Essential Statistics with R: Exercises

Exercise set 1

1. What's the mean 60-second pulse rate for all participants in the data?

```
## [1] 73.63382
```

2. What's the range of values for diastolic blood pressure in all participants? (Hint: see help for `min()`, `max()`, and `range()` functions, e.g., enter `?range` without the parentheses to get help).

```
## [1]   0 116
```

3. What are the median, lower, and upper quartiles for the age of all participants? (Hint: see help for `median`, or better yet, `quantile`).

```
##   0%  25%  50%  75% 100%
##    0   17   36   54   80
```

4. What's the variance and standard deviation for income among all participants?

```
## [1] 1121564068
```

```
## [1] 33489.76
```

Exercise set 2

1. Is the average BMI different in single people versus those in a committed relationship? Perform a t-test.
2. The `Work` variable is coded "Looking" (n=159), "NotWorking" (n=1317), and "Working" (n=2230). Examine how this variable is related to `Income`.
   a. Fit a linear model of `Income` against `Work`. Assign this to an object called `fit`. What does the `fit` object tell you when you display it directly?
   b. Run an `anova()` to get the ANOVA table. Is the model significant?
   c. Run a Tukey test to get the pairwise contrasts. (Hint: `TukeyHSD()` on `aov()` on the fit). What do you conclude?
   d. Instead of thinking of this as ANOVA, think of it as a linear model. After you've thought about it, get some `summary()` statistics on the fit. Do these results jive with the ANOVA model?
3. Examine the relationship between HDL cholesterol levels (`HDLChol`) and whether someone has diabetes or not (`Diabetes`).
   a. Is there a difference in means between diabetics and nondiabetics? Perform a t-test *without* a Welch correction (that is, assuming equal variances – see `?t.test` for help).
   b. Do the same analysis in a linear modeling framework.
   c. Does the relationship hold when adjusting for `Weight`?
   d. What about when adjusting for `Weight`, `Age`, `Gender`, `PhysActive` (whether someone participates in moderate or vigorous-intensity sports, fitness or recreational activities, coded as yes/no). What is the effect of each of these explanatory variables?

Exercise set 3

1. What's the relationship between diabetes and participating in rigorous physical activity or sports?
    a. Create a contingency table with Diabetes status in rows and physical activity status in columns.
    b. Display that table with margins.
    c. Show the proportions of diabetics and nondiabetics, separately, who are physically active or not.
    d. Is this relationship significant?
    e. Create two different visualizations showing the relationship.
2. Model the same association in a logistic regression framework to assess the risk of diabetes using physical activity as a predictor.
    a. Fit a model with just physical activity as a predictor, and display a model summary.
    b. Add gender to the model, and show a summary.
    c. Continue adding weight and age to the model. What happens to the gender association?
    d. Continue and add income to the model. What happens to the original association with physical activity?

Exercise set 4

1. You're doing a gene expression experiment. What's your power to detect a 2-fold change in a gene with a standard deviation of 0.7, given 3 samples? (Note - fold change is usually given on the log_2 scale, so a 2-fold change would be a `delta` of 1. That is, if the fold change is 2x, then $log_2(2) = 1$, and you should use 1 in the calculation, not 2).

```
## [1] 0.2709095
```

2. How many samples would you need to have 80% power to detect this effect?

```
## [1] 8.764711
```

3. You're doing a population study looking at the effect of a SNP on disease X. Disease X has a baseline prevalence of 5% in the population, but you suspect the SNP might increase the risk of disease X by 10% (this is typical for SNP effects on common, complex diseases). How many samples do you need to have 80% power to detect this effect, given that you want a statistical significance of $p < 0.001$?

```
## [1] 67946.82
```