Research Methods

Robert E Simpson

Singapore University of Technology & Design

robert_simpson@sutd.edu.sg

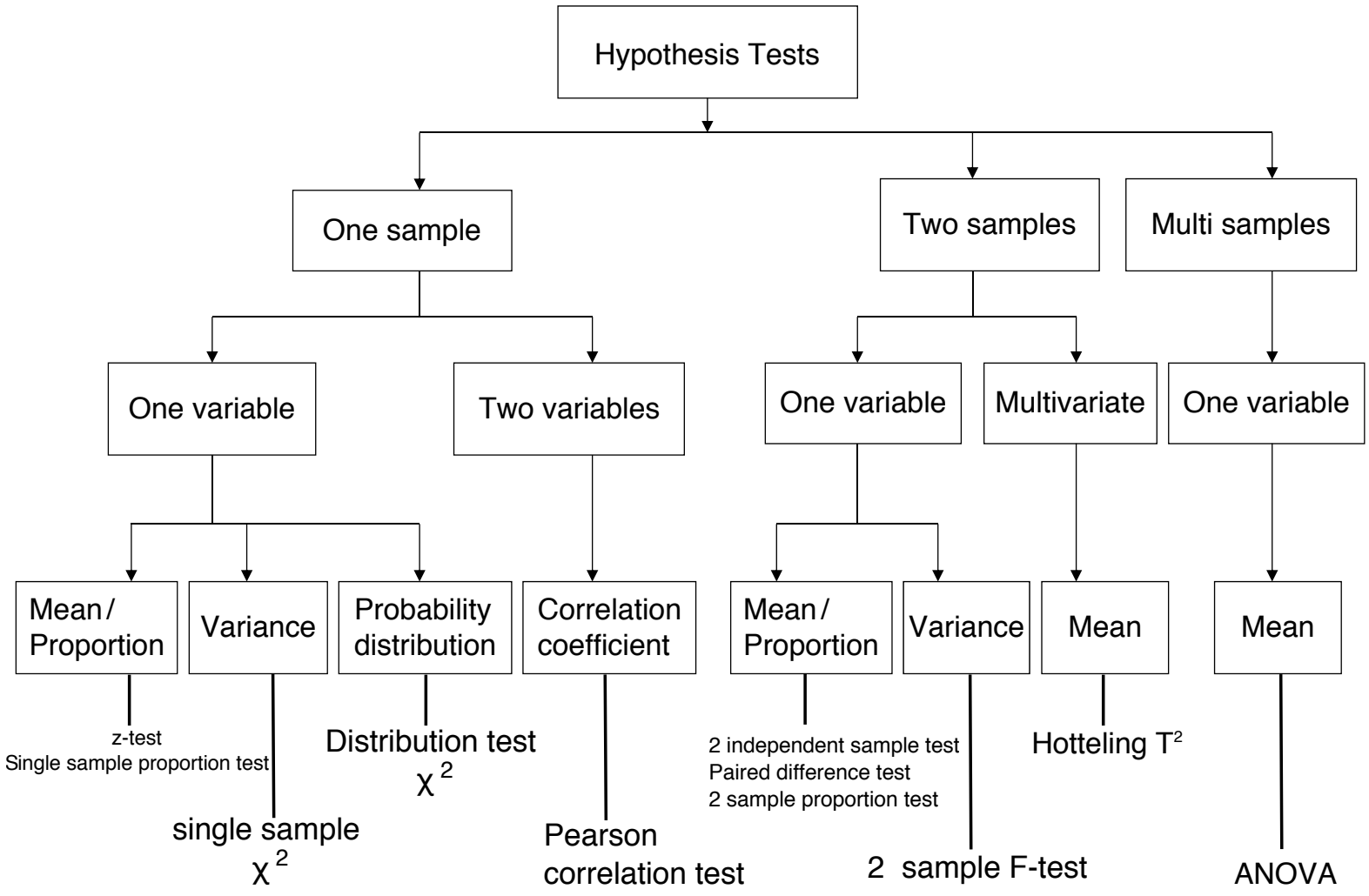
October 31, 2018

Objectives

You should be able to:

- Understand how the between-group and in-group variances influence the ability to group data
- Identify the *factors* and *levels* of an experiment
- Calculate the variances necessary for ANOVA(*in-group* , *between-group*, sum of squares)
- Perform hypothesis tests using single factor ANOVA
- Use Tukey's comaprison test to identify sources of large variation
- Calculate covariances and the Pearson Correlation coefficient
- Graphically identify correlations
- Set-up vectors to perform multivariate analysis
- Calculate the Hotelling T^2 and and use it to test whether two data set is significantly differ
- Use the Hotelling T^2 test to test whether two or more datasets of different factors come from the same dataset

Summary of the hypothesis tests



ANOVA

- Techniques used to identify and measure sources of variation within a data set
- Used to test the degree to which two or more groups vary/differ in an experiment

  (We have looked at testing for differences between two sample sets, ANOVA allows you to compare three or more different sample sets)

Applications of ANOVA

- Breaking down a distribution to see which sub-distributions create the overall population distribution
- The analysis of variance provides the formal tools to justify these intuitive judgments,

  i.e. provides a judgment in the confidence in an explanatory relationship.
- Dog breeds example from Wikipedia
  - Aim: Explain the weight distribution of dogs.
  - Method: Divide the dogs into groups that have a small *in-group* variance
  - Analysis: If the mean of each group is distinct, then the grouping is appropriate.

ANOVA– Wikipedia Dog show example

The weight of dogs at a dog show follows a complicated distribution. Possible hypothesises to explain the distribution may be:

1. The weight is related to the interaction of dog age and its hair length
2. The weight of the dog related to the interaction between the dogs purpose (pet v worker) and its build (athletic v skinny)
3. The weight of the dog is related to the breed.

However, we are unsure which grouping is most appropriate.

Age and hair length interaction

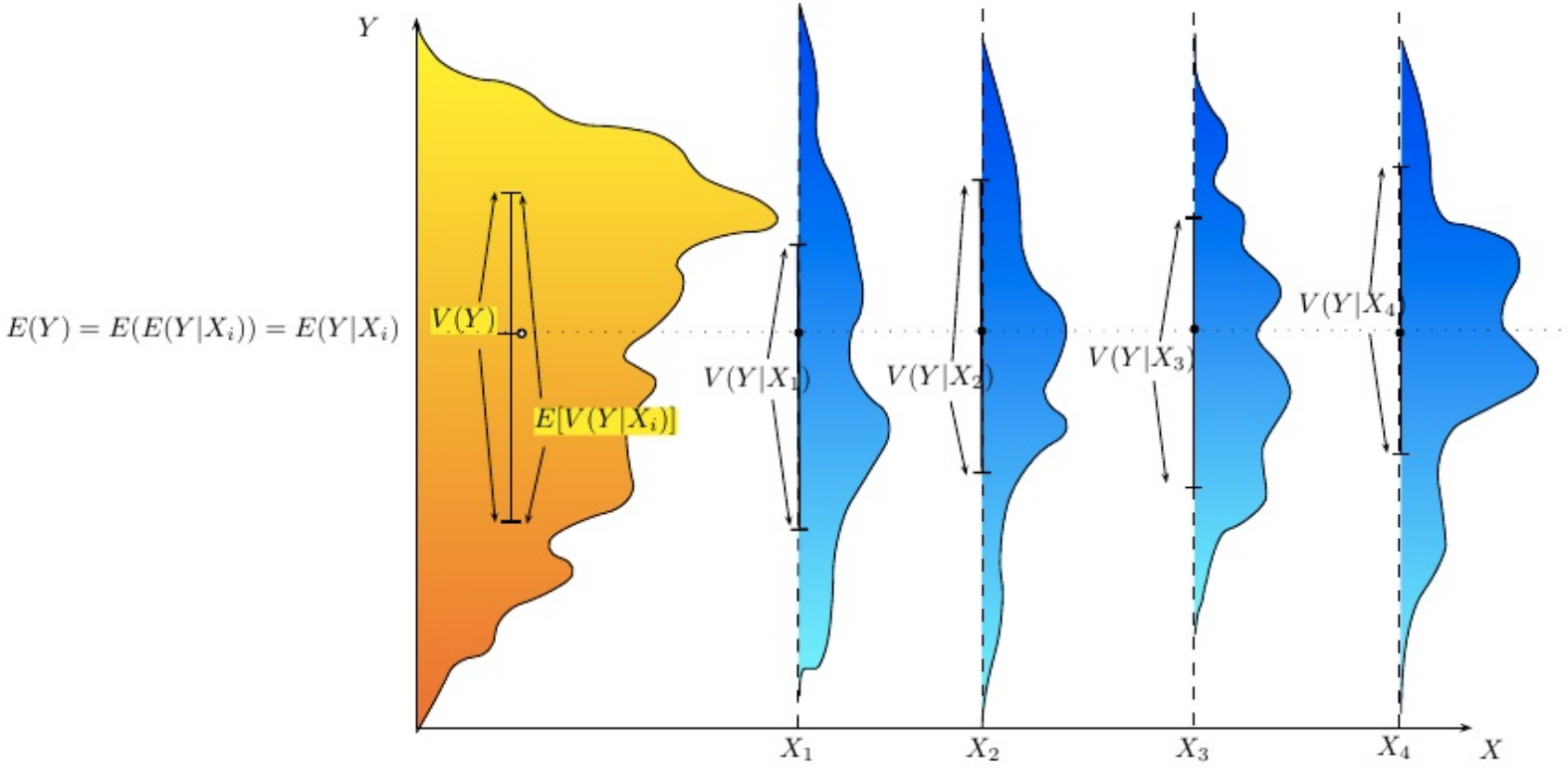


Figure 2: ANOVA : No fit

The dogs are divided in groups ($X_{1...4}$) according to the product (interaction) of two binary groupings: young vs old, and short-haired vs long-haired (thus, group 1 is young, short-haired dogs, group 2 is young, long-haired dogs, etc.).

Purpose and build interaction

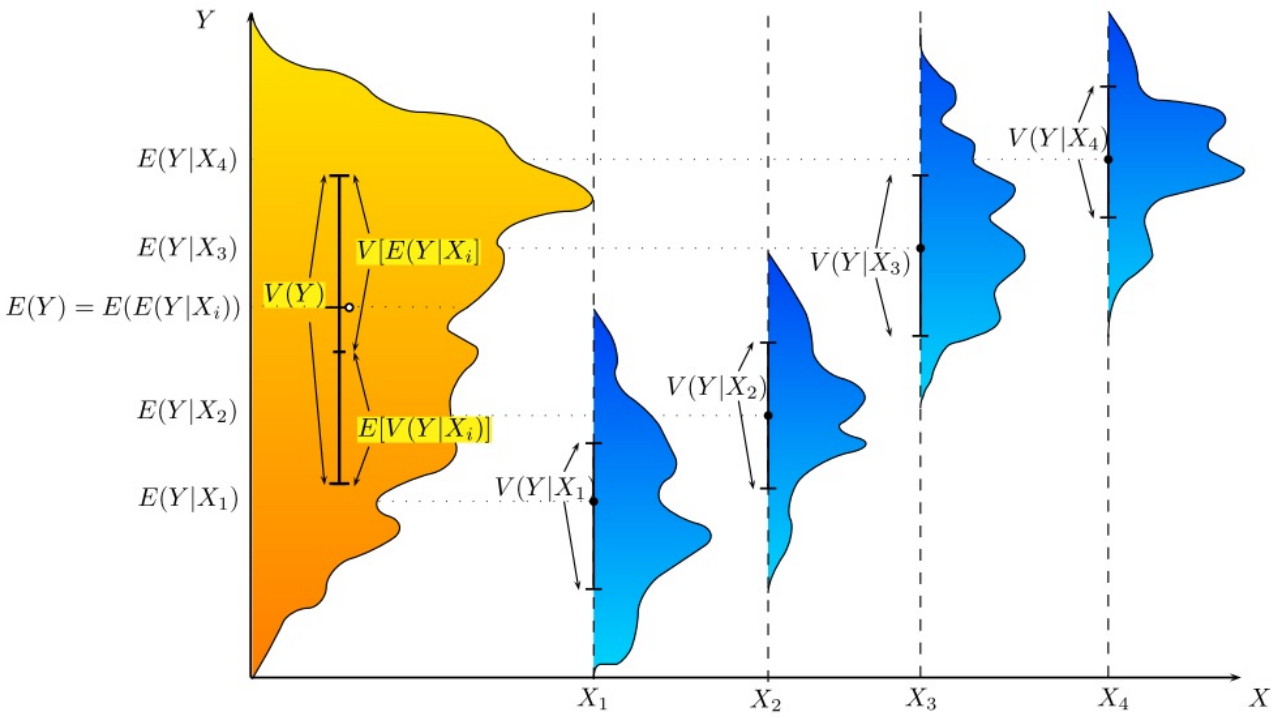


Figure 1: ANOVA : Fair fit

Dogs group by pet vs working breed, and less athletic vs more athletic. The heaviest dogs are likely to be big strong working breeds, while breeds kept as pets tend to be smaller and thus lighter. We now see fairly distinct groups.

Breed

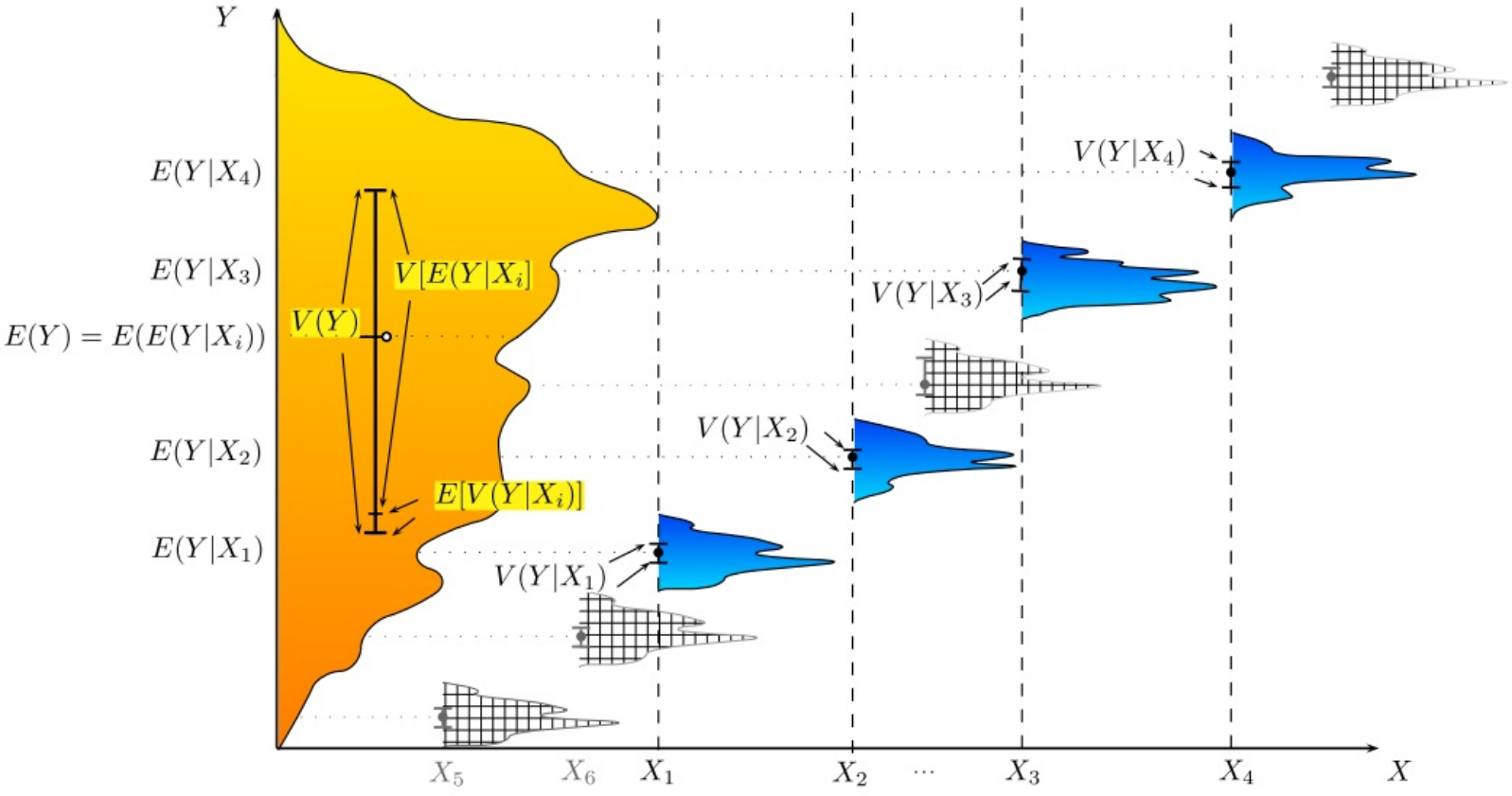


Figure 3: ANOVA : very good fit

Dog's weight grouped by breed produces a very good fit with small in group variation and well separated means.

F-Ratio

Recall, F-test is used to compare whether two variances are similar. A small F-value may indicate that the sample variances are different. We can use it to test whether the groups differ from one another. I.e. we can see which grouping is superior.

A successful group has well separated means and small *in-group* variance.
Thus a successful grouping will maximise: $\frac{\text{Between group variance}}{\text{In-group variance}}$

To perform an ANOVA, we need to compute both the *between group* variance and the *in-group* variance.

Single factor ANOVA

Often used to compare:

- Means sampled from three or more populations.
- Data from experiments when three or more different processes have been used.
  - The experimental variable is called the factor– it differentiates the sub-populations within the sample
  - The levels of a factor are the number of different sub-populations
  - Variation is assumed to be due to effects in the classification, with random error accounting for the remaining variation

Concept question 1: Factors and levels

Identify the factors and number of levels in the following examples:

(a) A study of the chip defect rate of different very large scale integration (VLSI) technologies for different node sizes (0.02 μm, 0.05 μm, 0.08 μm, and 0.1 μm).

(b) A study of the effect of temperature on the bacteria growth rate at different temperatures

(c) A study of the effect of five different motor bearings on the motor vibration

Concept question 1 solution

Identify the factors and number of levels in the following examples:

(a) A study the chip defect rate of different very large scale integration (VLSI) technologies for different node sizes (0.02 μm, 0.05 μm, 0.08 μm, and 0.1 μm).

- factor is the node size, and there are four different levels

(b) A study of the effect of temperature on the bacteria growth rate at different temperatures

- factor is the temperature, and number of levels is the number of different temperatures at which the bacteria is studied.

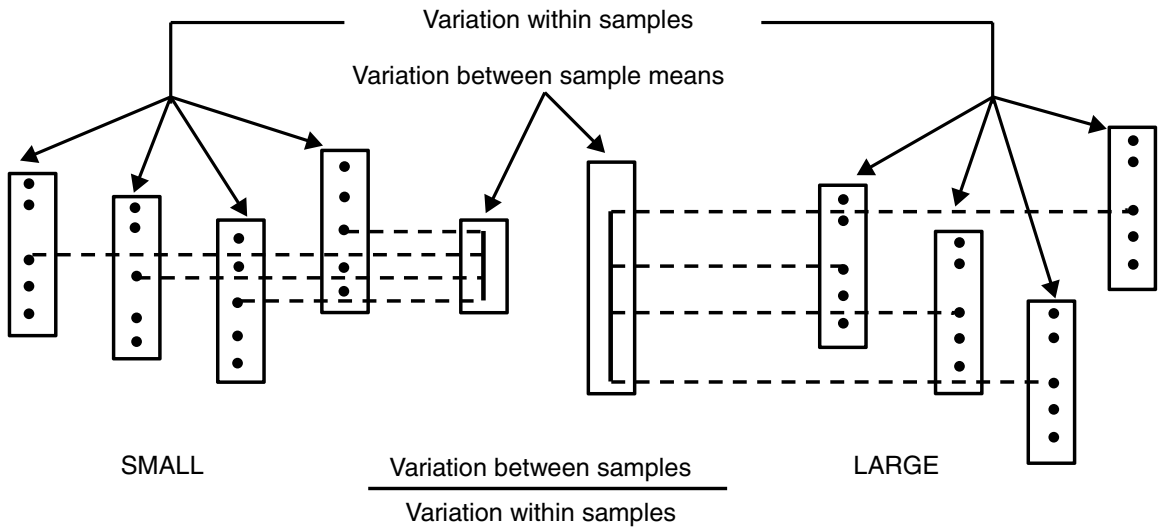
(c) A study of the effect of five different motor bearings on the motor lifetime

- factor is the motor bearing, and there are five levels

Single factor ANOVA

The variation of means of different samples suggests different parent populations.

Larger difference in sample means, $\bar{x}$ gives a larger *between-sample* variation. Hence samples are unlikely to have the same parent population



Single factor, 4 level ANOVA procedure

1. Assume four random samples have been selected, one from each population.
2. Hypothesise, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
3. Hypothesise, H_A: at least two of the μ_i's are different
4. Calculate the mean of each factor/group
5. Calculate the overall mean of the complete data set
6. Calculate the **between** group variation (SSB)
7. Calculate **within** group variation (SSW).
8. Calculate the mean sum of squares between the groups (MSB) and the mean sum of square Errors (MSW) by dividing the SSTr and SSE by their degrees of freedom.
9. Calculate the ratio of the MSB to MSW, and perform an F-test to see if the MSB is statistically significant.

In group variation

Sum of Square Errors, SSE

- Calculate the mean value of the level, $\bar{A}$,
- Compute the variability of the level: $\sum(A - \bar{A})^2$.
- Add the variability of each level to get within the group variablity, SSW= $\sum(A_1 - \bar{A}_1)^2 + \sum(A_2 - \bar{A}_2)^2 + \ldots$

$$\boxed{SSW = \sum_{i=1}^{k} n_i s_i^2} \qquad \text{d.f.=n-k} \qquad (1)$$

Where n_i is the number of samples for each level (group), i. This should be clear by considering the variance equation.

Between group variation

SSB

- Measures how different the groups are from each other
- We compare the mean at each levels to find size difference between each group mean in comparison to the *grand mean*
- Bigger differences between group means cause a larger between group variance, hence it is more likely that the treatment genuinely influenced them (i.e the grouping is correct)

Calculating SSB

Calculate the square of the difference between the mean of each level, $\bar{A}$, and the *grand mean*, $\bar{Y}$ of all levels. Sum these square differences and divide by the degrees of freedom to get the in-between group variance.

$$SSB = N_{A_1}(\bar{A}_1 - \bar{Y})^2 + N_{A_2}(\bar{A}_2 - \bar{Y})^2 + \dots$$

For k levels and n_i samples in level, i, we can we write this as:

$$\boxed{SSB = \sum_{i=1}^{k} n_i(\bar{A}_i - \bar{Y})^2} \tag{2}$$

With dof=k-1 (k is the total number of levels)

Sum of Squares Total

Variance of all data

Calculate the variance of all, N, data points and ignore the levels.
Method 1:

$$\boxed{SS_{\text{Total}} = \sum_{i}^{N} (Y_i - \bar{Y})^2} \tag{3}$$

Method 2:

$$\boxed{SS_{\text{Total}} = Ns^2} \tag{4}$$

Where s is the standard deviation of all N data points.
Method 3: The SS_{total} can be split into two different sources– between level variability and within level variability.

$$\boxed{SS_{\text{Total}} = SSB + SSW} \tag{5}$$

There are N-1 degrees of freedom ($dof = N - 1$).

Single factor ANOVA procedure

The *F-statistic* is then the ratio of the two variances:

$$\boxed{F = \frac{MSB}{MSW}} \tag{6}$$

Where:
Mean between-sample variation: $MSB = \frac{SSB}{k-1}$, with $dof = k - 1$
Mean total sum of squares: $MSW = \frac{SSW}{N-k}$, with $dof = N - k$

The F-test assumes normal populations and equal population variances.

ANOVA computation warning

In the previous calculations, I assumed that the *population* variance was calculated. This is because numpy variance (np.var) calculates the population variance by default. Therefore:

$$SSW = \sum_{i=1}^{k} n_i s_i^2$$

$$SS_{\text{Total}} = Ns^2$$

However, if the *sample* variance is known, then:

$$SSW = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

$$SS_{\text{Total}} = (N - 1)s^2$$

Note: Excel `Var` function computes the sample variance– beware!

Case Problem 8.1

Motor bearings

A motor manufacturer wishes to evaluate five different motor bearings for motor vibration (which adversely results in reduced life). Each type of bearing is installed on different random samples of six motors. The amount of vibration (in microns) is recorded when each of the 30 motors are running. The data obtained is assembled in the table below.

	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
1	13.1	16.3	13.7	15.7	13.5
2	15.0	15.7	13.9	13.7	13.4
3	14.0	17.2	12.4	14.4	13.2
4	14.4	14.9	13.8	16.0	12.7
5	14.0	14.4	14.9	13.9	13.4
6	11.6	17.2	13.3	14.7	12.3
mean	**13.68**	**15.95**	**13.67**	**14.73**	**13.08**
stdev	**1.19**	**1.17**	**0.82**	**0.94**	**0.48**

Vibration (μm) for five brands of bearings tested on six motor samples.

CP 8.1 Solution

$$\langle \bar{x} \rangle = \frac{((6 \times 13.86) + (6 \times 15.95) + (6 \times 13.67) + (6 \times 14.73) + (6 \times 13.08))}{6 \times 5}$$

$$= 14.22$$

$$SSB = 6 \times (13.68 - 14.222)^2 + 6 \times (15.95 - 14.222)^2 + 6 \times (13.67 - 14.222)^2 + \ldots$$

$$+ 6 \times (14.73 - 14.222)^2 + 6 \times (13.08 - 14.222)^2$$

$$= 30.86$$

$$MSB = \frac{SSB}{k-1} = \frac{30.68}{5-1} = 7.714$$

$$SSW = (6-1) * 1.194^2 + (6-1) * 1.167^2 + (6-1) * 0.816^2 + (6-1) * 0.940^2 + (6-1) * 0.479^2 = 22.84$$

$$MSW = \frac{SSE}{n-k} = \frac{22.84}{25} = 0.9135$$

$$F = \frac{MSB}{MSW} = 8.44$$

MSB has $d.f = k - 1$, MSE has $d.f = n - k$. From the F-tables, for $\alpha = 0.05$, the critical F-value is $F_c = 2.76$, which is less than $F = 8.44$ sowe reject H0 and conclude that the bearing brands have significant effect on the motor vibrations. Note: this conclusion is even valid at a significance level of $\alpha = 0.001$

In python:

```
import scipy.stats as ss import numpy as np b1= [13.1, 15.0, 14.0, 14.4 ,14.0,11.6]; b2=[16.3, 15.7,
17.2, 14.9, 14.4, 17.2]; b3=[13.7, 13.9, 12.4, 13.8, 14.9, 13.3]; b4=[15.7, 13.7, 14.4, 16.0,
13.9, 14.7]; b5=[13.5, 13.4, 13.2, 12.7, 13.4, 12.3]
F, p = ss.f_oneway(b1, b2, b3, b4, b5)
print F, p
>> 8.44395387871 0.000187149773642
```

CP 8.1 Solution

	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
1	13.1	16.3	13.7	15.7	13.5
2	15.0	15.7	13.9	13.7	13.4
3	14.0	17.2	12.4	14.4	13.2
4	14.4	14.9	13.8	16.0	12.7
5	14.0	14.4	14.9	13.9	13.4
6	11.6	17.2	13.3	14.7	12.3
mean	**13.68**	**15.95**	**13.67**	**14.73**	**13.08**
stdev	**1.19**	**1.17**	**0.82**	**0.94**	**0.48**
<xbar>	14.2				
sqdiff	0.29	2.98	0.31	0.26	1.30
n*sqdiff	1.75	17.89	1.86	1.56	7.80
SStr	30.86				
MSTR	7.71				
	7.13	6.81	3.33	4.41	1.15
SSE	22.84				
MSE	0.91				
F	8.44				
SST	53.69				

A limitation of ANOVA is that we cannot find the reason for a null hypothesis being rejected. For example, one sample could be very strange, yet the other samples may be very similar. Thus the one strange sample would cause the null hypothesis to be rejected.

ANOVA is rejects the null hypothesis but does not identify the strange sample.

Tukey's Comparison

ANOVA tells you there is a difference but it does not tell you the origin of the difference. Tukey significant difference test is based on paired comparisons and can be used to locate the origin of the difference.

- Must have equal sample sizes
- Separate tests are performed to check whether $\mu_i = \mu_j$ for each (i,j) pair of k populations means
- Works by comparing distance between two sample means $|\bar{x}_i - \bar{x}_j|$ to a threshold value of T that depends on a preset significance α and the MSW from the ANOVA test.

$$T = q_\alpha \sqrt{\frac{MSW}{n_i}} \tag{7}$$

Where n_i is the size of the sample drawn from each population.
q_α is the studentised range distribution, and it has its own probability table. Use $d.f. = (k, n-k)$.
If $|\bar{x}_i - \bar{x}_j| > T$, then we conclude that $\mu_i \neq \mu_j$ at the corresponding significance level.

Case problem 8.2

Using the same data as that in case problem 1, use Tukey's multiple comparison procedure to distinguish which of the motor bearing brands are superior to the rest.

Case problem 2 solution

$dof. = (k, n-k) = (5, 30-5) = (5, 25)$.
According to the **studentised range distribution**, $q_{0.05} = 4.153$.
From CP1, the MSW=0.91.
Thus,

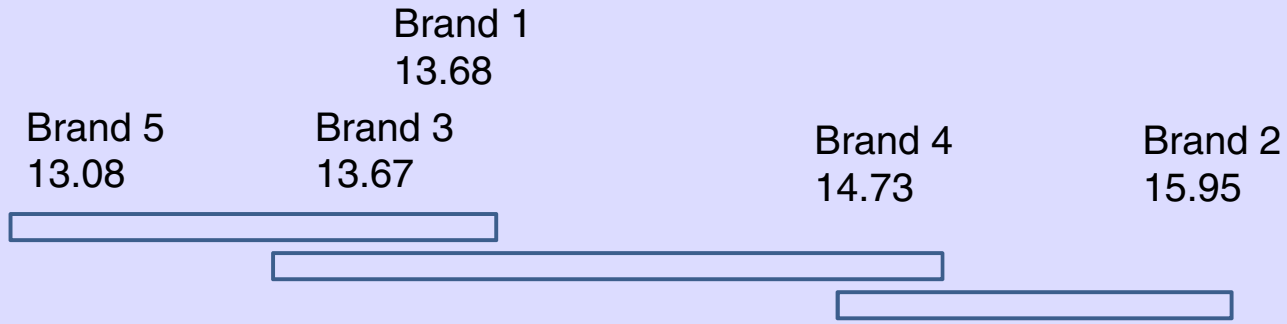
$$T = 4.153\sqrt{\frac{0.91}{6}} = 1.62 \tag{8}$$

Therefore, if the difference between any two means is greater than 1.62, we can conclude that the two populations are statistically different.

Case problem 8.2 solution

brand	brand	diff	ui/=uj
1	2	-2.27	X
1	3	0.02	
1	4	-1.05	
1	5	0.60	
2	3	2.28	X
2	4	1.22	
2	5	2.87	X
3	4	-1.07	
3	5	0.58	
4	5	1.65	X

Case problem 8.2 solution



Arranging the five sample means in ascending order and then drawing rows of bars connecting the pairs whose distances do not exceed T=1.62. It is now clear that though brand 5 has the lowest mean value, it is not significantly different from brands 1 and 3.

Therefore we would choose bearing brands 1,3, and 5, because these bearings are not statistically significantly different from one another, but they have a significantly lower vibration than brands 4 and 2.

Covariance and Pearson's correlation coefficient

Check if two data sets are correlated

We discussed covariance when deriving the propgation of error equation. The covariance represents the strength of the linear relationship between two variables:

$$cov(xy) = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{9}$$

The Pearson correlation coefficient removes the effect of magnitude in the variation of x and y, and therefore it is arguably more meaningful.

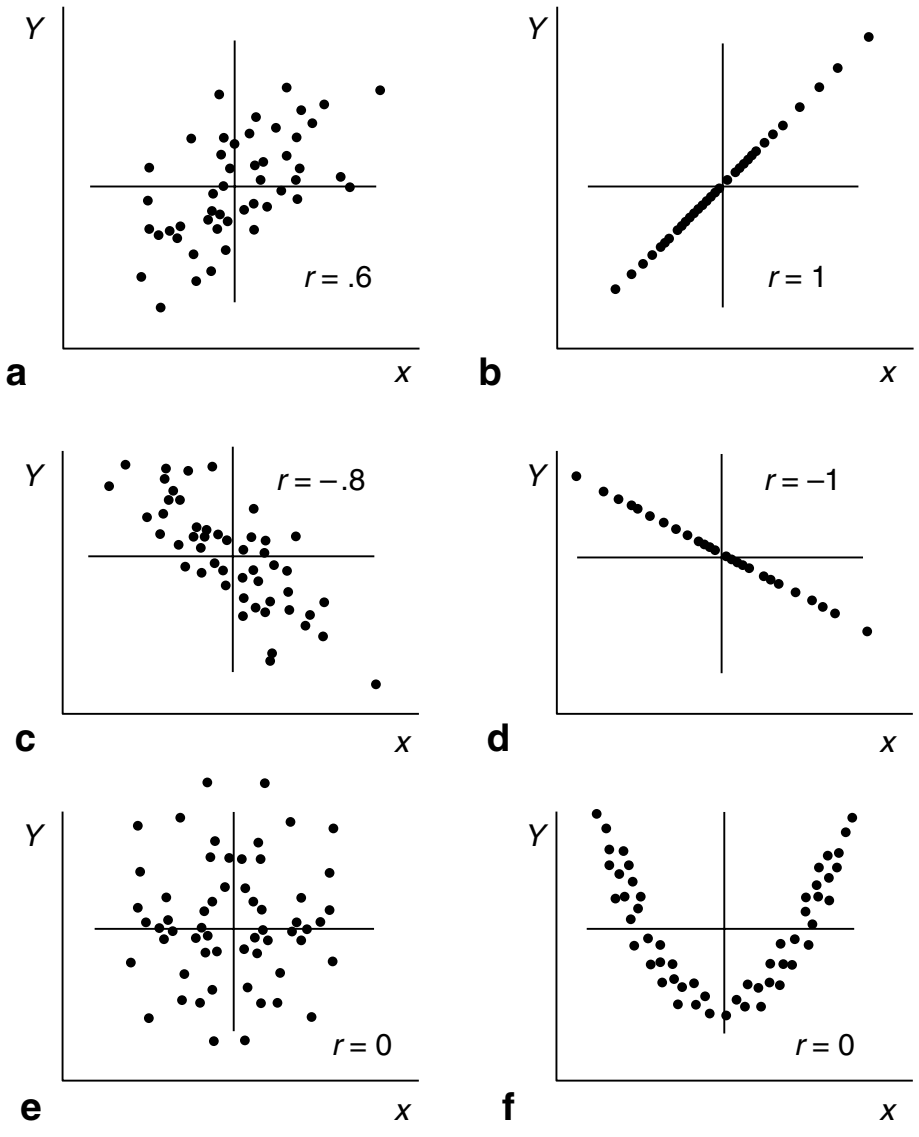
$$r = \frac{cov(xy)}{s_x s_y} \tag{10}$$

Covariance and Pearson's correlation coefficient

When $|r| > 0.9$, there is a strong linear correlation between x and y.
When $0.7 < |r| < 0.9$, there is a moderate linear correlation between x and y.
When $|r| < 0.7$, there is a weak linear correlation between x and y.

Warning
If $r = 0$, there is no linear correlation, but there might be a higher order correlation. Hence a poor Pearson correlation constant does not necessarily mean there is no correlation. In the following figure, **f** shows a second order relation between x and y, but no linear correlation, hence $r = 0$.

Correlation plots



Concept problem 8.1

The following measurements, presented in table 1, are taken for the extension of a spring under different loads. Does a correlation exist between the spring extension and the load applied to the spring?

Load (N)	2	4	6	8	10	12
Extension (mm)	10.4	19.6	29.9	42.2	49.2	58.5

Table 1: Extension of a spring with an applied load

Concept problem 1

CONCEPT QUESTION 2, WEEK 8

	LOAD	EXTENSION		X-MEAN(X)	Y-MEAN(Y)	X*Y
	2	10.4		-5	-24.566666666	122.833333333
	4	19.6		-3	-15.366666666	46.1
	6	29.9		-1	-5.0666666666	5.06666666666
	8	42.2		1	7.23333333333	7.23333333333
	10	49.2		3	14.2333333333	42.7
	12	58.5		5	23.5333333333	117.666666666
MEAN	7	34.966666666			**SUM(X*Y)=**	341.6
STDEV	3.74165738677	18.2978322942			**COV=**	68.32
					R=	0.99789350748

Multivariate methods

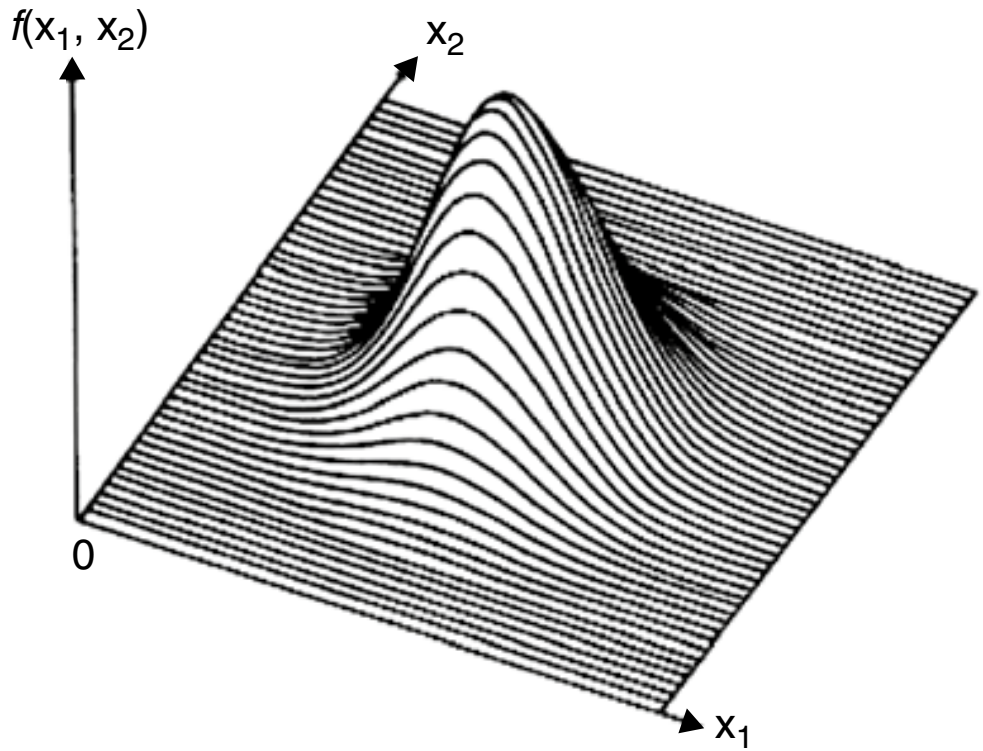
Multivariate analysis, AKA multifactor analysis deals with the statistical differences between samples of multiple different parameters. (e.g. considering multiple parameters such as dog weight, dog length, and hair length)

In contrast to t-tests and single factor ANOVA, multivariate methods allow multiple measures to be analysed simultaneously. **Inferences are made by considering the whole system of data and this results in more accurate inferences**.

Multivariate methods

The univariate probability distributions that were discussed in week 2 and 3 can be extended to be multivariate distributions. For example, let x_1 and x_2 be two discrete variables. Their joint PDF is given by:

$$f(x_1, x_2) \geq 0 \text{ and } \Sigma_{all(x_1,x_2)} f(x_1, x_2) = 1$$



Multivariate methods

Consider two multivariate datasets, each consisting of p variables. The data sets could consist of different number of observations, say n_1 and n_2. Let $\bar{\mathbf{X}}_\mathbf{1}$ and $\bar{\mathbf{X}}_\mathbf{2}$ be the sample mean vectors, each will have the same dimension, p.

$$\bar{\mathbf{X}}_\mathbf{1} = [\bar{x}_{11}, \bar{x}_{12}, \ldots, \bar{x}_{1i}, \ldots, \bar{x}_{1p}]$$
$$\bar{\mathbf{X}}_\mathbf{2} = [\bar{x}_{21}, \bar{x}_{22}, \ldots, \bar{x}_{2i}, \ldots, \bar{x}_{2p}]$$

Where $\bar{x}_{1i}$ is the sample average over n_1 observations of parameter i for the first dataset, and $\bar{x}_{2i}$ is the sample average over n_2 observations of parameter i for the second dataset.

Multivariate methods

Let $\mathbf{C}$ be the sample covariance matrix. That is, the covariance of different parameters within the same dataset. The covaricance is calculated as:

$$\boldsymbol{C} = COV(\boldsymbol{X_1}, \boldsymbol{X_2}) = \frac{\sum_{i=1}^{i=n}(X_{1,i} - \bar{X_1})(X_{2,i} - \bar{X_2})}{n-1} \tag{11}$$

Thus dataset will have covariance matrices, $\mathbf{C_1}$ and $\mathbf{C_2}$, and each will be of size $(p \times p)$.

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ \dots & \dots & \dots & \dots \\ c_{p1} & c_{p2} & \dots & c_{pp} \end{bmatrix}$$

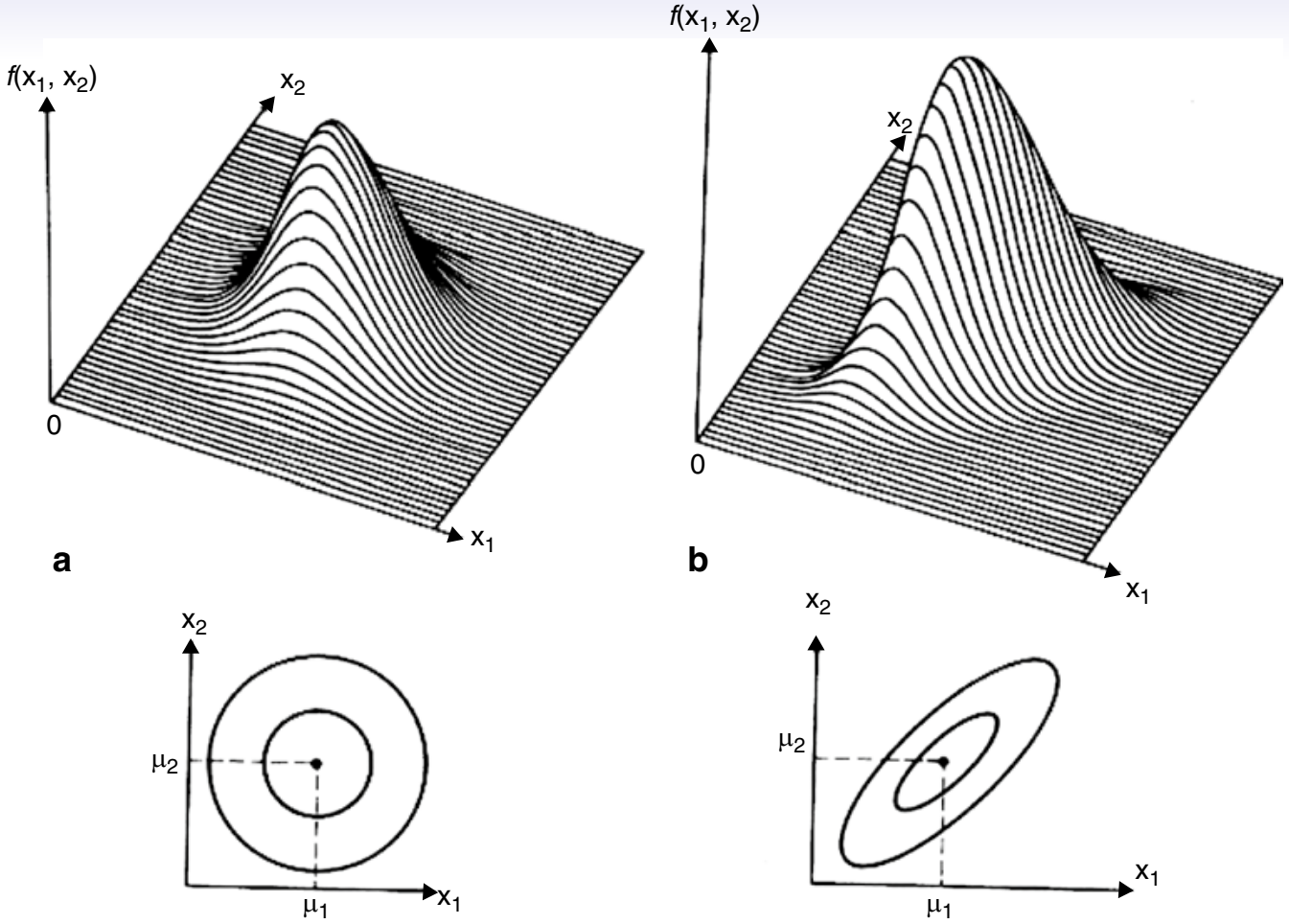
where c_{ii} is the variance of parameter i, and c_{ik} is the covariance of parameters i and k.

Multivariate methods

Similarly the correlation matrix is given by **R**:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

- Not affected by shift or scaling the data
- The observed variables can be standardised by subtracting the mean from the observation and dividing by the standard deviation.
    - A form that is easy to analyse



The left had data set is uncorrelated, and the contours are concentric. The right hand dataset has a correlation coefficient of 0.75, which results in elliptical contours.

Univariate t^2 test

Recall the t-statistic for univariate mean:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Now, consider t^2:

$$t^2 = \frac{(\bar{x} - \mu)^2}{s^2/n} \tag{12}$$

Now we have, a variance on the numerator and a variance on the denominator. Thus the F-statistic is applicable and H_0 is rejected at significance level α if t^2 is greater than the critical value from the F-table.

Multivariate Hotteling T^2Test

Equation 12 can be rewritten as:

$$t^2 = n(\bar{x} - \mu)\left(\frac{1}{s^2}\right)(\bar{x} - \mu)$$

But this only applies to a single variable dataset. However, it can be extended for a multivariate vector of means:

$$\boxed{T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})} \quad (13)$$

Where $\bar{\boldsymbol{X}}$ is a vector of sample means, $\boldsymbol{C}^{-1}$ is a vector of the corresponding variance-covariance matrix, and $\boldsymbol{\mu}$ is a vector of the corresponding population means.

So, Hotelling T^2 test is the multivariate variate version of a t-test. It tests whether two multivariate datasets are signifficantly different.

Shoes Example

A shoe company evaluates new shoe models based on five criteria: style, comfort, stability cushioning and durability, with each of the first four criteria evaluated on a scale of 1 to 20 and the durability criteria evaluated on the scale of 1 to 10.
The table below shows the goals for each criteria expected from new products.

	Goal	Mean	Stdev
Style	7	5.6	2.081666
Comfort	8	7.4	1.707825
Stability	5	5.08	2.722132
Cushion	7	5.04	2.169485
Durability	9	12.88	4.567275

Table 2: The acceptible goals for production, and the sample mean and standard dev from table 3

Shoe Example

Style	Comfort	Stability	Cushion	Durability
6	8	3	5	19
6	7	3	4	9
5	7	1	4	16
10	9	8	4	4
7	9	7	6	9
6	6	3	9	17
5	8	6	7	6
3	7	3	6	16
8	8	9	3	8
8	6	5	3	13
5	9	5	4	17
8	8	2	3	5
5	8	7	5	8
4	9	10	2	16
2	9	4	10	14
7	5	8	6	15
4	8	8	2	16
5	10	9	3	11
7	7	3	7	12
1	5	2	7	17
5	6	7	7	20
4	3	1	2	15
7	9	6	6	9
4	5	2	4	12
8	9	5	7	18

Table 3: Results of shoe survey from 25 different people

Shoe Example

Calculate Covariance matrix using equation 11 and the data in table 3.

$$C = \begin{vmatrix} 4.3333333 & 0.916667 & 1.533333 & -0.81667 & -4.46667 \\ 0.9166667 & 2.916667 & 2.425 & 0.025 & -2.40833 \\ 1.5333333 & 2.425 & 7.41 & -1.21167 & -2.57333 \\ -0.816667 & 0.025 & -1.21167 & 4.706667 & 1.963333 \\ -4.466667 & -2.40833 & -2.57333 & 1.963333 & 20.86 \end{vmatrix}$$

Shoe Example

Using equation 13 and the data in tables 2 and 3, we can calculate T^2.

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

$$T^2 = 25 \begin{bmatrix} 5.6-7 \\ 7.4-8 \\ 5.08-5 \\ 5.04-7 \\ 12.88-9 \end{bmatrix}^T \begin{bmatrix} 4.3333333 & 0.916667 & 1.533333 & -0.81667 & -4.46667 \\ 0.9166667 & 2.916667 & 2.425 & 0.025 & -2.40833 \\ 1.5333333 & 2.425 & 7.41 & -1.21167 & -2.57333 \\ -0.816667 & 0.025 & -1.21167 & 4.706667 & 1.963333 \\ -4.466667 & -2.40833 & -2.57333 & 1.963333 & 20.86 \end{bmatrix}^{-1} \begin{bmatrix} 5.6-7 \\ 7.4-8 \\ 5.08-5 \\ 5.04-7 \\ 12.88-9 \end{bmatrix}$$

$$T^2 = 52.6724$$

$$\boxed{F = \frac{n-k}{k(n-1)} T^2} = \frac{25-5}{5(25-1)} 52.6724 = 8.78$$

The critical F-value for $\alpha = 0.05$ is: $F_{0.05}(5, 20) = 2.7109$, which we find from the tables.

Since $8.78 >> 2.7109$, we reject the null hypothesis and conclude there is a significant difference between the mean scores in the sample and the stated goals.

Hotteling T^2Test

In the previous example, we compared a sample of means for different types of variable with desired true mean values (goals) to see if they are different. In that case, only the sample data had an uncertainty. In the next example, we compare two vectors of

different variables to see if they come from the same population. In this case, both vectors have uncertainty attached to their mean values.

Hotteling T^2Test

Determine whether multiple datasets come from the same population

Consider two samples of size n_2 and n_2. Compare the differences in P random variables from the two samples. Use a pooled variance to estimate (see week 5) the covariance matrix.

$$\mathbf{C} = \frac{(n_1 - 1)\mathbf{C_1} + (n_2 - 1)\mathbf{C_2}}{n_1 + n_2 - 2} \tag{14}$$

The Hotelling test is then defined as:

$$T^2 = \frac{n_1 n_2 (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T C^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)}{n_1 + n_2} \tag{15}$$

Hotteling T^2Test

A large value of T means that the two population vectors are statistically different.
The following 'transformed statistic' is used to test the significance of the difference:

$$F = \frac{(n_1 + n_2 - p - 1)T^2}{(n_1 + n_2 - 2)p} \tag{16}$$

The F-tables are used with $d.f = (n_1 + n_2 - p - 1)$

Summary

We have learnt to:

- Use ANOVA to test whether the variance of a set of observations indicate that the observations are different
- Use a pairwise Tukey difference test to identify significantly different data
- To measure correlations between observations
- Test multivariate dataset to see if they are statistically different using the Hotelling T^2 method