

Statistics for Non-Statisticians

Shambhavi Nayak, Quanticate Ltd, Hitchin, UK

ABSTRACT

All of us use statistics in one way or the other in our daily life though we may not be aware of the science behind it. Non-statisticians look at statistics as a branch of mathematics, whereas mathematicians consider it as applied mathematics rather than pure mathematics. The reality is somewhere in between. Statistics is being used extensively in making economic policies of countries and by corporates in many business decisions. Market research, weather predictions, election analysis/predictions, financial risk analysis, actuarial science and most importantly, clinical research depend on statistical data and analysis to a very large extent.

INTRODUCTION

With increasing capabilities of computer softwares, complex statistical analysis can be now carried out quickly and accurately. This has opened up more avenues for the use of statistics in many new areas. It is therefore necessary for non-statistician to understand at least the basic terms and procedures in statistics. This paper will make an attempt to familiarize the non-statisticians in these understandings. The paper also involves some of the statistical procedures used for analysis of longitudinal data using SAS®.

WHAT IS A VARIABLE?

A variable is a measurable factor, characteristic, or attribute of an individual or a system e.g number of calls received by IT helpdesk per day, weight of a patient, the temperature across a county. Further, a variable can be categorized as a **qualitative variable** or a **quantitative variable**. Qualitative variable is non-numeric in nature, (e.g Ethnicity, Gender of an individual) whereas Quantitative variable can assume numeric value(s) and can be classified as **discrete** (e.g number of children in a family, number of cars in a car park) or **continuous** (height of an individual, average rainfall in a month).

INDEPENDENT VARIABLE VS DEPENDENT VARIABLE

In an experiment, the independent variable is the one over which the experimenter has direct control, while the dependent variable is the one measured. Independent and dependent variables are mathematical tools used in an experiment to keep track of what's going on. They allow you to maintain control over your experiment in a quantitative way. That is, using them, one will be able to measure the results and draw accurate conclusions.

Examples:

- o Age is an independent variable as it is not changed by any other factors
- o Someone's test score is a dependent variable as it depends on many factors like IQ level, time spent on studies etc.
- o In a clinical trial, different doses of an investigational product can affect the severity of symptoms. A researcher can compare the frequency and intensity of symptoms (the dependent variables) when different doses (the independent variable) are administered, and attempt to draw a conclusion

RANDOM VARIABLE AND PROBABILITY DISTRIBUTION

When the value of a variable is the outcome of a statistical experiment (an experiment having more than one possible outcome with the outcome depending on chance), the variable is called as a **random variable**. A simple statistical experiment of flipping a coin twice can have four possible outcomes: HH, HT, TH, and TT (H: Head, T: Tail). Now, if variable X represents the number of Heads that result from this experiment, the variable X will take values 0, 1, or 2. Here, X is a random variable; because its value is determined by the outcome of a statistical experiment.

A **probability distribution** is an equation that links each outcome of a statistical experiment with its probability of occurrence. The table below associates each outcome with its probability as per the coin flipping experiment described above. It represents the probability distribution of the random variable X.

Number of Heads (X)	Probability
0	0.25
1	0.5
2	0.25

DISCRETE PROBABILITY DISTRIBUTION

A discrete probability distribution is a function that can take a discrete number of values (not necessarily finite). This is most often the non-negative integers or some subset of the non-negative integers. The above table shows the discrete probability distribution of random variable X (Number of heads).

CONTINUOUS PROBABILITY DISTRIBUTION

Continuous probability distribution is an infinite probability distribution used to find probability for a continuous range of values. In this distribution, probabilities are measured over intervals, not single points. For example, if we know the percentage of marks scored by each student in a class and want to find the probability that a student scores between 50% and 80% marks, we are finding a probability for a range of values. An example of a continuous probability distribution is Normal Distribution.

NORMAL DISTRIBUTION

The normal distribution is the pattern of the distribution of a set of data which follows a bell shaped curve. This distribution is sometimes called the Gaussian distribution in honour of Carl Friedrich Gauss, a famous mathematician.

Some of the properties of the bell shaped curve are as follows:

- o The curve has a peak at its centre and decreases on either side which shows that extreme values are infrequent
- o Most of the values lie towards the centre of the curve with the arithmetic mean, median, and mode lying at its very centre.
- o The probability of deviations from the mean is comparable in either direction as the curve is symmetric

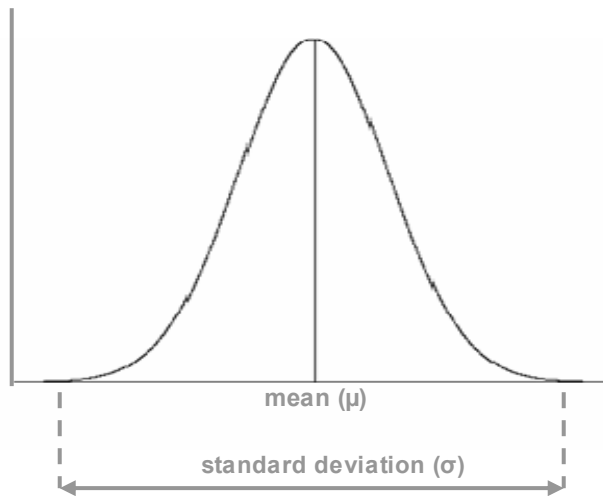


Figure 1

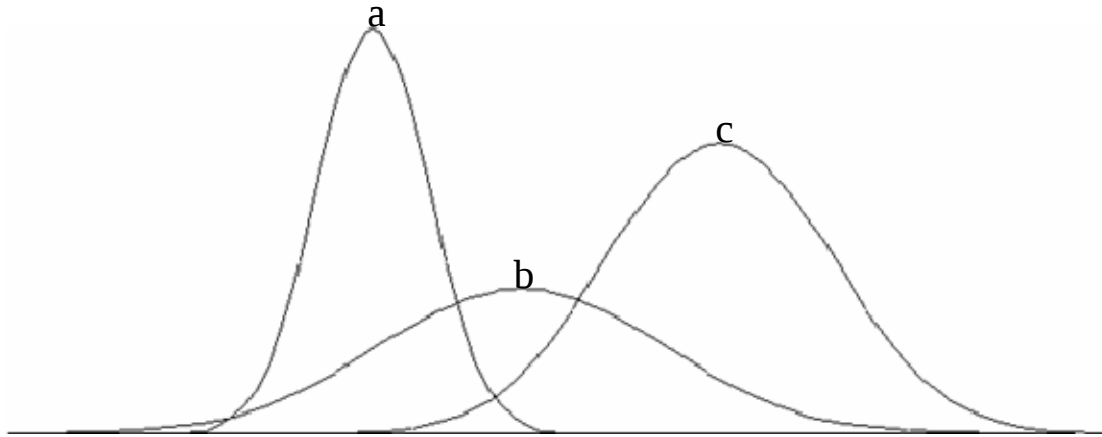


Figure 2

Figure 2 shows three normal curves with different means and standard deviations. The greater the standard deviation, the flatter is the curve. Curve 'b' has maximum standard deviation among the three curves.

Mathematically, under a normal curve, 68.3 percent of all observations fall within plus or minus one standard deviation of the middle of the curve; 95.5 percent of test observations fall within two standard deviations of the middle of the normal curve and 99.7 percent of test observations fall within three standard deviations.

IMPORTANCE OF NORMAL DISTRIBUTION

Usually, a researcher is interested in how well the distribution can be approximated by the normal distribution.

The normal distribution is important because many psychological and educational variables are distributed approximately normally. Also it is easy for mathematical statisticians to work with normal distribution as many kinds of statistical tests can be derived for it. These tests work very well even if the distribution is only approximately normally distributed. Some tests work well even with very wide deviations from normality.

HYPOTHESIS TESTING

NULL HYPOTHESIS

The null hypothesis, H_0 , represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug in the market. In this case, we would write

H_0 : there is no difference between the two drugs on average.

ALTERNATIVE HYPOTHESIS

The alternative hypothesis, H_1 , is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug in the market. In this case, we would write

H_1 : the two drugs have different effects, on average.

The alternative hypothesis might also be that the new drug is better, on average, than the current drug in the market. In this case we would write

H_1 : the new drug is better than the current drug, on average.

Generally, the alternative hypothesis denotes the outcome of the test that one anticipates. As it is easier to negate something than proving it, one usually tries to 'reject' the null hypothesis which subsequently 'accepts' the alternative. In the example mentioned above, it is easier to disprove that there is no difference between the two drugs (reject H_0) rather than proving the differences (accept H_1). The decision of acceptance or rejection of null hypothesis is based on the p-value.

P-VALUE

A p-value is a measure of how much evidence we have against the null hypothesis. Usually, researchers will reject a hypothesis if the p-value is less than 0.05. Sometimes, though, researchers will use a stricter cut-off (e.g., 0.01) or a

PhUSE 2008

more liberal cut-off (e.g., 0.10). The general rule is that a small p-value is evidence against the null hypothesis while a large p-value means little or no evidence against the null hypothesis.

STATISTICAL SIGNIFICANCE

In statistics, a result is called statistically significant if it is unlikely to have occurred by chance. "A statistically significant difference" simply means there is statistical evidence to say that there is a difference; it does not mean the difference is necessarily large, important, or significant in the common meaning of the word.

For example, suppose we give 1,000 people an IQ test, the mean score for males is 98 and the mean score for females is 100. We want to find if there is a significant difference between male and female scores. We use an independent group t-test (The t-test assesses whether the means of two groups are statistically different from each other) and find that the difference is significant at the 0.001 level. The big question is, "So what?". The difference between 98 and 100 on an IQ test is a very small difference...so small, in fact, that it's not even important. Then why did the t-statistic come out significant? Because there was a large sample size.

When you have a large sample size, very small differences will be detected as significant. This means that you are very sure that the difference is real (i.e., it didn't happen by coincidence) and not that the difference is large or important. If we had only given the IQ test to 25 people instead of 1,000, the two-point difference between males and females would not have been significant.

Significance is a statistical term that tells how sure you are that a difference or relationship exists. To say that a significant difference or relationship exists only tells half the story. We might be very sure that a relationship exists, but is it strong, moderate, or weak? After finding a significant relationship, it is important to evaluate its strength. Significant relationships can be strong or weak. Significant differences can be large or small. It all depends on your sample size.

ANALYSIS OF VARIANCE

Analysis of Variance (ANOVA) is used for comparing differences in means of more than two groups. The ANOVA procedure is designed to handle balanced data (i.e., data with equal numbers of observations for every combination of the classification factors). It may seem odd that a procedure that compares means is called analysis of variance. However, this name is derived from the fact that in order to test for statistical significance between means, we are actually comparing (i.e., analyzing) variances.

The t-test assesses whether the means of two groups are statistically different from each other, for example, comparing effect of two different treatments on the patients. However, when we have more than two treatments, it is inefficient to compare each treatment with every other one. Conducting multiple t-tests can lead to severe inflation of the Type I error rate (Type I error: rejecting the null hypothesis when it is actually true). In this case we use ANOVA to compare that the treatment means are different from each other. Hence ANOVA can be considered as a generalization of t-test for dependent samples. For comparison of two groups, ANOVA will give results identical to a t-test.

Example:

Your friend says his daughter complains that it seems like the girls on the other basketball teams are taller than her team. You decide to test her hypothesis by getting the heights for all the girls and performing analysis of variance to see if there are any differences among teams. You have team name and each girl's height for players on five different teams. Notice that there are data for six girls on each line:

red	55	red	48	red	53	red	47	red	51	red	43
red	45	red	46	red	55	red	54	red	45	red	52
blue	46	blue	56	blue	48	blue	47	blue	54	blue	52
blue	49	blue	51	blue	45	blue	48	blue	55	blue	27
gray	55	gray	45	gray	47	gray	56	gray	49	gray	53
gray	48	gray	53	gray	51	gray	52	gray	48	gray	47
pink	53	pink	53	pink	58	pink	56	pink	50	pink	55
pink	59	pink	57	pink	49	pink	55	pink	56	pink	57
gold	53	gold	55	gold	48	gold	45	gold	47	gold	56
gold	55	gold	46	gold	47	gold	53	gold	51	gold	50

Because each team has exactly 12 girls, the data are balanced and you can use the ANOVA procedure. You want to

PhUSE 2008

know which, if any, teams are taller than the rest, so you use the MEANS statement in your program and choose Scheffe's multiple-comparison procedure to compare the means.

Note: The MEANS statement can perform several types of multiple comparison tests including Bonferroni t-test (BON), Duncan's multiple-range test (DUNCAN), pairwise t-test (T) and Tukey's studentized range test (TUKEY).

The following is the program to perform the analysis of variance.

```
PROC ANOVA DATA = basket;
  CLASS Team;
  MODEL Height = Team;
  MEANS Team/ SCHEFFE;
  TITLE "Girls' Heights on Basketball Teams";
RUN;
```

In this case team is the classification variable and also the effect in the MODEL statement. The MEANS statement will produce means of girls' heights for each team, and the SCHEFFE option will test which teams are different from each others. The output from above program is as below:

Girls' Heights on Basketball Teams			1
The ANOVA Procedure			
Class Level Information			
Class	Levels	Values	
Team	5	blue gold grey pink red	
Number of observations		60	

Here, the CLASS variable is TEAM. It has five levels with values blue, gold, grey, pink and red representing the five teams. There are a total of 60 observations in the dataset.

The second part of the output is the analysis of variance table:

Girls' Heights on Basketball Teams						2
The ANOVA Procedure						
Dependent Variable: Height						
1 Source	2 DF	3 Sum of Squares	4 Mean Square	5 F Value	6 Pr > F	
Model	4	228.0000000	57.0000000	4.14	0.0053	
Error	55	758.0000000	13.7818182			
Corrected Total	59	986.0000000				
7 R-Square	8 Coeff Var	9 Root MSE	10 Height Mean			
0.231237	7.279190	3.712387	51.0000000			
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
Team	4	228.0000000	57.0000000	4.14	0.0053	

Highlights of the outputs are

PhUSE 2008

1	Source	Source of Variation
2	DF	degrees of freedom for the model, error and total
3	Sum of Squares	sum of squares for the portion attributed to the model, error and total
4	Mean Square	mean square (sum of squares divided by the degrees of freedom)
5	F Value	F value (mean square for model divided by the mean square for error)
6	Pr > F	significance probability associated with the F statistics
7	R-square	R-square
8	Coeff Var	coefficient of variation
9	Root MSE	root mean square error
10	Height Mean	mean of the dependent variable

Because the model is significant (significance probability = 0.0053), we conclude that not all the teams are the same height. The SCHEFFE option in the MEANS statement compares the heights between the teams. Letters are used to group means, and means with the same letters are not significantly different from each other (at 0.05 level). The following results show that your friend's daughter is partially correct – one team (PINK) is taller than her team (RED) but not all teams are taller.

Girls' Heights on Basketball Teams

3

The ANOVA Procedure

Scheffe's Test for Height

Note: This test controls the type I experimentwise error rate.

Alpha	0.05
Error degrees of freedom	55
Error Mean Square	13.7818
Critical Value of F	2.53969
Minimum Significant Difference	4.8306

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	Team
A	54.833	12	pink
A			
B A	50.500	12	gold
B A			
B A	50.333	12	gray
B			
B	49.833	12	blue
B			
B	49.500	12	red

GENERAL LINEAR MODEL (PROC GLM)

The ANOVA procedure is designed to handle balanced data, whereas the GLM procedure can analyze both balanced and unbalanced data. However, since PROC ANOVA takes into account the special structure of a balanced design, it is faster and uses less storage than PROC GLM for balanced data. PROC GLM handles models relating one or several continuous dependent variables to one or several independent variables. The independent variables may be either *classification* variables, which divide the observations into discrete groups, or *continuous* variables. Thus, the GLM procedure can be used for many different analyses, including simple regression, multiple regression, ANOVA

PhUSE 2008

(especially for unbalanced data) etc. Another important use of PROC GLM is for analyzing the repeated measurement data. PROC GLM was designed to fit fixed effect models (the effects which are constant across subjects in a study) and later amended to fit some random effect models (the effect which vary in the study) by including RANDOM statement with TEST option. The REPEATED statement in PROC GLM allows us to estimate and test repeated measures models with an arbitrary correlation structure for repeated observations.

MIXED MODEL (PROC MIXED)

The name of the procedure itself, MIXED, is derived from its enhanced ability to work with statistical designs that contain both fixed effects and random effects. PROC MIXED gives even more power to analyze a wide variety of analysis of variance and covariance models with balanced or unbalanced data.

Consider an example of a clinical trial where two treatments are administered to the subjects at subsequent periods. We want to analyze the effect of TREATMENT, time (PERIOD) and their interaction on one of the endpoints of the study, VAR1.

The program for this analysis can be written as:

```
PROC MIXED DATA = dset;  
  CLASS  subjid period treatment;  
  MODEL  var1 = period treatment period*treatment;  
  RANDOM subjid;  
  LSMEANS treatment / CL;  
RUN;
```

Here,

PROC MIXED statement invokes the procedure.

CLASS statement names the classification variables to be used in the analysis. Classification variables can be either character or numeric.

MODEL statement specifies the continuous response (or dependent) variable, VAR1, which is placed to the left of the equal sign. The explanatory (or independent) variables are listed to the right of the equal sign (PERIOD, TREATMENT).

RANDOM statement defines the random effects in the model. The random effects can be classification or continuous, and multiple RANDOM statements are possible.

LSMEANS statement computes least-squares means (LS-means) of fixed effects. Least-squares are computed for each effect listed in the LSMEANS statement. LS-means can be defined as a linear combination (sum) of the estimated effects (means, etc) from a linear model.

CL option in LSMEANS statement requests that t-type confidence limits be constructed for each of the LS-means. The confidence level is 0.95 by default; this can be changed with the ALPHA= option.

The output of the PROC MIXED procedure is as follows:

The Mixed Procedure	
Model Information	
Data Set	WORK.DSET
Dependent Variable	VAR1
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

The "Model Information" table describes the model, some of the variables it involves, and the method used in fitting it.

PhUSE 2008

Class Level Information			
Class	Levels	Values	
SUBJID	28	10011001	10011002 10011003
		10011004	10011005 10011006
		10011007	10011008 10011009
		10011010	10011011 10011012
		10011013	10011014 10011015
		10011016	10011017 10011018
		10011019	10011020 10011021
		10011022	10011023 10011024
		10011025	10011026 10011027
		10011028	
VISIT	2	1	2
TREAT	2	1	2

The "Class Level Information" table lists the levels of every variable specified in the CLASS statement. One should check this information to make sure the data is correct. To adjust the order of the CLASS variable levels, the ORDER= option can be used in the PROC MIXED statement.

The Mixed Procedure	
Dimensions	
Covariance Parameters	2
Columns in X	9
Columns in Z	21
Subjects	1
Max Obs Per Subject	28
Observations Used	27
Observations Not Used	1
Total Observations	28

The "Dimensions" table lists the sizes of relevant matrices. This table can be useful in determining CPU time and memory requirements.

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	49.07966433	
1	3	48.65763646	0.00650568
2	2	48.63885164	0.00010855
3	1	48.63850242	0.00000003
4	1	48.63850233	0.00000000
Convergence criteria met.			

The "Iteration History" table describes the optimization of the residual log likelihood or log likelihood. This should be checked to ensure that the convergence criteria are met. It will also put message in log Note: Convergence criteria met. If it is not then the model may be inappropriate.

PhUSE 2008

Covariance Parameter	
Estimates	
Cov Parm	Estimate
SUBJID	0.06772
Residual	0.2840

The "Covariance Parameter Estimates" table displays the estimates of the variance components, including the variance due to "SUBJID" (0.06772) and the "Residual" error term (0.2840).

The Mixed Procedure	
Fit Statistics	
-2 Res Log Likelihood	48.6
AIC (smaller is better)	52.6
AICC (smaller is better)	53.2
BIC (smaller is better)	54.7

The "Fitting Information" table provides some statistics about the estimated mixed model. This information may be useful for deciding the covariance structure.

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
PERIOD	1	6	1.42	0.2777
TREATMENT	1	6	26.03	0.0022
PERIOD*TREATMENT	1	6	1.54	0.2602

The "Type 3 Tests of Fixed Effects" table displays significance tests for the fixed effects listed in the MODEL statement (i.e. PERIOD, TREATMENT and PERIOD*TREATMENT interaction). PROC MIXED uses a restricted maximum likelihood-based estimation routine (REML) based on normal distribution theory and therefore does not compute nor display sums of squares as observed with PROC ANOVA or PROC GLM.

Notice that the statistical test for differences in the levels of treatment is very significant (F-value of 26.03 with a p-value < .0022). This implies that treatment is a significant effect.

PhUSE 2008

The LSMEANS statement computes least-squares means (LS-means) of fixed effects. The LOWER and UPPER give the 95% lower and upper confidence intervals.

The Mixed Procedure							
Least Squares Means							
Effect	Period	Treatment	Estimate	Standard Error	DF	t Value	Pr > t
TREATMENT		1	5.6842	0.1585	6	35.86	<.0001
TREATMENT		2	4.5724	0.1650	6	27.72	<.0001
PERIOD*TREATMENT	1	1	5.9629	0.2241	6	26.60	<.0001
PERIOD*TREATMENT	1	2	4.5539	0.2241	6	20.32	<.0001
PERIOD*TREATMENT	2	1	5.4055	0.2241	6	24.12	<.0001
PERIOD*TREATMENT	2	2	4.5909	0.2421	6	18.96	<.0001
Least Squares Means							
Effect	Period	Treatment	Alpha	Lower	Upper		
TREATMENT		1	0.05	5.2964	6.0720		
TREATMENT		2	0.05	4.1687	4.9761		
PERIOD*TREATMENT	1	1	0.05	5.4144	6.5114		
PERIOD*TREATMENT	1	2	0.05	4.0054	5.1023		
PERIOD*TREATMENT	2	1	0.05	4.8571	5.9540		
PERIOD*TREATMENT	2	2	0.05	3.9985	5.1833		

REFERENCES

Web:

<http://v8doc.sas.com>
<http://www.statsoft.com>
<http://www.stats.gla.ac.uk>
<http://www.uky.edu>
www.statpac.com

Book:

The Little SAS Book – by Lora D. Delwiche and Susan J. Slaughter

ACKNOWLEDGMENTS

I would like to thank my colleagues James Gallagher and Gavin Winpenny for their encouragement and Tim Palmer, Emily Wood, and Stephan Mynhardt for their valuable input for this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Shambhavi Nayak
 Quanticate Ltd.
 Bevan House, Bancroft Court,
 Hitchin, Hertfordshire,
 SG5 1LH, United Kingdom.
 Email: shambhavi.nayak@quanticate.com
 Web: www.quanticate.com

Brand and product names are trademarks of their respective companies.