STAT 430
SAS Examples SAS6
==========

ssh bnk@glue.umd.edu, tap sas913 (or sas82), sas
https://www.statlab.umd.edu/sasdoc/sashtml/onldoc.htm

a. t-Test
b. Wilcoxon Rank-Sum Test
c. Paired t-Test
d. Generate Random Data

a. t-Test
==========

0. First assign to treatment A or B at random

We have 25 milking cows to which we assign treatments A or B
using random number generator RANUNI(0) and PROC RANK.

```
OPTION PS=35 LS=70;

DATA ASSIGN;
INPUT NAME $ 5.;
/* GROUP will be assigned unif(0,1) random numbers*/;
GROUP=RANUNI(0);
/* Cows names */;
DATALINES;
C1
C2
C3
C4
C5
C6
C7
;
```

To verify what we have we print. Every cow is assigned a random number:

```
PROC PRINT DATA=ASSIGN;
VAR NAME GROUP;
RUN;
```

Obs	NAME	GROUP
1	C1	0.09277
2	C2	0.72997
3	C3	0.26708
4	C4	0.94593
5	C5	0.52810
6	C6	0.10918
7	C7	0.04104

If run again get different random numbers since we used the seed "0"!!! It depends on the time!!!

Obs	NAME	GROUP
1	C1	0.29978
2	C2	0.34289
3	C3	0.32744
4	C4	0.20692
5	C5	0.32888
6	C6	0.12573
7	C7	0.72896

Now use PROC RANK with option GROUPS=2 to divide the subjects and assign to '0' or '1' treatments, and create a new data set containing the treatments TREAT.

Apparently PROC RANK clips the random numbers in some fasion when making the assignment "0' or '1'!!!

```
PROC RANK DATA=ASSIGN GROUPS=2 OUT=TREAT;
VAR GROUP;
RUN;
```

```
To see the content of TREAT: Without FORMAT get 0-1 assignments
instead of A,B assignments!!!

PROC PRINT DATA=TREAT;
ID NAME;
VAR GROUP;
RUN;
```

NAME	GROUP
C1	1
C2	0
C3	1
C4	1
C5	0
C6	0
C7	1

```
Now use PROC FORMAT to change from 0-1 to A-B:

PROC FORMAT;
VALUE ZERONE 0='A' 1='B';
RUN;


PROC SORT DATA=TREAT; <---- Also DATA=ASSIGN  works!!!
BY NAME;
RUN;

PROC PRINT DATA=TREAT;
ID NAME;
VAR GROUP;
FORMAT GROUP ZERONE.;
RUN;
```

The SAS System

NAME	GROUP
C1	B
C2	A
C3	B
C4	B
C5	A
C6	A
C7	B

1. t-test

Milking cows were assigned randomly to treatment A or B, and then
the average daily milk production over a 3 week period was recorded.
Test equality of means using a t-test.

NOTE: The category or class or group are the indep variable!!!

```
OPTION PS=35 LS=70;

Ex. 1
======

DATA MILK;
INPUT DIET $ YIELD;
DATALINES;
A 44
A 44
A 56
A 46
A 47
A 38
A 58
A 53
A 49
A 35
```

```
A 46
A 30
A 41
B 35
B 47
B 55
B 29
B 40
B 39
B 32
B 41
B 42
B 57
B 51
B 39
;


PROC TTEST DATA=MILK;
CLASS DIET;
VAR YIELD;
RUN;
```

The TTEST Procedure

Statistics

Variable	DIET	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev
YIELD	A	13	40.32	45.154	49.987	5.7355
YIELD	B	12	36.697	42.250	47.803	6.1913
YIELD	Diff (1-2)		-4.02	2.9038	9.828	6.4985

Statistics

Variable	DIET	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum

YIELD	A	7.9984	13.203	2.2184	30	58
YIELD	B	8.7399	14.839	2.523	29	57
YIELD	Diff (1-2)	8.3613	11.729	3.3472		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
YIELD	Pooled	Equal	23	0.87	0.3946
YIELD	Satterthwaite	Unequal	22.3	0.86	0.3966

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
YIELD	Folded F	11	12	1.19	0.7621

Ex 2. (From the SAS book)
==========

```
DATA RESPONSE;
INPUT GROUP $ TIME;
DATALINES;
C  80
C  93
C  83
C  89
C  98
T 100
T 103
T 104
T  99
T 102
;

PROC TTEST DATA=RESPONSE;
```

```
CLASS GROUP;
VAR TIME;
RUN;
```

The TTEST Procedure

Statistics

Variable	GROUP	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev
TIME	C	5	79.535	88.6	97.665	4.3741
TIME	T	5	99.025	101.6	104.17	1.2424
TIME	Diff (1-2)		-20.83	-13	-5.173	3.6249

Statistics

Variable	GROUP	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
TIME	C	7.3007	20.979	3.265	80	98
TIME	T	2.0736	5.9587	0.9274	99	104
TIME	Diff (1-2)	5.3666	10.281	3.3941		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
TIME	Pooled	Equal	8	-3.83	0.0050
TIME	Satterthwaite	Unequal	4.64	-3.83	0.0141

The TTEST Procedure

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F

```
    TIME        Folded F         4         4      12.40    0.0318
```

b. Wilcoxon Rank-Sum Test
==========

Nonparametric two sample Wilcoxon Rank-Sum test (1945).

H_0: The two dists are identical
H_1: Dist A is SHIFTED to the right of the dist B.
     Dist A is SHIFTED to the left of the dist B.
     Dists A,B, are different.

Ex 1.

The mineral content of specimens from two locations A, B, are recorded.

A: 7.6, 11.1, 6.8, 9.8, 4.9, 6.1, 15.1 (n1=7=n_A)
B: 4.7, 6.4, 4.1, 3.7, 3.9 (n2=5=n_B)

Does location A have a higher mineral content?

Rank the combined data:

```
B    B   B   B   A   A   B   A   A   A   A    A
3.7 3.9 4.1 4.7 4.9 6.1 6.4 6.8 7.6 9.8 11.1 15.1
1   2   3   4   5   6   7   8   9   10  11   12  <---- RANKS
```

USE SMALLER SAMPLE:

W=W_B=1+2+3+4+7=17 <---- Sum of the ranks of B

```
Test equality vs second dist lies to the left of the first.
Thus we reject for small values of W.

Every 5-tuple out of {1,2,3,4,5,6,7,....12} has same chance.

P(W <= 17)=P(W=15)+P(W=16)+2*P(W=17)=0.0051

1+2+3+4+5=15
1+2+3+4+6=16   <--- Four 5's cases out of binomail coeef (12 C 5)=792 5's
1+2+3+5+6=17
1+2+3+4+7=17


 12
(  ) =12!/[5!*7!] = gamma(13)/(gamma(6)*gamma(8)) = 792
 5
> 4/792
[1] 0.005050505=0.0051 <--- P-value

Thus we reject the hypothesis of dist equality. Now let's see what
SAS does.


Note: E[W_B]= n_B*(n_A+n_B+1)/2=5*13/2=32.5
      E[W_A]= n_A*(n_A+n_B+1)/2=7*13/2=45.5

OPTION PS=35 LS=70;
DATA MINERAL;
INPUT LOCATION $ CONT @@; <-- Use "@@" to input data  horizontally.
DATALINES;
A 7.6 A 11.1 A 6.8 A 9.8 A 4.9 A 6.1 A 15.1
B 4.7 B  6.4 B 4.1 B 3.7 B 3.9
;


PROC NPAR1WAY DATA=MINERAL WILCOXON;
CLASS LOCATION;
VAR CONT;
EXACT WILCOXON;
RUN;
```

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable CONT
Classified by Variable LOCATION

LOCATION	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
A	7	61.0	45.50	6.157651	8.714286
B	5	17.0	32.50	6.157651	3.400000

Wilcoxon Two-Sample Test

Statistic (S)	17.0000	
Normal Approximation		
Z	-2.4360	
One-Sided Pr < Z	0.0074	<-- Good Approx.
Two-Sided Pr > |Z|	0.0149	
t Approximation		
One-Sided Pr < Z	0.0165	
Two-Sided Pr > |Z|	0.0331	
Exact Test		
One-Sided Pr <= S	0.0051	<---WHAT I GOT ABOVE!!!
Two-Sided Pr >= |S - Mean|	0.0101	

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	6.3363

DF	1
Pr > Chi-Square	0.0118

Ex 2.

Two fabrics tested for fire resistance. Measure damage in inches.

Fabric A: 5.7 7.3 7.6 6.0 6.5
Fabric B: 4.9 7.4 5.3 4.6 6.2

Any difference in flammability? Two sided test: H_1: F_1 neq F_2
Reject if W <= c1 or W >= c2.

B	B	B	A	A	B	A	A	B	A
4.6	4.9	5.3	5.7	6.0	6.2	6.5	7.3	7.4	7.6
1	2	3	4	5	6	7	8	9	10

W = 1+2+3+6+9 = 21 <--- Sum ranks of B

Sum ranks of A is 34.

```
DATA FLAME;
INPUT FABRIC $ DAMAGE @@;
DATALINES;
A 5.7 A 7.3 A 7.6 A 6.0 A 6.5
B 4.9 B 7.4 B 5.3 B 4.6 B 6.2
;

PROC NPAR1WAY DATA=FLAME WILCOXON;
CLASS FABRIC;
VAR DAMAGE;
EXACT WILCOXON;
RUN;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable DAMAGE

Classified by Variable FABRIC

FABRIC	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
A	5	34.0	27.50	4.787136	6.80
B	5	21.0	27.50	4.787136	4.20

Wilcoxon Two-Sample Test

Statistic (S)	34.0000
Normal Approximation	
Z	1.2534
One-Sided Pr > Z	0.1050
Two-Sided Pr > |Z|	0.2101
t Approximation	
One-Sided Pr > Z	0.1208
Two-Sided Pr > |Z|	0.2417 <-- Good Approx.
Exact Test	
One-Sided Pr >= S	0.1111
Two-Sided Pr >= |S - Mean|	0.2222 <---

Z includes a continuity correction of 0.5.

The NPAR1WAY Procedure

Kruskal-Wallis Test

Chi-Square	1.8436
DF	1
Pr > Chi-Square	0.1745

Dont reject!!!

c. Paired t-Test
==========

Simply use PROC MEANS on the difference with options: T and PRT (p-val).

Ex.1

A pill is used to reduce blood preasure. Have 15 subjects.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before (x):	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
After (y):	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74
d = x-y :	2	8	10	6	18	10	4	26	18	-8	0	32	0	-4	10

Test:

H_0: mu1=mu2 vs H_1: mu1 > mu2 <---- Note one sided!!!

Use the test stat: T= sqrt(n)*Dbar/S_D

```
DATA BP;
INPUT BEFORE AFTER;
DIFF=BEFORE-AFTER;
DATALINES;
70 68
80 72
72 62
76 70
76 58
76 66
72 68
78 52
```

```
82 64
64 72
74 74
92 60
74 74
68 72
84 74
;

PROC MEANS DATA=BP N MEAN STDERR T PRT;
VAR DIFF;
RUN;

Note: df=n-1=15-1=14!!!
Note: We get a two sided test. To get one sided, divide p-val by 2!!!
```

The MEANS Procedure

Analysis Variable : DIFF

N	Mean	Std Error	t Value	Pr > |t|
15	8.8000000	2.8338095	3.11	0.0077

```
P(T_{14} > 3.11 )=0.003839171 ===>            Reject in favor of m1 > m2.
P(|(T_{14}|>3.11)=2*0.003839171=0.007678342==>Reject in favor of m1
neq mu2.


More directly: Use PROC TTEST.

option ps=35 ls=70;

data Paired;
input diff;
datalines;
```

```
2
8
10
6
18
10
4
26
18
-8
0
32
0
-4
10
;

proc ttest data=Paired;
var diff;
run;
```

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev
diff	15	2.7221	8.8	14.878	8.0353	10.975

Statistics

Variable	Upper CL Std Dev	Std Err	Minimum	Maximum
diff	17.309	2.8338	-8	32

T-Tests

```
Variable      DF    t Value    Pr > |t|

diff          14       3.11      0.0077
```

d. Generate Random Data
==========

Normal N(0,1) Data

The SAS code below produces a data set called WORK.RANDOM containing 1000 N(0,1) observations.

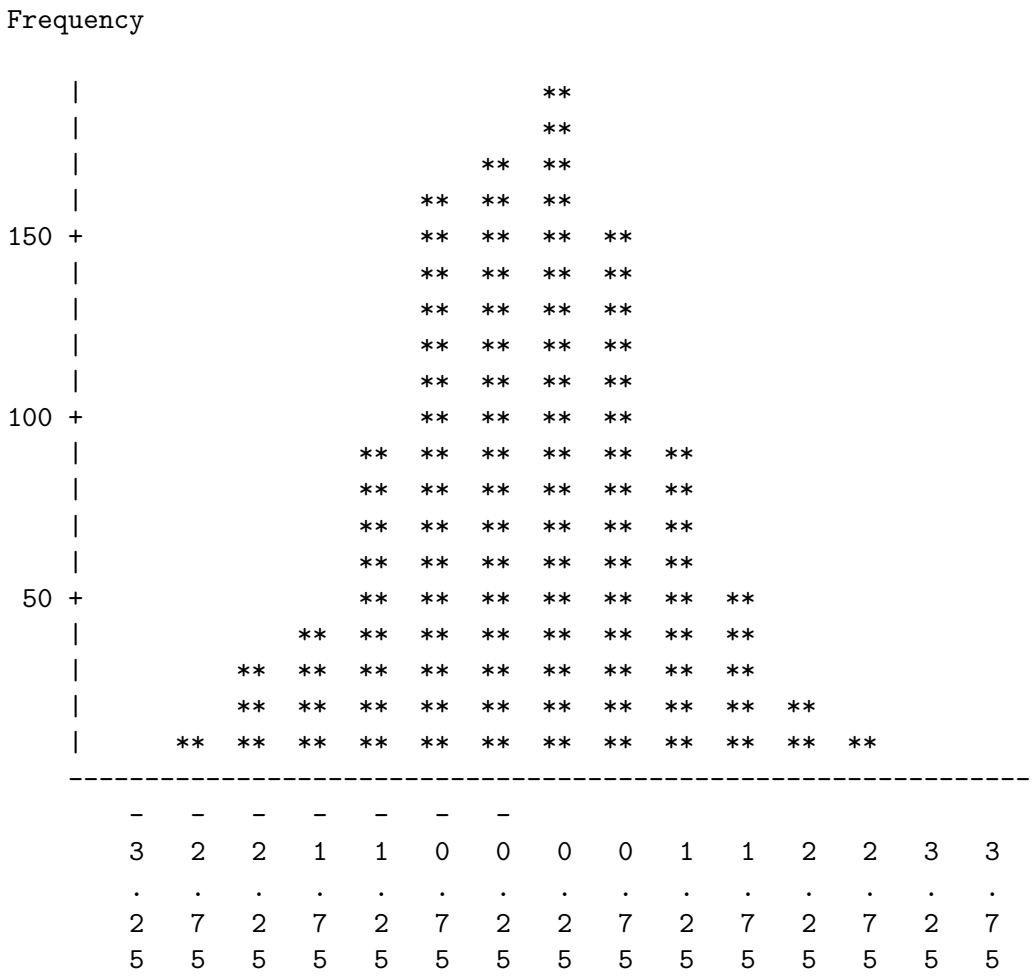
```
option ps=35 ls=70;

data random;
do i = 1 to 1000;
r = RANNOR(0);
output;
end;
run;
```

Get this message from the SAS Log window:

```
NOTE: The data set WORK.RANDOM has 1000 observations and 2 variables.
NOTE: DATA statement used:
      real time           4.94 seconds
      cpu time            0.17 seconds
```

To see a histogram of the generated data we use PROC CHART:

```
proc chart data=WORK.RANDOM;
var r;
run;
```



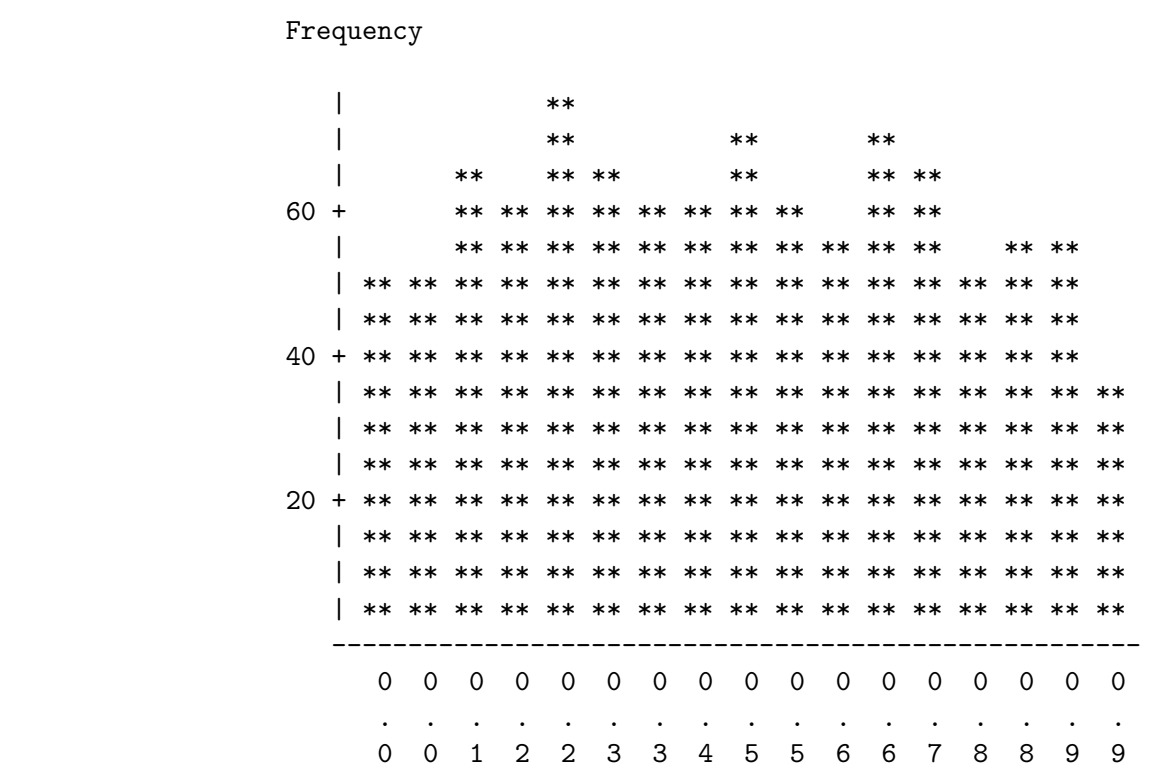
The MEANS Procedure

Analysis Variable : r

N	Mean	Std Dev	Minimum	Maximum
1000	-0.0307679	0.9932871	-3.1157183	2.6768072

Uniform Unif(0,1) Data

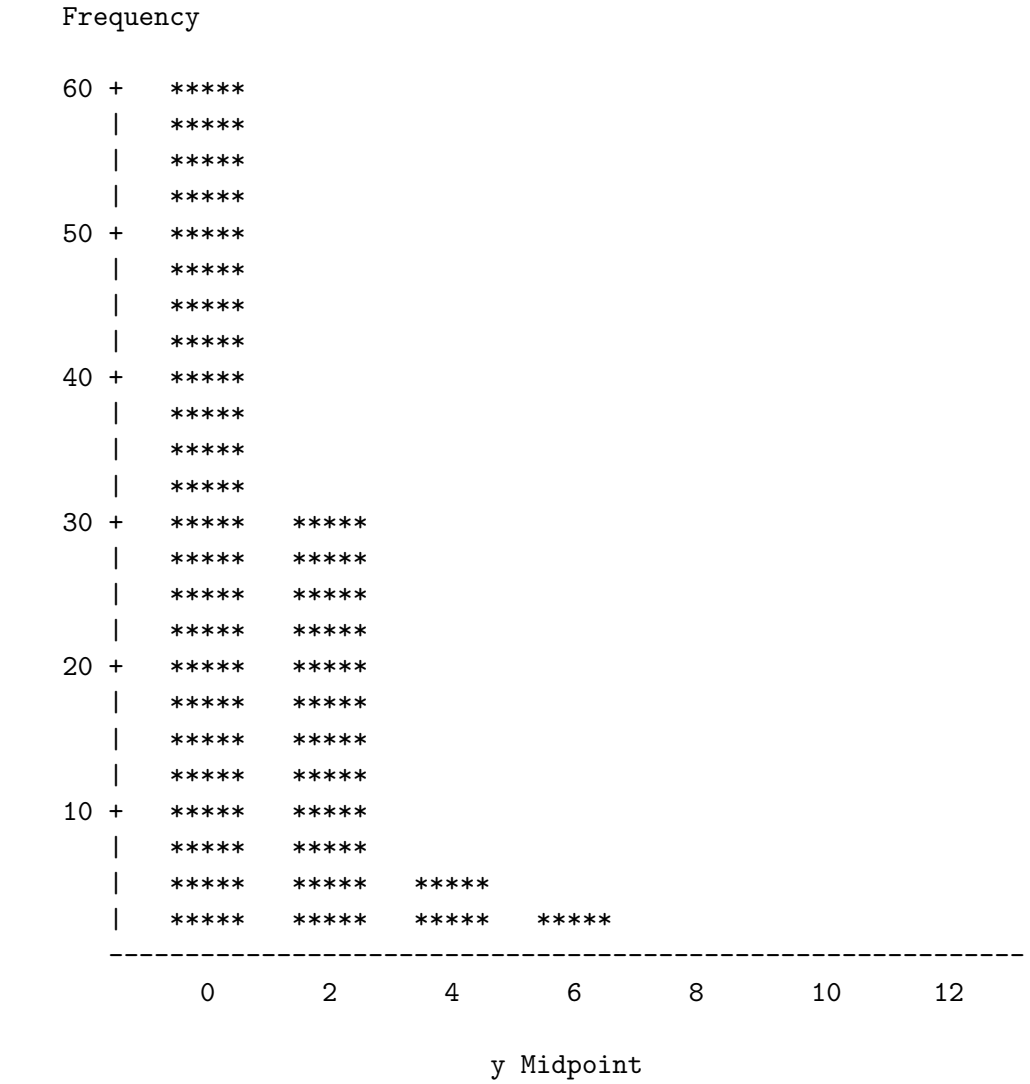
```
data random;
do i = 1 to 1000;
u = RANUNI(0);
output;
end;
run;

proc chart data=WORK.RANDOM;
var u;
run;
```



```
            3  9  5  1  7  3  9  5  1  7  3  9  5  1  7  3  9

                                u Midpoint

NOTE: The histogram midpoints are 0.03,0.09,....,0.99.


                        The MEANS Procedure

                        Analysis Variable : u

    N            Mean         Std Dev         Minimum         Maximum
 ------------------------------------------------------------------------
 1000       0.4953382       0.2797816       0.0033764       0.9991905
 ------------------------------------------------------------------------

ChiSq(1) Data
--------------------------
data random;
do i = 1 to 100;
r = RANNOR(0);
y = r**2;
output;
end;
run;

From the LOG Window: 3 variables!!! i,r,y!!!

NOTE: The data set WORK.RANDOM has 100 observations and 3 variables.
NOTE: DATA statement used:
      real time           0.14 seconds
      cpu time            0.00 seconds


proc print data=WORK.RANDOM;
id i;
run;

Prints all 3 variables!!!
```

i	r	y
1	-0.43033	0.1852
2	-1.62440	2.6387
3	-1.02342	1.0474
4	0.31375	0.0984
5	-0.62463	0.3902
6	-1.45657	2.1216
7	0.65087	0.4236
8	-1.82668	3.3368
9	1.53697	2.3623
10	-0.31588	0.0998
.....	.....	.....
.....	.....	.....
.....	.....	.....
85	0.19655	0.03863
86	-0.74097	0.54904
87	0.03033	0.00092
88	1.35246	1.82914
89	2.53814	6.44218
90	-0.60645	0.36779

```
proc chart data=WORK.RANDOM;
var y;
run;
```

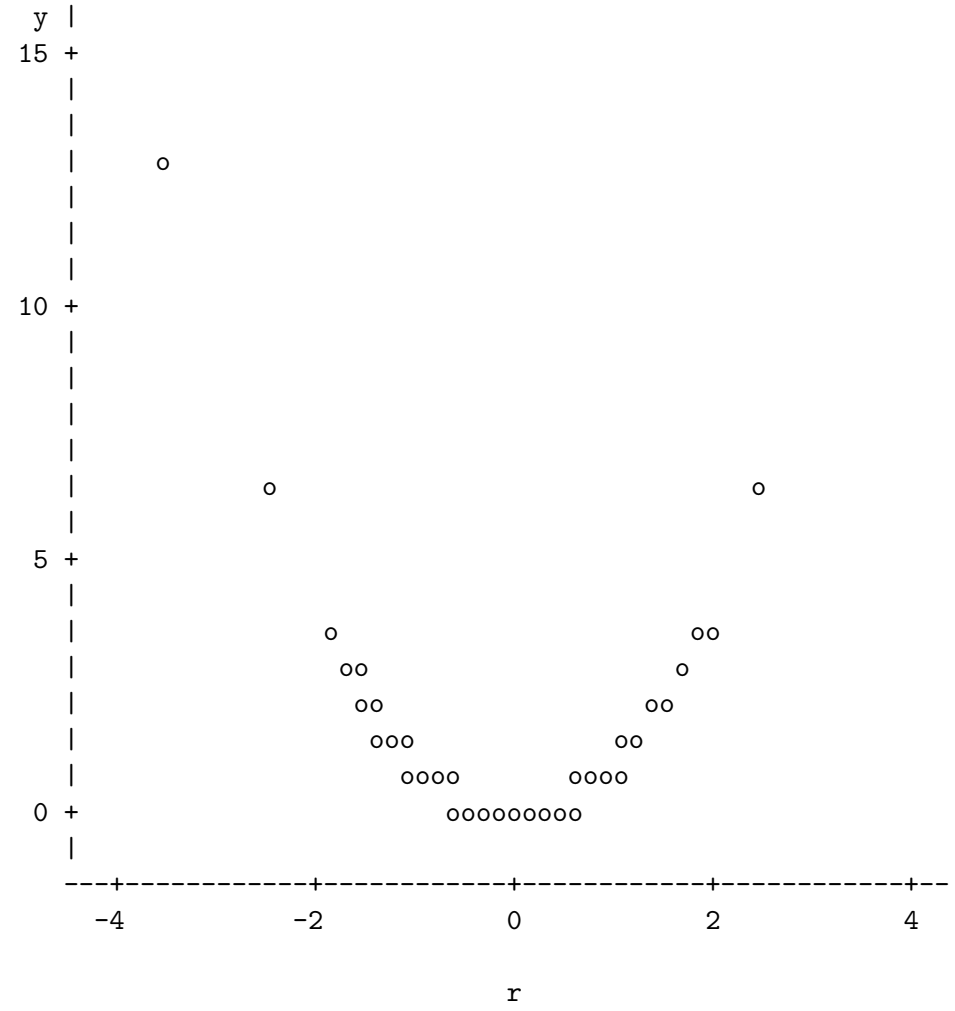



```
proc plot data=WORK.RANDOM;
plot y*r;
run;
```

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

     y |
    15 +
       |
       |
       |     A
       |
       |
       |
    10 +
       |
       |
       |
       |
       |
       |            A                                 A
       |
     5 +
       |
       |                D                         AA
       |                 AA                      A
       |                  BE                   CD
       |                   ADA               BC
       |                     ACIE        CBEA
     0 +                        BBCFFGDBB
       |
       ---+------------+------------+------------+------------+--
         -4           -2            0            2            4
```

```
proc plot data=WORK.RANDOM;
plot y*r='o';
run;
```

Plot of y*r. Symbol used is 'o'.



NOTE: 65 obs hidden.