Practice Exercise for PGPBABI-SMDM

Solution:

EXAMPLE 1 : SOLUTION

In this section of the tutorial, we will compare these Hyades stars with the remaining stars in the Hipparcos dataset on the basis of the color (B minus V) variable. That is, we are comparing the groups in the boxplot below:

```
   hip <- read.table("http://astrostatistics.psu.edu/datasets/HIP_star.dat",
      header=T, fill=T)
#   hip <- read.table("HIP_star.dat", header=T, fill=T)
   attach(hip)
   filter1 <-  (RA>50 & RA<100 & DE>0 & DE<25)
   filter2 <- (pmRA>90 & pmRA<130 & pmDE>-60 & pmDE< -10)
   filter  <-  filter1 & filter2 & (e_Plx<5)
   sum(filter)

   color <- B.V
   boxplot(color~filter,notch=T)
```

For ease of notation, we define vectors H and nH (for "Hyades" and "not Hyades") that contain the data values for the two groups.

```
   H <- color[filter]
   nH <- color[!filter & !is.na(color)]
```

In the definition of nH above, we needed to exclude the NA values.

A two-sample t-test may now be performed with a single line:

```
   t.test(H,nH)
```

Because it is instructive and quite easy, we may obtain the same results without resorting to the t.test function. First, we calculate the variances of the sample means for each group:

```
   v1 <- var(H)/92
   v2 <- var(nH)/2586
   c(var(H),var(nH))
```

The t statistic is based on the standardized difference between the two sample means. Because the two samples are assumed independent, the variance of this difference equals the sum of the individual variances (i.e., v1+v2). Nearly always in a two-sample t-test, we wish to test the null hypothesis that the true difference in means equals zero. Thus, standardizing the difference in means involves subtracting zero and then dividing by the square root of the variance:

```
tstat <- (mean(H)-mean(nH))/sqrt(v1+v2)
tstat
```

To test the null hypothesis, this t statistic is compared to a t distribution. In a Welch test, we assume that the variances of the two populations are not necessarily equal, and the degrees of freedom of the t distribution are computed using the so-called Satterthwaite approximation:

```
(v1 + v2)^2 / (v1^2/91 + v2^2/2585)
```

The two-sided p-value may now be determined by using the cumulative distribution function of the t distribution, which is given by the pt function.

```
2*pt(tstat,97.534)
```

Incidentally, one of the assumptions of the t-test, namely that each of the two underlying populations is normally distributed, is almost certainly not true in this example. However, because of the central limit theorem, the t-test is robust against violations of this assumption; even if the populations are not roughly normally distributed, the sample means are.

In this particular example, the Welch test is probably not necessary, since the sample variances are so close that an assumption of equal variances is warranted. Thus, we might conduct a slightly more restrictive t-test that assumes equal population variances. Without going into the details here, we merely present the R output:

```
t.test(H,nH,var.equal=T)
```