2.9 Problems (last update 21 June 2024)

1. Identify the unit on which each of the following variables is measured and classify the variable as qualitative or quantitative. If the variable is qualitative, classify it as nominal or ordinal. If the variable is quantitative, classify it as inherently discrete or conceptually continuous. Be sure to explain your reasoning. For example, for the variable "the age of a person", a person is a unit and the possible values of the variable are positive numbers representing possible ages. Since these ages are numbers on the number line this is a quantitative variable. In practice, age is usually treated as a discrete quantitative variable since we usually round our age down to our most recent birth date; however, thinking of age as the amount of time since birth, we can argue that age should be viewed as a continuous variable.

a) The number of children in a family;

b) The size of the memory (RAM) in a laptop computer (measured in GB);

c) The number of correct answers on a multiple choice exam;

d) The blood type (A, AB, B, O) of a person;

e) The letter grade obtained by a student in a statistics class.

Terminology note: Each of questions 2 and 3 involves two variables. One (the explanatory variable) is used to divide the units into two subgroups. The other (the response variable) is the variable of primary interest in the sense that we want to compare and contrast the distributions of the response variable for each subgroup.

2. The following table provides the number of households (in thousands) for the city of New York and the city of Los Angeles categorized by the language spoken at home for the household.

language spoken at home

language	**New York**	**Los Angeles**
English	3982	1384
Spanish	1884	1539
Other Indo–European	998	244
Asian and Pacific Island	546	277
total	7410	3444

a) There are 10,854 thousand units in this example. What is a unit for this example?

b) What is the explanatory (grouping) variable in this example? What are its possible values?

c) What is the response variable in this example? What are its possible values? Is this variable nominal qualitative, ordinal qualitative, discrete quantitative, or continuous quantitative?

Language spoken at home results by city (relative frequency distributions)

The FREQ Procedure

city=LosAngeles

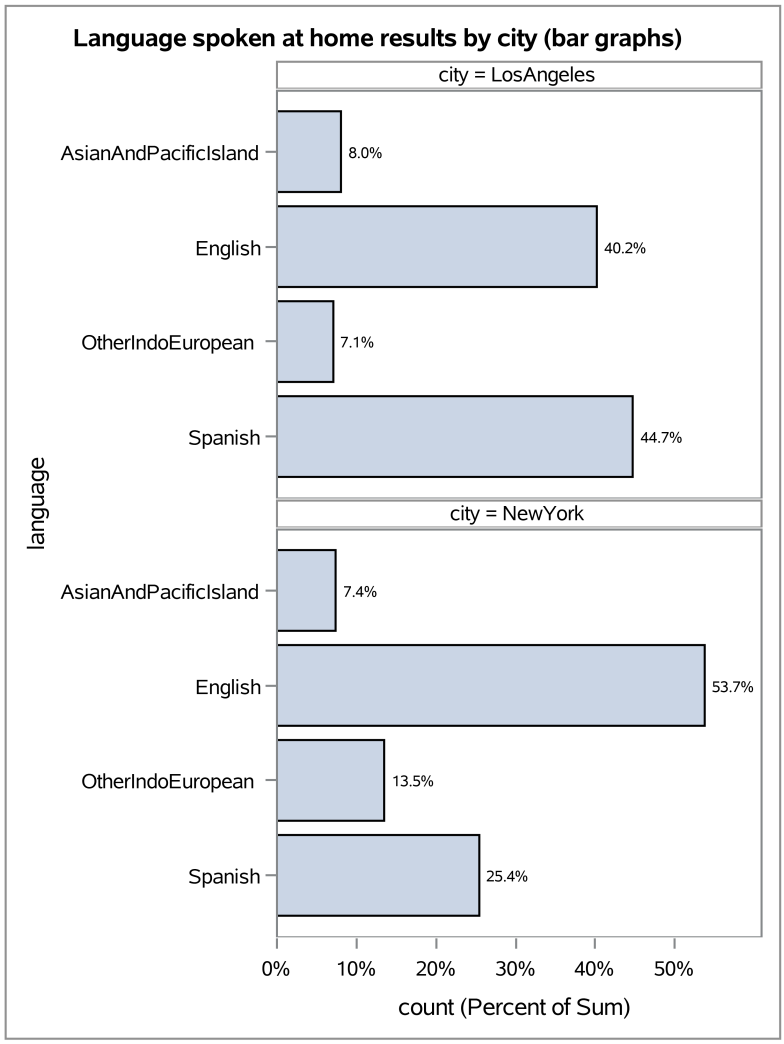
language	Frequency	Percent	Cumulative Frequency	Cumulative Percent
English	1384	40.19	1384	40.19
Spanish	1539	44.69	2923	84.87
OtherIndoEuropean	244	7.08	3167	91.96
AsianAndPacificIsland	277	8.04	3444	100.00

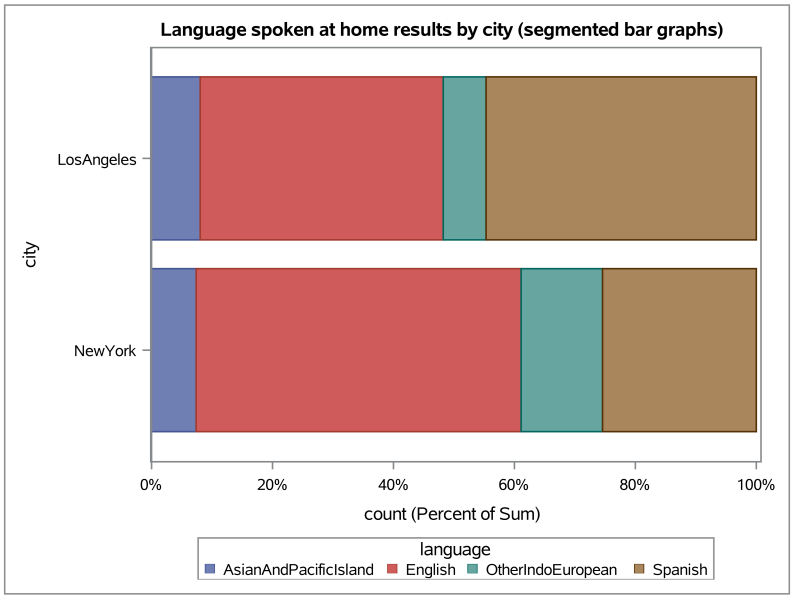
Language spoken at home results by city (relative frequency distributions)

The FREQ Procedure

city=NewYork

language	Frequency	Percent	Cumulative Frequency	Cumulative Percent
English	3982	53.74	3982	53.74
Spanish	1884	25.43	5866	79.16
OtherIndoEuropean	998	13.47	6864	92.63
AsianAndPacificIsland	546	7.37	7410	100.00


Language spoken at home results by city (bar graphs)
city = LosAngeles
AsianAndPacificIsland
English
OtherIndoEuropean
Spanish
8.0%
40.2%
7.1%
44.7%
city = NewYork
7.4%
53.7%
13.5%
25.4%
language
0%
10%
20%
30%
40%
50%
count (Percent of Sum)


Language spoken at home results by city (segmented bar graphs)
LosAngeles
NewYork
city
0%
20%
40%
60%
80%
100%
count (Percent of Sum)
language
AsianAndPacificIsland
English
OtherIndoEuropean
Spanish

3. Joe DiMaggio and Mickey Mantle were two well known baseball players. DiMaggio played center field for the New York Yankees for 13 years and was succeeded by Mantle who played center field for 18 years. There has been some argument about which of these two players was better at hitting home runs. The data given below are the numbers of home runs hit by the player during each of the seasons he played. For each player these numbers of home runs are listed in order by the seasons he played.

Home run data.

Joe DiMaggio:	29 46 32 30 31 30 21 25 20 39 14 32 12
Mickey Mantle:	13 23 21 27 37 52 34 42 31 40 54 30 15 35 19 23 22 18

a. There are 31 units in this example. What is a unit for this example?

b. What is the explanatory (grouping) variable in this example? What are its possible values?

c. What is the response variable in this example? What are its possible values? Is this variable nominal qualitative, ordinal qualitative, discrete quantitative, or continuous quantitative?

Joe Dimaggio home runs summary

The MEANS Procedure

Analysis Variable : runs									
N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Quartile Range	Mean	Std Dev
13	12.0000	21.0000	30.0000	32.0000	46.0000	34.0000	11.0000	27.7692	9.4353

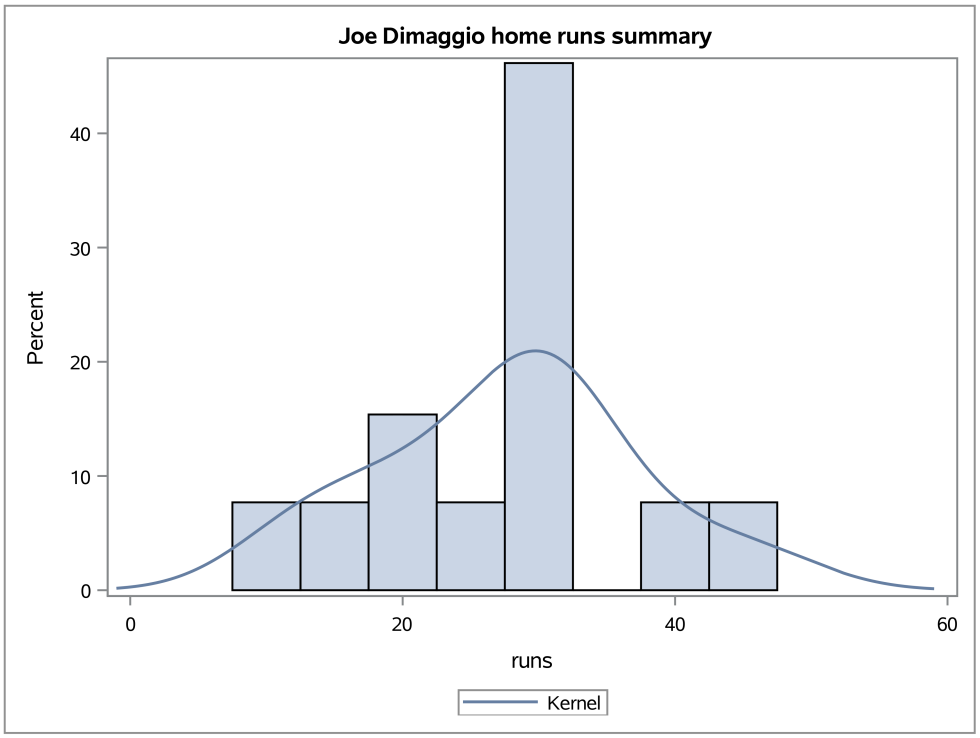
Analysis Variable : runs												
1st Pctl	5th Pctl	10th Pctl	20th Pctl	30th Pctl	40th Pctl	50th Pctl	60th Pctl	70th Pctl	80th Pctl	90th Pctl	95th Pctl	99th Pctl
12.0000	12.0000	14.0000	20.0000	21.0000	29.0000	30.0000	30.0000	32.0000	32.0000	39.0000	46.0000	46.0000

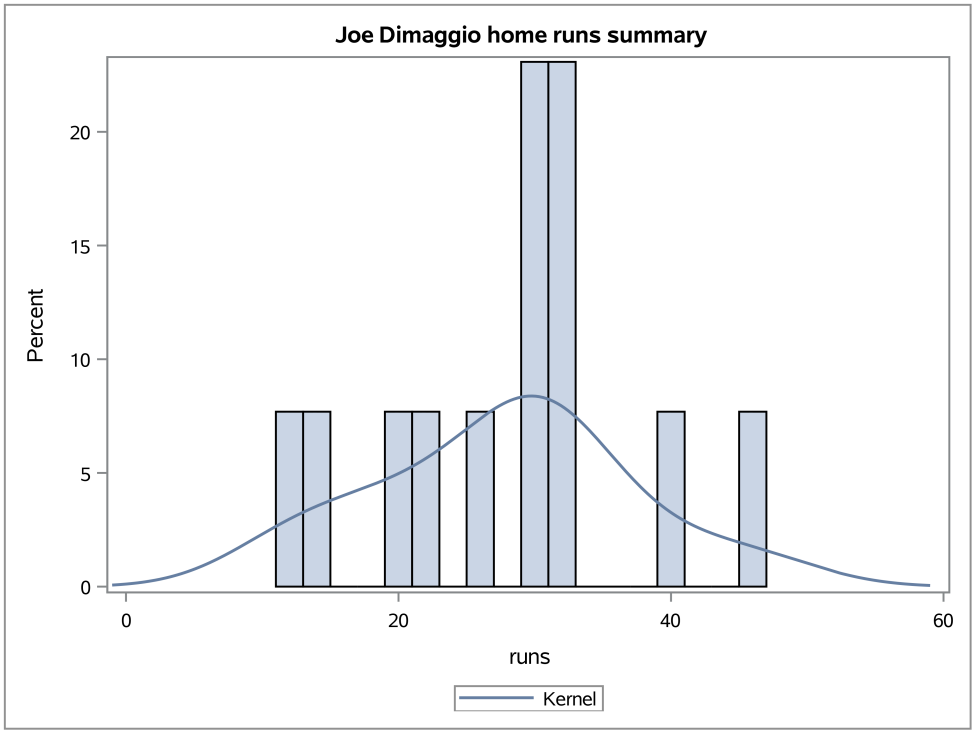
Mickey Mantle home runs summary

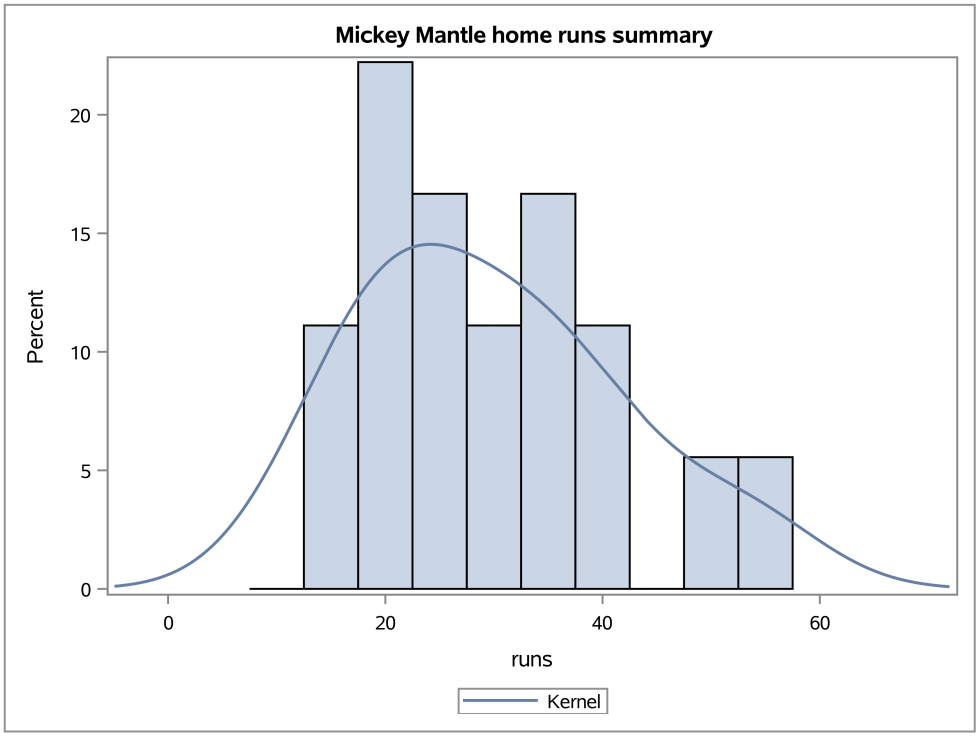
The MEANS Procedure

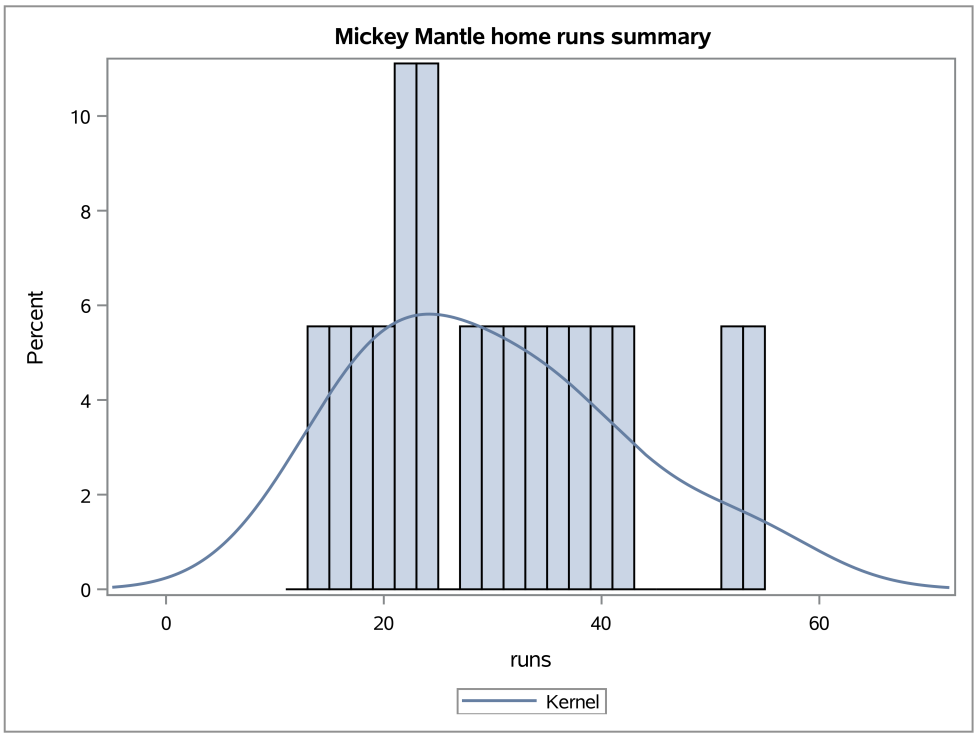
Analysis Variable : runs									
N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Quartile Range	Mean	Std Dev
18	13.0000	21.0000	28.5000	37.0000	54.0000	41.0000	16.0000	29.7778	11.9438

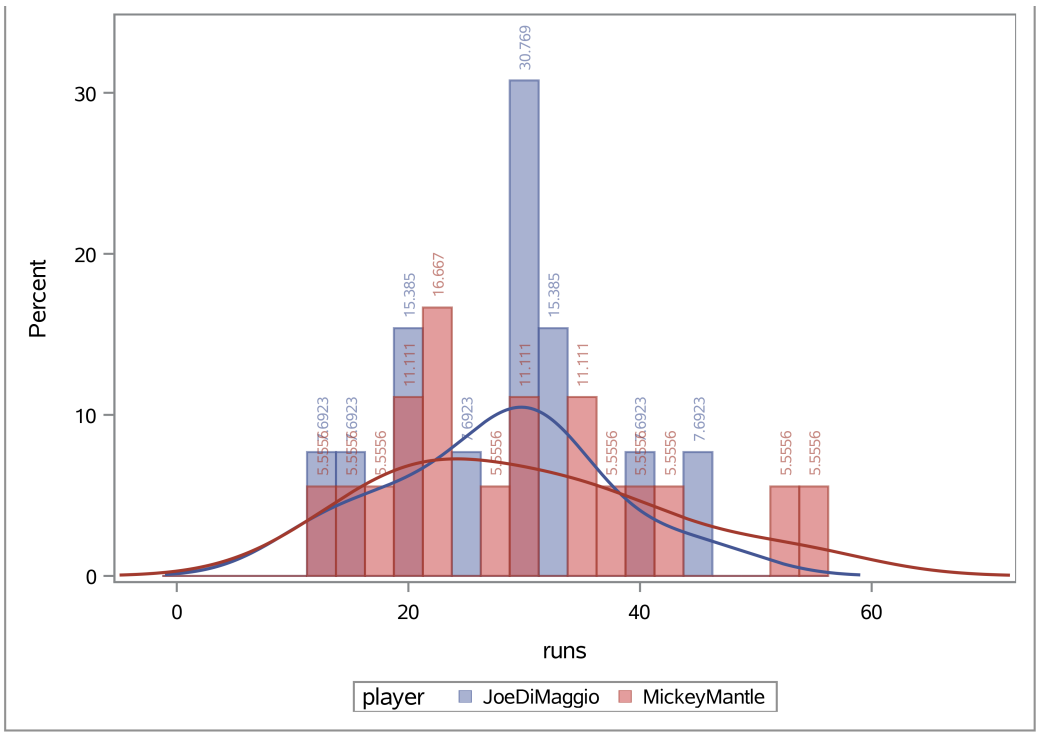
Analysis Variable : runs												
1st Pctl	5th Pctl	10th Pctl	20th Pctl	30th Pctl	40th Pctl	50th Pctl	60th Pctl	70th Pctl	80th Pctl	90th Pctl	95th Pctl	99th Pctl
13.0000	13.0000	15.0000	19.0000	22.0000	23.0000	28.5000	31.0000	35.0000	40.0000	52.0000	54.0000	54.0000

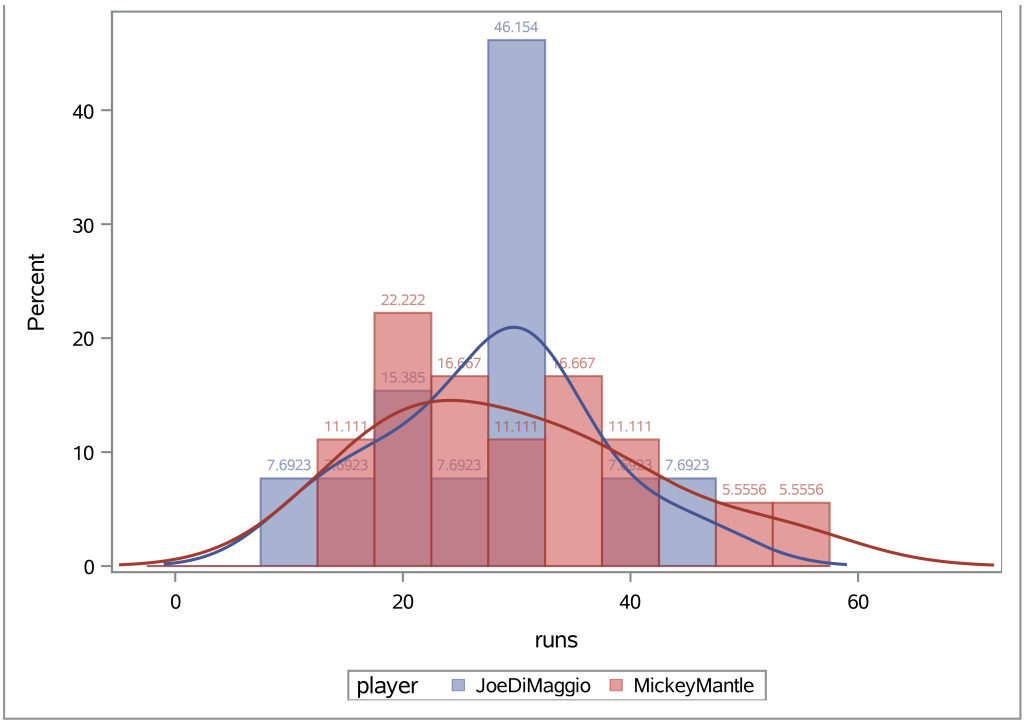

Joe Dimaggio home runs summary
Percent
0
10
20
30
40
0
20
40
60
runs
Kernel

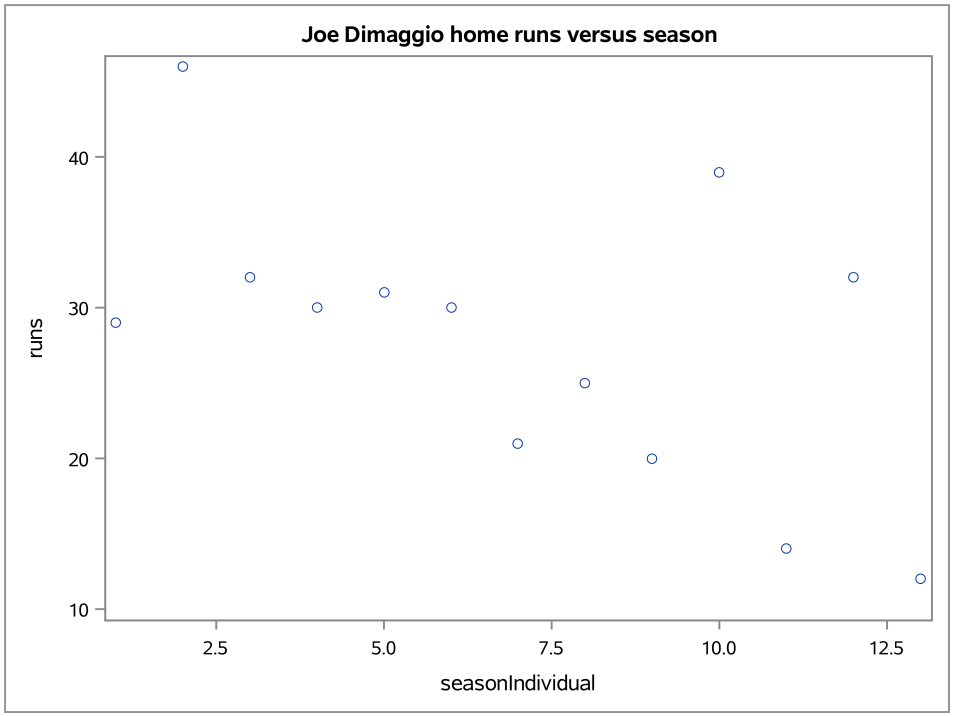

Joe Dimaggio home runs summary
Percent
0
5
10
15
20
0
20
40
60
runs
Kernel


Mickey Mantle home runs summary
Percent
0
5
10
15
20
0
20
40
60
runs
Kernel


Mickey Mantle home runs summary
Percent
0
2
4
6
8
10
0
20
40
60
runs
Kernel


Percent
30
20
10
0
0
20
40
60
runs
30.769
15.385
16.667
11.111
7.6923
5.5556
player
JoeDiMaggio
MickeyMantle

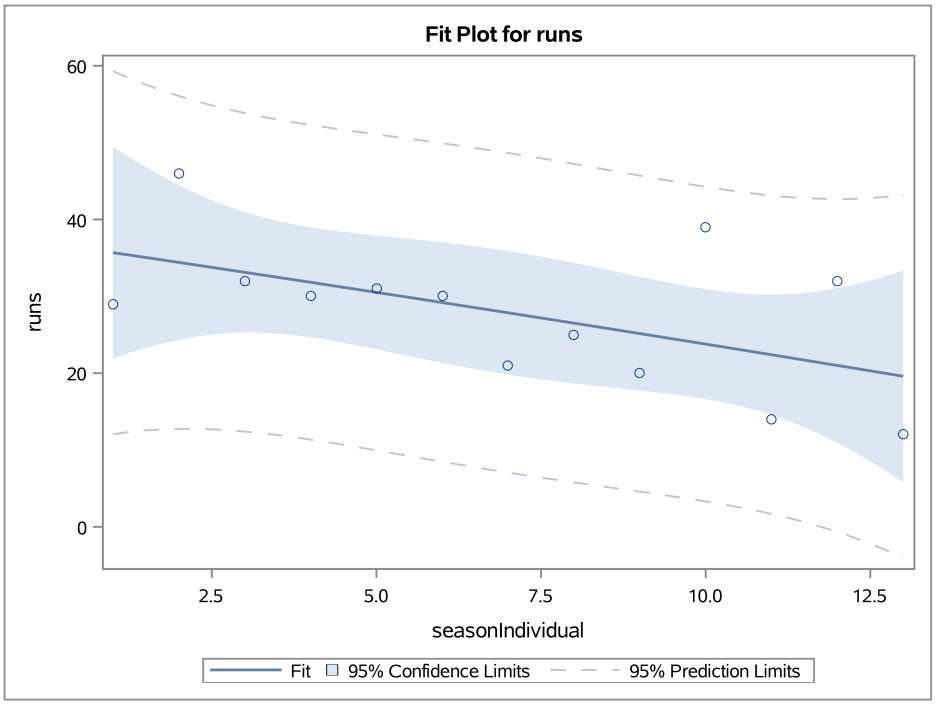

Percent
40
30
20
10
0
0
20
40
60
runs
46.154
22.222
16.667
15.385
11.111
7.6923
5.5556
player
JoeDiMaggio
MickeyMantle



Joe Dimaggio home runs versus season

The GLM Procedure

Dependent Variable: runs



Joe Dimaggio home runs versus season

The GLM Procedure

Dependent Variable: runs

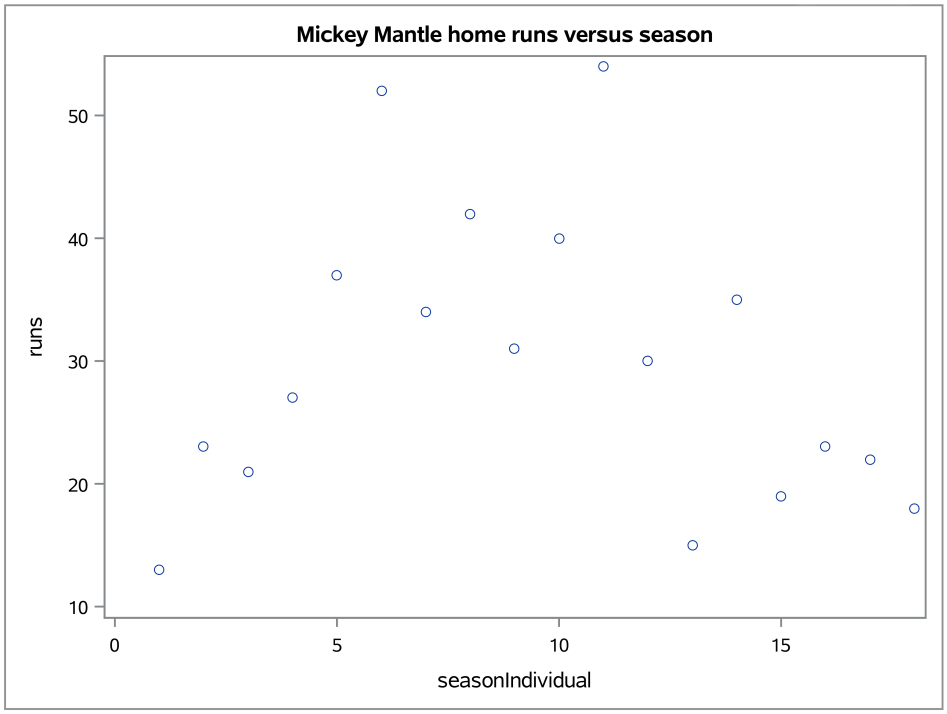
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	327.192807	163.596404	2.21	0.1607
Error	10	741.114885	74.111489		
Corrected Total	12	1068.307692			

R-Square	Coeff Var	Root MSE	runs Mean
0.306272	31.00123	8.608803	27.76923

Source	DF	Type I SS	Mean Square	F Value	Pr > F
seasonIndividual	1	327.1208791	327.1208791	4.41	0.0620
seasonInd*seasonIndi	1	0.0719281	0.0719281	0.00	0.9758

Source	DF	Type III SS	Mean Square	F Value	Pr > F
seasonIndividual	1	15.27519823	15.27519823	0.21	0.6595
seasonInd*seasonIndi	1	0.07192807	0.07192807	0.00	0.9758

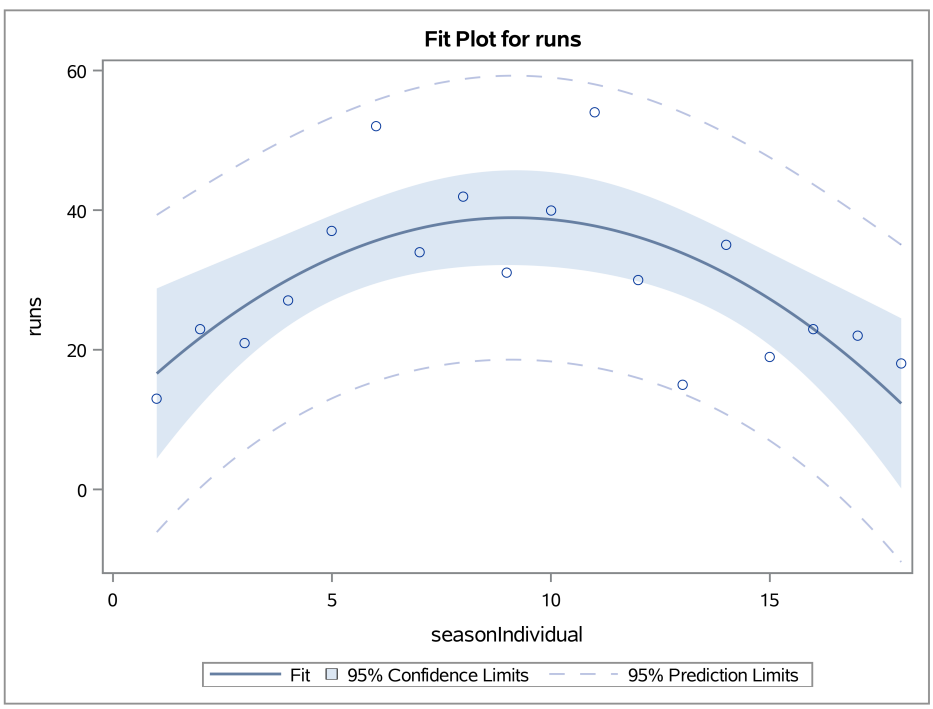
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	36.94405594	8.42626344	4.38	0.0014
seasonIndividual	-1.25674326	2.76819006	-0.45	0.6595
seasonInd*seasonIndi	-0.00599401	0.19240251	-0.03	0.9758



Mickey Mantle home runs versus season

The GLM Procedure

Dependent Variable: runs



Mickey Mantle home runs versus season

The GLM Procedure

Dependent Variable: runs

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1211.387685	605.693842	7.49	0.0056
Error	15	1213.723426	80.914895		
Corrected Total	17	2425.111111			

R-Square	Coeff Var	Root MSE	runs Mean
0.499518	30.20800	8.995271	29.77778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
seasonIndividual	1	30.720330	30.720330	0.38	0.5470
seasonInd*seasonIndi	1	1180.667355	1180.667355	14.59	0.0017

Source	DF	Type III SS	Mean Square	F Value	Pr > F
seasonIndividual	1	1029.075528	1029.075528	12.72	0.0028
seasonInd*seasonIndi	1	1180.667355	1180.667355	14.59	0.0017

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	10.76470588	7.13921922	1.51	0.1524
seasonIndividual	6.16976264	1.73005165	3.57	0.0028
seasonInd*seasonIndi	-0.33797730	0.08847855	-3.82	0.0017

For each of the following examples:

a. Describe the unit and give the number of units in the entire group.

b. Describe the variable(s) and indicate the possible values of the variable(s).

c. Classify each variable by type (nominal qualitative, ordinal qualitative, discrete quantitative, or continuous quantitative) and if there is more than one variable identify the explanatory variable and the response variable.

a) What is a unit?

b) If there is more than one group in the example, what is the explanatory (grouping) variable which determines the groups and what are its possible values?

c) What is the response variable in the example and what are its possible values?

d) Is the response variable nominal qualitative, ordinal qualitative, discrete quantitative, or continuous quantitative?

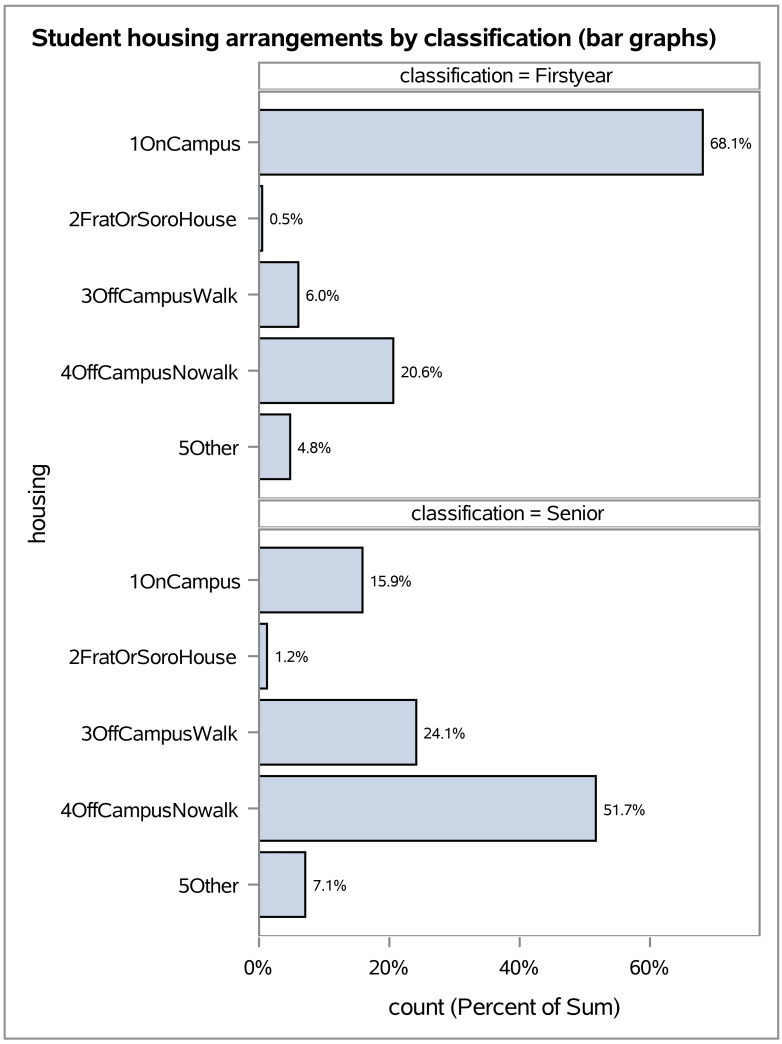
Example. Living arrangements of college students. One of the questions on the 2017 National Survey of Student Engagement is: Which of the following best describes where you are living while attending college?

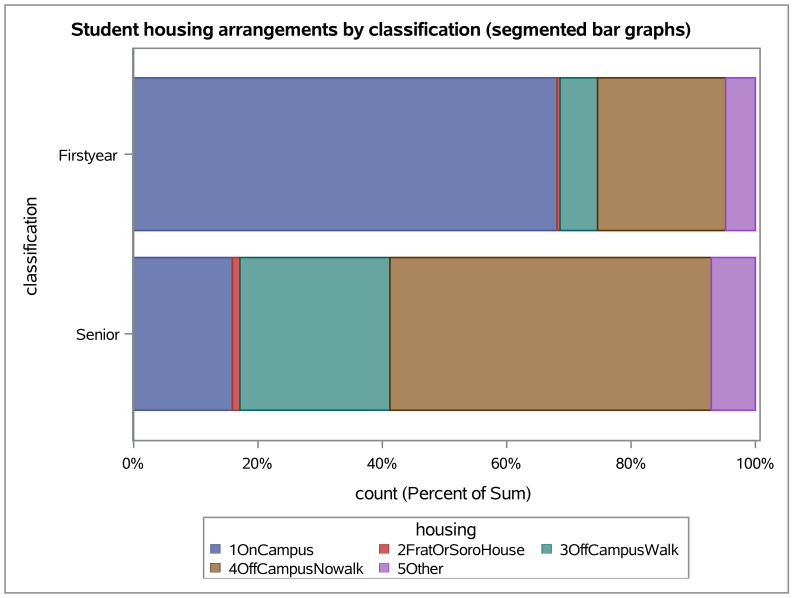
i) Residence hall, dormitory, or other campus housing (not fraternity or sorority house)
ii) Fraternity or sorority house
iii) Residence (house, apartment, etc.) within walking distance to the institution
iv) Residence (house, apartment, etc.) farther than walking distance to the institution
v) None of the above

Living arrangements of college students.

	first–year students		**seniors**	
response	**count**	**percent**	**count**	**percent**
i) On campus housing	86318	68.1	26888	15.9
ii) Fraternity or sorority house	610	.5	2067	1.2
iii) Off campus – within walking	7647	6.0	40866	24.1
iv) Off campus – beyond walking	26115	20.6	87535	51.7
v) None of the above	6055	4.8	12012	7.1
Total	126745	100	169368	100

Note: Through its student survey, The College Student Report, the National Survey of Student Engagement (NSSE) annually collects information at hundreds of four-year colleges and universities about first-year and senior students' participation in programs and activities that institutions provide for their learning and personal development. The results provide an estimate of how undergraduates spend their time and what they gain from attending college. In 2017, 725 colleges and universities participated and 517,850 students completed the survey.


Student housing arrangements by classification (bar graphs)
classification = Firstyear
1OnCampus
2FratOrSoroHouse
3OffCampusWalk
4OffCampusNowalk
5Other
68.1%
0.5%
6.0%
20.6%
4.8%
classification = Senior
15.9%
1.2%
24.1%
51.7%
7.1%
housing
0%
20%
40%
60%
count (Percent of Sum)


Student housing arrangements by classification (segmented bar graphs)
Firstyear
Senior
classification
0%
20%
40%
60%
80%
100%
count (Percent of Sum)
housing
1OnCampus
2FratOrSoroHouse
3OffCampusWalk
4OffCampusNowalk
5Other

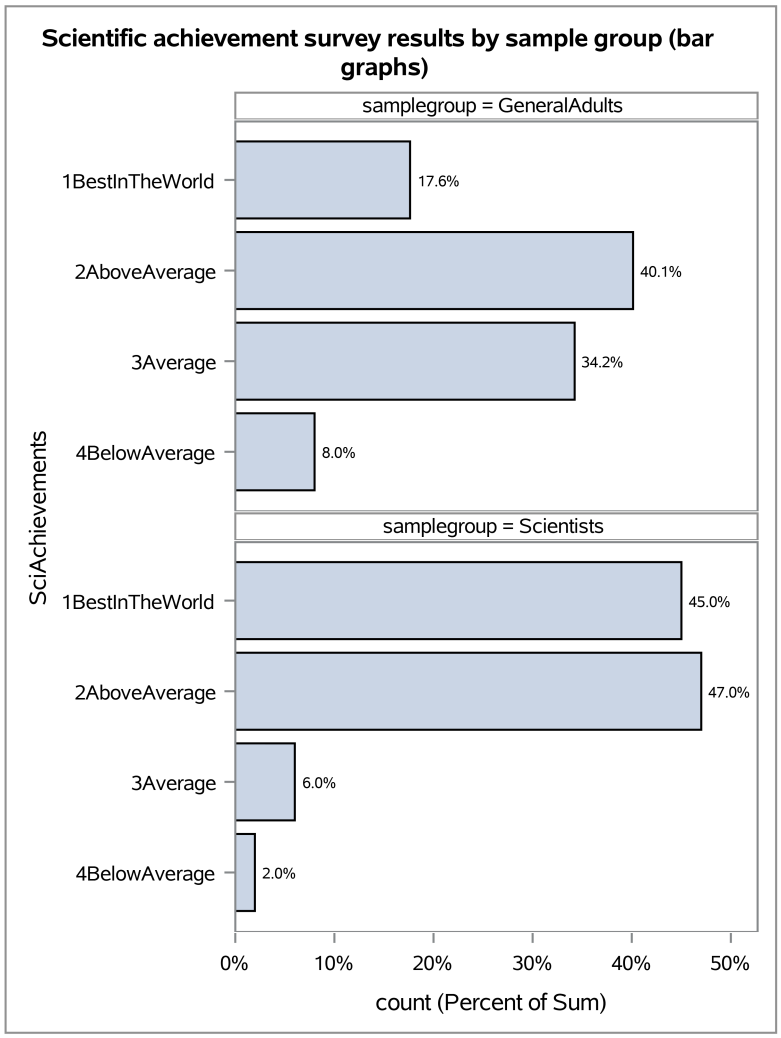
Example. Opinions regarding United States scientific achievements. The data which follow come from a Pew Research Center general public science survey conducted August 15–25, 2014 and an AAAS survey of scientists conducted September 11 – October 13, 2014. The question considered here is:

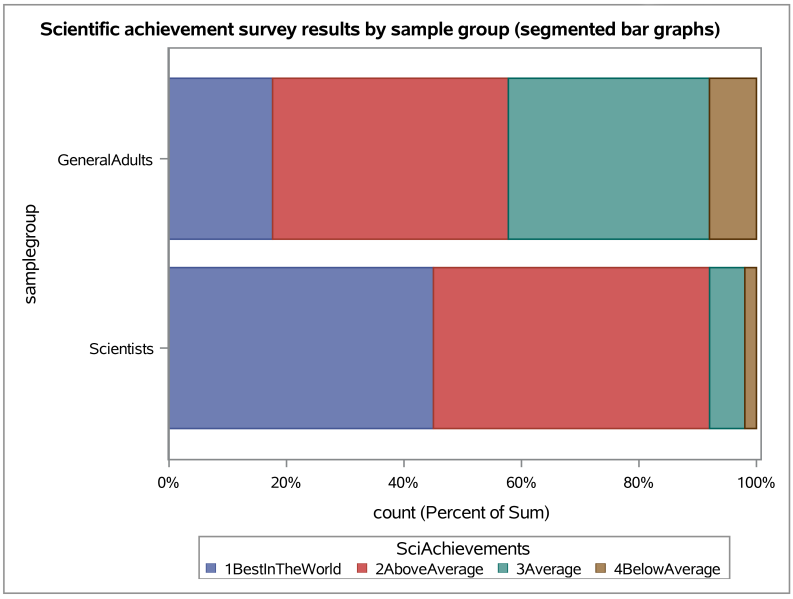
Do you think the U.S. is the best in the world, above average, average, or below average in its scientific achievements compared to other industrialized countries?

Opinions regarding United States scientific achievements.

	U.S. Adults		**AAAS Scientists**	
response	**count**	**percent**	**count**	**percent**
Best in the world	344	18	1687	45
Above average	783	40	1762	47
Average	668	34	225	6
Below average	156	8	74	2
Total	1951	100	3748	100

Note: The Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world. It conducts public opinion polling, demographic research, media content analysis and other empirical social science research. Pew Research Center does not take policy positions. It is a subsidiary of The Pew Charitable Trusts. The American Association for the Advancement of Science (AAAS) is the world's largest multidisciplinary scientific professional society. Established in 1848, the AAAS publishes Science magazine, one of the most widely circulated peer-reviewed scientific journals in the world. It is an open-member organization that brings together a wide segment of the scientific community.

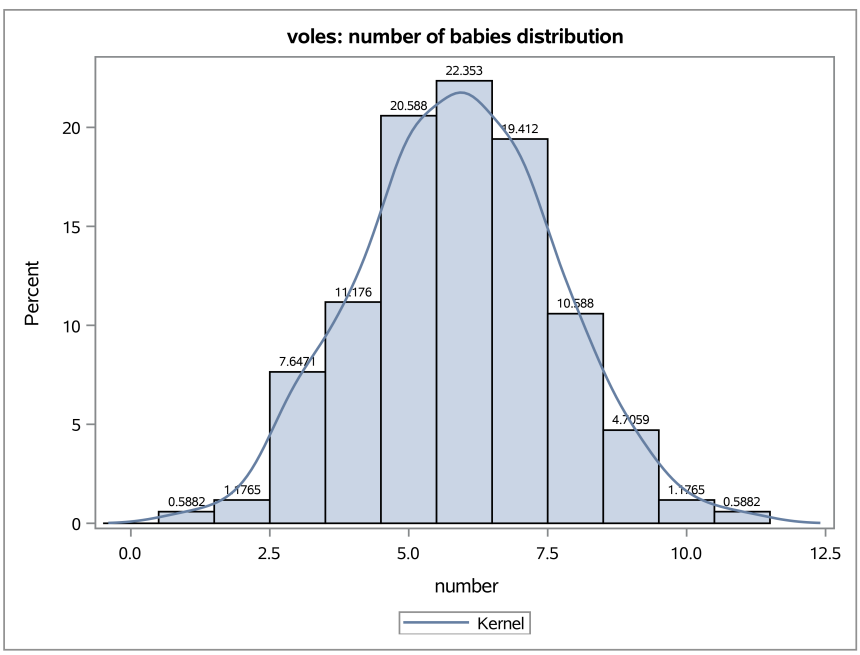

Scientific achievement survey results by sample group (bar graphs)
samplegroup = GeneralAdults
1BestInTheWorld
2AboveAverage
3Average
4BelowAverage
17.6%
40.1%
34.2%
8.0%
samplegroup = Scientists
45.0%
47.0%
6.0%
2.0%
SciAchievements
0%
10%
20%
30%
40%
50%
count (Percent of Sum)


Scientific achievement survey results by sample group (segmented bar graphs)
GeneralAdults
Scientists
samplegroup
0%
20%
40%
60%
80%
100%
count (Percent of Sum)
SciAchievements
1BestInTheWorld
2AboveAverage
3Average
4BelowAverage

Example. Vole reproduction. An investigation was conducted to study reproduction in laboratory colonies of voles. This example is taken from Devore and Peck, *Statistics*, (1997), 33; the original reference is the article "Reproduction in laboratory colonies of voles", *Oikos*, (1983), 184. The data summarized in the table below are the numbers of babies in 170 litters born to voles in a particular laboratory.

Vole baby frequency distribution.

number of babies	frequency
1	1
2	2
3	13
4	19
5	35
6	38
7	33
8	18
9	8
10	2
11	1
total	170



vole example number of babies in a litter

The FREQ Procedure

number	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	0.59	1	0.59
2	2	1.18	3	1.76
3	13	7.65	16	9.41
4	19	11.18	35	20.59
5	35	20.59	70	41.18
6	38	22.35	108	63.53
7	33	19.41	141	82.94
8	18	10.59	159	93.53
9	8	4.71	167	98.24
10	2	1.18	169	99.41
11	1	0.59	170	100.00

vole example number of babies in a litter

The MEANS Procedure

Analysis Variable : number									
N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Quartile Range	Mean	Std Dev
170	1.0000	5.0000	6.0000	7.0000	11.0000	10.0000	2.0000	5.8882	1.7488

Analysis Variable : number												
1st Pctl	5th Pctl	10th Pctl	20th Pctl	30th Pctl	40th Pctl	50th Pctl	60th Pctl	70th Pctl	80th Pctl	90th Pctl	95th Pctl	99th Pctl
2.0000	3.0000	4.0000	4.0000	5.0000	5.0000	6.0000	6.0000	7.0000	7.0000	8.0000	9.0000	10.0000

Example. Homophone confusion and Alzheimer's disease. A study was conducted to investigate the relationship between Alzheimer's disease and homophone spelling confusion. A homophone pair is a pair of words with the same pronunciation having different meanings and spellings. Twenty patients with Alzheimer's disease were asked to spell 24 homophone pairs (given in random order) and the number of homophone confusions, *e.g.* spelling doe given the context bake bread dough, was recorded for each patient. One year later, the same patients were again asked to spell the same 24 homophone pairs and the number of homophone confusions was again recorded. The data given below are the numbers of homophone confusions at the two times of measurement for the 20 Alzheimer's patients. This example comes from Neils, J., D.P. Roeltgen, and F. Constantinidou, "Decline in homophone spelling associated with loss of semantic influence on spelling in Alzheimer's disease", *Brain and Language*, **49**, (1995).

Alzheimer's homophone confusion data.

Patient	**Time 1**	**Time 2**	**Patient**	**Time 1**	**Time 2**
1	5	5	11	7	10
2	1	3	12	0	3
3	0	0	13	3	9
4	1	1	14	5	8
5	0	1	15	7	12
6	2	1	16	10	16
7	5	6	17	5	5
8	1	2	18	6	3
9	0	9	19	9	6
10	5	8	20	11	8

homophone confusions difference distribution

The FREQ Procedure

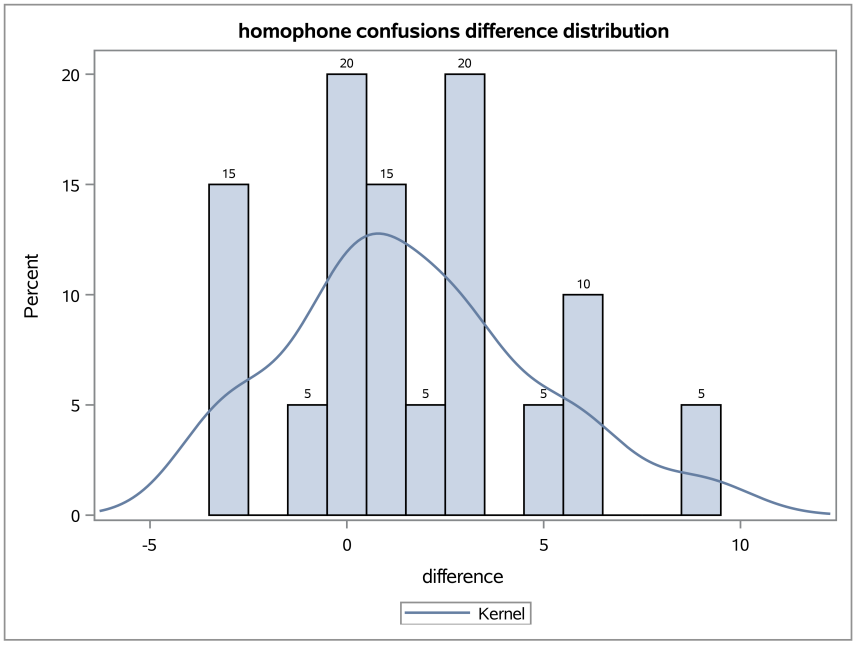
difference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-3	3	15.00	3	15.00
-1	1	5.00	4	20.00
0	4	20.00	8	40.00
1	3	15.00	11	55.00
2	1	5.00	12	60.00
3	4	20.00	16	80.00
5	1	5.00	17	85.00
6	2	10.00	19	95.00
9	1	5.00	20	100.00

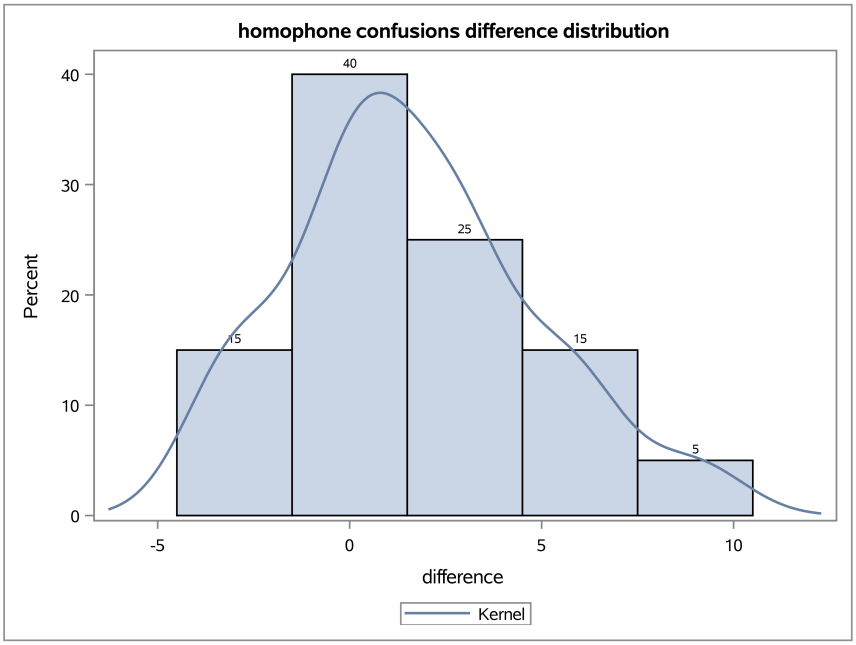
homophone confusions difference distribution

The MEANS Procedure

Analysis Variable : difference									
N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Quartile Range	Mean	Std Dev
20	-3.0000	0.0000	1.0000	3.0000	9.0000	12.0000	3.0000	1.6500	3.1999

Analysis Variable : difference												
1st Pctl	5th Pctl	10th Pctl	20th Pctl	30th Pctl	40th Pctl	50th Pctl	60th Pctl	70th Pctl	80th Pctl	90th Pctl	95th Pctl	99th Pctl
-3.0000	-3.0000	-3.0000	-0.5000	0.0000	0.5000	1.0000	2.5000	3.0000	4.0000	6.0000	7.5000	9.0000


homophone confusions difference distribution
Percent
20
15
10
5
0
15
5
20
15
5
20
5
10
5
-5
0
5
10
difference
Kernel


homophone confusions difference distribution
Percent
40
30
20
10
0
15
40
25
15
5
-5
0
5
10
difference
Kernel

Example. Wooly–bear caterpillar cocoons. A study was conducted to investigate the relationship between air temperature and the temperature inside a wooly–bear caterpillar cocoon. It seems quite reasonable to expect the temperature inside a cocoon to be higher than the air temperature (outside the cocoon). The data given below are pairs of air and cocoon temperatures made on 12 days at a location in the high arctic region. Each cocoon temperature is actually the average of two cocoon temperatures. This example comes from Kevan, P.C., T.S. Jensen, and J.D. Shorthouse, "Body temperatures and behaviorial thermoregulation of high arctic wooly–bear caterpillars and pupae (*Gynaephora rossii,* Lymantridae: Lepidoptera) and the importance of sunshine", *Arctic and Alpine Research*, **14**, (1982).

Wooly–bear temperature data.

Day	**Cocoon temp**	**Air temp**	**Day**	**Cocoon temp**	**Air temp**
1	15.1	10.4	7	3.6	1.7
2	14.6	9.2	8	5.3	2.0
3	6.8	2.2	9	7.0	3.0
4	6.8	2.6	10	7.1	3.5
5	8.0	4.1	11	9.6	4.5
6	8.7	3.7	12	9.5	4.4

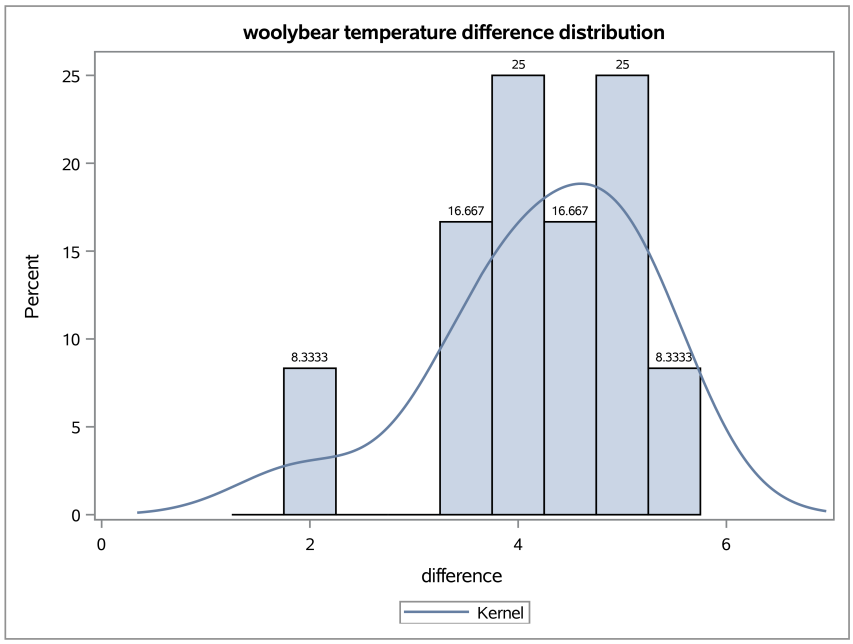
woolybear temperature difference distribution

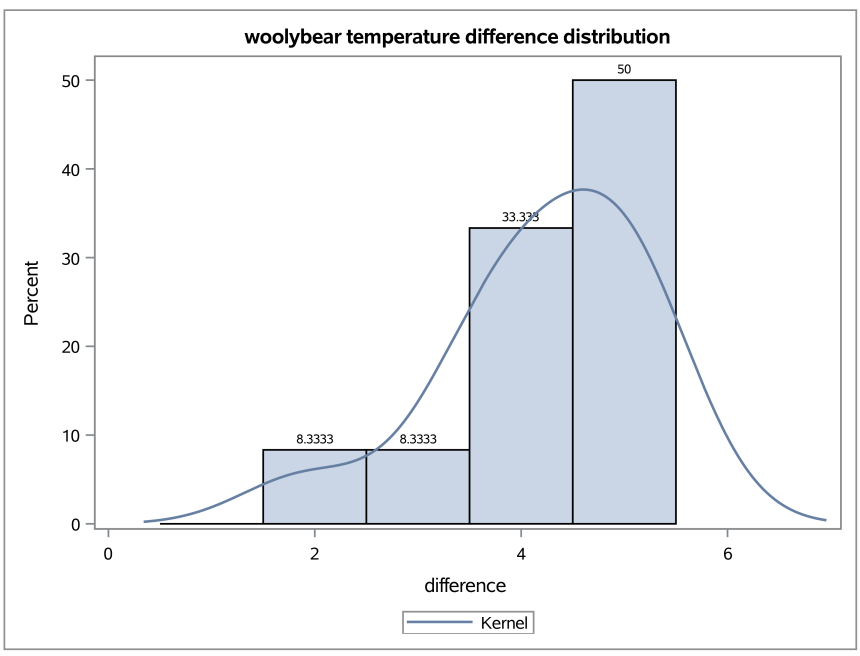
The FREQ Procedure

difference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.9	1	8.33	1	8.33
3.3	1	8.33	2	16.67
3.6	1	8.33	3	25.00
3.9	1	8.33	4	33.33
4	1	8.33	5	41.67
4.2	1	8.33	6	50.00
4.6	1	8.33	7	58.33
4.7	1	8.33	8	66.67
5	1	8.33	9	75.00
5.1	2	16.67	11	91.67
5.4	1	8.33	12	100.00

Analysis Variable : difference									
N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Quartile Range	Mean	Std Dev
12	1.9000	3.7500	4.4000	5.0500	5.4000	3.5000	1.3000	4.2333	0.9857

Analysis Variable : difference												
1st Pctl	5th Pctl	10th Pctl	20th Pctl	30th Pctl	40th Pctl	50th Pctl	60th Pctl	70th Pctl	80th Pctl	90th Pctl	95th Pctl	99th Pctl
1.9000	1.9000	3.3000	3.6000	3.9000	4.0000	4.4000	4.7000	5.0000	5.1000	5.1000	5.4000	5.4000


woolybear temperature difference distribution
Percent
25
20
15
10
5
0
8.3333
16.667
25
16.667
25
8.3333
0
2
4
6
difference
Kernel


woolybear temperature difference distribution
Percent
50
40
30
20
10
0
8.3333
8.3333
33.333
50
0
2
4
6
difference
Kernel

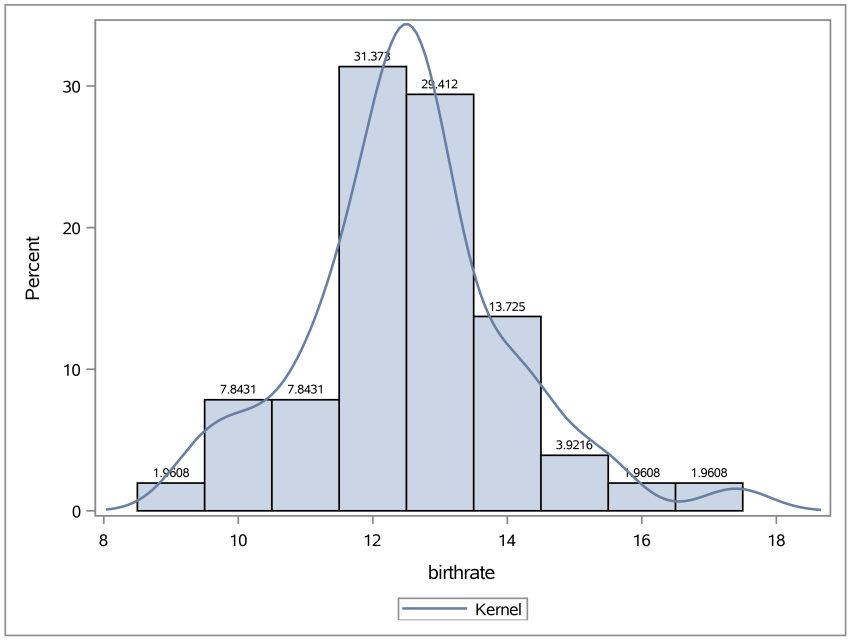
Example. Birth rates for the United States in 2014. The *National Vital Statistics Report* (Vol. 64, No. 12, December 23, 2015) presents 2014 data on U.S. births according to a wide variety of characteristics. The data used in this example are based on 100% of the birth certificates filed in all states and the District of Columbia (DC). These data are provided to the Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS) through the Vital Statistics Cooperative Program (VSCP).

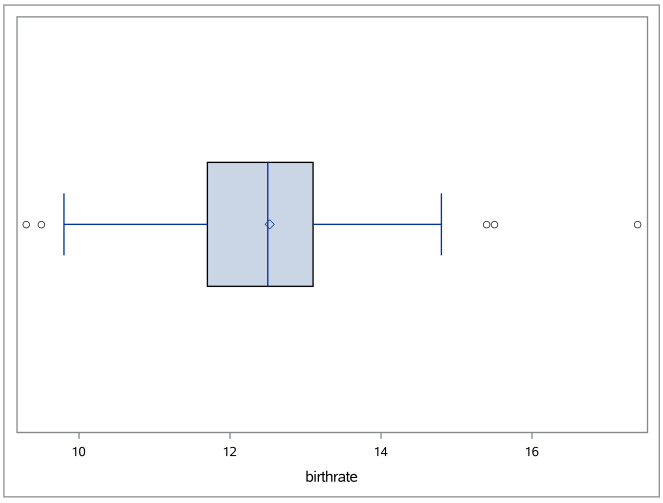
Birthrates 2014 (births per 1,000 women)

state	birthrate	state	birthrate
Alabama	12.3	Montana	12.1
Alaska	15.5	Nebraska	14.2
Arizona	12.9	Nevada	12.6
Arkansas	13.0	New Hampshire	9.3
California	13.0	New Jersey	11.6
Colorado	12.3	New Mexico	12.5
Connecticut	10.1	New York	12.1
Delaware	11.7	North Carolina	12.2
District of Columbia	14.4	North Dakota	15.4
Florida	11.1	Ohio	12.0
Georgia	13.0	Oklahoma	13.8
Hawaii	13.1	Oregon	11.5
Idaho	14.0	Pennsylvania	11.1
Illinois	12.3	Rhode Island	10.3
Indiana	12.7	South Carolina	11.9
Iowa	12.8	South Dakota	14.4
Kansas	13.5	Tennessee	12.5
Kentucky	12.7	Texas	14.8
Louisiana	13.9	Utah	17.4
Maine	9.5	Vermont	9.8
Maryland	12.4	Virginia	12.4
Massachusetts	10.7	Washington	12.5
Michigan	11.5	West Virginia	11.0
Minnesota	12.8	Wisconsin	11.7
Mississippi	12.9	Wyoming	13.2
Missouri	12.4		

birthrate relative frequency distributions
binstart is the left endpoint of the interval -- left endpoint is included in frequency
these intervals are the same as those of the histograms

Obs	binstart	frequency	percentage
1	8.5	1	1.9608
2	9.5	4	7.8431
3	10.5	4	7.8431
4	11.5	16	31.3725
5	12.5	15	29.4118
6	13.5	7	13.7255
7	14.5	2	3.9216
8	15.5	1	1.9608
9	16.5	1	1.9608





Analysis Variable : birthrate									
N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Quartile Range	Mean	Std Dev
51	9.3000	11.7000	12.5000	13.1000	17.4000	8.1000	1.4000	12.5255	1.5341

Analysis Variable : birthrate												
1st Pctl	5th Pctl	10th Pctl	20th Pctl	30th Pctl	40th Pctl	50th Pctl	60th Pctl	70th Pctl	80th Pctl	90th Pctl	95th Pctl	99th Pctl
9.3000	9.8000	10.7000	11.5000	12.0000	12.3000	12.5000	12.7000	13.0000	13.5000	14.4000	15.4000	17.4000