Univariate Graphics

STAT 133

Gaston Sanchez

Department of Statistics, UC–Berkeley

gastonsanchez.com
github.com/gastonstat/stat133
Course web: gastonsanchez.com/stat133

Looking at one single variable

Univariate Statistical Graphics

Getting started with graphics for exploration requires underdstanding charts and plots for single variables

Univariate Statistical Graphics

Getting started with graphics for exploration requires underdstanding charts and plots for single variables

Many multivariate graphics are extensions or combinations of univariate charts

Univariate graphics by type of variable

Qualitative Variable

- Bar chart
- Dot chart
- Pie chart
- Pareto chart

Quantitative Variable

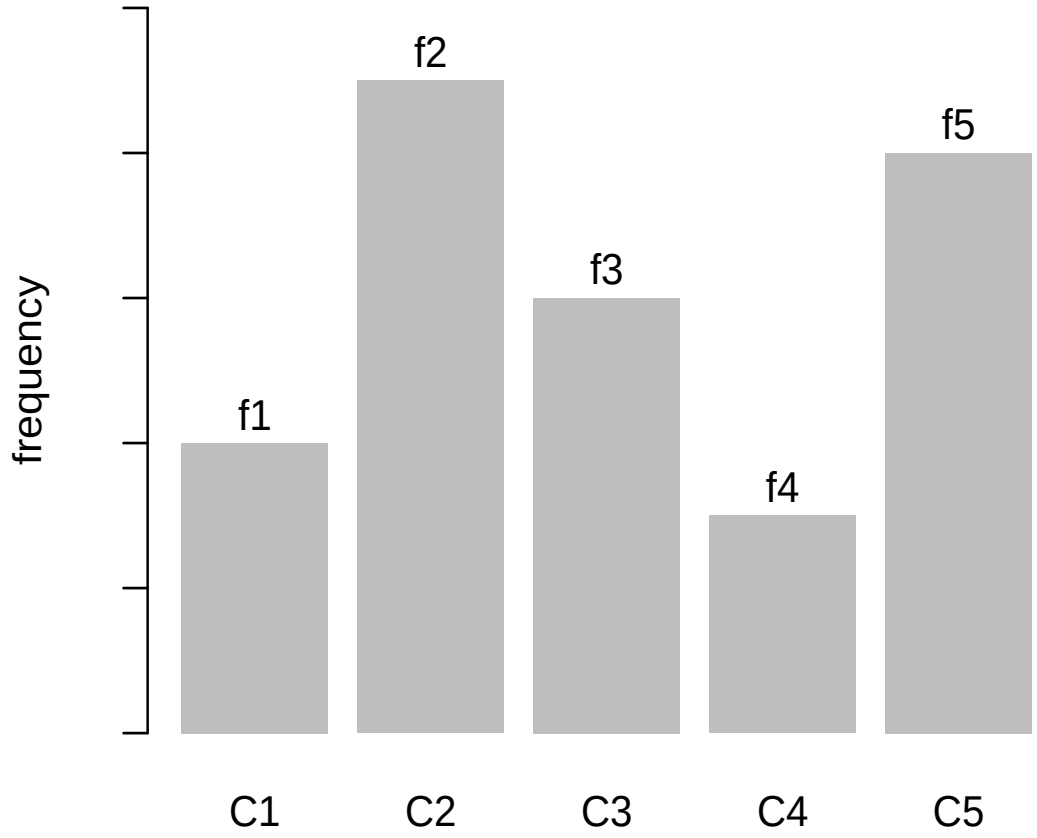
- All of qualitative
- Histogram
- Density curve
- Boxplot
- Ogive

Bar Charts

From Frequency Tables ...

Category	Absolute Frequency	Relative Frequency
C_1	f_1	f_1/n
C_2	f_2	f_2/n
C_3	f_3	f_3/n
$\ldots$	$\ldots$	$\ldots$
C_k	f_k	f_k/n
$total$	n	1

to Bar-charts



Bar-charts

Elements of vertical bar-charts

- categories on horizontal axis
- frequencies on vertical axis
- length of bar equal to frequency

(Note that you can also make a horizontal bar-chart, in which case the axes play inverse roles)

Bar-chart: predominant color in flags

Predominant Color in Flags

```
## Warning in file(file, "rt"):  cannot open file
'/Users/Gaston/Documents/stat133/stat133/datasets/flags.csv':
No such file or directory
## Error in file(file, "rt"):  cannot open the connection
## Error in table(flags$mainhue):  object 'flags' not found
## Error in nrow(flags):  object 'flags' not found
```

```
## Error in data.frame(color = names(hues), count =
as.numeric(hues), percent = round(100 * :  object 'hues'
not found
```

Bar-chart example

```
## Error in barplot(hues, border = NA, axes = FALSE, ylim =
c(0, 75)):  object 'hues' not found
## Error in axis(side = 2, at = seq(0, 75, 10), las = 2):
plot.new has not been called yet
## Error in text(x, y = as.vector(hues) + 2, labels =
hues):  object 'x' not found
```

Bar-chart: predominant color in flags

```
## Error in barplot(hues/nf, border = NA, axes = FALSE,
ylim = c(0, 0.4)):  object 'hues' not found
## Error in axis(side = 2, at = seq(0, 0.4, 0.1), las = 2,
labels = paste(seq(0, :  plot.new has not been called yet
## Error in text(x, y = as.vector(hues/nf) + 0.01, labels =
paste(100 * round(hues/nf, :  object 'x' not found
```

Bar-chart: predominant color in flags

```
## Error in sort(hues):  object 'hues' not found
## Error in axis(side = 2, at = seq(0, 75, 10), las = 2):
plot.new has not been called yet
## Error in text(x, y = as.vector(hues)[order(hues)] + 3,
labels = sort(hues)):  object 'x' not found
```

Bar-chart: predominant color in flags

```
## Error in sort(hues, decreasing = TRUE): object 'hues'
not found
## Error in axis(side = 2, at = seq(0, 75, 10), las = 2):
plot.new has not been called yet
## Error in text(x, y = as.vector(hues)[order(hues,
decreasing = TRUE)] + :  object 'x' not found
```

Bar-chart: predominant color in flags

```
## Error in sort(hues, decreasing = TRUE): object 'hues'
not found
## Error in axis(side = 1, at = seq(0, 75, 10)):  plot.new
has not been called yet
## Error in as.vector(hues):  object 'hues' not found
```

Dot charts

Dot charts

- Dot-charts are very similar to bar charts.
- Instead of using bars, dot-charts display frequencies with dots.
- They are simpler and cleaner than bar charts
- They are also less used than bar charts

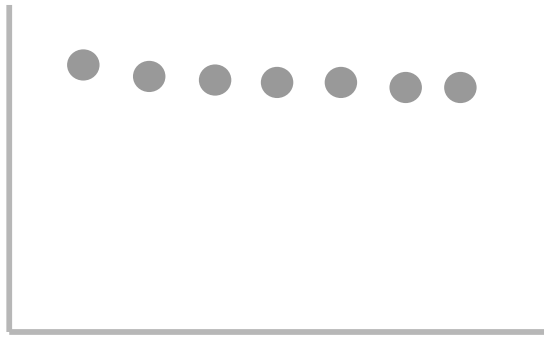
Dot-chart: predominant color in flags

```
## Error in as.vector(hues):  object 'hues' not found
## Error in as.graphicsAnnot(text):  object 'hues' not
found
## Error in axis(side = 1, at = seq(0, 80, 10), line =
0.5):  plot.new has not been called yet
## Error in segments(rep(0, 8), 1:8, rep(80, 8), 1:8, lty =
2, col = "grey"):  plot.new has not been called yet
## Error in as.vector(hues):  object 'hues' not found
```

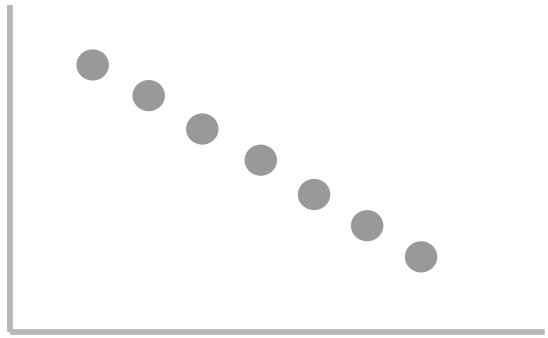
Ranked Dot-charts

```
## Error in as.vector(hues):  object 'hues' not found
## Error in as.graphicsAnnot(text):  object 'hues' not
found
## Error in axis(side = 1, at = seq(0, 80, 10), line =
0.5):  plot.new has not been called yet
## Error in segments(rep(0, 8), 1:8, rep(80, 8), 1:8, lty =
2, col = "grey"):  plot.new has not been called yet
## Error in as.vector(hues):  object 'hues' not found
```

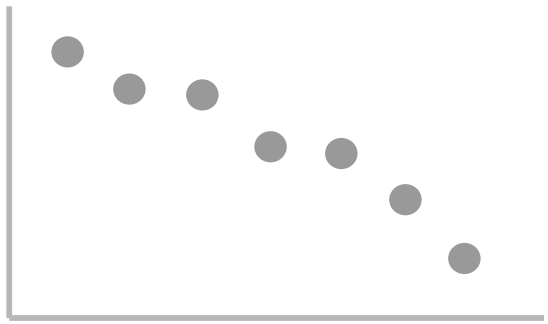
Ranked dot-chart patterns



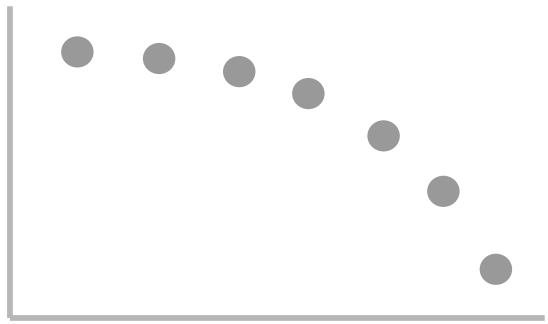
all values roughly the same



differences decrease by
roughly the same amount



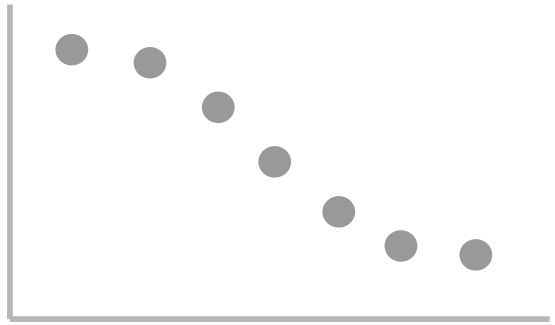
differences from one value to
the next vary significantly



differences from one value
to the next increase

Ranked dot-chart patterns

differences from one value
to the next decrease



shifting differences from
one value to the next

one or more values are extraordinarily
different from the rest

Pareto charts

Bar-chart with Pareto Line

```
## Error in eval(expr, envir, enclos):  object 'hues' not
found
## Error in as.vector(relhues):  object 'relhues' not found
## Error in sort(relhues, decreasing = TRUE): object
'relhues' not found
## Error in axis(side = 2, at = seq(0, 100, 10), las = 2,
cex.axis = 0.8, :  plot.new has not been called yet
```

Bar-chart with Pareto Line

```
## Error in sort(relhues, decreasing = TRUE): object
'relhues' not found
## Error in axis(side = 2, at = seq(0, 100, 10), las = 2,
cex.axis = 0.8, :  plot.new has not been called yet
## Error in points(x, y = cumhues, pch = 19, col =
"gray50"):  object 'x' not found
## Error in text(x, y = cumhues, labels = paste(cumhues,
"%", sep = ""), :  object 'x' not found
```

Bar-chart with Pareto Line

```
## Error in sort(relhues, decreasing = TRUE): object
'relhues' not found
## Error in axis(side = 2, at = seq(0, 100, 10), las = 2,
cex.axis = 0.8, :  plot.new has not been called yet
## Error in lines(x, y = cumhues, col = rgb(255/255,
99/255, 71/255, 0.8), :  object 'x' not found
## Error in points(x, y = cumhues, pch = 19, col =
"gray50"):  object 'x' not found
## Error in text(x, y = cumhues, labels = paste(cumhues,
"%", sep = ""), :  object 'x' not found
```

Pareto charts

- Pareto charts contains both bars and a line graph
- Individual values are representing in descending order
- Cumulative frequencies are represented by the line
- The left vertical axis is the frequency of occurrence

Pie charts

Pie Chart

```
## Error in pie(hues, col = names(hues)): object 'hues'
not found
```

Donut Chart

```
## Error in pie(hues, col = names(hues)):  object 'hues'
not found
## Error in plot.xy(xy.coords(x, y), type = type, ...):
plot.new has not been called yet
```

Pie charts disadvantages

- Pie charts force us to compare either 2-D areas formed by each slice or the angles formed
- Visual perception handles neitheir of these comparisons easily or accurately

Univariate Quantitative Charts

NFL Ticket prices (2013)

```
## Warning in file(file, "rt"):  cannot open file
'/Users/Gaston/Documents/stat133/stat133/datasets/tickets.csv':
No such file or directory
## Error in file(file, "rt"):  cannot open the connection
## Error in cbind(ticks[1:16, ], ticks[17:32, ]):  object
'ticks' not found
## Error in eval(expr, envir, enclos):  object 'ticks' not
found
```

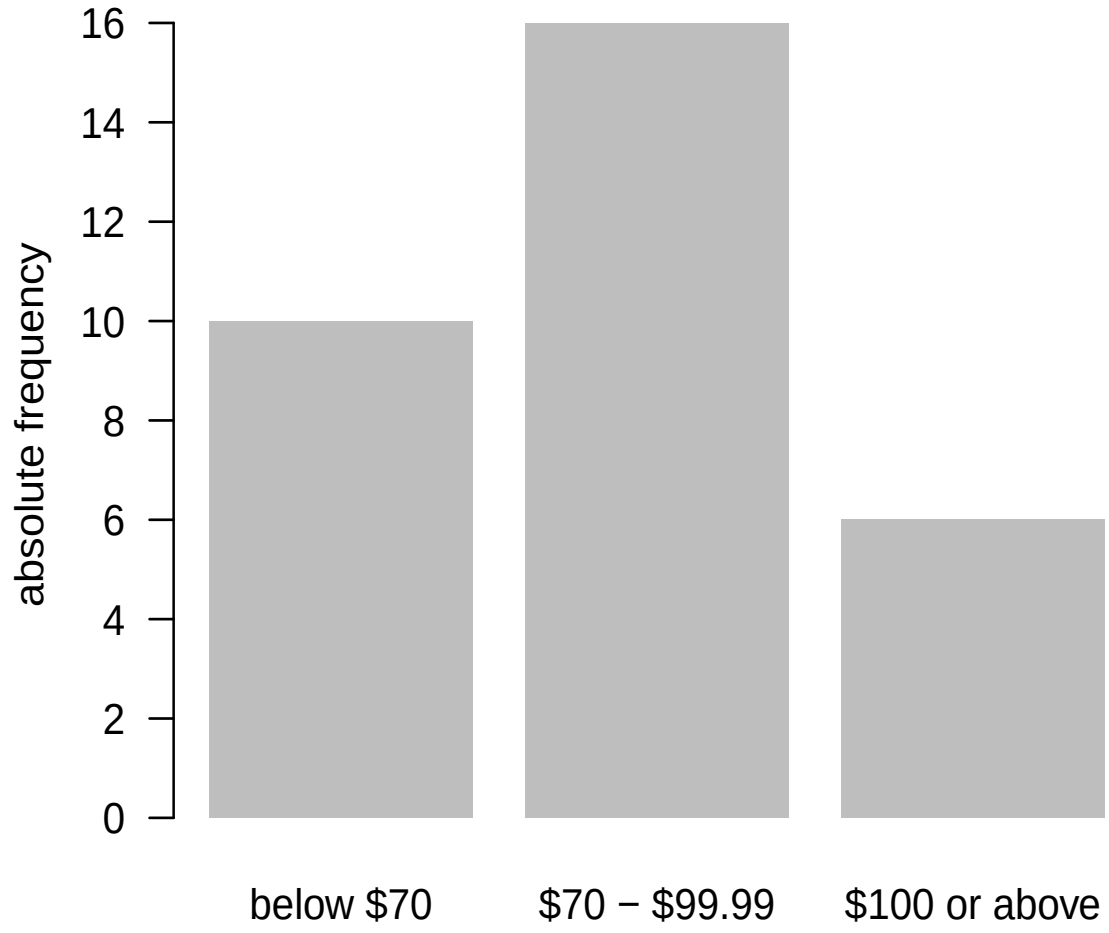
Bar charts for quantitative variables

- We can use bar charts with quantitative variables
- In this case we need to first categorize the variable, and then get a frequency table

Frequency Table of Ticket Prices

Category Name	Absolute Frequency	Relative Frequency
Below $70	10	0.3125
$70 - $99.99	16	0.5000
$100 or above	6	0.1875
Total	32	1.00

NFL Ticket prices (2013)



Histograms

Histograms

Histograms provide a way of viewing the general distribution of values in a quantitative variable

NFL Ticket prices (2013)

```
## Error in hist(tickets, breaks = seq(50, 120, 10), col =
"gray70", xlab = "", :  object 'tickets' not found
## Error in axis(side = 1, at = seq(50, 120, 10)):
plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in axis(side = 2, at = seq(0, 8, 2), las = 2):
plot.new has not been called yet
## Error in mtext("frequency", side = 2, line = 2, cex =
0.9):  plot.new has not been called yet
```

Building a Histogram

1. **Partition of values**: The range of the data values is partitioned into a number of non-overlapping "cells" or bins.
2. **Counting frequencies**: The number of data values falling into each cell is counted (either absolute or relative freqs)
3. **Drawing Bars**: The observations falling into a cell are represented as a "bar" drawn over the cell

About Histograms

- The bins represent ranges of values
- The bins (intervals) must be adjacent, and usually of equal size
- The bars are adjacent (not discontinuous)
- The areas of the bars are meaningful
- Height of bars equal to the frequency
- Width equal to the bin size
- The area of a bar gives the proportion of data values which fall in the bin

Histogram with 4 bins

```
## Error in hist(tickets, breaks = c(54, 70, 86, 102, 118),
col = "gray70", :  object 'tickets' not found
## Error in axis(side = 1, at = seq(54, 124, 8)):  plot.new
has not been called yet
## Error in axis(side = 2, at = seq(0, 12, 2), las = 2):
plot.new has not been called yet
```

Histograms with different bins

```
## Error in hist(tickets, breaks = c(50, 70, 90, 110, 130),
col = "gray70", :  object 'tickets' not found
## Error in axis(side = 1, at = seq(50, 130, 10)):
plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in axis(side = 2, at = seq(0, 14, 2), las = 2):
plot.new has not been called yet
## Error in mtext("frequency", side = 2, line = 2, cex =
0.9):  plot.new has not been called yet
## Error in hist(tickets, breaks = c(40, 60, 80, 100, 120,
140), col = "gray70", :  object 'tickets' not found
## Error in axis(side = 1, at = seq(40, 140, 20)):
plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in axis(side = 2, at = seq(0, 14, 2), las = 2):
plot.new has not been called yet
## Error in mtext("frequency", side = 2, line = 2, cex =
0.9):  plot.new has not been called yet
```

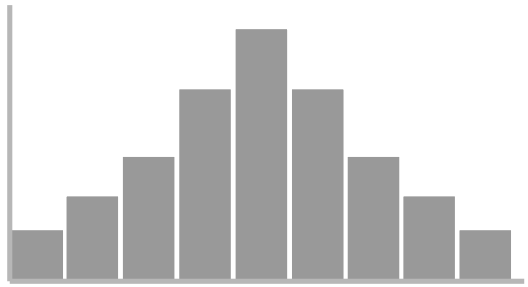
Avoid too few and too many bins

```
## Error in hist(tickets, breaks = c(50, 80, 110, 140), col
= "gray70", xlab = "", :  object 'tickets' not found
## Error in axis(side = 1, at = c(50, 80, 110, 140)):
plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in axis(side = 2, at = seq(0, 16, 2), las = 2):
plot.new has not been called yet
## Error in mtext("frequency", side = 2, line = 2, cex =
0.9):  plot.new has not been called yet
## Error in hist(tickets, breaks = seq(50, 120, 5), col =
"gray70", xlab = "", :  object 'tickets' not found
## Error in axis(side = 1, at = seq(50, 120, 5)):  plot.new
has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in axis(side = 2, at = seq(0, 6, 1), las = 2):
plot.new has not been called yet
## Error in mtext("frequency", side = 2, line = 2, cex =
0.9):  plot.new has not been called yet
```

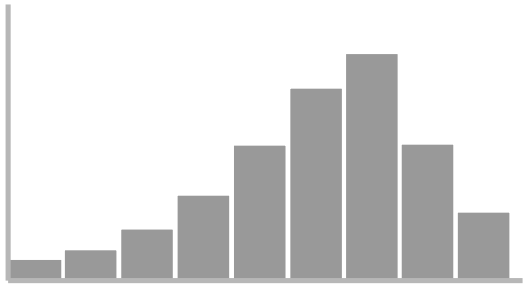
About Histograms

- The shape of a histogram depends on the chosen bins
- This suggests that there is a fundamental instability at the heart of its construction
- The bars are adjacent (not discontinuous)
- The areas of the bars are meaningful

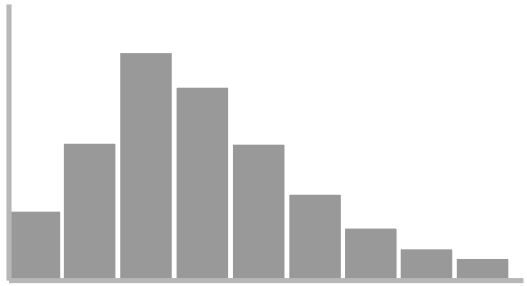
Histogram patterns



Symmetrical

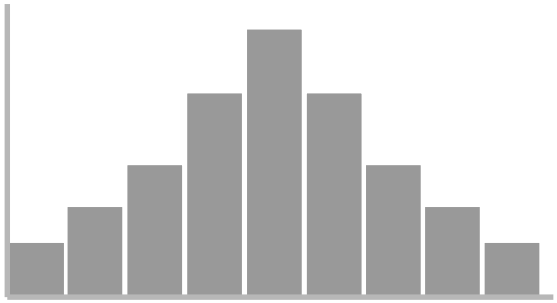


Skewed to the left

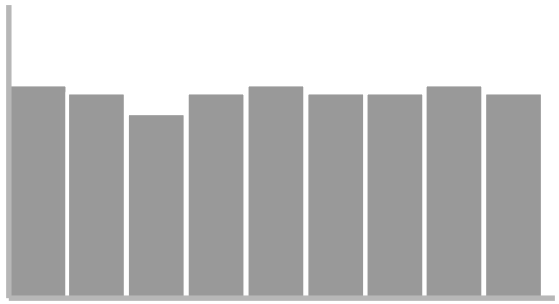


Skewed to the right

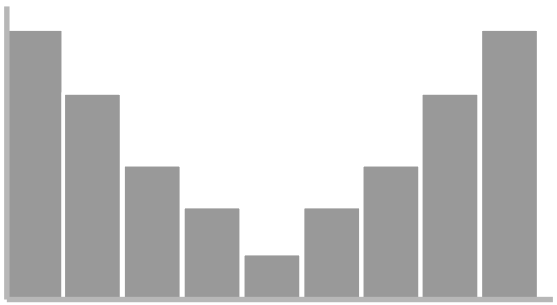
Histogram patterns



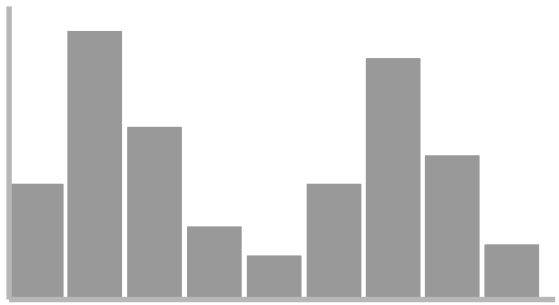
Curved



Flat or Uniform

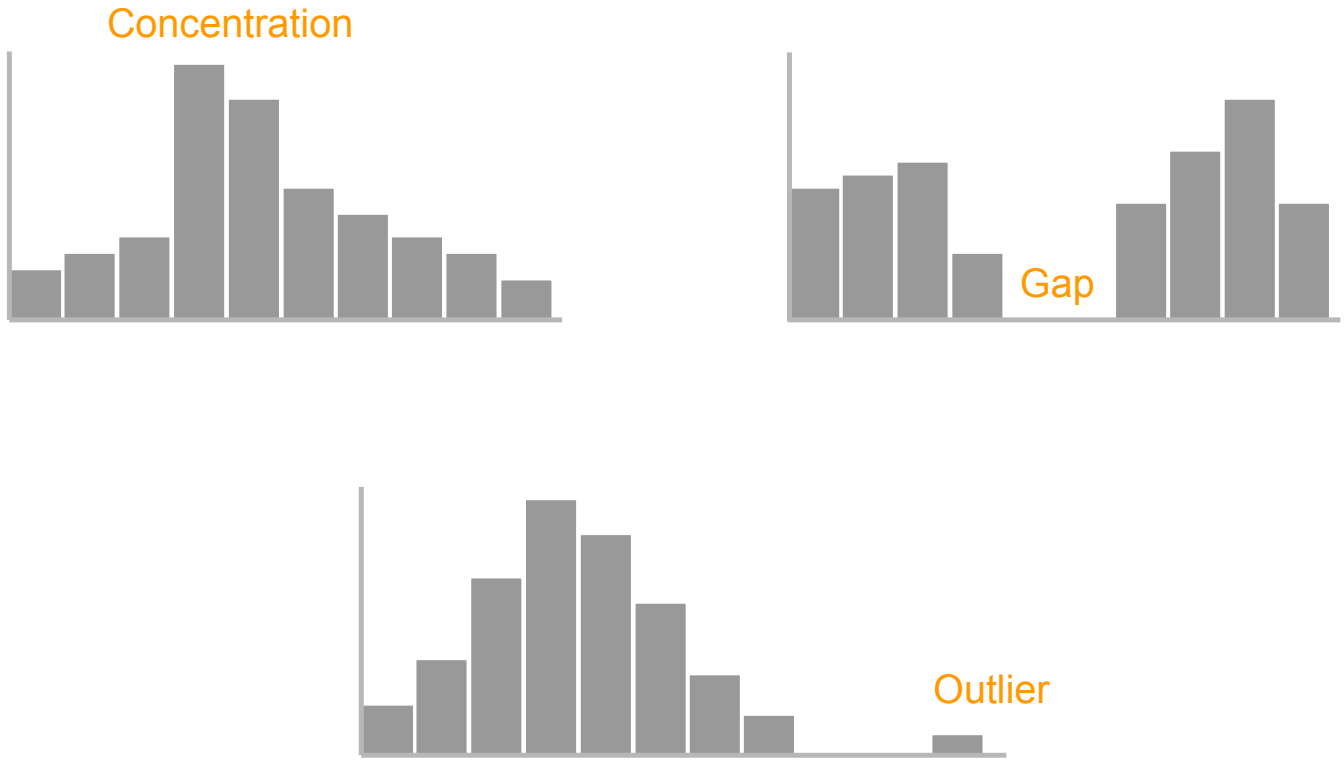


Curved Downward

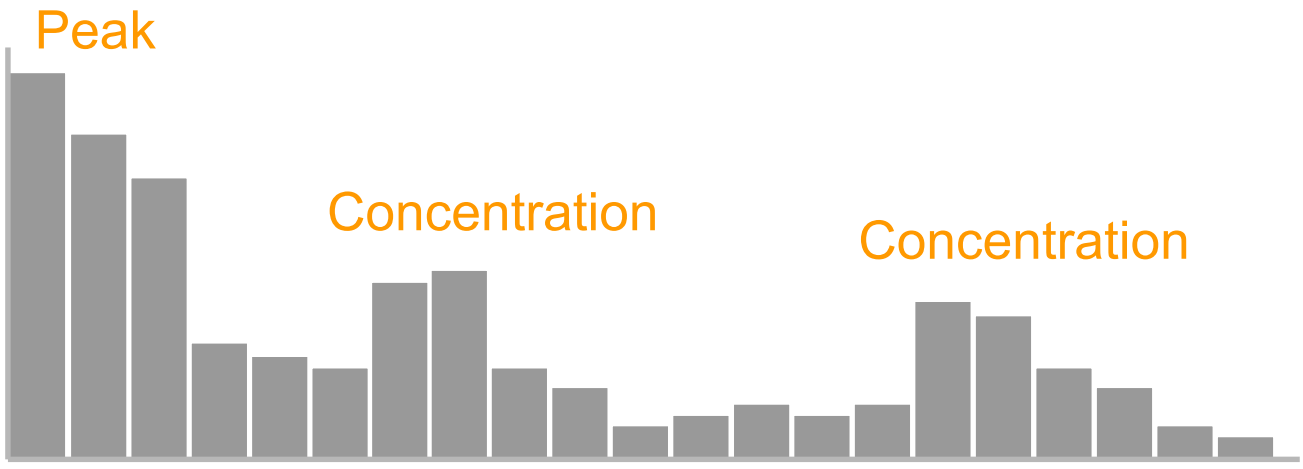


Multiple peaks
(e.g. bimodal, trimodal. etc)

Histogram patterns



Histogram patterns



Box plots

Building a Histogram

1. **Box-and-whisker plots**, most commonly known as "box plots"
2. created by John Tukey
3. simple and effective way to display the distribution of values
4. relies on the so-called **5-summary indicators**

Box plots based on 5-number summary

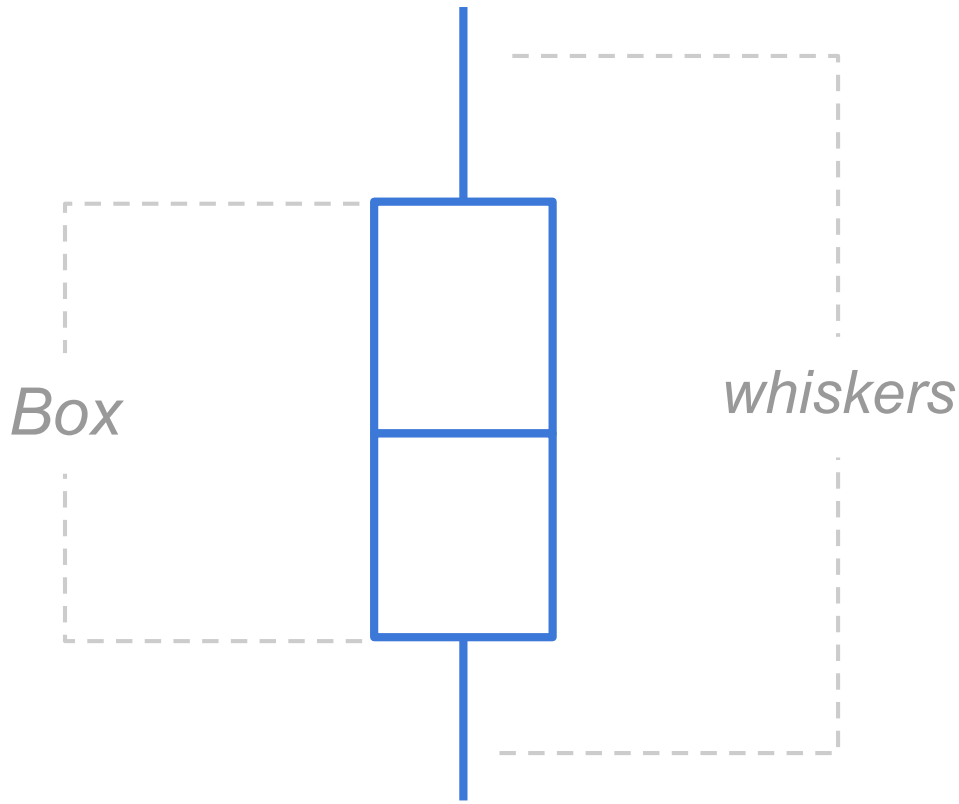
5 summary indicators

Box plots based on 5-number summary

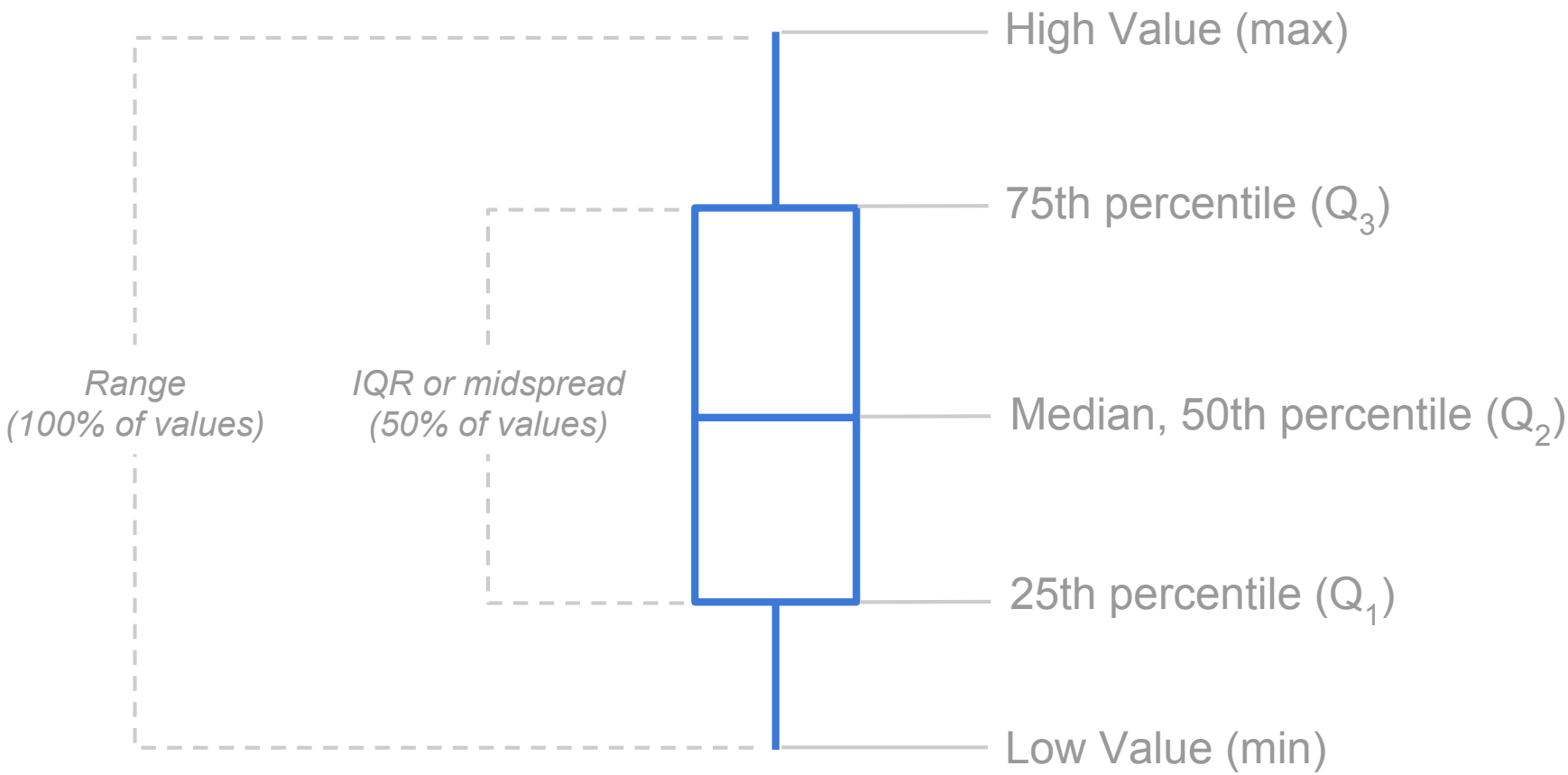
5 summary indicators

1. minimum
2. 25th percentile (1st quartile)
3. 50th percentile (2nd quartile, or median)
4. 75th percentile (3rd quartile)
5. maximum

Box plot basics



Box plot basics



NFL Ticket Prices

```
## Error in boxplot(tickets, horizontal = TRUE, axes =
FALSE, border = "white"):  object 'tickets' not found
## Error in axis(side = 1, at = seq(50, 120, 5), line = -1,
col.axis = "gray35"):  plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in points(tickets, rep(1, length(tickets)), pch =
19, col = hsv(0.6, :  object 'tickets' not found
```

5 number summary

```
## Error in boxplot(tickets, horizontal = TRUE, axes =
FALSE, border = "white"):  object 'tickets' not found
## Error in axis(side = 1, at = seq(50, 120, 5), line = -1,
col.axis = "gray35"):  plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in points(tickets, rep(1, length(tickets)), pch =
19, col = hsv(0.6, :  object 'tickets' not found
## Error in summary(tickets):  object 'tickets' not found
## Error in summary(tickets):  object 'tickets' not found
```

Box plot

```
## Error in boxplot(tickets, horizontal = TRUE, axes =
FALSE): object 'tickets' not found
## Error in axis(side = 1, at = seq(50, 120, 5), line = -1,
col.axis = "gray35"):  plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
## Error in points(tickets, rep(1, length(tickets)), pch =
19, col = hsv(0.6, :  object 'tickets' not found
```

Box plot

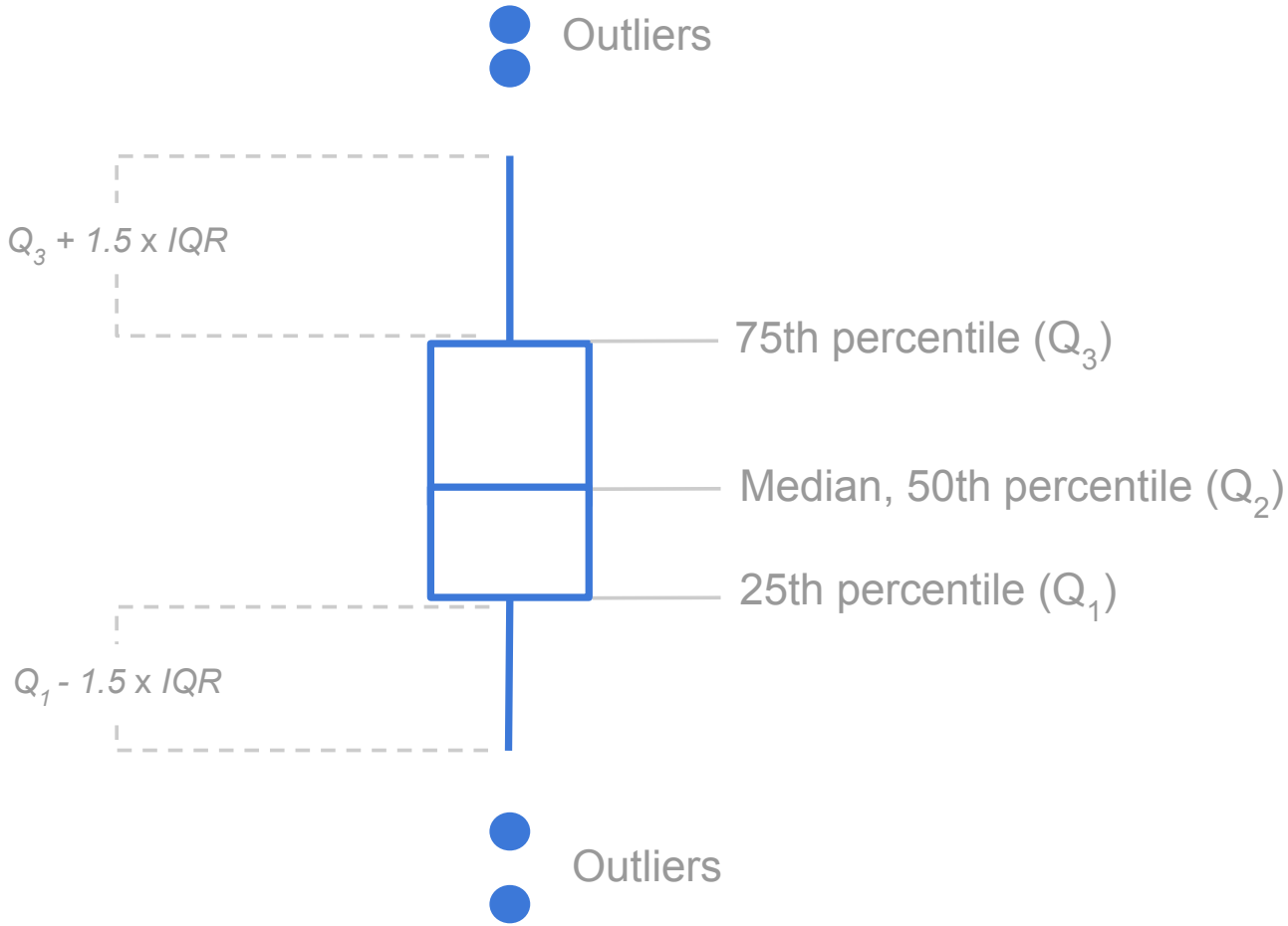
```
## Error in boxplot(tickets, horizontal = TRUE, axes =
FALSE, range = 0):  object 'tickets' not found
## Error in axis(side = 1, at = seq(50, 120, 5), line = -1,
col.axis = "gray35"):  plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2.5):  plot.new
has not been called yet
```

Box plot and outliers

The 1.5 x IQR rule for outliers

Call an observation a suspected outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile

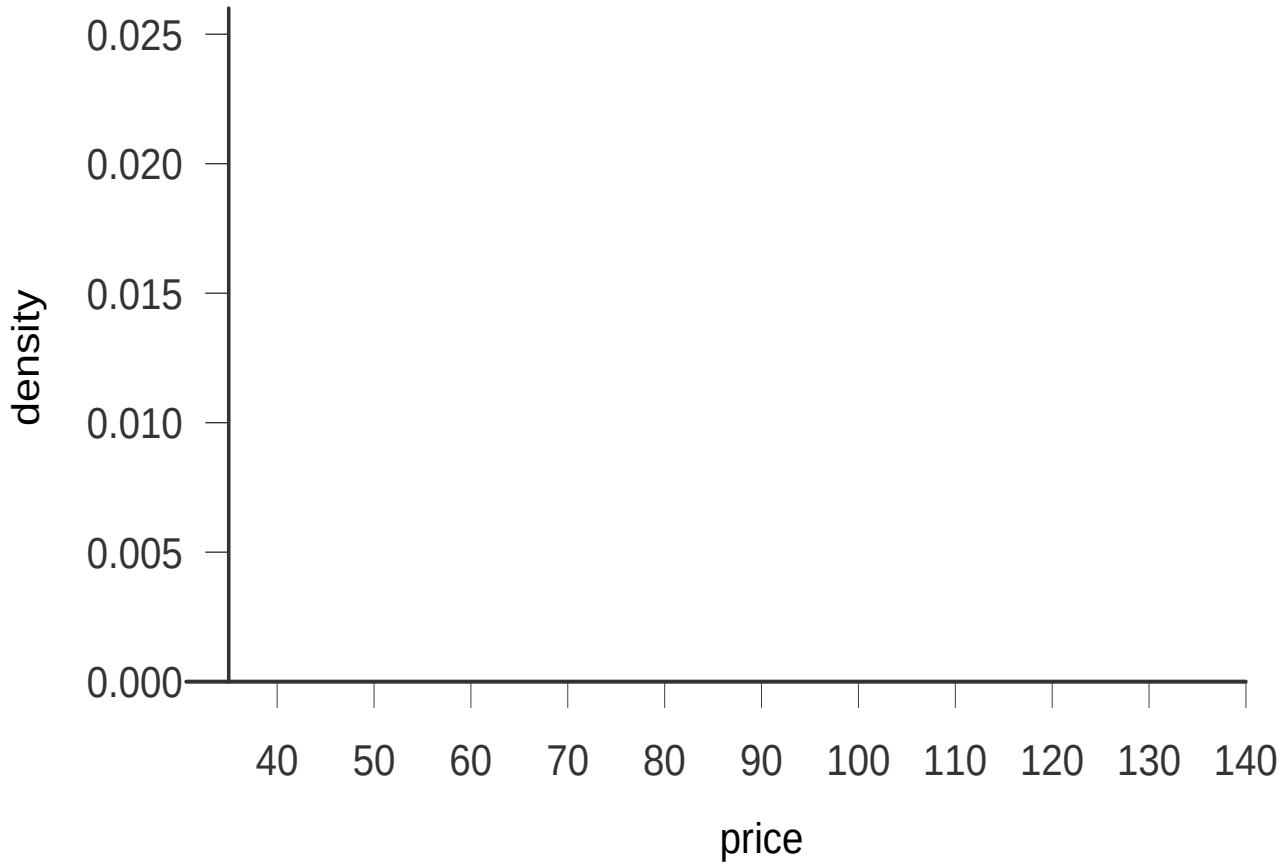
Modified Box plot



Density Curves

Density Curve

```
## Error in density(tickets):  object 'tickets' not found
```




```
## Error in xy.coords(x, y):  object 'dens' not found
```

Density Curve

A Density Curve

- Describes the distribution of values by a smooth curve
- Is always on or above the horizontal axis
- Has area equal to 1 underneath it
- Is an idealized distribution

Density Curve

About Density Curve

- The **mode** is the peak point of the curve (could be more than one or none)
- The **median** is the equal-areas point
- The **mean** is the balance point
- The median and the mean are always equal on a symmetric density curve

Ogives

Ogives

About Ogives

- Ogives help us examine the cumulative distribution of values in a quantitative variable
- An ogive tells us how many data are less than the indicated value on the horizontal axis
- An ogive shows how slowly or rapidly the data values accumulate over the range of the data

Frequency Table NFL Price Tickets

Bin	Interval	Mid-point	Frequency	Cum Freq
1	[50-60)	55	2	2
2	[60-70)	65	8	10
3	[70-80)	75	6	16
4	[80-90)	85	8	24
5	[90-100)	95	2	26
6	[100-110)	105	2	28
7	[110-120)	115	4	32

Ogive

```
## Error in hist(tickets, plot = FALSE): object 'tickets'
not found
## Error in eval(expr, envir, enclos):  object 'x' not
found
## Error in eval(expr, envir, enclos):  object 'abs_freq'
not found
## Error in eval(expr, envir, enclos):  object 'abs_freq'
not found
## Error in eval(expr, envir, enclos):  object 'rel_freq'
not found
## Error in plot(x$breaks, y = c(0, cum_absfreq[-8]), axes =
FALSE, main = "", :  object 'x' not found
## Error in axis(side = 1, at = seq(50, 120, 10), line =
-0.7):  plot.new has not been called yet
## Error in mtext("price", side = 1, line = 2):  plot.new
has not been called yet
## Error in axis(side = 2, at = seq(0, 32, 2), las = 2,
line = -1, cex.axis = 0.9):  plot.new has not been called
yet
## Error in mtext("cumulative frequency", side = 2, line =
```

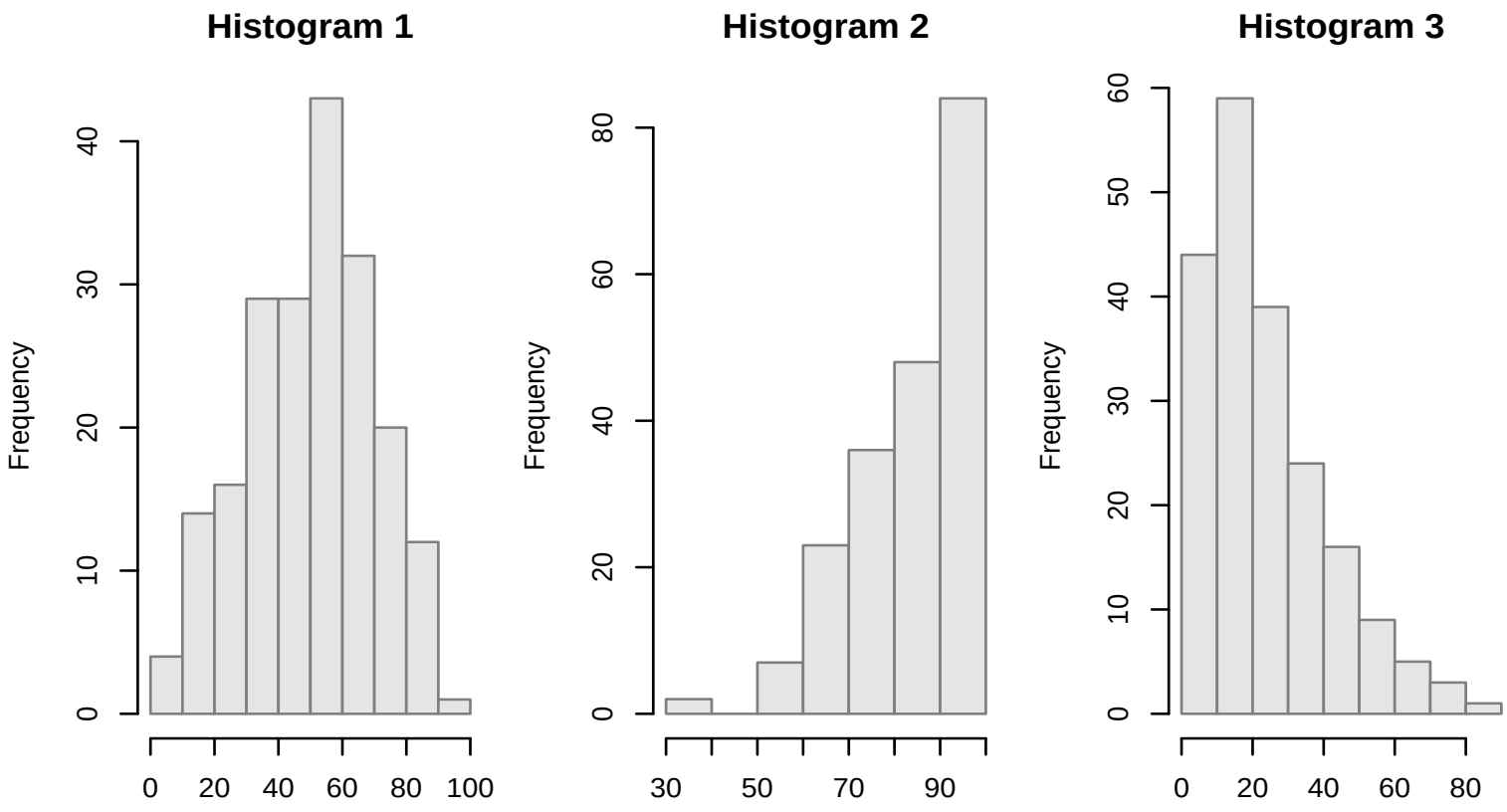
Ogives

Building an Ogive

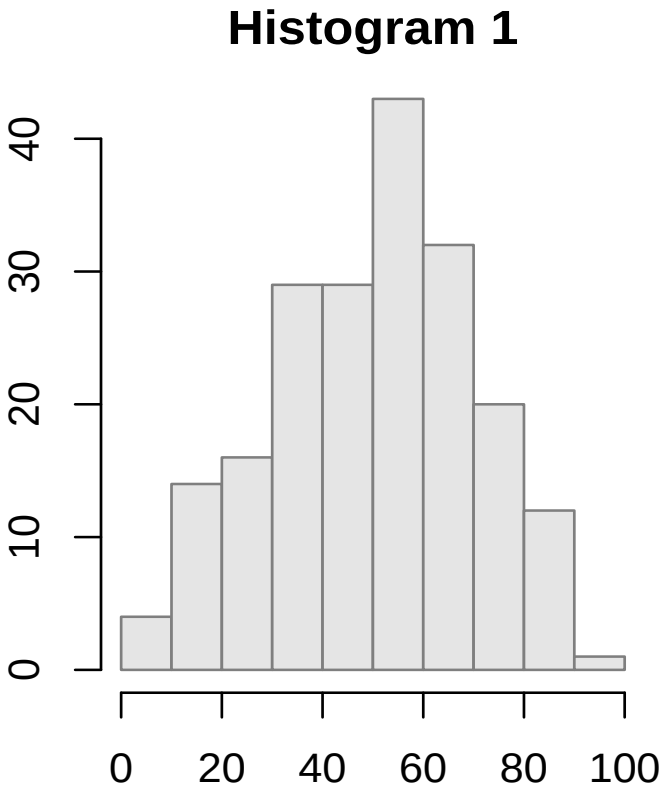
- Make a frequency table showing bin intervals and cumulative frequencies.
- An ogive begins on the horizontal axis at the lower boundary of the first bin.
- For each bin, make a dot over the upper interval limit at the height of the cumulative frequency.
- Connect the dots with line segments.

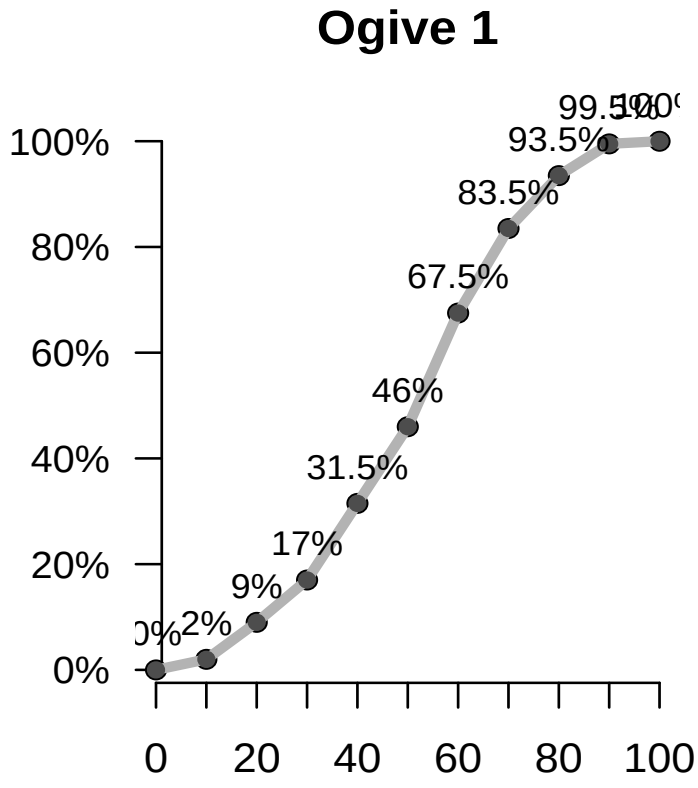
Distributions and Ogives

Three histograms

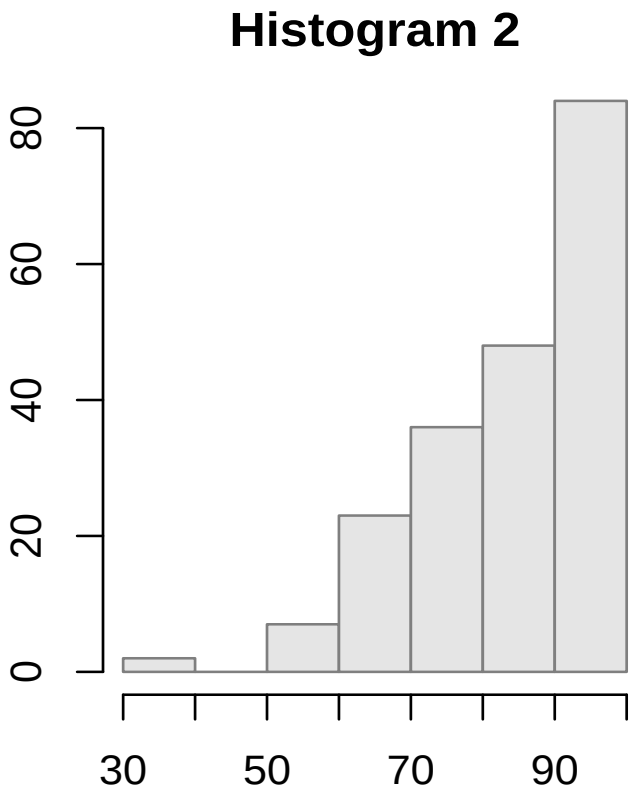


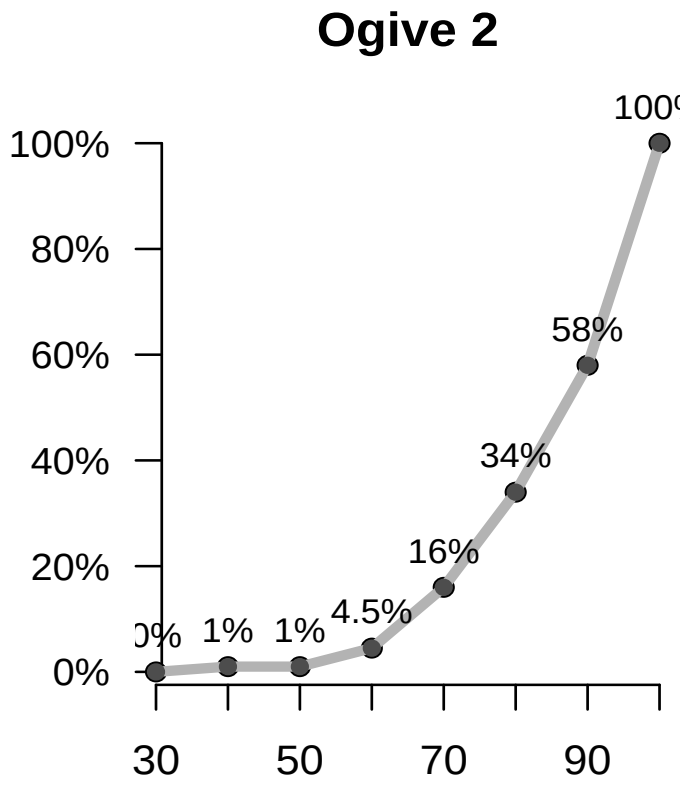
Symmetric Distribution





Skewed to the left Distribution





Skewed to the right Distribution