Chapter 2 Notes

Descriptive Analysis and Inference

Samples from populations can be used to obtain information about the underlying populations.

Descriptive statistics tell us about features of the data, which allow us to infer characteristics of the population the sample came from. Mean and median tell us about location. Standard deviation tells us about spread and skewness tells us about symmetry. Correlations and covariance tell us about the relation between components of multivariate data or between multiple univariate data sets.

Data visualizations can allow us to see features in the data. Scatter plots let us see the data as is. Histograms and box and whisker plots show the general shape of a distribution. Probability plots compare percentage points of a distribution and a data set (an empirical distribution).

Descriptive analysis can be useful as a preliminary exploratory step, a check of assumptions, the foundations for test statistics, and as the motivation for more advanced modeling.

proc univariate

The **proc univariate** procedure is a fairly general procedure for univariate descriptive analysis. By default, it will give us basic descriptive statistics, quantiles, extreme observations, and tests of location with the null hypothesis being that the true population location (the mean or median) is 0.

Additional statements can be added to get descriptive plots such as histograms and probability plots, and options can be added to do things including test distributional assumptions and change the null hypothesis for the test of location.

Note that **proc univariate** does many things and can generate lots of results, so we will usually want to use **ods select** or **ods exclude** statements to pick out the pieces we want or need for a particular analysis. Additionally, there are specific procedures for many of the results **proc univariate** generates, and it is generally true (in SAS and in other software) that functions or procedures for specific tasks will be better (e.g. easier to use, have fewer results that aren't of interest, and allow more flexibility for the particular analysis we are interested in). For instance, if we know we want to do a t-test, we can use **proc ttest** and if we know we just want a histogram we could use **proc sgplot** with a **histogram** statement.

proc corr

For correlations between variables we can use **proc corr** which will compute the correlations from the data and give p-values for the test for significant correlation (the null hypothesis is that the true correlation is 0, and the alternative hypothesis is that the true correlation is not 0). There are a number of correlations included. We will focus on the Pearson correlation (the correlation you first see in text books) and the Spearman correlation (a rank-based analogue which is more robust when extreme values are present).

An Example

Mortality and Water Hardness in the UK

The data is described in chapter 2 of *A Handbook of Statistical Analysis Using SAS, Third Edition*. Its contents can be summarized as follows:

- As the textbook describes, the data set has 61 data points.
- Each data point is for a town in England and contains the mortality rate per 100,000 males (averaged over 1958-1964) and the calcium concentration (higher concentration = harder water) in parts per million in the town's drinking water
- Towns classified as northern are marked with an asterisk (*)

Code for reading and processing the data is included in the **chapter2.sas** program file from the text and in the **Chapter2InitialCode.sas** file in the course space.

We have mortality data and calcium concentration data by region in the UK. What questions might we try to answer?

Two that the author asks are:

- How are mortality and water hardness related?
- Is there a geographic factor in the relationship?

Before we attempt to answer these questions, we would first want to look at descriptive measures and visualizations for the data. Before considering the geographic impact (e.g. differences between north and south) we can look at univariate measures and plots for the individual **mortality** and **hardness** variables, look at a scatter plot of these variables against each other, and look at the correlation between the variables.

Univariate Analysis

Univariate analysis can tell us about the mortality rates and calcium concentration, individually. We can get descriptive statistics; qualitatively and quantitatively test for normality (an important assumption for t-tests); see the distribution of the values in general, and perform tests of location (mean or median). The **Student's t** test results will assume an underlying normal population. The **Sign** test results are about the median—the assumption is that it is equally likely to observe a value above or below the null hypothesis value. The **Signed Rank** test is also about the median—there is an additional assumption of symmetry though (so ranks of large differences from the hypothesized median are equally likely above or below the median, and ranks of small differences from the median are also equally likely above or below the hypothesized median). See the notes on hypothesis tests at the end of this document for more information.

Bivariate Analysis

The scatter plot can give us some intuition about a relationship between mortality rates and calcium concentration in England in general. The correlations can give a quantitative measure of that relationship.

In Class Examples

- Univariate analysis on mortality rates and calcium concentrations (ignoring **location**)
- Univariate visualizations (ignoring **location**)
- Tests of normality for each variable (ignoring **location**)
- Tests for significant correlation between mortality rates and water hardness

In Class Exercises

To check for a geographic effect (e.g. differences between the north and the south) we will want to visualize the data by **location**, obtain description statistics by **location**, and test for differences between locations.

- Perform the same analyses (univariate analysis with distributional testing, scatter plotting, and correlation testing) **by location**. (Note: we will first need to sort the data by location). What conclusions do be draw from these analyses? How do these differ from the results ignoring **location**?
- The overall mean mortality is about 1500 and the overall mean hardness is close to 45. Use these as null hypothesis values and test for significant differences for each geographic location (see the **mu0** option for **proc univariate**). Which test (t, sign, signed-rank) should we trust for each? What would we conclude about difference from the overall means for each geographic location?

Now we want to compare populations (north vs. south). We want to test if there is a significant difference between mortality rates in the north and that in the south, and to test if there is a significant difference between calcium concentrations in the water in the north and the south.

We can use a t-test via **proc ttest** if normality is reasonable for each sample from the two populations. If not, we can use **proc npar1way** with option **wilcoxon** to perform a nonparametric ranked sum test to test for a difference between the populations. (See the notes on hypothesis tests at the end of this document for more information about the ranked sum test).

- Based on our distributional testing from before, choose either **ttest** or **npar1way** to test for a significant difference between northern and southern mortality rates. What do we conclude about the difference?
- Similarly, use either **ttest** or **npar1way** to test for a difference between northern and southern calcium concentrations.

Another alternative to nonparametric method is transformation of the data to get more normally distributed data. The authors suggest that a log transformation of the **hardness** variable may make it more normal.

- Create a variable called **lhardnes** which is the log of the **hardness** variable and obtain descriptive statistics, a histogram, and normality tests by **location** to see if that variable is reasonably normal.
- If the log values seem reasonably normal use **proc ttest** to test for a geographical difference in the log of calcium concentration.

Some Notes on Hypothesis Tests and Confidence Intervals

The General Testing Framework

When performing a hypothesis test, we start with a null hypothesis H_0 and an alternative hypothesis H_A. We have a test statistic which will be computed based on our data. From the test statistic, we will obtain a p-value, which is the probability of observing a test statistic at least as extreme as the one we got if the null hypothesis is in fact true.

Based on the p-value, we will determine whether or not to reject the null hypothesis. If the p-value is small (typically smaller than some pre-determined significance level called alpha, and .05 is a pretty common choice for alpha), this gives us strong evidence that the null hypothesis is unlikely to be true, and we reject the null hypothesis. If the p-value is not too small, we conclude that we do not have significant evidence against the null hypothesis and we therefore do not reject the null hypothesis.

Here is some specific information about tests we are working with.

Tests of Population Location

Tests of location allow us to test hypotheses about the mean or median of a population. There are **one-sample tests** and **two-sample tests**. There are also **one-sided tests** and **two-sided tests**.

A **one-sample test** is based upon a single sample from a population and allows us to test some characteristic of that population (e.g. is the true mortality rate in towns in the UK significantly different from 0). The tests of location returned by **proc univariate** for a variable are one-sample tests.

A **two-sample test** is based on samples taken from two populations and allows us to test something about differences in the two underlying populations (e.g. is there a significant difference between water hardness in northern towns and water hardness in the southern town). When we perform hypothesis tests **by location** to compare north and south data using **proc ttest** or **proc npar1way**, the location (location here meaning location for the variable of interest not geographic location) tests returned are two-sample tests.

A one-sided test is only concerned with one direction of difference (e.g. is the true calcium concentration significantly greater than 0). A two-sided test is interested in some difference (e.g. is the true calcium concentration significantly greater *or* significantly less than 0).

t-Test

The t-test is a parametric test of the population mean μ based on the sample mean $\bar{x}$. The test assumes normality of the underlying population. The one-sample two-sided test is given as:

H_0: = $\mu = \mu_0$

H_A: $\mu \neq \mu_0$

The one-sample one-sided test with an alternative hypothesis that the true population mean is actually greater is given as:

H_0: = $\mu = \mu_0$

H_A: $\mu > \mu_0$

We just flip the inequality in H_A if we want the alternative that the true population mean is less.

The two-sample test compares the population means of two populations.

H_0: = $\mu_1 - \mu_2 = \mu_0$

H_A: $\mu_1 - \mu_2$ differs from μ_0 in a specific way (they're not equal for a two-sided test or the difference is greater or less than μ_0 for a one-sided test)

Typically μ_0 will be 0 for a two-sample test indicating that the population means are the same (under the null hypothesis).

Nonparametric Tests

The **sign test** subtracts the hypothesized median from the data values and counts the positive values (or the negative values—if we know the number of positive values and the number of 0s, be know the number of negatives and vice versa). Under the null hypothesis we expect half the data points to be greater than the median, so if the number of positive signs is far from the number of negative signs, this will provide evidence that the true median is significantly different from the null hypothesis value. The sign test assumes nothing about the shape of the distribution.

The **signed rank test** adds an assumption of symmetry that the sign test does not have. The signed rank test subtracts the hypothesized median value from each of the observations, and gets the ranks for the absolute value of those differences. The test statistic is based on the sum of the signed ranks (e.g. if the original observation was less than the hypothesized median, we multiple the rank by -1 to get its signed rank, and if the observation was greater than the hypothesized median we multiply by 1). Under the assumptions, signed rank statistics far from 0 will provide evidence against the null hypothesis.

Wilcoxon's rank sum test is a nonparametric test (does not depend on a parameter of the distribution). It is more robust than a t-test and does not assume normality. The test statistic is based on ranks of the data. Here's the basic idea. If we combine the two data sets and put the values in numeric order, the positions in the sorted list will be the **ranks** of the individual data points. If we take the ranks for all the data points from one of the data sets and add up those ranks, this will give us the test statistic. Under the null hypothesis the values of neither population tend to be greater than those in the other. Under the alternative, one of the populations tends to have larger values.

Under the null hypothesis for the **Wilcoxon rank sum test**, we would expect the rank for any data point in one sample to be as likely to be large (or small) as for any data point in the other sample. Under the null, we would expect the sum of the ranks for data set 1 to be $\mathbf{n_1 /(n_1+ n_2)*S}$ where **S** is the sum of all ranks. The distribution of the rank sum based on the assumptions of the test which will give us the p-value for the statistic.

Tests of Population Correlation

The results from **proc corr** include p-values for two-sided tests for correlation. For these, we have

H_0: There is no correlation between the two variables (the true correlation ρ is 0).

H_A: There is a significant correlation between the two variables ($|\rho| > 0$).

The test statistic is the sample correlation **r** (we have used the Pearson and Spearman correlations), and the p-value comes from the distribution of the test statistic assuming H_0 is true.

Tests of Distributional Goodness of Fit

Distributional goodness of fit tests based on the empirical cdf are called EDF tests (empirical distribution function tests). These include the **Kolmogorov-Smirnov**, **Anderson-Darling**, and **Cramer-von Mises** tests we saw in class. They are based on measures of the distance between the theoretical cdf and the empirical cdf. Kolmogorov-Smirnov looks at the largest absolute distance, and the other two are based on weighted squared distances.

Our null hypothesis is that the data came from a distribution of the specified type (normal distribution for instance). The alternative is that the data did not come from a distribution of that type. Smaller distance would tell us that the curves (and hence the distributions) are similar. Greater distance would be an indication that the curves are farther apart (and hence the distributions are less similar).

We have a one-sided test with less similar curves providing evidence against the null hypothesis. If we get a small p-value for an EDF test, we reject the null hypothesis that the data were sampled from the chosen distribution. Failure to reject in the case of the normal, for instance, means that we can proceed with other analyses that carry with them an assumption of normality.

Shapiro-Wilk is specifically a test of normality, while the EDF tests mentioned above are valid for other continuous distributions as well.

Confidence Intervals

When dealing with parametric statistics (those based on some property of the underlying population such as the population mean), we may want to give a measure of how confident we are in the value of true population parameter. We can think of this as an analogue to a parametric hypothesis test.

In hypothesis testing, we have a significance level α at which we determine whether or not a test result is statistically significant. For confidence intervals, we have a **confidence level** (often taken to be 1-α) which tells us about where we think the true value lives given our sample and assumptions on the underlying population. Confidence intervals for the mean of a normally distributed population come

directly from a t-distribution and are symmetric around the sample mean, though confidence intervals of other parameters under other assumptions may not be symmetric.

What do confidence intervals tell us? We have a sample from a population and some assumptions about that population (and the way the data were sampled). The confidence intervals are constructed such that $(1- \alpha)x100\%$ of intervals constructed in the same way for samples from the assumed population would contain the true value of the parameter we wish to estimate. For instance, if we take 1000 samples of the same size from the same normal distribution, and construct 95% confidence intervals for the mean from each of those samples, we would expect that about 950 of those intervals would contain the population mean and the others would not.

A common misconception about confidence intervals is that they tell us the probability that the true value is in the interval. This is not true. The true value is the true value. It's in the interval or it isn't. The confidence interval tells us about how confident we can be that we have located the true value.

Confidence intervals can be used to determine significant differences as well. For instance, if we have a confidence interval that contains 0, we could infer that there isn't a statistically significant difference between the population value and 0. If 0 is outside the confidence interval, we could infer that there is a statistically significant difference.