A Little Stats Won't Hurt You

Nate Derby

Statis Pro Data Analytics
Seattle, WA, USA

Edmonton SAS Users Group, 11/13/09

Outline

What Can Statistical Methods Tell Us?

- Statistics can describe data, extract information from them:
  - Test hypotheses ("Is X correlated with Y?").
  - Extrapolate trends for forecasts ("What will X be in n weeks?").
  - Quantify what happened ("What is effect of X on Y?").
- First step: *Look at the data*, look for "interesting" features.
  - Meaning of "interesting" depends on context:
    - Preliminary steps: "Should this data point be included?"
    - More advanced stages: "If we change the data, how does that change the results?"
  - "If it looks interesting, it's probably wrong."
    - Investigate to determine if really "interesting" or just wrong.

Exploratory Data Analysis (EDA)

EDA = Looking at data to see what the data are telling us.

- Tells us which variables to look at, what models to use, etc.
- Prerequisites for more advanced methods:
  - ANOVA (`PROC ANOVA`, `PROC GLM`)
  - Linear/logistic regression (`PROC REG`, `PROC LOGISTIC`)
  - ARIMA (`PROC ARIMA`)
- We will focus on *univariate* methods.
  - Look at one variable at a time.
  - "How is that variable *distributed* (spread out) within our data set?"

Two Methods

- Data Visualization
  - Graphical techniques to quickly/easily see general trends.
  - Keep it simple! (Like/unlike a dashboard)
- Descriptive Statistics
  - Statistical measures to summarize characteristics of distribution.
  - Keep it short!

Example Data Sets

Six sets of 50 measurements of systolic blood pressure (mmHg):

113	122	106	131	130	112	132	122	114	117
108	103	117	120	117	126	116	128	124	123
126	143	118	110	103	119	136	109	113	116
127	97	144	108	121	128	115	115	124	115
120	98	115	107	131	126	112	118	121	126

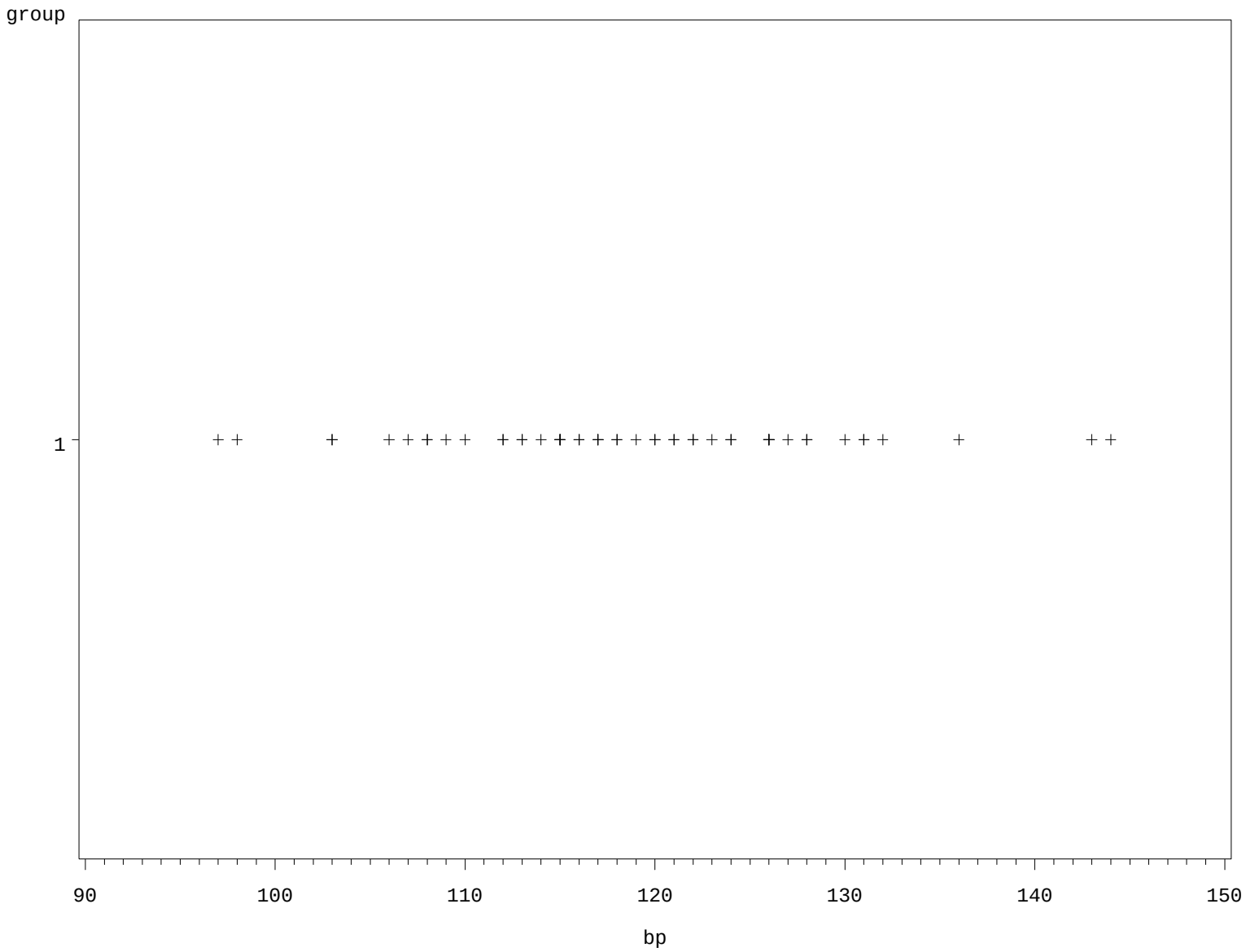
- Named `data1` - `data6`.
- Variables `bp`, `group` (values 1-6).
- Difficult to discern distribution information from the raw data!
  - Too much information!

Scatterplot I

First step: Look at one-dimensional *scatterplots*

Basic Scatterplot (page 2)

```
PROC GPLOT data=data1;
  PLOT group*bp;
RUN;
```


group
1
90
100
110
120
130
140
150
bp

Scatterplot II

Let's clean that up a bit:

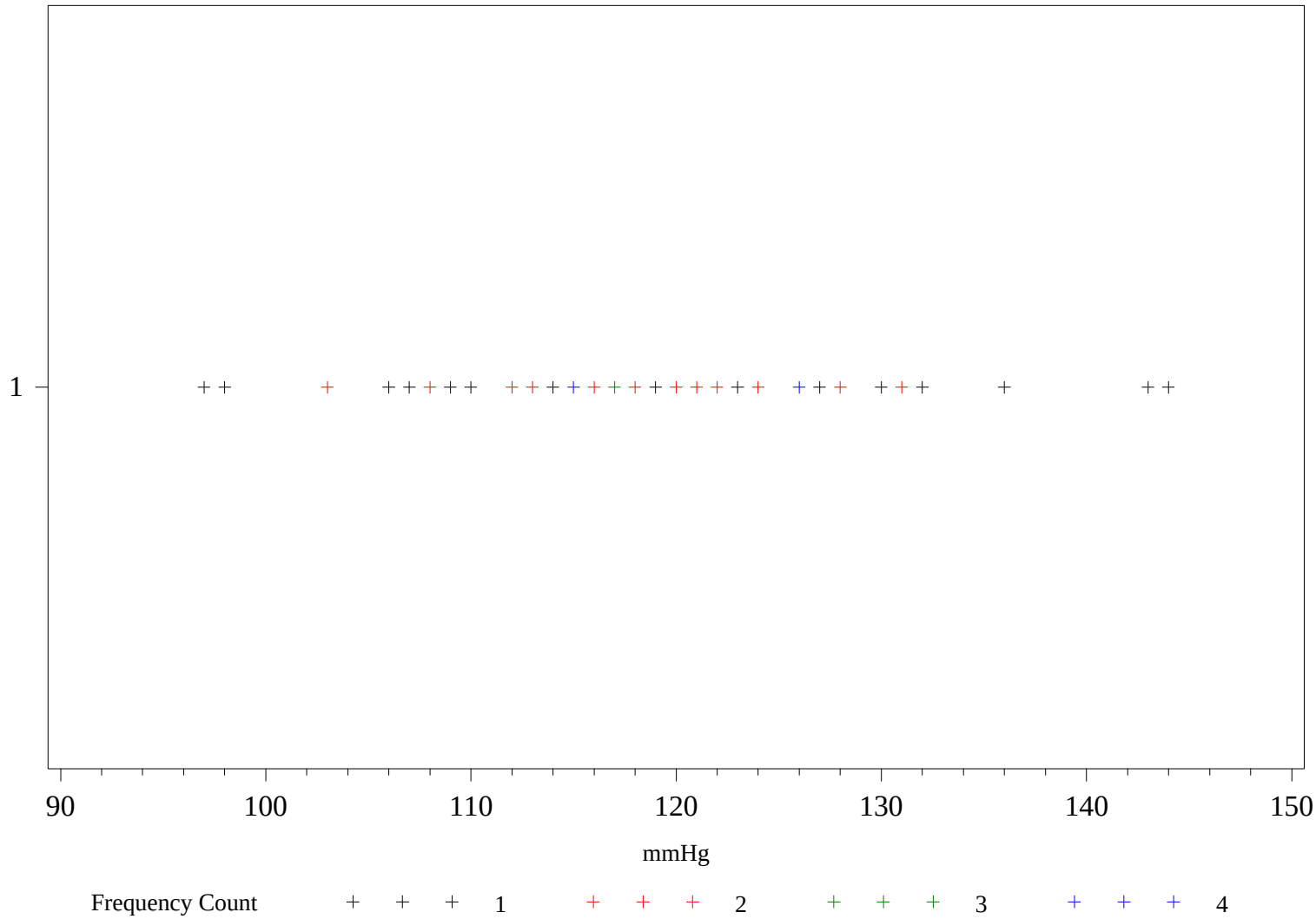
Custom-Formatted Scatterplot (page 2)

```
PROC FREQ data=data1 noprint;
  TABLES group*bp / out=data1stats ( keep = group bp count );
RUN;

goptions ftitle='Times/bold' ftext='Times';
  * Use Times New Roman font;
symbol1 c=red;
  * Plot the data points in red;
axis1 label=( 'mmHg' ) order=( 90 to 150 by 10 ) minor=( number=4 ) value=( height=1.2 );
  * Make the x-axis more readable;
axis2 label=( '' ) value=( height=1.2 );
  * Take off the label on the y-axis;
title 'Systolic Blood Pressure';
  * Give this graph a title;

PROC GPLOT data=data1stats;
  PLOT group*bp=count / haxis=axis1 vaxis=axis2;
RUN;
```

Systolic Blood Pressure



Bubble Plots

Let's try one variation:

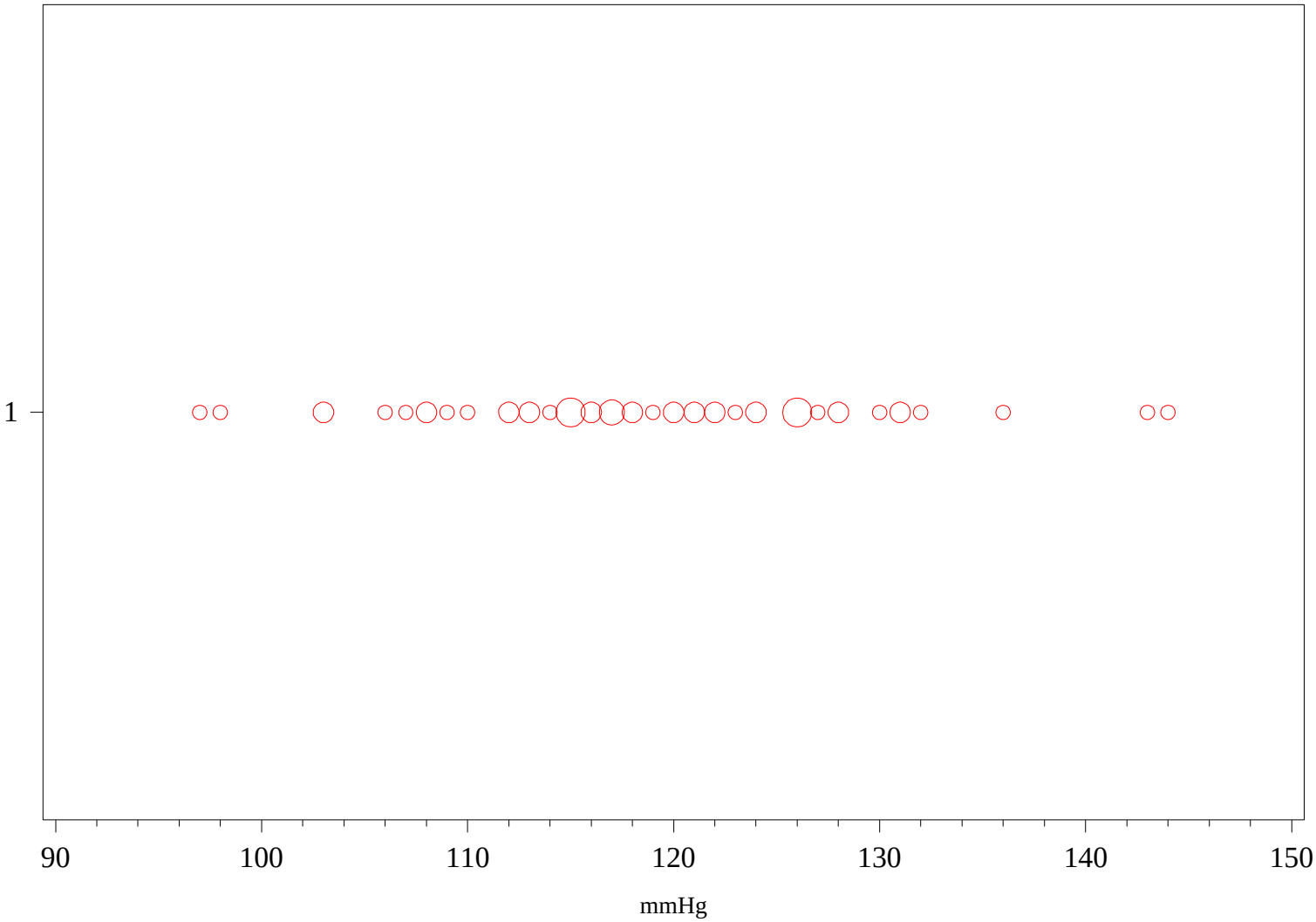
Custom-Formatted Scatterplot (page 2)

```
PROC FREQ data=data1 noprint;
  TABLES group*bp / out=data1stats ( keep = group bp count );
RUN;

goptions ftitle='Times/bold' ftext='Times';
  * Use Times New Roman font;
symbol1 c=red;
  * Plot the data points in red;
axis1 label=( 'mmHg' ) order=( 90 to 150 by 10 ) minor=( number=4 ) value=( height=1.2 );
  * Make the x-axis more readable;
axis2 label=( '' ) value=( height=1.2 );
  * Take off the label on the y-axis;
title 'Systolic Blood Pressure';
  * Give this graph a title;

PROC GPLOT data=data1stats;
  BUBBLE group*bp=count / haxis=axis1 vaxis=axis2 bcolor=red;
RUN;
```

Systolic Blood Pressure



Scatterplots and Bubble Plots: Conclusions

Scatterplots:

- Gives us a effective look at the range, distribution, outliers.
- Difficult to differentiate multiple points in the same space.

Bubble Plots:

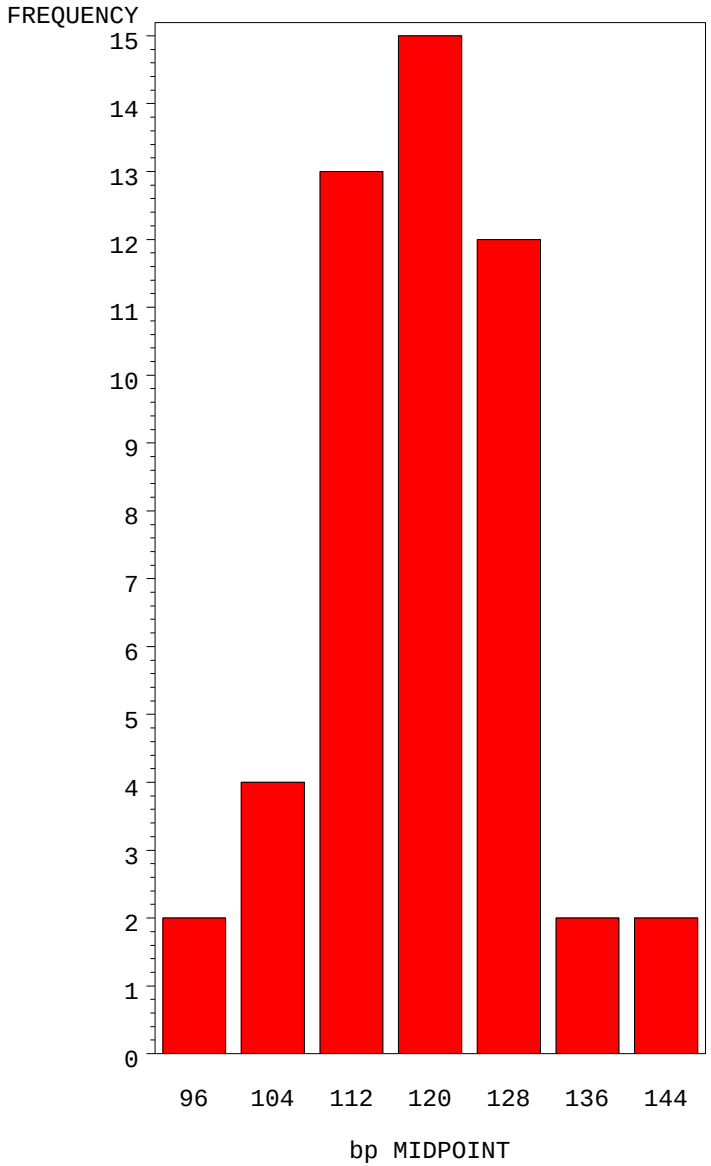
- Represents multiple points in the same space by a bubble, size relational to # points.
- More effective way to look at range, distribution, outliers.
- Other than that, not very useful.

Histogram I

Histogram: Chart of frequency/percentage counts for different ranges.

Basic Histogram (page 3)

```
PROC GPLOT data=data1;
  VBAR bp;
RUN;
```


FREQUENCY
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
96
104
112
120
128
136
144
bp MIDPOINT

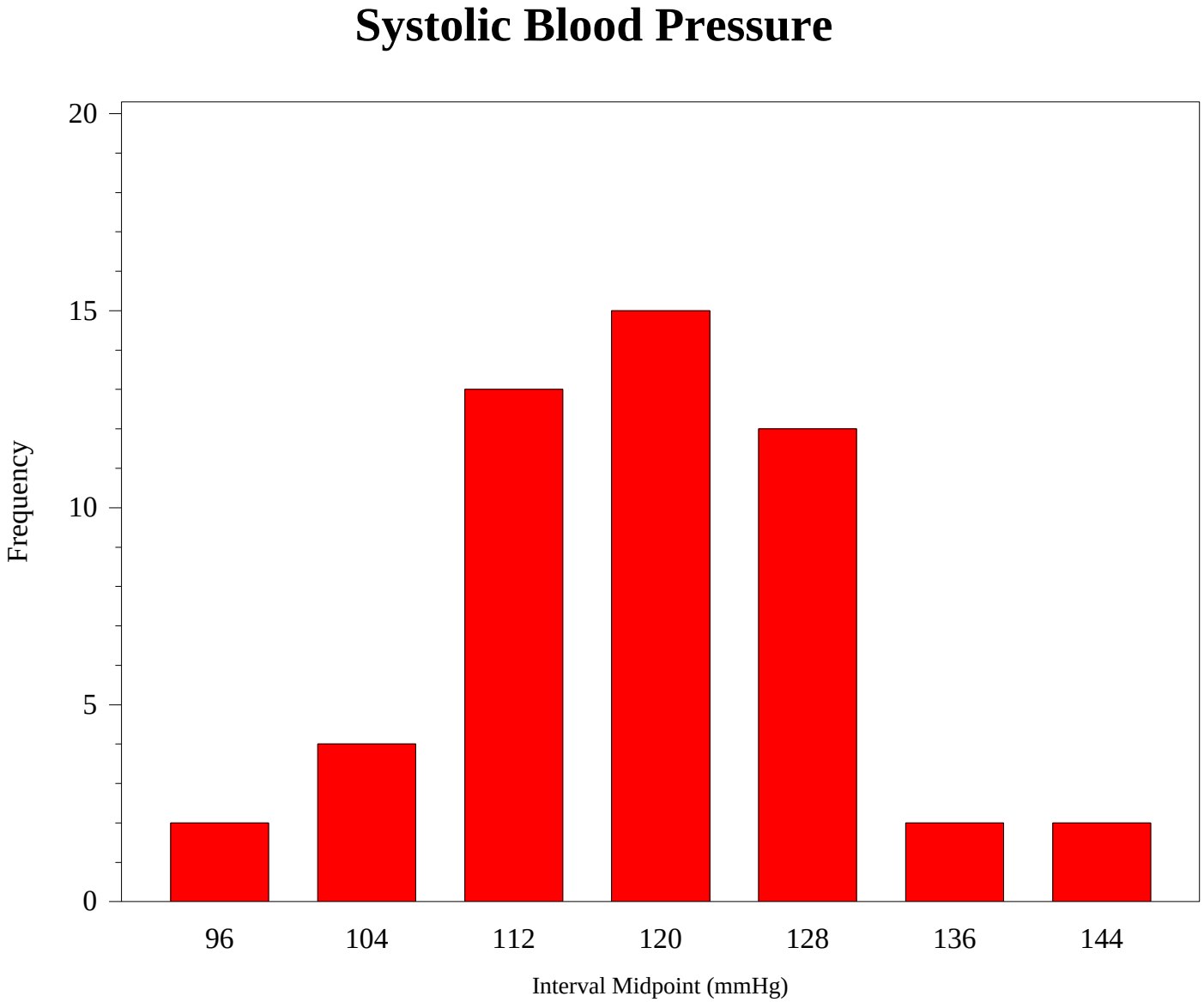
Histogram II

Again, clean it up:

Custom-Formatted Histogram (page 3)

```
goptions ftitle='Times/bold' ftext='Times';
axis1 label=( 'Interval Midpoint (mmHg)' height=1.2 )
  offset=( 4, 4 ) value=( height=1.2 );
axis2 label=( angle=90 height=1.2 'Frequency' )
  order=( 0 to 20 by 5 ) minor=( number=4 ) value=( height=1.2 );
title 'Systolic Blood Pressure';

PROC GCHART DATA=data1;
  VBAR bp / maxis=axis1 raxis=axis2 width=4 space=2;
RUN;
```


Systolic Blood Pressure
Frequency
0
5
10
15
20
96
104
112
120
128
136
144
Interval Midpoint (mmHg)

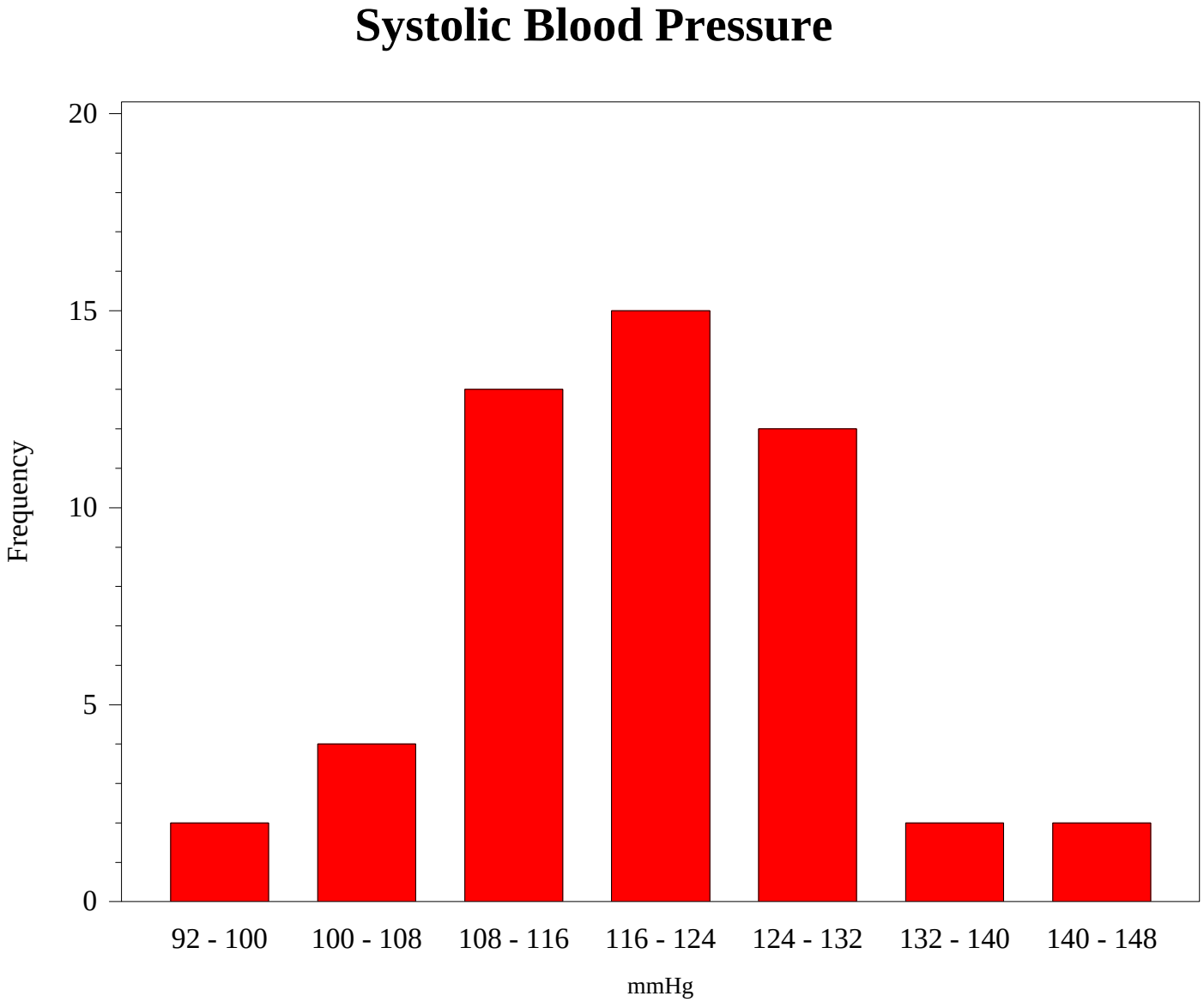
Histogram III

Still not quite right – want to make explicit ranges

Custom-Formatted Histogram, Customized Ranges (page 4)

```
axis1 label=( 'mmHg' ) value=( height=1.2 '92 - 100' '100 - 108'
  '108 - 116' '116 - 124' '124 - 132' '132 - 140' '140 - 148' )
  offset=( 4, 4 );
axis2 label=( angle=90 height=1.2 'Frequency' ) order=( 0 to 20 by 5 )
  minor=( number=4 ) value=( height=1.2 );

PROC GCHART DATA=data1;
  VBAR bp / maxis=axis1 raxis=axis2 width=4 space=2
    midpoints = 96 to 144 by 8;
RUN;
```


Systolic Blood Pressure
Frequency
20
15
10
5
0
92 - 100
100 - 108
108 - 116
116 - 124
124 - 132
132 - 140
140 - 148
mmHg

What Can We Deduce from Our Histogram?

- Data centered around 120 mmHg (easier with odd number of intervals),
- “Bulk” of data within 8-10 mmHg of 120 mmHg.

Answers to questions:

- *What is the central value of the data?*
- *How spread out are the data?*

Central Value of Data

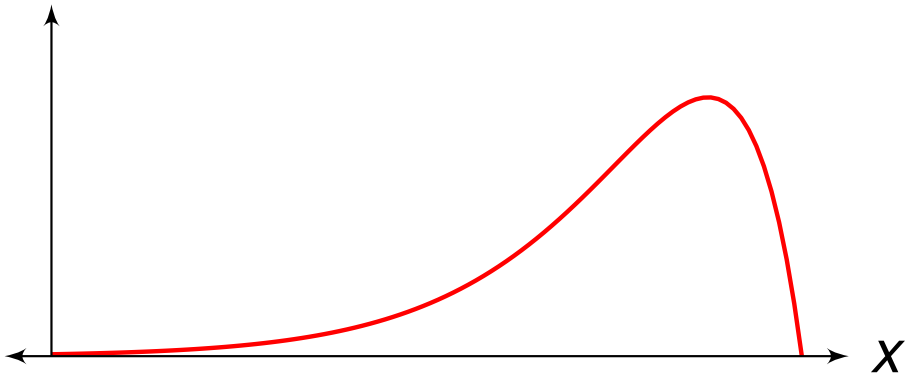
What is the central value of the data?

- *Mean*: Arithmetic average, or $\frac{x_1 + x_2 + \cdots + x_n}{n}$.
- *Median*: "Middle Value": 50% of data below this value.
- *Mode*: Value/category that is a maximum ("high point") in the distribution.

Relationship between median and mode can determine *skewness*: How lopsided the distribution is.

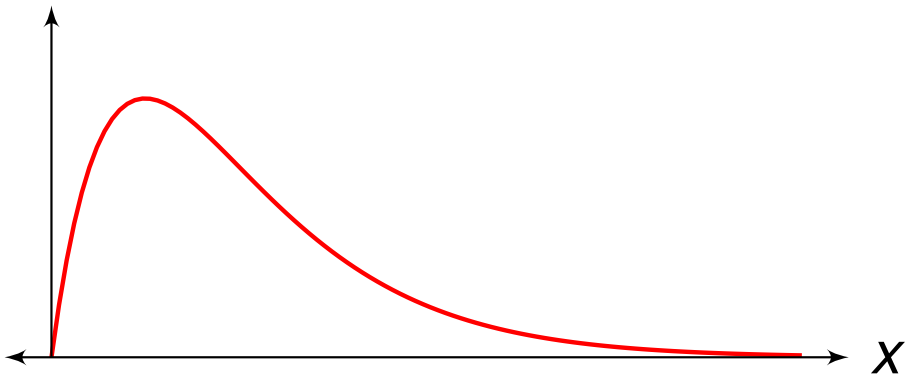
Skewness

Distribution



left-skewed (left-tailed)
mean $<$ median

Distribution



right-skewed (right-tailed)
mean $>$ median

Spread of Data

How spread out are the data?

- *Standard Deviation*: Average distance from the mean.
- *Minimum*: Smallest value
- *25th percentile*: 25% of data lie below this value.
- *75th percentile*: 75% of data lie below this value.
- *Interquartile Range*: Difference between 75th and 25th percentiles.
- *Maximum*: Largest value.

Note: Median = 50th percentile.

Number of Intervals

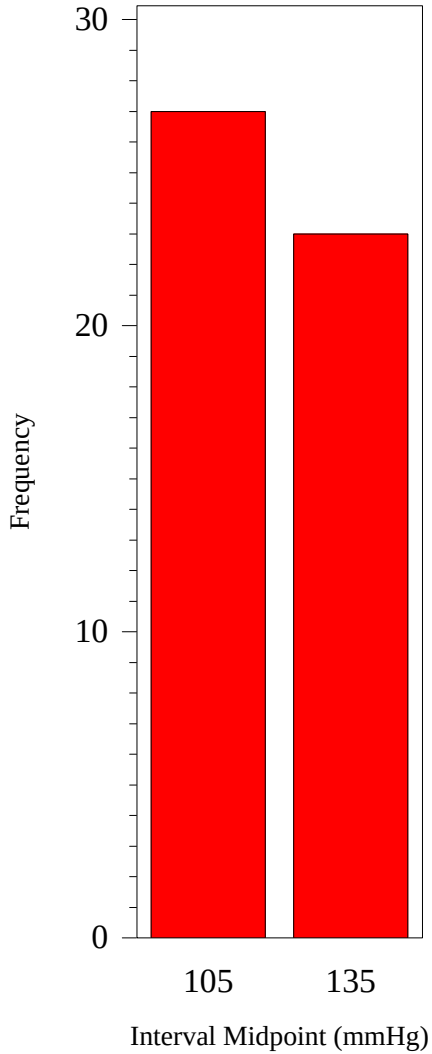
Beware of too many/few intervals!

Histogram with xx intervals (page 6)

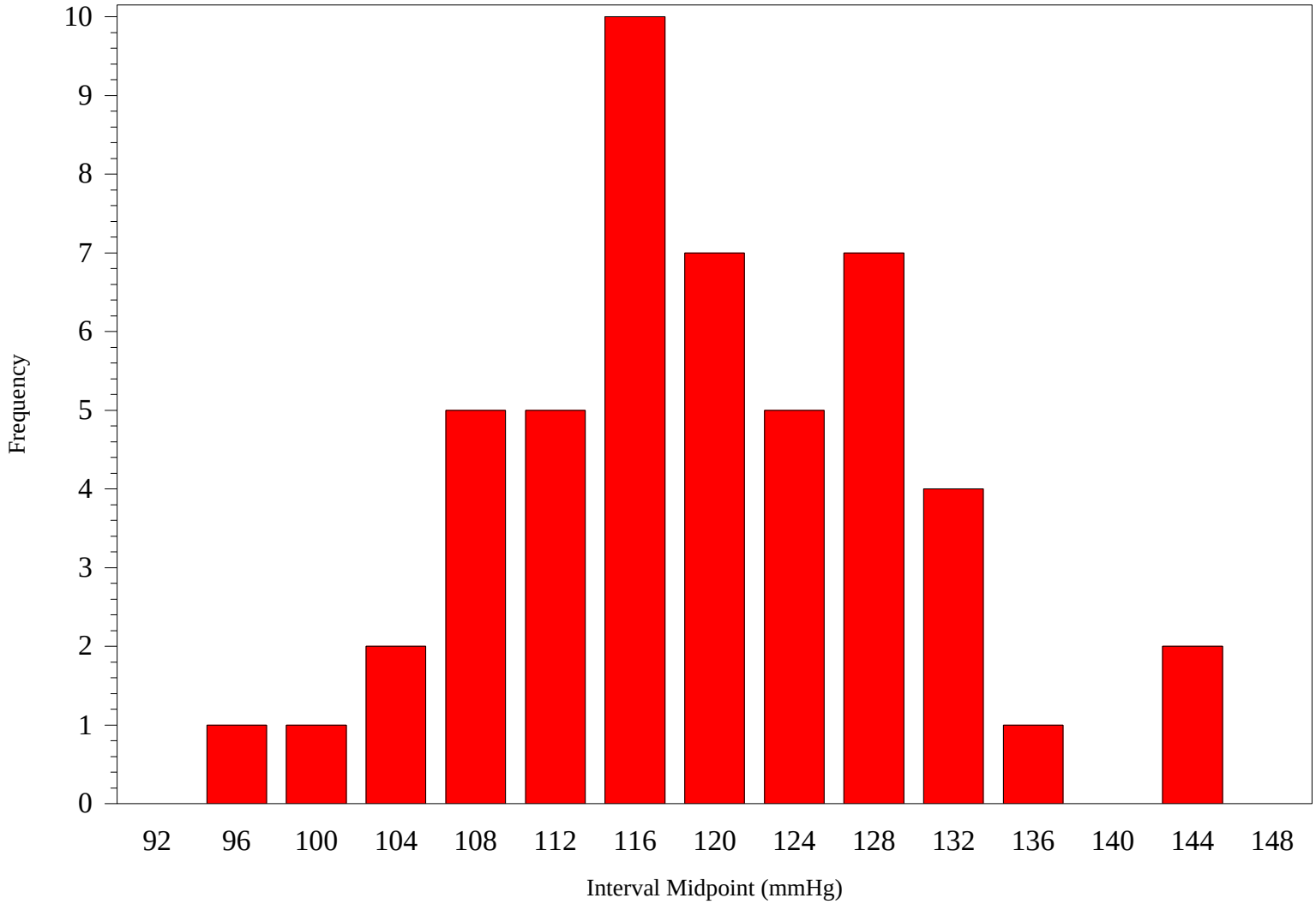
```
axis1 label=( 'Interval Midpoint (mmHg)' height=1.2 ) value=( height=1.2
axis2 label=( angle=90 'Frequency' height=1.2 ) value=( height=1.2 );
title 'Systolic Blood Pressure, xx Intervals';

PROC GCHART DATA=data1;
  VBAR bp / maxis=axis1 raxis=axis2 levels = xx;
RUN;
```

Systolic Blood Pressure, 2 Intervals



Systolic Blood Pressure, 15 Intervals



Number of Intervals

Rule of Thumb:

Number of Data Points	Number of Intervals
Under 50	5 to 7
50 to 100	6 to 10
100 to 250	7 to 12
over 250	10 to 20

Histogram: Conclusions

- Gives us a rough estimate of distribution.
- Can be misleading from too few/many intervals!
- (Can be misleading even from shifting intervals!)

Box Plot I

Box Plot: Graphical representation of six summary statistics:

- minimum,
- 25^{th} percentile,
- mean,
- 75^{th} percentile,
- maximum,
- mean.

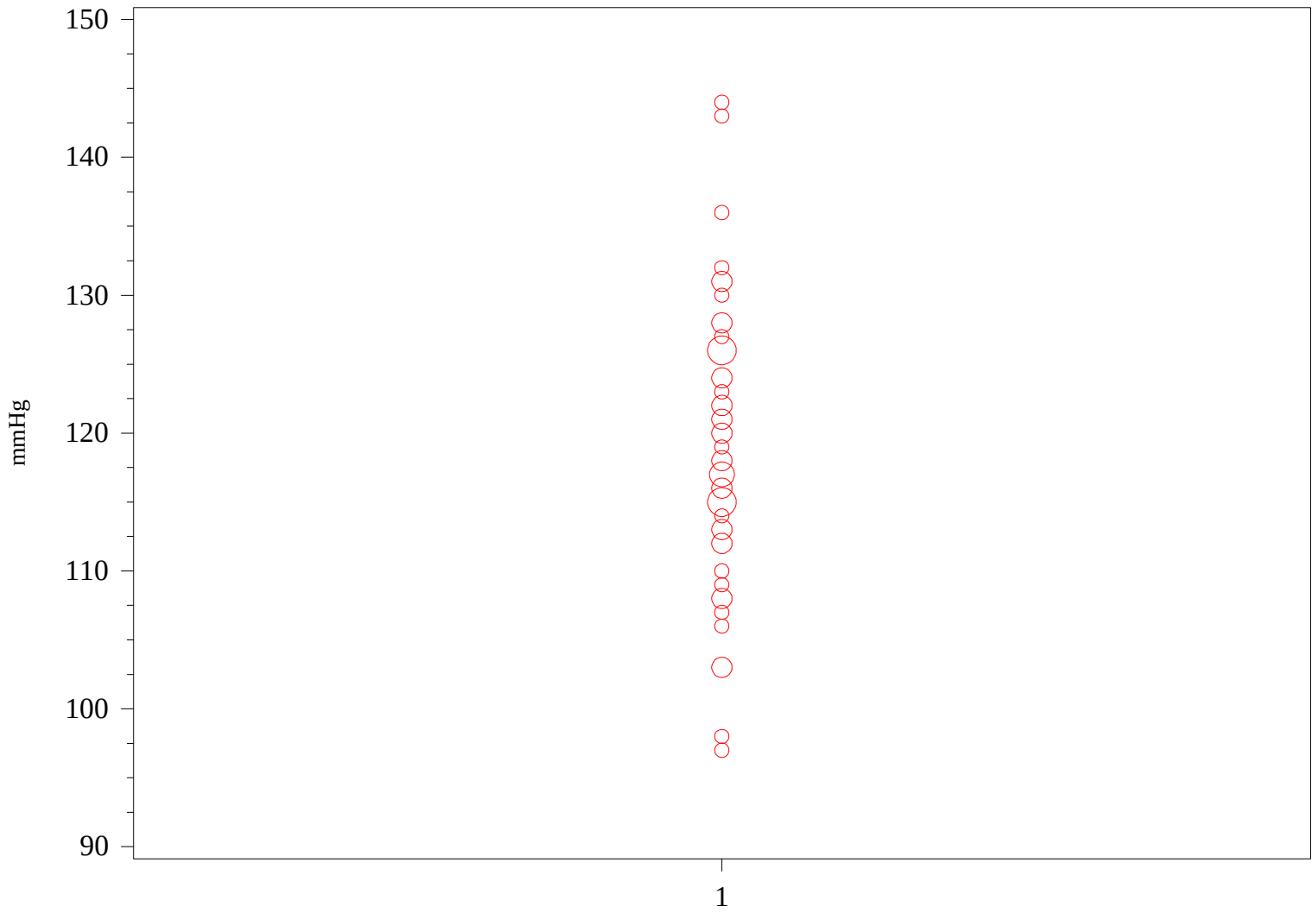
Box Plot II

Code: Similar to vertical-aligned `PROC GPLOT`

Vertical-Aligned Scatterplot (page 8)

```
goptions ftitle='Times/bold' ftext='Times';
symbol1 c=red;
axis1 label=( angle=90 'mmHg' height=1.2 )
  order=( 90 to 150 by 10 )
  minor=( number=3 ) value=( height=1.2 );
axis2 label=( '' ) value=( height=1.2 );
title 'Systolic Blood Pressure';

PROC GPLOT data=data1;
  BUBBLE bp*group=count / vaxis=axis1 haxis=axis2;
RUN;
```


Systolic Blood Pressure
150
140
130
120
110
100
90
mmHg
1

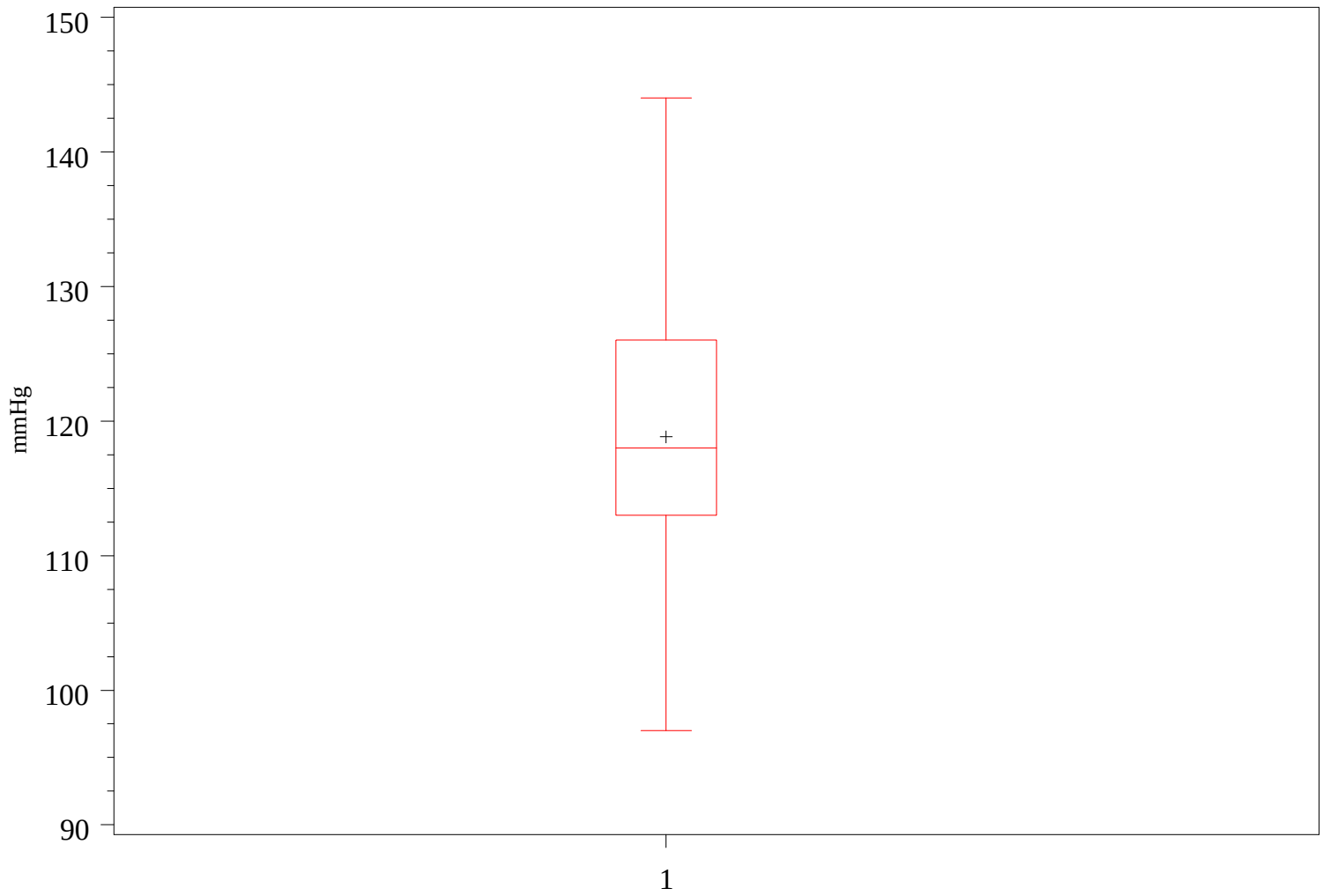
Box Plot III

Use `PROC BOXPLOT` instead of `PROC GPLOT`:

Basic Box Plot (page 9)

```
axis1 label=( 'mmHg' height=1.2 ) order=( 90 to 150 by 10 )
  minor=( number=3 ) value=( height=1.2 );
symbol1;

PROC BOXPLOT data=data1;
  PLOT bp*group / vaxis=axis1 haxis=axis2;
RUN;
```


Systolic Blood Pressure
150
140
130
120
110
100
90
mmHg
1

What Does the Box Plot Tell Us?

- "Bulk" of data between 108 mmHg and 132 mmHg, as in scatterplot.
- Now easier to see; 25^{th} and 75^{th} percentiles very clear!
- Mean $>$ Median, so definitely right-skewed.

Overall:

Box plots more reliable than histograms for percentiles or skewness.

Why? Because they incorporate *summary statistics*.

Box Plot IV

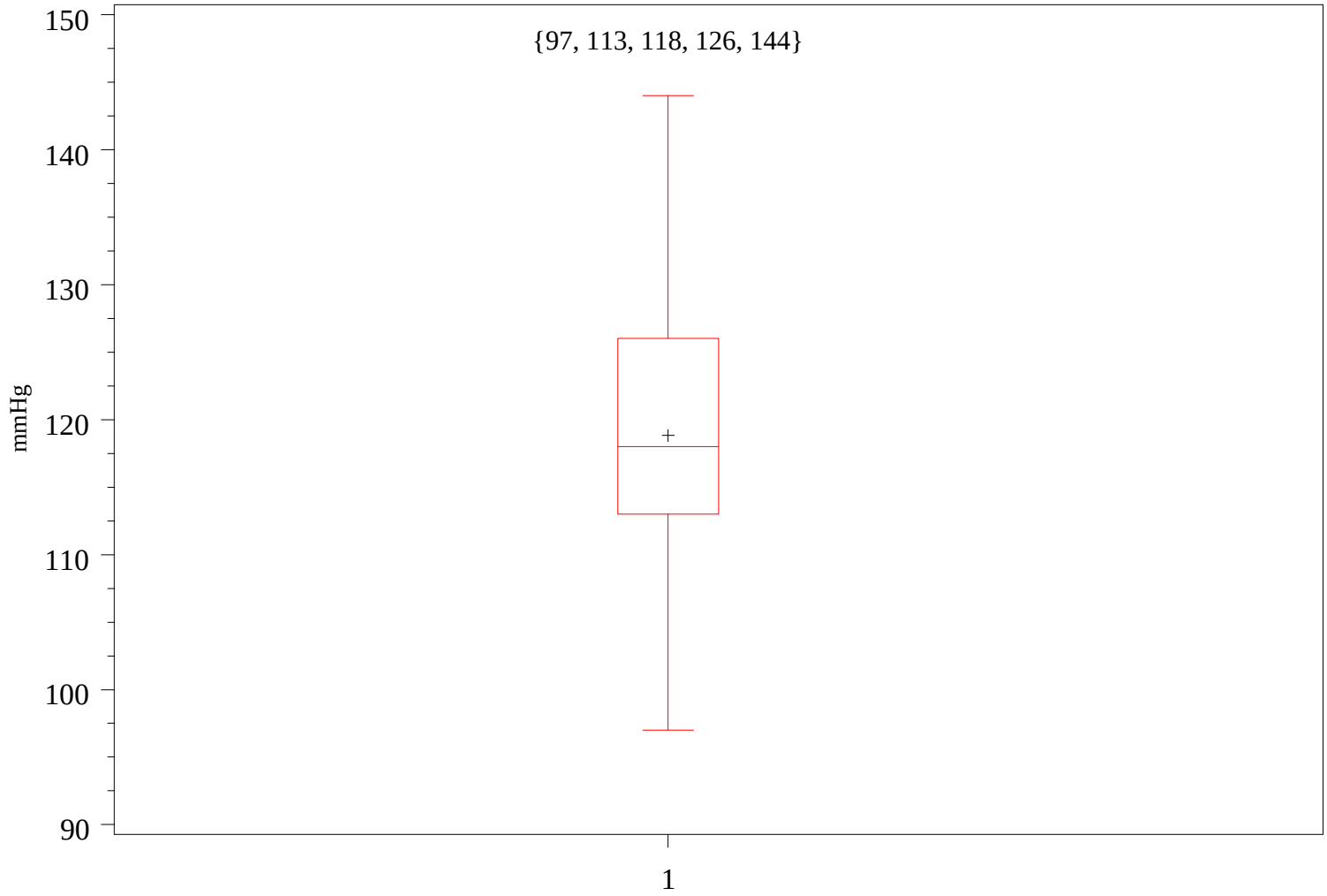
Adding Summary Statistics to Box Plot (page 9)

```
PROC UNIVARIATE noprint data=data1;
  VAR bp;
  BY group;
  OUTPUT min=min mean=mean q1=q1 median=med q3=q3 max=max out=stats;
RUN;

DATA anno1;
  SET stats;
  FORMAT function $8. text $50.;
  RETAIN when 'a';
  function = 'label';
  text = ''||trim( left( put( min, 5 ) ) )||', '||trim( left( put( q1, 5. ) ) )||
    ', '||trim( left( put( med, 5. ) ) )||', '||trim( left( put( q3, 5. ) ) )||
    ', '||trim( left( put( max, 5. ) ) )||'';
  position = '2';
  ...
  OUTPUT;
RUN;

PROC BOXPLOT data=data1;
  PLOT bp*group / vaxis=axis1 haxis=axis2 annotate=anno1;
RUN;
```

Systolic Blood Pressure



Comparing Multiple Data Sets

Box plots can be used to compare multiple data sets.

Comparing Multiple Data Sets with Box Plots (page 10)

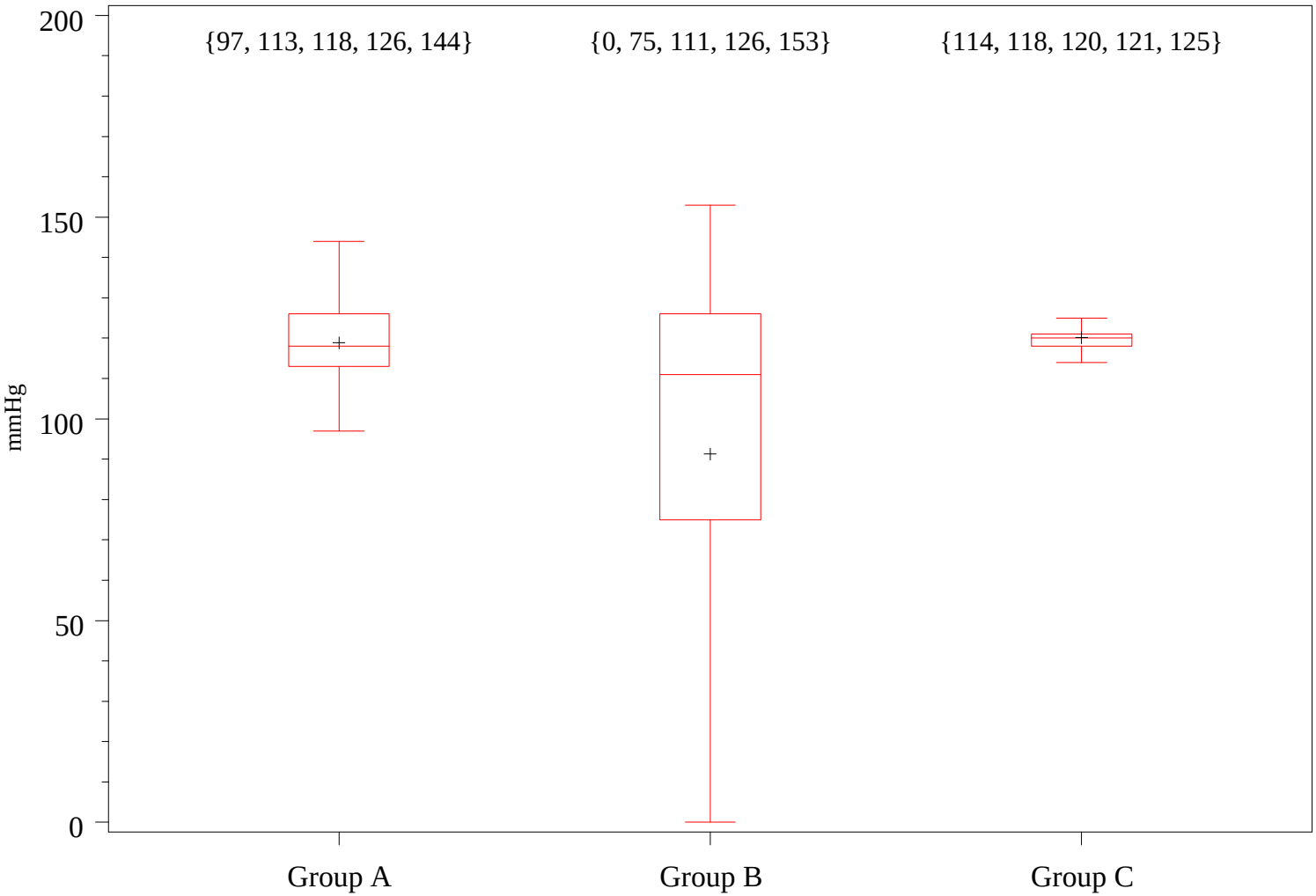
```
DATA data123;
  SET data1 data2a data3;
RUN;

PROC UNIVARIATE data=data123 noprint;
  VAR bp;
  BY group;
  OUTPUT min=min mean=mean q1=q1 median=med q3=q3 max = max out=stats;
RUN;

DATA anno123;
  SET stats;
  FORMAT function $8. text $50.;
  ...
RUN;

axis1 label=( 'mmHg' ) minor=( number=4 ) value=( height=1.2 );
axis2 label=( '' ) order=( 1 to 3 by 1 )
  value=( height=1.2 'Group A' 'Group B' 'Group C' ) minor=none;

PROC BOXPLOT data=data123;
  PLOT bp*group / vaxis=axis1 haxis=axis2 annotate=anno123;
RUN;
```


Systolic Blood Pressure
{97, 113, 118, 126, 144}
{0, 75, 111, 126, 153}
{114, 118, 120, 121, 125}
200
150
100
50
0
mmHg
Group A
Group B
Group C

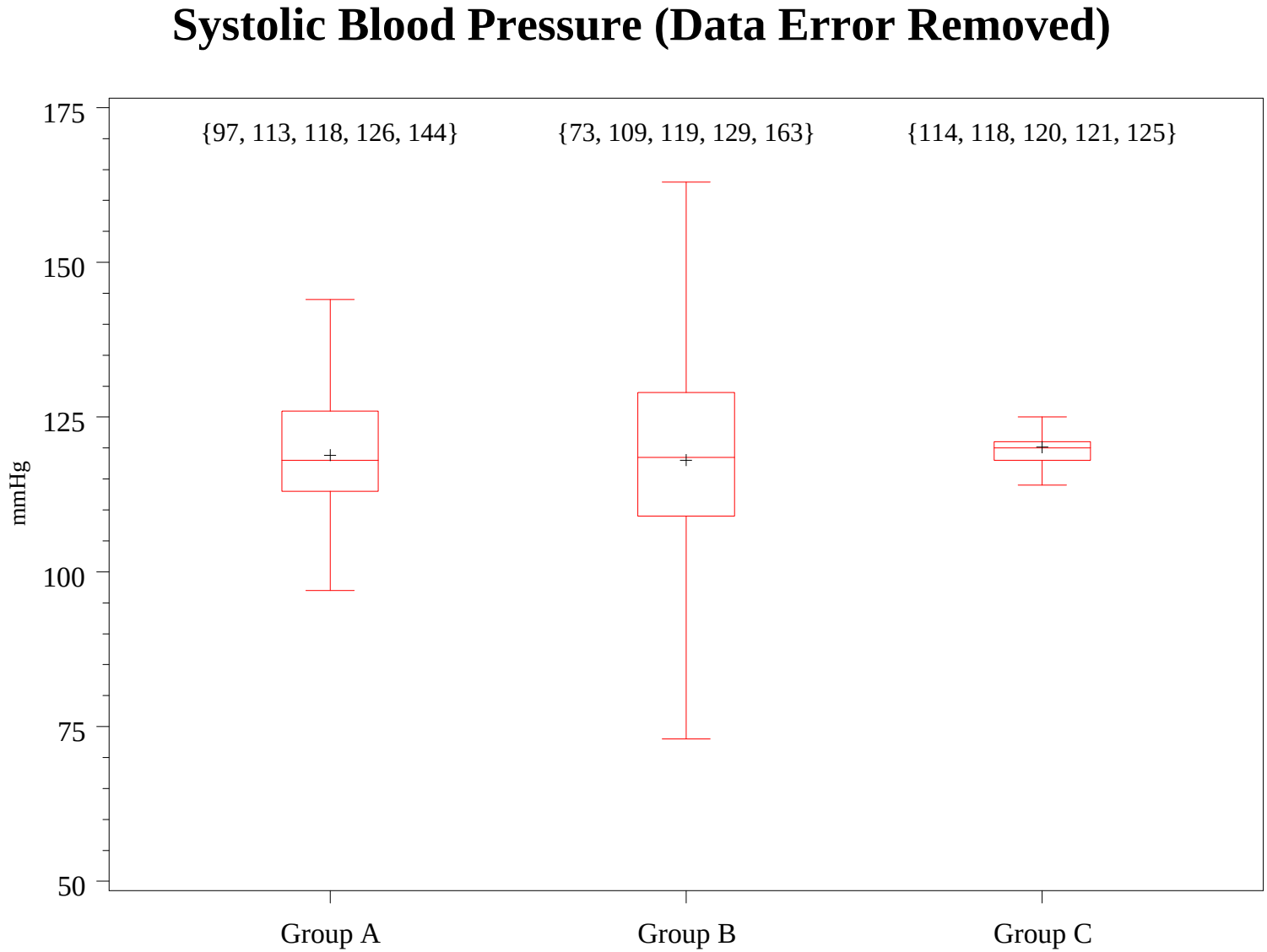
Exploratory Data Analysis

Obvious irregularity in Group B!

- Minimum value is zero (?!)
- Data heavily skewed toward that as well (mean $\ll$ median!).

Q: Does this make sense? (No!)

- Data error: 11 data points of zero.
- Re-run code after correcting this error.


Systolic Blood Pressure (Data Error Removed)
{97, 113, 118, 126, 144}
{73, 109, 119, 129, 163}
{114, 118, 120, 121, 125}
175
150
125
100
75
50
mmHg
Group A
Group B
Group C

Exploratory Data Analysis

With corrected data, what can we see about the data for these three groups?

- All have about the same central values (median and mean).
- All have (very) different spreads:
  - Highest for Group B.
  - Lowest for Group C.

Do these observations make sense?

- B composed of very diverse people?
- C composed of very similar people?

Answering these questions leads to deeper analysis.

Different Box Plots

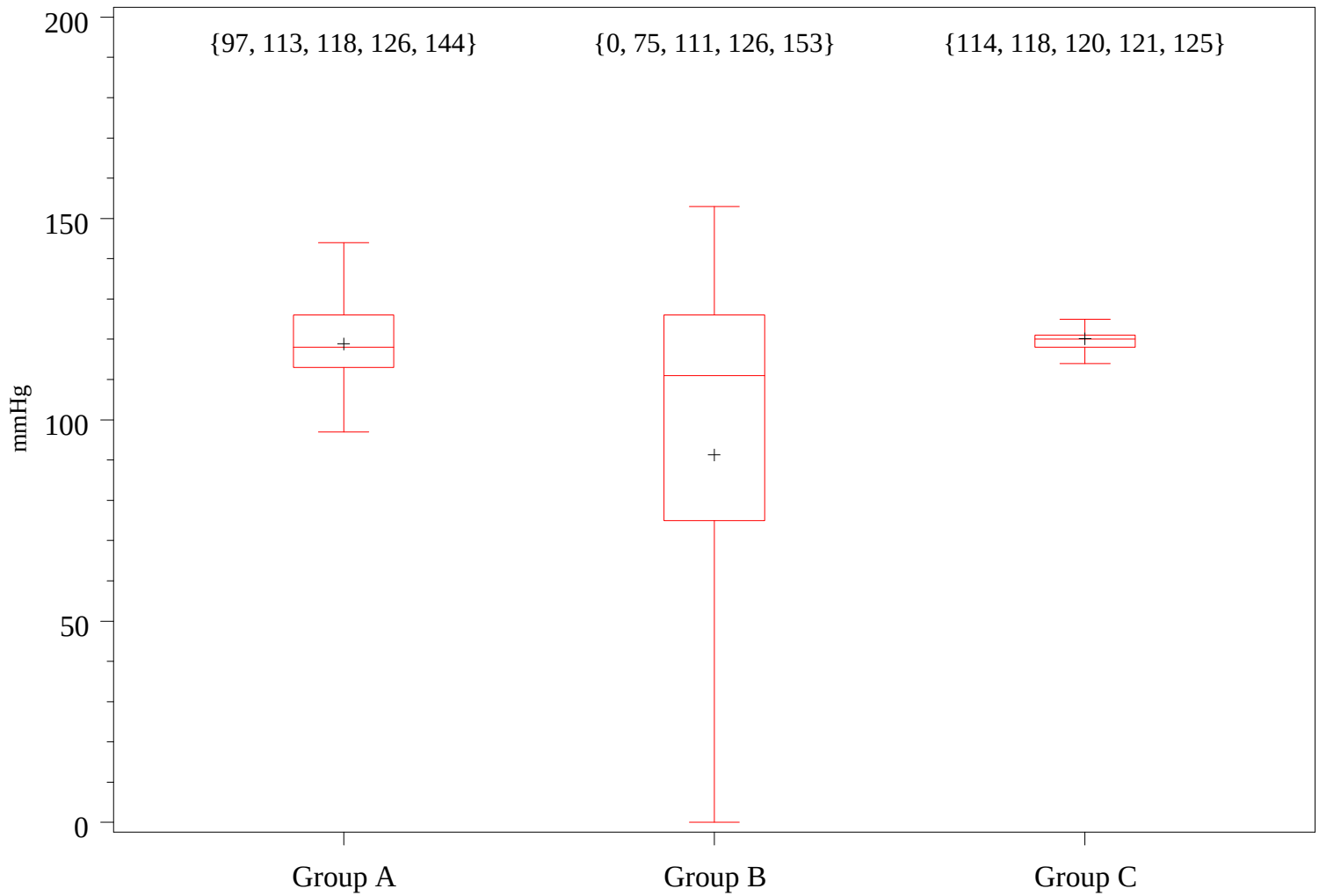
Try `boxstyle=schematic` option:

Box Plots with `boxstyle=schematic` (page 12)

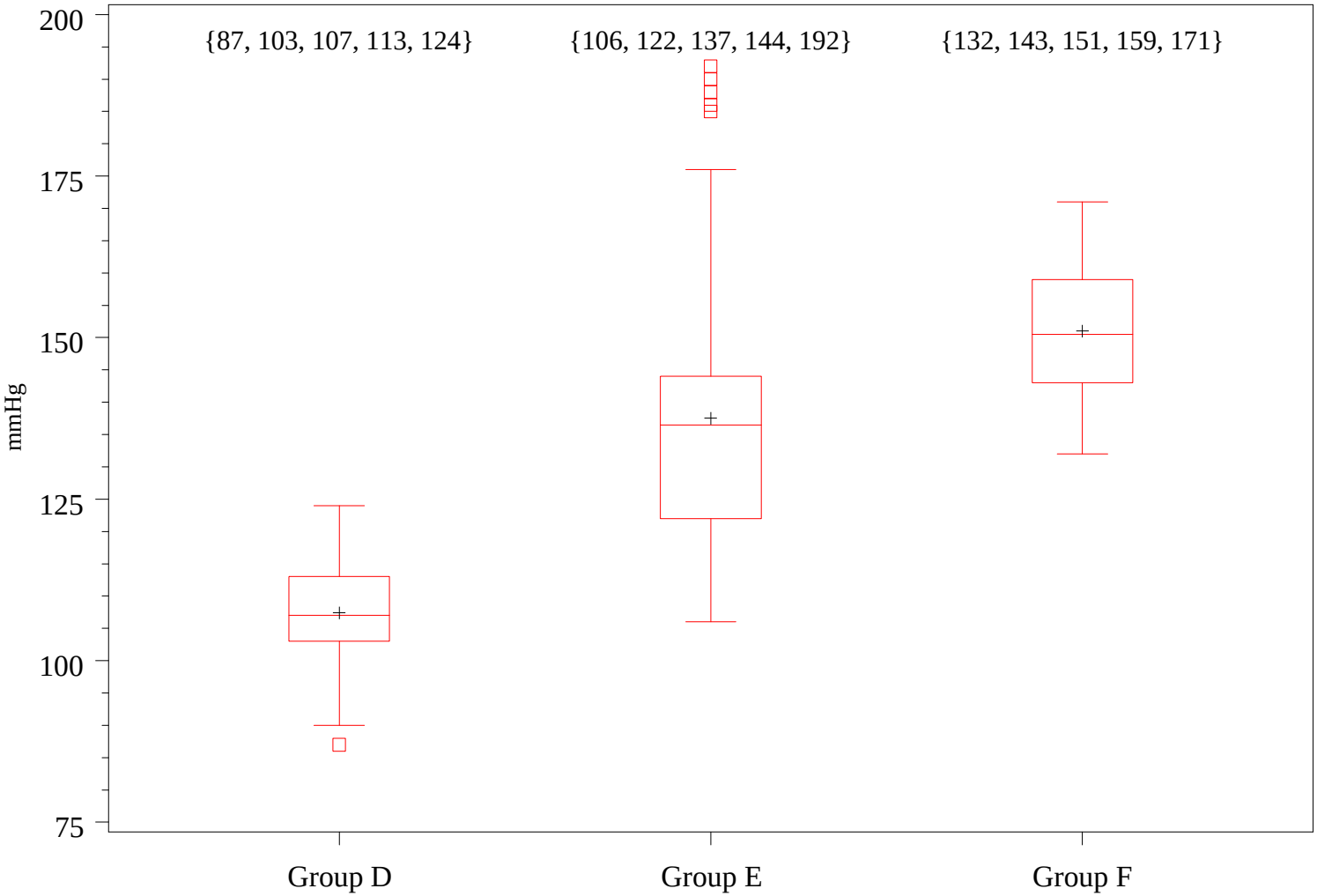
```
PROC BOXPLOT data=data456a;
  PLOT bp*group / vaxis=axis1 haxis=axis2
    annotate=anno456a boxstyle=schematic;
RUN;
```

Boxes now denote *outliers*:

- Points below the 25^{th} percentile - 1.5 $\times$ IQR.
- Points above the 75^{th} percentile + 1.5 $\times$ IQR.


Systolic Blood Pressure
{97, 113, 118, 126, 144}
{0, 75, 111, 126, 153}
{114, 118, 120, 121, 125}
200
150
100
50
0
mmHg
Group A
Group B
Group C

Systolic Blood Pressure



Exploratory Data Analysis

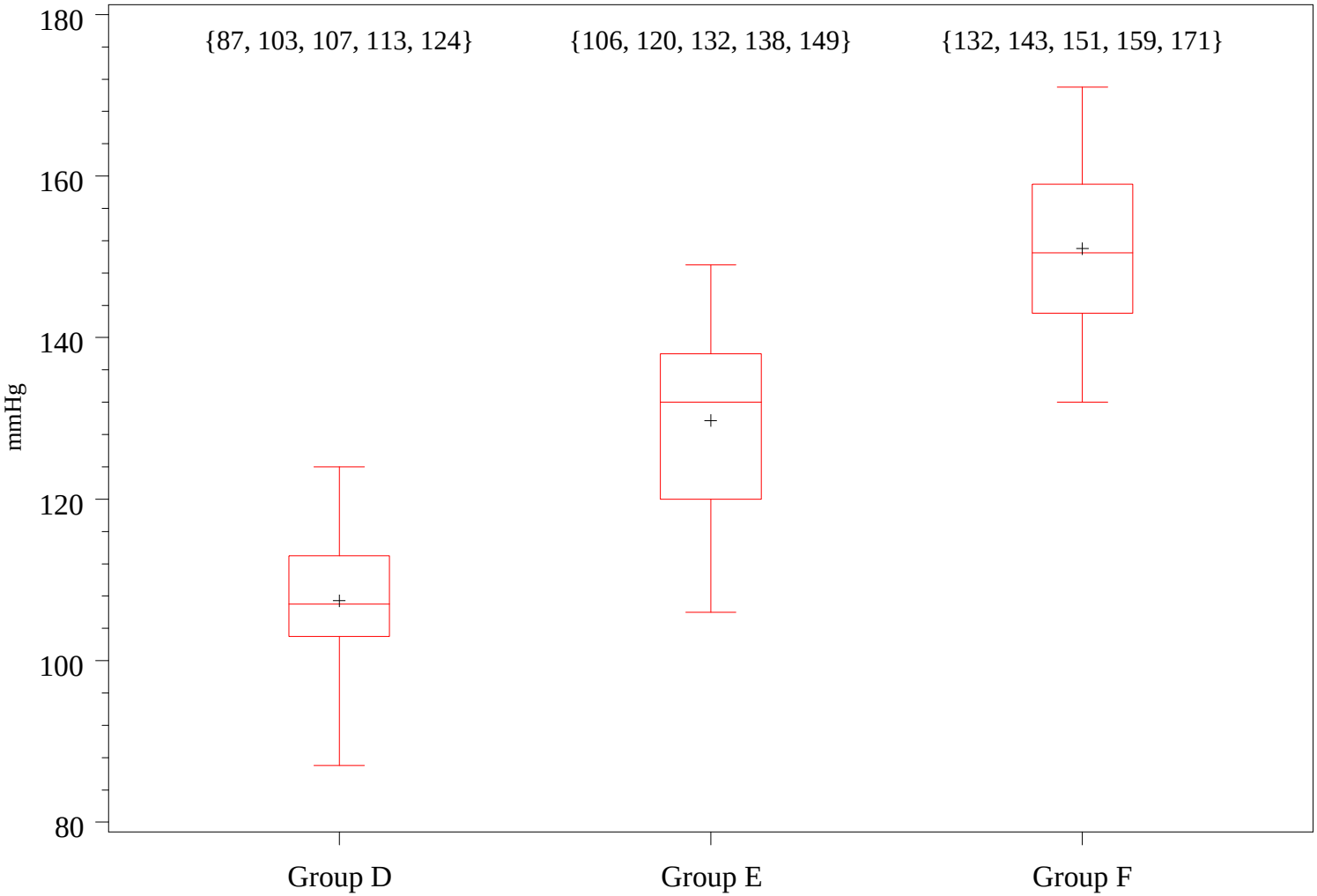
What's happening in Group E?

- Data heavily skewed upward (mean $>$ median).

Q: Does this make sense? (Maybe)

- Six values at/above 180 mmHg.
- Assume rational explanation for them (e.g., six people have extreme risk of heart attack).
- Not data errors; they are *outliers*.
- Do we keep them or throw them out? (Open question!)
  - What do we get if we throw them out?

Systolic Blood Pressure (Outliers Removed)



Exploratory Data Analysis (without outliers)

- All three groups have about the same spreads.
- Three groups have different central values.

Do these observations make sense?

- Outside realm of this paper.

With or Without Outliers?

Do we cut outliers out of the analysis?

- Yes: Including them distorts the general tendencies we are looking for.
- No: They are valid data points, and thus part of those general tendencies.

No simple answer, except that we can't just discard them. If taken out of the analysis,

- Mention them in sentence or footnote.
- Don't bury them in an end note or appendix.
- Show them in `PROC BOXPLOT` as isolated points (`boxstyle=schematic`).

Conclusions

- Box Plots show tendencies more effectively/robustly than histograms.
- Can be used to easily compare data sets.

Demystifying PROC UNIVARIATE

What do all these numbers mean?

Moments

N	50	Sum Weights	50
Mean	118.84	Sum Observations	5942
Std Deviation	10.14861	Variance	102.994286
Skewness	0.19322593	Kurtosis	0.26180882
Uncorrected SS	711194	Corrected SS	5046.72
Coeff Variation	8.53972571	Std Error Mean	1.4352302

Basic Statistical Measures

Location		Variability	
Mean	118.8400	Std Deviation	10.14861
Median	118.0000	Variance	102.99429
Mode	115.0000	Range	47.00000
		Interquartile Range	13.00000

NOTE: The mode displayed is the smallest of 2 modes with a count of 4.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	82.80205	Pr > |t|	<.0001
Sign	M	25	Pr >= |M|	<.0001
Signed Rank	S	637.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	144.0
99%	144.0
95%	136.0
90%	131.0
75% Q3	126.0
50% Median	118.0
25% Q1	113.0
10%	106.5
5%	103.0
1%	97.0
0% Min	97.0

Extreme Observations

-----Lowest-----		-----Highest----	
Value	Obs	Value	Obs
97	32	131	45
98	42	132	7
103	25	136	27
103	12	143	22
106	3	144	33

Demystifying PROC UNIVARIATE

- PROC UNIVARIATE provides summary statistics.
- Summary statistics are supposed to describe a data set with just a few numbers (*data reduction*).
- **PROC UNIVARIATE is using 46 numbers to summarize 50 numbers.**

$\Rightarrow$ PROC UNIVARIATE is not an effective data summary ... but it's not supposed to be!

Demystifying PROC UNIVARIATE

PROC UNIVARIATE has more information than we usually need.

- Not meant to be a data summary *per se*.
- Designed to be "one-stop shop" for anything we would ever want.
- Just because a statistic is listed does not mean we need to use it!

For completeness, we'll go through everything here (starting from the bottom).

Output of PROC UNIVARIATE

- **Extreme Observations** = five lowest/highest observations, with observation numbers.
- **Quantiles** = percentiles.
  - **Definition 5** = define percentiles a certain way (not very important).

Output of PROC UNIVARIATE

- **N** = total number of data points.
- **Sum Weights** = sum of data weights (almost always N).
- **Mean** = sample mean.
- **Sum Observations** = sum of data observations: $\sum_{i=1}^{\mathrm{N}} x_i$.
  - Equal to N × Mean.
- **Std Deviation** = sample standard deviation (mean distance from mean):

$$\sqrt{\frac{1}{\mathrm{N}-1}\sum_{i=1}^{\mathrm{N}}(x_i-\bar{x})^2}.$$

Output of PROC UNIVARIATE

- **Variance** = sample variance (mean squared distance from mean):

$$\frac{1}{\mathrm{N}-1}\sum_{i=1}^{\mathrm{N}}(x_i-\bar{x})^2.$$

  - Equal to (Std Deviation)^2.
- **Skewness** = measure of how skewed our data distribution is (negative = left, positive = right).
- **Kurtosis** = measure of how "peaked" our data distribution is at the mode.

Output of PROC UNIVARIATE

Fun Fact: $\sum_{i=1}^{\mathrm{N}}(x_i - \bar{x})^2 = \sum_{i=1}^{\mathrm{N}} x_i^2 - \mathrm{N} \cdot \bar{x}^2.$

- **Uncorrected SS** = *uncorrected sum of squares*: $\sum_{i=1}^{\mathrm{N}} x_i^2.$
  - Used in `Variance`, `Std Deviation`.
- **Corrected SS** = *corrected sum of squares*: $\sum_{i=1}^{\mathrm{N}}(x_i - \bar{x})^2.$
  - Used in `Variance`, `Std Deviation`.

Output of PROC UNIVARIATE

- **Coeff Variation** = *coefficient of variation* = scaled version of the spread:

$$\frac{100 \times \texttt{Std Deviation}}{\texttt{Mean}}.$$

- **Std Error Mean** = *standard error of the mean* = standard deviation of the distribution of the mean $\bar{x}$:

$$\frac{\texttt{Std Deviation}}{\sqrt{\texttt{N}}}.$$

Are we done?

Moments

N	50	Sum Weights	50
Mean	118.84	Sum Observations	5942
Std Deviation	10.14861	Variance	102.994286
Skewness	0.19322593	Kurtosis	0.26180882
Uncorrected SS	711194	Corrected SS	5046.72
Coeff Variation	8.53972571	Std Error Mean	1.4352302

Basic Statistical Measures

Location		Variability	
Mean	118.8400	Std Deviation	10.14861
Median	118.0000	Variance	102.99429
Mode	115.0000	Range	47.00000
		Interquartile Range	13.00000

NOTE: The mode displayed is the smallest of 2 modes with a count of 4.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	82.80205	Pr > |t|	<.0001
Sign	M	25	Pr >= |M|	<.0001
Signed Rank	S	637.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	144.0
99%	144.0
95%	136.0
90%	131.0
75% Q3	126.0
50% Median	118.0
25% Q1	113.0
10%	106.5
5%	103.0
1%	97.0
0% Min	97.0

Extreme Observations

-----Lowest-----		-----Highest----	
Value	Obs	Value	Obs
97	32	131	45
98	42	132	7
103	25	136	27
103	12	143	22
106	3	144	33

Output of PROC UNIVARIATE

Are all those decimal places necessary?

Std Deviation	10.14861	Variance	102.994286
Skewness	0.19322593	Kurtosis	0.26180882
Coeff Variation	8.53972571	Std Error Mean	1.4352302

- Yes, if used as intermediate value.
- No, if final value.

St Dev = 10.14861 Misleading accuracy, confusing!
St Dev = 10.1 Proper accuracy, intuitive.

Population vs Sample

Assumption:

Data is a representative *sample* from a much larger *population*.

- Statistics from `PROC UNIVARIATE` are estimates of (unknown) *population parameters*.
- Ex: Sample mean of 118.84 = estimate of population mean.
- Unbiased sample $\Rightarrow$ good estimates.
- Biased sample $\Rightarrow$ lousy estimates (1936, 1948 election polls).

Hypothesis Tests

A *hypothesis test* tests if population parameter significantly different from some value (usually 0).

- Ex: H_0: Population mean = 0.
- "Significantly different" = account for spread of data.
  - More spread = more volatile data = less reliable estimates.
  - Is our estimate reliable enough?
- We try to reject H_0.
  - Either reject or don't reject H_0.
  - Do we have enough evidence to reject H_0?
  - Never accept H_0.

Hypothesis Tests

How do we reject H_0?

1. Calculate a *test statistic* from the data.
2. Compare test statistic to an assumed distribution.
3. Calculate *p-value* $\approx$ probability of test statistic, assuming H_0.
4. Reject H_0 if *p*-value really small (less than 0.05 = 5%).

$\Rightarrow$ Assuming H_0, the probability of our result is really small.

$\Rightarrow$ We should have gotten that result less than 5% of the time.

$\Rightarrow$ Since we got that result, our assumption is probability wrong!

Hypothesis Tests in PROC UNIVARIATE

H_0: Population Mean = 0

Three different tests for H_0, using different test statistics:

- `Student's t`= *student's t test*.
- `Sign` = *sign test*.
- `Signed Rank` = *Wilcoxon signed rank test*.

Each test has different data assumptions. Which one to use?

$\Rightarrow$ Often not important! Reject H_0 if all three p-values < 0.05.

Hypothesis Tests in PROC UNIVARIATE

H_0: Population Mean = 0

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	82.80205	Pr > |t|	<.0001
Sign	M	25	Pr >= |M|	<.0001
Signed Rank	S	637.5	Pr >= |S|	<.0001

$\Rightarrow$ Reject H_0. (Trivial, since all values are > 0)

Conclusions

- This presentation shows most commonly used tools of exploratory data analysis.
- Methods: Data visualization and summarization.
- Simple, yet effective.
- They help us find data irregularities and “interesting” features.
- They give us guidance for further analysis, set stage for more complex methods.

Further Resources


Robert Adams,
Box Plots in SAS: UNIVARIATE, BOXPLOT or GPLOT?
Proceedings of the 21st NESUG Conference, np16, 2008.


Perry Watts,
Using SAS Software to Generate Textbook Style Histograms,
Proceedings of the 21st NESUG Conference, np03, 2008.

Nate Derby: `http://nderby.org`

`nderby@sprodata.com`