysis. Knowing why you're doing something is essential to deciding how to do it. A key word here is "modeling," the process of making choices about how to describe a system. Or, as Starfield, Smith, and Blelock consisely phrased it 20-years ago, a model is a "representation for a given purpose." In building a model, you have to decide what features of the system are essential to your purpose and which can be omitted. Models are always approximations, but the quality of the model depends on the purpose of the approximation. As George Box famously said, "All models are wrong. Some models are useful."

The sorts of decision-making that guide modeling are not typically emphasized in introductory statistics. In a typical course, there are few choices to be made and they are usually related simply to the setting: Is the data categorical or quantitative? Is the description a proportion or a mean? Is one group being described, or the difference between two groups, or the differences among more than two groups?

It's understandable that things developed this way. Egon Pearson laid out many of the factors at work in his 1926 review of R.A. Fisher's "Statistical Methods for Research Workers" (*Science Progress*, **20**:733-737 available [here](http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/esp.htm).)

Mr. Fisher has undertaken the very difficult task of attempting to put before research workers ... without any special mathematical training, a summary covering a great range of methods ..... The book is chiefly concerned with the best methods of handling small samples .... After conscientiously working through the examples, the student should feel able to apply the methods to exactly similar problems.... [[ E. Pearson, (1926) "Review of Statistical Methods for Research Workers (R. A. Fisher)", *Science Progress*, **20**:733-734.]]

Today's introduction to statistics should take into account the many ways the statistical environment the has changed since 1926. In the era of "big data," the problems of "handling small samples" are more peripheral. Statistics is applied to much more diverse and complicated problems, not just experiments but the interpretation of observational data in the social sciences, medicine, commerce, and government.

An experimentalist sets the conditions of the system her data are collected from. The analyst of observational data, however, needs to account for the conditions **such as they may be** including the variables that are thought to be confounding. At the start this requires identifying possible confounders, choosing which factors might be important and which not. In a word: modeling.

How to teach modeling to introductory students? Computation seems essential and, fortunately, excellent statistical software such as R is easily available to students. Beyond being free, R supports a notation that's well suited to modeling. We developed the mosaic package to build on this notation to make modeling accessible, even at an introductory level.

To illustrate, consider a dataset on human body shape, disease and mortality. These data were assembled from the [National Health and Nutrition Examination Survey](http://www.cdc.gov/nchs/nhanes/nh3data.htm), NHANES-III. The motivation for such a survey is to understand better the relationship between diet and health so that people can make helpful changes in their diets. As such, causation diet -> health is central. As an observational study, NHANES-III is not ideally suited to drawing causal inference. But experiment is not feasible and data such as these are the best we have to guide important lifestyle and medical decisions.

You may have heard about the link between overweight and type-II diabetes. Originally named "adult-onset diabetes," it is now considered an "epidemic" in children. (See, for instance, this [news report](http://www.washingtonpost.com/politics/fighting-the-growing-type-2-diabetes-epidemic-in-children/2013/01/06/361c3b00-5854-11e2-9fa9-5fbdc9530eb9_story.html).)

What's the relation between diabetes and overweight? With almost 30,000 people represented, of whom more than 1500 have diabetes, the NHANES data may provide insight. The body-mass index (BMI) is a measure of overweight included in the data. One of the standard descriptions in introductory statistics is to look at differences between the group means

require(mosaic)  
load('nhanes.rda')  
# require(DCF)  
# data(nhanes) # from DCF  
# Note in Draft  
# changes to make in the data when bringing it into the mosaic package  
# rename diab as diabetes  
# change the NaN to NA  
# make diabetes categorical with levels "diabetic" and "not"  
nhanes$diab <- ifelse( nhanes$diab==1, "diabetic",   
 ifelse( is.nan(nhanes$diab), NA, "not" ) )  
nhanes$sex <- ifelse( nhanes$sex==2, "F", "M")  
# missing data are encoded as NaN, not NA. Fix this.  
for (var in c('d','m','hgt','wgt','smoke'))  
 nhanes[[var]] <- ifelse( is.nan(nhanes[[var]]),NA, nhanes[[var]] )  
nhanes$smoke <- ifelse(nhanes$smoke==1,'smoker','not')  
nhanes$smoke <- as.factor(nhanes$smoke)  
# end of Note in Draft  
  
mean( bmi ~ diab, data=nhanes, na.rm=TRUE ) # need to kill the na.rm part.

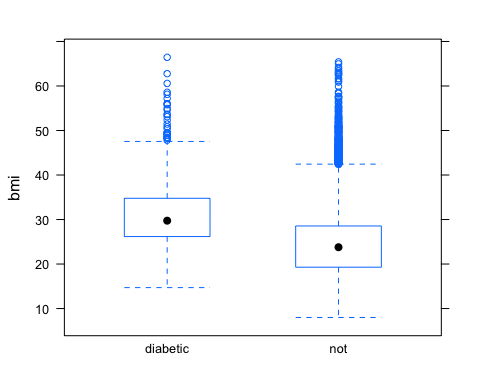
## diabetic not   
## 30.94 24.51

This notation, used in mosaic, clearly identifies the calculation being performed, the data on which the calculation will be based. The expression bmi ~ diab means "break bmi down by diabetes classification," or, "account for bmi by diabetes," or "model bmi by diabetes."

The same notation can be used for a t-test (use t.test( ) instead of mean( )) which gives a p-value on the difference of means .

A graphical depiction --- another sort of representation --- tells more of the story:

bwplot( bmi ~ diab, data=nhanes )



Clearly, the diabetics as a group tend to have higher BMI than the non-diabetics. But there is a lot of overlap. It is only for very low BMI --- say below 15 --- that there is what might be called "protection" against diabetes. But a BMI < 16 is classified by the World Health Organization as "severe thinness" [ref](http://apps.who.int/bmi/index.jsp?introPage=intro_3.html) and taken to indicate malnutrion, an eating disorder, or other health problems.

Another way to approach the relationship between diabetes and BMI is to look at the probability of having diabetes for high and low BMI. Setting the dividing point at BMI 30,

tally( diab ~ bmi>30, data=nhanes, margins=TRUE )

## bmi > 30  
## diab TRUE FALSE  
## diabetic 0.12296 0.03468  
## not 0.87704 0.96532  
## Total 1.00000 1.00000

The notation here, diab ~ bmi>=30 can be read: account for diabetes by whether or not the BMI is greater than 30.

For people with BMI > 30, diabetes is four times as prevalent as for lower BMI. Here, tally( ) is calculating conditional probabilities. The notation makes clear what is the conditioning variable. Compare the above to this:

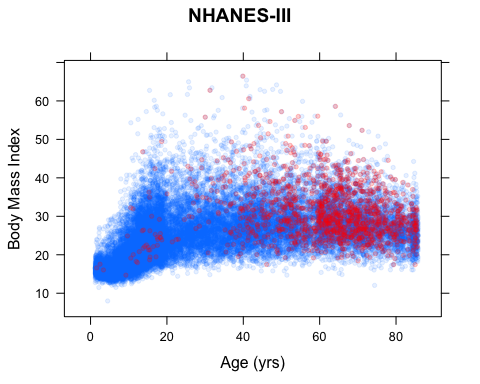
tally( bmi>30 ~ diab, data=nhanes, margins=TRUE )

## diab  
## bmi > 30 diabetic not  
## TRUE 0.4785 0.1904  
## FALSE 0.5215 0.8096  
## Total 1.0000 1.0000

(Exercise: Why are the two tallys so different?)

Each of these is a representation of the relationship between diabetes and bmi, but not necessarily the most informative representation. Here's another such representation:

# Note in draft: Change this to the new system  
plotPoints( bmi ~ jitter(age,factor=4), data=nhanes, alpha=.1, pch=20,  
 xlab="Age (yrs)", ylab="Body Mass Index",main="NHANES-III")  
plotPoints( bmi ~ jitter(age,factor=3), data=subset(nhanes,diab=='diabetic'),   
 col='red', pch=20, alpha=.2, add=TRUE)



This graphical model suggests a somewhat different story. First, you can see that BMI depends on age: Children have BMI less than 20; young and middle-aged adults have BMI > 20 and heading up toward much higher; in the elderly, BMI trends lower than for middle aged. A lot of the diabetes for people with BMI < 30 occurs in the elderly.

The graph is informative in some ways, but not in others. The most obvious purpose for such data is risk avoidance: to judge how BMI is related to risk of diabetes while adjusting for age. BMI is potentially able to be changed by diet, etc., but age is not under anyone's control. The graph portrays only a vague impression of this.

An alternative kind of model can be made to display directly the relationship of interest: risk of diabetes as a function of BMI adjusted for age. This can be expressed quantitatively as a *function*, a relationship between BMI risk and age.

Here, for instance, is an R statement for fitting a linear relationship.

mod <- glm( diab=="diabetic" ~ bmi\*age, data=nhanes)

You can read this as, "account for diabetes by a combination of age and BMI."

Such models are often interpreted using model coefficients as sample statistics. R and mosaic calculating inferential quantities as simple as this:

confint(mod)

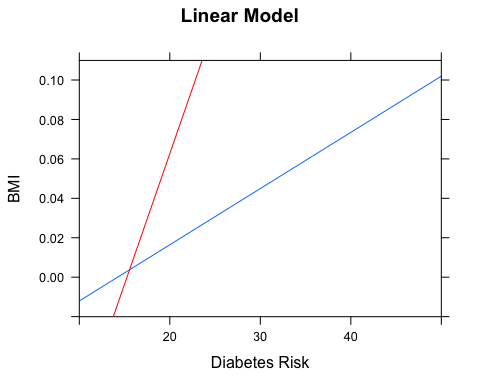
## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.0653735 0.095830  
## bmi -0.0056124 -0.004270  
## age -0.0045139 -0.003564  
## bmi:age 0.0002414 0.000278

Coefficients, though, can be hard to interpret at face value. For instance, here the coefficients are negative for bmi and age --- suggesting that diabetes risk is lowered with higher BMI and age.

It's easier to see what the fitted model has to say by plotting it as a *function*: diabetes risk versus BMI for different ages. In mosaic, you have direct access to the fitted model function itself.

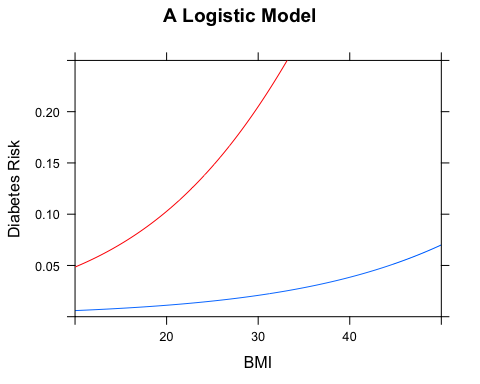
f <- makeFun(mod)  
# Simplify this statement for the article,   
# leaving out the labeling stuff  
plotFun( f(bmi, age=30)~bmi, bmi.lim=range(10,50),   
 xlab="Diabetes Risk", ylab="BMI",  
 main="Linear Model")  
plotFun( f(bmi, age=70)~bmi, add=TRUE, col='red')



# Put some text in this plot to label the lines ????

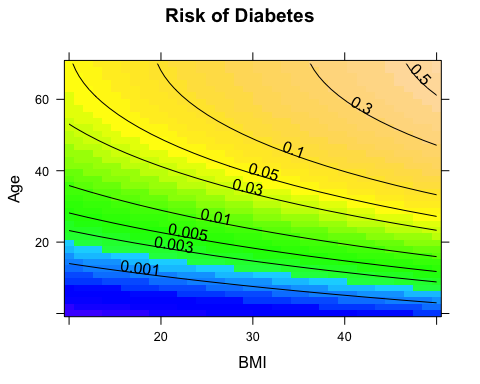
It's comparatively easy to elaborate the function adding more components or changing the form of the fitted function. For instance

mod2 <- glm( diab=='diabetic' ~ poly(age,2)\*bmi, data=nhanes,  
 family='binomial')  
f2 <- makeFun(mod2)  
plotFun( f2(bmi, age=30)~bmi, bmi.lim=range(10,50), ylab='Diabetes Risk', xlab='BMI',  
 main='A Logistic Model',ylim=c(0,0.25))  
plotFun( f2(bmi, age=70)~bmi, add=TRUE, col='red')



Or, risk can be plotted as a function of *both* age and BMI:

# Just a simple `plotFun()` will be shown in the paper.  
probLevels <- c(.001,.003,.005,.01,.03,.05,.1,.3,.5)  
plotFun( log(f2(bmi,age=age))~bmi&age,  
 bmi.lim=c(10,50), age.lim=c(0,70),  
 levels=log(probLevels), labels=as.character(probLevels),  
 xlab="BMI",ylab="Age",main='Risk of Diabetes')



The graph suggests that the risk of diabetes increases with BMI at any age.

You may ask, "Why that model and not some other? Why not include sex or other variables as a covariate? Why not split out the young and the elderly from the middle-aged?"

Our answer: Go and try it! Then you'll have to think about how to decide which of two models is better, a question which in part can be addressed through confidence intervals, etc.

For students, crafting models highlights the creative aspects of statistical reasoning. It's not just a matter of interpreting a p-value: you can examine many factors and build a model that addresses your actual purpose.

With software and comprehensible notation standing in for the missing "special mathematical training," students can be introduced to covariates and the process of judging whether a covariate contributes to the model in an important way. Indeed, without such capabilities, all one can do is preach caution about causation. But causation, even from observational data, is often central to the purpose of the statistical analysis.

To illustrate, consider what the NHANES data has to say about smoking and mortality.

tally( d ~ smoke, data=nhanes,   
 format="proportion", margins=TRUE)

## smoke  
## d not smoker  
## 0 0.92139 0.93435  
## 1 0.04828 0.04359  
## 2 0.03033 0.02206  
## Total 1.00000 1.00000

This tally suggests that smokers are somewhat less likely to die from cardiovascular causes. Indeed, according to this table, smokers are less likely to die from other causes too!

When considering sample statistics such as these, a student has few options. Perhaps the difference in proportion will turn out to be insignificant? (It's not: .) Perhaps observational data are inappropriate for drawing conclusions about causality? (Yes. But an experiment would be, to say the least, challenging to conduct.)

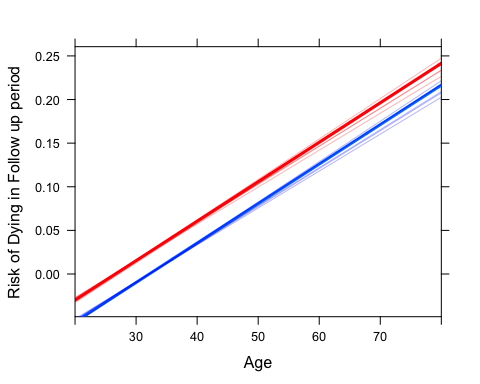
By starting with models, a path opens up to students. Here's a model corresponding to the simple tally of proportions:

mod1 <- lm( d!=0 ~ smoke, data=nhanes)

It can be important to examine confidence intervals and p-values. (Use confint(mod1) or summary(mod1).

But it's also important to examine covariates. The modeling framework makes this accessible. For instance:

mod2 <- lm( d!=0 ~ smoke + age, data=nhanes)  
f2 <- makeFun(mod2)  
plotFun( f2(age,smoke='smoker') ~ age,   
 age.lim=c(20,80), col='red', lwd=3,  
 ylab='Risk of Dying in Follow up period',  
 xlab='Age')  
plotFun( f2(age,smoke='not') ~ age, add=TRUE,  
 lwd=3)  
# IN DRAFT: move the following lines to another chunk  
# draw the confidence bands by resampling  
# In DRAFT: just a few iterations here to speed things up.  
do(5, parallel=FALSE) \* {  
 mod2 <- lm( d!=0 ~ smoke + age,   
 data=resample(nhanes))  
 f2 <- makeFun(mod2)  
 plotFun( f2(age,smoke='smoker') ~ age, add=TRUE,  
 col='red', alpha=.3)  
 plotFun( f2(age,smoke='not') ~ age, add=TRUE,  
 col='blue', alpha=.3)  
 }



## data frame with 0 columns and 0 rows

[[Note in Draft: It seems I have to use parallel=FALSE in do() when there are side effects. That's kind of nasty.]]

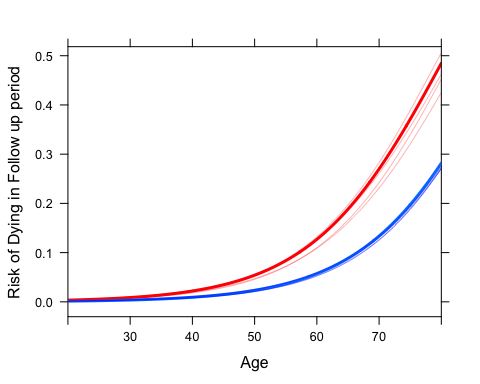
Bringing in statistical inference is straightforward using the resampling approach now coming to the fore in introductory textbooks. (Ref: Tintle book, Rossman book, Locke^5 book)

[[Note in Draft: This is where the bands would show up in the paper.]]

You can easily see that this model is wrong: the model risk is negative for younger people. But all models are wrong, and this one is a start. Francis Bacon (1562 -1626) said, "Truth comes out of error more readily than out of confusion." Or, as Brown and Cass put it: "Somehow, in emphasizing the logic of data manipulation, teachers of statistics are instilling excessive cautiousness. Students seem to develop extreme risk aversion, apparently fearing that the inevitable flaws in their analysis will be discovered and pounced upon by statistically trained colleagues." (E.N. Brown & R.E. Kass, (2009) "What is Statistics?", *The American Statistician*, **63**(2):105-110, <DOI:10.1198/tast.2009.0019>, available [here](http://www.stat.cmu.edu/~kass/papers/what.pdf))

The particular flaw here is easily addressed: use a less stiff model form, such as logistic regression:

mod3 <- glm( d!=0 ~ smoke + age,   
 family='binomial',data=nhanes)  
f3 <- makeFun(mod3)



## data frame with 0 columns and 0 rows

It's important to be able to recognize the flaws in any analysis, and to be able to address them if needed. A modeling approach to statistics emphasizes the need to represent the relationships within data, to explore different possibilities, and to focus on the purpose of the work (often, effect size after adjustment for covariates), rather than on p-values. R provides the power. The mosaic package, we hope, provides a helpful simplicity and consistency.

## Leavings [Ignore this]

This might be surprising to you, given what is widely believed (and correctly!) about the dangers of smoking. Perhaps there isn't enough data? Here's a confidence interval on the change in probability of dying associated with smoking:

confint( lm( d!=0 ~ smoke, data=nhanes))

## 2.5 % 97.5 %  
## (Intercept) 0.07378 0.083443  
## smokesmoker -0.02264 -0.003285

The confidence interval is all in negative territory. (Using summary() rather than confint() shows the *two-sided* p-value at .)

What's going on? Consider the covariates of sex and age:

confint( lm(d!=0 ~ smoke + age + sex, data=nhanes) )

## 2.5 % 97.5 %  
## (Intercept) -0.166623 -0.144531  
## smokesmoker 0.012364 0.030902  
## age 0.004293 0.004682  
## sexM 0.018516 0.034324

Now smoking is associated with an increase in mortality, as is age and being a male.

This is an instance of Simpson's paradox. The two models contradict each other. But the models are not equally plausible: it's silly to leave out age as a risk factor.

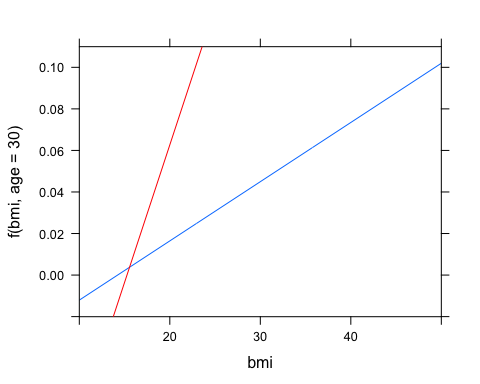
Emory Brown et al. Overcome the fear of getting an answer that can be challenged. Instill the value of being useful.

If students just have statistics and parameters such as the proportion, they are powerless to see how smoking and mortality are related.

IS THIS A NATURAL WAY TO DESCRIBE THE RELATIONSHIP. I submit that it's hard to interpret a graph like this. BETTER to quantify the relationship.

The graph doesn't quantify the relationship, but it does guide decision-making in constructing models: age seems important.

mod <- lm( diab=='diabetic' ~ bmi \* age, data=nhanes)  
f <- makeFun(mod)  
plotFun( f(bmi, age=30)~bmi, bmi.lim=range(10,50))  
plotFun( f(bmi, age=70)~bmi, add=TRUE, col='red')



Contour plot of diabetes probability versus age and bmi.

Do smoking and diabetes.

Then BMI versus smoking.

Diabetes using age, smoking, BMI.

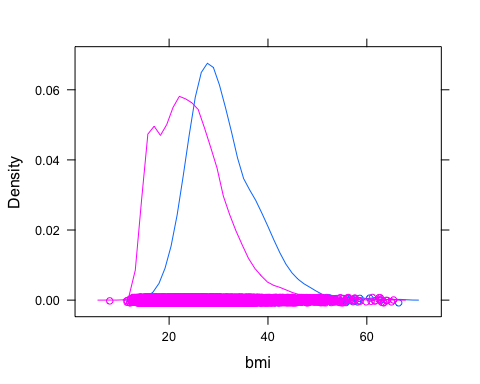
EXERCISE: Break out the children and quantify the relationship between diabetes, bmi, and age

tally( diab ~ bmi>30 & age<20 & age>60, data=nhanes, margins=TRUE )

## , , age < 20 = TRUE, age > 60 = TRUE  
##   
## bmi > 30  
## diab TRUE FALSE  
## diabetic   
## not   
## Total   
##   
## , , age < 20 = FALSE, age > 60 = TRUE  
##   
## bmi > 30  
## diab TRUE FALSE  
## diabetic 0.273145 0.157807  
## not 0.726855 0.842193  
## Total 1.000000 1.000000  
##   
## , , age < 20 = TRUE, age > 60 = FALSE  
##   
## bmi > 30  
## diab TRUE FALSE  
## diabetic 0.007767 0.002807  
## not 0.992233 0.997193  
## Total 1.000000 1.000000  
##   
## , , age < 20 = FALSE, age > 60 = FALSE  
##   
## bmi > 30  
## diab TRUE FALSE  
## diabetic 0.097065 0.033516  
## not 0.902935 0.966484  
## Total 1.000000 1.000000

NOTE IN DRAFT: We don't really want to do the following, since the syntax differs and the plot will be difficult for many readers and it takes up valuable space.

densityplot( ~ bmi, groups=diab, data=nhanes )



# maybe write showDensity() wrapper for package that integrates `densityplot()` into the same notation.

Nir Y. Krakauer, Jesse C. Krakauer, "A New Body Shape Index Predicts Mortality Hazard Independently of Body Mass Index" PloS one, 7 (7) PMID: 22815707, 2012. <http://www.plosone.org/article/info:doi/10.1371/journal.pone.0039504>

**USE the body-shape data from DCF:**

\*\* Or use the Galton data. \*\*

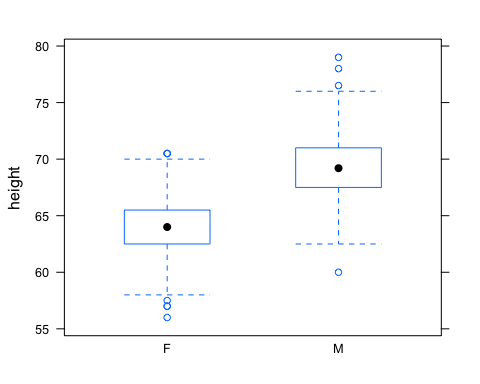
Consider this simple statistical computation in R/mosaic:

mean( height ~ sex, data=Galton )

## F M   
## 64.11 69.23

Means are actually an approach to modeling. The modeling choice here is that height is related to sex. Calculating means constructs a representation of how height relates to sex. Other representations might be of interest, for example this box-and-whiskers plot that shows variability in heights beyond that accounted for by sex:

bwplot( height ~ sex, data=Galton )



Even though means and box-and-whisker plots are very different, the R/mosaic puts front and center the common features:

1. What is the dataset?
2. What variables are to be included in the model and how should they be related?
3. What representation should be made?

Different representations serve different purposes. For instance, the concern might be whether there is more variation in female heights or in male heights. If so, then a representation of spread may be useful:

sd( height ~ sex, data=Galton)

## F M   
## 2.370 2.632

IQR( height ~ sex, data=Galton )

## F M   
## 3.0 3.5

There's nothing fundamentally different when more variables are added.

# mean( height ~ sex + diabetes, data=DCF )

But sometimes a different representation is called for. The lm() representation handles quantitative variables and gives a different view of the sex dependence (deltas rather than absolute levels).

Should there be an interaction term? This is a modeler's choice. One strategy in teaching statistical modeling is to let the students go wild with including terms. This makes it easier to illustrate why a modeler would choose to omit variables. ILLUSTRATE WITH SAT DATA.

MOVE ON TO BEING ABLE TO COMPUTE on the model

* Displaying differences: makeFun() and plotFun()
* Confidence intervals (Example: showing sampling variation in the mean)
* Hypothesis testing

There are some inconsistencies. Sometimes it's necessary to give a new name to a function to avoid interfering with existing function. Example: densityplot(). It's easy enough to define a new function, say, plotDensity() that works as expected. Trade-off between consistency to the notation and fidelity to operations that are widely used outside of mosaic.

Consistent notation for different levels of analysis: the tilde.

Implementation of important statistical operations: randomization and repetition.

Outputs that can

We hope the mosaic package can be as useful for teaching statistics as the many other packages that are so widely used in the display and analysis of data: ggplot, dply, lattice, knitr

## Full Quote

Mr. Fisher has undertaken the very difficult task of attempting to put before research workers [[omit?in biology and agriculture, who are]] without any special mathematical training, a summary covering a great range of methods [[omit? and results in the mathematical theory of statistics]]. The book is chiefly concerned with the best methods of handling small samples [[omit?, and purposely no general attempt is made to supply proofs of the results quoted, but forty-five]] [[omit? examples are worked out fully in the text illustrating different methods of attacking a variety of problems.]] After conscientiously working through the examples, the student should feel able to apply the methods to exactly similar problems [[omit?, but it appears a little doubtful whether without a thorough grasp of the underlying principles he could be safely trusted to tackle other problems where the conditions are somewhat altered.]]