6.9 Appendix: Conducting Tests in R

by EV Nordheim, MK Clayton & BS Yandell, October 23, 2003

R has a wide range of commands suited to testing. Here we focus on testing the population mean μ and the population variance σ^2 for normal data, and the population proportion p for binomial data. Section 6.9.1 shows examples of the T-test and Z-test under a variety of settings, depending on what you have available. Section 6.9.2 shows how to examine the distribution of p-values for the `t.test` with simulated data. Section 6.9.3 shows how to test the binomial proportion. Finally, Section 6.9.4 shows how to test the population variance for normal data.

6.9.1 Tests for Population Mean with Normal Data

The T-test introduced in this chapter tests a population mean for normal data when the variance is unknown. The `t.test` can be used for tests of the mean when you have all the data. `pt` and `qt` are useful to get probabilities and quantiles, respectively, when you only have the sample mean and sample variance. When the population variance is known, we can use the `pnorm` command in a manner similar to Appendix 4.4 to conduct tests on the population mean.

We demonstrate the use of this command for the white pine seedling data originally introduced in Chapter 2. In this chapter, we discussed hypothesis testing for these data in Subsection 6.3.2. Of interest was the null hypothesis $H_0 : \mu = 40$. In that subsection there were two white pine seedling data sets, collected from two different nurseries. For purposes of illustration we focus on data from the first nursery (Set A).

```
> whitepine = c(61, 17, 38, 32, 30, 38, 25, 46, 38, 27, 43, 31,
+     34, 41, 27, 22, 40, 22)
```

We enter the `t.test` command, specifying the null hypothesis of interest.

```
> t.test(whitepine, mu = 40)

        One Sample t-test

data:  whitepine
t = -2.4258, df = 17, p-value = 0.02669
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 28.78161 39.21839
sample estimates:
mean of x
       34
```

In Subsection 6.3.2 we concluded that the p-value was between 0.02 and 0.05. Using R, we have been able to specify the p-value more precisely as 0.027. Note that this additional precision does not change our conclusions about the strength of evidence against H_0. However, using the `t.test` command has saved us some work.

The `t.test` command has an option for conducting a one-sided test. Although we have argued that two-sided tests arise more frequently, there are occasions when one-sided tests are of interest. The R command for getting the left-sided test is

```
> t.test(whitepine, mu = 40, alternative = "less")

        One Sample t-test

data:  whitepine
t = -2.4258, df = 17, p-value = 0.01335
alternative hypothesis: true mean is less than 40
95 percent confidence interval:
     -Inf 38.30272
sample estimates:
mean of x
       34
```

The right-sided test uses option `alternative="greater"`, giving a p-value of 0.987

T-test using only sample mean and variance

Suppose we only knew that $\bar{y} = 34$ and $s = 10.49$ for the white pine seedling data. How could we conduct a T-test that $\mu = 40$? We can use the `pt` command. The command structure is similar to previous commands discussed in Appendices 4.4 (`pnorm`, `qnorm`) and 5.7.3 (`pchisq`, `qchisq`). For the white pine seedling T-test, first construct the T-value, $t = (\bar{y} - 40)/(s/\sqrt{n})$ = -2.43.

```
> ybar = mean(whitepine)
> ybar

[1] 34

> s = sd(whitepine)
> s

[1] 10.49370

> n = 18
> mu = 40
> t.value = (ybar - mu)/(s/sqrt(n))
> t.value
```

```
[1] -2.425823
```

Then look up the upper tail for the absolute value of t and double it to get the two-sided p-value. Here we use the `abs` command just to get the absolute value first, removing the minus sign to make sure we are looking in the extremes of the upper tail.

```
> 2 * pt(abs(t.value), n - 1, lower.tail = F)
```

```
[1] 0.02669288
```

Suppose you wanted instead to find the `t.value`, t, corresponding to the upper $\alpha = 0.10$ level, $P(T > t) = 0.10$, for the white pine problem. To find t you would use the `qt` command, much as in Appendix 4.4, and enter

```
> alpha = 0.1
> qt(alpha, n - 1, lower.tail = F)
```

```
[1] 1.333379
```

Z-test with known variance

R does not have a command for the significance testing of the mean of normal data with known variance. Instead we can use the `pnorm` command. Suppose we believed the variance of the white pine seedling heights was 100, hence the variance of the mean is 100/18. The left-sided p-value is found by

```
> pop.var = 100
> pnorm(ybar, mu, sqrt(pop.var/n))
```

```
[1] 0.005454749
```

The two-sided p-value is just double this, or 0.011.

6.9.2 Sampling Distribution of the p-value from T-tests

What is the distribution of the p-value for a test of mean when the null hypothesis is true? Suppose we draw samples of size 18 of simulated white pine seedlings from $N(40, 100)$ and suppose we use a T-test to find evidence whether $\mu = 40$ or not.

Line 1 of the simulation below is like Appendix 5.7.2, drawing 200 samples of size 18, one sample per column. The second line creates a command `get.p.value` to get the p-value (element `p.value`) from a `t.test` that the mean is `mu=40`. The third line uses `apply` to get the p-value by column from the `draws`. Finally, the last line finds out how many p-values are at or below 0.05.

```
> n.draw = 200
> alpha = 0.05
> draws = matrix(rnorm(n.draw * n, mu, sqrt(pop.var)), n)
> get.p.value = function(x) t.test(x, mu = mu)$p.value
> pvalues = apply(draws, 2, get.p.value)
> sum(pvalues <= alpha)

[1] 8
```

6.9.3 Test of Binomial Proportion

In addition, the prop.test command can be used for tests of proportions. The R command for directly testing hypotheses involving binomial data is prop.test. This uses the continuity correction, which we illustrate with the plant screening experiment.

```
> y = 83
> n = 100
> p = 0.75
> prop.test(y, n, p)

        1-sample proportions test with continuity correction

data:  y out of n, null probability p
X-squared = 3, df = 1, p-value = 0.08326
alternative hypothesis: true p is not equal to 0.75
95 percent confidence interval:
 0.7389130 0.8950666
sample estimates:
   p
0.83
```

6.9.4 Test of Population Variance for Normal Data

R commands for testing hypotheses about σ^2 for normal data can be calculated using the methods of this chapter and the probability commands described in the Appendix to Chapter 5 to find the p-value. For instance, a test whether $H_0 : \sigma^2 = 100$ vs. $H_A : \sigma^2 > 100$ for the white pine data has the following p-value:

```
> n = 18
> sample.var = var(whitepine)
> sample.var

[1] 110.1176
```

```
> pop.var = 100
> pchisq((n - 1) * sample.var/pop.var, n - 1, lower.tail = FALSE)

[1] 0.3448386
```