SAS® GLOBAL FORUM 2018

USERS PROGRAM

An unusual remedy using the usual "nbins" option to rectify anomalous histograms in SAS

Rachana Lele
Graduate Student
MS Biostatistics, University of Louisville, KY
rachanak.lele@louisville.edu

April 8 – 11 | Denver, CO
#SASGF

An unusual remedy using the usual “nbins” option to rectify anomalous histograms in SAS

Rachana Lele

Graduate Student, MS Biostatistics, University of Louisville, KY

rachanak.lele@louisville.edu

OUTLINE

Background

- **Histograms** are used for assessing **normality** of data
- In **SAS®**, **SGPLOT/SGPANEL** or **UNIVARIATE** procedures can be used for generating histograms
- Histograms plotted in SAS® using SGPLOT/ SGPANEL procedures show an **anomaly** when the **largest value** in a set of data coincides with one of the **tick points on the x-axis of the histogram**

Present Work

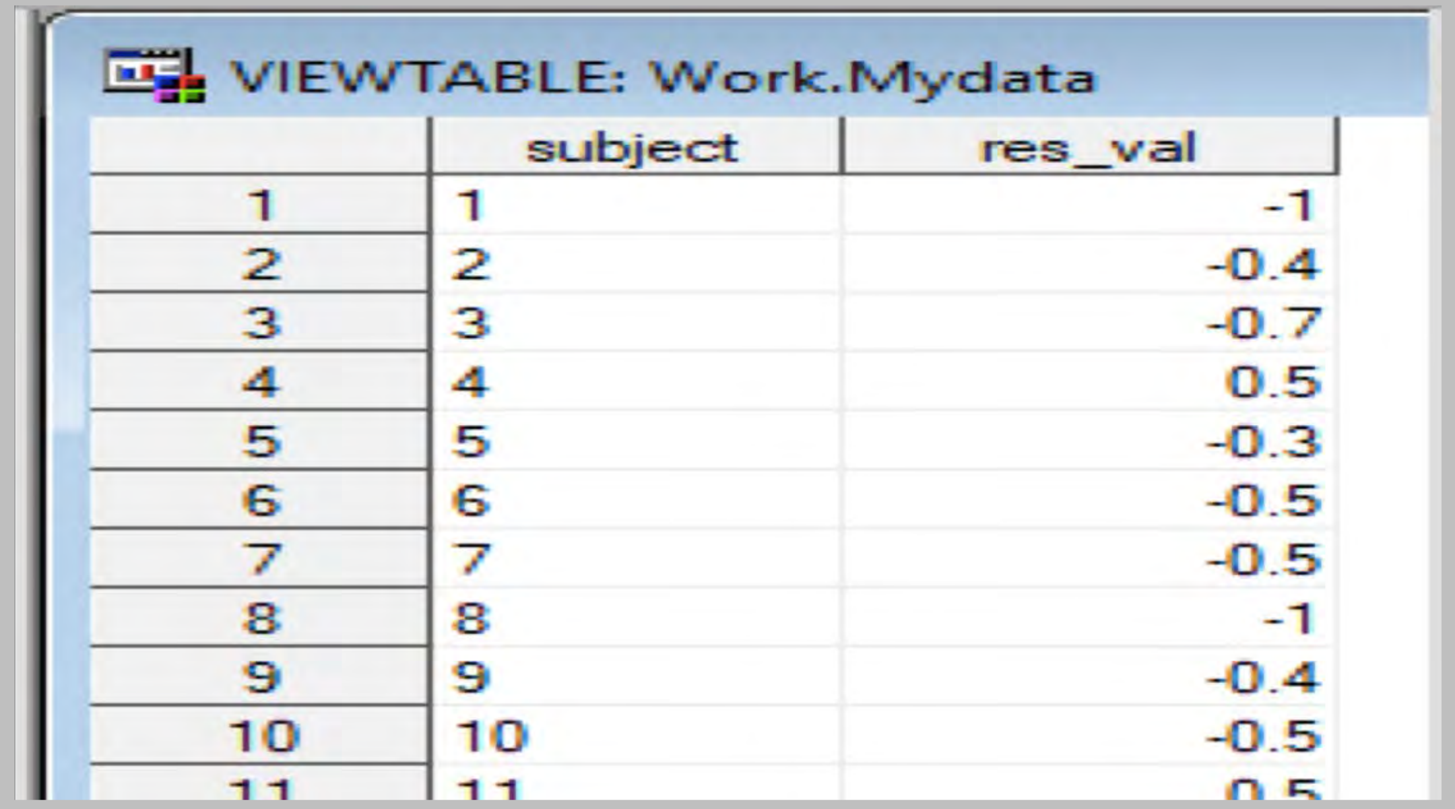
- Discusses this anomaly and suggests a **remedy** for solving it
- Suggests that **UNIVARIATE** procedure can be used to **validate** the histograms produced using the **SGPLOT/ SGPANEL** procedures
- **Alternative method** for calculating ***binstart*** in the SGPLOT/SGPANEL procedures

INTRODUCTION

- **Normality check** is the first step for analysis of data
- **Normal** data → **Symmetric** bell-shaped histogram
- **Parametric or non-parametric** methods based on normality of data
- Hence plotting **correct** histograms is important
- In clinical research, often outcome variable is ‘*change from baseline (CFB)*’

Example

- A hypothetical situation where a dietary supplement is being tested for its effectiveness in reducing weight
- Data collected for 40 subjects, subject number and Change from baseline values were recorded
- Change from baseline was calculated as ‘Post-baseline weight – Baseline weight’
- Data were of the following form:

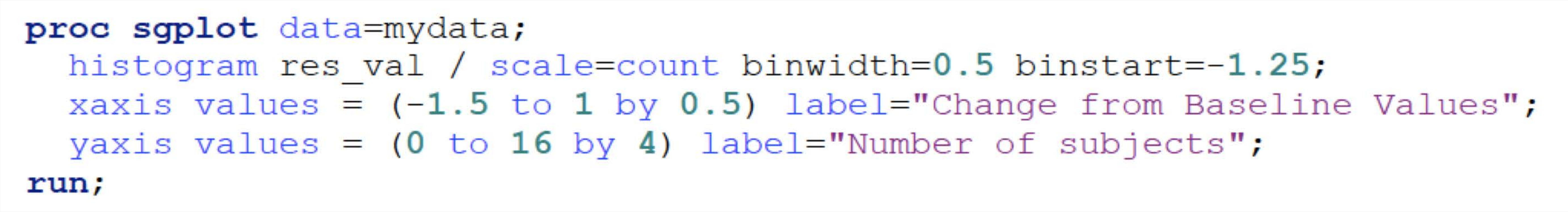


VIEWTABLE: Work.Mydata

	subject	res_val
1	1	-1
2	2	-0.4
3	3	-0.7
4	4	0.5
5	5	-0.3
6	6	-0.5
7	7	-0.5
8	8	-1
9	9	-0.4
10	10	-0.5
11	11	0.5

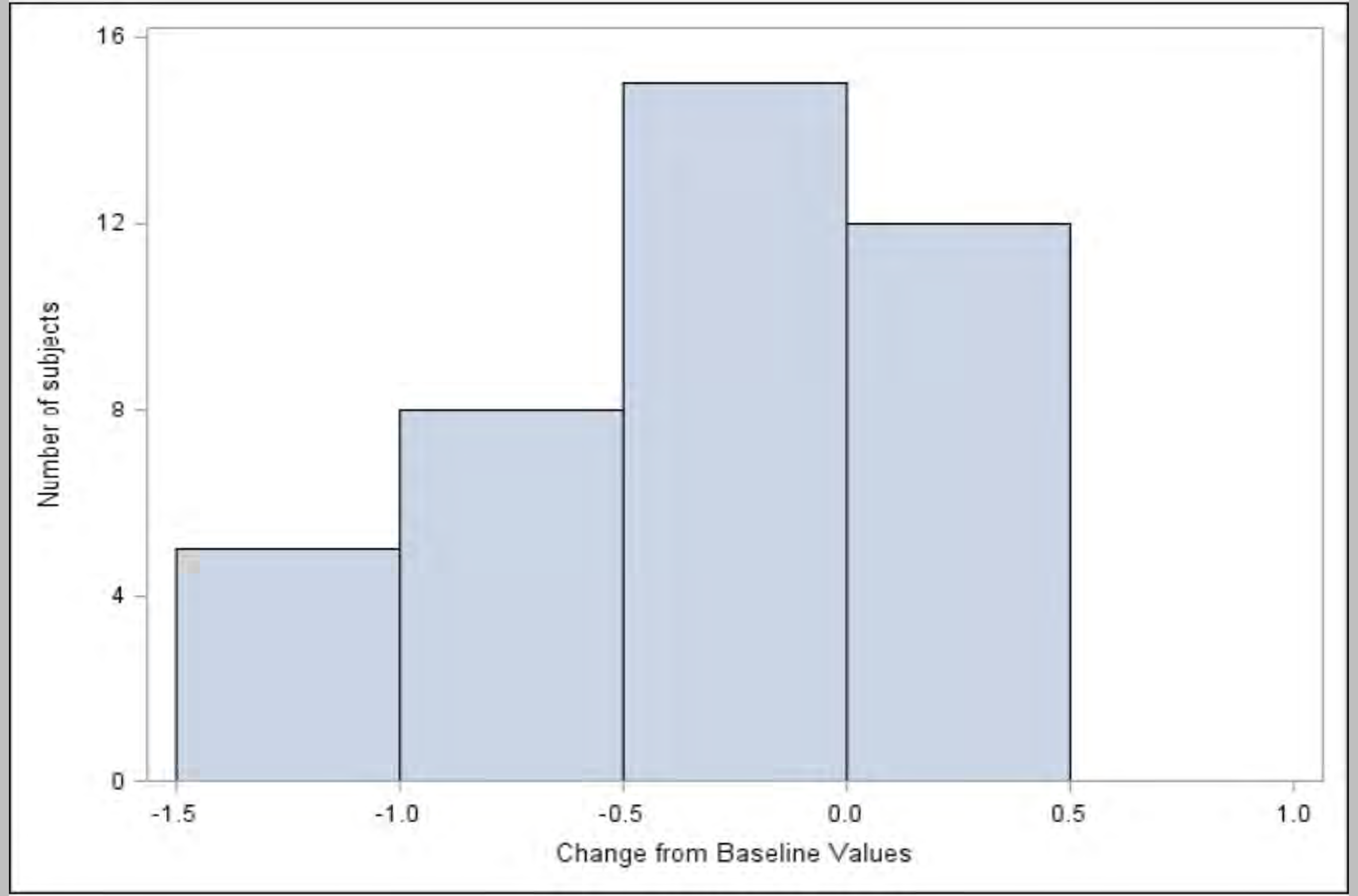
METHODS AND RESULTS

- **SGPLOT** procedure used



```
proc sgplot data=mydata;
  histogram res_val / scale=count binwidth=0.5 binstart=-1.25;
  xaxis values = (-1.5 to 1 by 0.5) label="Change from Baseline Values";
  yaxis values = (0 to 16 by 4) label="Number of subjects";
run;
```

- The above code results in the following **incorrect** histogram:



- Note that the **largest** *CFB* value in the dataset is **0.5** which is also the **last tick-point** specified in *xaxis* statement of the SGPLOT procedure

The Issue

- **5 subjects** with CFB value of **0.5** are included in the **incorrect** bin **(0.0, 0.5)**
- These 5 subjects should actually be included in the bin **(0.5, 1.0)** which is **missing** in the histogram

An unusual remedy using the usual “nbins” option to rectify anomalous histograms in SAS

Rachana Lele

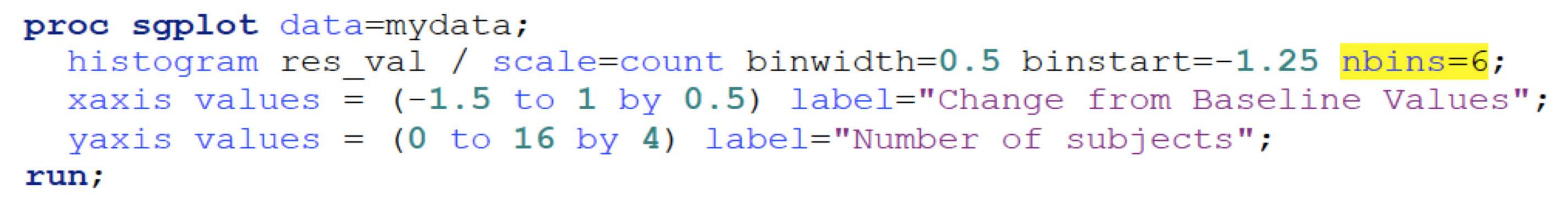
Graduate Student, MS Biostatistics, University of Louisville, KY

rachanak.lele@louisville.edu

METHODS AND RESULTS (continued)

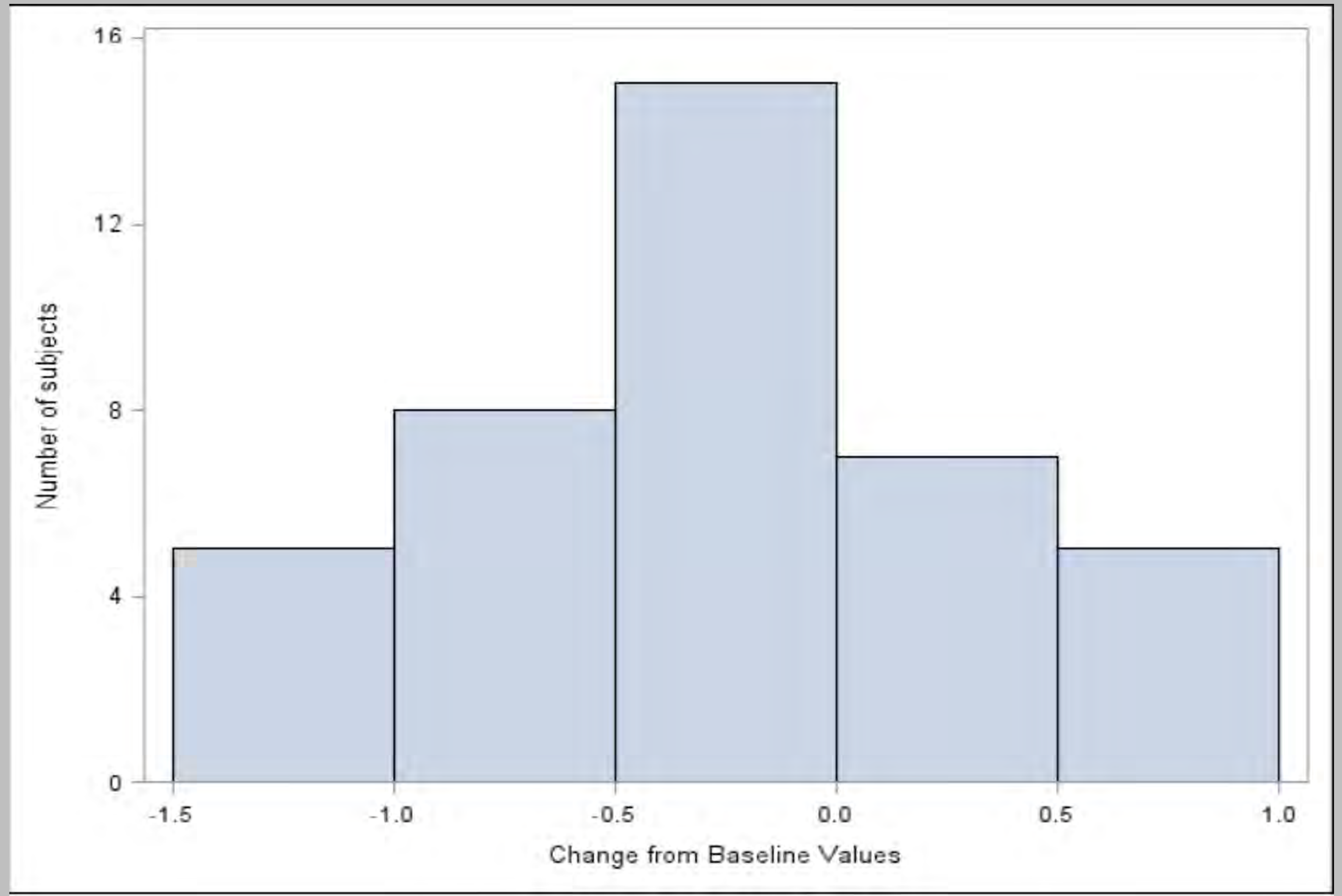
The Solution

- Include ***nbins*** option in the histogram statement of the SGPLOT procedure
- the ***nbins*** specified should be **greater than or equal** to the tick points on the x-axis of the histogram



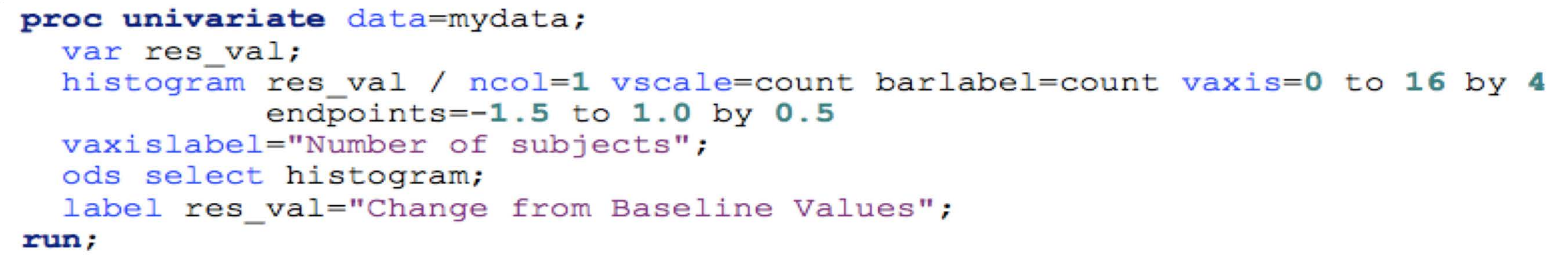
```
proc sgplot data=mydata;
  histogram res_val / scale=count binwidth=0.5 binstart=-1.25 nbins=6;
  xaxis values = (-1.5 to 1 by 0.5) label="Change from Baseline Values";
  yaxis values = (0 to 16 by 4) label="Number of subjects";
run;
```

- The above code results in the following **correct** histogram:



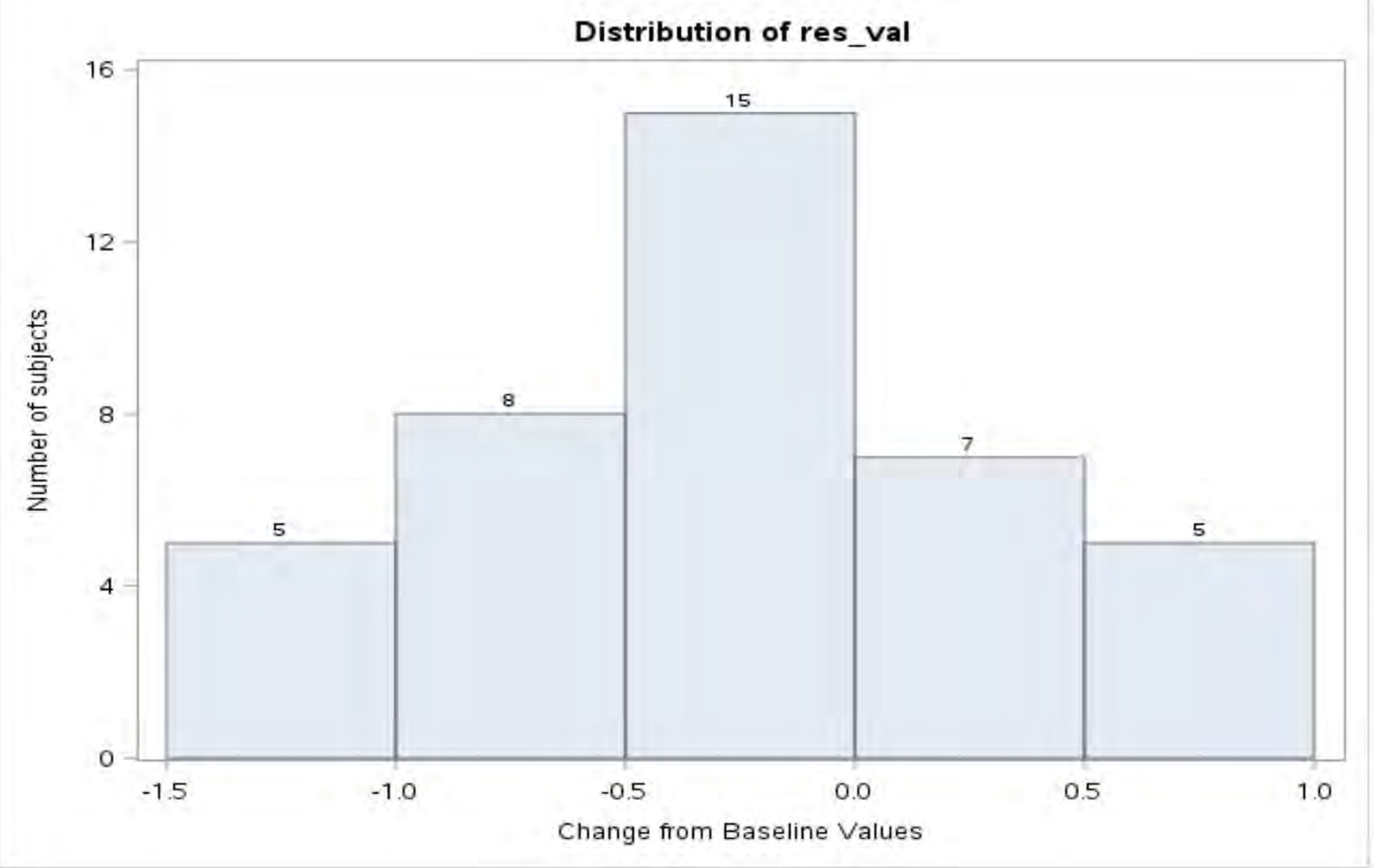
ALTERNATIVE WAYS

To produce histograms:



```
proc univariate data=mydata;
  var res_val;
  histogram res_val / ncol=1 vscale=count barlabel=count vaxis=0 to 16 by 4
            endpoints=-1.5 to 1.0 by 0.5
  vaxislabel="Number of subjects";
  ods select histogram;
  label res_val="Change from Baseline Values";
run;
```

- The above code results in the following **correct** histogram:



To calculate binstart in SGPLOT procedure:

- a, b and c are defined as follows: *xaxis values = (a to b by c)*
- For calculating binstart:

  Usual way: ***a + (c/2)***

  Alternative way: *a - (c/2)*
- This results in the same correct histogram as the one produced using *nbins* option

An unusual remedy using the usual “nbins” option to rectify anomalous histograms in SAS

Rachana Lele

Graduate Student, MS Biostatistics, University of Louisville, KY

rachanak.lele@louisville.edu

CONCLUSIONS

nbins option in SGPLOT procedure

- The SGPLOT procedure **without the *nbins* option** produces a histogram with **one bin less** than the expected number of bins when the **largest observation** in the data matches **any tick point** on the x-axis of the histogram.
- Hence, using the ***nbins* option** in the histogram statement of the SGPLOT procedure along with the *binstart* and *binwidth* options helps to create the histogram with **appropriate number of bins**.
- The value specified with the nbins option should be **greater than or equal to the number of tick points** specified on the x-axis of the histogram.

UNIVARIATE procedure

- An alternative way to **create a histogram** with the **expected and correct number of bins** is to use the **UNIVARIATE** procedure.
- This procedure can also be useful for **validating the histograms** produced using the **SGPLOT** procedure.

binstart option in SGPLOT procedure

- The ***binstart* option** in the histogram statement of the SGPLOT procedure can alternatively be calculated as ***a-(c/2)*** and would produce the same histogram which is produced when the binstart is calculated as *a+(c/2)*, which is the usual method.

NOTES

- All the discussion in the previous sections of this paper majorly focuses on the **SGPLOT procedure** in SAS®.
- However, it must be noted that the *nbins* and *binstart* options described for the histogram statement in the SGPLOT procedure work the same way in the **histogram statement** of the **SGPANEL procedure**.
- Also, the histograms produced by the **SGPANEL** procedure can be **validated** using those produced by the **UNIVARIATE** procedure.
- Thus, all the methods, results, discussions and recommendations given in this work can be extended to the **SGPANEL procedure**.

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF

Paper 2589-2018

An Unusual Remedy Using the Usual NBINS Option to Rectify Anomalous Histograms in SAS®

Rachana Lele, Department of Bioinformatics and Biostatistics, University of Louisville, KY

ABSTRACT

Data visualization is a strong tool for understanding the nature and distribution of collected data. A histogram is one such data visualization tool that can be used to assess normality of the data. Hence, plotting the correct histograms is important since the decision regarding additional analytical methods (parametric or nonparametric) is based on whether the data follows normal distribution. In SAS®, different procedures, such as SGPLOT, SGPANEL, or UNIVARIATE, can be used to generate histograms. However, histograms plotted in SAS using the SGPLOT or SGPANEL procedures show an anomaly when the largest value in a set of data coincides with one of the tick points on the X axis of the histogram. This paper discusses this anomaly and suggests a remedy for solving it. This paper also suggests that the UNIVARIATE procedure can be used to validate the histograms produced using the SGPLOT and SGPANEL procedures. Furthermore, an alternative method for calculating the value to be specified for the BINSTART option in the SGPLOT and SGPANEL procedures that does not alter the histogram produced by the usual method is also suggested. All of the procedures described above were performed using SAS® 9.3.

INTRODUCTION

In any kind of research that involves statistical analysis of data, it is generally recommended to check whether the data follows normal distribution and to decide any further analytical procedures based on the results of the normality check. A histogram is a visual tool which helps in understanding the true nature of the data. A symmetric bell-shaped histogram indicates that the data are fairly normal and any kind of skew (positive or negative) indicates that the data may not be normally distributed. The normality test is important for deciding whether parametric or non-parametric testing should be carried out. Hence, if normality of the data is being predicted on the basis of histograms, getting the correct histograms is an important step in the analysis.

In SAS®, histograms can be produced using procedures such as SGPLOT/ SGPANEL or UNIVARIATE. This paper discusses a unique issue which arises when histograms are plotted using SGPLOT/ SGPANEL procedures in SAS® for a set of data in which the largest value coincides with one of the tick points on the X axis of the histogram. It suggests usage of a simple and usual option NBINS to solve this issue. It also suggests validating histograms produced by the SGPLOT/ SGPANEL procedures using those produced by the UNIVARIATE procedure. Furthermore, it has also been suggested that if the BINSTART in the SGPLOT/ SGPANEL procedures is calculated in an alternative way, it does not alter the histogram which is produced using the usual method for calculating BINSTART.

BACKGROUND

In clinical research, ‘change from baseline’ is a common response variable which is used in the statistical analysis to assess treatment effect. This variable is calculated by subtracting the baseline values for a particular variable such as the weight or blood pressure, from the values obtained at a post baseline time point.

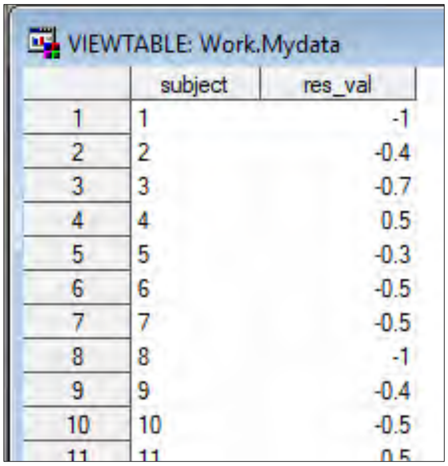
In this paper a hypothetical case study where a dietary supplement is being tested for its effectiveness in reducing weight was considered. In this hypothetical case study, data were collected for 40 subjects and their weights (in kg) before and after taking the dietary supplement were reported up to 1 decimal place. Subsequently, change from baseline was calculated for each subject and a dataset named *‘mydata’* was created using the collected data.

In the dataset *mydata*, the subject number is captured in the variable named *‘subject’* and the change from baseline values are captured in the variable named *‘res_val’*.

The dataset ‘*mydata*’ was created using the following SAS® code:

```
DATA mydata;
  input subject $ res_val;
  datalines;
  1 -1
  2 -0.4
  3 -0.7
  4 0.5
  5 -0.3
  6 -0.5
  7 -0.5
  8 -1
  9 -0.4
  10 -0.5
  11 0.5
  12 -0.6
  13 -0.7
  14 -1.4
  15 -0.1
  16 -0.3
  17 -0.1
  18 -0.1
  19 -0.7
  20 -0.9
  21 -0.3
  22 0.5
  23 0.4
  24 -0.2
  25 -0.8
  26 -0.4
  27 0.3
  28 0.4
  29 0.5
  30 -1.3
  31 0
  32 -0.5
  33 0.4
  34 -1.5
  35 -0.1
  36 0.5
  37 -1.5
  38 -1.4
  39 0
  40 0.2
  ;
RUN;
```

The created SAS® dataset is shown in **Figure 1**:



VIEWTABLE: Work.Mydata

	subject	res_val
1	1	-1
2	2	-0.4
3	3	-0.7
4	4	0.5
5	5	-0.3
6	6	-0.5
7	7	-0.5
8	8	-1
9	9	-0.4
10	10	-0.5
11	11	0.5

Figure 1. Dataset *'mydata'*

In order to decide whether parametric or non-parametric methods should be used to analyze the change from baseline values, normality of the data was assessed using a histogram. A histogram for these data were plotted using the SGPLOT procedure in SAS®. The following section discusses the procedure in greater detail.

METHODS

The following SAS® code was used to create the required histogram:

```
PROC SGPLOT data=mydata;
  histogram res_val / scale=count binwidth=0.5 binstart=-1.25;
  xaxis values = (-1.5 to 1 by 0.5) label="Change from Baseline Values";
  yaxis values = (0 to 16 by 4) label="Number of subjects";
RUN;
```

The following histogram (**Figure 2**) was created after running the above SAS® code:

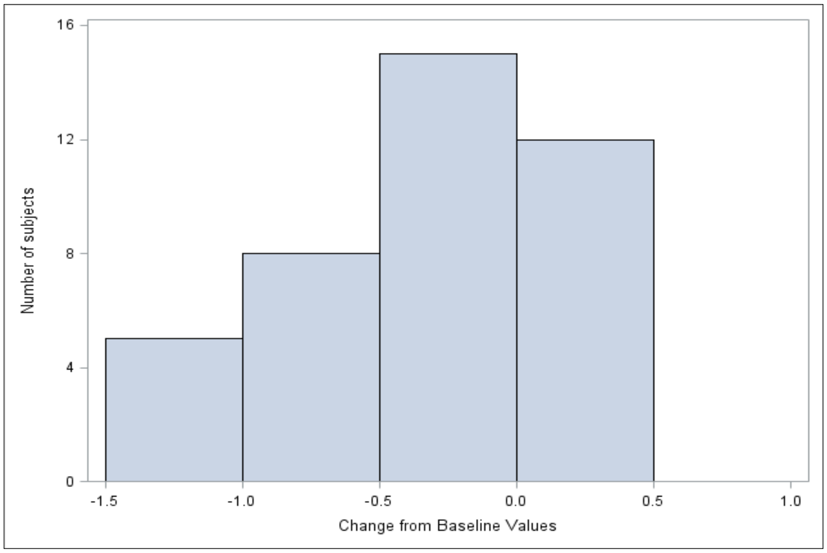


Figure 2. Histogram using PROC SGPLOT

THE ISSUE

Some of the change from baseline values in the dataset *mydata* coincide with the tick points on the X axis of the histogram. The points which coincide are as follows: -1.5, -1.0, -0.5, 0.0, 0.5. From the histogram, we observe that we do not get a bin which starts at 0.5 and ends at 1 even when the X axis statement in the SGPLOT procedure specifies the last tick point as 1.0. In the dataset '*mydata*', there are five subjects for whom the change from baseline value equals 0.5. From the theory of histograms, we know that the lower limit of a bin is included in the bin and the upper limit is not included in the bin. Consequently, the total number of points corresponding to the data value 0.5 should have been plotted in the bin starting at 0.5 and ending at 1.0. However, after careful observation of the histogram, we may note that the data value 0.5 gets incorrectly plotted in the bin (0.0, 0.5) instead of the bin (0.5, 1.0).
This will lead us to incorrectly conclude that the data does not follow normal distribution.

THE SOLUTION

To correct the anomaly observed in the histogram above, the simple NBINS option can be specified in the histogram statement of the SGPLOT procedure. The NBINS option generally has a different functionality in the SGPLOT procedure. It is used only to specify the number of bins. However, it was observed that when the NBINS option is specified in the histogram statement of the SGPLOT procedure, the largest data value 0.5 gets plotted in the correct bin.
The following SAS® code, thus, results in the correct histogram:

```
PROC SGPLOT data=mydata;
  histogram res_val / scale=count binwidth=0.5 binstart=-1.25 nbins=6;
  xaxis values = (-1.5 to 1 by 0.5) label="Change from Baseline Values";
  yaxis values = (0 to 16 by 4) label="Number of subjects";
RUN;
```

It must be noted that the only difference in the previous code and the new code is the inclusion of NBINS=6 in the histogram statement and that the NBINS specified should be greater than or equal to the tick points on the X axis of the histogram. If the value specified in the NBINS option is less than the number of tick points on the X axis of the histogram (in this case, 6), it results in the same incorrect histogram as produced by the SAS® code which does not include the NBINS option.
Inclusion of the NBINS option results in the following correct histogram as shown in **Figure 3**:

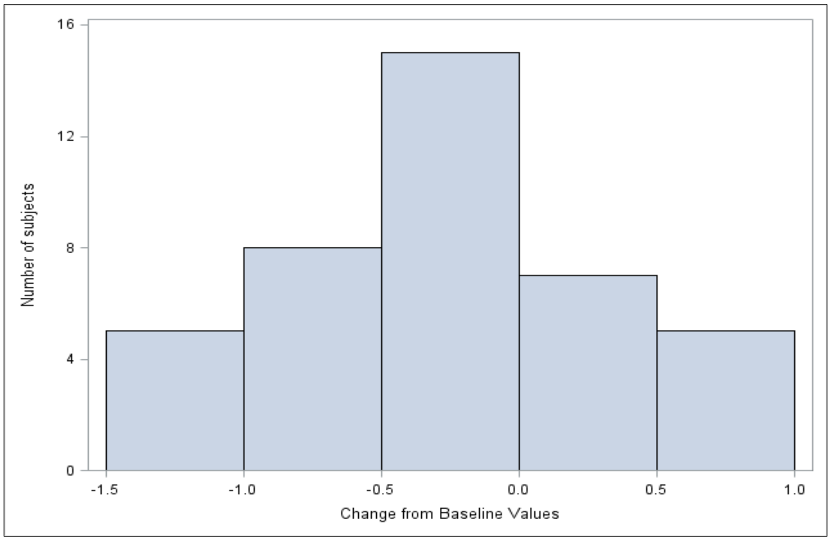


Figure 3. Histogram using PROC SGPLOT and NBINS option

We can observe from the histogram in **Figure 3**, that the data follows normal distribution.

ALTERNATIVE WAY

Different procedures can be used to produce a histogram in SAS®. One such procedure is the UNIVARIATE procedure. The following SAS® code can be used to produce a histogram using the UNIVARIATE procedure:

```
PROC UNIVARIATE data=mydata;
  var res_val;
  histogram res_val / ncol=1 vscale=count barlabel=count vaxis=0 to 16 by 4
            endpoints=-1.5 to 1.0 by 0.5
  vaxislabel="Number of subjects";
  ods select histogram;
  label res_val="Change from Baseline Values";
RUN;
```

Following is the histogram produced by the UNIVARIATE procedure as shown in **Figure 4**:

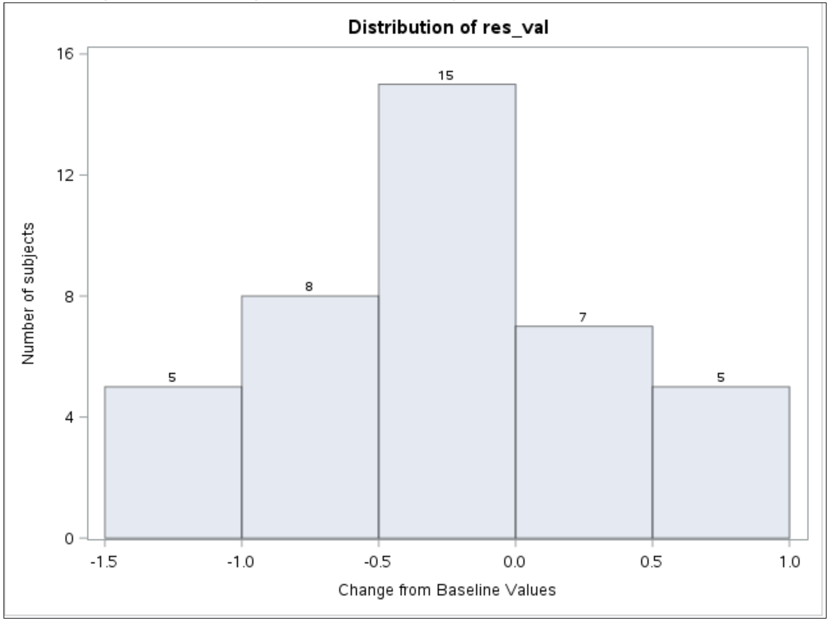


Figure 4. Histogram using PROC UNIVARIATE

OBSERVATIONS AND RECOMMENDATIONS

It can be observed from the above histogram (**Figure 4**) that five data points corresponding to the data value 0.5 have been correctly plotted in the bin starting at 0.5 and ending at 1.0. Hence, the UNIVARIATE procedure can be alternatively used to create the required histogram. Also, validation of all outputs such as tables, listings and figures with the help of double programming is a common procedure in clinical research. Thus, the UNIVARIATE procedure can serve as an excellent tool for validating histograms produced by the SGPLOT procedure and any issues in the histograms produced using the SGPLOT procedure could easily be checked by comparing them with the histograms produced using the UNIVARIATE procedure.

ALTERNATIVE WAY FOR CALCULATING BINSTART

Generally, while creating histograms, the value specified in the BINSTART option in the histogram statement of the SGPLOT procedure in SAS® is calculated as *a+(c/2)*; where a and c are defined as follows: X axis values = (a to b by c). However, it was observed that even if the BINSTART is calculated as *a-(c/2)*, the resulting histogram is the same as shown in **Figure 4a** and **Figure 4b**.

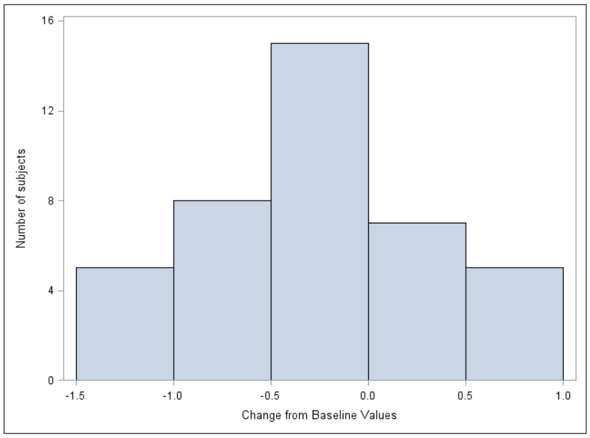


Figure 4a. Histogram using BINSTART = a+(c/2)

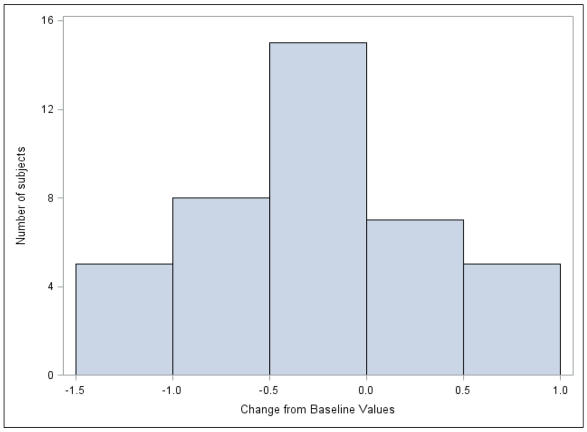


Figure 4b. Histogram using BINSTART = a-(c/2)

We can clearly observe that the histograms in **Figure 4a** and **Figure 4b** are same regardless of whether the BINSTART was calculated as *a+(c/2)* or *a-(c/2)*.

CONCLUSION

The SGPLOT procedure without the NBINS option produces a histogram with one bin less than the expected number of bins when the largest observation in the data matches any tick point on the X axis of the histogram. Hence, using the NBINS option in the histogram statement of the SGPLOT procedure along with the BINSTART and BINWIDTH options helps to create the histogram with appropriate number of bins. The value specified with the NBINS option should be greater than or equal to the number of tick points specified on the X axis of the histogram.
An alternative way to create a histogram with the expected and correct number of bins is to use the UNIVARIATE procedure. This procedure can also be useful for validating the histograms produced using the SGPANEL procedure.
The BINSTART option in the histogram statement of the SGPLOT procedure can alternatively be calculated as *a-(c/2)* and would produce the same histogram which is produced when the BINSTART is calculated as *a+(c/2)*, which is the usual method.

All the discussion in the previous sections of this paper majorly focuses on the SGPLOT procedure in SAS®. However, it must be noted that the NBINS and BINSTART options described for the histogram statement in the SGPLOT procedure work the same way in the histogram statement of the SGPANEL procedure. Also, the histograms produced by the SGPANEL procedure can be validated using those produced by the UNIVARIATE procedure. Thus, all the methods, results, discussions and recommendations given in this work can be extended to the SGPANEL procedure.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:

Rachana Lele
Graduate Student, Biostatistics,
Department of Bioinformatics and Biostatistics,
University of Louisville, KY
rachanak.lele@louisville.edu