

Week 3

Shu Guo

Thursday, February 05, 2015

Population, Samples, and Estimates

Introduction

we are now ready to describe the mathematical theory that permits us to compute p-values in practice. We will also learn about confidence intervals and power calculations.

Population parameters

In the mouse weight example, we have two populations; female mice on control diet and female mice on high fat diet, and the outcome of interest was weight. We consider this population to be fixed, and the randomness comes from the sampling. One reason we have been using this dataset as an example is because we happen to have the weights of all the mice of this type. Read the data:

We can then access the population values and determine, for example, how many we have. Here is the control population:

```
controlPopulation <- dat[dat$Sex == "F" & dat$Diet == "chow", 3]  
length(controlPopulation)
```

```
## [1] 225
```

Denote these values as x_1, \dots, x_m . In this case $m = 225$. Now we can do the same with the high fat diet population

```
hfPopulation <- dat[dat$Sex == "F" & dat$Diet == "hf", 3]  
length(hfPopulation)
```

```
## [1] 200
```

and denote with $y_1, \dots, y_n, n = 200$.

Define summaries of interest for these population such as the mean and variance.

the mean:

$$\mu_X = \frac{1}{m} \sum_{i=1}^m x_i \text{ and } \mu_Y = \frac{1}{n} \sum_{i=1}^n y_i$$

the variance:

$$\sigma_X^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_x)^2 \text{ and } \sigma_Y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mu_y)^2$$

with the standard deviation being the square root of the variance. We refer to such quantities, that can be obtained from the population, as *population parameters*.

The question we started out asking can now be written mathematically: $\mu_Y - \mu_X = 0$? We take a sample and try to answer the questions with the sample. This is the essence of statistical inference.

Sample estimates

In the previous section, we obtained samples of 12 mice from each population. We represent these with capital letters to indicate that they are random. This is common practice in statistics, although it is not always followed. So the samples are X_1, \dots, X_M and Y_1, \dots, Y_N and in this case $N = M = 12$. Since we want to know what $\mu_Y - \mu_X$ is we consider the sample version: $\bar{Y} - \bar{X}$ with

$$\bar{X} = \frac{1}{M} \sum_{i=1}^M X_i \text{ and } \bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

Here we described the mathematical theory that mathematically relates \bar{X} to μ_X and \bar{Y} to μ_Y , which will in turn help us understand the relationship between $\bar{Y} - \bar{X}$ and $\mu_Y - \mu_X$.

Central Limit Theorem in practice