Some Examples

Sample Mean and Sample Variance are Independent under Normality

Suppose that $X_1, \ldots, X_n$ are independent normal random variables with mean μ and variance $sigma^2$. Define the sample mean and the sample variance by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

respectively. It is well known that $\bar{X}$ and S^2 are independent. Confirm this fact by simulation using S-Plus.

Here is a simple functions that generates values from $\bar{X}$ and S^2.

```
sample.mean.var   <- function(mu, sigma, sim, sample.size)
{
xbar     <- rep(NA, sim)
Ssquare  <- rep(NA, sim)
for (i in 1:sim)
{
sample     <- rnorm(sample.size, mean=mu, sd=sigma)
xbar[i]    <- mean(sample)
Ssquare[i] <- var(sample, unbiased=T)
}
return(cbind(xbar, Ssquare))
}
```

The output of this function is a matrix with `sim` rows and 2 columns. Hence a plot of the first column against the second shows that the two variables are independent. Similar results are reached by calculating the correlation of the two columns or by testing whether or not there is correlation (use the function `cor` and `cor.test` respectively). However it should be noted that if the correlation coefficient between two random variables is close to 0 then the variables are not necessarily independent unless they are normally distributed.

Data Manipulation

Create a vector x with elements equal to 0 to 2 in increments of .01. Compute the value for all values of x and plot x versus y, where $y = \sqrt{x^2 + x + 4}$. Obtain the values of x for which y < 2.5.

The following simple functions lead to the result.

```
> x <- seq(0,2, by=0.01)
> y <- sqrt(x^2+x+4)
> plot(x,y,type="l")
> x[y < 2.5]
```

Data description and simple inference: IQ scores of children of depressed and non–depressed women

Description of Data

These data consist of the IQ scores of children of age five, labeled according to whether or not their mother have suffered an episode of post–natal depression. We focus on answering the question whether or not the two groups of children have different IQ scores.

Significance tests for assessing differences between groups

Most commonly a t–test is used to assess the hypothesis that the two groups have the same population mean, the most appropriate alternative hypothesis usually being that they do not. The test assumes that the observations:

1. are independent of one another;
2. come from populations having normal distribution
3. come from populations having the same variance

In addition to the p–value obtained from a significance test, a confidence interval for the difference in the two means is generally required.

Data Analysis

Before applying a significance test an attempt must be made to check that the assumptions on which the test are based are satisfied by the data under investigation. An initial assessment is usually based on some simple plots: histograms, boxplots and probability plots are particularly useful for indicating departures from normality, the presence of outliers, etc. The resulting diagrams are shown in Figures 1 to 3.

The most obvious feature of all the diagrams is that both groups of IQ scores contain a clear outlier or "wild" observation, corresponding in each case to a child with very low IQ. Such observations can greatly distort summary measures such as means and variances and are also likely to affect significance tests based on the assumption of normality. For the moment, however, suppose no action is taken about the two outliers, and a t-test is applied to the IQ scores. This results in the information given below:

```
> t.test(scorend,scored)

    Standard Two-Sample t-Test

data:  scorend and scored
t = 2.4637, df = 92, p-value = 0.0156
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.27152 21.16477
sample estimates:
 mean of x mean of y
  112.7848  101.0667
```

The difference in mean IQ in the two groups of children is highly significant, and the confidence interval quantifies this difference. However, the presence of outliers makes this result suspect. The variances in each group, for example, are 205.50 (children of non–depressed mothers) and 729.21 (children of depressed mothers). There certainly appears to be a considerable difference in the variance of the IQ scores in the two groups, thus violating one of the assumptions of the previously performed t-test. A formal test of hypothesis that the population variances are equal can be made and the result is:

```
> var.test(scorend,scored)

    F test for variance equality

data:  scorend and scored
F = 0.2818, num df = 78, denom df = 14, p-value = 0.0003
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1089903 0.5746361
sample estimates:
 variance of x variance of y
      205.5044      729.2095
```

The hypothesis that of the equality of the two variances must clearly be rejected. But what happens if the test is repeated with the outlying observation in each group removed? We can apply the variance test again, but now excluding the two children's with the very low IQs:

```
    F test for variance equality

data:  scorend[scorend > 50] and scored[scored > 50]
F = 0.5664, num df = 77, denom df = 13, p-value = 0.1283
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2102836 1.1765891
sample estimates:
 variance of x variance of y
       152.967      270.0659
```

This gives the two variances as 152.97 and 270.07, and the F-test for the equality hypothesis is non-significant. So removing the two outliers makes the data more appropriate for the t-test; it also makes sense in the context of this particular study since one of the children involved was autistic, and the other brain–damaged at birth. The t-test can be repeated without the outliers:

```
Standard Two-Sample t-Test

data:  scorend[scorend > 50] and scored[scored > 50]
t = 1.8242, df = 90, p-value = 0.0714
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -0.6148208  14.4170186
sample estimates:
 mean of x mean of y
  113.6154  106.7143
```

It is clear that without the two outliers the evidence for a difference in the average IQs of children of depressed and non–depressed mothers diminishes.

To perform the analysis in S-plus as was described above use the following by noticing first that the data were entered through the `Import` facility and the column names were V1 (for d=depressed and nd=non-depressed) and V2 for the IQ score.

```
> attach(iqdata)
> scorend <- iqdata$V2[V1=="nd"]
> scored <- iqdata$V2[V1=="d"]
> boxplot(scorend, scored, names=c("Non-depressed mothers","Depressed mothers"))
> par(mfrow=c(2,1))
> hist(scorend, xlab="IQ of children from Non-depressed mothers")
> hist(scored, xlab="IQ of children from depressed mothers")
> qqnorm(scorend)
> qqline(scorend)
> title(main="Normal Probability plot for IQ of children
```

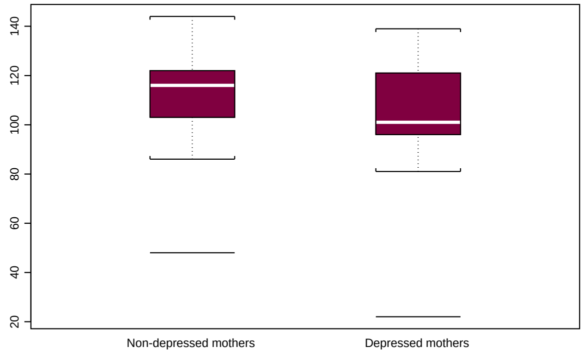

Figure 1: Boxplots of children's IQ scores.

```
Continue string: from Non-depressed mothers")
> qqnorm(scored)
> qqline(scored)
> title(main="Normal Probability plot for IQ of children
Continue string: from dpressed mothers")
```

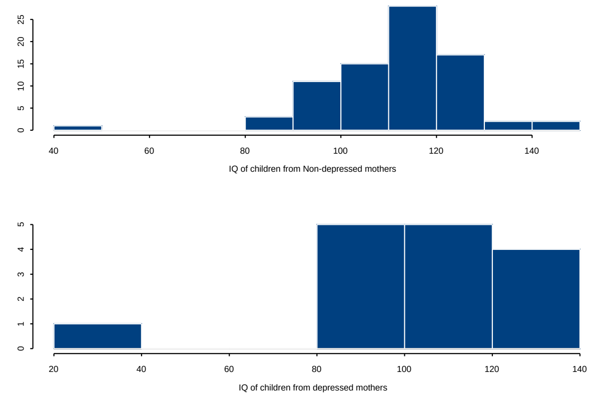

Figure 2: Histograms of children’s IQ scores

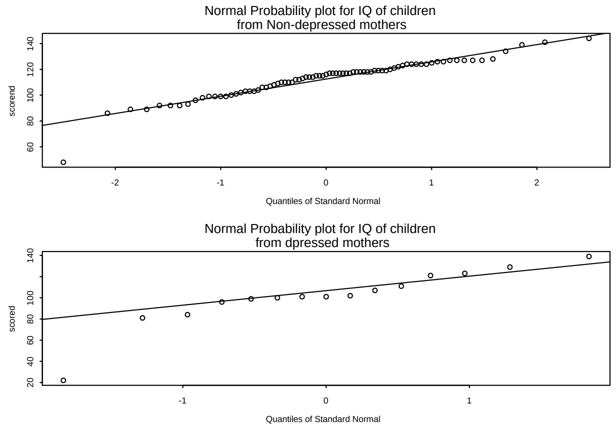

Figure 3: Normal probability plots of children’s IQ scores