edX

Phenotypes Assessment

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).
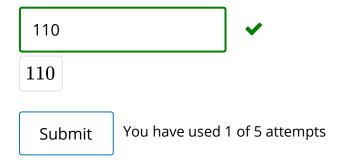
Tabulating gender representation

1/1 point (graded)
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?:

110

✔

110

Submit   You have used 1 of 5 attempts

Descriptive statistics on smoking

1/1 point (graded)
What is the median of the distribution of pack years smoked in this cohort (women and men)?

40

✔

40

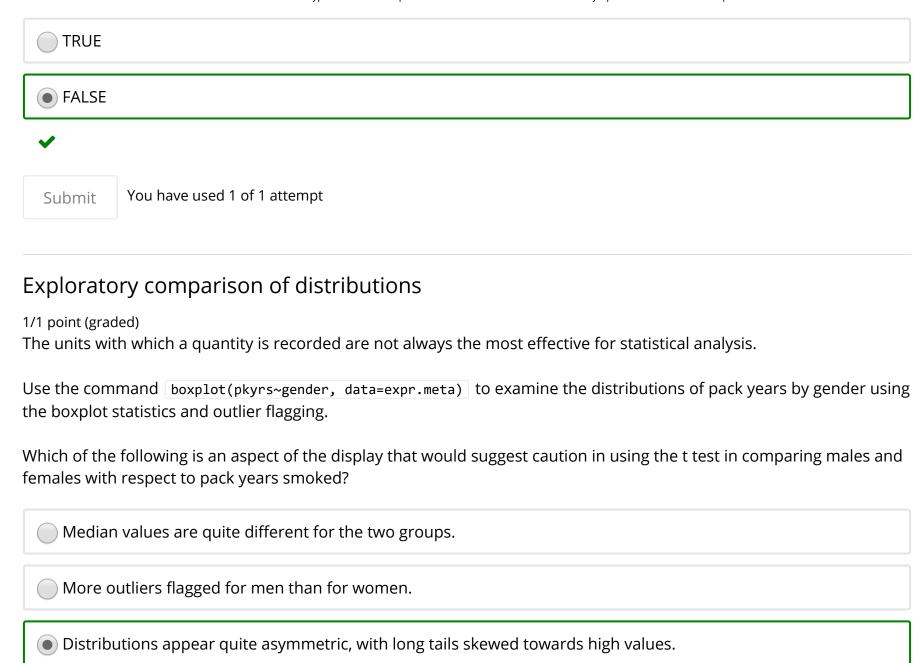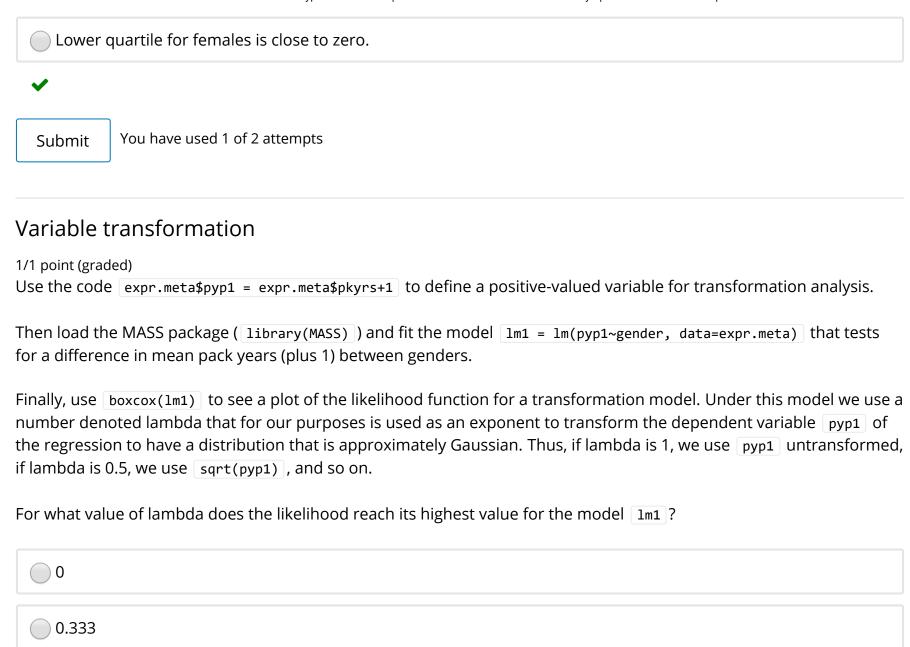Submit   You have used 1 of 5 attempts

---

EDA on distributional modeling

1/1 point (graded)
True or False: The distribution of pack-years smoked is well-approximated by a Gaussian (Normal) probability distribution.

Select one:

○ TRUE

◉ FALSE

✔

| Submit | You have used 1 of 1 attempt |

Exploratory comparison of distributions

1/1 point (graded)
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.

○ Lower quartile for females is close to zero.

✔

Submit   You have used 1 of 2 attempts

Variable transformation

1/1 point (graded)

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

✔

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.

Submit   You have used 1 of 2 attempts