MANOVA

MANOVA stands for Multivariate Analysis of Variance.

Introduction.

What is the difference between a regular ANOVA and MANOVA?

In the case of ANOVA we have one measurement (and however many factors, arranged in whatever way we're interested in).

For example, we measure the petal length of three species of flower to see if there's a difference in our flowers.

We would set up a regular one way ANOVA, and use, for example, the following data (this is part of famous data set used many, many times in statistics):

We have 50 measurements of the petal length for flowers of three different species of iris, and we get the following summary (in cm):

	Iris setosa	*Iris versicolor*	*Iris virginica*
$\bar{y}$	1.462	4.260	5.552
s	0.1736640	0.4699110	0.5518947

We perform a regular ANOVA and get the following result (output from R):

```
               Df Sum Sq Mean Sq F value Pr(>F)
iris$Species    2  437.1  218.55    1180 <2e-16 ***
Residuals     147   27.2    0.19
```

And we follow up with the usual Tukey's:

```
                      diff     lwr     upr   p adj
versicolor-setosa    2.798 2.59422 3.00178     0
virginica-setosa     4.090 3.88622 4.29378     0
virginica-versicolor 1.292 1.08822 1.49578     0
```

So far everything looks good.

But suppose we now had decided to take two measurements. Assume, we'd measured petal length and sepal width.

So here are the summaries for sepal width (also in cm):

	Iris setosa	*Iris versicolor*	*Iris virginica*
$\bar{y}$	3.428	2.770	2.974
s	0.3790644	0.3137983	0.3224966

And again the ANOVA (and Tukey's) results:

```
               Df Sum Sq Mean Sq F value Pr(>F)    
iris$Species    2  11.35   5.672   49.16 <2e-16 ***
Residuals     147  16.96   0.115                   

                       diff         lwr        upr     p adj
versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

And again we see that everything is different.

So what's the point?

Notice that we did two ANOVA's? We are suddenly doing two tests for exactly the same thing:

Is there a difference between the three species of flower?

As it turns out, in the original dataset, four variables were actually measured (the dataset also includes petal width and sepal length).

So we'd wind up doing 4 ANOVA's.

In a similar case to doing multiple *t*-tests, we are now longer using α.

(As a practical note, a Bonferroni adjustment wouldn't be a problem here, particularly given the very low *p*-values).

The same question comes up as did when we introduced ANOVA.

Is there one test that we can do, instead of 4??

Obviously, the answer is yes.

Multiple measurements on the same thing:

When we have several (more than one) measurement on the same thing, we have several "variables" that we want to use.

This gives rise to a branch of statistics known as "Multivariate" statistics.

Multivariate statistics deals with cases where we have several variables that we want to analyze at the *same time*.

Basic ideas in multivariate statistics

First, it's important to realize that we can't do more than introduce the basics.

To do multivariate statistics correctly, we need to introduce matrix and vector algebra.

But back to our flowers. Now we want to test both petal length and sepal width *at the same time*.

What kind of hypotheses would we have? For a regular one way ANOVA we'd have:

H_0: $\mu_1 = \mu_2 = \mu_3$ H_1: at least one mean is different

μ_1 = mean petal length for *Iris setosa*
μ_2 = mean petal length for *Iris versicolor*
μ_3 = mean petal length for *Iris virginica*

We could set up a similar situation for sepal width.

BUT, this is not what we want to test anymore.

We want to test:

H_0: $\mu_1 = \mu_2 = \mu_3$ *and* H_0: $\mu_4 = \mu_5 = \mu_6$ *at the same time!*

where μ_4 = mean sepal length for *Iris setosa*, μ_5 = mean sepal length for *Iris versicolor*, and μ_6 = mean sepal length for *Iris virginica*.

Note that if we reject, we reject both H_0's at once (because, really, there is only one H_0).

(If you know a little matrix algebra, we could write:

H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$

where $\boldsymbol{\mu}_1 = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}$, and μ_a = mean petal length for *Iris setosa*

μ_b = mean sepal length for *Iris setosa*

(and then so on for $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$)

and we would note that there is really just one H_0).

All right, so now we know what we want to do. How do we do it?

MANOVA - the math:

Well, the title here is wrong. We simply can't do the math without matrix algebra. But let's lay out the basic ideas in words.

When we calculate t^* for a regular t-test, what we're doing is something like:

$$\frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}}$$

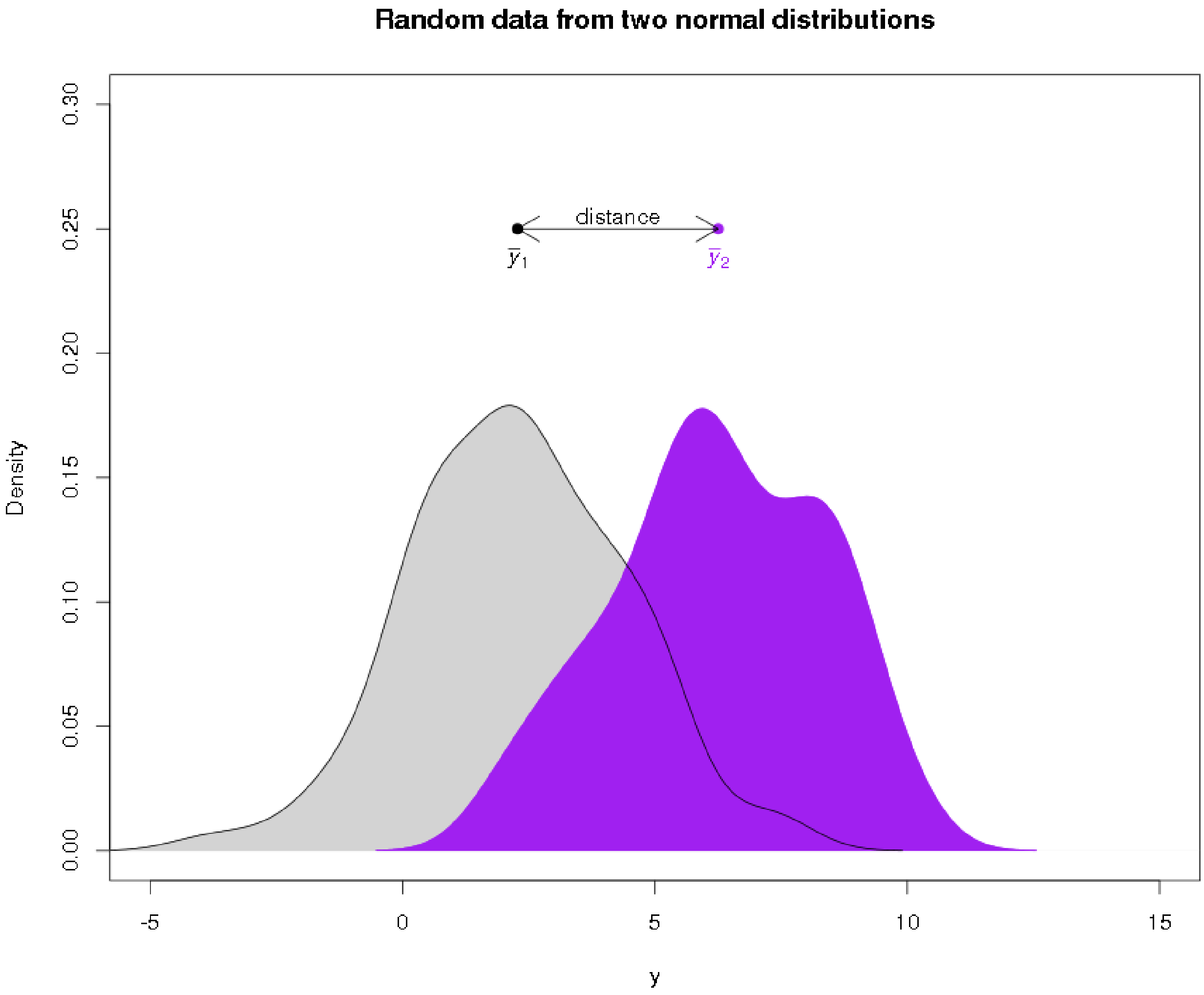
Think about this as a difference or “distance” in the numerator, divided by a way to figure out if this “distance” is important.

The denominator “standardizes” the distance:

If the SE is large, then the distance doesn't mean that much.

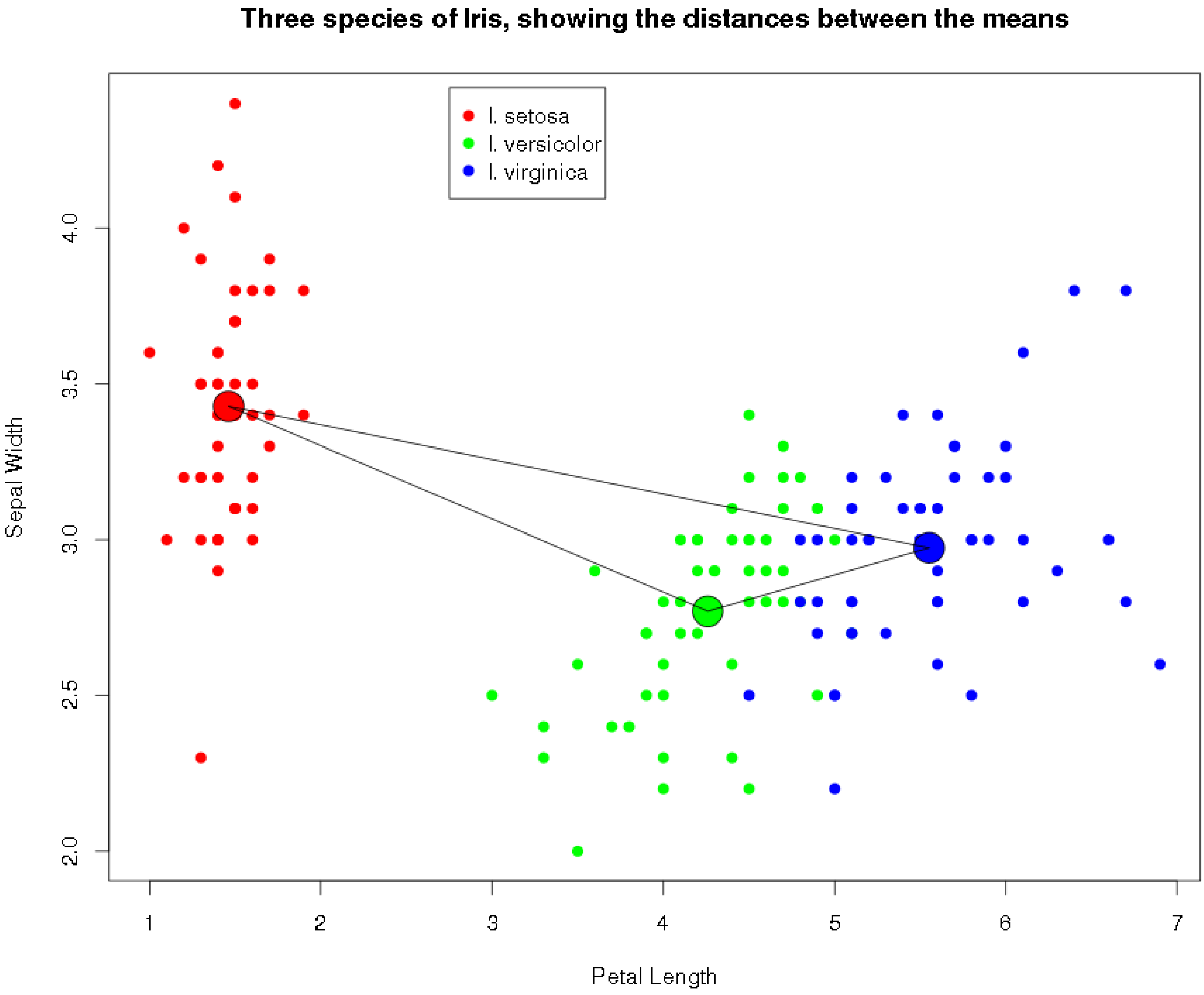
If the SE is small, then the distance is more important.

The following graph shows roughly what happens:



What we do now is we calculate distances in more than one dimension, and then “standardize” these distances based on the variability in more than one dimension.

For example, consider the following diagram:



What we do is calculate the “statistical” distances between the means of our groups, and then divide this by the variation (in two (or more) dimensions). For two samples, that would give us something very similar to the t-test:

$$T^2* = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{S}_{pooled}\right]^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

and if this is greater than some critical value for T^2, we reject H_0. Note the similarities to our regular *t**, but you should realize that all the bold stuff above is either a vector or matrix.

We do something similar once we get to more than two groups, and start using something similar to the F-statistics to determine our significance (most of the measurements listed below can be “converted” to F.

We'll have to stop here, as we really can't do much more unless we take the time to learn matrix algebra.

MANOVA - how to do it in R:

First, set up your data as usual. The only difference is that you would now have one or more columns containing measurements. For example:

Petal length	Petal width	...other measurements...	Species
xx	xx		a
xx	xx		a
...	...		...
xx	xx		b
xx	xx		b
...etc.			

And so on (just type “iris” at the R command line to see an example - the “iris” dataset is built into R).

Now, for our iris data, we would use the following commands:

```
iris3 <- manova(cbind(iris$Petal.Length,iris$Sepal.Width) ~ iris$Species)
summary(iris3,test = "Pillai")
```

Let's explain a few things:

1 - we need to use the “`cbind`” statement to tell R which variables we want to use (at the same time). We could easily use three, four, five or more variables here.

(It is possible to combine your variables into a single matrix, but let's not get carried away).

2 - the right side of the “~” still contains the factor(s) you're interested in (yes, you can do two way MANOVA and so on.)

3 - the “`test`” statement in the summary tells R what kind of multivariate test statistic to use. There are actually four in common use.

The one you use depends a bit on personal preference and/or the situation. They are (the R names given in quotes):

Pillai's trace (“`Pillai`”)
Wilk's lambda (“`Wilks`”)
Hotteling's trace (“`Hotelling-Lawley`”)
Roy's greatest root (“`Roy`”)

Your text has some more information on which to use when, but Pillai's is probably the best if you want an all around recommendation (the others will do better depending on the circumstances).

For just two samples, you can also use Hotelling's T^2

Let's look at the output, using Hotteling's trace:

```
              Df Hotelling-Lawley approx F num Df den Df    Pr(>F)    
iris$Species   2           21.861   792.46      4    290 < 2.2e-16 ***
Residuals    147
```

And you should be able to figure out how to interpret the results.

You'll see that you get slightly different result based on the method you use. Usually (but not always), all four methods will give you the same result.

What about the assumptions?

Random data is the usual assumption

Multivariate normality:

This means that the variables, when considered together, make up a “multivariate normal distribution”

For example, in two dimensions your distribution should look like a real bell (perhaps oval, but that's okay).

See figure 16.1, p. 317 in your text.

You can not verify that this is true by doing q-q plots on each variable.

Just because variable a ~ N, and b ~ N, does not mean that a and b considered together have a multivariate normal distribution.

Still, it is also true that if a is not normal or b is not normal then a and b will definitely not have a multivariate normal distribution.

To verify this assumption, you need to do a chi-square plot.

We don't have the time to go into this, but we can do a demo in R.

Interpreting this plot is is similar to interpreting normal qq plots.

Homogeneous variance-covariance matrices.

Okay, now you're worried. Let's explain:

“Homogeneous” simply means “similar” or “same”. A little like “homozygous”.

Variances should be obvious.

Covariances? This is a measure of the variability *between* variables.

These covariances should be roughly the same for any two groups.

For example, the covariance for length and width between flower 1 should be the same as the covariance for length and width for flower 2, for flower, 3, and so on.

This will be more obvious after we look at correlation.

Matrix? Because the variances and covariances are arranged into a matrix.

How do you verify this?

If your n_k's are roughly similar, it's not big deal.

There is no really good way to do this; you'll just have to “assume” this

Obviously, if you look at your variances and covariances and they're seriously different you have a problem!