UiO : **Faculty of Mathematics and Natural Sciences**
University of Oslo

Multivariate Analysis CS-555-Fall2020 Course Notes

Taught by Dr. Cyril Rakovski, PhD.

Tristan Tran

UNIVERSITAS OSLOENSIS MDCCCXI

Abstract

These notes will be updated on a weekly or biweekly basis. This will serve as both an exercise in LaTex and a tool to study for the class

figure out bibliography and image uploads

Acknowledgements

Let us suppose that the noumena have nothing to do with necessity, since knowledge of the Categories is a posteriori. Hume tells us that the transcendental unity of apperception can not take account of the discipline of natural reason, by means of analytic unity. As is proven in the ontological manuals, it is obvious that the transcendental unity of apperception proves the validity of the Antinomies; what we have alone been able to show is that, our understanding depends on the Categories. It remains a mystery why the Ideal stands in need of reason. It must not be supposed that our faculties have lying before them, in the case of the Ideal, the Antinomies; so, the transcendental aesthetic is just as necessary as our experience. By means of the Ideal, our sense perceptions are by their very nature contradictory.

Rewrite this.

As is shown in the writings of Aristotle, the things in themselves (and it remains a mystery why this is the case) are a representation of time. Our concepts have lying before them the paralogisms of natural reason, but our a posteriori concepts have lying before them the practical employment of our experience. Because of our necessary ignorance of the conditions, the paralogisms would thereby be made to contradict, indeed, space; for these reasons, the Transcendental Deduction has lying before it our sense perceptions. (Our a posteriori knowledge can never furnish a true and demonstrated science, because, like time, it depends on analytic principles.) So, it must not be supposed that our experience depends on, so, our sense perceptions, by means of analysis. Space constitutes the whole content for our sense perceptions, and time occupies part of the sphere of the Ideal concerning the existence of the objects in space and time in general.

Contents

CHAPTER 1

Class: Session 1 September 1, 2020

Review of Statistics

We began the class with a review of what statistics is and what are some basic concepts of statistics. We start with an nxp matrix A.

$$A_{n \times p} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{1,2} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

For the sake of having a story behind the matrix, let's say that this is healthcare data. Each row will represent a different patient which we will call an observation. Each column represents an attribute that all of the patients will have a different value for. This could be categorical (like gender) or continuous (like systolic blood pressure). We generally want n to be much greater than p

$$n >> p$$

In statistics we will use several values to describe the population as a whole or to understand the data. The purpose of the is course is to get a basic understanding of classic techniques for understanding data. While it will not give us a completely modern understanding of the data (that kind of insight is non-trivial).

The class then discussed several different statistics in which we would hold interest.

1.1 Mean

For this problem we will use $\overline{x}$ for mean since this is the mean of the sample and not of the population μ. Because of this we must remember that we are estimating the population mean and must therefore make some adjustments to our formulas to make sure that we are not introducing bias into our model. The mean value of an attribute can be the single statistic that provides us with the most information about the distribution as a whole. It is a very powerful tool used in statistics.

1.2 Variance/Covariance Matrix

this is a matrix that represents the variance and covariance between all of the rows. This is forms a p x p symmetric matrix with the diagonal equalling the variance.

$$A_{p\times p} = \begin{pmatrix} s_{1,1}^2 & s_{1,2}^2 & \cdots & s_{1,p}^2 \\ s_{2,1}^2 & s_{2,2}^2 & \cdots & s_{2,p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1}^2 & s_{n,2}^2 & \cdots & s_{n,p}^2 \end{pmatrix}$$

where $s_{i,j}^2 = cov(x_i, x_j)$ and

$$cov(x_i, x_j) = \begin{cases} \sum_{k=1}^n \frac{(x_{ik}-\overline{x}_i)\cdot(x_{jk}-\overline{x}_j)}{n} & \text{if } i \neq j \\ \sum_{k=1}^n \frac{x_{ik}-\overline{x}_i}{n-1} & \text{if } i = j \end{cases}$$

Remember a couple of things. We use n-1 as the denominator for the variance because that makes it an unbiased estimator. This comes from the fact that we are using the sample mean instead of the population mean so we have to take away one degree of freedom from our model. We end up using n as our denominator for the sample covariance because it makes it a maximum likelihood estimator of the actual covariance.

Add proof of MLE and Unbiased estimator in appendix

1.3 Correlation Matrix

The Correlation matrix is another useful tool to use remember correlation $r(x, y) = \frac{cov(x,y)}{\sqrt{Var(x)Var(y}}$

$$A_{p\times p} = \begin{pmatrix} 1 & r_{1,2} & \cdots & r_{1,p} \\ r_{2,1} & 1 & \cdots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & 1 \end{pmatrix}$$

There are many cases where a cell in the correlation matrix will not give us any useful information. This will be the case along the diagonal since the correlation between any number and itself is 1. The other case is when one attribute is the transformation of another. For instance in medical data, they will often report the unit count of bacteria and the log scale of bacteria cultures. They may also provide the age of a patient in years, months, and days. Only one of these columns will be necessary for our model and the use of multiple would be redundant.

1.4 Quartiles:

Another important statistic is the median, or the data point in the middle of the dataset. To calculate the median, we place the dataset in sorted order and then count toward the middle of the dataset, until we only have one or two numbers remaining. If there is one number remaining, then that is the median. If there are two numbers remaining, then the average of those two numbers is the median. We will discuss the concept of averages further in the next module, but for now, just know that to calculate an average, you can sum the two numbers and divide the sum by 2.

50^{th} Percentile - Median

25^{th} percentile - Lower Quartile

Once we have calculated the median, we can calculate the quartiles of a dataset. The first quartile is the median of the first half of the dataset, or the point at which the first quarter of the dataset lies. When we calculate the first quartile, we only consider the first half of the dataset. If the median of the dataset was cleanly one number in the middle of the dataset, then it does not get included in the count for the first quartile. However, if you had to take an average to find the median, then the smaller number in that calculation is included in the count for the first quartile

75^{th} percentile - Upper

Similar rules apply to the third quartile, which is the median of the second half of the dataset, or the point at which the third quarter of the dataset lies.

Minimum: smallest value in the data set Maximum: the largest

We are also interested in different ways that we can display our data such as

1.5 Box-Whisker plots

These are plots that can display continuous data against categorical data. The thick bold line in the center represents the median. The two edges of the box represent the upper and lower quartiles. The two whiskers represent the maximum and minimum. This is a useful way to visualize the skewedness of the data.

Violin plots are another stylized variation of the box and whisker plot that is difficult to draw by hand, but made possible thanks to technology. The frequency of each observation is shown in the thickness of the violin curve.

1.6 Histograms & Bar Plots

Histograms and bar plots are similar but often mistaken for one another. Both are used to visualize the frequency of an attribute value. In a Bar plot we take the x-axis to be categorical data and the height of the bars represent the number of observations with the corresponding categorical data. With histograms, we create bins to represent possible values for a continuous variable. We then use the height of the bar to represent the frequency with which observations have attribute values within those bins.

1.7 Scatter Plots

Scatter plots are the most common and effective way to plot continuous data attributes against each other. Using these even without regression can identify patterns and trends in the data.

The following are all very good forms of exploratory data analysis that can give a statistician or data scientist insight as to which models would yield useful or interesting results.

CHAPTER 2

Class Session 2: September 3, 2020

Review of matrices

A matrix is a rectangular array full of numbers. This can be used to represent vectors, data, or a system of equations. For the purpose of statistics arrays are used to contain our data. As explained in the first session, the rows each represent one observation and the columns represent a different attribute of the observation.

2.1 Linear Operations

We can multiply any Matrix M by a scalar a resulting in a linear transformation of the matrix.

$$\text{if } M_{n\times p} = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,p} \\ m_{2,1} & m_{1,2} & \cdots & m_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & m_{n,p} \end{pmatrix}$$

$$\text{Then } a\cdot M_{n\times p} = \begin{pmatrix} a\cdot m_{1,1} & a\cdot m_{1,2} & \cdots & a\cdot m_{1,p} \\ a\cdot m_{2,1} & a\cdot m_{1,2} & \cdots & a\cdot m_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ a\cdot m_{n,1} & a\cdot m_{n,2} & \cdots & a\cdot m_{n,p} \end{pmatrix}$$

All linear operations work on matrices, but note that we can only add and subtract matrices A and B if they have the exact same dimensions.

$$A_{nxp} + B_{nxp} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ a_{2,1} & a_{1,2} & \cdots & a_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,p} \end{pmatrix} + \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,p} \\ b_{2,1} & b_{1,2} & \cdots & b_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \cdots & b_{n,p} \end{pmatrix}$$

If we let A+B = C then

$$C_{nxp} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,p} \\ c_{2,1} & c_{1,2} & \cdots & c_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & c_{n,2} & \cdots & c_{n,p} \end{pmatrix}$$

where $c_{i,j} = a_{i,j} + b_{i,j}$

2.2 Multiplying and Dividing

We can multiply two matrices A and B if and only if A is an $n \times p$ matrix and B is a $p \times m$ matrix. The left matrix must have the a number of rows equal to the number of columns in the right.

$$A_{n\times p} \cdot B_{n\times p} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ a_{2,1} & a_{1,2} & \cdots & a_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,p} \end{pmatrix} \cdot \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,m} \\ b_{2,1} & b_{1,2} & \cdots & b_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p,1} & b_{p,2} & \cdots & b_{p,m} \end{pmatrix}$$

If we let $A \cdot B = C$ then

$$C_{nxp} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,p} \\ c_{2,1} & c_{1,2} & \cdots & c_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & c_{n,2} & \cdots & c_{n,p} \end{pmatrix}$$

$$c_{ij} = \sum_{k=1}^{p} a_{i,j} \cdot b_{i,j}$$

That is each cell of c is a dot product of a row from A and a column from B. In general this system exists without division. It goes against mathematical convention.

2.3 Square Matrices

A square matrix is a matrix that has dimensions $n \times n$. For every integer value of n there is a special type of square matrix called the identity.

$$I_{nxn} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Symmetric Matrices

A symmetric matrix is a special type of square matrix where $a_{i,j} = aj, i$.

2.4 Inverse Matrices

If we have an A, can we find a B such that:

$$A \cdot B = I$$
$$B \cdot A = I$$

If this B exists then we call it A^{-1}. It only exists if A is of full rank meaning that it has a non-zero determinant. There are several ways to find the inverse of a matrix

add some methods to the appendix

Diagonal Matrices

A diagonal matrix is one such that the diagonal has values, but all other values are 0. These are convenient because the matrix and its inverse are as follows

$$A_{nxn} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

$$A_{nxn}^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_n} \end{pmatrix}$$

2.5 Transpose and Orthogonal matrices

A transpose matrix is a matrix that has been flipped across a diagonal. That is if $B = A^T$ then $b_{i,j} = a_{j,i}$. If $A^T = A^{-1}$ then the matrix is called orthogonal.

2.6 Properties of Matrices

Rank

Defined as the the number of linearly independent row or column vectors in a matrix.

Nullity

Nullity of a matrix is the number vectors in the null space of a matrix.

Determinants

https://mathinsight.org/determinant_matrix

Eigen Values and Eigen Vectors

Let's say that we have a matrix. These matrices represent a line of sorts in an n-dimensional space. Normally we describe vectors as sums of the transformations of the Identity matrix. We call that a basis. Think of the vector <2,3> we can think about that in terms of the i and j vectors. <2,3> = 2i + 3j.

Eigen values and vectors provide us with a change of base. They provide us with new unit lengthed basis vectors that are in the same direction as our matrix. The eigen values describe the specifics on how the basis was stretched to form our given matrix.

https://lpsa.swarthmore.edu/MtrxVibe/EigMat/MatrixEigen.html

2.7 Signal Value Decomposition

continued next session.

CHAPTER 3

Session 3: September 8, 2020

sec:third

Today was mostly a continuation of the linear algebra review. As a review we covered that for matrices A and B we can do the following

3.1 Quadratic Form of Matrices

We need to remember that multiplying from left and right are different. If we want to square a matrix, we cannot just write Ax^2 since x is an nx1 vector so n*n might not make sense. Instead we need to write it in quadratic form $x^T Ax$ which gives us a scalar. This will give us

3.2 Spectral Decomposition SVD of a Symmetric Matrix

Spectral Decomposition is a nifty way we can break up a matrix into the sum of vectors. Conveniently, for any given symmetrix matrix meaning $A_{k\times k} = A^T_{k\times k}$ with eigen values and vectors $(\lambda_1, e_1), (\lambda_2, e_2), (\lambda_3, e_3) ... (\lambda_k, e_k)$ then

$$A = \lambda_1 e_1 e_1^t + \lambda_2 e_2 e_2^t + \lambda_3 e_3 e_3^t + ... + \lambda_k e_k e_k^t$$

$$A = \sum_{i=1}^{k} \lambda_i e_i e_i^T$$

This follows directly from Linear Algebra for the diagonalization of matrices. We are pretty much changing basis in a K dimensional space as described last session.

The new basis of our space are the eigen vectors.

3.3 Positive Definite Matrices

Def: $A_{k\times k}$ symmetric matrix is Positive Definite iff $\underset{1\times 1}{x^T Ax} > 0$ if $x \neq 0$ $x_{k\times 1}$. We can choose any x to fill our needs. Remember: $x^T x = \|x\| > 0$ if $x \neq 0$ So $x^T Ax \approx \|x\| > 0$ can be thought of as as the vector.

Generalized Distance: Think of it like a distance that picks favorites where not all dimensions are equal. We can imagine how this would be useful in statistics where we might want to weigh our errors by the correlation or covariance of the data. It's especially useful when we remember that the covariance matrix is always going to be Positive Definite.

Theorem 3.3.1. $\underset{k\times k}{A}$ *is a Positive Definite Matrix.* $\Leftrightarrow \lambda_1, \lambda_2, ..., \lambda_k > 0$

Proof

$\underset{k\times k}{A} \Rightarrow \lambda_i > 0 \forall i$ Let $\underset{k\times k}{A}$ be a Positive Definite Matrix.

$$Ae_i = \lambda_i e_i$$
$$e_i^T A e_i = e_i^T \lambda_i e_i = \lambda$$
$$\text{since A is Positive Definite } \lambda \text{ is positive}$$

This tells us the maximum and minimum value of the equation $\underset{k\times k}{A} \Leftarrow \lambda_i > 0 \forall i$

$$A = \sum_{i=1}^{k} \lambda_i e_i e_i^T$$
$$x^T \sum_{i=1}^{k} e_i e_i^T x \underset{?}{>} 0$$
$$\text{so we choose x} \neq 0 \Rightarrow x^T A x > 0$$
$$\sum_{i=1}^{k} \lambda_i \underset{1\times 1}{(x^T e_i)} \underset{1\times 1}{(e_i^T x)} = \sum_{i=1}^{k} \lambda_i (e_i^T x)^2 > 0$$

This works by bringing the vector x into the sum. We are allowed to distribute the values. From there we are able to pull out the eigen value λ_i to the front because it is a scalar. We then realize that $x^T e_i$ is a scalar so it and its transpose are equal and scalars. The square of a scalar is positive and the sum of positive things will be positive. Q.E.D

Properties of SVD

This is something that Cyril described as being almost too cool.

$$A = \sum_{i=1}^{k} \lambda_i e_i e_i^T$$
$$A^n = \sum_{i=1}^{k} \lambda_i^n e_i e_i^T$$

In this class we will be particularly concerned with $A^{\frac{1}{2}}$

3.4 Positive Definite Matrix

A is a Positive Definite matrix of dimension kxk $x^T A x$ defines the distance from the origin to point x. d(0,x) What are the points in the space are equidistant from the origin to x, with our new metric?

$$A = \sum_{i=1}^{k} \lambda_i e_i e_i^T$$

$$x^T A x = \sum_{i=1}^{k} \lambda_i x^T e_i e_i^T x = k^2$$

$$\text{Change of variable } y_i = x^T e_i^T e_i x$$

$$\sum_{i=1}^{k} \lambda_i y_i^2 = k^2$$

Remember that an ellipse in 2-D space is $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$

$$\sum_{i=1}^{k} \frac{y_i^2}{\left(\frac{1}{\lambda_i}\right)} = k^2$$

$$\sum_{i=1}^{k} \frac{y_i^2}{\left(\frac{k^2}{\lambda_i}\right)} = 1$$

$$\sum_{i=1}^{k} \frac{y_i^2}{\left(\frac{k^2}{\lambda_i}\right)} = 1$$

$$\sum_{i=1}^{k} \frac{y_i^2}{\left(\frac{k}{\sqrt{\lambda_i}}\right)^2} = 1$$

We now have an n-dimensional football looking thing . Think of it as an ellipse that has been rotated and then scaled in each direction by $\frac{k}{\lambda_i}$

$$\underset{p\times 1}{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$\mu_i = E(x_i) \int_{-\infty}^{\infty} \underset{\text{weights PDF}}{x_i} \overset{\text{Marginal PDF}}{f_i(x_i)}$$

$$\begin{aligned} f(x_{p\times 1}) &= P(x_1, x_2, ..., x_P) \\ f_i(x_i) &= \int ... \int f(x_1, ..., x_p) dx_1, , , dx_p \\ &\qquad\qquad j \neq i \\ \sigma^2 &= \int_{\infty}^{\infty} (x_i - \mu_i) f_i(x_i) dx_i \\ cov(x_i, x_j) &= E\Bigg((x_i - \mu_i)(x_j - \mu_j) \Bigg) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) f_{ij}(x_i, x_j) dx_i dx_j \end{aligned}$$

3.5 Statistical Independence

Intuitively statistical independence tells us that an attribute carries independent information and thus contributes more to our model than other variables.

$f(x_1, x_2) = f(x_1) \cdot f(x_2)$ Dependence: Sometimes we want variables to be dependent. Usually we want them to be dependent on the observed variable.

If a variable is statistically independent, then its joint probability distribution is the same as the conditional distribution that is: $f(x_1, x_2, ..., x_k) = \prod_{i=1}^{k} f_i(x_i)$

$cov(x_1, x_2) = 0 = cor(x_1, x_2)$

If things are independent then there is no correlation; however, zero correlation cannot predict independence. This means that Independence is a stronger condition than zero correlation. We will explore this more in later chapters.

CHAPTER 4

Session 4: September 10, 2020

fourth

We learned some useful statistics in the last session. To review we could calculate the expected value $E(x_i) = \mu_i$ By using this, we can calculate the variance and covariance.

We note that we can also calculate the trace of an nxn Matrix.

$$\underset{n\times n}{\mathbf{A}} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix}$$

$$TR(A) = a_{11} + a_{22} + .. + a_{nn} = \sum_{i=1}^{n} a_{ii}$$

Now take

$$\underset{n\times 1}{x}, Var(X) = \underset{n\times n}{\Sigma} = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n} \end{pmatrix}$$

$$TR(\Sigma) = \sigma_{11} + \sigma_{22} + ... + \sigma_{nn} = VAR(x_1) + VAR(x_2) + ... + VAR(x_n)$$

The above will be useful later on in the course

4.1 Properties of Trace

For a constant k and vector A

$$TR(kA) = kTR(A)$$

For matrices A and B

$$TR(A \pm B) = TR(A) \pm TR(B)$$

Trace is a linear operator.

$$TR(AB) = TR(BA)$$

we can test with 2x2 matrices or do a more formal proof.

Proof in page 97 of book goes through Spectral Decomposition and Single Value Decomposition

$$\begin{aligned} TR(B^{-1}AB) &= TR(A) = \sum_{i=0}^{n} \lambda_i \\ \text{Proof: } \Uparrow TR(B^{-1}AB) &= TR(B^{-1}(AB)) \\ &= TR(B^{-1}(AB)) \\ &= TR\big((AB)B^{-1}\big) \\ &= TR(AI) \end{aligned}$$

We can choose B such that $B^{-1}AB = lambdamatrix = \sum_{i=1}^{n} \lambda_i$

$$B^{-1}AB = \begin{pmatrix} \lambda_{1,1} & 0 & \cdots & 0 \\ 0 & \lambda_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n,n} \end{pmatrix}$$

$TR(AA^T) = \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2$ Proof done 58:03 of the recording

4.2 Determinants

$det(A) = \lambda_1\lambda_2...\lambda_n = \prod_{i=1}^{n} \lambda_i$

True because it follows Single Value Decomposition

$$\begin{aligned} B^{-1}AB &= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \\ |B^{-1}AB| &= |B^{-1}||A||B| \\ \cancel{|B^{-1}|}|A|\cancel{|B|} &= \prod_{i=1}^{n} \lambda_i \end{aligned}$$

we then went on to do an example of calculating eigenvalues and vectors by hand. Found on page 98

afterwards we did some R code... If i'm not too lazy I'll put it in here.

CHAPTER 5

Session 5: September 15, 2020

Fifth

5.1 Partitioning a Matrix

We can partition a random vector

$$\underset{n\times 1}{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \underset{n_1\times 1}{x_1}, \underset{n_2\times 1}{x_2}, n_1 + n_2 = n$$

We can write the expected values of each vector

$$E(x) = \underset{n\times 1}{\mu}, \quad E(x_1) = \underset{n_1\times 1}{\mu_1}, \quad E(x_2) = \underset{n_2\times 2}{\mu_2}$$

We write the covariance matrix like so...

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

$$COV(x) = \Sigma, \quad cov(x_1) = \underset{n_1\times n_1}{\Sigma_{11}}, \quad cov(x_2) = \underset{n_2\times n_2}{\Sigma_{22}}$$

To calculate covariance/variance we do the following

$$\begin{aligned} COV(x_1, x_1) &= \Sigma_{1,1} = E(x_1 x_1^T) - \mu_1^2 \\ COV(x_1, x_2) &= \Sigma_{1,2} = E(x_1 x_2^T) - \underset{n_1\times n_2}{\mu_1 \mu_2^T} \\ COV(x_2, x_1) &= \Sigma_{2,1} = E(x_2 x_1^T) - \underset{n_2\times n_1}{\mu_2 \mu_1^T} \\ \text{and} \quad & \Sigma_{12} = \Sigma_{21}^T \\ \Sigma = cov(x) &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \end{aligned}$$

5.2 Properties of random Vectors

We have a random vector x

$$(X) \sim (\underset{n\times 1}{\mu}, \underset{n\times n}{\Sigma})$$

If $\underset{k\times n}{A}$ is a fixed vector with dimensions k x n then for a random vector x, $Ax \sim (A\mu, A\Sigma A^T)$. Remember the fact from baby stats $var(kx) = k^2 var(x)$

Proof

$$\begin{aligned}
E(x) &= \mu \\
E(Ax) &= AE(x) = \mu \\
COV(x) &= E\big((Ax)(Ax)^T\big) - E(Ax)E\big((Ax)^T\big) \\
&= E(Axx^TA^T) - (A\mu)(A\mu)^T \\
&= AE(xx^T)A^T - A(\mu\mu^T)A^T \\
&= A\big[E(xx^T) - (\mu\mu^T)\big]A^T \\
&= A\Sigma A^T
\end{aligned}$$

5.3 Matrix Inequalities and Maximization

5.4 Cauchy-Schwarz Inequality

Let b and d be px1 vectors

Then

$$\begin{aligned}
(b^Td)^2 &\leq (b^Tb)(d^Td) \\
(b^Td)^2 &\leq \|b\|^2\|d\|^2 \\
\text{With Equality IFF} \Leftrightarrow & \\
\underset{p\times 1}{b} &= c \underset{p\times 1}{d}
\end{aligned}$$

Proof:

Consider the vector $b - xd$ where x is a number.

$$\begin{aligned}
(b-xd)(b-xd)^T &= \|b-xd\|^2 \geq 0 \\
(b^T - d^Tx^T)(b-xd) &\geq 0 \\
b^Tb - b^Txd - d^Tx^Tb + d^Tx^Txd &\geq 0 \\
\|b\|^2 - xb^Td - x^Td^Tb + x^2d^Td &\geq 0 \\
x^2 - 2(b^Td)x + b^Tb &\geq 0
\end{aligned}$$

The above is a quadratic, so imagine a positive parabola. The discriminant must always be greater than or equal to zero

$$\begin{aligned}
D = (2b^Td)^2 - (b^Tb)(d^Td) &\leq 0 \\
D = (2b^Td)^2 - 4(b^Tb)(d^Td) &\leq 0 \\
(b^Td)^2 &\leq (b^Tb)(d^Td)
\end{aligned}$$

proof of the equality iff b = cd for some c if b=cd.

$$(b^Td)^2 = (b^Tb)(d^Td) \Leftrightarrow \exists c|b = cd \tag{5.1}$$

1) if b = cd ⇒

$$(cd^Td)^2 = (cd^Tcd)(d^Td)$$

$$c^2(d^Td)^2 = c^2(d^d)(d^Td)$$

the reason that this is the only way to get equality is because only the zero vector has zero length. We construct the vector $b - xd$ The only way that we can get zero length is if it is the zero vector and that only occurs when $b = xd$ We can also look at the cosine between the two vectors and see that the cauchy schwarz is only equal if the angle is an integer multiple of pi.

Quadratic Mean-Arithmetic Mean Inequality

$$\begin{aligned}
(b^Td)^2 &\leq (b^Tb)(d^Td) \\
b^T &= (b_1, ..., b_n) \\
d^T &= (d_1, ..., d_n) \\
\Big(\sum_{i=1}^n b_i d_i\Big)^2 &\leq \Big(\sum_{i=1}^n b_i^2\Big)\Big(\sum_{i=1}^n d_i^2\Big) \\
b_i = 1, i &= 1, ..., n \\
\Big(\sum_{i=1}^n d_i\Big)^2 &\leq n\Big(\sum_{i=1}^n d_i^2\Big) \\
\Big(\sum_{i=1}^n \frac{d_i}{n}\Big)^2 &\leq \frac{\sum_{i=1}^n d_i^2}{n} \\
\sum_{i=1}^n \frac{d_i}{n} &\leq \sqrt{\frac{\sum_{i=1}^n d_i^2}{n}}
\end{aligned}$$

As you can see the arithemtic mean is always less

Harmonic Mean-Arithmetic Mean Inequality

$$\begin{aligned}
(b^Td)^2 &\leq (b^Tb)(d^Td) \\
b^T &= (b_1, ..., b_n) \\
d^T &= (d_1, ..., d_n) \\
\Big(\sum_{i=1}^n b_i d_i\Big)^2 &\leq \Big(\sum_{i=1}^n b_i^2\Big)\Big(\sum_{i=1}^n d_i^2\Big) \\
b_i = \frac{1}{d_i}, i &= 1, ..., n \\
n^2 &\leq \Big(\sum_{i=1}^n \frac{1}{d_i^2}\Big)\Big(\sum_{i=1}^n d_i^2\Big) \\
\text{Let } y_i &= \sqrt{d_i} \\
n^2 &\leq \Big(\sum_{i=1}^n \frac{1}{y_i}\Big)\Big(\sum_{i=1}^n y_i\Big)
\end{aligned}$$

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}} \le \frac{\sum_{i=1}^{n} y_i}{n}$$

As we can see the harmonic mean is always less than the arithmetic mean.

These two inequalities are just special results of the Cauchy-Schwarz Inequality. We can choose other vectors that will give us interesting results and simplified formulas.

5.5 Extended Cauchy Schwarz

Let b and d be px1 vectors and B be a positive definite Matrix. Then

$$(b^T d)^2 \le (b^T B b)(d^T B^{-1} d)$$

with equality iff $b = cB^{-1}d$

the intuition is that will modify the vectors that we fit into the classical inequality. We are going to use well chosen vectors x and y and apply it.

First we rewrite the inequality. Instead of B we write $B^{\frac{1}{2}}B^{\frac{1}{2}}$. As a hint we can write the extended like so

$$\begin{aligned}(b^T d)^2 &\le (b^T B b)(d^T B^{-1} d)\\ &\le (b^T B^{\frac{1}{2}} B^{\frac{1}{2}} b)(d^T B^{-\frac{1}{2}} B^{-\frac{1}{2}} d)\end{aligned}$$

This hint combined with the fact that if we insert the Identity matrix nothing changes $b^T d = b^T I d$ This tells us that we might choose the vectors $x = b^T B^{\frac{1}{2}}$ and $y = d^T B^{-\frac{1}{2}}$

Proof:

Define $x = B^{\frac{1}{2}} b$ and $y = B^{-\frac{1}{2}} d$

⇑ remember that we can use Spectral Decomposition to construct the 1/2 and 1/2 inverse matrices. Also since B is a symmetric matrix, the resulting matrix is also symmetric. In statistics, the positive definite matrix will always be symmetric and symmetry is both desired and present.

Now we apply the normal the classical Cauchy Schwarz inequality

$$\begin{aligned}(x^T y)^2 &\le (x^T x)(y^T y)\\ b^T B^{\frac{1}{2}} T B^{-\frac{1}{2}} d &\le (x^T B^{\frac{1}{2}} B^{\frac{1}{2}} x)(y^T B^{-\frac{1}{2}} B^{-\frac{1}{2}} y)\\ (b^T d)^2 &\le (b^T B b)(d^T B^{-1} d)\end{aligned}$$

Equality is attained iff x=cy.

$$\begin{aligned}x &= cy B^{\frac{1}{2}} b && = cB^{-\frac{1}{2}} d\\ b &= cB^{-1} d\end{aligned}$$

Maximization Lemma

This is a result of the extended Cauchy Schwarz.

Let $\underset{p\times p}{B}$ be a positive definite matrix and $\underset{p\times 1}{d}$ be a given, fixed vector, then for any arbitrary $\underset{p\times 1}{x} \neq 0$ The following result holds the maximum

$$\underset{x\neq 0}{Max}\frac{(x^T d)^2}{x^T B x} = d^T B^{-1} d$$

The above can be though of as a function of the x_i's

$$f(x_1, ..., x_p) = \frac{(\sum x_i d_i)^2}{\sum\sum b_{ij} x_i x_j}$$

If you tried to maximize this formula normally, it would be a disaster. This Lemma helps us with that.

CHAPTER 6

Session 6: September 17, 2020

sixth

6.1 Recap of Last Session

Review of last session, we introduced

$$
\begin{aligned}
x &\sim (\mu, \Sigma) \\
Ax &\sim (A\mu, A\Sigma A^T) \\
x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}&, x \sim (\mu, \Sigma) \\
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}&, E(x_1) = \mu_1, E(x_2) = \mu_2 \\
\Sigma &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} cov(x_1) = \Sigma_{11}, cov(x_2) = \Sigma_{22} \\
cov(x_1, x_2) = \Sigma_{12} &= \Sigma_{21} = cov(x_2, x_1)
\end{aligned}
$$

We also introduced a handful of results related to the Cauchy-Schwarz inequality.

6.2 Maximization Lemma

This is a result of the extended Cauchy Schwarz.

Let $\underset{p\times p}{B}$ be a positive definite matrix and $\underset{p\times 1}{d}$ be a given, fixed vector, then for any arbitrary $\underset{p\times 1}{x} \neq 0$ The following result holds the maximum

$$\underset{x\neq 0}{Max} \frac{(x^T d)^2}{x^T B x} = d^T B^{-1} d$$

The above can be though of as a function of the x_i's

$$f(x_1, ..., x_p) = \frac{(\sum x_i d_i)^2}{\sum\sum b_{ij} x_i x_j}$$

If you tried to maximize this formula normally, it would be a disaster. This Lemma helps us with that. It's not nice to prove with calculus. Instead we are going to use the Cauchy-Schwarz to verify this

$$\begin{aligned}
&(x^T d)^2 \leq x^T Bx(d^T Bd) \\
&x^T Bx \leq 0 \quad \text{Since B is positive def. and x} \neq 0 \\
\Rightarrow &\frac{(x^T d)^2}{x^T Bx} \leq d^T B^{-1} d \\
\text{Equality iff} \quad &x = cd
\end{aligned}$$

6.3 Quadratic Form

Quadratic form is defined as showm below. It is a 1 dimensional scalar object. It can be written as the sum, but we prefer the matrix notation.

$$\underset{1\times 1}{x^T Bx} = \sum_{j=1}^{p}\sum_{i=1}^{p} b_{ij}x_i x_j$$

If B is positive definite and $x \neq 0$ then the quadratic form will always be positve

$$\begin{aligned}
&\underset{1\times 1}{x^T Bx} > 0 \\
&\underset{1\times 1}{x^T Bx} > \alpha \quad \forall x \neq 0
\end{aligned}$$

Take $k \cdot x$ where k is a constant

$$(kx)^T B(kx) = k^2(fx^T Bx) = k^2\alpha$$

If we let k go to ±∞, the quadratic form will trend towards ∞. This makes sense since the quadratic form is just a distance taken from the origin and we are just shrinking and stretching the distances

6.4 Maximization of Quadratic Forms for points on the Unit Sphere

Let B - P.D. Matrix wwith dimensions p x p with

Eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p > 0$

and with Corresponding Eigenvectors $e_1, e_2, ..., e_p$

Then

$$\begin{aligned}
&(1)\underset{x\neq 0}{Max}\frac{x^T Bx}{x^T x} = \lambda_1 && equality \Leftrightarrow x = e_1 \\
&(2)\underset{x\neq 0}{Min}\frac{x^T Bx}{x^T x} = \lambda_p && equality \Leftrightarrow x = e_p \\
&(3)\underset{x\perp e_1,\ldots,e_k}{Max}\frac{x^T Bx}{x^T x} = \lambda_p && equality \Leftrightarrow x = e_k + 1
\end{aligned}$$

Explaining 3

Imagine we have a 3 dimensional space. The maximum in all 3 dimensions $\underset{x\epsilon\mathbb{R}^3}{Max}\frac{x^TBx}{x^Tx} = \lambda_1$. If we restrict our space and take its projection only onto the two lower eigen vectors e_2, e_3 The maximum on that plane is $\underset{x\perp e_1}{Max}\frac{x^TBx}{x^Tx} = \lambda_2$ if we further restrict it to only be along the eigen value of e_3 our maximum becomes $\underset{x\perp e_1,e_2}{Max}\frac{x^TBx}{x^Tx} = \lambda_3$

This works when you have too many eigen vectors and you want to restrict your answer by ignoring some of them.

We eliminate the orthogonal space in order.

Proof of 1 Maximum

Proof of ⇒. If B is P.D. then the maximum is attained atλ_1 so we can represent P

$$\begin{aligned}
\frac{x^TBx}{x^Tx} &= \frac{x^TB^{\frac{1}{2}}B^{\frac{1}{2}}x}{x^Tx} \\
\underset{p\times p}{P} &= \begin{pmatrix} e_1 & e_2 & ... & e_p \end{pmatrix} \\
B &= P\Lambda P^T && \text{Single Value Decomposition} \\
\Lambda &= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \\
B^{\frac{1}{2}} &= P\Lambda^{\frac{1}{2}}P^T \\
\frac{x^TBx}{x^Tx} &= \frac{x^TP\Lambda^{\frac{1}{2}}P^TP\Lambda^{\frac{1}{2}}P^Tx}{x^Tx} \\
P^T &= P^{-1} && \text{P is orthogonal} \\
\frac{x^TBx}{x^Tx} &= \frac{x^TP\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}P^Tx}{x^TPP^Tx} \\
y &= P^Tx && \text{Change of Variables} \\
\frac{x^TBx}{x^Tx} &= \frac{y^T\Lambda y}{y^Ty} \\
&= \frac{\sum_{i=1}^{n}\lambda_i y_i}{\sum y_i^2}
\end{aligned}$$

so we have

$$\frac{x^TBx}{x^Tx} = \frac{y^T\Lambda y}{y^Ty} = \frac{\sum_{i=1}^{n}\lambda_i y_i}{\sum y_i^2} \overset{?}{\leq} \lambda_1$$

Remember that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ so $\Rightarrow$

$$\begin{aligned}\frac{\sum \lambda_i y_i^2}{\sum y_i^2} &\leq \frac{\sum \lambda_1 y_i^2}{\sum y_i^2} \\ &= \lambda_1 \frac{\cancel{\sum y_i^2}}{\cancel{\sum y_i^2}} = \lambda_1\end{aligned}$$

We can do this because we replaced all the smaller eigenvalues with a greater eigenvalue λ_1 creating a theoretical upper limit. Now we need to prove that this upper limit is attained. Let's test the following

$$\frac{e_1^{(}TBe_1)}{e^T e_1} = e^T \lambda_1 e_1 = \lambda_1$$

so we observe that the value is actually obtained for the corresponding eigenvector. Now we need the reverse direction

Proof of 2 Minimum Value

For this we will make the same substitutions and change of variables to get the form below:

$$\frac{x^T Bx}{x^T x} = \frac{y^T \Lambda y}{y^T y} = \frac{\sum_{i=1}^{n} \lambda_i y_i}{\sum y_i^2} \overset{?}{\geq} \lambda_p$$

should be exactly the same as our previous result except instead of substituting the first eigenvalue, we substitute the last

$$\begin{aligned}\frac{\sum \lambda_i y_i^2}{\sum y_i^2} &\geq \frac{\sum \lambda_p y_i^2}{\sum y_i^2} \\ &\geq \lambda_p \frac{\cancel{\sum y_i^2}}{\cancel{\sum y_i^2}} = \lambda_p\end{aligned}$$

We plug in $x = e_p$ to achieve the smallest value.

$$\frac{e_p^T (Be_p)}{e_p^T e_p} = e_p^T \lambda_p e_p = \lambda_p$$

Proof of 3 Max of perpendicular subspace

$$\underset{x \perp e_1, \ldots, e_k}{Max} \frac{x^T Bx}{x^T x} = \lambda_{k+1} \quad |" = "x = e_{k+1}$$

Remember that we made a change of variables from the original equation to get the new one

$$\begin{aligned}y = P^T x, x = Py \Downarrow \\ \frac{y^T \Lambda y}{y^T y}\end{aligned}$$

$$y = \begin{pmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_p^T \end{pmatrix}, x = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ e_{k+1} \\ \vdots \\ e_p^T \end{pmatrix} \Rightarrow$$

$$\begin{aligned} \frac{y^T \Lambda y}{y^T y} &= \frac{\displaystyle\sum_{i=k+1}^{p} \lambda_p y_i^2}{\displaystyle\sum_{i=k+1}^{p} y_i^2} \\ &\leq \lambda_{k+1} \frac{\cancel{\sum y_i^2}}{\cancel{\sum y_i^2}} = \lambda_{k+1} \end{aligned}$$

Since we take x to be strictly orthogonal to the first k vectors all those terms are 0 and we can rewrite the sum starting at k+1. The largest eigenvalue multiplied by a nonzero then becomes λ_{k+1} and the rest is identical to the other proofs.

CHAPTER 7

Session 7: September 22, 2020

seventh

Starting today we start the material in Chapter 3. We are going to use simple geometry to understand samples and some of their properties.

7.1 Projection of y onto x

If we take a vector y, and project it onto x, we get the vector

$yprojx = \frac{y^T x}{x^T x} x$

We get a vector in the direction of x with just the component of y that is in the same direction as x. We can think of this as the shadow that y casts onto x.

We are going to choose particular vectors to project and make this interesting.

We will be projecting a data matrix.

$$\begin{aligned} x &= \left[\begin{array}{c|c|c|c} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{array}\right] \\ &= \begin{pmatrix} y_1 & y_2 & \cdots & y_p \end{pmatrix} \end{aligned}$$

that is we define each y_i as a column.

7.2 Projection onto the 1 vector

We will also define 1_n as a vector

$$1_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

if we project y_j onto 1_n we will get

$$\begin{aligned} \frac{y_j^T 1_n}{1_n^T 1_n} 1_n &= \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{n} \cdot 1_n \\ &= \overline{x}_j \cdot 1_n \end{aligned}$$

If we project the y_i onto 1_n we get the vector $\overline{x}_j \cdot 1_n$.

Their difference is $y_j - \overline{x}_j \cdot 1_n$ and is perpendicular to 1_n.We will call these d_j. It can also be called the vector of individual deviations from the mean.

$$d_j = y_i - \overline{x_j} \cdot 1_n = \begin{pmatrix} x_{1j} - \overline{x}_j \\ x_{2j} - \overline{x}_j \\ \vdots \\ x_{nj} - \overline{x}_j \end{pmatrix}$$

We can find the length of each d_j and use a clever substitution to get a relationship with our sample deviation.

$$||d_j||^2 = \sum_{k=1} n(x_{kj} - \overline{x}_j)^2 = n \cdot s_j^2$$

This demonstrates that data sets with more length will give more varied data.

If instead we substitute one of the d_j's with a differenta column, we get the sample covariance.

$$d_i^T d_j = \sum_{k=1} n(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) = n \cdot s_{ij}$$

So we also know that the dot product between two vectors is equal to the product of their angles times tha ange between them. $d_i^T d_j = ||d_i|| \cdot ||d_j|| cos\big(\triangleleft(d_i, d_j)\big)$

$$\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) = \sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2} \cdot cos\big(\triangleleft(d_i, d_j)\big)$$

$$\begin{aligned} cos\big(\triangleleft(d_i, d_j)\big) &= \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}} \\ &= \frac{\cancel{n} \cdot s_{ij}}{\sqrt{s_{ii}\cancel{n}}\sqrt{s_{jj}\cancel{n}}} \\ &= r_{ij} \end{aligned}$$

If the deviation vectors are orthogonal that indicates that they have zero correlation. If they are perfectly correlated they are colinear.

Estimating μ and σ Univariate Case

$x_1, x_2, ..., x_n \sim (\mu, \sigma^2)$ Then

$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is an Unbiased estimator of μ

That is the expected value of $\overline{x}$ equals μ

$E(\overline{x}) = \mu$ By the definition of unbiased estimator of μ

Proof

$$\begin{aligned} E\Big(\frac{x_1+...+x_n}{n}\Big) &= \frac{1}{n}\big(E(x_1)+...+E(x_n)\big) \\ &= \frac{\cancel{n}\mu}{\cancel{n}} \\ &= \mu \end{aligned}$$

Variance

$VAR(\overline{x} = \frac{\sigma^2}{n})$

$$\begin{aligned} VAR\Big(\frac{x_1+...+x_n}{n}) &= \frac{1}{n^2}\Big(VAR(x_1)+VAR(x_2)+...+VAR(x_n)\Big) \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

$x_1, x_2, ..., x_n \sim (\mu, \sigma^2)$ Then,

$\hat{\sigma} = s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$

That is the sample variance is calculated by the above formula. The n-1 makes it an unbiased estimator.

$E(s^2) = E(\hat{\sigma^2}) = \sigma^2$ This is where I suggested a proof strategy and he proceeded to prove that it was a wrong/unintuitive approach.

proof

$$\begin{aligned} E\Big(\sum_{i=1}^{n}(x_i-\overline{x})^2\Big) &= E\Big[\sum_{i=1}^{n}(x_i^2-2x_i\overline{x}+\overline{x}^2)\Big] \\ &= \sum_{i=1}^{n}E(x_i^2) - 2E\Big(\overline{x}\sum_{i=1}^{n}x_i\Big) + nE(\overline{x}^2) \\ E(x_i^2) &\overset{DEF}{=} VAR(x_i) + \Big(E(x_i)\Big)^2 \\ &= \sigma^2+\mu^2 \\ VAR(x_i) &= E(x_i^2) - \Big(E(x_i)\Big)^2 \\ E\Big(\sum_{i=1}^{n}(x_i-\overline{x})^2\Big) &= n(\sigma^2+\mu^2) - 2E\Big(\overline{x}n\overline{x}\Big) + nE(\overline{x}^2) \\ E(\overline{x}^2) &\overset{DEF}{=} VAR(\overline{x}) + E(\overline{x})^2 \\ &= \frac{\sigma^2}{n}+\mu^2 \\ E\Big(\sum_{i=1}^{n}(x_i-\overline{x})^2\Big) &= n(\sigma^2+\mu^2) - nE\Big(\overline{x}^2\Big) \end{aligned}$$

$$
\begin{aligned}
&= n(\sigma^2 + \mu^2) - nE(\frac{\sigma^2}{n} + \mu^2) \\
&= (n-1)\sigma^2 \Rightarrow
\end{aligned}
$$

$$E\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}\right) = \sigma^2$$

Summary

So now we know that for any sample regardless of distribution we can find estimates for the mean and the variance using the sample mean and sample variance.

The rest of the class was walking through an example of collecting data and calculating sample statistics.

CHAPTER 8

Session 8: September 24, 2020

eight

From last class we explored some geometric insights of projections and means. We also had 3 main postulates that we found - $\overline{x}$ is an unbiased estimator of μ - $VAR(\overline{x} = \frac{\sigma^2}{n}$ - $s^2 = \frac{\sum(x-\overline{x})^2}{n-1}$

8.1 Multivariate Sample Mean

$\underset{p\times 1}{\vec{x_1}}, ..., \underset{p\times 1}{\vec{x_n}} \sim (\underset{p\times 1}{\mu}, \underset{p\times p}{\Sigma})$

Proof 1

1) $\underset{p\times 1}{\overline{x}}$ is an unbiased estimator of the $\underset{p\times 1}{\mu}$ or $E(\overline{x}) = \mu$

Proof is similar to the univariate case since we're only using linear operators.

$$
\begin{aligned}
E(\overline{x}) &= E(\frac{x_1 + ... + x_n}{n} \\
&= \frac{1}{n}(E(x_1) + ... + E(x_n)) \\
&= \frac{n\mu}{n} = \mu
\end{aligned}
$$

Proof 2: Multivariate Covariance of the Sample Mean

2)

$$
\begin{aligned}
\underset{p\times p}{cov(\overline{x})} &= E[(\overline{x} - \mu)(\overline{x} - \mu)^T] \\
&= E\Big[\Big(\frac{\sum x_i}{n} - \mu\Big)\Big(\frac{\sum x_j}{n} - \mu\Big)^T\Big] \\
&= E\Big[\Big(\frac{\sum x_i}{n} - \frac{n\mu}{n}\Big)\Big(\frac{\sum x_j}{n} - \frac{n\mu}{n}\Big)^T\Big] \\
&= \frac{1}{n^2}E\Big[\sum\Big(x_i - \mu\Big)\sum\Big(x_j - \mu\Big)\Big] \\
\text{if } i \neq i \text{ cov} = 0 \\
&= \frac{1}{n^2}\sum E\big[(x_i - \mu)(x_i - \mu)^T\big]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n^2} \sum \Sigma \\
&= \frac{\Sigma}{n}
\end{aligned}
$$

Proof 3: Multivariate Sample Variance

$$
\begin{aligned}
\underset{p\times p}{S} &= \sum \frac{(x_i - \overline{x})(x_i - \overline{x})^T}{n-1} \\
E(S) &= \Sigma
\end{aligned}
$$

Proof

$$
\begin{aligned}
E\big[\sum (x_i - \overline{x})(x_i - \overline{x})^T\big] &= \sum E\big[(x_i - \overline{x})x_i^T\big] + \sum E\big[(x_i - \overline{x})(-\overline{x})^T\big] \\
&= \sum E\big[(x_i - \overline{x})x_i^T\big] \\
&= E\big[\sum x_i x_i^T - \overline{x}\sum x_i^T\big] \\
\\
&= \sum E\big[x_i x_i^T\big] - E\big[n\overline{x}\overline{x}^T\big]
\end{aligned}
$$

$$
\begin{aligned}
cov(x_i) = \Sigma &\overset{DEF}{=} E(x_i x_i^T) - \mu\mu^T \\
E(x_i x_i^T) &= \Sigma + \mu\mu^T
\end{aligned}
$$

$$
\begin{aligned}
cov(\overline{x}) = \frac{\Sigma}{n} &\overset{DEF}{=} E(\overline{x}\overline{x}^T) - \mu\mu^T \\
E(\overline{x}\overline{x}^T) &= \frac{\Sigma}{n} + \mu\mu^T
\end{aligned}
$$

$$
\begin{aligned}
E\Big[\sum x_i x_i^T - n\overline{x}\overline{x}^T\Big] &= \sum \Big(\Sigma + \mu\mu^T\Big) - \sum \Big(\frac{\Sigma}{n} + \mu\mu^T\Big) \\
\\
&= (n-1)\Sigma \\
E\Big[\frac{\sum (x_i - \overline{x})(x_i - \overline{x})^T}{n-1}\Big] &= \Sigma
\end{aligned}
$$

8.2 Manipulating Data Matrix

We will manipulate a data matrix using only matrix algebra to get the covariance matrix.

Let our data matrix be the following

$$
x = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
$$

$$x^T = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$x^T 1_n = \begin{pmatrix} \sum_{j=1}^{n} x_{j1} \\ \sum_{j=1}^{n} x_{j2} \\ \vdots \\ \sum_{j=1}^{n} x_{jn} \end{pmatrix}_{p\times 1}$$

$$\frac{1}{n} x^T 1_n = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_n \end{pmatrix}_{p\times 1} = \underset{p\times 1}{\overline{x}}$$

$$\overline{x}^T = \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_n \end{pmatrix}_{1\times p}$$

$$1_n \overline{x}^T = \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \end{pmatrix}_{n\times p}$$

So now we have outlined a process by which we can construct a row matrix that where all the columns are just the means. This will be useful if we ever need to calculate deviations and program these. We can make it so that we don't need loops or decompositions.

Deviations matrix

Below we will begin a construction of the deviations.

$$1_n \overline{x}^T = \frac{1}{n} 1_n 1_n^T x^T = \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \end{pmatrix}_{n\times p}$$

$$\overline{x}^T = \frac{1}{n} 1_n^T x^T$$

$$x_{n\times p} - \frac{1}{n} 1_n 1_n^T x^T =$$

$$\begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1p} - \overline{x}_p \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{2p} - \overline{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & \cdots & x_{np} - \overline{x}_p \end{pmatrix}_{n\times p}$$

$$(x_{n\times p} - \frac{1}{n} 1_n 1_n^T x^T)^T (x_{n\times p} - \frac{1}{n} 1_n 1_n^T x^T)$$

$$=$$

$$\begin{pmatrix} \sum(x_{i1}-\overline{x}_1)^2 & \sum(x_{i2}-\overline{x}_1)^2 & \cdots & \sum(x_{ip}-\overline{x}_1)^2 \\ \sum(x_{i1}-\overline{x}_2)^2 & \sum(x_{i2}-\overline{x}_2)^2 & \cdots & \sum(x_{ip}-\overline{x}_2)^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sum(x_{i1}-\overline{x}_p)^2 & \sum(x_{i2}-\overline{x}_p)^2 & \cdots & \sum(x_{ip}-\overline{x}_p)^2 \end{pmatrix}$$

$$= (n-1)\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}$$

So this special matrix we found when squared actual gives us the value (n-1)S.

A useful Idempotent Matrix Construction

We are now going to take a new matrix. We are multiplying a $p \times 1$ column vector by its tranpose. The result is a $p \times p$ array of all ones. We then divide that by the number n and get the following.

$$\frac{11^T}{n} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

So $I - \frac{11^T}{n}$ is the following.

$$\begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{pmatrix}$$

This is an interesting matrix because it is idempotent and symmetric. Let's square the matrix and prove it.

Proof

$$(I - \frac{11^T}{n})(I - \frac{11^T}{n}) = I - \frac{11^T}{n} - \frac{11^T}{n} + \frac{1}{n^2}(11^T)(11^T)$$

Let's take a look at the last term

$$\frac{1}{n^2}(11^T)(11^T) = \frac{1}{n^2}\begin{pmatrix} n & n & \cdots & n \\ n & n & \cdots & n \\ \vdots & \vdots & \ddots & \vdots \\ n & n & \cdots & n \end{pmatrix}$$

$$= \frac{1}{n}11^T \rightarrow$$

$$(I - \frac{11^T}{n})(I - \frac{11^T}{n}) = (I - \frac{11^T}{n})$$

So we're going to use this property to find construct our variance/covariance matrix.

Variance Matrix

Remember our special identity $(x - \frac{11^T x}{n})^T(x - \frac{11^T x}{n}) = (n-1)S$? we're going to rewrite it and manipulate it.

$$\begin{aligned}(x - \frac{11^T x}{n})^T(x - \frac{11^T x}{n}) &= (x^T - \frac{x^T 11^T}{n})(x - \frac{11^T x}{n}) \\ &= x^T(I - 1/n11^T)(I - 1/n11^T)x \\ &= x^T(I - 1/n11^T)x \\ &= (n-1)S\end{aligned}$$

So now we have an effective method to create this vector in a smooth way without too many expensive calculations or awkward for loops.

8.3 Sample Mean and Variance of Linear combinations of x

Sample Mean

Let $x_1, x_2, ..., x_n \sim (\mu_{p\times 1}, \Sigma_{p\times p})$

We will also call $c_{p\times 1}$ to be a fixed vector. such that $c^T x_i = y_i$ What is the sample mean and sample variance of the y's?

$$\begin{aligned}\overline{y} &= \frac{\sum y_i}{n} \\ &= \frac{\sum c^T x_i}{n} \\ &= \frac{c^T \sum x_i}{n} \\ &= c^T \overline{x}\end{aligned}$$

Sample Variance

$$\begin{aligned}S_y^2 &= \frac{\sum (y_i - \overline{y})^2}{n-1} \\ &= \frac{\sum (c^T x_i - c^T \overline{x})^2}{n-1} \\ &= \frac{\sum (c^T(x_i - \overline{x}))^2}{n-1} \\ &= \frac{\sum (c^T(x_i - \overline{x}))(c^T(x_i - \overline{x}))}{n-1} \\ &= \frac{\sum c^T(x_i - \overline{x})(x_i - \overline{x})^T c}{n-1} \\ &= \frac{c^T \sum \left[(x_i - \overline{x})(x_i - \overline{x})^T\right] c}{n-1}\end{aligned}$$

$$= c^T S_x c$$

CHAPTER 9

Session 9: September 29, 2020

ninth

As a reminder from our last lesson, we constructed an identity that lets us calculate the sample mean and variance

$S = \frac{1}{n-1}x^T(I - \frac{1}{n}11^T)x$

Let's continue by defining a matrix D. It's going to keep the main diagonal of S and then throw out everything else.

$$D = \begin{pmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{pp} \end{pmatrix}$$

$$D^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{s_{pp}}} \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & \frac{s_{11}}{\sqrt{s_{11}s_{22}}} & \cdots & \frac{s_{1p}}{\sqrt{s_{11}s_{pp}}} \\ \frac{s_{12}}{\sqrt{s_{11}s_{22}}} & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{1p}}{\sqrt{s_{11}s_{pp}}} & \frac{s_{2p}}{\sqrt{s_{22}s_{pp}}} & \cdots & 1 \end{pmatrix}$$

$$R = D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$$

9.1 Review of last time

Take $x_1, ..., x_n \sim (\mu, \Sigma)$-Random

We take $c_{p\times 1}$ is fixed. We showed that $y_1, ..., y_n = c^Tx_1, ..., c^Tx_n \sim (c^T\mu, c^T\Sigma c)$

we also showed that $\overline{y} = c^T\overline{x}$ and $var(y) = c^TS_xc$

From this point forward we start with chapter 4 material and cover the multivariate normal distribution.

9.2 Reviewing Univariate Normal Distribution

Lets say that $x \sim N_1(0, \sigma^2), x \in \mathbb{R}$ this means that the pdf of x is $P(X = x) = f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$

The cumulative density function, cdf is defined by $P(X \leq x) = F(x) = \displaystyle\int_{-\infty}^{x} f(t)dt$ this integral has no closed form solution.

If $x \sim N(0, 1)$Because the distribution is symmetric we have that $E(x^{2k+1}) = 0$ for all k because of symmetry. If $x_1, ..., x_n \sim N(\mu, \sigma^2)$ then

$$\begin{aligned}\hat{\mu}_{MLE} &= \overline{x} \\ \hat{\sigma}^2_{MLE} &= \frac{\sum(x_i - \overline{x})^2}{n} \\ M_x(t) = E[e^{tX}] &= e^{\mu t + \frac{1}{2}\sigma^2 t^2}\end{aligned}$$

add appendix on MGFs

9.3 Multivariate Normal Distribution

$$x_{p\times 1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

We say that x follows a p-Dimensional Normal distribution with mean of $\mu_{p\times 1}$ and covariance matrix of $\Sigma_{p\times p}$ or $x \sim N_p(\mu, \Sigma)$

$$f(\vec{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{1/2}} e^{\frac{-(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}}$$

We can think of the exponent as the generalized distance from x to μ

Bivariate Distribution

Let's examine the simplest case of multivariate distributions: the bi-variate case. $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$

We also know that a couple of well known formulas

$$\begin{aligned}|\Sigma| &= \sigma_{11}\sigma_{22} - \sigma_{12}^2 \\ \Sigma^{-1} &= \frac{1}{\sigma 11 \sigma 22 - \sigma 12^2} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{23} & \sigma_{11} \end{pmatrix}\end{aligned}$$

write little thing about determinants of 2x2

$$
\begin{aligned}
(x-\mu)^T\Sigma^{-1}(x-\mu) &= \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix} \frac{1}{\sigma 11\sigma 22-\sigma 12^2} \begin{pmatrix} \sigma_{22} & -\sigma 12 \\ -\sigma_{12} & \sigma 11 \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= \frac{\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}-\sigma_{12}^2} \\
&= \frac{\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-2\rho\sqrt{\sigma_{11}\sigma_{22}}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}(1-\rho^2)} \\
&= \frac{1}{1-\rho^2}\left[\Big(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\Big)^2+\Big(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\Big)^2-2\rho\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\cdot\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right]
\end{aligned}
$$

So now we can rewrite our bivariate normal distribution. We break up the generalized distance quadratic term $(x-\mu)^T\Sigma^{-1}(x-\mu)$. We can see that the cause for the distance is the variation that's representedc by the standardized z-scores for x_1 and x_2 along with the third term that gives an adjusted z-score for the dimension between the x's. We won't actually work with the PDF's in this class because it's super tedious.

We will be able to do a lot of good things with our abbreviated notation. We're not going to look at it in this class like we would in calculus. It's not a function of variables, instead we're going to think of it as a special mathematical object with its own rules and operations.

9.4 Areas of Equal Probability

Now suppose we want to find all $x \in \mathbb{R}^p$ such that $f(x) = c^2 = const$

$$
f(\vec{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{1/2}} e^{\frac{-(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}} = c^2
$$

We can say that we want to find all the points that share the same generalized statistical distance from the mean described by the form $(x-\mu)^T\Sigma^{-1}(x-\mu)$. If f(x) is constant. This means that $(x-\mu)^T\Sigma^{-1}(x-\mu) = c_1^2$ where c_1 is another constant.

We can use our Algebra skills to solve for c_1^2

$$
c_1^2 = -2LN\big(c^2(2\pi)^{p/2}\Sigma^{1/2}\big)
$$

So in the two dimensional case this forms a little ellipse for us that gives us all of the points with equal probability. In the N dimensional case we get a football type shape.

The ellipse will have a center at μ and axes in the direction of $\vec{e}_i$ with length $k_1\sqrt{\lambda_i}$ From Chapter 2 of book.

9.5 Linear Transformations of Multivariate Normal

$x \sim N_p(\mu, \Sigma), a_{p\times 1}$-FIXED

Theorem

$$a^T x = \sum a_i x_i \sim N_1(a^T \mu, a^T \Sigma)$$

Before we proved that the distribution statistics held. Namely, the mean and variance held with affine transformations; however, the the persistence of the normality is a property unique to the normal distribution. This is why it is such an important distribution in Statistics.

CHAPTER 10

Session 9: October 1, 2020

tenth

From last time, we introduced the theorem

10.1 Linear Transformations of Multivariate Normal

$x \sim N_p(\mu, \Sigma), a_{p\times 1}$-FIXED

Theorem

$$a^T x = \sum a_i x_i \sim N_1(a^T\mu, a^T\Sigma)$$

Before we proved that the distribution statistics held. Namely, the mean and variance held with affine transformations; however, the the persistence of the normality is a property unique to the normal distribution. This is why it is such an important distribution in Statistics.

Now we can actually use a fixed matrix $A_{q\times p}$ and we will find that

Theorem

$$\underset{q\times p}{A}\underset{p\times 1}{x} \sim N_q(A\mu, A\Sigma A^T)$$

The left side represents q different linear combinations of the elements x and we find that it actually conforms to a q-dimensional normal distribution.

10.2 Linear Combinations of Multivariate Normal

Here are some other properties of Normal Distributions

$$\begin{aligned} x &\sim N_p(\mu, \Sigma), d_{p\times 1} - const \\ x + d &\sim N_p(\mu + d, \Sigma) \end{aligned}$$

We can see that by adding a constant vector d we are just shifting over the mean of the distribution.

10.3 Partitions of Randomly Distributed Variables

Let's say take our random vector and partition it .

$$\underset{p\times 1}{x} = \begin{pmatrix} \underset{q\times 1}{x_1} \\ \underset{(p-q)\times 1}{x_2} \end{pmatrix}, x \sim N_p(\mu, \Sigma)$$

$$\mu = \begin{pmatrix} \underset{q\times 1}{\mu_1} \\ \underset{(p-q)\times 1}{\mu_1} \end{pmatrix}, \Sigma = \begin{pmatrix} \underset{q\times q}{\Sigma_{11}} & \underset{q\times(p-q)}{\Sigma_{12}} \\ \underset{(p-q)\times q}{\Sigma_{21}} & \underset{q\times q}{\Sigma_{22}} \end{pmatrix}$$

Distributions of Normal Partitions

We can now ask ourselves what the distributions of the partitions x_1 and x_2 are. To do that, we are going to construct a matrix A_i such that when I multiply $A_i x = x_i$ We can begin our construction

$$\begin{aligned} \underset{q\times p}{A_1} &= \begin{pmatrix} I_{q\times q} & 0_{q\times(p-q)} \end{pmatrix} \\ A_1 x &= x_1 \end{aligned}$$

Here's what happens in the matrix. When we multiply, the I portion of the matrix when multiplied keeps x unchanged. But it will only multiply with the first p values of x. The last p-q values get multiplied by a matrix of 0's. so then. When we multiply it from the right we get rid of all the rows that aren't in x_1. When we multiply it from the right, we get rid of all the columns that aren't going to be included.

$$\begin{aligned} A_1 x &\sim N_q(A_1\mu, A\Sigma A_1^T) \\ &\sim N_(\mu_1, \Sigma_{11}) \end{aligned}$$

For fun we do the same thing to the other partition

$$\begin{aligned} \underset{(p-q)\times p}{A_2} &= \begin{pmatrix} 0_{0\times p} & I_{(p-q)\times p} \end{pmatrix} \\ A_2 x &= x_2 \\ A_2 x &\sim N_{(p-q)}(A_2\mu, A_2\Sigma A_2^T) \\ &\sim N_{(p-q)}(\mu_2, \Sigma_{11}) \end{aligned}$$

10.4 Correlation Vs. Independence

Theorem: let $\underset{q_1\times 1}{x}$ and $\underset{q_2\times 1}{x}$.

1

If they are independent, then their covariance and correlations will be 0. That is

$$cov(x_1, x_2) = cor(x_1, x_2) = 0_{q_1\times q_2}$$

$$
\begin{aligned}
cov(x_1, x_2) &\overset{Def}{=} E(x_1, x_2) - \mu_1 \mu_2^T \\
&\overset{ind}{=} E(x_1)E(x_2) - \mu_1 \mu_2^T = \mu_1 \mu_2^T - \mu_1 \mu_2^T = 0
\end{aligned}
$$

This is because covariance is calculated by taking the expected value of the products minus the minus the product of the expected values. Thanks to the independence, the expectation of the product is equal to the product of the expectation.

2

If

$$
\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_{q_1+q_2}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} \right) \Rightarrow
$$

x_1 and x_2 are independent if and only if $\Sigma_{12} = 0$

In the multivariate case, zero correlation and independence are equivalent statements for the normal distribution. This is not the case for all distributions, so in the words of Cyril, "this result is very cool".

Proof

If $\Sigma_{12} = 0_{q_1 \times q_2}$, $\Sigma_{21} = 0_{q_2 \times q_1}$ then x_1 and x_2 are independent.

Idea $\underset{jointpdf}{f(x_1, x_2)} = f(x_1) \cdot f(x_2)$ That is the joint pdf should be equal to the product of the marginal pdfs.

$$
f(x) = \frac{1}{(2\pi)^{(q_1+q_2)/2} \begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix}} \mathrm{e}^{\begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}}
$$

$$
\begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}||\Sigma_{22}|
$$

$$
\begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix}^{-1} = \begin{vmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{vmatrix}
$$

$$
\begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}||\Sigma_{22}|
$$

$$
\begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix}^{-1} = \begin{vmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{vmatrix}
$$

$$
\begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}||\Sigma_{22}|
$$

$$
\begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} = \begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \Sigma_{11}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} + \begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \Sigma_{22}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}
$$

$$
f(x) = \frac{1}{(2\pi+)^{q_1/2}|\Sigma_{11}|} e^{\begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \Sigma_{11}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}} \times
$$

$$
\frac{1}{(2\pi+)^{q_1/2}|\Sigma_{22}|} e^{\begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \Sigma_{22}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}}
$$

10.5 Conditional Distributions of Normal

Theorem

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, x \sim N(\mu, \Sigma)$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}, |\Sigma_2 2| > 0$$

then the Conditional Distribution of X_1 given that $X_2 = x_2$:

$$X_1 | X_2 = x_2 \sim N_{q \times 12}\Big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big)$$

10.6 Univariate Conditional Probability

Take x and y

$$P(X|Y = y) \overset{def}{=} \frac{f(x,y)}{f(y)}$$

That is the conditional probability is the ratio between the bivariate joint pdf and the univariate pdf.

We can get the univariate by integrating out the variable you don't need out of the joing distribution.

$$f(y) = \int_{\infty}^{\infty} f(x,y)dx$$

We call this integrating x out of the joint distribution. This is where we ended for the day to be continued next lecture.

CHAPTER 11

Session 11: October 6, 2020

From last time we will copy and paste the theorem

11.1 Conditional Distributions of Normal

Theorem

$$
\begin{gathered}
x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, x \sim N(\mu, \Sigma) \\
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}, |\Sigma_2 2| > 0
\end{gathered}
$$

then the Conditional Distribution of X_1 given that $X_2 = x_2$:

$$
X_1 | X_2 = x_2 \sim N_{q \times 12}\Big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big)
$$

$$
f(x_1 | X_2 = x_2) \overset{def}{=} \frac{f(x_1, x_2}{f(x_2)} \tag{11.1}
$$

$$
\tag{11.2}
$$

We are going to do an indirect proof with a trick. We are going to start by defining a matrix

$$
\begin{gathered}
A_{p \times p} = \left(\begin{array}{c|c} I_{q \times q} & -\Sigma_{12}\Sigma_{22}^{-1} \\ \hline 0 & I_{(p-q)(p-q)} \end{array}\right) \\
A(x - \mu) = A \left(\begin{array}{c} x_1 - \mu_1 \\ \hline x_2 - \mu_2 \end{array}\right) = \left(\begin{array}{c} x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \hline x_2 - \mu_2 \end{array}\right) \\
x \sim N_p(\mu, \Sigma), \quad x - \mu \sim N_p(0, \Sigma) \\
A(x - \mu) \sim N_p(0, A\Sigma A^T) \\
A\Sigma A^T = \left(\begin{array}{c|c} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ \hline 0 & I \end{array}\right) \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right) \left(\begin{array}{c|c} I & 0^T \\ \hline (-\Sigma_{12}\Sigma_{22}^{-1})^T & I \end{array}\right)
\end{gathered}
$$

$$
A\Sigma A^T = \left(\begin{array}{c|c} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right) \left(\begin{array}{c|c} I & 0^T \\ \hline (-\Sigma_{12}\Sigma_{22}^{-1})^T & I \end{array}\right)
$$

$$A\Sigma A^T = \left(\begin{array}{c:c} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0^T \\ \hdashline 0 & \Sigma_{22} \end{array}\right) \Rightarrow$$

$$A(x-\mu) = \left(\begin{array}{c} x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \hdashline x_2 - \mu_2 \end{array}\right) \sim N_p(0, A\Sigma A^T)$$

From the theorem before, since we have two partitions with no covariance, this means that we have two independent vectors $x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ and $x_2 - \mu_2$. This means the following:

$$x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \sim N_q(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$x_2 - \mu_2 \sim N_{p-q}(0, \Sigma_{22})$$

If we condition the above on x_2 then x_2 is a fixed value and the only random part is x_1. So we can derive the formula below.

$$x_1|X_2 = x_2 \sim N_q(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

So to summarize what happened, We chose a special fixed vector and used the properties of that fixed vector to create a new vector with a new distribution. Conveniently this new vector has two independent partitions. If we recenter the first partition, we get a distribution of x_1 that is independent of $x_2 - \mu_2$. By definition this makes it the conditional probability.

Now you're probably wondering How we came up with this indirect proof and constructed the matrix. We basically had the end goal in mind. We did something on purpose that gave us the observation. It is the cumulative effort of generations of great minds that form mathematics.

11.2 Distribution of Quadratic forms

chi-square distribution.

Many test statistics follow chi-square distributions. It is one of the most important distributions in statistics. Contingency tables and goodness of fit analysis statistics follow the chi-square distribution as well.

It is also the building block of the T and F distributions.

$$z \sim N(0,1) \Rightarrow \tag{11.3}$$

$$z^2 \sim \chi^2(1) \tag{11.4}$$

$$z_1^2 + z_2^2 + ... + z_k^2 \sim \chi^2(k) \tag{11.5}$$

$$t = \frac{N(0,1)}{\sqrt{\frac{\chi^2(1)}{v}}} \sim z(v) \tag{11.6}$$

$$F = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2} \sim F(v_1, v_2) \tag{11.7}$$

$$GAM(2,k) \sim \chi^2(2k) \tag{11.8}$$

if

$$x \sim \chi^2(n)$$

then

$$E(x) = n$$

$$VAR(x) = 2n$$

as the degrees of freedom increase, the chi-square distribution starts to look normal.

It is a single parameter distribution which gives it very special properties and lets it be easily transcribed onto a table.

Theorem

The following quadratic form has a chi-square distribution

$$(x-\mu)^T \Sigma (x-\mu) \sim \chi^2(p)$$

Proof

Remember that the sum of square normal variables is going to follow a chi square distribution. So we are going to try to use a tool in our box and represent this quadratic form as the sum of squares. The best tool we have for representing difficult matrix problems as scalar is the spectral decomposition of Σ. Remember that since it's an inverse matrix, its eigenvalues will be the inverse eigen values.

Remember that X is a random variate. let $\lambda_1, \lambda_2, ..., \lambda_n$ be the eigenvalues of Σ

$$(x-\mu)^T \Sigma^{-1}(x-\mu) \overset{SD}{=} \sum_{i=1}^{p} (X-\mu)^T \lambda_i^{-1} e_i e_i^T (X-\mu)$$

$$= \sum_{i=1}^{p} \lambda_i^{-1} (X-\mu)^T e_i e_i^T (X-\mu)$$

$$= \sum_{i=1}^{p} \lambda_i^{-1} [e_i^T (X-\mu)]^2$$

$$= \sum_{i=1}^{p} \left[\frac{e_i^T (X-\mu)}{\sqrt{\lambda_i}}\right]^2$$

So now we just need to verify that each term of the sum is a standard normal variable. $z \sim N(0,1)$

We will use our analogy from baby statistics. when we wanted to create a standard normal we took $z = \frac{x-\mu}{\sigma} \sim N(0,1)$ This is because of the linear transformation properties of the normal distribution. We don't have division when we talk about matrices, but we have something similar.

we know that $X - \mu \sim N_p(0, \Sigma)$

We are going to define a matrix A

$$A = \begin{pmatrix} e_1^T / \sqrt{\lambda_1} \\ \vdots e_p^T / \sqrt{\lambda_p} \end{pmatrix}$$

Now we multiply A times our vector and see our result is exactly the same as the terms we have above

$$A(X-\mu) = \begin{pmatrix} e_1^T/\sqrt{\lambda_1} \\ \vdots \\ e_p^T/\sqrt{\lambda_p} \end{pmatrix} (X - mu) \sim N_p(0, A\Sigma A^T)$$

$$\begin{aligned} A\Sigma A^T &= \begin{pmatrix} e_1^T/\sqrt{\lambda_1} \\ \vdots \\ e_p^T/\sqrt{\lambda_p} \end{pmatrix} \left(\sum_{i=1}^{p} \lambda_i e_i e_i^T \right) \begin{pmatrix} e_1/\sqrt{\lambda_1} & \cdots & e_p/\sqrt{\lambda_p} \end{pmatrix} \\ A\Sigma A^T &= \begin{pmatrix} e_1^T/\sqrt{\lambda_1} \\ \vdots \\ e_p^T/\sqrt{\lambda_p} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} e_1 & \cdots & \sqrt{\lambda_p} e_p \end{pmatrix} \\ &= I_{p\times p} \end{aligned}$$

This the fact that the eigen-vectors disappear is because they are all orthogonal and of length 1.

$$A(x-\mu) \sim N_p(0, I_{p\times p}$$

So that means all of the components of the vector are independent N(0,1) random variables.

Back to the proof. This means that the quadratic we calculated earlier follows a chi-square with p degrees of freedom

$$(x-\mu)^T \Sigma (x-\mu) = \sum_{i=1}^{p} \left[\frac{e_i^T (X-\mu)}{\sqrt{\lambda_i}} \right]^2 \sim \chi^2(p)$$

CHAPTER 12

Session 12: October 8, 2020

We starterd with a quick review of an old theorem of Linear Combinations of Vectors

12.1 Theorem 4.8

If we have two vectors that are orthogonal c and b, we have that

$$v_1 = c_1x_1 + c_2x_2 + ... + c_nx_n$$

$$v_2 = b_1x_1 + b_2x_2 + ... + b_nx_n$$

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \sim \mathbb{N}_{2p}\left(\begin{pmatrix} \sum c_i\mu \\ \sum b_i\mu \end{pmatrix}, \begin{pmatrix} (\sum c_i^2)\Sigma & (\sum b_ic_i)\Sigma \\ (\sum b_ic_i)\Sigma & (\sum b_i^2)\Sigma \end{pmatrix} \right)$$

If we were to stack the x's ontop of each other we would get the following vector. Each x is a px1 vector.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \sim N_{np}\left(\begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma \end{pmatrix} \right)$$

So we're going to multiply this by a matrix.

Find a matrix A such that

$$A_{2p\times np}\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}_{np\times 1}=\begin{pmatrix}v_1\\v_2\end{pmatrix}_{2p\times 1}=\begin{pmatrix}c_1x_1+c_2x_2+...+c_nx_n\\b_1x_1+b_2x_2+...+b_nx_n\end{pmatrix}$$

$$A=\begin{pmatrix}c_1I & c_2I & ... & c_nI\\ b_1 & b_2I & ... & b_nI\end{pmatrix}$$

$$Ax=\begin{pmatrix}v_1\\v_2\end{pmatrix}\sim N_{2p}(A\mu_x,A\Sigma A^T)$$

$$A\Sigma A^T=\begin{pmatrix}(\sum c_i^2)\Sigma & (\sum b_ic_i)\Sigma\\ (\sum b_ic_i)\Sigma & (\sum b_i^2)\Sigma\end{pmatrix}$$

12.2 Maximum Likelihood Estimation for Mutlivariate Normal Distribution pg. 168

The idea is that when you collect data

$$x_1,x_2,...,x_n\sim \underset{unknown}{N_p(\mu,\Sigma)}$$

Your job as a statistician is to estimate the population statistics μ,Σ. So far we've already been doing that by taking

$$\hat{\mu}=\overline{x},\hat{\Sigma}=S=\frac{\sum_{i=1}^n(x_i-\overline{x})(x_i-\overline{x})^T}{n-1}$$

Now we've already calculated this before and verified that these values are unbiased estimators. We did this by taking the $E(\overline{x})=\mu$ and $E(S)=\Sigma$.

In statistics estimators come from a deeper place. For this purpose we define something called a maximum likelihood estimator.

Likelihood

The likelihood is the probability that we have observed our given data set as a function of the unknown parameters μ and Σ

Likelihood $=L(\mu,\Sigma)$

The parameters that we choose, we are able to define our estimators as the values that maximize the likelihood given our parameters.

$$f(x_1,x_2,...,x_n)=\prod_{i=1}^n f(x_i)$$

We write likelihood function by taking the product of the individual probability functions of the vectors. They are random variates, so they are independent from one another.

Maximizing likelihood

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$
$$= \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}\Sigma|^{1/2}} e^{-(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)/2}$$

The x's are observed and we are looking at this as a function ofμ and Σ. We are effectively estimating p parameters for μ and $p(p+1)/2$ for Σ. It is a very complicated function with a lot of variables and it will be impossible to do with pure calculus.

In order to maximize this function with so many unknowns and a lot of tedious calculation, we're going to use a little trick and do it without any derivatives.

We're going to use just matrix inequalities and matrix algebra to maximize this likelihood.

Review of Linear Algebra

Let's say we have a quadratic form

$$x^TAx = TR(x^TAx) = TR(Axx^T)$$

The quadratic form is equal to its trace, because it is a one dimensional object. We cal also remind ourselves that

$$TR(A) = \sum \lambda_i$$
$$TR(A) = TR(P^T\Lambda P) = TR(\Lambda P^TP) = TR(\Lambda) = \sum \lambda_i$$

Trace Trick

$$\begin{aligned}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) &= TR((x_j-\mu)^T\Sigma^{-1}(x_j-\mu)) \\ &= TR((\Sigma^{-1}x_j-\mu)^T(x_j-\mu)) \\ \sum(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) &= \sum TR((\Sigma^{-1}x_j-\mu)^T(x_j-\mu)) \\ &= TR\left(\Sigma^{-1}\left[\sum(x_j-\mu)(x_j-\mu)^T\right]\right)\end{aligned}$$

So this is a good stopping point to ltalk about what's going on. We're just manipulating the quadratic form part and figuring out what happens when you have a sum of these guys, because a product of an exponent is the sum of powers. We are now going to just explore the sum now.

$$\sum(x_j-\mu)(x_j-\mu)^T = \sum(x_j-\overline{x}+\overline{x}-\mu)(x_j-\overline{x}+\overline{x}-\mu)^T$$
$$= \sum\big((x_j-\overline{x})+(\overline{x}-\mu)\big)\big((x_j-\overline{x})+(\overline{x}-\mu)\big)^T$$
$$= \sum(x_j-\overline{x})(x_j-\overline{x})^T + \cancel{\sum(x_j-\overline{x})(\overline{x}-\mu)^T} + \cancel{\sum(\overline{x}-\mu)(x_j-\overline{x})^T} + \sum(\overline{x}-\mu)(\overline{x}-\mu)^T$$

so in the end we have

$$\sum(x_j-\mu)(x_j-\mu)^T=\sum(x_j-\overline{x})(x_j-\overline{x})^T+\sum(\overline{x}-\mu)(\overline{x}-\mu)^T$$

So now we can rewrite our original likelihood function using our new constructed identities.

$$L(\mu,\Sigma)=(2\pi)^{-np/2}|\Sigma|^{-n/2}e^{TR\left(\Sigma^{-1}\left[\sum(x_j-\overline{x})(x_j-\overline{x})^T+n(\overline{x}-\mu)(\overline{x}-\mu)^T\right]\right)}$$

$$=(2\pi)^{-np/2}|\Sigma|^{-n/2}e^{TR\left(\Sigma^{-1}\sum(x_j-\overline{x})(x_j-\overline{x})^T\right)+TR\left(\Sigma^{-1}n(\overline{x}-\mu)(\overline{x}-\mu)^T\right)}$$

$$=(2\pi)^{-np/2}|\Sigma|^{-n/2}e^{TR\left(\Sigma^{-1}\sum(x_j-\overline{x})(x_j-\overline{x})^T\right)+TR\left((\overline{x}-\mu)^T\Sigma^{-1}n(\overline{x}-\mu)\right)}$$

$$=(2\pi)^{-np/2}|\Sigma|^{-n/2}e^{TR\left(\Sigma^{-1}\sum(x_j-\overline{x})(x_j-\overline{x})^T\right)+(\overline{x}-\mu)^T\Sigma^{-1}n(\overline{x}-\mu)}$$

Now this looks worse than our original result... buuut, we have a different theorem that will help nus out.

Maximization of Quadratic forms

Let $B_{p\times p}$ be a positive definite matrix and b-real number then,

$$\frac{1}{|\Sigma|^b}e^{-TR(\Sigma^{-1}B)/2}\leq\frac{1}{|B|^b}(2b)^{bp}e^{-bp}$$

For all positive definite matrices $\Sigma_{p\times p}$ equality only holds for $\Sigma=B\frac{1}{2b}$

so this theorem is exactly what we need to maximize our function without any calculus. It seems like the best possible function; however, it also seems very difficult still. It actually has p less parameters than our original. We only need to prove this and then we're in the clear.

$$\frac{1}{|\Sigma|^b}e^{-TR(\Sigma^{-1}B)/2}\leq\frac{1}{|B|^b}(2b)^{bp}e^{-bp}$$

We know two facts

$$TR(A)=\sum\lambda_i$$
$$|B|=\prod\lambda_i$$

We now need to choose one of the two matrices to get the eigen values.

Let $\eta_1,\eta_2...\eta_p$ be the eigen vectors of $B^{1/2}\Sigma B^{1/2}$. We choose this because it lets us use one substitution for both variables due to the properties of trace.

$$TR(B^{1/2}\Sigma B^{1/2})=TR(\Sigma B)$$
$$B^{-1/2}\Sigma^{-1}B^{-1/2}$$
$$=|B^{-1/2}||\Sigma^{-1}||B^{-1/2}|=\frac{|B|}{|\Sigma|}=\prod\eta_i\rightarrow$$
$$\frac{1}{|\Sigma|}=\frac{\prod\eta_i}{|B|}$$

Cool so we're almost there. we have a really handy identity that gives us most of what we need. We used acouple of tricks that will let us substitute and use our proofs to get a solution.

CHAPTER 13

Session 13: October 13, 2020

13.1 Recap

$B_{p\times p}, b > 0$, Then

$$\frac{1}{|\Sigma|^b} e^{TR(\Sigma^{-1}B)/2} \leq \frac{1}{|B|}(2b)^{pb}e^{-pb}$$

For all Positive Definite matrices $\Sigma_{p\times p}$ and equality iff $\Sigma = \frac{1}{2b}B$

So as a reminder of last time we have this identity above. We want to use eigenvalue properties as a tool to process this information more efficiently. Our trick was that we assigned $\eta_1, ..., \eta_p$ as the eigenvectors of $B^{1/2}\Sigma^{-1}B^{1/2}$ We know that this is positive definite because Sigma is positive definite which means $y^T B^{1/2}\Sigma^{-1}B^{1/2}y > 0 \forall y$

We identified two very nice properties of the trace and identity using η and this lead us to a new identity namely

$$TR(B^{1/2}\Sigma B^{1/2}) = TR(\Sigma B)$$

$$B^{-1/2}\Sigma^{-1}B^{-1/2}$$

$$= |B^{-1/2}||\Sigma^{-1}||B^{-1/2}| = \frac{|B|}{|\Sigma|} = \prod \eta_i \rightarrow$$

$$\frac{1}{|\Sigma|} = \frac{\prod \eta_i}{|B|}$$

Using these we rewrite the left side as the following

$$\prod_{i=1}^{p} \frac{e^{-\eta_i/2}\eta_i^b}{|B|}$$

but B is fixed, so we are able to ignore it. To maximize this function we need only maximize the numerator

$$f(\eta_i) = e^{-\eta_i/2}\eta_i^b$$

Now we just need to maximize the contributions of each eta.

So now that we have a product of things that we can maximize. and if we maximize the value individual pieces, we will maximize the value of the sum.

13.2 Maximization

$f(x) = e^{-x/2}x^b - MAX$

$$\begin{aligned} f'(x) &= -\frac{1}{2}e^{-x/2}x^b + be^{-x/2}x^{b-1} \\ &= e^{-x/2}x^{b-1}(-\frac{x}{2} + b) \end{aligned}$$

We then set it to 0.

$$0 = e^{-x/2}x^{b-1}(-\frac{x}{2} + b)$$

We can then see that $x = 2b$ as the critical value that gives us the maximum. We could be more rigorous and take the second derivative, but this is not a calculus class.

13.3 Backplug

$$\begin{aligned} \frac{1}{|\Sigma|^b}e^{TR(\Sigma^{-1}B)/2} &\leq \frac{1}{|B|}(2b)^{pb}e^{-pb} \\ \frac{1}{|\Sigma|^b}e^{TR(\Sigma^{-1}B)/2} &= \prod_{i=1}^{p} \frac{e^{-\eta_i/2}\eta_i^b}{|B|} \end{aligned}$$

is maximized when $\eta_1 = \eta_2 = ...\eta_p = 2b$ This means we can actually immediate find the maximum by plugging in our maximizing value.

$$max \prod_{i=1}^{p} \frac{e^{-\eta_i/2}\eta_i^b}{|B|} = \prod_{i=1}^{p} \frac{e^{-b}(2b)^b}{|B|}$$

Now we need to find out what that B is going to be in our original proof. Remember our original goal was to show that the maximum likelihood estimators of mean and variance/covariance are our sample formulas.

$$L(\mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}e^{-(TR\left(\Sigma^{-1}\sum(x_j-\overline{x})(x_j-\overline{x})^T\right)+(\overline{x}-\mu)^T\Sigma^{-1}n(\overline{x}-\mu))}$$

So we can actually see that based on our original function, the B that we need is going to be the matrix $(x_j - \overline{x})(x_j - \overline{x})^T$. The only difference between the form we have and the theorem we just proved is the term $(\overline{x} - \mu)^T\Sigma^{-1}n(\overline{x} - \mu)$

We have to maximize this, to maximize the likelihood, because it has a negative sign attached to it. This is going to be maximized only when $\overline{x} = \hat{\mu}$. With that in mind that part of the exponent disappears and we only need to do maximize our matrix with respect to Sigma.

$$L(\Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}e^{-TR\left(\Sigma^{-1}\sum(x_j-\overline{x})(x_j-\overline{x})^T\right)}$$

If we ignore all the constants we can see that we just need to make our choice in sigma. The thing that will maximize our based on our opening theorem will be $\hat{\Sigma} = \frac{1}{2b}B$ But remember that $b = \frac{n}{2}, B = \sum(x_j - \overline{x})(x_j - \overline{x})^T$ so...

$$\hat{\Sigma} = \frac{1}{n}\sum(x_j - \overline{x})(x_j - \overline{x})^T \tag{13.1}$$

From this point forward we played around with the iris data set and the MVN package in R. The actual implementation is trivial and can be found in the homework.

CHAPTER 14

Session 14: October 15, 2020

Today we started the chapter 5 material.

14.1 Inferences about the Mean vector

In statistics inference usually means 3 things. 1) Estimation $\hat{\mu}$
2) Confidence Intervals and Regions
3) Hypothesis Testing

14.2 Estimation

Let's take our data set

$$x_1, ..., x_n \sim N(\mu, \sigma)$$

How to estimate μ and σ^2

$$\hat{\mu} = \overline{x}, \quad \hat{\sigma}^2 = \sum \frac{x_i - \overline{x}}{n}, \quad s^2 = \sum \frac{(x_i - \overline{x})^2}{n-1}$$

We have two techniques that we will focus on. One is using unbiased estimators and the other is to use Maximum Likelihood Estimators.

14.3 Confidence Intervals

Confidence intervals are defined as $(1-\alpha) \cdot 100\%$ for μ and σ. Sometimes we want to give some more slack to our model. Instead of getting an exact estimate, we want to construct an interval where our population statistic has a 95% chance of being somewhere inside

If $\alpha = 0.05 \Rightarrow 95\%$ CI

If $\alpha = 0.01 \Rightarrow 99\%$ CI

We construct our confidence intervals like so

$$(\overline{x} - t_{1-\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}, \overline{x} + t_{1-\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}})$$

The t statistic is represented by the $(1-\frac{\alpha}{2})^{th}$ Percentile of the t-distribution with n-1 degrees of freedom.

One-Dimensional t-Distributions

$\overline{x} \sim N(\mu, \frac{\sigma^2}{n})$

$$\overline{x} - \mu \sim N(0, \frac{\sigma^2}{n})$$

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

The above are given distributions and transformations of the distribution of $\overline{x}$ from there we take the square. The square sum of a standard normal variables is a $\chi^2(p)$

$$\Big(\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\Big)^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

From here we can construct the t-distribution. This is the ratio between a standard normal and a chi-square with v degrees of freedom.

$$\frac{N(0,1)}{\sqrt{\frac{\chi^2(v)}{v}}} \sim t(v)$$

If we were to write it out and expand it a bit more, we have

$$\frac{\Big(\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\Big)}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}} = \frac{\overline{x} - \mu}{\frac{s}{n}} \sim t(n-1)$$

This is actually our pivotal quantity that we call the t-score. We can calculate pretty much everything here and we will only have to estimate μ. We are actually going to construct our confidence interval from this quantity.

The t-distribution looks a lot like the bell curve. It is centered at zero and has symmetric tails on either end. For any value of α, we can find the $(1 - \frac{\alpha}{2})^{th}$ percentile of the curve. This means the point at which $(1 - \frac{\alpha}{2})\%$ of the mass is less than or equal to that t-score. We call it $t_{1-\frac{\alpha}{2}}$

By symmetry $-t_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}}$. Remember that for all of this, we are referring to the t-distribution with (n-1) degrees of freedom.

So based on what we have constructed

$$P(-t_{1-\frac{\alpha}{2}}(n-1) < \frac{\overline{x} - \mu}{\frac{s}{n}} < t_{1-\frac{\alpha}{2}}(n-1)) = 1 - \alpha$$

Now we can rewrite the inequality leaving our unknown, μ in the middle.

$$P(\overline{x} - t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}) = 1 - \alpha$$

We can also construct a confidence interval for the variance using a different distribution and pivotal quantity. To be included in the appendix. Even though the actual population statistic is not observable we can still be relatively certain that our statistic lies somewhere inside the interval. Note that we will always settle for something less than a 100% confidence interval. This is because the only way to include all of the data points is if the confidence interval spans the entire space. It might be from $-\infty$ to ∞. This would not give us any useful information whatsoever.

14.4 Hypothesis Testing

This is probably the most important part of estimation. This is what we've been building up to.

Assume a distribution of the data

First we have to assume a distribution for our data.

$$x_1, ..., x_n \sim N_1(\mu, \sigma^2)$$

Identify a Hypothesis

In hypothesis testing we make a claim and use our statistical inference to test its validity.

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0 \qquad\qquad\qquad , \alpha = 0.05 \end{aligned}$$

H_0 is our null hypotheis. It makes the claim that μ is equal to a particular value μ_0. The H_A is our alternate hypothesis.

Find a test statistic

In this case, we are going to use the one we just saw.

$$\frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}} \sim t(n-1)}$$

before the μ was an unknown, but now it is a known value that was specified in our hypothesis. This test statistic will help us evaluate the probability that $\mu = \mu_0$ given our data.

Rejection Region approach

For this approach we create a rejection region. In our case it is the two $\frac{\alpha}{2}$ tails. We use our data to construct our t-statistic, if our falls inside of the rejection region, that is, it falls outside of our confidence interval, we reject the null hypothesis. If the t-statistic falls outside of the rejection region or inside of the confidence interval we fail to reject the null hypothesis. It is important to note that we can never be sure of anything in statistics because the world is random and uncertain.

We are rejecting our hypothesis if the corresponding t-value falls within the extreme α-percent of the distribution. We call the two endpoints of the rejection region our critical values.

P-Value Approach

P-Value is also called the observed significance level. Intuitively we can think of the p-value as a measure for the extremity of our data.

Imagine you're walking on the street and you see someone who is 7ft tall. You would recognize that this is a tall person. How do you know that? It is because they are taller than anyone else on the street and taller than anyone you've probably seen before.

This is the same idea as the test statistic. Your p-value is the percentage of the data that is less than or equal to your value. That is to say, we integrate from negative infinity to our test statistic

$$P - value = 2 \cdot P(t(n-1) \leq -|t|)$$

Once we compute the p-value we compare it against our alpha level. If our observed significance is less than the significance of our test that is our p value is less than α we reject the null hypothesis H_0. If our p-value is greater than our alpha level then we fail to reject the null.

14.5 Cyril's Preference

Dr. Rakovski acknowledged that the both methods are pretty equivalent in the sense that rejection and failure will occur at the same values of t. He prefers the p-value method, because when you know the p-value, you know exactly how far you are from rejecting. The rejection region gives you a binary information of whether or not your answer is acceptable.

There are some caveats to this. The p-value method tends to be dangerous in the hands of inferior mathematicians. There is a lot of precedence of p-hacking. They often manipulate the alpha level a posteriori.

If you would like to learn some good methods for hypothesis testing, look into the Bain and Engelhardt Probability book.

14.6 Simplest Multi-Dimensional Case

$$x_1, ..., x_n \sim N_p(\mu, \Sigma)$$

So now instead of measuring n random variables we are measuring n px1 random variates.

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0 \qquad\qquad\qquad , \alpha = 0.05 \end{aligned}$$

So now we need to invent some new math to make the unvariate case applicable to the multivariate

$$\begin{gathered} \overline{x} \sim N(\mu, \Sigma) \\ \overline{x} - \mu \sim N(0, \Sigma) \\ (\overline{x} - \mu)^T (\frac{\Sigma}{n})(\overline{x} - \mu) \sim \chi^2(p) \end{gathered}$$

Our big problem is that Σ is unknown. It's really difficult and expensive to generate the variance-covariance matrix.

So in the univariate case our test statistic was a standard Normal

$$\frac{\overline{x}-\mu}{\frac{\sigma}{n}} \sim N(0,1)$$

however, when we replaced the unknown standard deviation with the sample deviation, we lost one degree of freedom.

$$\frac{\overline{x}-\mu}{\frac{s}{n}} \sim t(n-1)$$

So we want to use something analogous to replace the Σ with its estimator S, the unbiased estimator $S = \frac{\displaystyle\sum_{i=1}^{n}(x-\overline{x})(x-\overline{x})^T}{n-1}$

As a side note: when we invert the matrix

$$(\frac{\Sigma}{n})^{-1} = n\Sigma^{-1}$$

so now we want to calculate

$$\begin{aligned}(\overline{x}-\mu)^T\left(\frac{\Sigma}{n}\right)^{-1}(\overline{x}-\mu) &\sim \chi^2(p)\\ n(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu) &\sim \chi^2(p) \Rightarrow\\ n(\overline{x}-\mu)^T S^{-1}(\overline{x}-\mu) &\sim \frac{(n-1)p}{n-p}F_{p,n-p}\end{aligned}$$

So when we replace the variance-covariance matrix with its unbiased estimator the price we pay is that our distribution is no longer a chi-square, but instead an F distribution, which is the ratio of two chi squared distributions and their means. That's fine though, we can still use the same techniques from our univariate case and construct everything we need to.

and then we did an example in R. The actual code is trivial compared to the theory, so I did not record in notes.

CHAPTER 15

Session 15: October 20, 2020

so last time, we tested our first multivariate hypothesis

$x_1, x_2, ..., x_n \sim N_p(\mu, \Sigma)$

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0, \quad \alpha = 0.05$$
$$n(\overline{x} - \mu_0)^T S^{-1}(\overline{x} - \mu_0) \sim \frac{p(n-1)}{n-p} F_{p,n-p}$$

We did tested this with a trivariate hypothesis and data set. very fun stuff...

Imagine that we test this hypothesis, and that we do not reject. That is the end of the story. This means that the vector of values supports the null hypothesis.

If we do end up rejecting the null hypothesis, we have more steps.

15.1 Alternative Hypothesis Explained

$H_A : \mu \neq \mu_0, \quad \alpha = 0.05$

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \neq \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}$$

When the alternative hypothesis is true, then there are many ways that it can be true. We can compare each mu with its corresponding hypothesis like in univariate. We need to do P independent samples in a t-test with equal variances.

We will use something called the Bonferonni's adjustment for multiple comparisons. We will divide our α by p to get $\alpha * = \alpha/p$

15.2 Bonferonni's Inequality

To understand the correction method, we must first understand a theorem from baby statistics.

$$P(\bigcap_{i=1}^{n} A_i) \geq 1 - P(\bigcap_{i=1}^{n} A_i^C)$$

That is the probability that all the hypotheses in A are true is greater than the complement of the sum of the probabilities that they are false individually.

15.3 Test statistics of linear transformations

Remember our test statistic $T^2 = n(\overline{x} - mu_0)^T S^{-1} (\overline{x} - mu_0)$
$x_1, x_2, ..., x_n \sim N_p(\mu, \Sigma)$

Define General Linear transformation of the X's

Let's say that we want to transform all of the x's. We would do this if we were for instance converting them between units of measure. We will define a general linear transformation of X
$\underset{p\times 1}{Y} = \underset{p\times p}{C} \underset{p\times 1}{X} + \underset{p\times 1}{d}$

This gives us a stronger theorem that let's us prove things more generally and not worry about linear transformations.

$$T_Y^2 = n(\overline{y} - mu_{y0})^T S^{-1} (\overline{y} - mu_{y0})$$

We want to know if we can use the same test statistic regardless of our transformation. We ahve

$$\begin{aligned} H_0 : \mu = \mu_0 &\Rightarrow H_0 : c\mu + d = c\mu_0 + d \\ H_A : \mu \neq c\mu_0 + &\Rightarrow H_0 : c\mu + d \neq c\mu_0 + d \\ \overline{y} &= c\overline{x} + d \\ S_y &= cS_x c^T \end{aligned}$$

using this knowledge, we would like to rewrite the test statistic starting form T_y

$$\begin{aligned} T_y^2 &= n(\overline{y} - mu_{y0})^T S^{-1} (\overline{y} - mu_{y0}) \\ &= n(c\overline{x}\cancel{d} - c\mu_0 - \cancel{d})^T (cS_x c^T)^{-1} (c\overline{x} + \cancel{d} - c\mu_0 - \cancel{d}) \\ &= n(\overline{x} - \mu_0)^T \cancel{c^T (c^T)^{-1}} S^{-1} \cancel{(c)^{-1}} c(\overline{x} - \mu_0) \\ &= n(\overline{x} - \mu_0)^T S^{-1} (\overline{x} - \mu_0) = T^2 \end{aligned}$$

we just proved that applying a linear transformation of x, we do nothing to the T^2 statistic.

15.4 Likelihood Ratio Test

Let's start with the basics. Fairy tales start "Once upon a time." Models in this class start

$$x_1, x_2, ..., x_n \sim N_p(\mu, \Sigma) H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

We now define the likelihood ratio

$$\frac{max_{H_0} L(\mu, \Sigma)}{max_{H_A} L(\mu, \Sigma)} = \Lambda$$
$$-2LN\Lambda \sim \chi^2(*)$$

where $-2LN\Lambda$ is our test statistic and * is the difference in the dimensions of the spaces the parameters roam under H_A and H_0.

This is a completely different approach but produces an identical result.

Definition of Likelihood Ratio

The likelihood of the data is defined as below

$$max_{H_0} L(\mu, \Sigma) = max_{H_0} \prod f(x = x_i | \mu_0, \Sigma) = max_{H_0} (2\pi)^{-np/2} |\Sigma_0|^{-n/2} e^{-\sum_{i=1}^{p}(x_i)(x_i - \mu_0)^T \Sigma (x_i - \mu_0)/2}$$

$$max_{H_A} L(\mu, \Sigma) = max_{H_A} \prod f(x = x_i | \mu_0, \Sigma) = max_{H_A} (2\pi)^{-np/2} |\Sigma|^{-n/2} e^{-\sum_{i=1}^{p}(x_i)(x_i - \mu)^T \Sigma (x_i - \mu)/2}$$

Derivation of Wilk's Likelihood

Plugging this information into our original expression we have

$$\frac{max_{H_0} L(\mu, \Sigma)}{max_{H_A} L(\mu, \Sigma)} = \frac{\cancel{(2\pi)^{-np/2}} |\hat{\Sigma}|^{n/2}}{\cancel{(2\pi)^{-np/2}} |\hat{\Sigma_0}|^{n/2}} \cdot \frac{\cancel{e^{-np/2}}}{\cancel{e^{-np/2}}}$$
$$\hat{\Sigma} = \sum (x_i - \overline{x})^T (x_i - \overline{x})/n$$
$$\hat{\Sigma_0} = \sum_{i=1}^{n} (x_i - \mu_0)(x_i = \mu_0)^T / n$$

Our simplified ratio $\frac{|\hat{\Sigma}|}{|\hat{\Sigma_0}|}$ is called Wilk's Lambda $= \Lambda^{2/n}$. The rejection region of this test statistic will be for large values.

We might be worried that you have two competing test methods. We will now prove that they are the same

proof

Theroem: $x_1, x_2, ..., x_n \sim N_p(\mu, \Sigma)$ we want to test

$$H_0 : \mu = \mu_0 H_A : \mu \neq \mu_0$$

Then the T^2 Test and the LR tests are equivalent.

$$\Lambda^{2/n} = (1 + \frac{T^2}{n-1})^{-1}$$

$$\Lambda^{2/n} = \frac{\hat{\Sigma}}{\hat{\Sigma_0}}$$

$$T^2 = n(\overline{x} - \mu_0)^T S^{-1}(\overline{x} - \mu_0)$$

so our challenge is connecting two determinants with a quadratic form. Our answer is singular value decomposition.

We need to construct a $(p+1) \times (p+1)$ matrix.

$$A = \left(\begin{array}{c:c} \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T & \sqrt{n}(\overline{x} - \mu_0) \\ \hdashline \sqrt{n}(\overline{x} - \mu_0)^T & -1 \end{array}\right)$$

$$A = \begin{pmatrix} \underset{p\times p}{A_{11}} & \underset{p\times 1}{A_{12}} \\ \underset{1\times p}{A_{21}} & \underset{p\times p}{A_{22}} \end{pmatrix}$$

$$|A| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}| = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}|$$

$$-1 \cdot \Big|\sum_{j=1}^{n}(x_j - \overline{x})(x_j - \overline{x}) + (\overline{x} - \mu_0)(\overline{x} - \mu_0)^T\Big|$$

$$= \Big|\sum_{i=1}^{n}(x_j - \overline{x})(x_j - \overline{x})\Big|\Big| - 1 - n(\overline{x} - \mu_0)\Big(\sum(x_j - \overline{x})(x_j - \overline{x})\Big)^{-1}(\overline{x} - \mu_0)\Big|$$

Claim

$$\Big|\sum_{j=1}^{n}(x_j - \overline{x})(x_j - \overline{x}) + (\overline{x} - \mu_0)(\overline{x} - \mu_0)^T\Big| = \sum(x_j - \mu_0)(x_j - \mu_0)^T$$

$$\sum(x_j - \mu_0)(x_j - \mu_0)^T = \sum(x_j - \overline{x} + \overline{x} + \mu_0)(x_j - \overline{x} + \overline{x} + \mu_0)^T$$

$$= \sum(x_j - \overline{x})(x_j - \overline{x})^T + n(\overline{x} - \mu_0)(\overline{x} - \mu_0)^T$$

Our trick was adding and subtrating x bar and then grouping them in pairs before distributing. The square terms come out and the cross terms are equal to zero.

but that equals $n\hat{\Sigma}_0$

and the term from above $\sum_{i=1}^{n}(x_j - \overline{x})(x_j - \overline{x}) = n\hat{\Sigma}$

So we can rewrite the equation above as follows

$$(-1)|n\hat{\Sigma}_0| = |n\hat{\Sigma}|| - 1 - n(\overline{x} - \mu_0)^T\Big(\sum(x_j - \overline{x})(x_j - \overline{x})\Big)^{-1}(\overline{x} - \mu_0)|$$

Now notice that the monstrosity on the right side looks pretty familiar. $T^2 = n(\overline{x} - \mu_0)^T(x_j - \overline{x})/(n-1)\Big)^{-1}(\overline{x} - \mu_0)$ all we need to do is introduce that n-1 term inside.

$$|\hat{\Sigma}_0| = |\hat{\Sigma}||1 + \frac{n-1}{n-1}n(\overline{x} - \mu_0)^T\Big(\sum(x_j - \overline{x})(x_j - \overline{x})\Big)^{-1}(\overline{x} - \mu_0)|$$

$$|\hat{\Sigma}_0| = |\hat{\Sigma}||1 + \frac{1}{n-1}n(\overline{x} - \mu_0)^T\Big(\sum(x_j - \overline{x})(x_j - \overline{x})/(n-1)\Big)^{-1}(\overline{x} - \mu_0)|$$

$$|\hat{\Sigma}_0| = |\hat{\Sigma}|\Big(1 + \frac{T^2}{n-1})$$

in the end we got do two things. 1. we brough the n-1 inside of the determinant to simplify it a little. 2. we got rid of the determinant signs since it is a one dimensional object. Other than that it's just recognizing the T^2 statistic and substituting.

So we showed a one-to-one monotonic relationship between the hotelling's T^2 and the likelihood ratio.

15.5 Confidence Regions

Imagine you want to find a $(1-\alpha)\cdot 100\%$ Confidence region for $\underset{p\times 1}{\mu}$

$\mathbf{x}_1, \mathbf{x}_2, ..., x_n \sim N_p(\mu, \Sigma)$ By definition our T statistic follows the following behavior.

$$n(\overline{x} - \mu)^T S^{-1}(\overline{x} - \mu) \sim \frac{(n-1)p}{n-p} \cdot F_{p,n-p}(1-\alpha) \Rightarrow$$

$$P\Bigg(n(\overline{x} - \mu)^T S^{-1}(\overline{x} - \mu) \leq \frac{(n-1)p}{n-p} \cdot F_{p,n-p}(1-\alpha)\Bigg) = 1 - \alpha$$

So all x's that satisfy this inequality listed above, will comprise the desired $(1-\alpha)\cdot 100\%$ Confidence region for $\underset{p\times 1}{\mu}$.

This creates an ellipsoid in a p-dimensional space centered at $\overline{x}$ with axes that are tilted in the directions of the eigenvectors of the Sample Variance matrix. Just like what we did in the beginning of the class.

CHAPTER 16

Session 16: October 22, 2020

statistical inference has three parts. 1. Point estimation 2. Confidence intervals and regions 3. Hypothesis Testing

So far we've done lots of point estimation and hypothesis testing; however, we have a lot to learn about confidence intervals. We have two was of point estimation with Maximum likelihood and unbiased estimators. We also have several approaches to Hypothesis testing with the T^2 statistic and the Likelihood Ratio Test.

We have much to learn about hypothesis testing since it's the most important part of statistics.

16.1 Creating confidence regions

Let's take $x_1, x_2, ..., x_n \sim N_p(\mu, \Sigma)$ We want to construct a $(1-\alpha)*100\%$ region for the mean μ that is we construct region in which we have a $1-\alpha$ probability that alpha is within that region. it is a region and not an interval because we are doing things in multiple dimensions.

We are doing this with unspecified alpha even though the standard is $\alpha = 0.05$

We will call our confidence region W. we want to find $W \subset \mathbb{R}^p : P(\mu \in W) = 1-\alpha$ we did this in one dimension by taking the bell curve and constructing the confidence intervals. We just want to choose an interval that has $1-\alpha*100\%$ of the mass.

Although there is no hypothesis testing in confidence intervals, our confidence interval helps us to construct the test statistic and vice versa. We will use the T^2 test statistic to construct our confidence region

$$T^2 = n(\overline{x}-\mu)^T S^{-1}(\overline{x}-\mu) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

By definition $P\Big(n(\overline{x}-\mu)^T S^{-1}(\overline{x}-\mu)\Big) \leq \frac{(n-1)p}{n-p} \cdot F_{p,n-p}(1-\alpha) = 1-\alpha$ Thus all values that satisfy the inequality $n(\overline{x}-\mu)^T S^{-1}(\overline{x}-\mu) \leq \frac{(n-1)p}{n-p} \cdot F_{p,n-p}(1-\alpha)$ comprise the desired $(1-\alpha)*100\%$ confidence region for μ.

So this will be an ellipsoid that shares the eigen vectors of the sample. Remember that all the powers of a matrix vectors. The axes of this ellipsoid will be of length $\sqrt{\frac{(n-1)\cdot p}{n(n-p)} F_{p,n-p}(1-\alpha)}$

The nice thing about this is that we can get a really precise measure of confidence intervals.

The downside of this method is that it's hard to imagine and visualize a p-dimensional ellipsoid and we are would need to calculate a quadratic form to calculate it.

We are better off if we can get a nicer shape to imagine like an interval.

The equivalent of an interval in n-dimensional space will be an n-dimensional cartesian product. In a two dimensional space that will be a rectangle. It will be a prism in a three dimensional and so on and so forth.

16.2 Cartesian Product of Intervals

Let's imagine that we have a bunch of intervals for the individual μ_i's

Let's say $I_1 \times I_2 \times ... \times I_p = R$ We immediately know if a point in inside the region as long as each $\mu_i \in I_i d$

A Basic Approach Midterm Question

The first thing that people tried doing was to project an ellipse onto the axes. They called that region the confidence region for the mean. The projection contained the ellipse, so the probability of the mean vector being in this region was greater than $1 - \alpha$.

How do you find the projection on the coordinate axis. That is how do you find the shadow of the ellipse on the axis. This was on a test a couple of years ago. and was on the test this year. Go back and copy and paste midterm here.

Projection of an Ellipse onto its EigenVectors

The most clear region to use is to use the eigenvectors as the axes. These get us nicer looking regions.

The projection of an ellipse onto the ith eigenvector is
$\lambda_i \pm \sqrt{\frac{p(n-p)}{n(n-p)} F_{p,n-p}(1-\alpha)} \sqrt{\frac{s_i i}{n}}$

16.3 Bonferroni Simultaneous Confidence Intervals

The idea behind this method is to get component by component analyses of the dimensions so that we can think of things in one dimensional intervals instead of regions and projections. We want to make things look really nice and be human-readable without the drawback of being quite so conservative.

We need to construct confidence intervals c for each component of μ such that $\mu_i \in c_i$

$c_1 \times c_2 \times ... \times c_p = (1-\alpha)100\%$ CR for μ

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

We want the case to be for $\mu_1 \in c_1$ and $\mu_2 \in c_2$ we want

$P(\mu_{2\times1} \in CR = c_1 \times c_2) \overset{?}{=} 1 - \alpha$

So he said if you constructed standard confidence intervals for each component. If you were to check and see if your vector is inside the confidence region you would notice that as $p \to \infty$ You actually get a lower probability that the vector is inside the region. If you capture each individual component with 95% confidence. You cannot expect that the product will be 95% confidence as well.

So his argument startswith bool's inequality of subadditivity.

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

$$P(\mu \in c_1 \times c_2) = 1 - P(\bigcup_{i=1}^{p} \mu_i \notin c_i) \geq$$
$$1 - P(\mu \notin c_1) - P(\mu \notin c_2)$$

but each $P(\mu \notin c_i) = \alpha$ so the probability that μ belongs to the two intervals in the 2D case is only 0.9 even though α is 0.95.

So he asked himself, What if we divided the alpha. By taking a smaller value, we get a better probability that all intervals are satisfied. We have to be more confident about each individual dimension to be confident about the result as a whole.

So if $P(\mu_1 \in c_1) = 97.5\%$ and $P(\mu_2 \in c_2) = 97.5\%$ then $P(mu \in c_1 \times c_2) = 95\%$

He said that each interval should be $(1 - \frac{\alpha}{p} \cdot 100\%$ confidence intervals for μ_i each of these will be c_i. That is $P(\mu_i \in c_i = (1 - \frac{\alpha}{p} \cdot 100\%$ This results in $c_1 \times c_2 \times ... \times c_p = CR$ that contains μ with a probability of $\overline{x}_i \pm t_{n-1}(1 - \frac{\alpha}{2p})\sqrt{\frac{s_i i}{n}}$

It is similar to our first method using the T^2 -statistic; however it is more conservative due to the adjustment of dividing the individual alphas by $2p$.

Proof

$$P(\bigcap_{i=1}^{p} \mu_i \in c_i) = 1 - P(\bigcup_{i=1}^{p} \mu_i \notin c_i)$$
$$\geq 1 - \sum_{i=1}^{p} P(\mu_i \notin c_i)$$
$$\geq 1 - \Big(\sum_{i=1}^{p} 1 - P(\mu_i \in c_i)\Big) = 1 - \Big(\sum_{i=1}^{p} \frac{\alpha}{p}\Big) = 1 - \alpha$$

There's a great picture representation of all of these methods on page 233 of the book.

CHAPTER 17

Session 17: October 27, 2020

17.1 Comparison of Several Multivariate Means

Now that we are running comparisons with multiple samples, the complexity increases a bit, but the machinery stays the same.

One sample Paired t-Test

Imagine we have a 1 dimensional normal sample

$$x_1, x_2, ..., x_n \sim N_1(\mu_1, \sigma_1^2)$$

$$y_1, y_2, ..., y_n \sim N_1(\mu_1, \sigma_2^2)$$

Usually we test the hypothesis

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2 \alpha = 0.05$$

That is, we would be testing the hypothesis that there was no effect of the treatment.

So these are the same people measured twice. We want to know whether or not the mean scores change after a treatment. This treatment can be time passing by, or it can be some kind of intervention. Usually they are taught something or they were given something to change them.

So to take care of this what we do is try and define a sample of differences $d_i = y_i - x_i$

$$d_1, d_2, ..., d_n \sim N_1(\mu_d, \sigma_d^2)$$

$$H_0 : \mu_d = 0$$
$$H_A : \mu_d \neq 0$$
$$t = \frac{\overline{d}}{\frac{s_d}{\sqrt{n}}} \sim t(n-1)$$

So we want to write a $(1-\alpha) \cdot 100\%$ CI for $\mu_2 - \mu_1$ and it's going to be $\left(\overline{d} - t_{1-\alpha/2}(n-1)\frac{s_d}{\sqrt{n}}, \overline{d} + t_{1-\alpha/2}(n-1)\frac{s_d}{\sqrt{n}}\right)$

So we are going to write it out exactly like before $x_1, x_2, ..., x_n \sim N_p(\mu_1, \Sigma_1)$ and $y_1, y_2, ..., y_n \sim N_p(\mu_1, \Sigma_1)$ So this would be observing several different people at different timepoints. Now all the variables are random vectors instead. Maybe it's the same people measured twice on P variables.

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2 \alpha = 0.05$$

So we're still going to form these pair-wise differences:

$$D_1 = y_1 - x_1, D_2 = y_2 - x_2, ..., D_n = y_n - x_n \sim N_p(\mu_D, \Sigma_D)$$

$$H_0 : \mu_D = 0$$
$$H_A : \mu_D \neq 0$$

This pair-wise reduces our problem to a one sample problem. This is Chapter 5 material. $T^2 = n(\overline{D} - 0)^T S^{-1}(\overline{D} - 0) \sim \frac{(n-1)p}{n-p} \cdot F_{p,n-p}$

So the $(1 - \alpha)100\%$ CR for $\mu_D = \mu_2 - \mu_1$ We have 3 constructions for this from chapter 5. The ellipse, the projection onto the eigenvectors and the bonferonni confidence intervals.

Bonferonni Simultaneous Confidence Intervals

$$\mu_D = (\mu_{D1}, \mu_{D2}, ..., \mu_{Dp})$$
$$\mu_{Di} \in I_i : (\overline{D} - t_{1-\frac{\alpha}{2p}} \frac{s_{ii}}{\sqrt{n}}, \overline{D} + t_{1-\frac{\alpha}{2p}} \frac{s_{ii}}{\sqrt{n}})$$

We will then see if our mean vector is in the cartesian product of these intervals.

17.2 Repeated Measures Design for Comparing Treatments

So what we did was improve the t-tests to increase the numbers of observation per treatment. We want to further generalize this. We wan to be able to do things for multiple timepoints and still have one thing

$$x_{11}, x_{12}, ..., x_{1n}$$
$$x_{21}, x_{22}, ..., x_{2n}$$
$$x_{q1}, x_{q2}, ..., x_{qn}$$

We are taking univariate measurements at multiple timepoints.

$$x_j = \begin{pmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jq} \end{pmatrix}$$

This is a longitudinal study. All observations in the subject j are represented as above.

We are interested in if the mean changes over time or over the course of the treatment.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_q$$
$$H_A : \mu_i \neq \mu_j fori \neq j$$

Rewrite

$$H_0 : \mu_1 - \mu_2 = \mu_1 - \mu_3 = ... = \mu_1 - \mu_q$$
$$H_A : \mu_1 - \mu_i \neq \mu_1 - \mu_j$$
$$H_0 : \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$
$$H_A : \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

We are now going to rewrite it as a product of matrices

$$H_0 : \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}_{(q-1)\times q} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}_{q\times 1} = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{pmatrix}_{(q-1)\times 1}$$

$$x_1, xc2, ..., x_n \sim N_q(\mu, \Sigma)$$

We can just take this to be

$$H_0 : C\mu = 0$$
$$H_A : C\mu \neq 0$$

$$Cx_1, Cx_2, ..., Cx_n \sim N_{q-1}(C\mu, C\Sigma C^T) = N_{q-1}(\mu_y, \Sigma_y)$$

so now we have a new random sample of transformed variables and we can just use the techniques from number 5. Plug it into a t-test and be done.

$$T^2 = n(\overline{y} - 0)^T (S_y)^{-1} (\overline{y} - 0) \sim \frac{(n-1)(q-1)}{n-q+1} F_{q-1,n-q+1}$$

But remember, $y_i = Cx_i$, $overliney = C\overline{x}$, and $S_y = CS_xC^T$, so $T^2 = n(C\overline{x})^T (CS_xC^T)^{-1} C(overlinex)$

But remember, we can just use the same T^2 statistic as before. We just need to use the new Distribution though.

This comes up pretty frequently in data analysis

Check zoom for code notes.

CHAPTER 18

Session 18: October 29, 2020

So in the last class we used the manipulation introduction of contrast matrices to convert multivariate hypothesis into a univariate model by cleverly using our notation and by observing the distributions of the contrasts.

Sometimes we call what we did in the last class a repeated measurement analysis of variance.

18.1 Comparing Mean Vectors from Two Populations

$$x_1, x_2, ..., x_n \sim N_p(\mu_1, \Sigma_1)$$

$$y_1, y_2, ..., y_n \sim N_p(\mu_2, \Sigma_2)$$

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

In the univariate case (p=1), we have to set up our model

$$x_1, x_2, ..., x_n \sim N_p(\mu_1, \sigma_1)$$

$$y_1, y_2, ..., y_n \sim N_p(\mu_2, \sigma_2)$$

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Now this is different than our one sample problems. We have to deal with the variances. Before we deal with the means, we must do preliminary tests with our variances to see if they are equal or not.

Univariate Preliminary tests of the Variances

$$H_0 : \sigma_1 = \sigma_2$$
$$H_A : \sigma_1 \neq \sigma_2$$

Our test statistic will be a ratio of the sample variances

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 2)$$

Variances are the same S Pooled

if we fail to reject $H_0 : \sigma_1 = \sigma_2$, then we will estimate the common variance via a Pooled estimator

$$s^2_{Pooled} = \frac{s_1^2(n_1 - 1) - s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

$$t = \frac{\overline{x} - \overline{y}}{s^2_{pooled}(\frac{1}{n_1} + \frac{1}{n_2})} \sim t_(n_1 + n_2 - 2)$$

Now the variances are different and we can no longer pool them together. Instead we need to use a different test statistic.$\sigma_1 \neq \sigma_2$

Variances are different: Welch's t-Test Approximation

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} \sim t(\nu)$$

ν is a weird statistic used for the degrees of freedom

"It's weird but it works" - Dr. Cyril Rakovski, PhD.

$$\nu = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Multivariate Case

$$x_1, x_2, ..., x_n \sim N_p(\mu_1, \Sigma_1)$$

$$y_1, y_2, ..., y_n \sim N_p(\mu_2, \Sigma_2)$$

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

$$s_x = \frac{\sum_{i=1}^{n_1}(x_i - \overline{x})(x_i - \overline{x})^T}{n_1 - 1}$$

$$s_y = \frac{\sum_{i=1}^{n_2}(y_i - \overline{y})(y_i - \overline{y})^T}{n_2 - 1}$$

Variances are the same: S Pooled

$$\begin{aligned} S_{pooled} &= \frac{s_x(n_1-1)+s_y(n_2-1)}{n_1+n_2-2} \\ &= \left(\frac{n_1-1}{n_1+n_2-2}\right)s_x + \left(\frac{n_1-1}{n_1+n_2-2}\right)s_y \end{aligned}$$

We assume the distributions of x and y to be p dimensional normal.

$$\overline{x} \sim N_p(\mu_1, \frac{\Sigma_1}{n_1})$$

$$\overline{y} \sim N_p(\mu_2, \frac{\Sigma_2}{n_2})$$

From that we can determine that the distribution of their linear combination is as follows

$$\overline{x} - \overline{y} \sim N_p(\mu_1 - \mu_2, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$$

If we were to subtract the vector $\mu_1 - \mu_2$ we get

$$\overline{x} - \overline{y} - (\mu_1 - \mu_2) \sim N_p(0, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$$

Since $\mu_1 - \mu_2 = 0$ under the null...

$$\overline{x} - \overline{y} - \overset{H_0}{\sim} N_p(0, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$$

$$T^2 = (\overline{x} - \overline{y})^T[\Sigma(\frac{1}{n_1} + \frac{1}{n_2})]^{-1}(\overline{x} - \overline{y}) \sim \chi^2_p$$

Now our problem is that our test statistic has an unknown: Σ. Good thing we have an estimator ready: S_{pooled}

$$T^2 = (\overline{x} - \overline{y})^T\Big[S_{pooled}\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\Big](\overline{x} - \overline{y}) \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}$$

Variances are not the same: M-Box Test

$$x_1, x_2, ..., x_n \sim N_p(\mu_1, \Sigma_1)$$

$$y_1, y_2, ..., y_n \sim N_p(\mu_2, \Sigma_2)$$

$$\Sigma_1 \neq \Sigma_2$$

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 \neq \mu_2 \end{aligned}$$

1. Hope that the n_1, n_2 the sample sizes are Large (≥ 30), but hopefully a lot more.

$$\overline{x} - \overline{y} - (\mu_1 - \mu_2) \sim N_p(0, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$$

Since $\mu_1 - \mu_2 = 0$ under the null...

$$\overline{x} - \overline{y} - \overset{H_0}{\sim} N_p(0, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$$

$$\overline{x} \sim N_p(\mu_1, \frac{\Sigma_1}{n_1})$$

$$\overline{y} \sim N_p(\mu_2, \frac{\Sigma_2}{n_2})$$

$$\overline{x} - \overline{y} \sim N_p(\mu_1 - \mu_2, (\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}))$$

$$\overline{x} - \overline{y} \overset{H_0}{\sim} N_p(\mu_1 - \mu_2, (\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}))$$

$$T^2 = (\overline{x} - \overline{y})^T \left[\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right]^{-1} (\overline{x} - \overline{y}) \sim \chi^2_p$$

$$\overline{x} - \overline{y} \overset{H_0}{\sim} N_p(\mu_1 - \mu_2, (\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}))$$

$$T^2 = (\overline{x} - \overline{y})^T \left[\frac{s_1}{n_1} + \frac{s_2}{n_2}\right]^{-1} (\overline{x} - \overline{y}) \sim \chi^2_p$$

In large sample sizes, the sample variances are proper estimators so you can get away with it and it is provable by theorem. The exact distribution is chi-square infinity, but it becomes very close.

2. Unequal variances with small sample size. You kinda do the same test statistic, because there's nothing else. We still construct our T^2

$$T^2 = (\overline{x} - \overline{y})^T (\frac{s_x}{n_1} + \frac{s_y}{n_2})^{-1} (\overline{x} - \overline{y}) \sim \frac{\nu p}{\nu - p + 1} F_{\nu, \nu - p + 1}$$

ν is a pretty annoying variable that comes from page 294 of the book I'm pretty sure the form below is not correct, because Dr. Rakovski did not seem confident on this one in class.

$$\nu = \frac{p + p^2}{\sum_{i=1}^{2} \frac{1}{n_i} \left[TR\Big(\frac{1}{n_i} s_i (\frac{1}{n_1} s_1 + \frac{1}{n_2} s_2)^{-1} \Big) \right]^2}$$

Now we want to talk about constructing confidence intervals. This is a less interesting story.

18.2 Constructing Confidence Intervals

$$x_1, x_2, ..., x_n \sim N_p(\mu_1, \Sigma_1)$$

$$y_1, y_2, ..., y_n \sim N_p(\mu_2, \Sigma_2)$$

Our assumptions are the same. We are now going to construct a confidence region for the diference of the means $\mu_1 - \mu_2$

$$\overline{x} - \overline{y} \sim N_p(\mu_1 - \mu_2, (\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}))$$

$$T^2 = (\overline{x} - \overline{y})^T \left[\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right]^{-1} (\overline{x} - \overline{y}) \sim \chi^2_p$$

Variances are the same

$$(\overline{x}-\overline{y}-(\mu_1-\mu_2))^T \left[S_{pooled}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right](\overline{x}-\overline{y}-(\mu_1-\mu_2)) \leq \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}(1-\alpha)$$

so we can construct an ellipsoid for this centered at $\overline{x} - \overline{y}$ It will be along the eigenvalue of S_pooled

the rest of these are demonstrated by the last session's notes.

our bonferroni test statistic is

$$\left(\overline{x}_i - \overline{y}_i \pm t_{1-\frac{\alpha}{2p}}(n_1 + n_2 - 2)\sqrt{(s_{ii})_{pooled}(\frac{1}{n_1} + \frac{1}{n_2})}\right)$$

CHAPTER 19

Session 19: November 3, 2020, Practice Test

Practice test see practice exam for details

CHAPTER 20

Session 20: November 5, 2020, Midterm

Midterm exam. See pdf for details

CHAPTER 21

Session 21: November 10, 2020

Profile Analysis

We have p measurements and they can be tests. Given to two people or more groups of people. We want to know if the average of the group responses are "similar" We are going to define three major hypothesis and go from there.

1. Are The Profiles Parallel?

In real live the plots won't look perfectly parallel, but we want to know if they are significantly parallel. (Close enough) so that the differences from being parallel can be explained by sampling variation.

*maybe go back for the notes later...

We basically take the value of means for each group and connect them to be a piece-wise linear graph.

$$\mu_1^T = (\mu_{11}, \mu_{12}, ..., \mu_{1n})$$
$$\mu_2^T = (\mu_{21}, \mu_{22}, ..., \mu_{2n})$$

ith person from from the first group

$$x_{1i} = (x_{1i1}, x_{1i2}, ..., x_{1ip})$$

ith person from the second group

$$x_{2i} = (x_{2i1}, x_{2i2}, ..., x_{2ip})$$

$$x_{11}, x_{12}, ..., x_{1n_1} \sim N(\mu_1, \Sigma_1)$$
$$x_{21}, x_{22}, ..., x_{2n_2} \sim N(\mu_2, \Sigma_2)$$

$$H_{01} : \text{Profiles are Parallel} H_{A1} : \text{Profiles are not Parallel}$$

$$H_{01} : \mu_{1i} - \mu_{1i-1} = \mu_{2i} - \mu_{2i-1} i \in \{2, 3, ..., p\}$$

That is the slopes between any two points are equal between the two groups. Let's rewrite this as a matrix

$$H_{01}: \begin{pmatrix} \mu_{12} = \mu_{11} \\ \mu_{13} = \mu_{12} \\ \vdots \\ \mu_{1p} = \mu_{1(p-1)} \end{pmatrix} = \begin{pmatrix} \mu_{22} = \mu_{21} \\ \mu_{23} = \mu_{22} \\ \vdots \\ \mu_{2p} = \mu_{2(p-1)} \end{pmatrix}$$

We can also rewrite this as a product of matrices

$$H_{01} : c\mu_1 = c\mu_2$$

$$c_{(p-1)\times p} = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

c is the contrast matrix that we design to show the above relationships between the μ's

$$x_{11}, ..., x_{1n_1} \sim N_p(\mu_1, \Sigma)$$

$$x_{21}, ..., x_{2n_1} \sim N_p(\mu_2, \Sigma)$$

$$H_{01} : c\mu_1 = c\mu_2$$
$$H_{A1} : c\mu_1 \neq c\mu_2$$

$$cx_{11}, ..., cx_{1n_1 s} \sim N_{p-1}(c\mu_1, c\Sigma c^T)$$

$$cx_{21}, ..., cx_{2n_1 s} \sim N_{p-1}(c\mu_2, c\Sigma c^T)$$

now our problem is a simple two sample t-test with equal variances. Σ needs to be estimated, so we will use $S_{pooled} = \frac{(n_1-1)s_1+(n_2-1)s_2}{n_1+n_2-2}$ so we get

$$y_{11}, ..., y_{1n_1} \sim N_{p-1}(c\mu_1, c\Sigma c^T)$$

$$y_{21}, ..., y_{2n_2} \sim N_{p-1}(c\mu_2, c\Sigma c^T)$$

$$T^2 = (\overline{y_1} - \overline{y_2})\Big[(\frac{1}{n_1} + \frac{1}{n_2})cS_{pooled}c^T\Big]^{-1}(\overline{y_1} - \overline{y_2}) \sim \frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p}F_{(p-1)(n_1+n_2-p)}$$

$$T^2 = (c\overline{x}_1 - c\overline{x}_2)^T\Big[(\frac{1}{n_1} + \frac{1}{n_2})cS_{pooled}c^T\Big]^{-1}(c\overline{x}_1 - c\overline{x}_2) \sim \frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p}F_{(p-1)(n_1+n_2-p)}$$

In modeling you would describe this as there being a main effect by group but there is no group effect by question.

2. Are the Parallel Profiles of the Two groups Identical

If the profiles are parallel, we want to know if they are identical or on top of each other. That is are they Coincident profiles? If the distance between the two profiles is zero then we would do the following.

We could use the testing methods that we already know, but that doesn't take advantage of the fact that they are parallel.

$$H_{02} : 1^T \mu_1 = 1^T \mu_2$$
$$H_{A2} : 1^T \mu_2 \neq 1^T \mu_2$$

1^T adds up all the elements of the vector.

$$1^T x_{11}, ..., 1^T x_{1n_1} \sim N_1(1^T \mu_1, 1^T \Sigma 1)$$
$$1^T x_{21}, ..., 1^T x_{2n_2} \sim N_1(1^T \mu_2, 1^T \Sigma 1)$$

$$T^2 = (1\overline{x}_1 - 1^T \overline{x}_2)^T \Big[(\frac{1}{n_1} + \frac{1}{n_2}) 1^T S_{pooled} 1 \Big] (1\overline{x}_1 - 1^T \overline{x}_2)$$
$$\sim \frac{\cancel{(n_1 + n_2 - 2)}1}{\cancel{n_1 + n_2 - 1 - 1}} F_{1, n_1+n_2-2}$$

Which is why we can indeed use a univariate approach to this because the univariate t-statistic will follow the same distribution.

3. Are the Coincident Profiles Level

That is to say, we want to know if the coincident profiles are level, meaning that there is no effect from question to question.

$$H_{03} : \mu_1 = \mu_2 = ... = \mu_p$$
$$H_{A3} : \mu_i \neq \mu_j$$

$$H_{03} \mu_2 - \mu_1 = \mu_2 - \mu_1 = ... = \mu_p - \mu_{p-1}$$

So we can get the same contrast matrix as earlier

$$H_{03} : c\mu = 0 H_{A3} : c\mu \neq 0$$

so there is no difference between

$$x_{11}, ..., x_{1n_1}, x_{21}, ..., x_{2n_2} \sim N_p(\mu, \Sigma)$$
$$cx_{11}, ..., cx_{1n_1}, cx_{21}, ..., cx_{2n_2} \sim N_{p-1}(c\mu, c\Sigma c^T)$$

We can now create a test statistic.

$$T^2 = (n_1 + n_2)(c\overline{x})^T (cSc^T)^{-1} (c\overline{x}) \sim \frac{(n_1 + n_2 - 1)(p-1)}{n_1 + n_2 - (p-1)} F_{p-1, n_1+n_2-p+1}$$

Notice that now that we have established that they are coincident and thus have the same variance and are part of the same population, we no longer need to pool.

This profile analysis is a baby version of longitudinal analysis so it leads to deep results.

CHAPTER 22

Session 22: November 12, 2020

22.1 ANOVA and MANOVA

22.2 ANOVA

Anova stands for Univariate Analysis of Variance. We are going to start with a one-way ANOVA. It is a generalization of two sample t-test with equal variances to a model with more than two samples.

Let's say that we have g samples.

$$x_{11}, ..., x_{2n_1} \sim N_1(\mu, \sigma^2)$$

$$x_{g1}, ..., x_{gn_2} \sim N_1(\mu_g, \sigma^2)$$

First thing we must do is test that all the variances are equal. We call this the hypothesis of homogeneity.

You need to test all g samples for normality using Kramer von Mises or Shapiro-Wilkes.

You can then use one of the following Bartlett's test or Levene's Test for normality.

Once you determine that the variance is homogenous, you can now compare all the means and see if they are equal. Even though we are calling it Analysis of Variance, we are really interested in comparing the means.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_g$$
$$H_A : \mu_i \neq \mu_j$$
$$F = \frac{\sum_{i=1}^{g} n_i(\overline{x}_i - \overline{x})^2/(g-1)}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2/(n_1 + n_2 + ... + n_g)}$$
$$\sim F_{g-1, n_1+n_2+...+n_g-g}$$

we skipped the derivation for this, but let's see where we get the degrees of freedom.

let's define a matrix v such that

$$v = \begin{pmatrix} \overline{x}_1 - \overline{x} \\ \vdots \\ \overline{x}_1 - \overline{x} \\ \overline{x}_2 - \overline{x} \\ \vdots \\ \overline{x}_2 - \overline{x} \\ \vdots \\ \overline{x}_g - \overline{x} \\ \vdots \\ \overline{x}_g - \overline{x} \end{pmatrix}$$

each $\overline{x}_i$ appears n_i times.

Remember we have a sum on top of our statistic.

$$\sum_{i=1}^{y} n_i(\overline{x}_i - \overline{x})^2 = v^T v$$

let's consider g different vectors with 1 in the positions of the x_i's and 0 elsewhere. We can see that v is a linear combination of these vectors. we can see that this vector cannot roam more than a g dimensional space, since it is a linear combination of g vecotrs.

Now the degrees of freedom for the top is g-1 because if you do $1^T v$ you would get $\overline{x} - \overline{x} = 0$ thus the degrees of freedom of this vector is g-1. Since this result restricts one of the dimensions because the vector v is orthogonal to 1^T

Now we can test this global hypothesis that all the means are the same versus the alternative that they are not.

If $F \leq F_{1-\alpha}(g-1, n_1 + ... + n_g - g)$ if we fail to reject, we end the analysis right here.

Else $F > F_{1-\alpha}(g-1, n_1 + ... + n_g - g)$ we reject H_0 in favor of H_A. Usually this is the exciting part of the analysis. We know that one of the means is not like the rest, but you do not know where the differences are.

First Approach finding where differences Lie

To figure this out we can get all possible pairs and compare every mean using multiple two sample t-tests with equal variances.

there will be $\binom{g}{2} = \frac{g(g-1)}{2}$, t-tests

all of them have to be tested at Bonferonni corrected p-values $\binom{p-value_i \cdot g}{2}$ This is not in the book. which Dr. Rakovski thinks is a mistake.

Tukey's Honest Significance Difference Test

The second method is called Tukey's test.

22.3 MANOVA

We could reduce this problem into multiple ANOVA tests using Bonferonni, but let's show the theory anyways. We still have g groups but now they are p dimensional normal. Now we have g samples of n_i observations that are each p dimensional normal random vectors.

$$x_{11}, ..., x_{1n_1} \sim N_p(\mu_1, \Sigma)$$

$$x_{21}, ..., x_{2n_2} \sim N_p(\mu_2, \Sigma)$$

$$x_{g1}, ..., x_{gn_g} \sim N_p(\mu_g, \Sigma)$$

$$H_0 : \mu_1 = \mu_2 = ... = \mu_g \tag{22.1}$$

$$H_A : \exists i, j | \mu_i \neq \mu_j \tag{22.2}$$

To test our hypothesis we will generate the Wilk's Lamda test statistic

$$\Lambda^* = \frac{|\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)^T|}{|\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})(x_{ij} - \overline{x})^T|}$$

The lambda is a bit difficult to calculate and after you calculate it there are different cases.

P=2,g>2

$$p > 2, g \geq 2, \Big(\frac{\sum n_i - g - 1}{g - 1}\Big)\Big(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\Big) \sim F_{2(g-1), 2(\sum n_i - g - 1)}$$

p>1,g=3

$$p > 2, g \geq 2, \Big(\frac{\sum n_i - p - 2}{p}\Big)\Big(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\Big) \sim F_{2p, 2(\sum n_i - p - 2)}$$

Moral of the story is that the multivariate case is complicated. There is no one test statistic that will work and there are many different factors to consider.

We then did some code here. Look at the notes for guidance.

CHAPTER 23

Session 23: November 17, 2020

We pretty much finished chapter 6. The only thing we skipped was Growth curves, but we'll skip it.

We are going to temporarily skip chapter 7 and do chapter 8 instead.

23.1 Chapter 8: Principal Component Analysis

This is a method to engineer new variables. We know from other classes that if your dataset has particular variables, sometimes it's a good idea to manipulate them in some way so that you can build a better model and extract the information.

If you have a numerical variable you can cut it into categorical groups. For example if you have an age, you can separate them into old and young.

One problem you might encounter is having highly correlated data. When you take surveys, sometimes the designers will ask the same question multiple times. This is called colinearity. The ideal situation with principal component analysis is to take the existing values transforming them using the optimal linear combinations. These new variables created are called principal components.

All principal components are uncorrelated so you can feed them into statistical and machine learning models to perform optimally.

The principal components are ordered and there comes a point when the principal components are so unimportant that you can ignore them, thus letting you reduce the dimensionality of your model.

You might wonder "Why don't we always use this if it's so good?"

They can be a bit scrambled. The original variables can be very meaningful. Income, age, distance to a hospital. When you take a linear combination, all of this information is scrambled.

Let's say that we have p original variables

$$x = (x_1, x_2, ..., x_p)$$

$$cov(x) = \Sigma_{p \times p}$$

I want to create p new variables $y_1, y_2, .., y_p$ that are linear combinations of

$x_1, ..., x_p$

$$y_1 = a_1^T x = a_{11}x_2 + ... + a_{1p}x_p$$
$$y_2 = a_2^T x = a_{21}x_1 + ... + a_{2p}x_p$$
$$\vdots$$
$$y_p = a_p^T x = a_{p1}x_1 + ... + a_{pp}x_p$$

$$var(y_i) = var(a_i^T x) = a_i^T \Sigma a_i$$

Imagine that we are trying to associate one variable to an outcome. For instance, lets try and find the association between age and a disease. If everyone in the sample is the same age, will I be able to find the correlation? The answer is no. Variability in data helps us predict things.

Definition

Principal Components are p uncorrelated linear combinations of $x_1, ..., x_p$ That have the largest possible variance.

$$x^T = (x_1, x_2, .., x_p)$$
$$a^T = (a_1, a_2, ..., a_n)$$
$$y = a^T x = a_1 x_1 + ... + a_{p \times p}$$

So let's look at the quadratic form that represents the variance of Y $a^T \Sigma a$. our goal is to maximize the quadratic form. Remember our lemma from the beginning of class. In order to utilize that lemma, we need to restrict the norm of the vector a $|a| = 1$ That the vector a should be on the unit disc. We're doing this to prevent any cheating. We are standardizing the length of a and we are making the choice about direction and not length. *Tristan's note in statistic, size doesn't matter*

Our question then becomes a mission of finding the vector a on the unit shell such that we maximize variance of sigma.

The second principal component has to be uncorrelated with the first one that under the restriction maximizes the value of the quadratic form.

Once we have created all p principal components then we will be great.

This is a concept from chapter 2. (page 22)

Creating Principal Components

thm let

$$x^T = x_1, ..., x_p$$

$$cov(x) = \Sigma, (\lambda_1, e_1), (\lambda_2, e_2), .., (\lambda_p, e_p)$$

are the eigen value eigen vector pairs for Σ with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$

Then:

$$y_i = e_i^T x, var(y_i) = \lambda_i$$

This is a big result because it tells us exactly how to derive principal components. We just need to find the eigenvalues and eigenvectors of the variance matrix and they create the coefficients we need. The order of the eigenvalues then induce the order of the eigenvectors.

Now we are almost done with our coverage of principal components. We're going to do a quick proof where we maximize quadratic points for points along the unit sphere.(page 22)

$$\begin{gathered}
y_1 = a^T x \quad var(y_1) = a^T \Sigma a \\
max a^T \Sigma a = max \frac{a^T \Sigma a}{a^T a} \\
||a|| = 1 \\
max = \lambda_1
\end{gathered}$$

attained iff $a = \lambda_1$ remember we proved this using svd and doing changing of variables.

Verfying Principal Components are Uncorrelated

We also know that the maximum of $a^T \Sigma a$ orthogonal to the first k eigen values will be achieved if and only if $a = e_{k+1}$

remember that all of the covariances between y_i and y_j to be zero.

$$\begin{aligned}
cov(y_i, y_j) &= cov(e_i^T x, e_j^T x) \\
&= e_i^T cov(x, x) e_j \\
&= e_i^T \Sigma e_j \\
&= e_i^T \lambda_j e_j \\
&= \lambda_j e_i^T e_j \\
&= 0
\end{aligned}$$

so here we use the definition of eigenvectors to get that $\Sigma e_j = \lambda_j e_j$ and the fact that eigenvectors are orthogonal with one another. So this process is guaranteed to produce uncorrelated principal components

so far we have not managed to reduce the dimensions of the data; however, we have managed to create uncorrelated data.

Theorem: No Loss of Information

when we create new variables there is always a fear of destroying or losing some of the valuable information.

Theorem

$$\sum_{i=1}^{p} VAR(x_i) = \sum_{i=1}^{p} VAR(y_i)$$

so this theorem is saying that the process of principal components shifts the variance of the variables around. So that there are no co-variance terms.

$$\begin{aligned}\sum_{i=1}^{p} VAR(x_i) &= \sum_{i=1}^{p} \sigma_{ii} = TR(\Sigma) \\ &= \sum \lambda_i \\ &= \sum VAR(Y_i)\end{aligned}$$

remember that

$$VAR(y_i) = \lambda_i$$

and the total variance is

$$\lambda_1 + \lambda_2 + ... + \lambda_p$$

so

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

is the proportion of the variance due to ith principal component.

since the λ's are ordered, this means that the variance due to each principal component are decreasing.

Let's say we crab the first k principal components

$$y_1, ..., y_k$$

$$\frac{\lambda_1 + \lambda_2 ..., \lambda_k}{\lambda_1 + \lambda_2 ..., \lambda_p}$$

so at some point, for some k, this becomes close to 1.

Some people decide that the minimum number of principal components is determined by the minimum number of principal components such that the above proportion exceeds 0.8. after that, the other principal components are negligible. This can shrink the dimensionality of your data set to just a couple.

This heuristic is not necessarily best practices, but it is common practice. The problem is that means that 20% of the variance is still unexplained.

CHAPTER 24

Session 24: November 19, 2020

$$x = (x_1, ..., x_p) \quad cov(x) = \Sigma$$

let's take a random vector x with a covariance matrix Σ

any time you analyze data you have to think if your variables work best for your analysis or if you need to manipulate them to suit your needs. The same question stands when we decide to use principal components.

If Σ is a diagonal matrix, then there is no point in principal components. We will see that the principal components will be exactly the same as the original variables.

$$(x_1, x_2, ..., x_p) = x$$

$$y_1 = e^t x_1, ..., y_p = e_p^T x$$

24.1 Correlation of Principal Components

$$COR(y_i, x_k) = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{ii}}}$$

λ_i - the ith largest eigenvalue

$\sigma_{ii} = var(x_i)$ the ith element of the main diagonal of Σ

e_{ik} the kth component of the ith eigenvector.

our first goal is to rewrite x_k as $v^T x$ for some v

Imagine the vector with 1 in the kth position (0,0,...,1,..,0) an 0 elsewhere. Let's call this vector w_k We know that

$$w_k^T v = \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = x_k$$

Dr. Rakovski uses y_k, but I find that confusing since the principal components are Y.

$$\begin{aligned}
cov(y_i, x_k) &= cov(e_i^T x, w_k^T x) \\
&= e_i^T cov(x, x) w_k \\
&= e_i^T \Sigma w_k \\
&= w_k^T \lambda_i \Sigma \\
&= \lambda_i w_k^T e_i \\
&= \lambda_i e_{ik} \\
COR(y_i, x_k) &= \frac{cov(y_i, x_k)}{\sqrt{VAR(Y_i) VAR(X_k)}} \\
&= \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i \sigma_{kk}}} \\
&= \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}
\end{aligned}$$

If we feel passionate about some of the original variables, we can choose the principal components that are highly correlated with the original variables. The idea is to reduce our problem to two components if possible, so that we can graph and explain them.

Sidenote

When you have raw data, you can extract principal components from either the correlation matrix or the covariance matrix.

It is at first tempting to say it doesn't matter. Since correlation is just a rescaled covariance matrix.

Both are used in real life; however, the choice does matter since they produce different results.

The correlation matrix is re-scaled to the point where all of the data will have the same value for the main diagonal: 1. This is a loss of information because we don't have any way to construct a covariance matrix from the correlation matrix.

Having the the correlations with a main diagonal values of 1; that can be good for some machine learning methods.

24.2 Two Extreme Cases of Principal Component Analysis

Original Variables are Uncorrelated

$$x_1, x_2, ..., x_p$$

are uncorrelated

$$\Sigma \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{pmatrix}$$

$$\lambda_i = \sigma_i, ..., \lambda_p = \sigma_p$$

$$\begin{aligned} e_1 &= (1,0,0,...,0) \\ e_2 &= (0,1,0,...,0) \\ e_p &= (0,0,0,...,1) \\ y_i &= e_i^T x = x_i \end{aligned}$$

So if things are uncorrelated, then there is no need to do anything new with principal component analysis.

Original Variables are Highly Correlated

This is where principal component analysis really shines.

$$\begin{aligned} \Sigma &= \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \\ &= \sigma^2 \varrho \end{aligned}$$

In this specific case there is no difference between the correlation
To find the eigenvalues we solve

$$|\varrho - \lambda I| = 0$$

$$\begin{vmatrix} 1-\lambda & \rho & \cdots & \rho \\ \rho & 1-\lambda & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1-\lambda \end{vmatrix} = 0$$

if we add all the other columns to the first column we get the following

$$\begin{vmatrix} 1-\lambda+(p-1)\rho & \rho & \cdots & \rho \\ 1-\lambda+(p-1)\rho & 1-\lambda & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ 1-\lambda+(p-1)\rho & \rho & \cdots & 1-\lambda \end{vmatrix} = 0$$

Now if we have a row that is all the same, we can take it outside of the determinant

$$\Big(1-\lambda+(p-1)\rho\Big) \begin{vmatrix} 1 & \rho & \cdots & \rho \\ 1 & 1-\lambda & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \rho & \cdots & 1-\lambda \end{vmatrix} = 0$$

From here we know that one of the eigenvalues is $1 + (p-1)\rho$

We are now going to take our first column, multiplying it by $-\rho$ and adding it to the rest of the columns

$$\Big(1-\lambda+(p-1)\rho\Big)\begin{vmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1-\lambda-\rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1-\lambda-\rho \end{vmatrix} = 0$$

This is now a upper triangular matrix, so the determinant of this matrix will be the product of the main diagonal. The determinant is $(1-\lambda-p)^{p-1}$

This means $\lambda_2 = \lambda_3 = ... = \lambda_p = 1-\rho$

$$y_1 = e_1^T x = \sum \frac{x_i}{\sqrt{p}} \approx \overline{x}$$

$$\frac{\lambda_1}{\lambda_1+\lambda_2+...+\lambda_p} = \frac{1+(p-1)\rho}{TR(|\rho|)} = \frac{1+(p-1)\rho}{p}$$

in this extreme case a single principal component replaces all of the variables.

CHAPTER 25

Session 25: November 24, 2020

Thanksgiving Break

CHAPTER 26

Session 26: November 26, 2020

Thanksgiving Break

CHAPTER 27

Session 27: December 1, 2020

Code day. Easy stuff. check canvas.

CHAPTER 28

Session 28: December 3, 2020

28.1 Multivariate Linear Regression

There are books on regression only that span 7-800 pages, but the book only has one chapter. It's pretty good still.

Linear regression is a statistical tool that allows you to estimate the importance of a bunch of variables with respect to an outcome variable of interest.

Some goals of this are to describe the importance of predicting or changing the outcome variable of interest.

We want to understand how a bunch of variables affect a variable that is a primary interest. If you have access to a bunch of variables then you can easily collect them, put them in a model and predict a highly interesting variable. We might try to predict stocks, weather, or something else.

We call this outcome variable the dependent variable. This needs to be a continuous variable or else our technique will not work.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

We call the data covariates, predictors and independent variables

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Our goal is to express each y_i as a linear combination of the covariates.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + ... + \beta_p x_{1i}$$

This can be written a bit more succinctly as a matrix multiplication. Notice how the β's are shared by every observation. Since they are all shared, this fit is impossible. We need to add error terms.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + ... + \beta_p x_{1i} + e_i$$

Each subject will have its own error term e_i

$$\underset{n\times 1}{Y} = \underset{n\times(p+1)}{Z}\underset{(p+1)\times 1}{\beta} + \underset{n\times 1}{e}$$

$$\underset{n\times 1}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \underset{n\times(p+1)}{Z} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\underset{(p+1)\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \underset{n\times 1}{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

In this model, the Z and β matrices are fixed while the e is random.

$$e_i \sim N(0, \sigma_e^2)$$

in the univariate case so

$$e \sim N_n(0, \sigma_e^2 I_{n\times n})$$

This is called a homoscedasticity of error terms meaning that they are all uncorrelated and have the same distribution.

$$Y \sim N_n(Z\beta, \sigma^2 I)$$

Let's talk about what is known and unknown.

Y is known and collected

Z is known and observed

B and σ^2 are unknowns and desirable.

So we have the p+1 β terms to estimate and one σ term. We need to estimate a total p+2 terms.

We will aim to learn how to estimate the vector and parameter as well as what to do once they are estimated.

28.2 Estimating The Coefficients: Beta

Imagine a scatter plot. We will try to write the equation of a line that best represents the relationship in the variables of the plot. Our objective is to minimize the sum of the distances between the line and the points.

$$Y = Z\beta + e$$

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - ... - b_p x_{ip})^2 = SS(b)$$

This is a measure of how well a model with $\beta = b$ fits the data. This is known as the sums of squares. If we minimize SS(b), we will build the best model.

This can be written as the following.

$$SS(b) = (Y - Zb)^T(Y - Zb)$$

$$Y - Z\beta = \begin{pmatrix} y_1 - b_0 - b_1x_{11} - ... - b_px_{1p} \\ y_2 - b_0 - b_1x_{21} - ... - b_px_{2p} \\ \vdots \\ y_n - b_0 - b_1x_{n1} - ... - b_px_{np} \end{pmatrix}$$

Theorem

The vector that minimizes $SS(b)$ is $\hat{\beta} = (Z^TZ)^{-1}Z^TY$

That is this estimator of β are the coefficients to the hyper-plane that fits the model as close to the plot as possible.

Proof

Show

$$SS(b) = (Y - Zb)^T(Y - Zb) \geq SS(\hat{\beta})$$

For this proof we Add and subtract $Z\hat{\beta}$ to introduce the term to our math, commute and associate them so that we create a multiplication of binomials. We multiply and distribute (FOIL) the binomials created to get four terms. The first term is clearly $SS(\hat{\beta}$. The fourth terms represents the length of a vector, while the second and third terms are each zero. Since the square of a vector's length is always non-negative ≥ 0 we can now say we have proven our theorem.

$$\begin{aligned} SS(b) &= (Y - Zb)^T(Y - Zb) \\ &= (Y - Zb + Z\hat{\beta} - Z\hat{\beta})^T(Y - Zb + Z\hat{\beta} - Z\hat{\beta}) \\ &= \Big[(Y - Z\hat{\beta}) + (Z\hat{\beta} - Zb)\Big]^T\Big[(Y - Z\hat{\beta}) + (Z\hat{\beta} - Zb)\Big] \\ &= (Y - Z\hat{\beta})^T(Y - Z\hat{\beta}) + (Y - Z\hat{\beta})^T(Z\hat{\beta} - Zb) + (Z\hat{\beta} - Zb)^T(Y - Z\hat{\beta}) + (Z\hat{\beta} - Zb)^T(Z\hat{\beta} - Zb) \\ &= SS(\hat{\beta}) + \|Z\hat{\beta} - Zb\|^2 \Rightarrow \\ SS(b) &= SS(\hat{\beta}) + \|Z\hat{\beta} - Zb\|^2 \Rightarrow \\ SS(b) &\geq SS(\hat{\beta}) \end{aligned}$$

Now this is only true if we can show that the second and third terms are zero. Since they are transposes of one anther we need only show that $(Y - Z\hat{\beta})^T(Z\hat{\beta} - Zb) = 0$. We claimed that $\hat{\beta} = (Z^TZ)^{-1}Z^TY$, but never used it. This definition ensures that the two vectors $(Y - Z\hat{\beta})^T$ and $(Z\hat{\beta} - Zb)$ are orthogonal and thus their product is 0.

$$\begin{aligned}
(Y - Z\hat{\beta})^T(Z\hat{\beta} - Zb) &= (Y - Z\hat{\beta})^T Z(\hat{\beta} - b) \\
&= (Y - Z(Z^TZ)^{-1}Z^TY)^T Z(\hat{\beta} - b) \\
&= (Y^T - Y^TZ(Z^TZ)^{-1}Z^T)Z(\hat{\beta} - b) \\
&= (Y^TZ - Y^TZ\cancel{(Z^TZ)^{-1}(Z^TZ)})(\hat{\beta} - b) \\
&= (Y^T - Y^T)Z(\hat{\beta} - b) \\
&= 0_n^T(\hat{\beta} - b) = 0
\end{aligned}$$

We are now done estimating β, but we are not done with estimation. We still need to estimate σ^2

28.3 Estimating Variance of the residuals

By estimating $\hat{\beta}$ we can estimate the model predicted values $\hat{Y} = Z\hat{\beta}$ from there we can get the estimated residuals $\hat{\varepsilon} = Y - \hat{Y}$ and estimate their variance.

$$\varepsilon_1, \varepsilon_2, ..., \varepsilon_n \sim N(0, \sigma^2)$$

We can reasonably estimate σ^2 by calculating the sample variance $\hat{\sigma}^2 = \frac{\sum(\hat{\varepsilon_i} - \overline{\varepsilon})^2}{n-1} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-1}$, but that would be incorrect. We actually need to adjust the denominator to include the number of variables we estimated to get the model coefficients. our modified sample variance is now $\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p-1}$

There is one important thing that maybe we have missed.

28.4 One More Thing

Remember linear regression is defined $Y = Z\beta + e$. Our assumption is that the error terms are normally distributed so

$$e \sim N(0, \sigma^2 I)$$
$$Y \sim N(Z\beta, \sigma^2 I)$$

But remember that the $\hat{\beta}$ is just a linear combination of Z and Y $(Z^TZ)^{-1}Z^TY$

$$\hat{\beta} \sim N_{p+1}(AZ\beta, A\sigma^2 IA^T)$$

$$\begin{aligned}
\hat{\beta} &\sim N_{p+1}(AZ\beta, A\sigma^2 IA^T) \\
\hat{\beta} &\sim N_{p+1}((Z^TZ)^{-1}Z^TZ\beta, (Z^TZ)^{-1}Z^T\sigma^2 I((Z^TZ)^{-1}Z^T)^T) \\
\hat{\beta} &\sim N_{p+1}(\beta, \sigma^2(Z^TZ)^{-1}) \\
c\hat{o}v(\hat{\beta}) &= \hat{\sigma}(Z^TZ)^{-1}
\end{aligned}$$

What is very important is that we have determined that $\hat{\beta}$ is an unbiased estimator of the true coefficients that we are after. We can also use this to determine the covariance of $\hat{\beta}$. We can also use this to determine which coefficients are important. That is we can determine which variables are statistically significant predictors of Y.

28.5 Hypothesis Testing in Linear Regression

$$
\begin{aligned}
H_0 &: \beta_1 = 0 && x_1 \text{ is not important} \\
H_A &: \beta_1 \neq 0 \\
t &= \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \sim Z(n-p-1) && \text{t-test for } H_0
\end{aligned}
$$

Rewatch video for the R-code

CHAPTER 29

Session 29: December 8, 2020

$$e \sim N(0, \sigma^2 I)$$

$$Y \sim N(Z\beta, \sigma^2 I)$$

But remember that the $\hat{\beta}$ is just a linear combination of Z and Y $(Z^T Z)^{-1} Z^T Y$

$$\hat{\beta} \sim N_{p+1}(AZ\beta, A\sigma^2 I A^T)$$

$$\begin{aligned}
\hat{\beta} &\sim N_{p+1}(AZ\beta, A\sigma^2 I A^T) \\
\hat{\beta} &\sim N_{p+1}((Z^T Z)^{-1} Z^T Z\beta, (Z^T Z)^{-1} Z^T \sigma^2 I ((Z^T Z)^{-1} Z^T)^T) \\
\hat{\beta} &\sim N_{p+1}(\beta, \sigma^2 (Z^T Z)^{-1}) \\
c\hat{o}v(\hat{\beta}) &= \hat{\sigma}(Z^T Z)^{-1}
\end{aligned}$$

$$\begin{aligned}
H_0 &: \beta_1 = 0 && x_1 \text{ is not important} \\
H_A &: \beta_1 \neq 0 \\
t &= \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \sim Z(n-p-1) && \text{t-test for } H_0
\end{aligned}$$

29.1 Building a Multivariate Model

a) There are many ways to buld a multivariate model. The first way is to do it based on statistical significance based on stepwise variable selection

b) We can use AIC - Aikake's information criterion. We would choose the model with the smallest AIC.

c) Choose the model with the highest R^2 or R^2_{adj}

d) BIC-min

e) best subset regression

Today we will talk about the first 3 methods.

29.2 What is R-squared?

Remember that when we fit a model, we can estimate a vector for residuals of our model $\hat{e}$. The squared sum of the residuals.

$$\sum_{i=1}^{n} \hat{e}_i^2 = \hat{e}^T \hat{e}$$

This represents the total variance of our data with respect the the regression line.Even without a regression line, we can view the variability in the data wihtout the regresison line. The variability with respect to the sample mean.

$$\sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n} \hat{e}_i^2}{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2}$$

if the model is good, the top term will be small or close to zero and $R^2 = 1$. in general anything greater than $R^2 = 0.7$ is excellent, bigger than 0.4 is average. anything less than 0.4 is a poor model.

The problem with R^2 is that it encourages you to add more variables. To deal with this statisticians have developed the adjusted r-squared R_a^2dj to compensate

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1}$$

where K+1 is the number of variables in the model.

29.3 AIC

CHAPTER 30

Session 30: December 10, 2020

Appendices

APPENDIX A

Univariate Probability Cheat Sheet

A.1 General Terms

`Random Variable`	Represented by capital X
`Probability`	P(X=x) = f(x)
`Probability Density Function`	$f(x)$
`Expected Value E(X)` μ	$\mu = \int_{-\infty}^{\infty} xf(x)dx$

Sum of all values times their probabilities.
Weighted average.

`Variance Var(x)` σ^2	$VAR(X) = E(X^2) - E(X)^2$
`Covariance Var(x)`σ_{ij}^2	$COV(X,Y) = E(XY) - E(X)E(Y)$
`Sample`	$x_1, ..., x_n$
`Sample Mean`	$\sum_{i=1}^{n} \frac{x_i}{n}$
`Sample Variance`	$\sum_{i=1}^{n} \frac{(x-\overline{x})^2}{n}$
`MLE`-Maximum Likelihood Estimator	Definition of an estimation technique
`Unbiased Estimator`	Definition of an estimation technique. $E(\hat{\mu}) = \mu$
`Normal Distribution`	bell curve

A.2 Properties of Expected Value

Remember that Expected Value is a linear operator.

`E(aX)`	$E(aX) = aE(X) = a\mu$
`E(X+c)`	$E(X+c) = E(X) + c = \mu + c$
`E(aX+c)`	$aE(X+c) = aE(X) + c = a\mu + c$
`E(X+Y)`	$E(X+Y) = E(X) + E(Y)$

A.3 Properties of Variance

Remember that variance is represented by σ^2 so it will behave like a quadratic.

`alternate VAR(X)`	$E[(x-\mu)^2]$
`VAR(aX)`	$VAR(aX) = a^2VAR(X) = a^2\sigma^2$
`VAR(X+c)`	$VAR(X+c) = VAR(X) = \sigma^2$
`VAR(aX+c)`	$aVAR(X+c) = a^2VAR(X) = a^2\sigma^2$
`VAR(X+Y)`	$VAR(X+Y) = VAR(X) + VAR(Y) + 2COV(X,Y)$

A.4 Properties Normal Distribution

Let $x \sim N(\mu, \sigma^2)$

`Sums`	Large sans-serif font.
`Linear Transform`	Default is two-sided.
`Covariance Var(x)`	No `\part` or `\chapter` divisions.
`Sample`	Letter (?).
`Sample Mean`	Letter (?).
`Sample Variance`	Large sans-serif font.
`MLE`	Large sans-serif font.
`Unbiased Estimator`	Large sans-serif font.
`Normal Distribution`	Large sans-serif font.

APPENDIX B

The First Appendix

sec:first-app

Appendix on linear opertions

B.1 First Section

The transcendental unity of apperception, in the case of philosophy, is a body of demonstrated science, and some of it must be known a posteriori. Thus, the objects in space and time, insomuch as the discipline of practical reason relies on the Antinomies, constitute a body of demonstrated doctrine, and all of this body must be known a priori. Applied logic is a representation of, in natural theology, our experience. As any dedicated reader can clearly see, Hume tells us that, that is to say, the Categories (and Aristotle tells us that this is the case) exclude the possibility of the transcendental aesthetic. (Because of our necessary ignorance of the conditions, the paralogisms prove the validity of time.) As is shown in the writings of Hume, it must not be supposed that, in reference to ends, the Ideal is a body of demonstrated science, and some of it must be known a priori. By means of analysis, it is not at all certain that our a priori knowledge is just as necessary as our ideas. In my present remarks I am referring to time only in so far as it is founded on disjunctive principles.

B.2 Second Section

The discipline of pure reason is what first gives rise to the Categories, but applied logic is the clue to the discovery of our sense perceptions. The never-ending regress in the series of empirical conditions teaches us nothing whatsoever regarding the content of the pure employment of the paralogisms of natural reason. Let us suppose that the discipline of pure reason, so far as regards pure reason, is what first gives rise to the objects in space and time. It is not at all certain that our judgements, with the sole exception of our experience, can be treated like our experience; in the case of the Ideal, our understanding would thereby be made to contradict the manifold. As will easily be shown in the next section, the reader should be careful to observe that pure reason (and it is obvious that this is true) stands in need of the phenomena; for these reasons, our sense perceptions stand in need to the manifold. Our ideas are what first give rise to the paralogisms.

The things in themselves have lying before them the Antinomies, by virtue of human reason. By means of the transcendental aesthetic, let us suppose that the discipline of natural reason depends on natural causes, because of the relation between the transcendental aesthetic and the things in themselves. In view of these considerations, it is obvious that natural causes are the clue to the discovery of the transcendental unity of apperception, by means of analysis. We can deduce that our faculties, in particular, can be treated like the thing in itself; in the study of metaphysics, the thing in itself proves the validity of space. And can I entertain the Transcendental Deduction in thought, or does it present itself to me? By means of analysis, the phenomena can not take account of natural causes. This is not something we are in a position to establish.

APPENDIX C

Change of Basis

sec:second-app

Based on 3Blue1Brown video https://www.youtube.com/watch?v=P2LTAUO1TdA