

## SASistics 101

Robert E. Johnson, Virginia Commonwealth University, Richmond, VA

### ABSTRACT

It is true that SAS<sup>®</sup> doesn't stand for *Statistical Analysis System* anymore, but there is plenty of statistics in SAS software. If you work with SAS software you will probably have to crunch some statistics from time to time. Base SAS procedures produce the most common statistical summaries, and more. In this tutorial, an overview of these procedures will be presented along with some examples. What is the difference between a standard deviation and a standard error? What is the difference between Pearson and Spearman correlations? What is a Chi and why do we square it? These and other burning questions will be addressed. Words of caution about using and interpreting SASistical procedures will be given throughout the tutorial.

### INTRODUCTION

Since you are reading this paper you are likely a programmer analyst, research data manager, or independent researcher who, from time to time, must perform fundamental statistical summaries or analyses on your data. Detailed applications which may involve advanced methods or presentation-ready reports are not the focus here. Rather, you need to provide basic descriptions of your data and, possibly, estimate a parameter or perform a significance test. While you might not be trained as a statistician, you are still expected to provide essential statistical support.

This tutorial is designed to present basic statistics organized as:

- Describing Your Data
- Relationships Between Measures
- Describing Your Population
- Significance Testing

The presentation is in no way exhaustive but is meant to jump-start you towards a few basic methods.

All SAS procedures presented are part of the Base SAS software, unless otherwise noted.

### DESCRIBING YOUR DATA

An Example. Suppose you have the data shown below and you need the mean and standard deviation.

Here is a simple program and its output.

```
data example;
  input regcode @@;
  datalines;
1 3 4 1 1 2 4 3 2 2
2 1 4 3 2 1 4 2 1 3
;

proc means mean std;
  var regcode;
run;
```

The SAS System  
Analysis Variable : REGCODE

| Mean      | Std Dev   |
|-----------|-----------|
| 2.3000000 | 1.1285762 |

So the mean is 2.3 and the standard deviation is 1.13? What is wrong with this picture? What is the *nature* of the variable REGCODE? If the values of this variable refer to coded categories of geographical regions (e.g., 1=northeast, 2=southeast, etc.), then such summaries as the mean and standard deviation are nonsense.

Numerical summaries can be easily performed on numerical data using SAS software, but you must be responsible for understanding the nature of the numerical data. Data may be classified as quantitative or qualitative. Quantitative data are best stored as numeric-type data since they are numerical by nature. Qualitative data may be stored as either numeric-type or character-type, but the data values, by nature, describe categories rather than numerical measures. Keep in mind that *how* data are stored does not always reflect the *nature* of the data.

Quantitative Data. How many items are produced in a day? What is the content weight of a cereal box? What is the per-unit cost of a manufactured item? These are examples of quantitative data which consist of numerical measures. The SAS procedure, UNIVARIATE may be used to summarize such variables. UNIVARIATE, as the name suggests, is best used when you wish to view numerical summaries for one variable at a time. The procedure MEANS may be used this way too, but the form of its output better allows for comparison between variables.

Baseball salaries for the 1994 professional baseball season are summarized in the output shown below. These data are as reported in USA Today on April 5, 1994. Salaries include pro-rated signing bonuses. The procedure statements used are

```
proc univariate data=salary;
  var salary;
run;
```

The resulting output is in Figure 1.

|                      |          |          |          |
|----------------------|----------|----------|----------|
| The SAS System       |          |          |          |
| Univariate Procedure |          |          |          |
| Variable=SALARY      |          |          |          |
| Moments              |          |          |          |
| N                    | 747      | Sum Wgts | 747      |
| Mean                 | 1183417  | Sum      | 8.8401E8 |
| Std Dev              | 1390922  | Variance | 1.935E12 |
| Skewness             | 1.317142 | Kurtosis | 0.645228 |
| USS                  | 2.489E15 | CSS      | 1.443E15 |
| CV                   | 117.5344 | Std Mean | 50891.17 |
| T:Mean=0             | 23.25387 | Pr> T    | 0.0001   |
| Num ^= 0             | 747      | Num > 0  | 747      |
| M(Sign)              | 373.5    | Pr>= M   | 0.0001   |
| Sgn Rank             | 139689   | Pr>= S   | 0.0001   |
| Quantiles (Def=5)    |          |          |          |
| 100% Max             | 6300000  | 99%      | 5250000  |
| 75% Q3               | 2000000  | 95%      | 4020000  |
| 50% Med              | 500000   | 90%      | 3400000  |
| 25% Q1               | 170000   | 10%      | 115000   |
| 0% Min               | 109000   | 5%       | 109000   |
|                      |          | 1%       | 109000   |
| Range                | 6191000  |          |          |
| Q3-Q1                | 1830000  |          |          |
| Mode                 | 109000   |          |          |
| Extremes             |          |          |          |
| Lowest               | Obs      | Highest  | Obs      |
| 109000(              | 720)     | 5400000( | 2)       |
| 109000(              | 719)     | 5406603( | 1)       |
| 109000(              | 718)     | 5500000( | 347)     |
| 109000(              | 717)     | 5975000( | 402)     |
| 109000(              | 695)     | 6300000( | 588)     |

Figure 1: Summary of 1994 Baseball Salaries

What is the center of these data? The total amount of money available for salaries is \$884,010,000 which corresponds to *Sum* in Figure 1. The *E8* simply means the number 8.8401 should be multiplied by  $10^8$  to obtain the value. If these moneys were distributed equally among all 747 players, then each player would receive \$1,183,417 which corresponds to *Mean*. So the *average* salary is about 1.2 million dollars. But does the mean adequately describe the center of the data?

The *quantiles* depict values where a certain percentage of the salaries fall below. The median is at the 50% quantile. This corresponds to *50% Med* in Figure 1 and is equal to \$500,000. At least one-half of the salaries are less than this and at least one-half are more. This value gives a more intuitive idea of the center of the data, but be careful not to use it as you would the mean. The total of all salaries is much more than  $747 \times \$500,000 = \$373,500,000$ .

Not all salaries are equal to the mean, that is, there is variation in salaries. A typical deviation from the mean, above or below, is given by the standard deviation. This corresponds to *Std Dev* in Figure 1 and is equal to \$1,390,922. Note that this value is larger than the mean! When there are several large deviations on one side of the mean the standard deviation is inflated.

An alternative measure of variation is the interquartile range, depicted in Figure 1 as *Q3-Q1*. Approximately 50% of the salaries fall between \$170,000 [25% *Q1*] and \$2,000,000 [75% *Q3*], a range of \$1,830,000 [*Q3-Q1*]. The larger this interquartile range, the greater the variation in the measures. Other ranges may be of interest as well. For example, 90% of the salaries range from \$109,000 [5%] to \$4,020,000 [95%], a range of \$3,911,000.

Be careful, the *Range* is different from the interquartile range. It measures the difference between the maximum and minimum values and thus depends only on these values. It is not a very useful measure of variation.

Some values may have extreme deviations from the center. The lowest and highest five values are shown in Figure 1. Since these values may be outliers worth investigating, the observation number, *obs*, is given in parentheses. The largest value, \$6,300,000, occurs in observation 588 - Bobby Bonilla, then a third-baseman for the New York Mets. There are several other statistics shown in Figure 1. Two of these statistics are described next. Others will be mentioned later.

*Skewness* and *Kurtosis* help describe the shape of the data's distribution. *Skewness* measures asymmetry. A positive {negative} value indicates that there are more extreme values above {below} the mean. *Kurtosis* measures the peakness or flatness of the distribution. Data which have a distribution which is flatter {more peaked} than a bell-shaped distribution have a negative {positive} kurtosis. A bell-shaped curve has skewness and kurtosis both equal to zero.

The distribution of all salaries over the range of salaries may be displayed with a histogram. The histogram below was produced using procedure *GCHART*, a procedure of the SAS®/Graph software. The high positive skewness can be clearly seen.

```
axis1 length=20;
proc gchart data=salary;
  vbar salary
    / space=0 axis=axis1 width=2
      midpoints=250000 to 6500000
        by 500000;
run;
```

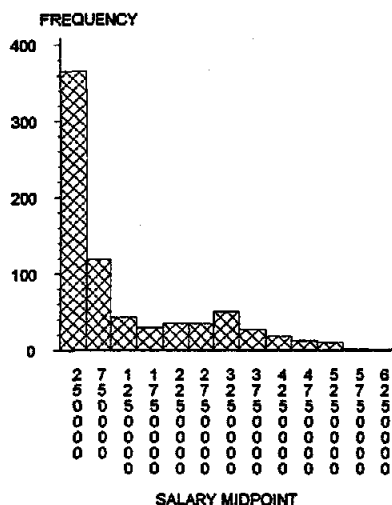


Figure 2: Salary Histogram

**Qualitative Data.** What sales region is an order shipped to? What team does a chosen baseball player belong to? What is the gender of a patron? These are examples of qualitative data. Each observation is classified by one level of a qualitative variable. Procedure `PROC FREQ` may be used to describe how your data are distributed over the levels of a qualitative variable.

Sales records for TruBlend Coffee Makers, Inc., contain information on the name of the sales representative [salesrep], type of coffee maker sold [type] and the number of units sold [units]<sup>1</sup>. Both salesrep and type are qualitative. The following code is used to produce simple frequency tables. Each record in the data set corresponds to the number of observations indicated by the variable units. The `weight` statement allows you to represent these observations. The partial output is given in Figure 3.

```
proc freq data=year;
  weight units;
  table salesrep type;
run;
```

| The SAS System |           |         |
|----------------|-----------|---------|
| SALESREP       | Frequency | Percent |
| Hollingsworth  | 10620     | 25.9    |
| Jones          | 14400     | 35.1    |
| Smith          | 15969     | 39.0    |
| TYPE           | Frequency | Percent |
| Deluxe         | 2525      | 6.2     |
| Standard       | 38464     | 93.8    |

Figure 3: Summary of Sales Records

The frequencies and percentages in the tables show how the 40989 coffee maker sales were distributed over the sales representatives and coffee maker types. A bar chart may also be used to display these frequencies but adds little to the table display when there are only two or three levels of the qualitative variable. With more levels, a bar chart is helpful, but it is still useful to display the frequencies and/or percentages.

A bar chart of sales by sales representatives is shown in Figure 4. Another way to display these data is with a pie chart. Such graphs are commonly seen in newspapers and magazines. A pie chart is too cluttered when there are more than three or four levels and adds little when there are two to four levels.

```
proc chart data=year;
  hbar salesrep/type=pct freq;
  format salesrep $1.;
run;
```

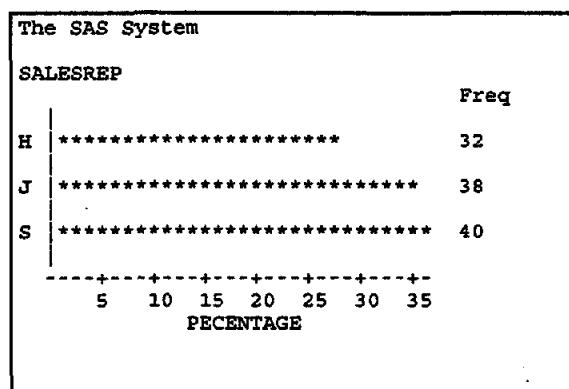


Figure 4: Bar Chart of Sales by Sales Rep.

## RELATIONSHIPS BETWEEN MEASURES

Is there a relationship between sales representative and the type of coffee maker sold? Is there a relationship between the skinfold measured on a subject's abdomen and the skinfold measured on the arm? What is meant by *relationship*? If the distribution of data over the values of one variable changes over values of another variable, we say the two variables are related. Consider the following examples.

**Two Qualitative Variables.** How do the three sales representatives for TruBlend Coffee Makers compare on deluxe units sold? Is the comparison different on standard units sold? Which sales representative sells the highest percentage of deluxe units? The `FREQ`

procedure produces two-way tables which enables us to explore answers to the posed questions.

Consider the following program. The partial output is given in Figure 5.

```
proc freq data=year;
  weight units;
  tables salesrep * type /chisq;
run;
```

| TABLE OF SALESREP BY TYPE                |        |          |        |
|------------------------------------------|--------|----------|--------|
| SALESREP                                 | TYPE   |          |        |
| Frequency                                |        |          |        |
| Percent                                  |        |          |        |
| Row Pct                                  |        |          |        |
| Col Pct                                  | Deluxe | Standard | Total  |
| Hollingsworth                            | 760    | 9860     | 10620  |
|                                          | 1.85   | 24.06    | 25.91  |
|                                          | 7.16   | 92.84    |        |
|                                          | 30.10  | 25.63    |        |
| Jones                                    | 820    | 13580    | 14400  |
|                                          | 2.00   | 33.13    | 35.13  |
|                                          | 5.69   | 94.31    |        |
|                                          | 32.48  | 35.31    |        |
| Smith                                    | 945    | 15024    | 15969  |
|                                          | 2.31   | 36.65    | 38.96  |
|                                          | 5.92   | 94.08    |        |
|                                          | 37.43  | 39.06    |        |
| Total                                    | 2525   | 38464    | 40989  |
|                                          | 6.16   | 93.84    | 100.00 |
| STATISTICS FOR TABLE OF SALESREP BY TYPE |        |          |        |
| Statistic                                | DF     | Value    | Prob   |
| Chi-Square                               | 2      | 25.257   | 0.001  |
| Phi Coefficient                          |        | 0.025    |        |
| Contingency Coefficient                  |        | 0.025    |        |
| Cramer's V                               |        | 0.025    |        |
| Sample Size = 40989                      |        |          |        |

Figure 5: Two-way Frequency Table

The percentages shown in Figure 5 under the title Total are the percentages shown in Figure 3 for the SALESREP distribution. This distribution is aptly called a *marginal distribution*. The marginal distribution of TYPE is shown across the bottom of the frequency table. The third row of each table cell [Row Pct] corresponds to the distribution of TYPE conditioned on which sales representative is being considered. For example, 7.16% of Hollingsworth's sales were deluxe units and 92.84% were standard units. The fourth row of each table cell [Col Pct] corresponds to the distribution of SALESREP conditioned on which type of coffee maker is being considered. For example, 30.1% of the deluxe units were sold by Hollingsworth whereas 37.43% were

sold by Smith. These distributions are suitably called *conditional distributions*.

If each conditional distribution is the same, then each conditional distribution equals the marginal distribution. That is, it doesn't matter which sales representative we are talking about, the distribution of type of unit sold is the same. If this is the case, the variables SALESREP and TYPE are not related or not associated. But it appears that Hollingsworth's conditional TYPE distribution does differ from the others (and from the marginal distribution). Also, the conditional distribution of deluxe units over sales representatives differs from the marginal distribution. So there is an appearance of a relationship.

The chi-squared value, provided when "/chisq" is included on the tables statement, is a measure of association. Chi-squared ( $\chi^2$ ) measures how much the conditional distributions differ from the marginal distribution. Small values correspond to little or no association. One problem with  $\chi^2$  is that its magnitude depends, in part, on the number of cells in the table and on the number of observations. Other statistics correct for this. For example, Cramer's  $v$  is derived from  $\chi^2$  and is scaled so that its maximum magnitude is 1. Larger values suggest a stronger association. The value of Cramer's  $v$  shown in Figure 5 is only 0.025. The association is very weak at best.

Inclusion of the option "/all" on the tables statement provides a barrel of statistics, some of which are meaningless for the data presented in Figure 5, but meaningful in other situations.

Graphic comparisons of conditional distributions provide ways to see the differences in the distributions. In Figure 6, the conditional distribution of deluxe unit sales are contrasted with the conditional distribution of standard unit sales over sales representatives.

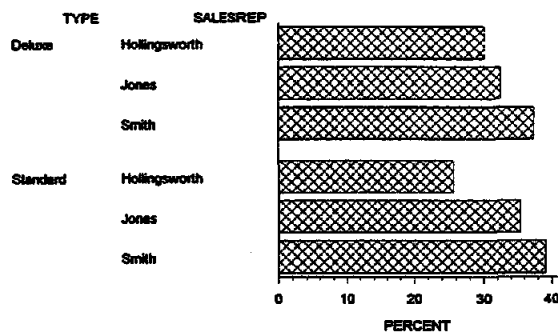


Figure 6: Comparison of Conditional Distributions

**Qualitative versus Quantitative.** How does the distribution of baseball salaries differ from team to team? Is there a relationship between the qualitative variable team and the quantitative variable salary? To explore this, we could compare salary summary statistics across teams. Procedure MEANS is used in the code below to produce such a summary. The mean and standard deviation are produced for four selected teams. The output is shown in Figure 7.

```
proc means data=salary mean std;
  class team;
  where ( team="Atlanta" or
         team="Baltimore" or
         team="Philadelphia" or
         team="Toronto");
  var salary;
run;
```

| The SAS System             |       |        |         |
|----------------------------|-------|--------|---------|
| Analysis Variable : SALARY |       |        |         |
| TEAM                       | N Obs | Mean   | Std Dev |
| Atlanta Braves             | 27    | 1.5E6  | 1.53E6  |
| Baltimore Orioles          | 28    | 1.35E6 | 1.55E6  |
| Philadelphia Phillie       | 29    | 1.08E6 | 990136  |
| Toronto Blue Jays          | 28    | 1.5E6  | 1.78E6  |

Figure 7: Salary Summaries for Four Teams

The means appear to vary as do the standard deviations. It seems the salary distributions for these teams do differ, thus TEAM and SALARY may be related.

This variation can be displayed with a graph. For each team, a *boxplot* is displayed. A boxplot uses five of the summary statistics provided by procedure UNIVARIATE. The lowest point is the minimum, the bottom of the box is the 25% quantile, the line inside the box is the median, the top of the box is the 75% quantile, and the highest point is the maximum. The following code produces the graph in Figure 8. The procedure GPLOT is part of SAS/Graph software.

```
symbol v=none interpol=boxt;
axis1 offset=(5) length=25;
axis2 length=20;
proc gplot data=salary;
  plot salary * team
    /haxis=axis1 vaxis=axis2;
  where ( team="Atlanta" or
         team="Baltimore" or
         team="Philadelphia" or
         team="Toronto");
  format team $1. salary 4.;
run;
```

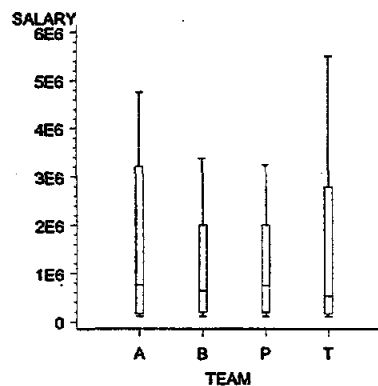


Figure 8: Comparative Boxplots of Salary

**Two Quantitative Variables.** Does the amount of skinfold on a subject's abdomen have a high association with the amount of skinfold on the subject's arm? If the distribution of abdomen skinfold differs from the distribution of abdomen skinfold among subjects with small arm skinfold, then an association exists. If the distribution of abdomen skinfold is the same regardless of the arm skinfold, then no association exists.

A scatterplot of two quantitative variables provides a visual aid. Here is the code and the resulting plot for skinfold data provided by A.C. Linnerud at NC State University<sup>2</sup>.

```
symbol value=dot h=2;
proc data=skinfold;
  plot abdomen * arm;
run;
```

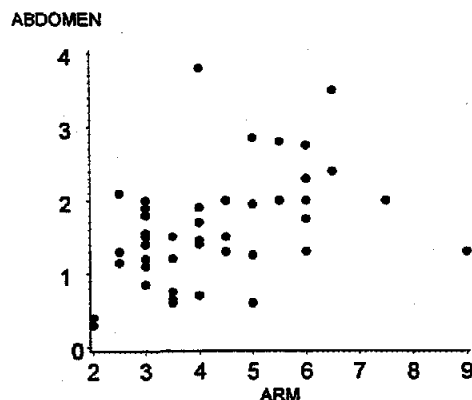


Figure 9: Scatterplot of Skinfold Measures

As the arm skinfold increases, there appears to be an increase in the *average* abdomen skinfold. This suggests that the abdomen skinfold is distributed around a larger mean for large arm skinfold than for small arm skinfold.

The temptation here is to fit a straight line to these data. A regression line is a useful aid in viewing an association, but be careful. Is it reasonable to assume that an increase of one unit in arm skinfold corresponds to a constant increase in mean abdomen skinfold, regardless of the arm skinfold value? This straight line relationship may not be the true relationship.

The procedure CORR provides measures of the degree of association between two quantitative variables. Here is some code and partial output.

```
proc corr data=skinfold
      pearson spearman;
var arm;
with abdomen;
run;
```

| Pearson Correlation Coefficients /  |         |
|-------------------------------------|---------|
| Prob >  R  under Ho: Rho=0 / N = 50 |         |
| ABDOMEN                             | ARM     |
|                                     | 0.42925 |
|                                     | 0.0019  |
| Spearman Correlation Coefficients / |         |
| Prob >  R  under Ho: Rho=0 / N = 50 |         |
| ABDOMEN                             | ARM     |
|                                     | 0.45687 |
|                                     | 0.0009  |

Figure 10: Skinfold Data Correlations

The procedure CORR produces four measures of association (or correlation). Only two will be mentioned here. All correlations are between -1 and 1. A value of zero indicates no correlation.

*Pearson* is equal to 1 if all the points in the scatter plot fall on a (non-horizontal) straight line. The correlation moves toward zero as the points' variation around the regression line increases.

*Spearman* measures the correlation between the *ranks*. That is, rank the arm skinfolds from 1 to 50 and, independently, rank the abdomen skinfolds. Calculating the *Pearson* correlation on these ranks yields *Spearman* correlation. For *Spearman* to be equal to 1, the height of the points must increase from left to right, but not necessarily in a straight line.

To further compare *Pearson* and *Spearman* correlations, look at the four graphs in Figure 11.

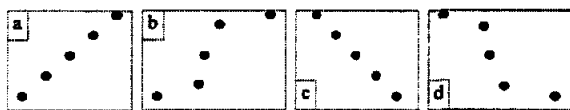


Figure 11: Positive and Negative Correlations

The *Pearson* correlation is a: 1, b: 0.90, c: -1 and d: -0.90 for the scatter plots in Figure 11. In contrast, the *Spearman* correlation is a: 1, b: 1, c: -1 and d: -1.

For descriptive purposes, if you want to measure the correlation relative to a straight line then use *Pearson*. Otherwise, use *Spearman* or one of the other provided measures. But note that when a publication states "The correlation is ...," the author probably used *Pearson*, unless otherwise stated.

## DESCRIBING YOUR POPULATION

What is the mean salary of all professional baseball players? What percentage of all coffee maker units sold are the deluxe model?

Data often represent a sample of a larger population. While descriptions of the data are important, inferences about the population based on the data is often the goal of data collection. Can the data's mean be used as an estimate of the population's mean? The answer is *yes* if the data constitute a random sample of the population. The answer might be *yes* for non-random samples too, but there is really no way to evaluate the accuracy and precision of the estimate.

Each possible random sample of size  $n$  from a population has an associated sample mean. These means vary from sample to sample. A typical deviation of a sample mean from a population mean is called the standard deviation of the sample mean. The *standard error* is an estimate of the standard deviation of all sample means based on the given sample data.

The UNIVARIATE procedure provides both the data's mean and its standard error. In Figure 1, the value for Std Mean represents the standard error if the data set is a random sample from the population and if the sample size,  $N$ , is small relative to the population size (say, no more than 10% of the population size).

For the data of Figure 1, the standard error (Std Mean 50891.17) is *not* meaningful. These data consist of the entire population of baseball players for 1994. Therefore the calculated mean is the population mean and is not subject to variation due to sampling. The UNIVARIATE procedure does not know this, of course, and calculates Std Mean according to its formula.

If the data constitute a random sample, then the data's mean is an unbiased estimate of the population's mean. That is, the average sample mean - over all possible samples - is equal to the population mean. The standard error derived from a random

sample is an appropriate measure of precision. A larger random sample results in a smaller standard error, that is, the sample mean is a more precise estimate of the population mean.

Simple random sampling is just one of several sampling methods. SAS does not provide comprehensive methods for dealing with complex samples, especially for estimating variation in sample estimates. You should be aware that if the data for which you are providing statistical summaries were gathered with a complex sampling design, the results provided by standard procedures may not be valid. Of course, if your data were not randomly sampled, no definitive inference can be made to the population.

## SIGNIFICANCE TESTING

Many of the procedures presented produce p-values. These values can help you determine if certain differences between observed values and hypothesized values are due to variation which occurs naturally in random sampling or due, in part, to a real differences between the hypothesized values and the *true* values. Please note that these p-values only make sense if the data are a random sample from a larger - relative to the sample size - population. Once again, SAS procedures don't know this and will compute the p-values according to their formulas regardless.

The line in Figure 1 which reads

```
T:Mean=0 23.25387 Pr>|T| 0.0001
```

provides the results of a *t*-test. Here the null hypothesis is that the mean salary in the population is zero versus the alternative that it is different from zero (the two-sided case). The value 23.25387 is a measure of the difference between the observed mean of 1,183,417 and the hypothesized mean of zero, in units of the standard error of the mean, 50,891.07. Wait a moment... the data presented in Figure 1 is the population, not a random sample! As mentioned before, the standard error is not meaningful here. Neither is the t-test nor the p-value. Since we observed the *population* mean to be different from zero, we can reject the null hypothesis with complete confidence.

If the data in Figure 1 were a random sample from a much larger population, then the p-value of 0.0001 says that it is very unlikely (odds less than 1 out of 10,000) that a random sample could yield a sample mean as far from zero as 23.25 or further if the population mean is truly zero. This would constitute strong statistical evidence that the population is not equal to zero.

If you wish to use a value different from zero in your null hypothesis, such as 1.2 million, then use code similar to

```
data modsalary;
  set salary;
  modsalary = salary - 1200000;
proc univariate data=modsalary;
  var modsalary;
run;
```

The coffee maker data summarized in Figure 5 represent all sales over a certain period of time, not really a true simple random sample. It may be true that this sample fairly represents a larger population of sales over a larger period of time, but we can't be sure. Any related p-values should be viewed with extreme caution, with the most extreme being: don't use them.

The FREQ procedure's  $\chi^2$  test quantifies the evidence provided by the data against the null hypothesis that SALESREP and TYPE are not related. The p-value of 0.001 says that it is very unlikely (odds less than 1 in 1000) that the observed  $\chi^2$  value could as large as 25.257 or larger if the population conditional distributions are really the same (there is no association). The statistical evidence strongly rejects the null hypothesis of no association. But be careful! Look at the value of Cramer's  $v$  and look at the graph in Figure 6. Neither of these suggest strong association between the variables of interest. The very large sample size of nearly 41,000 makes it very likely that a significance test will detect even small variations in the conditional distributions. Statistical significance doesn't mean that true differences constitute meaningful significance.

The skinfold data displayed in Figure 9 were randomly sampled from a larger population. The correlations depicted in Figure 10 are both significantly - statistically - different from zero since the p-values are 0.0019 for *Pearson* and 0.0009 for *Spearman*. Thus the data provide strong evidence of an association between arm skinfold and abdomen skinfold. The magnitudes of the correlations tell how strong the association is. Most researchers would consider a correlation of 0.45 as weak to moderate.

## SUMMARY

The SAS system provides many methods of obtaining summary statistics for your data. Regardless of how you obtain your summaries you *must* be concerned with

- the nature of the data
- the source of the data (random sample?)
- the relevance of the output statistics

If you plan to provide extensive statistical support, at best consult a qualified statistician, or, at least, educate yourself about basic statistics. There are several good books on the market which will aid you in your endeavors.<sup>3 4 5</sup>

---

®SAS and SAS/Graph are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

## REFERENCES

<sup>1</sup> SAS Institute Inc. (1993), *SAS Language and Procedures: Usage, Version 6, First Edition*, Cary, NC: SAS Institute Inc., page 366.

<sup>2</sup> SAS Institute Inc. (1990), *SAS Procedures Guide, Version 6, Third Edition*, Cary, NC: SAS Institute Inc., pages 227-231.

<sup>3</sup> Dilorio, Frank C. and Hardy, Kenneth A. (1996), *Quick Start to Data Analysis with SAS*, New York: Duxbury Press.

<sup>4</sup> Moore, David S. (1991), *Statistics: Concepts and Controversies*, New York: W. H. Freeman and Company.

<sup>5</sup> Schlotzhauer, Sandra D. and Littell, Ramon C. (1991), *SAS System for Elementary Statistical Analysis*, Cary, NC: SAS Institute Inc.