STAT COMPUTING
BANA 6043

Lecture 3

Statistical graphics, tests and models

You should know…

- How to combine data sets (set and merge statements)
- How to use built-in functions (mathematics, probability and statistics functions)
- LABEL and TITLE statements
- PROC UNIVARIATE and MEANS (descriptive statistics, normality tests, boxplots, histograms and Q-Q plots)
- PROC FREQ (tables and tests for categorical data)
- How to perform simple simulations (DO loops and random observation generators)

Today's Topics

- PROC CHART
- X-Y Plot (PROC PLOT)
- One-Sample T-test (PROC MEANS)
- Two-Sample T-test (PROC TTEST)
- One-Way ANOVA (PROC ANOVA & PROC GLM)
- Two-Way ANOVA
- Regression analysis (PROC REG)
- Model checking

Data set for future use

Step 1. Create a dataset "simulation" by simulating 200 observations from the following linear model:

Y = alpha + beta1 * X1 + beta2 * X2 + noise

where

- alpha=1, beta1=2, beta2=-1.5
- X1 ~ N(1, 4), X2 ~ N(3,1), noise ~ N(0,1)

Step 2. Define a new binary variable Y_bin such that Y_bin=1 if Y>0 and Y_bin=0 otherwise.

Step 3. Make the final data contain only 4 variables: X1, X2, Y and Y_bin.

```
data simulation;
  do i=1 to 200;
   x1=1+2*rannor(12); x2=3+rannor(12); error=rannor(12);
   alpha=1; beta_1=2; beta_2=-1.5;
   y=alpha+beta_1*x1+beta_2*x2+error;
   if y>0 then y_bin=1;
   else y_bin=0;
   output;
  end;
  keep x1 x2 y y_bin;
run;

proc print data=simulation;
 title Simulated Data Set;
run;
```

PROC CHART

- PROC CHART is used to produce histograms.
  - The histograms are more sophisticated and informative than the ones produced from PROC UNIVARIATE.

- **General Form**

PROC CHART data=dataset;
    VBAR *variables* < / options >;
    HBAR *variables* < / options >;

VBAR produces a vertical bar chart.
HBAR produces a horizontal bar chart.

Compare the two sets of code

- Set A

```
proc chart data=simulation;
 vbar x1;
run;
```

- Set B

```
proc chart data=simulation;
 hbar x1;
run;
```

```
Frequency

   |                                *****
40 +                                *****
   |                                *****
   |                                *****    *****
   |                                *****    *****
   |                                *****    *****
35 +                                *****    *****
   |                                *****    *****    *****
   |                                *****    *****    *****
   |                                *****    *****    *****
   |                       *****    *****    *****    *****
30 +                       *****    *****    *****    *****
   |                       *****    *****    *****    *****
   |                       *****    *****    *****    *****
   |                       *****    *****    *****    *****
   |                       *****    *****    *****    *****
25 +                       *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
20 +                       *****    *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
   |                       *****    *****    *****    *****    *****
15 +              *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****
10 +              *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****    *****
   |              *****    *****    *****    *****    *****    *****    *****
 5 +     *****    *****    *****    *****    *****    *****    *****    *****    *****
   |     *****    *****    *****    *****    *****    *****    *****    *****    *****
   |     *****    *****    *****    *****    *****    *****    *****    *****    *****
   |     *****    *****    *****    *****    *****    *****    *****    *****    *****
   |     *****    *****    *****    *****    *****    *****    *****    *****    *****
   --------------------------------------------------------------------------------------
          -3.6     -2.4     -1.2      0.0      1.2      2.4      3.6      4.8      6.0

                                        x1 Midpoint
```

```
    x1                                                                    Cum.             Cum.
Midpoint                                                           Freq   Freq   Percent   Percent

  -3.6   *****                                                        5      5      2.50      2.50
  -2.4   ***************                                             15     20      7.50     10.00
  -1.2   *******************************                             31     51     15.50     25.50
   0.0   *****************************************                   41     92     20.50     46.00
   1.2   **************************************                      38    130     19.00     65.00
   2.4   **********************************                          34    164     17.00     82.00
   3.6   ************************                                    24    188     12.00     94.00
   4.8   *******                                                      7    195      3.50     97.50
   6.0   *****                                                        5    200      2.50    100.00

             5    10    15    20    25    30    35    40

                       Frequency
```

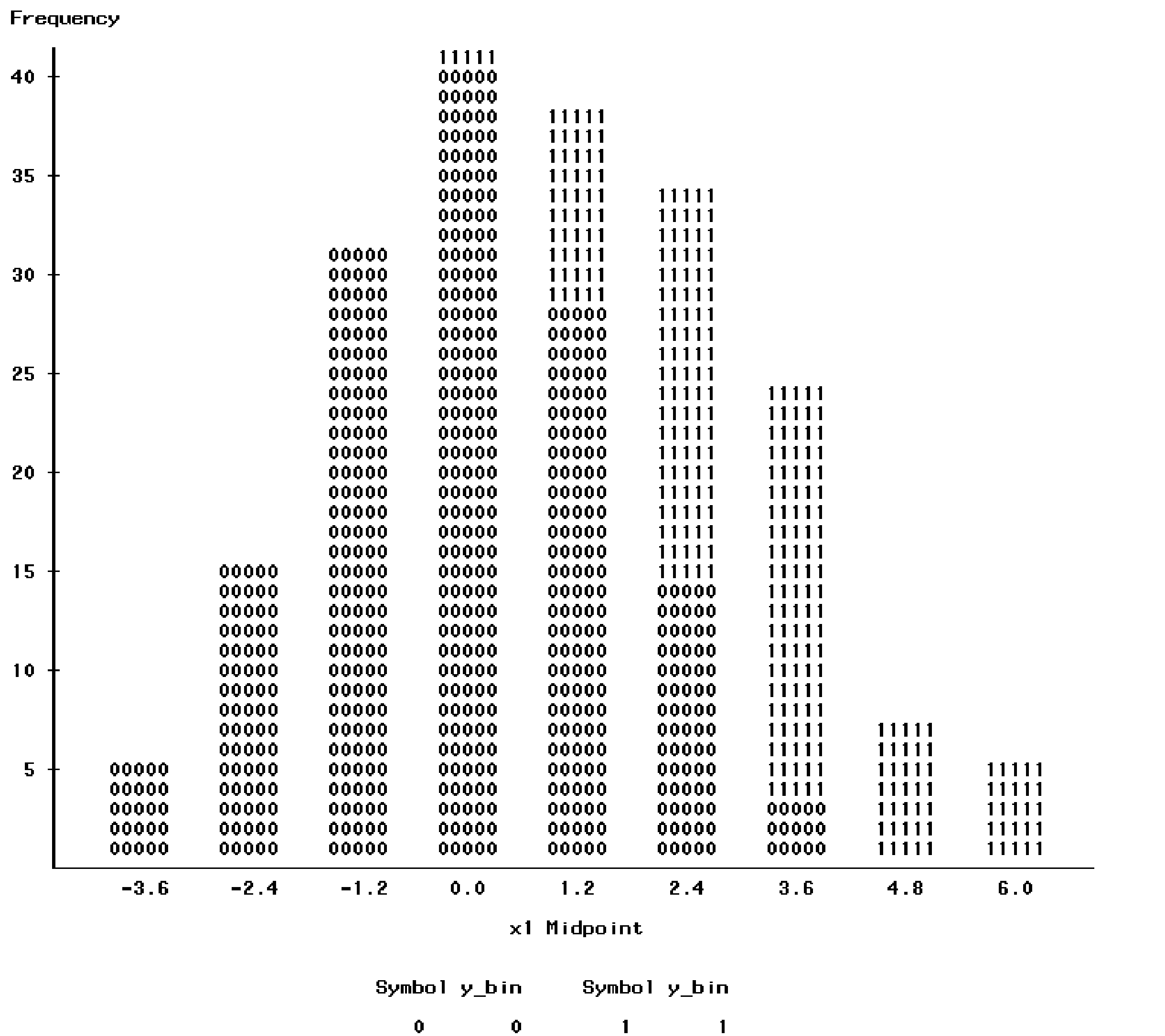
Compare the two sets of code

- Set A

```
proc chart data=simulation;
 vbar x1;
run;
```

- Set B

```
proc chart data=simulation;
 vbar x1 / subgroup = y_bin;
run;
```


Frequency
40
35
30
25
20
15
10
5
-3.6
-2.4
-1.2
0.0
1.2
2.4
3.6
4.8
6.0
x1 Midpoint
Symbol y_bin Symbol y_bin
0 0 1 1

Options for PROC CHART

SUBGROUP = variable -- The bar can be divided into parts representing the values of the specified variable. The first character of variable is used.

LEVELS = # of midpoints -- Specify the number of bars on the chart. SAS will automatically choose the number of the bars unless you specify it.

MIDPOINTS =list -- specify the values for the midpoints. Valid format for MIDPOINTS option is as the following:

MIDPOINTS=10 20 30

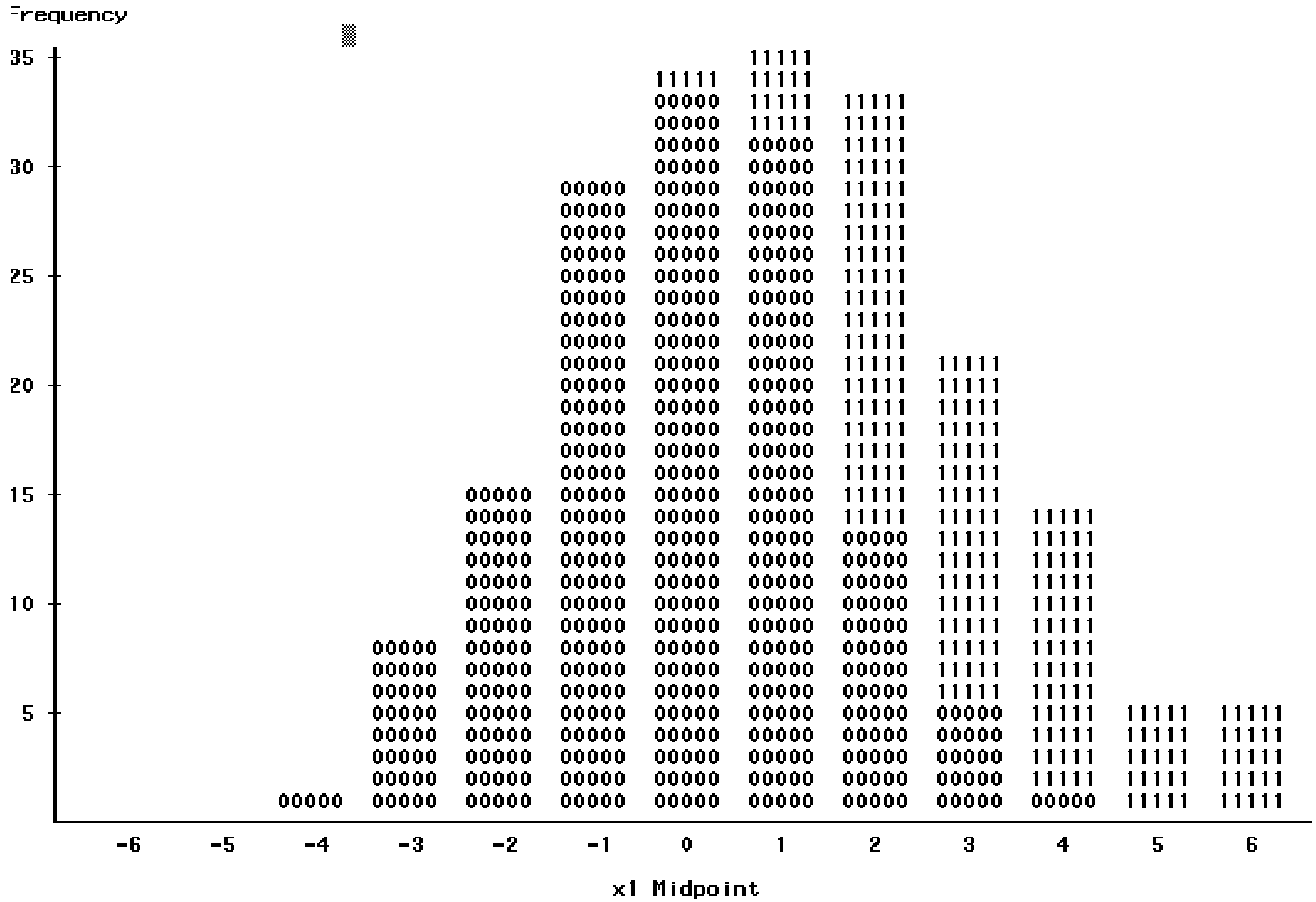
MIDPOINTS=10 TO 100 BY 10

NOTE: DO not use LEVELS and MIDPOINTS options simultaneously.

Try these

```
proc chart data=simulation;
  vbar x1 / subgroup=y_bin midpoints= -6 to 6 by 1;
run;
```

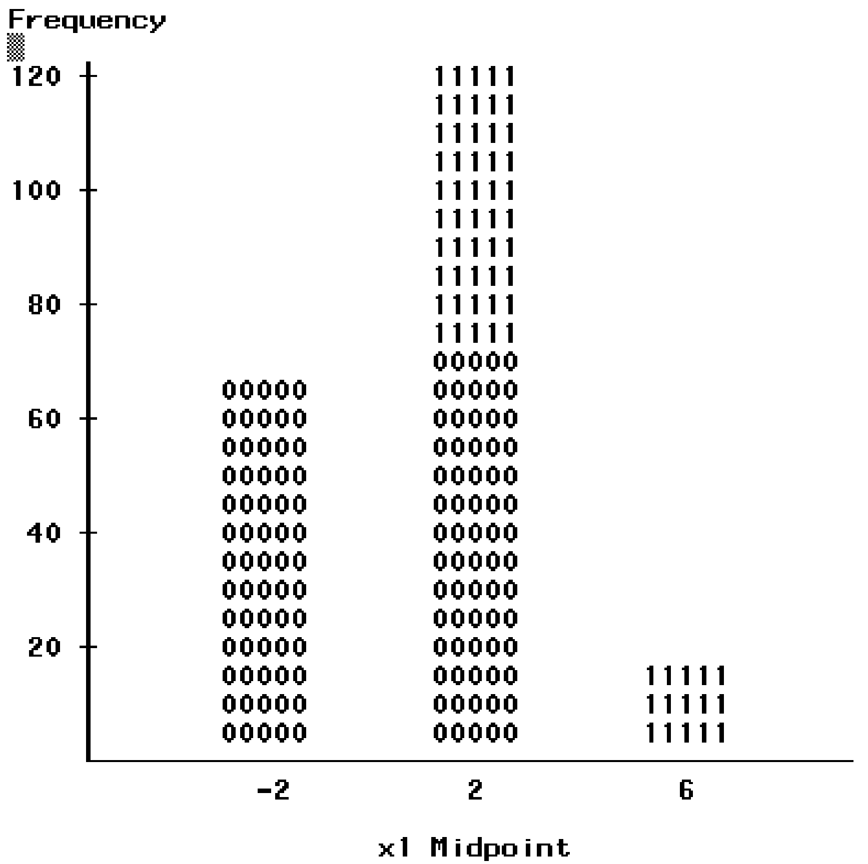
```
proc chart data=simulation;
  vbar x1 / subgroup=y_bin levels=3;
run;
```



Symbol	y_bin	Symbol	y_bin
0	0	1	1

Simulated Data Set

16:27 Satur



Symbol	y_bin	Symbol	y_bin
0	0	1	1

Exercise 1

<u>Step 1</u>. Create a data set "clinical_trial" containing observations of two random variables X and Y. For the first 200 observations, let X = "placebo" and Y ~ $N(0,1)$. For the last 200 observations, let X = "drug" and Y ~ $N(1,1)$. <u>The random variable Y can be thought of as certain "effectiveness score".</u>

<u>Step 2</u>. Compare the distribution of Y in the two groups where X = "placebo" and X = "drug" .

Code for generating the data

```
DATA clinical_trial;
        DO K = 1 TO 200;
                X = "placebo";
                Y = RANNOR(23);
                OUTPUT;
        END;
        DO K = 1 TO 200;
                X = "drug";
                Y = 1+RANNOR(23);
                OUTPUT;
        END;
DROP K;
RUN;
```

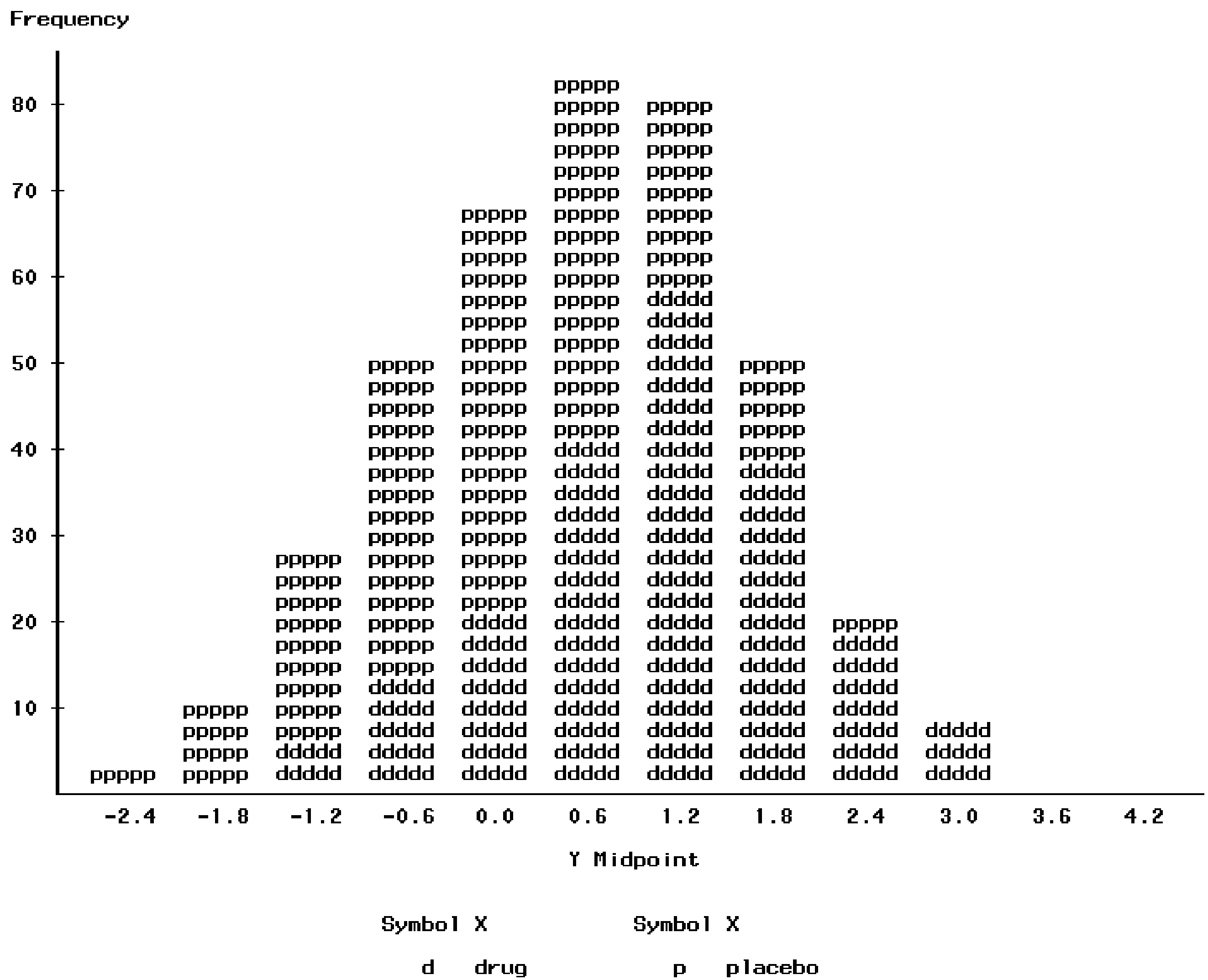
Code for comparing the distributions of Y in the two groups where X = “placebo” and X = “drug”

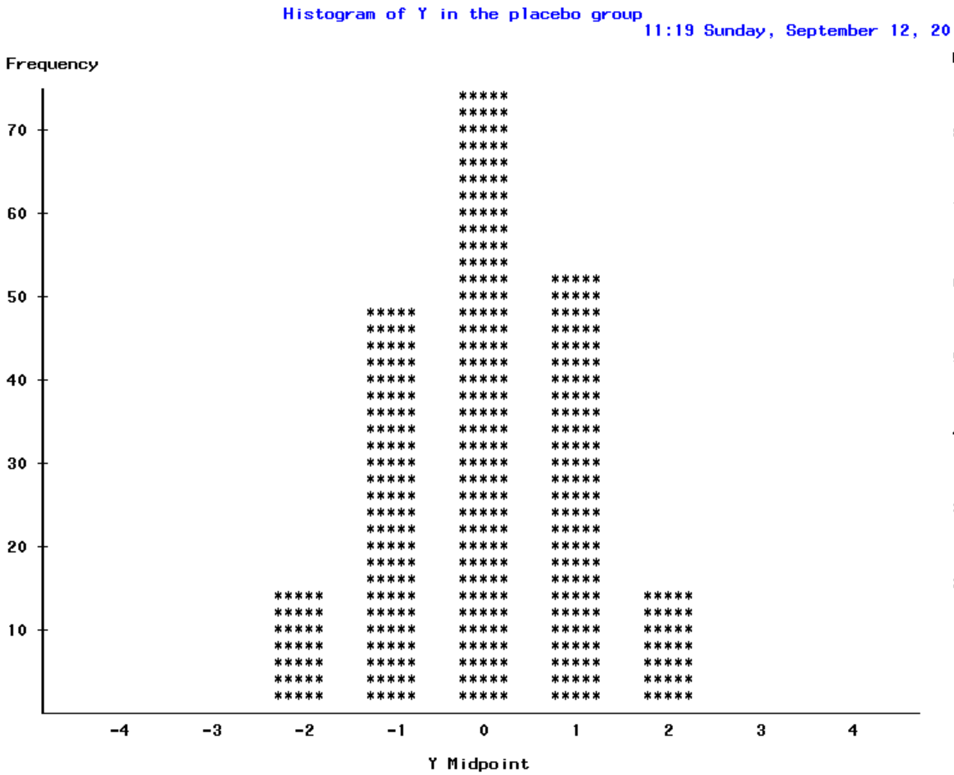
GRAPHICAL COMPARISON

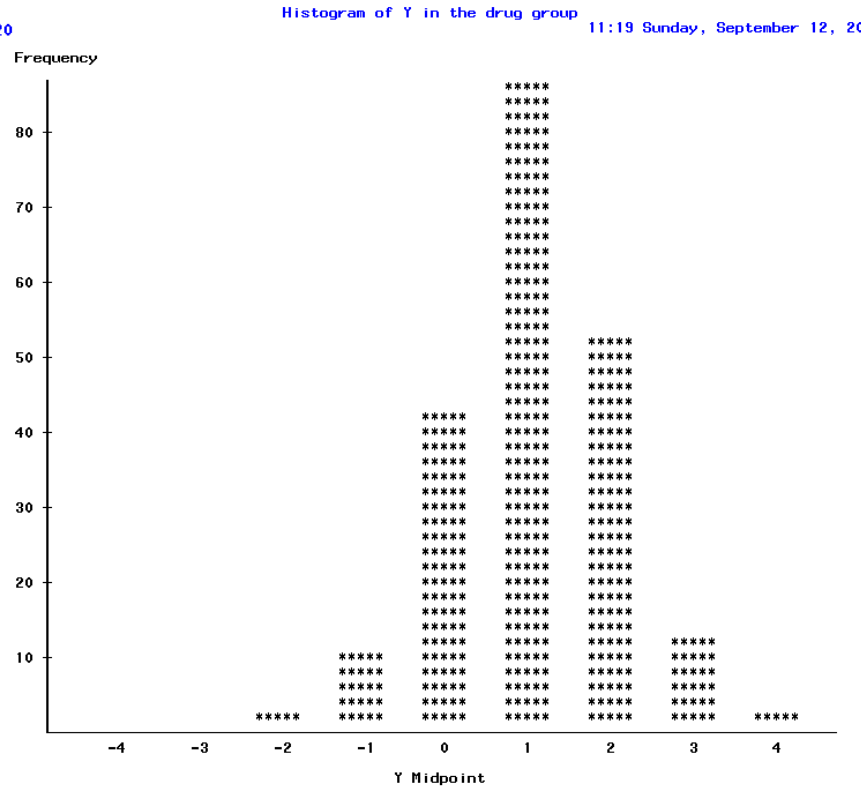
```
PROC CHART DATA=clinical_trial;
 VBAR Y / SUBGROUP = X; title Histogram of Y by groups;
RUN;

DATA PLACEBO;
 SET CLINICAL_TRIAL;  IF X='placebo';
RUN;
PROC CHART DATA=PLACEBO;
 VBAR Y / midpoints=-4 to 4 by 1; title Histogram of Y in the placebo group;
RUN;

DATA DRUG;
 SET CLINICAL_TRIAL;  IF X='drug';
RUN;
PROC CHART DATA=DRUG;
 VBAR Y / midpoints=-4 to 4 by 1; title Histogram of Y in the drug group;
RUN;
```

Histogram of Y by groups

11:19 Sunday, September 12, 2




Histogram of Y in the placebo group
11:19 Sunday, September 12, 20
Frequency
70
60
50
40
30
20
10
-4
-3
-2
-1
0
1
2
3
4
Y Midpoint


Histogram of Y in the drug group
11:19 Sunday, September 12, 2(
Frequency
80
70
60
50
40
30
20
10
-4
-3
-2
-1
0
1
2
3
4
Y Midpoint

What else we need to compare?

Code for comparing the distributions of Y in the two groups where X = “placebo” and X = “drug”

NUMERIC COMPARIASON

PROC MEANS DATA=PLACEBO N NMISS MEAN STD MIN MAX RANGE;
TITLE SUMMARY STATISTICS FOR PLACEBO;
RUN;

PROC MEANS DATA=DRUG N NMISS MEAN STD MIN MAX RANGE;
TITLE SUMMARY STATISTICS FOR DRUG;
RUN;

SUMMARY STATISTICS FOR PLACEBO

11:19 Sunday, September 12, 20

The MEANS Procedure

Analysis Variable : Y

N	N Miss	Mean	Std Dev	Minimum	Maximum	Range
200	0	0.0332881	0.9554329	-2.4248602	2.3204428	4.7453029

SUMMARY STATISTICS FOR DRUG

11:19 Sunday, September 12, 20

The MEANS Procedure

Analysis Variable : Y

N	N Miss	Mean	Std Dev	Minimum	Maximum	Range
200	0	1.0908419	0.9676618	-2.2896456	4.0053988	6.2950444

How **confident** can we say that the distribution of Y (effectiveness score) in the drug group is different from that in the placebo group?

We need to do **hypothesis testing**.

Two-Sample T-test

- **PROC TTEST** tests whether two means are equal or not (two-sided test).
  - It reports test results for two scenarios: when the variances in the two groups are unequal or equal.
  - It also report results for testing whether the two variances are equal for deciding which scenario we should look into.
- General Form

```
PROC TTEST data=data set;
      CLASS   variable;
      VAR     variables;
```

Syntax

<u>Example</u>

```
PROC TTEST DATA=CLINICAL_TRIAL;
 CLASS X;
 VAR Y;
 TITLE T-TEST FOR COMPARING THE MEANS OF Y IN PLACEBO AND DRUG GROUPS;
 RUN;
```

- The **CLASS** statement identifies the variable the divides the data set into two groups. The CLASS variable (e.g. X in the example) <u>must only have two values</u>, which can be either numeric or character.

Tips and Tricks

- PROC TTEST also produces an *F test* to test whether the variances from the two groups are equal or not.

- PROC TTEST, in fact, conducts two t-test under the assumptions of unequal variance or equal variance. So first check the F test to see whether the variances are equal or not. Then refer to the corresponding t-test.

The TTEST Procedure

Statistics

Variable	X	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
Y	drug	200	0.9559	1.0908	1.2258	0.8812	0.9677	1.0731	0.0684
Y	placebo	200	-0.1	0.0333	0.1665	0.8701	0.9554	1.0595	0.0676
Y	Diff (1-2)		0.8685	1.0576	1.2466	0.8992	0.9616	1.0334	0.0962

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
Y	Pooled	Equal	398	11.00	<.0001
Y	Satterthwaite	Unequal	398	11.00	<.0001

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
Y	Folded F	199	199	1.03	0.8578

http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_ttest_a0000000116.htm

Figure 92.4 Simple Statistics

The TTEST Procedure

Variable: Score

Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
f	7	76.8571	2.5448	0.9619	73.0000	80.0000
m	7	82.7143	3.1472	1.1895	78.0000	87.0000
Diff (1-2)		-5.8571	2.8619	1.5298		

Simple statistics for the two populations being compared, as well as for the difference of the means between the populations, are displayed in Figure 92.4. The Gender column indicates the population corresponding to the statistics in that row. The sample size (N), mean, standard deviation, standard error, and minimum and maximum values are displayed.

Confidence limits for means and standard deviations are shown in Figure 92.5.

Figure 92.5 Simple Statistics

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev		95% UMPU CL Std Dev	
f		76.8571	74.5036	79.2107	2.5448	1.6399	5.6039	1.5634	5.2219
m		82.7143	79.8036	85.6249	3.1472	2.0280	6.9303	1.9335	6.4579
Diff (1-2)	Pooled	-5.8571	-9.1902	-2.5241	2.8619	2.0522	4.7242	2.0019	4.5727
Diff (1-2)	Satterthwaite	-5.8571	-9.2064	-2.5078					

For the mean differences, both pooled (assuming equal variances for males and females) and Satterthwaite (assuming unequal variances) 95% intervals are shown. The confidence limits for the standard deviations are of the equal-tailed variety.

The test statistics, associated degrees of freedom, and p-values are displayed in Figure 92.6.

One-Sample T-test

- **PROC MEANS** will do a one-sample t-test for the true mean of a *normally distributed* random variable.

- By default, PROC MEANS computes the t-statistic and p-value associated with

$$H0: \mu = 0 \quad \text{vs.} \quad H1: \mu \neq 0$$

PROC MEANS data=dataset T PRT;
VAR *variables*;

Note: The options T and PRT return the t-statistic and the p-value, respectively.

Try these

- **One-sample T-test for the placebo group**

```
PROC MEANS DATA=PLACEBO T PRT;
VAR Y;
TITLE T-test for the placebo group;
RUN;
```

- **One-sample T-test for the drug group**

```
PROC MEANS DATA=DRUG T PRT;
VAR Y;
TITLE T-test for the drug group;
RUN;
```

T-test for the placebo group

The MEANS Procedure

Analysis Variable : Y

t Value	Pr > |t|
0.49	0.6228

T-test for the drug group

The MEANS Procedure

Analysis Variable : Y

t Value	Pr > |t|
15.94	<.0001

Exercise 2

What if we want to test

$$H0: \mu = 1 \quad \text{vs.} \quad H1: \mu \neq 1$$

Please Google search "SAS t test" to figure out how to do that yourself.

SAS/STAT(R) 9.2 User's Guide, Second Edition

Search this document | Search

Previous Page | Next Page

One-Sample *t* Test

A one-sample *t* test can be used to compare a sample mean to a given value. This example, taken from Huntsberger and Billingsley (1989, p. 290), tests whether the mean length of a certain type of court case is more than 80 days by using 20 randomly chosen cases. The data are read by the following DATA step:

```
data time;
   input time @@;
   datalines;
 43  90  84  87  116   95  86   99   93  92
121  71  66  98   79  102  60  112  105  98
;
run;
```

The only variable in the data set, *time*, is assumed to be normally distributed. The trailing at signs (@@) indicate that there is more than one observation on a line. The following statements invoke PROC TTEST for a one-sample *t* test:

```
ods graphics on;

proc ttest h0=80 plots(showh0) sides=u alpha=0.1;
   var time;
run;

ods graphics off;
```

The ODS GRAPHICS statement requests graphical output. The VAR statement indicates that the *time* variable is being studied, while the H0= option specifies that the mean of the *time* variable should be compared to the null value 80 rather than the default of 0. The PLOTS(SHOWH0) option requests that this null value be displayed on all relevant graphs. The SIDES=U option reflects the focus of the research question, namely whether the mean court case length is ***greater than*** 80 days, rather than ***different than*** 80 days (in which case you would use the default SIDES=2 option). The ALPHA=0.1 option requests 90% confidence intervals rather than the default 95% confidence intervals. The output is displayed in Figure 92.1.

Figure 92.1 One-Sample *t* Test Results

The TTEST Procedure

Variable: time

N	Mean	Std Dev	Std Err	Minimum	Maximum
20	89.8500	19.1456	4.2811	43.0000	121.0

proc ttest data=drug h0=**1**; var y; **run**;

T-test for the drug group

15:48 Sunday, September 12, 20

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
Y	200	0.9559	1.0908	1.2258	0.8812	0.9677	1.0731	0.0684

T-Tests

Variable	DF	t Value	Pr > |t|
Y	199	1.33	0.1858

Summary

One-sample t-test

Compare the mean of one sample to a specified fixed value (PROC MEANS OR TTEST)

Two-sample t-test

Compare the means of two samples (PROC TTEST)

Question: how to compare the means of multiple samples?

One-Way ANOVA

- One-way ANOVA (Analysis of Variation) is a simple extension of two-sample t-test.
- In the two-sample t-test, one is testing whether the means of two groups are equal.
- In one-way ANOVA, we can test whether the means of two or *more* groups are equal. **PROC ANOVA** will provide the p-value of F-test instead of t-test. The null hypothesis is all means are equal.

Example

Suppose we have three types of corn and corresponding yeilds for four years in a row.

X	Y	X	Y	X	Y
a	6	b	4	c	7
a	5	b	5	c	5
a	5	b	4	c	6
a	6	b	6	c	7

X is a variable indicating different types of corn.

We will ask: does the type matter?

PROC ANOVA

- General Form

  PROC ANOVA data = *data set*;

  CLASS *variable*;

  MODEL *response var = explanatory var*;

  MEANS *effects < / options>*;

- PROC ANOVA is valid only if the data is *balanced*. In other words, the number of observations in each group is the same. If the data is unbalance, use PROC GLM.

Code

```
DATA CORN;
    INPUT X $ Y @@;
    LABEL X=TYPE Y=YIELD;
CARDS;
a 6 b 4 c 7 a 5 b 5 c 5
a 5 b 4 c 6 a 6 b 6 c 7
;
RUN;
PROC ANOVA DATA=CORN;
  CLASS X;
  MODEL Y=X;
  MEANS X;
  TITLE ANOVA OF CORN;
RUN;
```

The ANOVA Procedure

Dependent Variable: Y   YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.50000000	2.25000000	3.12	0.0937
Error	9	6.50000000	0.72222222		
Corrected Total	11	11.00000000			

R-Square	Coeff Var	Root MSE	Y Mean
0.409091	15.45157	0.849837	5.500000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
X	2	4.50000000	2.25000000	3.12	0.0937

ANOVA OF CORN   15:48 Sunday, S

The ANOVA Procedure

Level of X	N	Y Mean	Y Std Dev
a	4	5.50000000	0.57735027
b	4	4.75000000	0.95742711
c	4	6.25000000	0.95742711

Break

Observations of two variables Y and X.

How to study the relationship between Y and X?

Examples:

1. Genetic research
2. Aviation studies
3. Pharmaceutical studies

Example

<u>Step 1</u>. Create a dataset "simulation" by simulating 200 observations from the following linear model:

$$Y = alpha + beta1 * X1 + beta2 * X2 + noise$$

where

- alpha=1, beta1=2, beta2=-1.5
- X1 ~ N(1, 4), X2 ~ N(3,1), noise ~ N(0,1)

<u>Step 2</u>. Define a new binary variable Y_bin such that Y_bin=1 if Y>0 and Y_bin=0 otherwise.

<u>Step 3</u>. Make the final data contain only 4 variables: X1, X2, Y and Y_bin.

Step 1. Do the plots!

X-Y plots

- **PROC PLOT** generates X-Y plots using a PLOT statement to specify which variable is to be used on the X-axis and which is to be on the Y-axis.

- **General Form**

```
PROC PLOT data=…;
   PLOT   yvar * xvar;
```

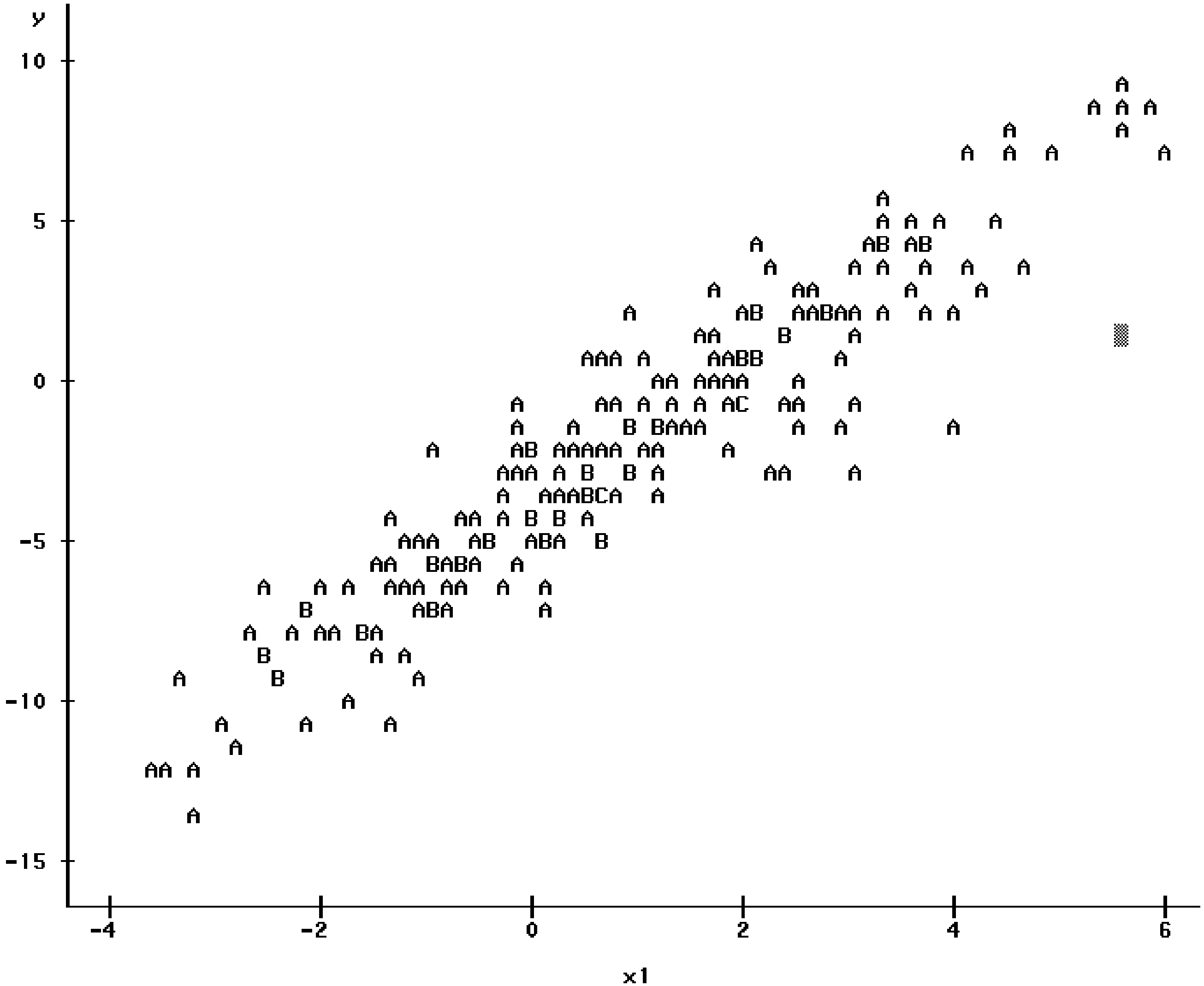
Try these

```
proc plot data=simulation;
 plot y*x1;
 plot y*x2;
 plot x1*x2;
run;
```

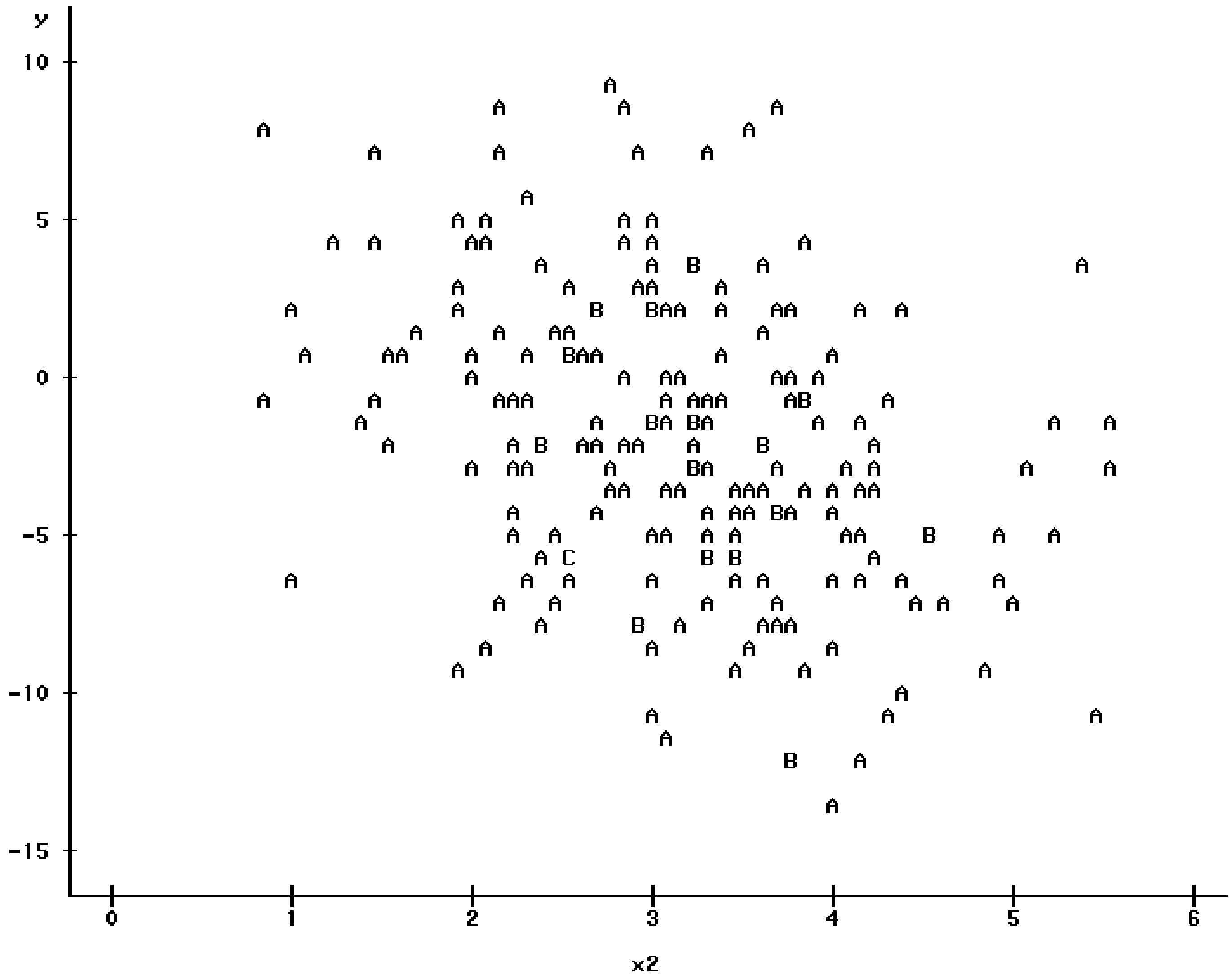
Question: what conclusion can you draw from the three plots?

Simulated Data Set   10:11 Monday, September 13, 2014

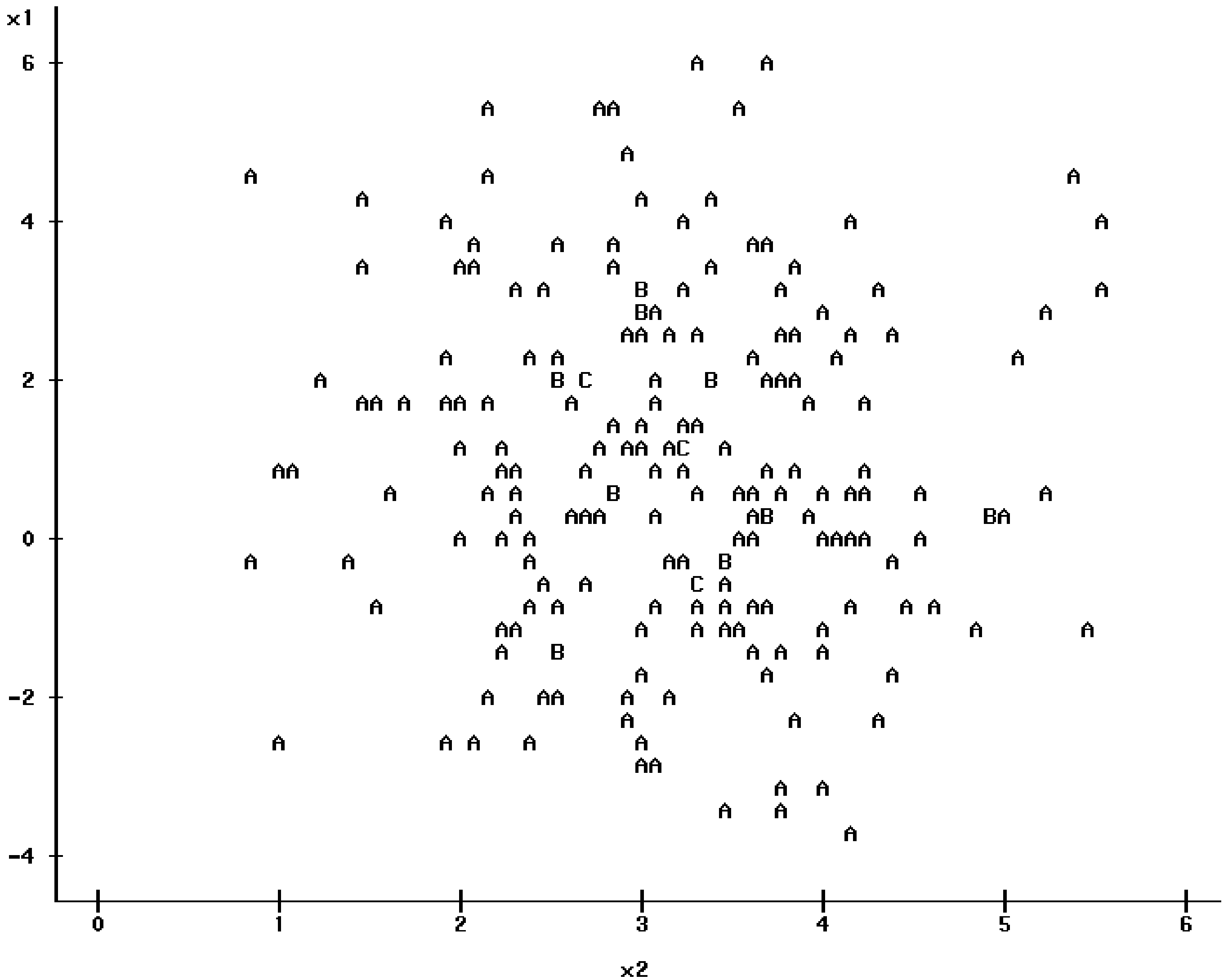
Plot of y*x1. Legend: A = 1 obs, B = 2 obs, etc.



Plot of y*x2. Legend: A = 1 obs, B = 2 obs, etc.



Plot of x1*x2. Legend: A = 1 obs, B = 2 obs, etc.



Tips and Tricks

- yvar * xvar = 'char'

  Observations are plotted using the character specified, such as '+', '*', or '.'.

- yvar1 * xvar1 = 'char1'

  yvar2 * xvar2 = 'char2'   /   OVERLAY

  Two plots yvar1*xvar1 and yvar2*xvar2 appear on the same plot. They are overlaid.

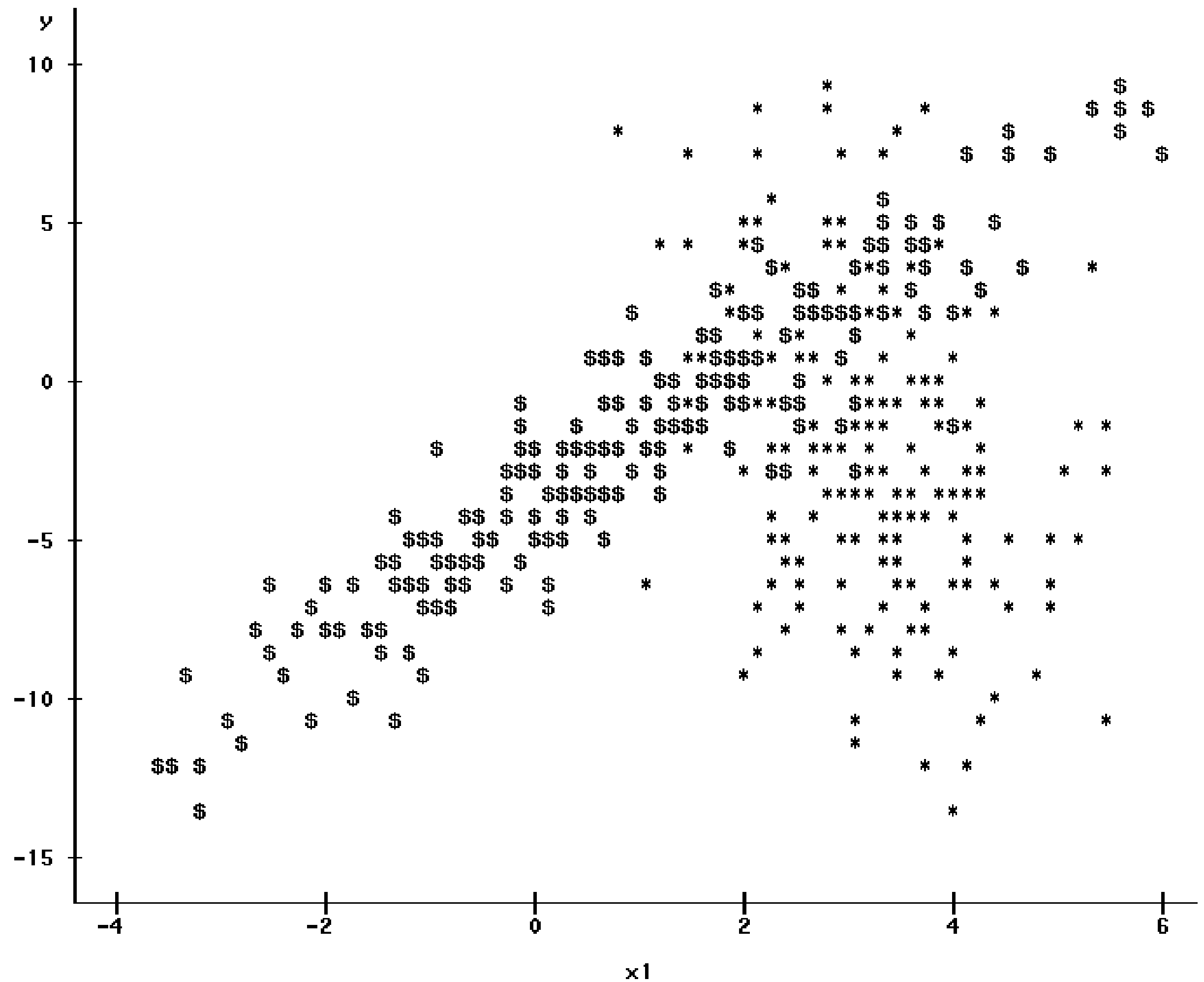
Example

```
proc plot data=simulation;
 plot y*x1=‘@' y*x2='*' / overlay;
run;
```

What message can you tell from the plot?

Simulated Data Set   10:24 Monday, September 13, 201

Plot of y*x1. Symbol used is '$'.
Plot of y*x2. Symbol used is '*'.



NOTE: 76 obs hidden.

Step 2. Calculate the correlation matrix

PROC CORR

PROC CORR computes the Pearson correlation coefficient for all pairs of variables listed in the VAR statement. The correlations are given in a matrix form.

- General Form

  PROC CORR data = *data set*;

  VAR *variable*;

  WITH *variables;*

Try this

```
proc corr data=simulation;
 var y x1 x2;
 title Pairwise correlation coefficients;
run;
```

Question: what messages can you tell from the output tables?

The CORR Procedure

3 Variables: y x1 x2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
y	200	-2.01070	4.69742	-402.13987	-13.61345	9.53415
x1	200	0.91373	2.09444	182.74577	-3.61074	6.04519
x2	200	3.15529	0.94250	631.05861	0.84937	5.52165

Pearson Correlation Coefficients, N = 200
Prob > \|r\| under H0: Rho=0

	y	x1	x2
y	1.00000	0.93595 <.0001	-0.33612 <.0001
x1	0.93595 <.0001	1.00000	-0.05277 0.4580
x2	-0.33612 <.0001	-0.05277 0.4580	1.00000

Try this

```
proc corr data=simulation;
 var x1 x2;
 with y;
 title Correlaiton coefficients with Y;
run;
```

Figure out how different the output is from the previous one.

The CORR Procedure

```
1 With Variables:    y
2      Variables:    x1        x2
```

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
y	200	-2.01070	4.69742	-402.13987	-13.61345	9.53415
x1	200	0.91373	2.09444	182.74577	-3.61074	6.04519
x2	200	3.15529	0.94250	631.05861	0.84937	5.52165

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	x1	x2
y	0.93595 <.0001	-0.33612 <.0001

Step 3. Do regression analysis

Simple Linear Regression

- Regression is the area of statistics that is concerned with finding a model that describes the relationship between a response variable and several predictor (explanatory) variables.
- In a simple linear regression model, there is only one predictor variable.
- Both PROC REG and PROC GLM can do regression analysis. But PROC REG has more options that are useful for regression analysis.

PROC REG

- **General Form**

PROC REG DATA=*data set*;

MODEL *response* = *predictors* < / OPTIONS >;

PLOT yvar * xvar < / OPTIONS >;

- Model Options

P ---- prints the observed value, the predicted value and the residuals.

R ---- prints everything in P-option plus standard errors, studentized residuals and so on.

Try this

```
proc reg data=simulation;
 model y=x1 x2;
 title Regression analysis of the simulated data set;
run;
```

To-do list:

1. Look for the parameter estimates and their variability.
2. Compare the parameter estimates with the true value of the parameters.

Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4208.61467	2104.30734	2271.87	<.0001
Error	197	182.47024	0.92624		
Corrected Total	199	4391.08491			

Root MSE	0.96242	R-Square	0.9584
Dependent Mean	-2.01070	Adj R-Sq	0.9580
Coeff Var	-47.86475		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.62409	0.24197	2.58	0.0106
x1	1	2.06512	0.03262	63.31	<.0001
x2	1	-1.43307	0.07249	-19.77	<.0001

Try this

```
proc reg data=simulation;
 model y=x1 x2 / r;
 title Regression analysis of the simulated data set;
run;
```

To-do lists:

1. Look for the predicted values and the residuals
2. How to check the distribution of residuals?

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
44	2.8384	4.5240	0.1189	-1.6856	0.955	-1.765	***¦	0.016
45	-3.5135	-2.2510	0.0719	-1.2625	0.960	-1.315	**¦	0.003
46	-7.5854	-8.2313	0.1484	0.6459	0.951	0.679	¦*	0.004
47	-0.3456	-1.5322	0.0685	1.1866	0.960	1.236	¦**	0.003
48	-2.0475	-2.3204	0.0729	0.2729	0.960	0.284	¦	0.000
49	-5.8204	-4.7055	0.1074	-1.1149	0.956	-1.166	**¦	0.006
50	2.3542	3.0030	0.1450	-0.6488	0.951	-0.682	*¦	0.004
51	3.1843	1.6111	0.0867	1.5732	0.959	1.641	¦***	0.007
52	-4.1738	-3.8239	0.0826	-0.3499	0.959	-0.365	¦	0.000
53	-0.6275	0.2829	0.0997	-0.9104	0.957	-0.951	*¦	0.003
54	-5.5172	-4.5478	0.0980	-0.9693	0.957	-1.012	**¦	0.004
55	-2.5160	-2.5797	0.1009	0.0637	0.957	0.0666	¦	0.000
56	-6.9552	-6.1463	0.1504	-0.8089	0.951	-0.851	*¦	0.006
57	1.3566	1.7439	0.1283	-0.3873	0.954	-0.406	¦	0.001
58	-4.1774	-4.9914	0.0815	0.8140	0.959	0.849	¦*	0.002
59	3.7250	2.4874	0.2168	1.2375	0.938	1.320	¦**	0.031
60	2.4210	1.1969	0.0854	1.2240	0.959	1.277	¦**	0.004
61	-0.6256	-1.3954	0.0693	0.7698	0.960	0.802	¦*	0.001
62	7.3079	6.9582	0.1527	0.3497	0.950	0.368	¦	0.001
63	5.6384	4.1484	0.1191	1.4899	0.955	1.560	¦***	0.013
64	-7.6449	-8.0959	0.1134	0.4510	0.956	0.472	¦	0.001
65	-5.3275	-5.2337	0.1002	-0.0938	0.957	-0.0980	¦	0.000
66	-2.2752	-1.1959	0.0701	-1.0794	0.960	-1.125	**¦	0.002
67	-1.6247	-1.2294	0.0750	-0.3953	0.959	-0.412	¦	0.000
68	4.4188	2.0174	0.1176	2.4014	0.955	2.514	¦*****	0.032
69	-3.3302	-4.2722	0.1006	0.9420	0.957	0.984	¦*	0.004
70	-5.5856	-6.0382	0.1133	0.4526	0.956	0.474	¦	0.001
71	-0.3404	-0.6213	0.0852	0.2809	0.959	0.293	¦	0.000
72	4.9548	5.4123	0.1320	-0.4575	0.953	-0.480	¦	0.001
73	-3.0297	-2.1755	0.0682	-0.8542	0.960	-0.890	*¦	0.001
74	-6.6415	-6.1990	0.0934	-0.4425	0.958	-0.462	¦	0.001
75	-2.6816	-2.2020	0.0954	-0.4796	0.958	-0.501	*¦	0.001
76	-0.7264	-0.9410	0.1855	0.2146	0.944	0.227	¦	0.001
77	-3.5166	-3.9907	0.0903	0.4741	0.958	0.495	¦	0.001
78	-3.8472	-2.5468	0.0771	-1.3004	0.959	-1.356	**¦	0.004
79	-4.5702	-4.3911	0.0789	-0.1791	0.959	-0.187	¦	0.000
80	-6.8768	-7.6862	0.1345	0.8095	0.953	0.849	¦*	0.005
81	1.8858	1.2014	0.0856	0.6844	0.959	0.714	¦*	0.001
82	3.2896	2.7834	0.1031	0.5062	0.957	0.529	¦*	0.001
83	-7.9430	-8.3488	0.1277	0.4058	0.954	0.425	¦	0.001
84	1.7391	0.9104	0.1022	0.8288	0.957	0.866	¦*	0.003
85	2.8101	1.7807	0.0886	1.0295	0.958	1.074	¦**	0.003
86	-0.5916	-1.1078	0.0903	0.5163	0.958	0.539	¦*	0.001

Step 4. Model Checking

What needs to be checked?

In regression analysis, assumptions are made about the error (noise) terms. They are assumed to be

1. Independent.
2. Normally distributed.
3. Mean 0.
4. Constant variance.

These assumptions need to be checked after we obtain a fitted model.

Output residuals to a new dataset

1. We can create a <u>new</u> data set containing the analysis results (such as the residuals) obtained from PROC REG.
2. Then we can use PROC UNIVARIATE, PROC PLOT… to perform analysis of residuals.

```
proc reg data=simulation;
 model y=x1 x2 / r;
 output out=diagnostics r=residual;
run;
```

Question

How can you check the following assumptions for the residual?

1. Independent.
2. Normally distributed.
3. Mean 0.
4. Constant variance.

A check list for model diagnostics

1. Independent.
   – Plot the residuals versus X1 and X2 (PROC PLOT).
2. Normally distributed.
   – Plot the histogram of the residuals (PROC CHART)
   – Perform hypothesis testing (PROC UNIVARIATE).
3. Mean 0.
   – Check the residual plots.
   – Perform hypothesis testing (PROC TTEST)
4. Constant variance.
   – Check the residual plots

A check list for data analysis

Observations of two variables Y and X.

How to study the relationship between Y and X?

Step 1. Do the plots!

Step 2. Calculate the correlation matrix.

Step 3. Do regression analysis

Step 4. Model Checking