

SESUG Paper SD-124-2017

Two Sample Parametric and Nonparametric Statistical Tests for a Large Number of Comparisons Made Easy

Dennis J. Beal, Leidos, Oak Ridge, Tennessee

ABSTRACT

Parametric and nonparametric two sample statistical tests have been in SAS® for a long time. However, for environmental data there often are many constituents to compare their mean or median concentrations to see if they are statistically the same or not. Typically, hundreds of chemical or radionuclide constituents need to be statistically tested between two sampling locations. This paper provides SAS code that efficiently calculates the summary statistics and output from two sample statistical tests by constituent including p -values and a final determination whether they are significantly different at a given significance level. Parametric t-tests are performed when the data are normally or lognormally distributed in both samples. The nonparametric two sample Wilcoxon Rank Sum (WRS) test is performed when the data are neither normally nor lognormally distributed. Output delivery system (ODS) is used to export p -values from statistical tests into a data set so conclusions can be made easily whether two sample locations have significantly different levels of constituents. This paper is for intermediate SAS users of Base SAS®.

Key words: two sample tests, t-test, Wilcoxon Rank Sum, environmental data, dynamic data exchange

INTRODUCTION

Often in environmental applications, project managers want to know whether two sampling locations have significantly different levels of concentrations from each other. For example, a project may want to compare the data from one well that is located upgradient from a landfill with another well that is located downgradient from the landfill. When samples are collected on a quarterly or semi-annual basis, the laboratory often analyzes for hundreds of constituents. Once the data are validated from the laboratory, the statistician is asked to evaluate the data to determine if the two wells are statistically similar or different from each other. If the statistician determines the data is normally distributed for a particular constituent in both wells, then the parametric two sample t-test can be used to test the null hypothesis of equal means for a given significance level. The distribution is determined using the Shapiro-Wilk test for normality. If the statistician determines the data is lognormally distributed in both wells, then the parametric t-test can be used on the log transformed concentrations to test the null hypothesis of equal medians for a given significance level. If the data are neither normal nor lognormal, the nonparametric Wilcoxon Rank Sum test can be used to test the null hypothesis of equal medians.

This paper will focus on using efficient SAS code and the ODS output from `PROC TTEST` and `PROC NPAR1WAY` to save important statistics from the t-test and WRS test into SAS data sets. These statistics will then be merged with summary statistics and output into a table in Microsoft Excel. An example will be shown using environmental data. The SAS code presented in this paper uses the SAS System for personal computers version 9.4 running on a Windows® 7 Professional platform.

EXAMPLE WITH SAS CODE

This paper will use an environmental data set collected from two groundwater wells A and B as an example. The project required each well to be sampled quarterly over a period of several years and analyzed for inorganics and radionuclides. For each constituent, the project sought to determine whether the two wells have statistically similar mean or median concentrations at the 0.05 significance level.

After assimilating any duplicate samples in order to ensure statistical independence and unbiasedness between sampling events, the statistician calculated summary statistics for each well. These summary statistics include the frequency of detection (the number of “detected” results out of the total number of results), mean, median, standard deviation and maximum detected concentration. Kaplan-Meier means, medians, and standard deviations were calculated for inorganics when non-detects (left-censored data)

were present with a sufficient number of detected concentrations (Beal 2010, Beal 2015, Helsel 2005, Helsel 2012). For radionuclides, the statistician used the actual reported concentrations from the laboratory to calculate summary statistics whether detected or not, even if negative or zero.

In order to determine a distribution for each constituent, half the detection limits were used as proxy values (SAS variable `proxyres`) for non-detect inorganics. Otherwise, the proxy value is the reported concentration. For radionuclides and inorganics with all detected concentrations the data were summarized using `PROC UNIVARIATE`. The variable `hit` has the value 0 if the concentration is a non-detect and 1 if the concentration is detected. The variable `logproxyres` is the natural logarithm of the proxy concentration when the proxy concentration is positive. The variable `anagroup` is the analysis group of the constituent (e.g., metals or radionuclides). The variable `chemical` is the constituent name (e.g., arsenic, lead, uranium-234). The variable `units` is the corresponding units for the concentration (e.g., $\mu\text{g/L}$, pCi/L). The variable `station` is the well identification (e.g., Well A, Well B).

```
proc sort data=dat; by station anagroup chemical units;

proc univariate data=dat;
  var proxyres hit logproxyres;
  by station anagroup chemical units;
  output out=statsout(drop=x y) min=min mean=mean median=median n=n std=std
    max=max sum=x hitsum probn=probn y probl;
run;
```

The variable `probn` is the Shapiro-Wilk p -value that tests the data for normality. The variable `probl` is the Shapiro-Wilk p -value that tests the data for lognormality. For inorganics that have non-detects and a sufficient number of detected concentrations, the Kaplan-Meier mean, median, and standard deviation are calculated (Beal 2010, Beal 2015). These statistics are stored in a SAS data set `KM_SUMSTATS` with variable names `km_mean`, `km_median`, and `km_std`.

These summary statistics are then merged together. The variable `DIST` has the value 'N' if the data are determined to be normally distributed, 'L' if the data are lognormally distributed and 'X' if the data are neither normal nor lognormal (i.e., nonparametric) at the 0.05 significance level.

```
proc sort data=km_sumstats; by station anagroup chemical units;

data withsumstats;
  merge statsout km_sumstats;
  by station anagroup chemical units;
  if probn > 0.05 then DIST = 'N'; ** data are normally distributed;
  else if probl > 0.05 then DIST = 'L'; ** data are lognormally distributed;
  else DIST = 'X'; ** data are nonparametric distributed;
run;
```

The SAS code `ods trace on / listing` shows the name of the SAS tables created by the nonparametric procedure `NPAR1WAY` so that pertinent SAS variables can be extracted from the SAS table. The SAS option `Wilcoxon` specifies the WRS test be used because there are only two wells to compare. See Table 83.6 in the SAS documentation for a list of all the ODS table names generated by `PROC NPAR1WAY`. The variable `pt2_wil` is the two-sided WRS p -value using the t approximation. The variable `p_kw` is the Kruskal-Wallis (KW) p -value. These p -values are output into the SAS data set `tabout` and will be used when the data from one or both of the wells is nonparametric or the wells do not have the same distribution (e.g., Well A is normally distributed and Well B is lognormally distributed).

```
proc sort data=withsumstats; by anagroup chemical units station;
ods trace on / listing;
proc npar1way data=dat wilcoxon;
  class station;
  var proxyres;
  by anagroup chemical units;
```

```

output out=tabout(keep=anagroup chemical units pt2_wil p_kw);
run;
ods trace off;

```

PROC TTEST will perform the parametric F-test testing the null hypothesis of equal variances when the data from both wells A and B are normally distributed for a given constituent. Using the ODS output, the SAS table name `equality` is created and renamed `ttest_eq_var`. The variable `probf` contains the F distribution *p*-value. When the F *p*-value is small (e.g., below 0.05 significance level), then the conclusion is there is 95 percent confidence that the variances are not statistically equal. If the F *p*-value exceeds the 0.05 significance level, then the conclusion is there is 95 percent confidence that the variances are statistically equal. This is important to know so the statistician uses the proper t-test assumptions of equal or unequal variances because the test statistics are different for the two assumptions. The SAS option `dist=normal` specifies the data are assumed to be normally distributed for both wells.

```

ods trace on / listing;
ods output equality=ttest_eq_var(keep=anagroup chemical units probf);
proc ttest data=dat dist=normal ;
  class station;
  var proxyres;
  by anagroup chemical units;
run;
ods output close;
ods trace off;

```

Now PROC TTEST will perform the parametric t-test testing the null hypothesis of equal means when the data from both wells A and B are normally distributed for a given constituent. The SAS data set `ttest_pvals` contains the *p*-values for both the equal variances and unequal variances tests.

```

ods trace on / listing;
ods output ttests=ttest_pvals;
proc ttest data=dat dist=normal;
  class station;
  var proxyres;
  by anagroup chemical units;
run;
ods output close;
ods trace off;

```

The *p*-values from the t-test assuming equal variances are stored in the SAS data set `eqvar` and renamed `probtn_eq_pvalue`.

```

data eqvar;
  set ttest_pvals;
  where variances = 'Equal';
  rename probt=PROBTN_EQ_PVALUE;
  keep anagroup chemical units probt; run;

```

Similarly, the SAS data set `uneqvar` stores the *p*-values from the t-test assuming unequal variances and the variable `probt` is renamed `probtn_uneq_pvalue`.

```

data uneqvar;
  set ttest_pvals;
  where variances = 'Unequal';
  rename probt=PROBTN_UNEQ_PVALUE;
  keep anagroup chemical units probt; run;

```

PROC TTEST will perform the parametric F-test testing the null hypothesis of equal variances when the data from both wells A and B are lognormally distributed for a given constituent. Using the ODS output, the SAS table name `equality` is created and renamed `ttestl_eq_var`. The variable `probf` contains the F

distribution p -value and is renamed `probf1`. The SAS option `dist=lognormal` specifies the data are assumed to be lognormally distributed for both wells.

```
ods trace on / listing;
ods output equality=ttestl_eq_var(keep=anagroup chemical units probf
rename=(probf=PROBFL));
proc ttest data=dat dist=lognormal;
  class station;
  var proxyres;
  by anagroup chemical units;
run;
ods output close;
ods trace off;
```

Now PROC TTEST will perform the parametric t-test testing the null hypothesis of equal medians when the data from both wells A and B are lognormally distributed for a given constituent. When the t-test is used on log transformed data, the t-test actually becomes a test on the medians due to the transformation back to original units. The SAS data set `ttest_pvals` contains the p -values for both the equal variances and unequal variances tests.

```
ods trace on / listing;
ods output ttests=ttestl_pvals;
proc ttest data=dat dist=lognormal;
  class station;
  var proxyres;
  by anagroup chemical units;
run;
ods output close;
ods trace off;
```

The p -values from the t-test assuming equal variances are stored in the SAS data set `eqvar1` and renamed `probt1_eq_pvalue`.

```
data eqvar1;
  set ttestl_pvals;
  where cvs = 'Equal';
  rename probt=PROBT1_EQ_PVALUE;
  keep anagroup chemical units probt; run;
```

Similarly, the p -values from the t-test assuming unequal variances are stored in the SAS data set `uneqvar1` and renamed `probt1_uneq_pvalue`.

```
data uneqvar1;
  set ttestl_pvals;
  where cvs = 'Unequal';
  rename probt=PROBT1_UNEQ_PVALUE;
  keep anagroup chemical units probt; run;
```

Next the summary statistics from each well need to be stored in a SAS data set and renamed with the well designation in the variable name. The variable `PDETC` is the frequency of detection (number of detected samples / total number of samples) as a character variable.

```
data well_A;
  length PDETC $9;
  set withsumstats;
  where station = 'Well A';
  PDETC_A = trim(left(put(hitsum, 3.0))) || ' / ' || trim(left(put(n, 3.0)));
  rename dist=DIST_A min=MIN_A mean=MEAN_A km_mean=KM_MEAN_A median=MEDIAN_A
        km_median=KM_MEDIAN_A max=MAX_A hitsum=HITSUM_A n=N_A;
```

```

keep anagroup chemical units pdetc_A dist mean km_mean median km_median
max;

data well_B;
  length PDETC $9;
  set withsumstats;
  where station = 'Well B';
  PDETC_B = trim (left (put (hitsum, 3.0)))||' / '||trim(left(put(n, 3.0)));
  rename dist=DIST_B min=MIN_B mean=MEAN_B km_mean=KM_MEAN_B median=MEDIAN_B
         km_median=KM_MEDIAN_B max=MAX_B hitsum=HITSUM_B n=N_B;
  keep anagroup chemical units pdetc_B dist mean km_mean median km_median
max;

```

SAS code then merges these summary SAS data sets to create the data set that will be exported into Excel. Constituents with enough detected concentrations for organics or enough total number of samples for radionuclides are kept. A minimum of eight detected concentrations for inorganics or at least eight samples for radionuclides are needed to perform two sample statistical tests. Logic is used to assign the *p*-value from the appropriate statistical test for the resulting table. If non-detects are present for any inorganic, then the Kaplan-Meier mean and median are displayed in the table. Otherwise, for radionuclides or inorganics with all detected concentrations, the usual arithmetic mean and median are displayed.

```

proc sort data=well_A; by anagroup chemical units;
proc sort data=well_B; by anagroup chemical units;

data alldat2;
  merge well_A well_B tabout ttest_eq_var eqvar uneqvar ttestl_eq_var eqvarl
  uneqvarl;
  by anagroup chemical units;
  if pdetc_A = ' ' or pdetc_B = ' ' then delete;
  if anagroup = 'Radionuclides' and (n_A < 8 or n_B < 8) then delete;
  if anagroup ^= 'Radionuclides' and (hitsum_A < 8 or hitsum_B < 8) then
  delete;
  if dist_A = 'L' and dist_B = 'L' and min_A > 0 and min_B > 0 and probfl >
  0.05 then PVALUE_USE = probtl_eq_pvalue;
  else if dist_A = 'L' and dist_B = 'L' and min_A > 0 and min_B > 0 and 0 <
  probfl <= 0.05 then PVALUE_USE = probtl_uneq_pvalue;
  else if dist_A = 'N' and dist_B = 'N' and probf > 0.05 then PVALUE_USE =
  probtn_eq_pvalue;
  else if dist_A = 'N' and dist_B = 'N' and 0 < probf <= 0.05 then
  PVALUE_USE = probtn_uneq_pvalue;
  else PVALUE_USE = pt2_wil;
  if pvalue_use <= 0.05 then SIGNIFICANT = 'Yes';
  else significant = 'No';
  if km_mean_A ^=. and 0 < hitsum_A < n_A then do;
    MEAN_TABLE_A = km_mean_A; MEDIAN_TABLE_A = km_median_A; end;
  else if hitsum_A = n_A then do;
    MEAN_TABLE_A = mean_A; MEDIAN_TABLE_A = median_A; end;
  if km_mean_B ^=. and 0 < hitsum_B < n_B then do;
    MEAN_TABLE_B = km_mean_B; MEDIAN_TABLE_B = km_median_B; end;
  else if hitsum_B = n_B then do;
    MEAN_TABLE_B = mean_B; MEDIAN_TABLE_B = median_B; end;

proc print data=alldat2; run;

```

EXPORT THE RESULTS TO EXCEL

After storing the table of results into the SAS data set `alldat2`, the results are ready to be exported into Microsoft Excel. The fastest way to send the data to Excel is to use `PROC EXPORT` as seen in the following SAS code. The `retain` statement ensures the variables are exported in the order desired.

```

data exportit;
  retain ANAGROUP CHEMICAL UNITS PDETC_A DIST_A MEAN_TABLE_A MEDIAN_TABLE_A
    PDETC_B DIST_B MEAN_TABLE_B MEDIAN_TABLE_B SIGNIFICANT;
  set alldat2;
run;

proc export data=exportit outfile='C:\Myfiles\Well comparison results.xlsx'
  dbms=xlsx;
run;

```

However, Dynamic Data Exchange (DDE) is a powerful tool that will export the resulting SAS data set into a high quality formatted table in Excel. A SAS program can be written that customizes every detail of the table including centering and merging cells, italicizing cells or even parts of cells, bolding words or parts of words in specific cells, fixing widths of columns, and highlighting results in any color for the reader. While providing the DDE code is not the purpose of this paper, other papers provide the information to construct the SAS code (Beal 2004, Beal 2006). Table 1 shows the finished table formatted using DDE. The SAS code for sending the results to Excel using DDE can be provided by the author upon request.

Constituent	Well A				Well B				Signifi- cant? ^a
	Freq. of Detection	Dist.	Mean ^b	Median ^b	Freq. of Detection	Dist.	Mean ^b	Median ^b	
Metals (mg/L)									
Aluminum	11 / 14	L*	0.573	0.091	27 / 47	X*	0.039	0.015	Yes
Barium	14 / 14	N	0.265	0.265	47 / 47	X	0.22	0.211	Yes
Iron	13 / 14	L*	0.512	0.223	46 / 47	X*	0.244	0.137	No
Lithium	14 / 14	N	0.01	0.01	41 / 41	X	0.012	0.012	Yes
Magnesium	14 / 14	N	5.34	5.33	47 / 47	X	10.1	10.2	Yes
Manganese	14 / 14	X	0.072	0.052	47 / 47	X	0.017	0.015	Yes
Phosphorus	14 / 14	X	0.043	0.036	10 / 27	X*	0.007	0.007	Yes
Potassium	14 / 14	X	1.03	0.906	47 / 47	X	2.04	2.02	Yes
Sodium	14 / 14	X	5.88	5.78	47 / 47	X	49.4	9.64	Yes
Strontium	14 / 14	N	0.095	0.096	47 / 47	N	0.665	0.67	Yes
Titanium	8 / 14	X*	0.011	0.003	19 / 32	X*	0.002	6E-04	Yes
Radionuclides (pCi/L)									
Europium-152	1 / 14	N	2.35	0.655	1 / 35	N	0.59	0.2	No
Europium-155	1 / 14	L	-2.19	-0.885	0 / 35	L	-3.34	-1.76	No
Strontium-90	3 / 14	N	0.531	0.639	10 / 48	L	0.319	0.23	Yes
Technetium-99	5 / 14	N	1.41	1.47	9 / 48	N	0.837	0.99	No
Thorium-230	7 / 14	N	0.277	0.265	30 / 48	L	0.395	0.255	No
Thorium-232	1 / 14	X	0.026	0.015	11 / 48	L	0.091	0.06	Yes
Thorium-234	2 / 14	L	0.086	0.065	2 / 34	L	0.056	0.045	No
Uranium-234	6 / 14	L	0.164	0.105	27 / 48	L	0.206	0.18	No
Uranium-235	1 / 14	N	0.06	0.045	12 / 46	L	0.082	0.06	No
Uranium-238	2 / 14	L	0.086	0.065	12 / 48	L	0.089	0.065	No

^a significant at the 0.05 significance level

^b This summary statistic is calculated using both detects and non-detects

Dist. = distribution. Distribution flags are defined as:

L = lognormal

N = normal

X = neither normal, lognormal nor gamma

-- = insufficient data to calculate the statistic

* The mean and median were calculated using Kaplan-Meier because there are non-detects present with at least two distinct detected results.

Table 1. Statistical comparison of Well A with Well B using data from 2003 through 2016

CONCLUSION

Two sample statistical tests on the mean or median concentrations are common statistical tools that determine whether two samples have significantly different concentrations in the center of their distributions. Using ODS output within the SAS system, output from the parametric F test for equality of variances, t-test on means and nonparametric Wilcoxon Rank Sum test on medians can be stored into SAS data sets and used with appropriate logic to create a table that efficiently summarizes the results with minimal human intervention. SAS code is shown using an example from environmental data that efficiently calculates the necessary statistics for hundreds of constituents and exports a table into Microsoft Excel. DDE can be used to customize the table directly from within SAS.

REFERENCES

- Beal, Dennis J. 2006. "Using Advanced Features of DDE to Customize Formatted Tables in Excel," *Proceedings of the 14th Annual Conference of the SouthEast SAS Users Group*.
- Beal, Dennis J. 2004. "Using Dynamic Data Exchange to Customize Formatted Reports in Microsoft Excel," *Proceedings of the 12th Annual Conference of the SouthEast SAS Users Group*.
- Beal, Dennis J. 2010. "A Macro for Calculating Summary Statistics on Left Censored Environmental Data using the Kaplan-Meier Method," *Proceedings of the 18th Annual Conference of the SouthEast SAS Users Group*.
- Beal, Dennis J. 2015. "A Macro for Calculating Percentiles on Left Censored Environmental Data using the Kaplan-Meier Method," *Proceedings of the 23rd Annual Conference of the SouthEast SAS Users Group*.
- Helsel, Dennis R. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. New Jersey: John Wiley & Sons, Inc.
- Helsel, Dennis R. 2012. *Statistics for Censored Environmental Data Using Minitab and R*. New Jersey: John Wiley & Sons, Inc. (Second Edition).

CONTACT INFORMATION

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

Dennis J. Beal, Ph.D.
Senior Statistician/Risk Scientist
Leidos
301 Laboratory Road
P.O. Box 2502
Oak Ridge, Tennessee 37830

phone: 865-481-8736
e-mail: beald@leidos.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are registered trademarks or trademarks of their respective companies.