

Math 4397/6397

Problem Set 8, due Thu, Nov 5, 2009

Solutions

Problem 1. Let μ_d be the expected value for the difference between initial and follow-up FEV. Then we test $H_0 : \mu_d = 0$ versus $H_a : \mu_d < 0$ (one-sided) with $\alpha = 0.05$.

The test statistic is

$$\frac{\bar{x}_d - 0}{S_d/\sqrt{n_d}} = \frac{-0.147}{0.15/\sqrt{10}} = -3.1$$

which gives a p -value of $\text{pt}(-3.1, 9) = 0.00636 < 0.05$, so we reject the null hypothesis that there is no decline in the FEV.

Problem 2. Since we assume a "large" study, we start with the normal approximation to the test statistic. In that case, $\alpha = 0.05$ and a power of 0.8 require

$$0.8 = P_{\mu_a} \left(\frac{\bar{X}_d - 0}{\sigma/\sqrt{n}} \leq -1.96 \right) = P \left(Z \leq -1.96 - \frac{\mu_a}{\sigma/\sqrt{n}} \right)$$

where σ is the true standard deviation of the difference. We boldly replace σ_d by S_d from the small study and get

$$-1.96 - \frac{\mu_a}{S_d/\sqrt{n}} = z_{0.8} = .8416$$

solving for n gives that we need

$$n = \left(\frac{.15}{.147} \right)^2 (.8416 + 1.96)^2 = 8.23$$

which means we need $n = 9$ patients.

Since the number is not large, our conclusion is not valid and we have to repeat this with the the t -distribution instead of the standard normal. We follow Brittany's suggestion and simulate to check which size is needed.

For $n = 7$, we would use the t -quantile $t_{6,0.05} = -1.943$ in the one-sided t -test. To compute the power, we simulate

```
> simData<-matrix(rnorm(10000*7, mean=-.147, sd=0.15),10000,7)
> mns<-apply(simData,1,mean)
> sds<-apply(simData,1,sd)
> tStats<-(mns)/sds*sqrt(7)
> mean(tStats<=qt(0.05,6))
[1] 0.7386
```

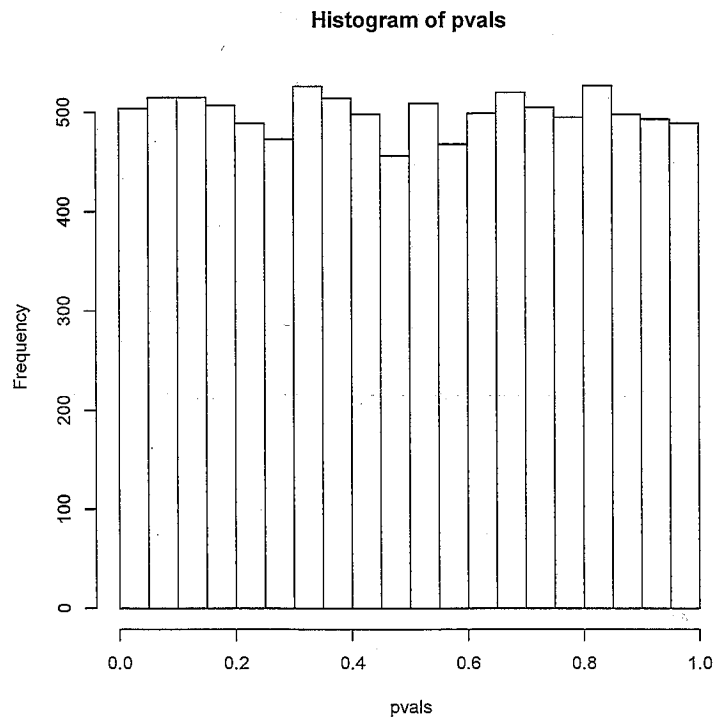
which gives a power of 0.739.

For $n = 8$, we obtain (simply replacing 7 by 8 and 6 by 7 in the above) a power of 0.8013. We know that for larger n , the power goes up. This means, a group of $n = 8$ patients is sufficient.

The reason why using the t -test leads to a smaller n to obtain the same power is that the rejection region is smaller, because the t -quantiles (for $\alpha = 0.05$) are smaller than the corresponding standard normal quantiles.

Problem 3. We simulate

```
> X=rnorm(160000, mean=5,sd=1)
> X=matrix(X,nrow=10000,ncol=16)
> means<-apply(X,1,mean)
> sds<-apply(X,1,sd)
> test.stat=(means-5)/(sds/4)
> pvals=pt(test.stat, 16)
> hist(pvals)
```



The resulting histogram indicates that the p -values are uniformly distributed in the interval $[0, 1]$.

With hindsight, this is clear: If the p -value is smaller than our preferred α , then we reject. Assuming the null hypothesis (used to generate the data in our simulation), we want this to occur with probability α . This means, we want a cumulative probability of p -values smaller than 0.05 equal to 0.05. More generally, we want that the cumulative

probability of p -values smaller than α is equal to α . This is precisely the property of the uniform distribution.

Problem 4. Project 1. We begin with the data reduction.

- We read the data from `wing_xy.dat` using the following command:

```
wing_data=read.table('wing_xy.dat')
```

Then we convert this dataframe to a matrix:

```
Y=as.matrix(wing_data)
```

- The mean of each column of \mathbf{Y} represents the x or y coordinate of the mean location of a landmark. All of these together represent a “mean wing-shape”.
- The matrix \mathbf{X} is obtained with the following R code:

```
X=Y-matrix(1,138,1)%*% t(apply(Y,2,mean))
```

- The matrix \mathbf{C} is obtained via:

```
C=(t(X)%*%X)/137
```

\mathbf{C} is a 30×30 matrix and the i, j -th entry is the covariance between the i -th and the j -th entries of the vector describing the wing shape.

- We find the eigen values of \mathbf{C} :

```
wing_eigen<-eigen(C)
```

The object `wing_eigen` contains the matrix of eigenvectors, `wing_eigen$vector`, and the list of eigenvalues, `wing_eigen$values`. The eigenvalues are arranged in descending order:

```
> wing_eigen$values
 [1] 9.223072e-02 4.048230e-02 1.162364e-02 5.413526e-03 4.182046e-04
 [6] 2.250910e-04 8.825268e-05 6.336823e-05 3.860906e-05 3.519571e-05
[11] 2.143574e-05 2.022550e-05 1.392090e-05 1.242741e-05 1.087593e-05
[16] 7.977010e-06 7.383532e-06 5.919557e-06 4.094817e-06 3.876182e-06
[21] 3.098840e-06 2.891545e-06 2.456649e-06 2.139290e-06 1.917019e-06
[26] 1.640990e-06 1.308479e-06 1.008610e-06 8.527374e-07 6.673642e-07
```

Thus the eigenvector \mathbf{v}_{max} corresponding to the largest eigenvalue α_{max} is the first column of `wing_eigen$vector`. We store this vector in the variable `comp1`:

```
comp1<-wing_eigen$vector[,1]
```

The vector \mathbf{v}_{max} represents the direction in which the multivariate data in the matrix \mathbf{Y} shows the maximum variance. This maximum variance is the given by the eigenvalue α_{max} .

- The components in the score vector \mathbf{Z} are:

```
Z<-X%*%comp1
```

These z_i -values are good descriptors because they represent the projection along the direction in which the data shows the most variance. Thus, they can be used for discriminating between wing shapes. We have assumed that most of this variance is due the difference in wing shapes and not due to noise.

Now we test the hypotheses.

- Assuming equal variance gives

```
> t.test(Z[1:42],Z[43:90],var.equal=TRUE)
```

Two Sample t-test

data: Z[1:42] and Z[43:90]

t = -4.6546, df = 88, p-value = 1.140e-05

alternative hypothesis: true difference in means is not equal to 0 .

95 percent confidence interval:

-0.3428719 -0.1376926

sample estimates:

mean of x mean of y

0.0001783903 0.2404606482

- Similarly,

```
> t.test(Z[1:42],Z[91:138],var.equal=TRUE)
```

Two Sample t-test

data: Z[1:42] and Z[91:138]

t = 5.8624, df = 88, p-value = 7.807e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.1591686 0.3224217

sample estimates:

mean of x mean of y

0.0001783903 -0.2406167397

- Again,

```
> t.test(Z[43:90],Z[91:138],var.equal=TRUE)
```

Two Sample t-test

data: Z[43:90] and Z[91:138]

t = 9.67, df = 94, p-value = 9.243e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.3822988 0.5798560

sample estimates:

mean of x mean of y

0.2404606 -0.2406167

- Since there is a significant difference between each pair among the three sample means, we conclude that the expected value of z_i is different for the three species.