Statistical Inference

Statistical Inference is the branch of statistics concerned with using sample data to make inferences about the population. Populations are characterised by numerical descriptive measures, referred to as *parameters*.

Note: A *parameter* is a value, usually unknown (and which therefore has to be estimated), that is used to represent a certain population characteristic.

1: Hypothesis Testing

The mechanism for hypothesis testing follows a series of general steps, in this course we will:

1. **Determine the appropriate statistical test;**
2. **State the null and alternative hypotheses of interest;**
3. **Check the assumptions regarding the variables of interest are satisfied;**
4. **Perform the analysis using SAS;**
5. **Document conclusions on the basis of the analysis performed.**

The conclusions that are derived in step 5 are dictated by the significance level (α). This is decided upon prior to the analysis and is the level at which the analyst decides to reject the null hypothesis. The outcomes of hypothesis testing can be expressed as follows:

Hypothesis	**True outcome**	
	H_0	**H_1**
H_0 (Null hypothesis)	Correct decision $(1 - \alpha)$	False negative decision (Type II Error (β))
H_1 (Alternative hypothesis)	False positive decision (Type I error α (or p-value)	Correct decision (1 – β)

A general rule of thumb that is applied by statisticians is to reject the null hypothesis if the p-value (α) that is obtained in the output is less than 0.05, and is referred to as ***rejecting the null hypothesis at 5% level of significance***.

The table below provides a summary of the relevant statistical tests for alternative situations where hypothesis testing is required.

Hypothesis	**Normally distributed**	**Non-normal/rank/scores**
Compare a variable to a hypothetical value	One-sample t-test	Wilcoxon-test
Compare two independent variables	Two-sample t-test	Wilcoxon Mann-Whitney test
Compare two variables with same response for the same subjects	Paired t-test	Wilcoxon-test (Signed rank test)
Compare three or more variables	One-way ANOVA	Kruskal-Wallis test
Relationship between two variables	Pearson correlation coefficient	Spearman correlation coefficient
Linear relationship between two variables	Simple linear regression	Non-parametric linear regression (outside of scope for this course)
Linear relationship between a dependent variable and two or more independent variables	Multiple linear regression	

1.1 Comparing a variable to a hypothetical value

This form of hypothesis testing is directed at answering the question, **“Is the population mean $\boldsymbol{\mu = c}$?”,** with c being a pre-specified constant and μ being the population mean for the variable of interest.

When the sample size is small (less than 30 observations) and the distribution of the population from which the sample is selected is assumed to be Normal, a **one-sample t-test** would be appropriate. The null hypothesis for this would be:

$$H_o: \mu = c$$

$$H_1: \mu \neq c$$

However if this is not the case, then the variable must be assumed to be distribution free and a non-parametric test is considered appropriate, with the alternative in this case being the **Wilcoxon one-sample signed rank test** (another alternative is the one-sample sign test).

In this case the null hypothesis is based on the population median, *m*:

$$H_o: m = c$$

$$H_1: m \neq c$$

1.2 Comparing two independent variables

The question that one requires to answer in this scenario is **“Do the population means for the samples say $\boldsymbol{\mu_1}$and $\boldsymbol{\mu_2}$ differ significantly”**.

When the two samples are drawn independently, from Normal distributions and have equal variances ($\sigma_1^2 = \sigma_2^2$), we would use a **two-sample t-test** (if the variances are unequal an alternative t-test exists where a pooled estimate is used, assuming the other two assumptions are satisfied). The null hypothesis in such a situation is:

$$H_o: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

A test that requires less stringent assumptions is the Wilcoxon Rank Sum test (also known as Mann-Whitney U), with the sole assumption being that we have independent random samples from two populations.

The null hypothesis in this situation is based upon the distributions (p.d.f's) of the two samples, say *f(x₁)* and *f (x₂)*, and can be expressed as:

$$H_o: f(x_1) = f(x_2)$$

$$H_1: f(x_1) \neq f(x_2)$$

1.3 Comparing two variables with the same response for the same subjects

The previous method is only valid where we have independent samples from two populations. However, in some experimental situations we have each measurement in one sample being ***matched*** **or** ***paired*** with a particular measurement in another sample.

Let $D_0 = \mu_1 - \mu_2$represent the mean difference between the matching data points in each sample; the null hypothesis can then be expressed as:

$$H_o: D_0 = 0$$

$$H_1: D_0 \neq 0$$

The non-parametric alternative in this case is again the Wilcoxon Signed Rank test.

H_o: the difference between the pairs follows a symmetric distribution around zero

H_1: the difference between the pairs does not follow a symmetric distribution around zero

2: CORRELATION

The correlation coefficient (*r)* measures the strength of association between two variables.

A correlation coefficient near to 1 implies a strong positive relationship, near 0 implies the variables are independent and near –1 implies a negative relationship.

If we were asked to study the relationships between the scores obtained in the different IQ domains and the MRI count (found in the IQ data studied in a previous lecture), an appropriate method of analysis would be to construct a correlation matrix. In essence this involves recording all the correlations in a single matrix as opposed to individually.

Pearson's correlation coefficient is the most commonly used on continuous variables and requires that the variables follow a Normal distribution. However, if the data is skewed or if one of the variables is on an ordinal scale and the other is not on an ordinal scale, then a more appropriate measurement is Spearman's rho.

2.1 How to test if a variable is normally distributed in SAS?

In order to determine which correlation coefficient should be used within the analysis we first need to determine if the variables we wish to analyse are Normally distributed.

One way we can carry out a **test for normality** is as follows:

- Click on **Tasks** then **Distribution Analysis.**
- Select the variables for analysis.
- Under the **Options tab** select to **Add normal curve**.
- Under this section, you can select **Histogram and Goodness of fit tests.**

The hypotheses in this case are:

H_o: there is no difference between the distribution of the variable and that of the normal distribution;
H_1: there is a difference between the distribution of the variable and that of the normal distribution.

If we revisit the IQ Data studies previously. The following table is achieved for the results of the fitted Normal Distribution for FISQ.

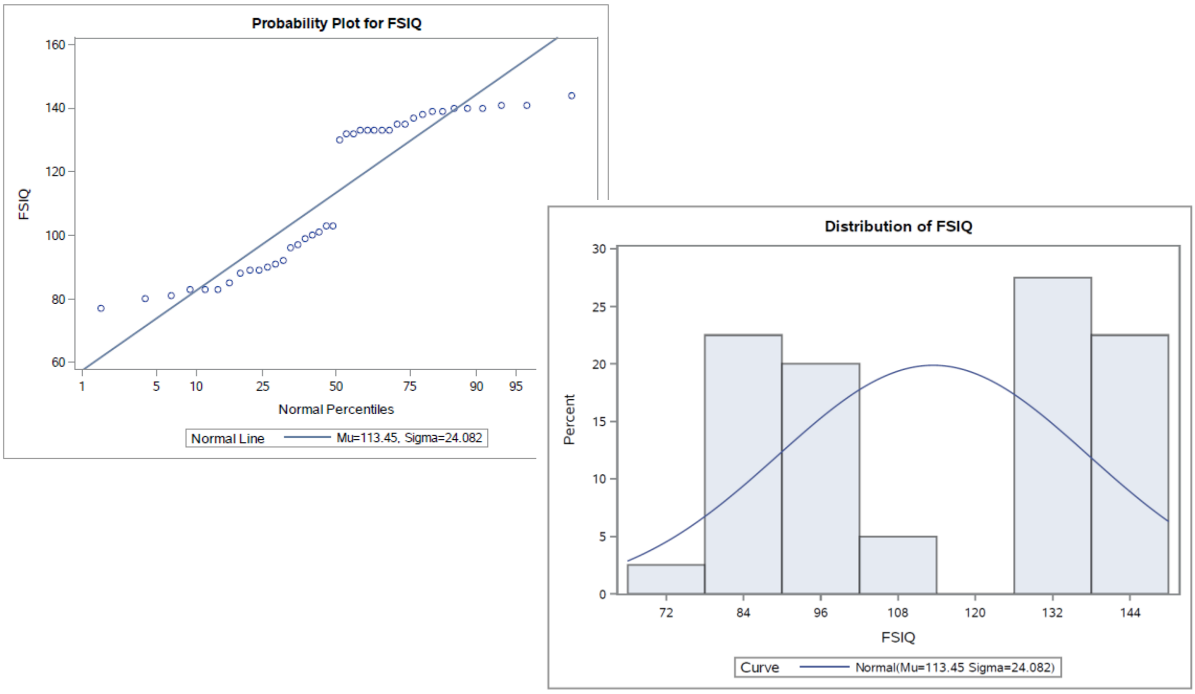
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	D	0.25443386	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.51127276	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.89955516	Pr > A-Sq	<0.005

When you fit a parametric distribution, **PROC UNIVARIATE** provides a series of goodness-of-fit tests based on the empirical distribution function (EDF), these are the Kolmogorov-Smirnov statistic, the Anderson-Darling statistic, and the Cramér-von Mises statistic.

The EDF tests offer advantages over a traditional chi-square goodness-of-fit test, including improved power and invariance with respect to the histogram midpoints.

By studying the output for all three tests we can conclude that all 3 p-values are significant at the 5% level, therefore the distribution of the FISQ variable is signficantly different to that of a Normal distribution.

This is also supported by the probability plot and the histogram which both differ from the plots that would be expected if the data followed the desired distribution.



Example

As our data is not normally distributed, we therefore need to use the Spearman's rho correlation coefficient to determine if a significant relationship (correlation) exists between the IQ variables present in the IQ data.

We perform a **correlation** as follows:

- Select **Tasks, Statistics** then **Correlation Analysis.**
- Select all variables for analysis.
- Under **Options, STATISTICS, Nonparametric Correlations**, select Spearman's rank**.**
- You can also select a scatter plot under **PLOTS** if you wish.

The hypotheses in this case are:

H_0: there is no linear relationship between the IQ variables;
H_1: there is a linear relationship between the IQ variables.
The following output table is achieved by generating a correlation matrix on FSIQ, VIQ, PIQ and MRI_Count using the Spearman Correlation Coefficient.

Spearman Correlation Coefficients, N = 40 Prob > |r| under H0: Rho=0

	FSIQ	**VIQ**	**PIQ**	**MRI_Count**
FSIQ	1.00000	0.91805	0.87869	0.47214
		<.0001	<.0001	0.0021
VIQ	0.91805	1.00000	0.71498	0.39981
	<.0001		<.0001	0.0106
PIQ	0.87869	0.71498	1.00000	0.41246
	<.0001	<.0001		0.0082
MRI_Count	0.47214	0.39981	0.41246	1.00000
	0.0021	0.0106	0.0082	

By studying the correlation matrix we find that all of the measurements are positively correlated and are significant at the 5% level. However, the strength of the association varies considerably.

Determining the strength of the association is subjective, and can vary from analyst to analyst, however for the purpose of this course, we shall apply the following rules:

- -1.0 to -0.6 strong negative association
- -0.6 to -0.3 weak negative association
- -0.3 to +0.3 little or no association
- +0.3 to +0.6 weak positive association
- +0.6 to +1.0 strong positive association.

The results obtained are expected in practise due to the responses of interest all measuring the same characteristic (intelligence), consequently we would expect the measurements to increase simultaneously.

Comment on the outcomes of these tests.

Example

Researchers at a concrete company are interested in determining whether an additive in the mixing process affects the strength of concrete. Two operators working for the company measured each batch of concrete for strength and the company is also interested in whether the measurements by the two operators are consistent.

The data for the study is stored in the **Concrete** data table, which has the following variables:

- **Strength1** – measured strength of concrete by operator 1 (primary operator on duty);
- **Strength2** – measured strength of concrete by operator 2 (the trainee on duty);
- **Brand** – Graystone, Consolidated, EZ Mix;
- **Additive** – presence of reinforcement material (reinforced, standard);
- **Humidity** – observed humidity.

The analysis focuses on two research questions and uses two types of **t-test**:

1. Does the type of additive have any effect on concrete strength for the primary operator?
    - **An independent t-test is used.**
2. Are there differences between operators in strength measures?
    - **A paired-samples t-test is used.**

Typically, the first order of business in data analysis is to test for the assumptions of the statistics you are using. Recall the assumptions of 2 sample t-tests: **normally distributed sample means and equal group variances**. These will be selected whilst carrying out the test.

To carry out a **paired sample t-test** we need to:

- Select **Tasks** then, **Statistics**, **t Tests.**
- From the t-Test drop down list select **Paired tests**.
- Select the Group 1 and Group 2 variables as **Strength 1 and Strength 2.**
- Under options observe and select the variety of test options.
- Under plots you also have the option to include a range of plots.

The hypotheses in this case are:

H_o: there is no difference between the strength measures of the two operators;
H_1: there is a difference between the strength measures of the two operators.

The output obtained from the paired-samples t-test is as follows.

The first table in the output report provides the outcome of the tests for normality.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.80558	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.268429	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.368979	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2.167378	**Pr > A-Sq**	<0.0050

Looking at the p-values for each of these tests, it is clear that the null hypothesis should be rejected and that there is evidence to suggest that this data is not normally distributed.

The next section shows descriptive statistics for the difference between Strength1 and Strength2 and 95% confidence interval bounds.

Mean	**95% CL Mean**		**Std Dev**	**95% CL Std Dev**	
-0.0533	-0.2039	0.0972	0.4032	0.3211	0.5420

The confidence interval for the mean includes 0; therefore, there is evidence to suggest that the measures collected by the two operators are not significantly different from one another.

A more direct result is given by the results of the Wilcoxon signed rank test. This analysis shows a p-value of 0.259. The p-value is greater than our alpha level of .05, so we conclude that the strength measures for the operators are **not significantly different**.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	-0.72449	**Pr > |t|**	0.4746
Sign	**M**	1	**Pr >= |M|**	0.8555
Signed Rank	**S**	-55.5	**Pr >= |S|**	0.2529

To carry out a **Two-sample t-test** we need to:

- Select **Tasks, Statsitcis,** then **t Tests.**
- From the t-Test drop down select **Two-Sample test.**
- Select the **analysis and grouping variables**.
- Under plots you should also include a box plot.

The hypotheses in this case are:

- H_o: there is no difference between the strength measure for the primary operator when comparing reinforced to standard concrete;
- H_1: there is a difference between the strength measure for the primary operator when comparing reinforced to standard concrete.

The output obtained from the two sample t-test was as follows.

After checking the assumptions for normality have been met. The assumption of homogeneity of variances needs to be examined. This assumption is tested automatically with the t-test analysis. Scroll to the bottom of the t-test report and look at the **Equality of Variances** report. This tests the null hypothesis that the group variances are equal against the alternative that they are different.

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	1.16	0.7900

If this test is significant, there is evidence that the group variances are different (in other words, the equality of variances assumption has not been met). In this case, the p-value is high (>0.05), so you can conclude that the assumption of equal group variances is reasonably appropriate.

In the t-test report, two tests are reported; the difference between them is in the degrees of freedom. The first (Pooled) should be interpreted when the assumption of equal group variances has been met. The second (Satterthwaite) should be interpreted when the assumption of equal group variances has not been met. Note: when variances are equal both methods will report very similar values.

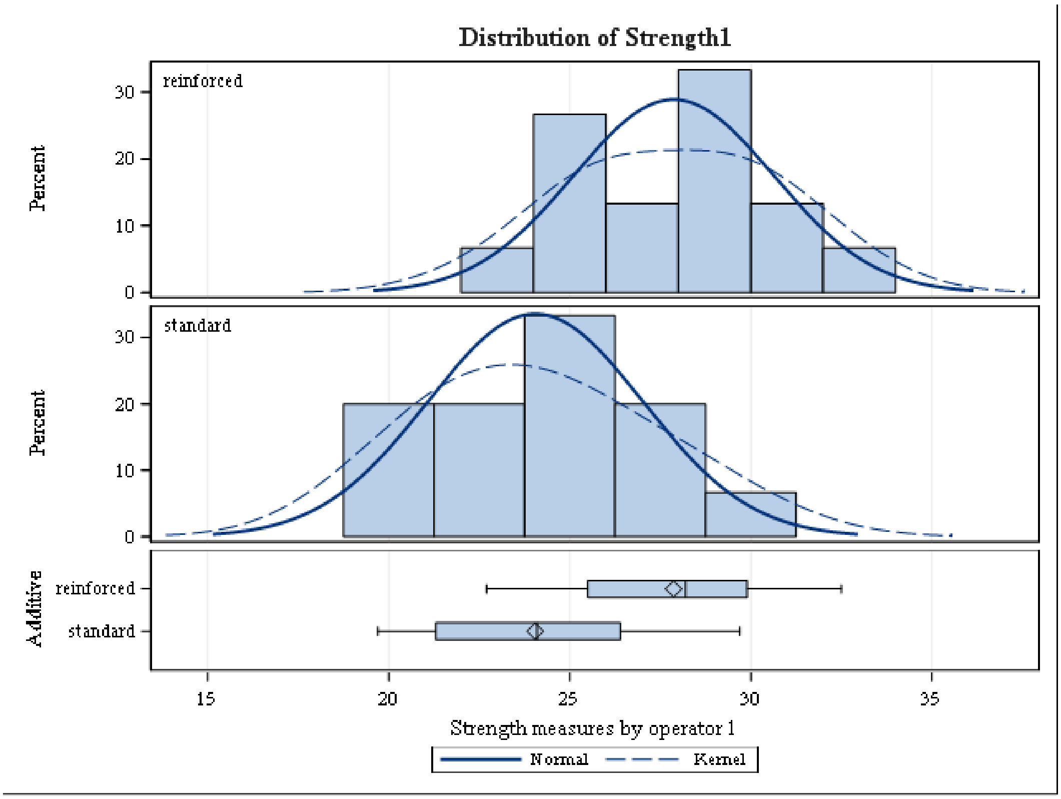
The first table in the output report shows descriptive statistics for the difference between the type of additive and 95% confidence interval bounds.

Additive	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
reinforced		27.8667	26.3369	29.3964	2.7624	2.0224	4.3566
standard		24.0467	22.4019	25.6914	2.9701	2.1745	4.6841
Diff (1-2)	Pooled	3.8200	1.6747	5.9653	2.8681	2.2761	3.8790
Diff (1-2)	Satterthwaite	3.8200	1.6742	5.9658			

The confidence interval for the difference does not include 0; therefore, you can conclude that the strength of the concrete by additive are significantly different from one another.

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	3.65	0.0011
Satterthwaite	Unequal	27.854	3.65	0.0011

The p-value associated with the t-test for equal variances (Pooled) is less than the α of 0.05, so again you can conclude that the strength measures for the two types of concrete are significantly different. The box and whisker plot supports these findings.



Comment on the outcomes of these two tests.

Tutorial Questions

Save the following data files from Blackboard – ***carbonmonoxide, eye_experiment, pain_experiment, sportshoes and concrete.***

Each of the files contains data that requires a hypothesis test to be performed, use your notes to determine the appropriate test, and then apply the test and interpret the output.

You must write a **short summary of your findings** for each question, including your reasoning for the test and its assumptions and the output achieved.

1. Sport Shoes

Twenty-five athletes were randomly selected from a regional running association. They were asked to try two alternative leading brands in trainers when running, and record the period in weeks until the trainers were regarded as being unusable (which was predefined to avoid bias). Is there a significant difference in the durability of the two trainer types?

2. Pain Experiment

After a study on an anti-depressant it was discovered by researchers that patients with lower back pain experienced a decrease in radicular pain after 6 to 8 weeks of daily treatment.

Following this discovery a new study was conducted to determine if this response was drug related or coincidental. Subjects were thus enrolled for a follow-up study and were randomly allocated to either the drug or placebo. Following the conclusion of the trial subjects were asked to rank their rating of pain in comparison to baseline. The positive values indicate less pain (the higher the better), and negative values indicate increased pain. Are there significant differences between the pain experienced by the subjects in the two groups?

3. Eye Experiment

Patients with vision problems enrolled for a study to compare the use of two varieties of eye droplets available for their condition. The subjects were instructed to add a drop of each drug to a different eye, four times a day for a period of four weeks. At the end of the trial the patients recorded marks between 1 and 4 for four different regions of the eye. Comparisons were then made between the total scores, with higher score denoting poorer visibility. Were there statistically significant differences between the varieties of eye droplets?

4. Carbon Monoxide

Periodic tests are undertaken to determine the level of carbon monoxide emissions of two of the major cigarette brands in a region of South East Asia. An investigator collects two random samples for each of these cigarettes and records the CO contents in milligrams for them. The local authorities wish to determine if there are significant differences between the brands, perform an appropriate statistical test to determine this.

5. Concrete

Using the concrete data discussed in lecture this week, determine if a relationship exists between the measured strength of concrete by operator 1 and operator 2.

6. Salary

A random sample of ten schools in the UK were taken; the average wage of the A-Level teachers were ranked, and the average number of points obtained by the students in the previous year's GCSE were recorded:

Salary	3	2	1	4	10	6	8	7	9	5
Points	81	75	73	71	69	57	48	41	31	29

Determine if the collective outcome of a pupils GCSE exams has an association with the salary of the teacher. You will need to create a new data set containing this information.