Phenotypes Assessment

Stephen Blatti

November 21, 2017

In the video we talk about phenotypes. Here we show some examples of what we mean by phenotypes, how they can be coded in R objects, and how we compute with them. Later in the course we will perform analyses that statistically connect these phenotypes to measured molecular outcomes. Here we explore the use of data frames to store phenoypes (columns) from several individuals (rows).

```
# library(BiocInstaller)
# biocLite("COPDSexualDimorphism.data")
```

Tabulating gender representation

Install and attach the COPDSexualDimorphism.data package using biocLite. Use the commands

```
library(COPDSexualDimorphism.data)
data(lgrc.expr.meta)
```

to add the object expr.meta to your workspace. The variable pkyrs in the expr.meta data.frame represents pack years smoked. Other variables include gender (self-explanatory) and diagmaj (disease status). What is the number of female participants in this study?:

```
names(expr.meta)
```

```
## [1] "tissueid"    "sample_name" "newid"       "GENDER"      "age"
## [6] "cigever"     "pkyrs"       "diagmaj"     "gender"
```

```
head(expr.meta)
```

```
##     tissueid     sample_name  newid   GENDER age       cigever pkyrs
## 1 LT001098RU LT001098RU_COPD 161745 2-Female  46 2-Ever (>100)    35
## 2 LT001796RU LT001796RU_CTRL 212671   1-Male  48 2-Ever (>100)    19
## 3 LT005419RU LT005419RU_COPD 291396   1-Male  70 2-Ever (>100)    43
## 4 LT007392RU LT007392RU_COPD 169067   1-Male  46 2-Ever (>100)    45
## 5 LT009615LU LT009615LU_CTRL  49801 2-Female  49 2-Ever (>100)    45
## 6 LT010491LL LT010491LL_COPD 180409   1-Male  78 2-Ever (>100)    51
##            diagmaj   gender
## 1 2-COPD/Emphysema 2-Female
## 2        3-Control   1-Male
## 3 2-COPD/Emphysema   1-Male
## 4 2-COPD/Emphysema   1-Male
## 5        3-Control 2-Female
## 6 2-COPD/Emphysema   1-Male
```

```
table(expr.meta$gender)

##
##   1-Male 2-Female
##      119      110
```

Descriptive statistics on smoking

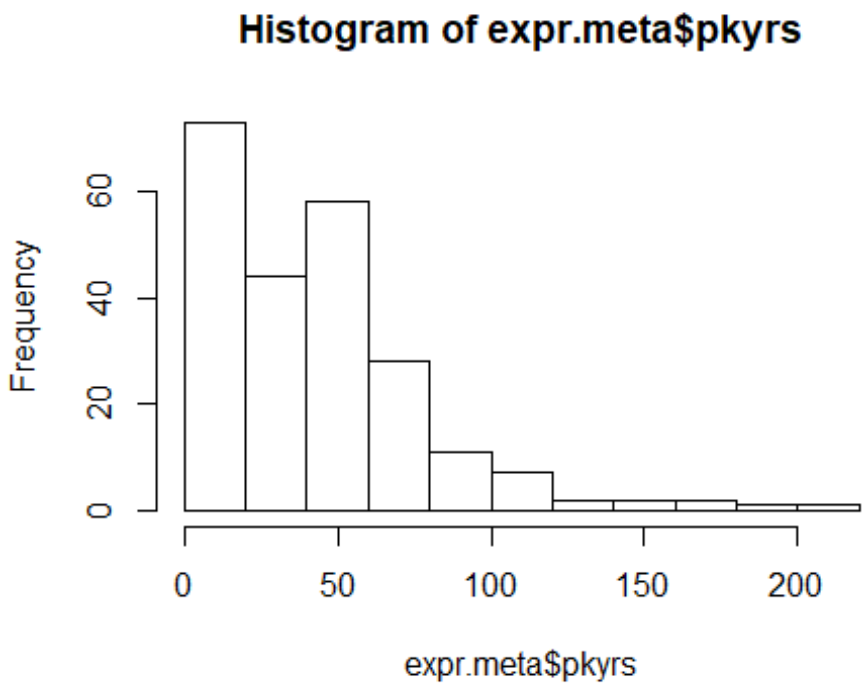
What is the median of the distribution of pack years smoked in this cohort (women and men)?

```
median(expr.meta$pkyrs)

## [1] 40

summary(expr.meta$pkyrs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   15.00   40.00   44.17   60.00  212.00
```

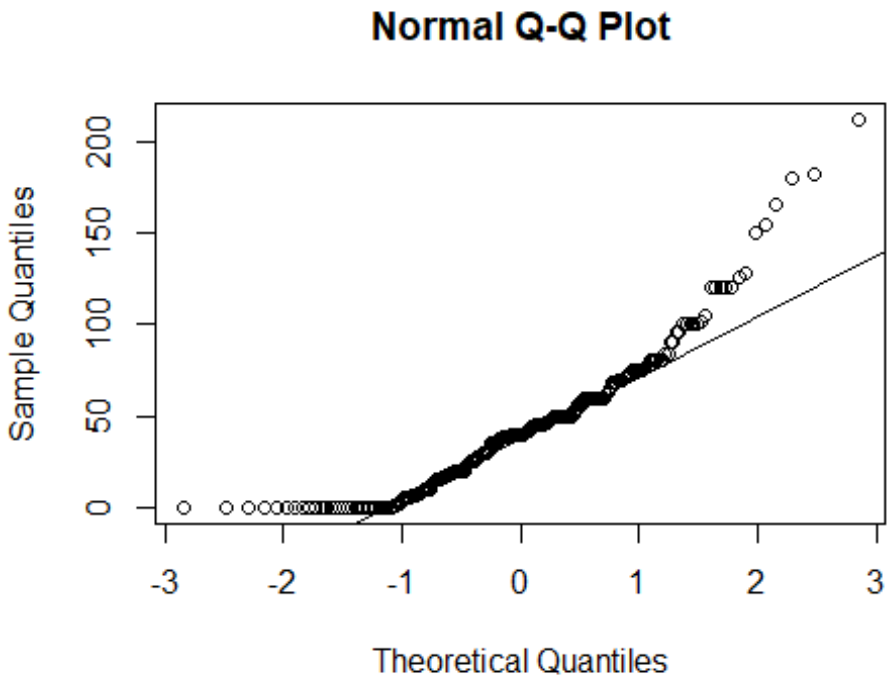
EDA on distributional modeling

True or False: The distribution of pack-years smoked is well-approximated by a Gaussian (Normal) probability distribution. Select one:

```
TRUE
FALSE

hist(expr.meta$pkyrs)
```




```
qqnorm(expr.meta$pkyrs)
qqline(expr.meta$pkyrs)
```



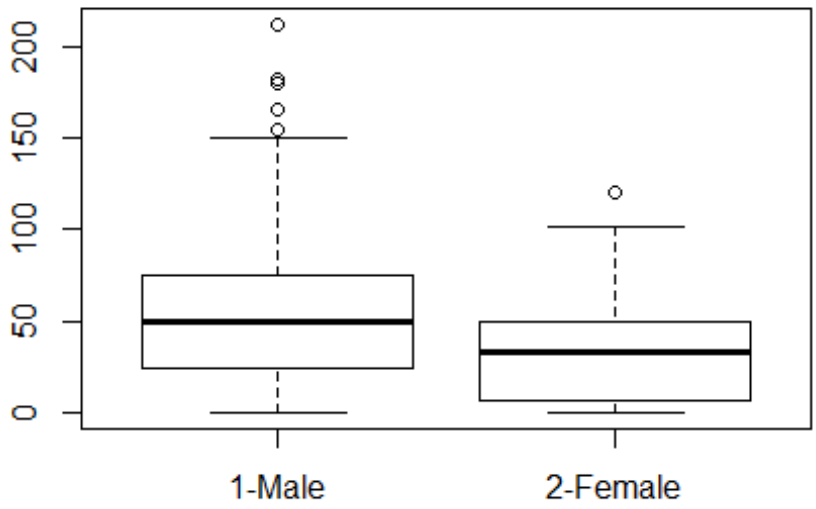
False. The substantial number of zero values renders a Gaussian model for pack years formally implausible.

Exploratory comparison of distributions

The units with which a quantity is recorded are not always the most effective for statistical analysis.

Use the command boxplot(pkyrs~gender, data=expr.meta) to examine the distributions of pack years by gender using the boxplot statistics and outlier flagging. Which of the following is an aspect of the display that would suggest caution in using the t test in comparing males and females with respect to pack years smoked?

```
Median values are quite different for the two groups.
More outliers flagged for men than for women.
Distributions appear quite asymmetric, with long tails skewed towards high
values. C
Lower quartile for females is close to zero.

boxplot(pkyrs~gender, data=expr.meta)
```



Explantion: All of the propositions among the choices are true, but validity of the t test is compromised when applied to variables with skewed distributions. The test's validity does not depend on separation of median values or magnitude of quartile, so first and last choices are incorrect. Difference in number of outliers flagged can be important, but the more systematic concern with asymmetry of distributions is the target response for this problem.

Variable transformation

Use the code expr.meta$pyp1 = expr.meta$pkyrs+1 to define a positive-valued variable for transformation analysis.

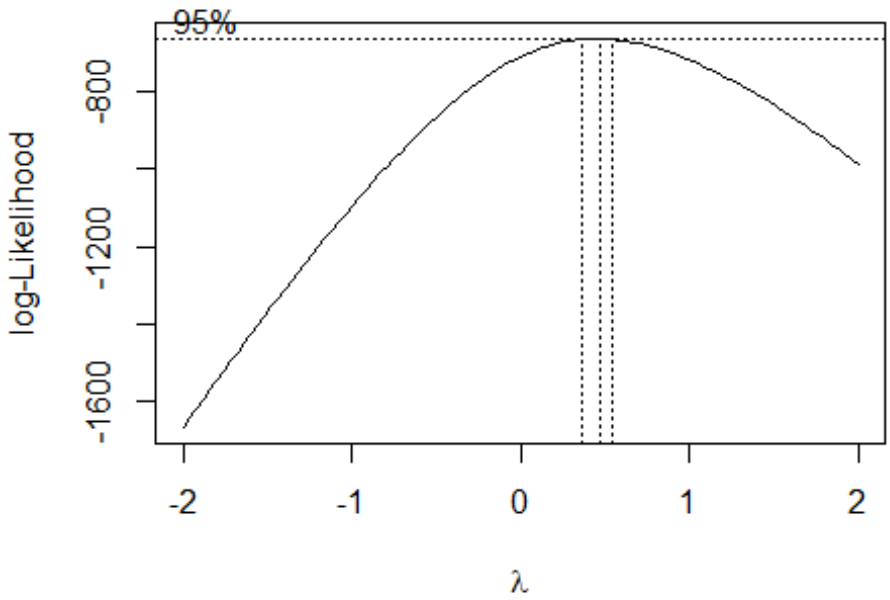
Then load the MASS package (library(MASS)) and fit the model lm1 = lm(pyp1~gender, data=expr.meta) that tests for a difference in mean pack years (plus 1) between genders.

Finally, use boxcox(lm1) to see a plot of the likelihood function for a transformation model. Under this model we use a number denoted lambda that for our purposes is used as an exponent to transform the dependent variable pyp1 of the regression to have a distribution that is approximately Gaussian. Thus, if lambda is 1, we use pyp1 untransformed, if lambda is 0.5, we use sqrt(pyp1), and so on.

For what value of lambda does the likelihood reach its highest value for the model lm1?

```
0
0.333
0.5
1
```

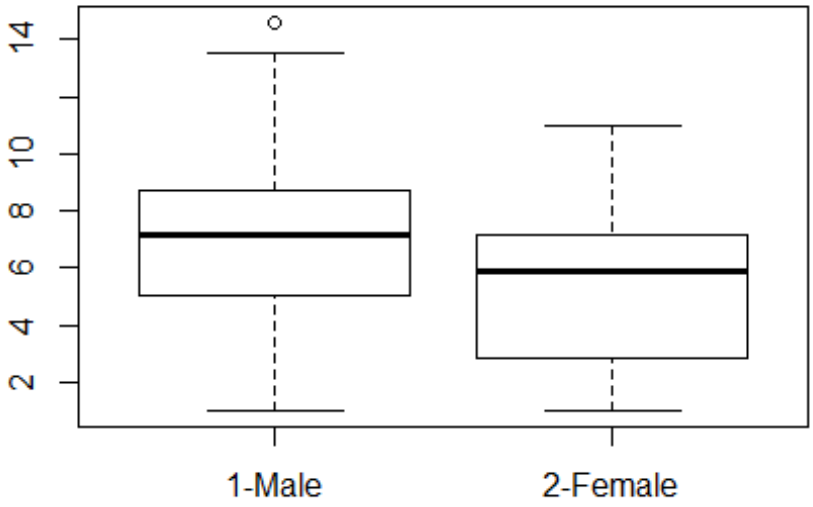
```
expr.meta$pyp1 = expr.meta$pkyrs+1
library(MASS)
lm1 = lm(pyp1~gender, data=expr.meta)
boxcox(lm1)
```



Once you have read the plot to obtain the value of lambda at which the transformation model likelihood is maximized, use the code boxplot(I(pyp1^lambda)~gender, data=expr.meta) to see the effects of the transformation on symmetry and presence of outliers.

Transformations with similar intent will be important in various aspects of statistical analysis of genome-scale data.

```
lambda = 0.5
boxplot(I(pyp1^lambda)~gender, data=expr.meta)
```


14
10
8
6
4
2
1-Male
2-Female