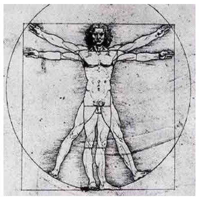

Human Proportions

1 The shape of body part measurements

The human body comes in various shapes and sizes. However, as daVinci knew, there are certain proportions that are consistent throughout. For this project two data sets are used which contain various measurements of human bodies.

To download the data sets issue these commands:

```
> source("http://www.math.csi.cuny.edu/st/R/normtemp.R")
> source("http://www.math.csi.cuny.edu/st/R/fat.R")
```

The `normtemp` data set[1] contains measurements of normal body temperature for 300 healthy adults in the variable `temperature`. The variable `gender` records the gender of the subject, and `hr` the heart rate in beats per minute.

The `fat` data set[2] contains many measurements of human bodies that can be done with a tape measure (circumference measurements), for instance the variable `wrist` contains measurements of wrist size in centimeters. Additionally, the variable `body.fat` contains body fat measurements.

After downloading the data sets, they may be attached so that the variable names are visible from the command line.

```
> attach(normtemp)
> attach(fat)
```

1.1 two-sample tests

Is there a statistically significant difference between the body temperatures for males and females? Let μ_m be the population mean for males, and μ_f be the population mean for females. If we assume that the two populations are normally distributed, then a significance test of

$$H_0 : \mu_1 = \mu_2, \qquad H_A : \mu_1 \neq \mu_2$$

(or $H_A : \mu_1 < \mu_2$, or $H_A : \mu_1 > \mu_2$) can also be done using a t-distributed statistic.

The `t.test()` will still perform this test. In the two-sample case, the null hypothesis is not specified as it is always the same.

For example, the two-sample test of equivalence of means for the temperature data can be performed as follows

[1]This data set was contributed to the *Journal of Statistical Education* by Allen L. Shoemaker, `http://www.amstat.org/publications/jse/v4n2/datasets.shoemaker.html`

[2]This data set was contributed to the *Journal of Statistical Education* by Roger W. Johnson, `http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html`.

```
> males = temperature[gender == "Male"]
> females = temperature[gender == "female"]
> t.test(males, females, alt = "two.sided")


        Welch Two Sample t-test

data:  males and females
t = -2.2854, df = 127.51, p-value = 0.02394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

The small p-value indicates that there is a statistically significant difference.

Question 1: The normality of the two populations (sampled in `males` and `females`) was not verified. Comment on the validity of this assumption. (A large sample size is also sufficient to use the t-test.)

Question 2: If the population variances can be assumed to be equal, then the extra argument `var.equal=TRUE` to `t.test()` will generally give smaller p-values, as the sampling distribution of the test statistic has smaller degrees of freedom in most cases. First check if this is a reasonable assumption about our data, and then check if it makes a difference in the t-test.

An alternate syntax (the model formula syntax) for performing a two-sample test when there is one variable (`gender`) which is used to indicate which level of some factor the subject has may be used. (These variables are referred to as *indicator variables.*) The t-test above, may have been carried out with the syntax:

```
> t.test(temperature ~ gender)
```

Question 3: Verify that using the model formula gives the same output and also that the default value for `alt=` is `"two.sided"`.

Question 4: Perform a two-sample t-test of the resting heart rate (`hr`) for males and females. Is the difference statistically significant at the $\alpha = 0.10$ level?

One can define indicator variables to break up a data set. For instance, a test to see if age has an effect on body fat percentage could be done with

```
> over40 = age > 40
> t.test(body.fat ~ over40)
```

Question 5: Perform the t-test described above.

1. What is the p-value reported?
2. What is the null and alternative hypotheses tested?
3. You can readily check the assumptions on the population by creating side-by-side boxplots. These are created with the same model formula syntax:

```
> boxplot(body.fat ~ over40)
```

Question 6: The age 40 splits the data on weights (in `weight`) into two groups. Perform a significance test of

$$H_0 : \mu_{\text{40 or under}} = \mu_{\text{over 40}}, \qquad H_A : \mu_{\text{40 or under}} < \mu_{\text{over 40}}$$

where each μ is, as usual, the population mean for the population the sample comes from.

1. What is the p-value?
2. Justify you choice of test statistic.

Question 7: Are taller people less susceptible to high body fat? Define an indicator variable as follows:

```
> tall = height > 72
```

Using this, perform a significance test of

$$H_0 : \mu_{\text{6 feet or under}} = \mu_{\text{over 6 feet}}, \qquad H_A : \mu_{\text{6 feet or under}} < \mu_{\text{over 6 feet}}$$

where μ is the population mean of the body fat for the respective populations. (The data is in the variable `body.fat`.)

1. What is the p-value of the significance test?
2. Justify you choice of test statistic.

The function `wilcox.test()` for two samples performs a Wilcoxon rank-sum test of the equivalence of medians. The assumption on the populations is that they have the same shape, although perhaps different medians. They need not be normally distributed. As the population of weights is typically long-tailed, it is often best to compare sub-populations using this test instead of the t-test.

Question 8: Repeat the significance test investigating if age and weight are somehow related using the Wilcoxon rank-sum test. Compare your p-value to that returned by a t-test for the same data.

Question 9: Does a person's density (recorded in `density`) decrease with age? Use the `old` indicator variable to investigate this significance test

$$H_0 : M_{40 \text{ or under}} = M_{\text{over } 40}, \qquad H_A : M_{40 \text{ or under}} < M_{\text{over } 40},$$

where M is the subpopulation median of densities.

1. Decide if the Wilcoxon rank-sum test is appropriate for this data set. If so, compute the p-value.
2. Decide if the t-test is appropriate for this data set (where for a symmetric population the population mean and median are identical). If so, compute the p-value.
3. If both tests are appropriate, compare the p-values. Are they similar? Different?