Descriptive Statistics: Histograms

Gaston Sanchez

Visualizing variability
by means of graphical
displays


NEW YORK KNICKS
CLEVELAND CAVALIERS
CHICAGO BULLS
DETROIT PISTONS
ATLANTA HAWKS
BOSTON CELTICS
TORONTO RAPTORS
DENVER NUGGETS
ORLANDO MAGIC
LOS ANGELES CLIPPERS
HOUSTON ROCKETS
MINNESOTA TIMBERWOLVES
GOLDEN STATE WARRIORS
PORTLAND TRAIL BLAZERS
PHOENIX SUNS
WIZARDS
THUNDER OKC
MIAMI HEAT
NBA
NEW ORLEANS HORNETS
CHARLOTTE BOBCATS
LOS ANGELES LAKERS
Pacers
UTAH JAZZ
NETS
BROOKLYN
MILWAUKEE BUCKS
SACRAMENTO KINGS
MEMPHIS GRIZZLIES
SAN ANTONIO SPURS
DALLAS MAVERICKS
76ers
PHILADELPHIA

NBA season 2015-2016

player	team	player_num	birthdate	age	country	position	height	weight	experience	salary
Al Horford	ATL	15	6/3/86	29	do	center	82	245	8	12000000
Dennis Schroder	ATL	17	9/15/93	22	de	point guard	73	172	2	1763400
Jeff Teague	ATL	0	6/10/88	27	us	point guard	74	186	6	8000000
Justin Holiday	ATL	7	4/5/89	26	us	shooting guard	78	185	2	NA
Kent Bazemore	ATL	24	7/1/89	26	us	small forward	77	201	3	2000000
Kirk Hinrich	ATL	12	1/2/81	35	us	point guard	76	190	12	2870000
Kris Humphries	ATL	43	2/6/85	30	us	power forward	81	235	11	388025
Kyle Korver	ATL	26	3/17/81	34	us	shooting guard	79	212	12	5746479
Lamar Patterson	ATL	13	8/12/91	24	us	shooting guard	77	225	0	525093
Mike Muscala	ATL	31	7/1/91	24	us	center	83	240	2	947276
Mike Scott	ATL	32	7/16/88	27	us	power forward	80	237	3	3333333
Paul Millsap	ATL	4	2/10/85	30	us	power forward	80	246	9	19000000
Shelvin Mack	ATL	8	4/22/90	25	us	point guard	75	203	4	NA
Thabo Sefolosha	ATL	25	5/2/84	31	ch	small forward	79	220	9	4000000
Tiago Splitter	ATL	11	1/1/85	31	br	center	83	245	5	8500000
Tim Hardaway	ATL	10	3/16/92	23	us	shooting guard	78	205	2	1304520
Walter Tavares	ATL	22	3/22/92	23	cv	center	87	260	0	1000000
Amir Johnson	BOS	90	5/1/87	28	us	power forward	81	240	10	12000000
Avery Bradley	BOS	0	11/26/90	25	us	shooting guard	74	180	5	7730337
Coty Clarke	BOS	63	7/4/92	23	us	small forward	79	232	0	61776

Data file "nba_players.csv" available in the course's github repository

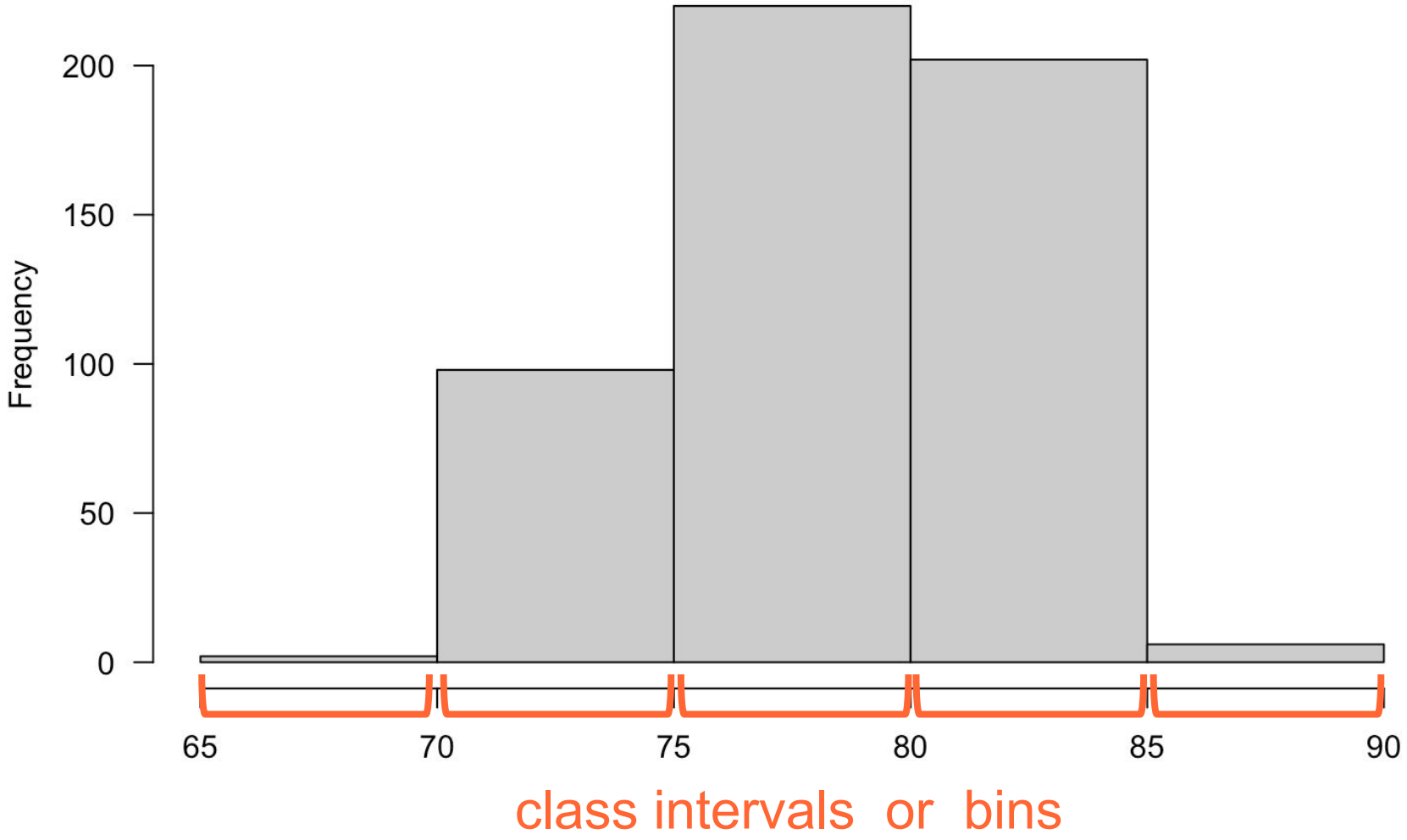
Histograms

Histograms
(not *Instagrams*)

Height (measured in inches)

player	team	player_num	birthdate	age	country	position	height	weight	experience	salary
Al Horford	ATL	15	6/3/86	29	do	center	82	245	8	12000000
Dennis Schroder	ATL	17	9/15/93	22	de	point guard	73	172	2	1763400
Jeff Teague	ATL	0	6/10/88	27	us	point guard	74	186	6	8000000
Justin Holiday	ATL	7	4/5/89	26	us	shooting guard	78	185	2	NA
Kent Bazemore	ATL	24	7/1/89	26	us	small forward	77	201	3	2000000
Kirk Hinrich	ATL	12	1/2/81	35	us	point guard	76	190	12	2870000
Kris Humphries	ATL	43	2/6/85	30	us	power forward	81	235	11	388025
Kyle Korver	ATL	26	3/17/81	34	us	shooting guard	79	212	12	5746479
Lamar Patterson	ATL	13	8/12/91	24	us	shooting guard	77	225	0	525093
Mike Muscala	ATL	31	7/1/91	24	us	center	83	240	2	947276
Mike Scott	ATL	32	7/16/88	27	us	power forward	80	237	3	3333333
Paul Millsap	ATL	4	2/10/85	30	us	power forward	80	246	9	19000000
Shelvin Mack	ATL	8	4/22/90	25	us	point guard	75	203	4	NA
Thabo Sefolosha	ATL	25	5/2/84	31	ch	small forward	79	220	9	4000000
Tiago Splitter	ATL	11	1/1/85	31	br	center	83	245	5	8500000
Tim Hardaway	ATL	10	3/16/92	23	us	shooting guard	78	205	2	1304520
Walter Tavares	ATL	22	3/22/92	23	cv	center	87	260	0	1000000
Amir Johnson	BOS	90	5/1/87	28	us	power forward	81	240	10	12000000
Avery Bradley	BOS	0	11/26/90	25	us	shooting guard	74	180	5	7730337
Coty Clarke	BOS	63	7/4/92	23	us	small forward	79	232	0	61776

Histogram of players height



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

About Histograms

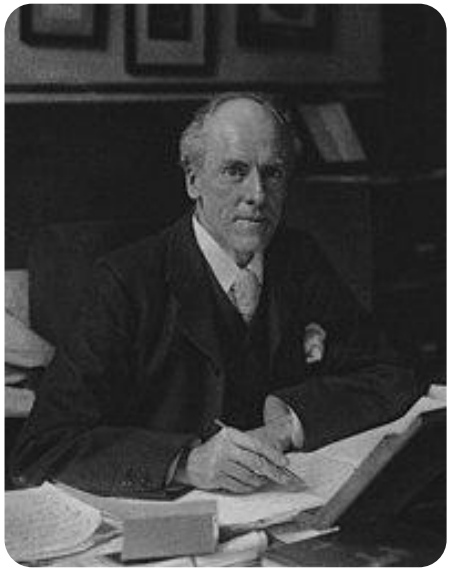
histograms $\neq$ bar charts

Histograms are very similar to bar charts, but the way bars are constructed is different.

Etimology

Histo: from Greek *"isto-s"* = "mast" (vertical)

Gram: from Greek *"gram-ma"* = draw/graph



Term coined by English statistician Karl Pearson (~ 1891), inspired by French economist and geographer Emile Levasseur (1885)

Histograms

"A bar graph plot of the data, with the bars placed adjacent to to each other, typically used in frequency distributions."

Histograms are statistical charts, usually for technical use (not for "mass" consumption: you won't see a histogram on TV)

About histograms

The bins represent ranges of values.

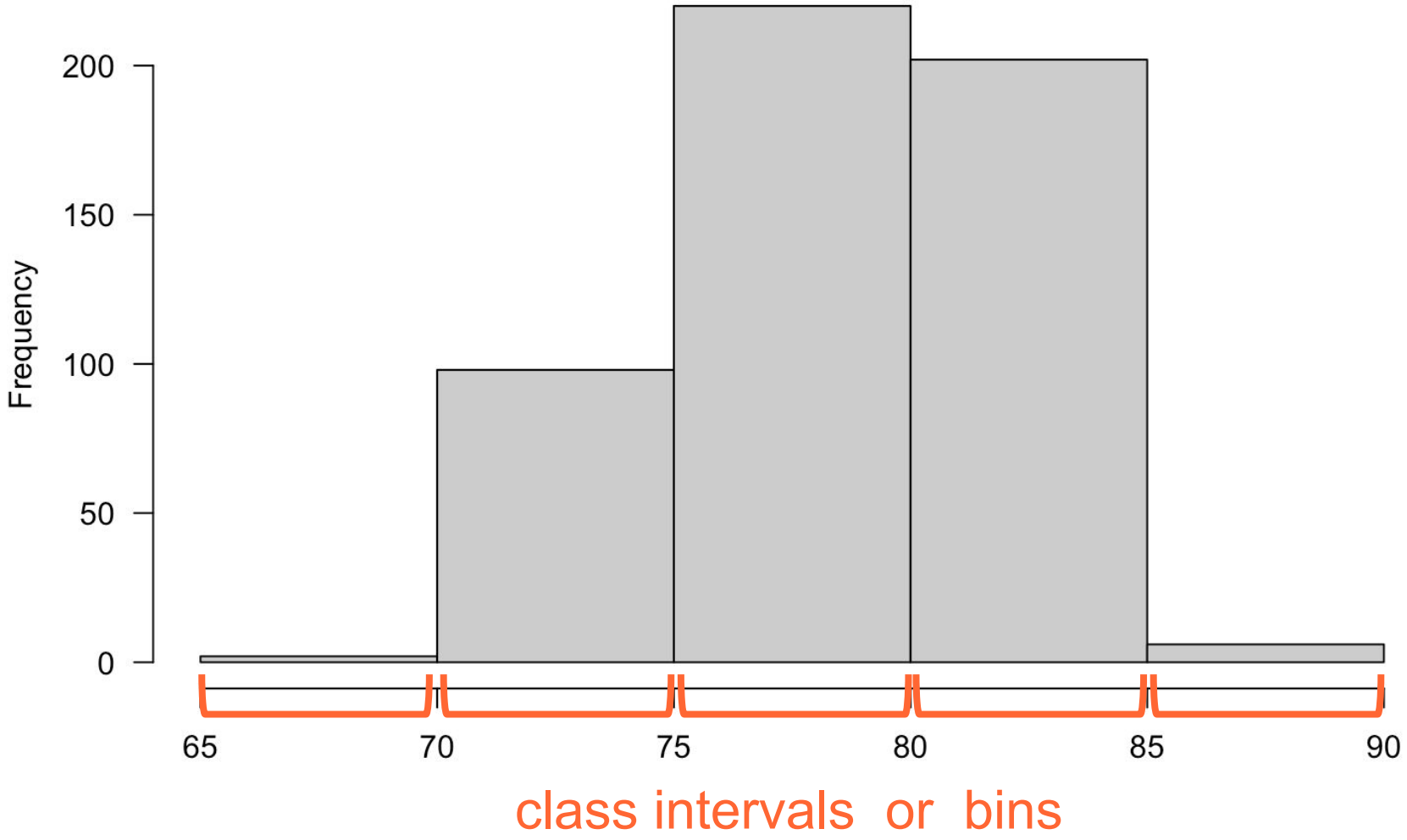
The bins (intervals) must be adjacent, and usually of equal size.

The length of the bar is not that important.

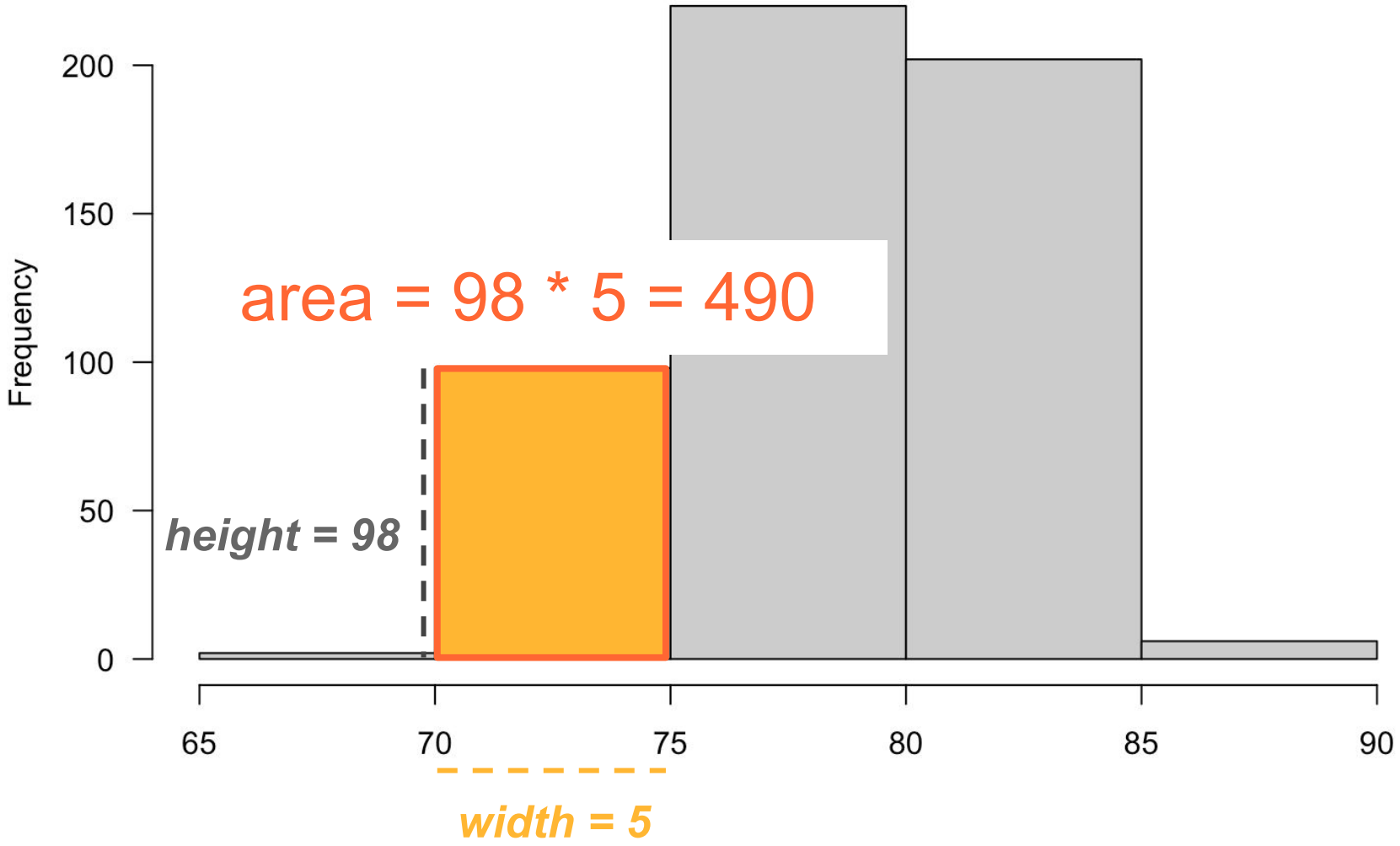
What really matters is the area of the bars: they are proportional to the relative frequencies.

Reading a histogram

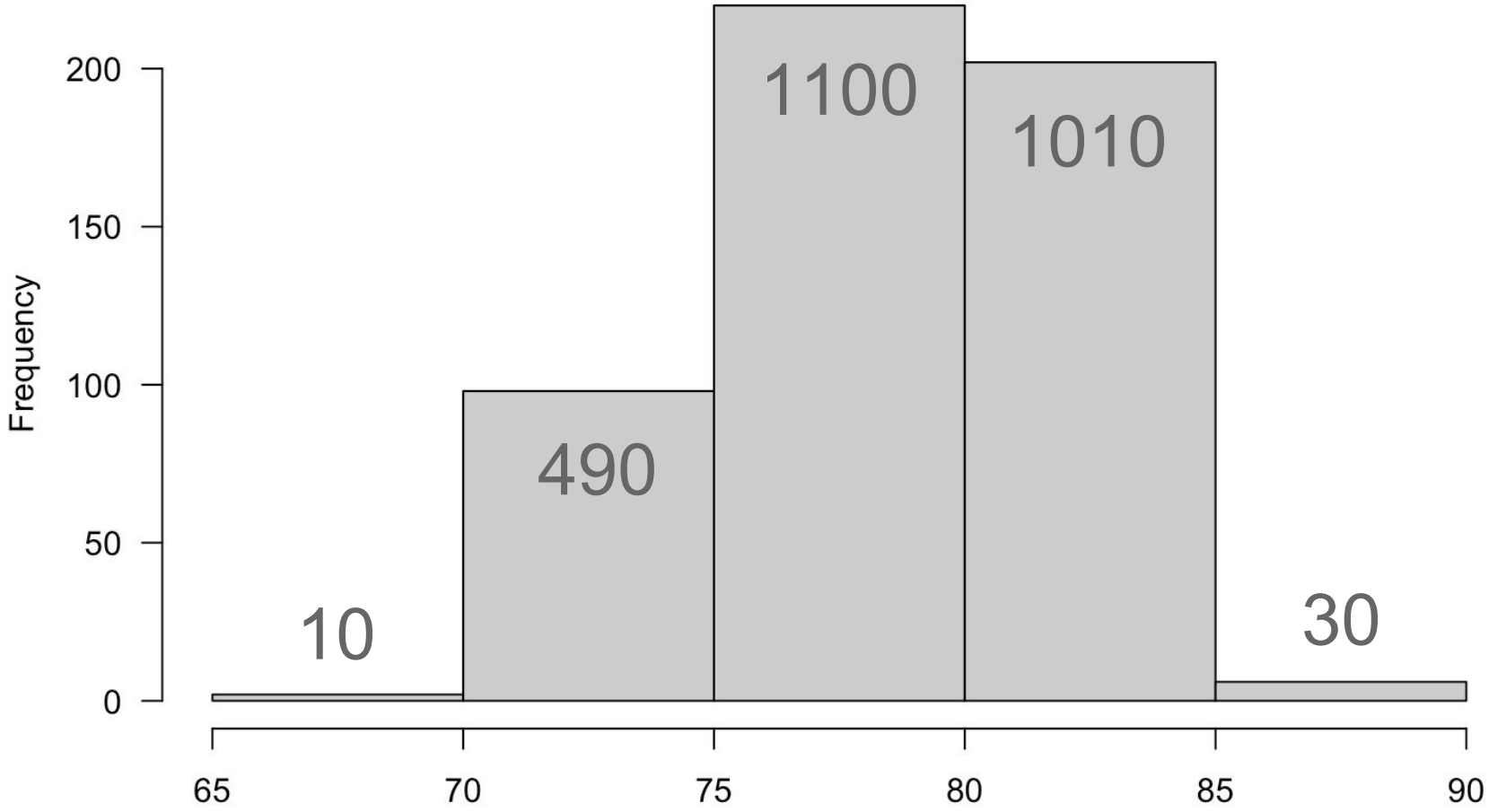
Histogram of players height



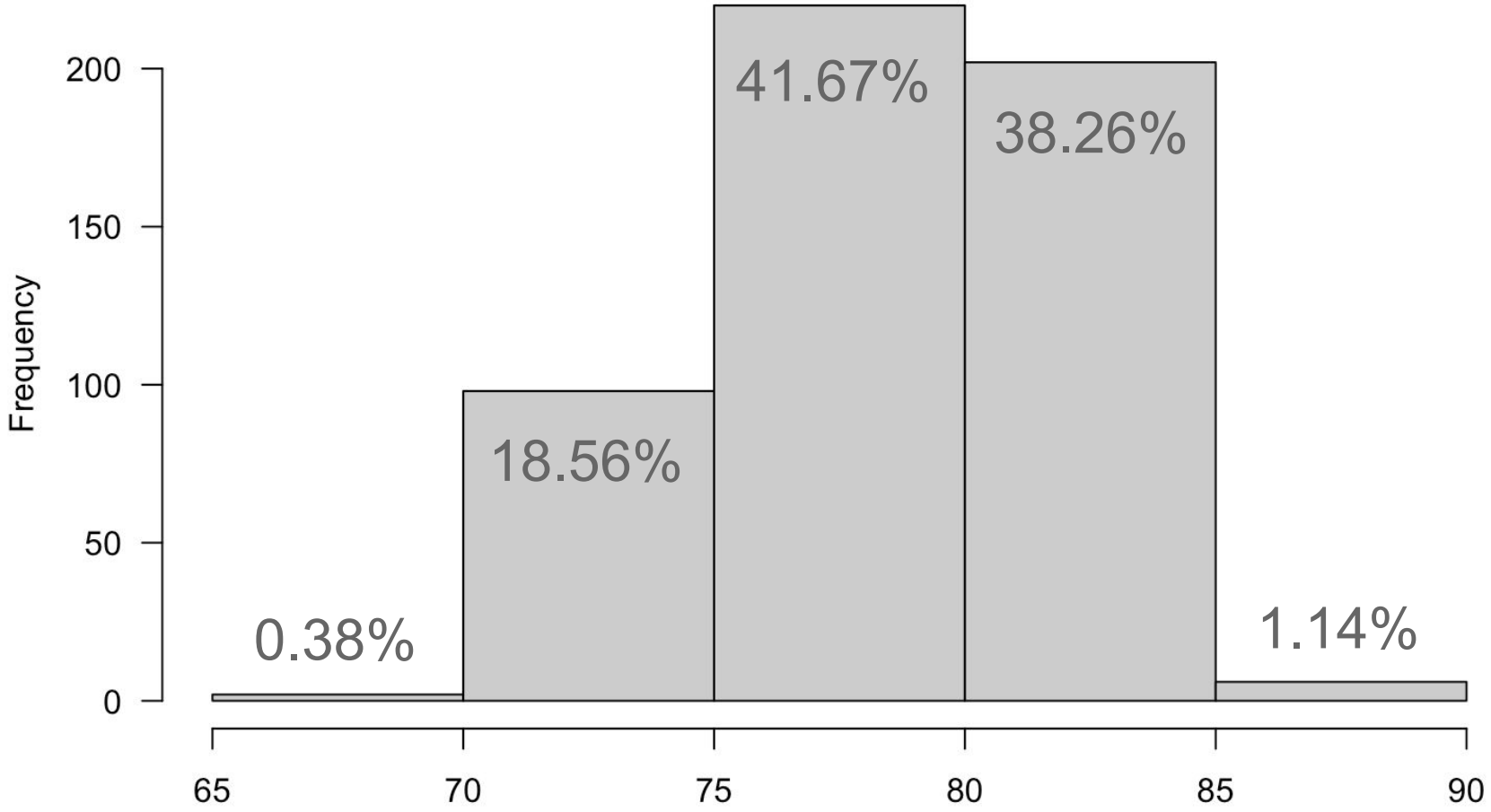
Histogram of players height (5 bins)



Total Area = 2640

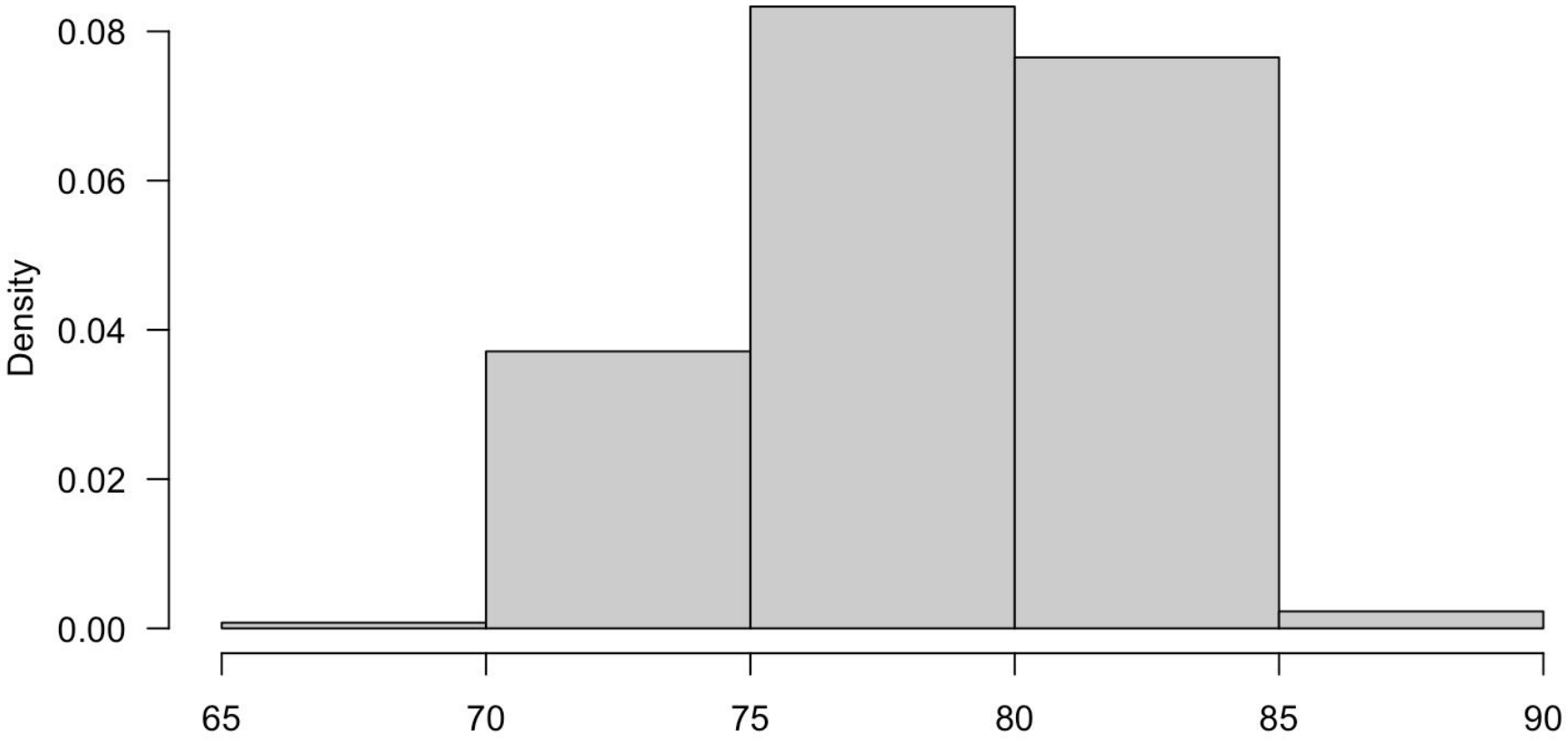


Total Area = 100%



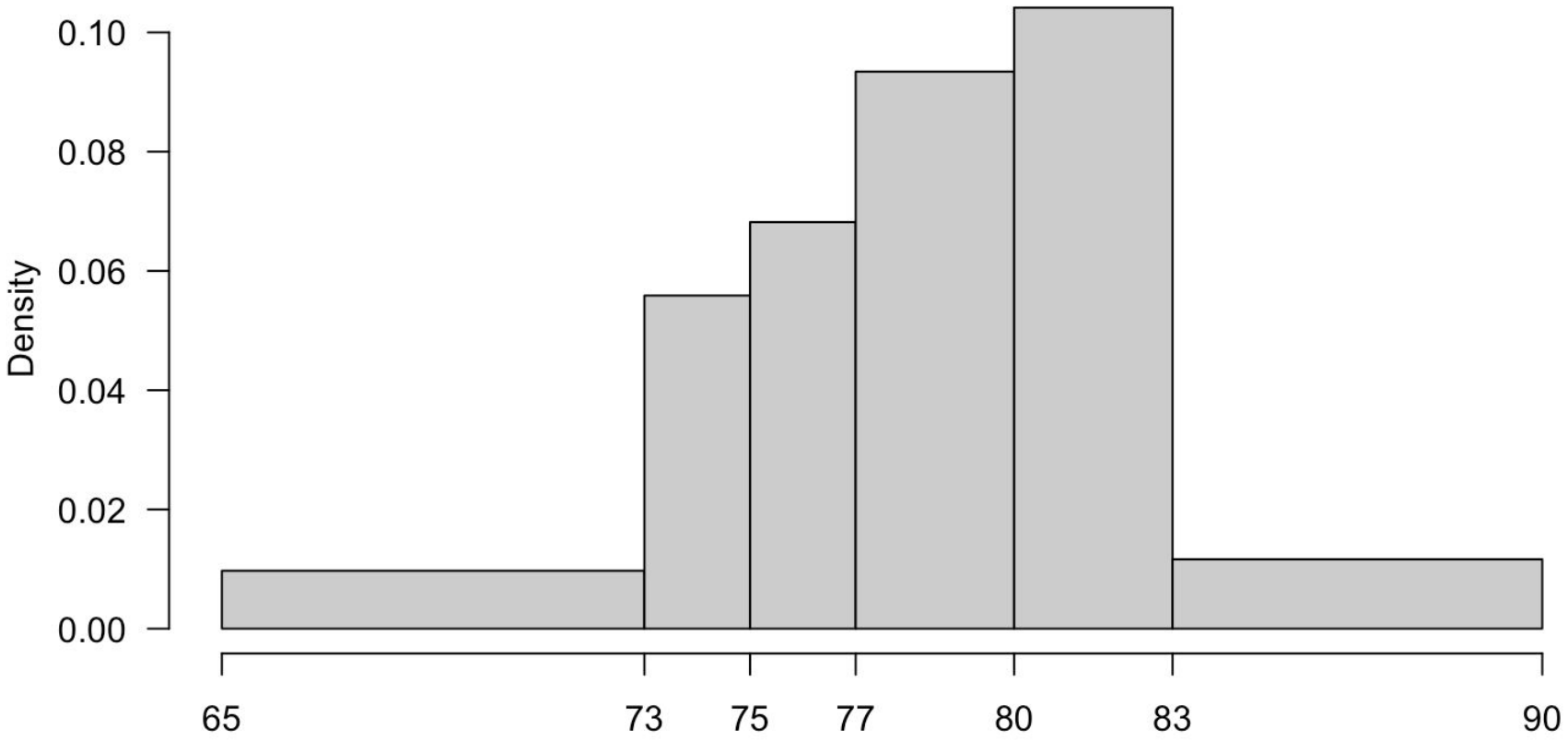
The area of a bar gives the proportion of data values which fall in the bin

Histogram with density scale

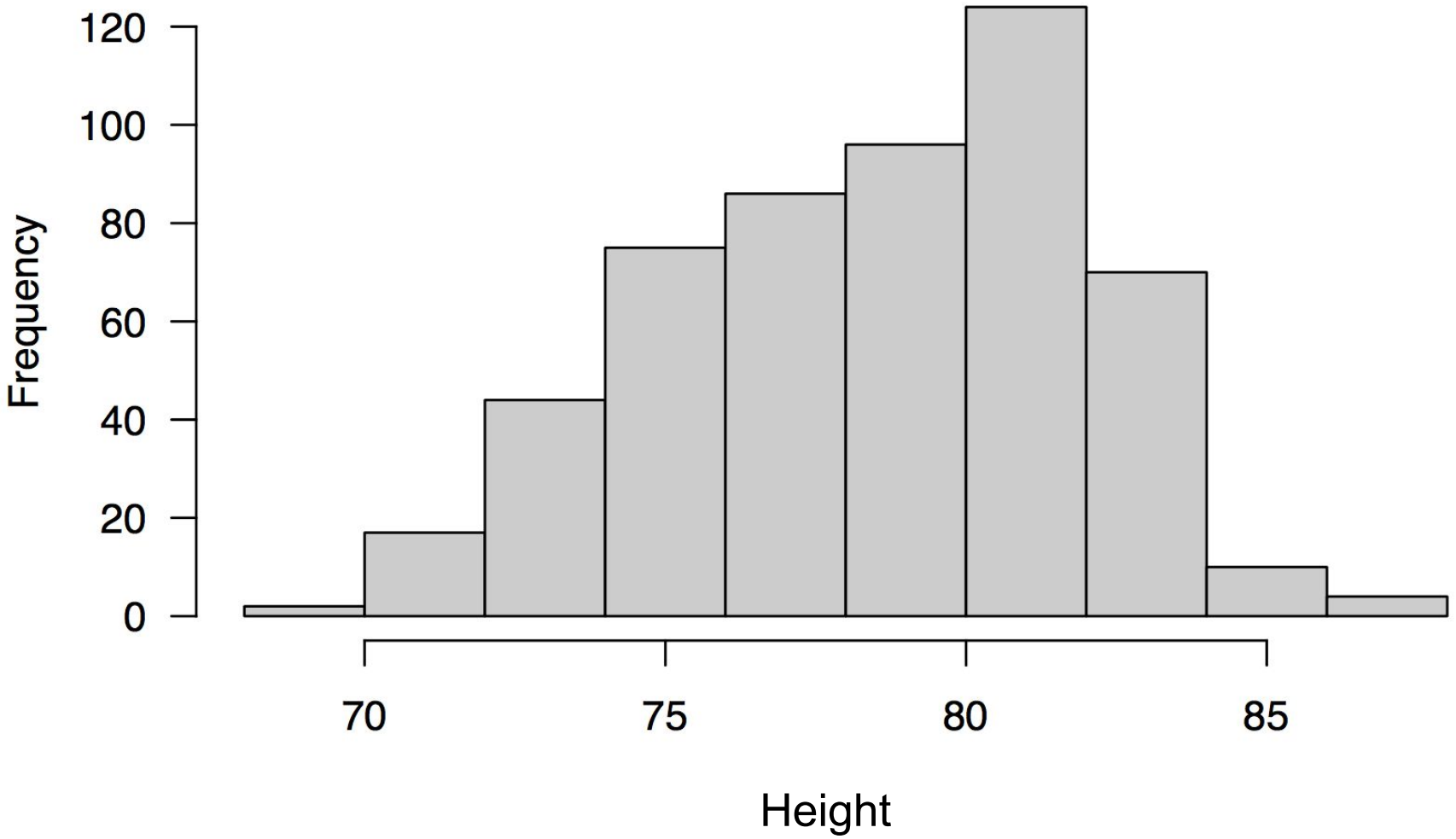


The area of a bar gives the proportion of data values which fall in the bin

Class intervals of different width

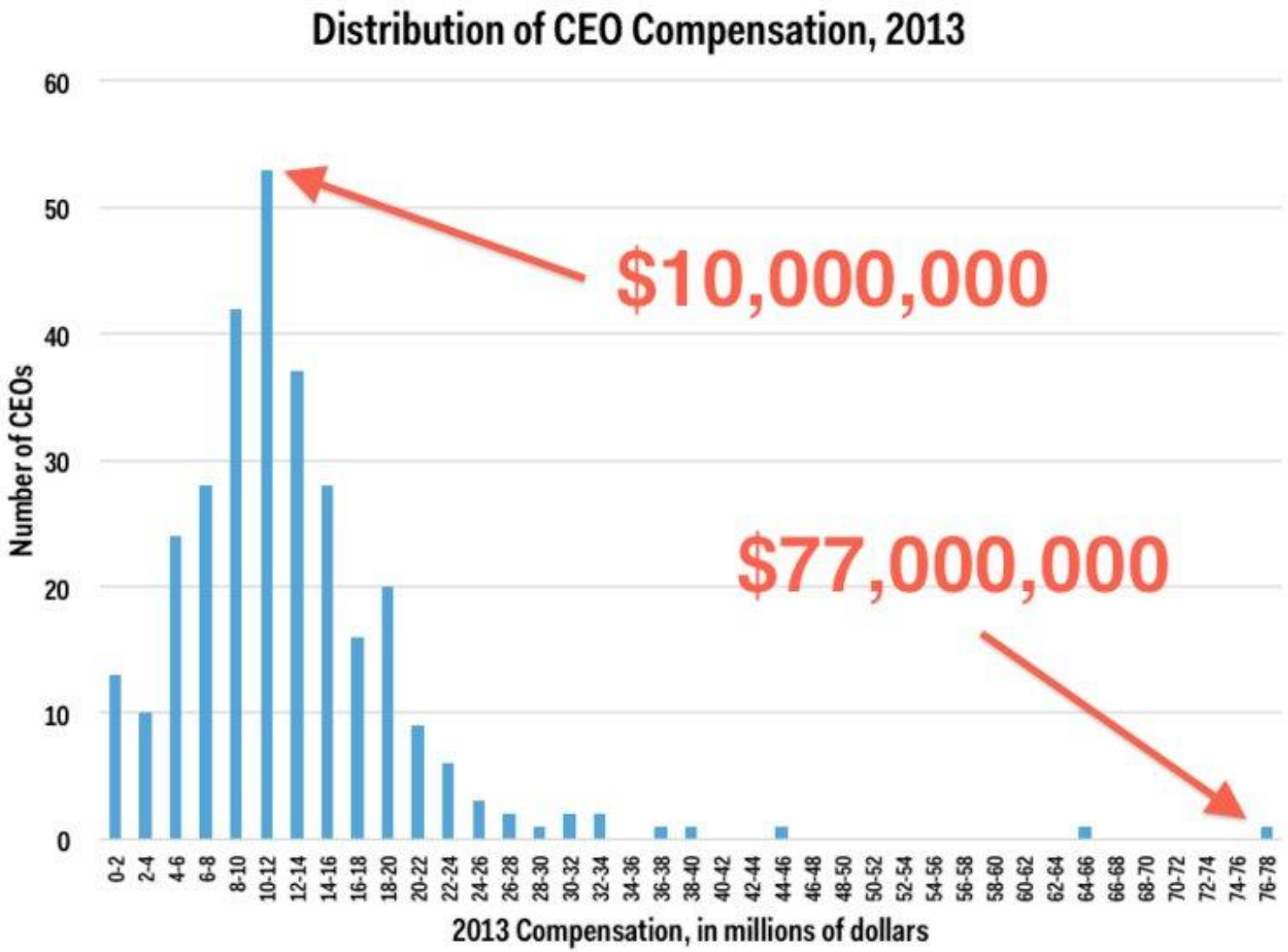


Histogram with more bins



Some Examples

We took the CEO pay data and made a histogram chart showing the distribution of compensation amounts. Each column shows the number of CEOs whose compensation fell into each $2,000,000 wide bracket:

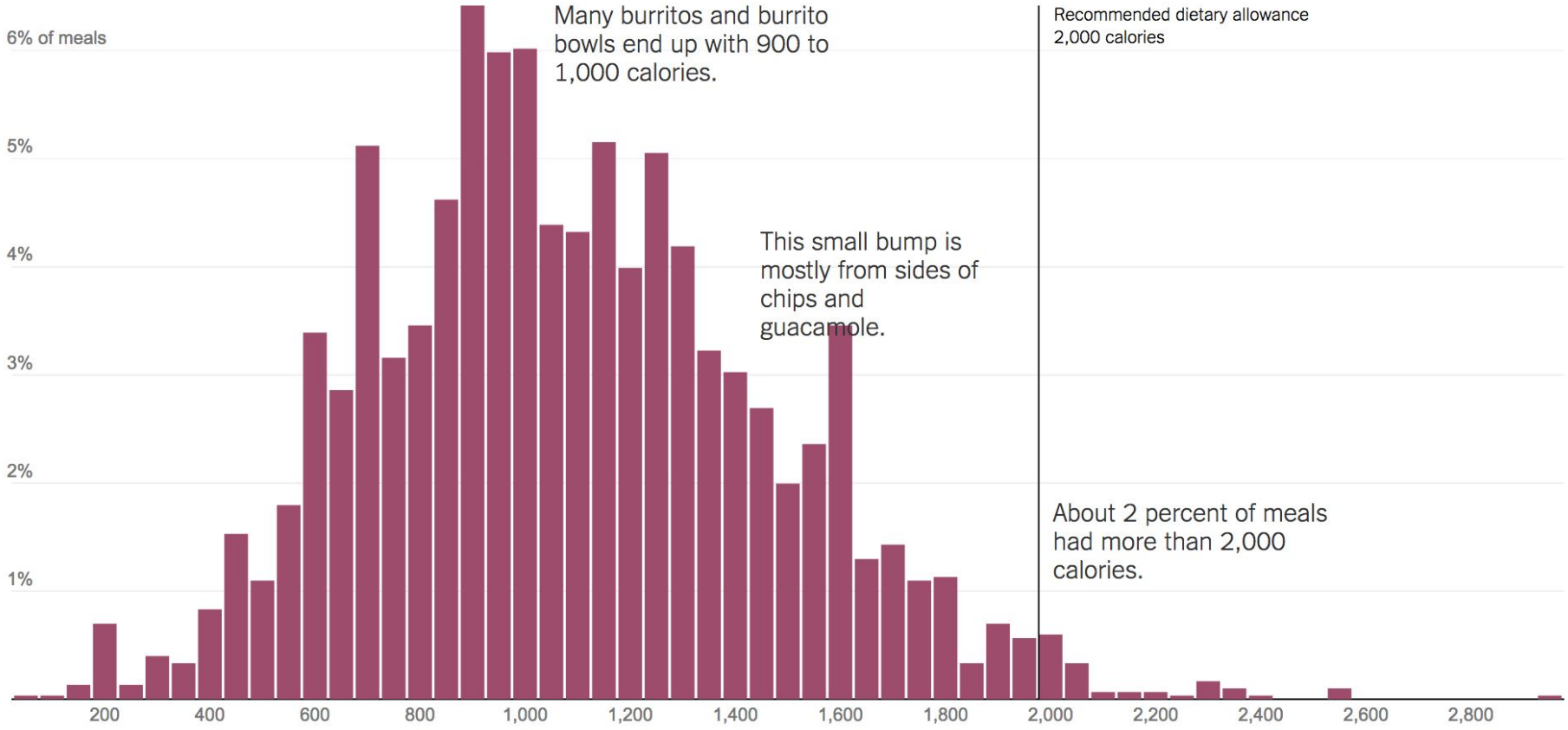


https://www.businessinsider.com/ceo-compensation-histogram-chart-2014-5

At Chipotle, How Many Calories Do People Really Eat?

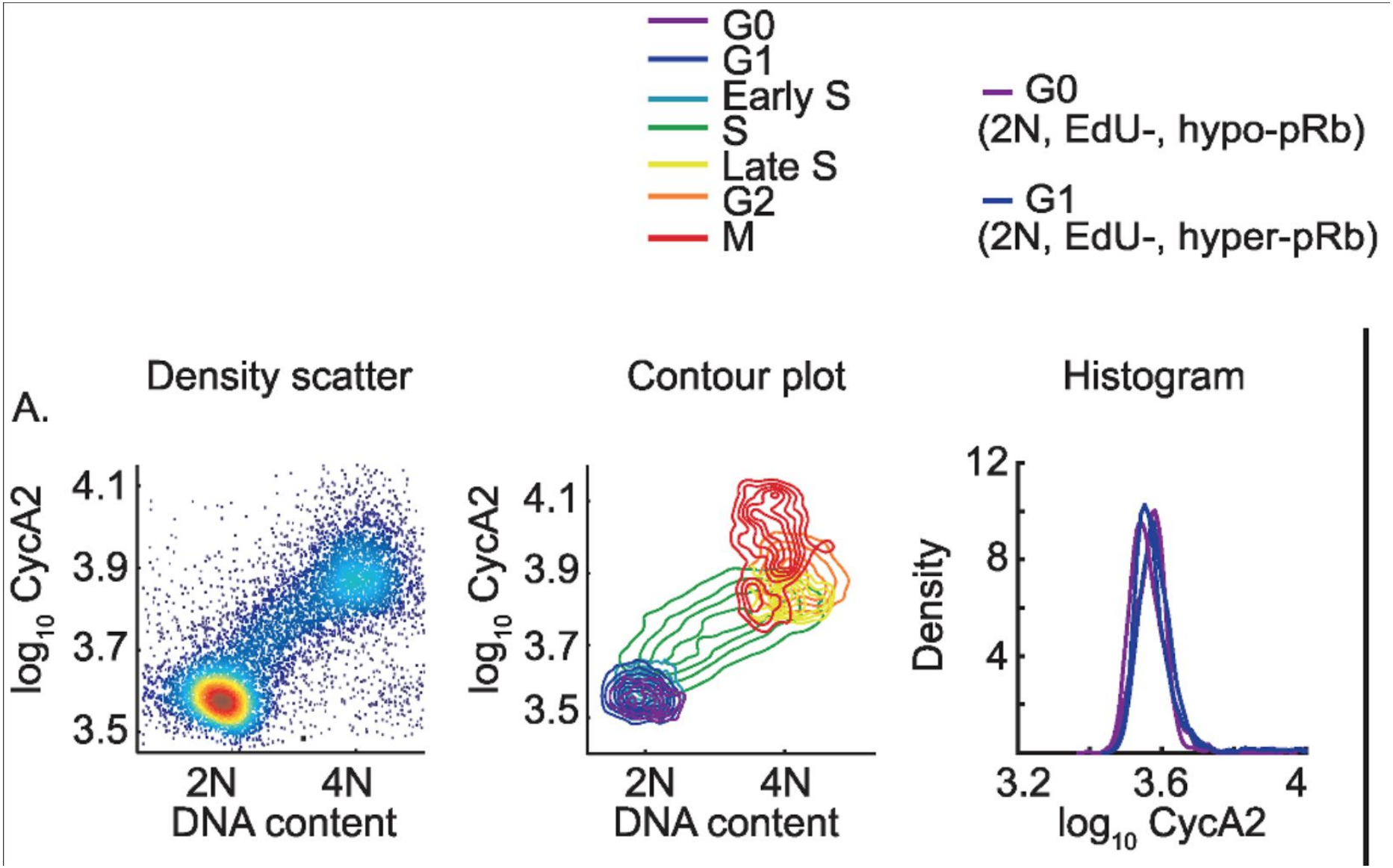
By **KEVIN QUEALY, AMANDA COX and JOSH KATZ** FEB. 17, 2015

Most meals have more than 1,000 calories and almost a full day's worth of sodium. **RELATED ARTICLE**



https://www.nytimes.com/interactive/2015/02/17/upshot/what-do-people-actually-order-at-chipotle.html

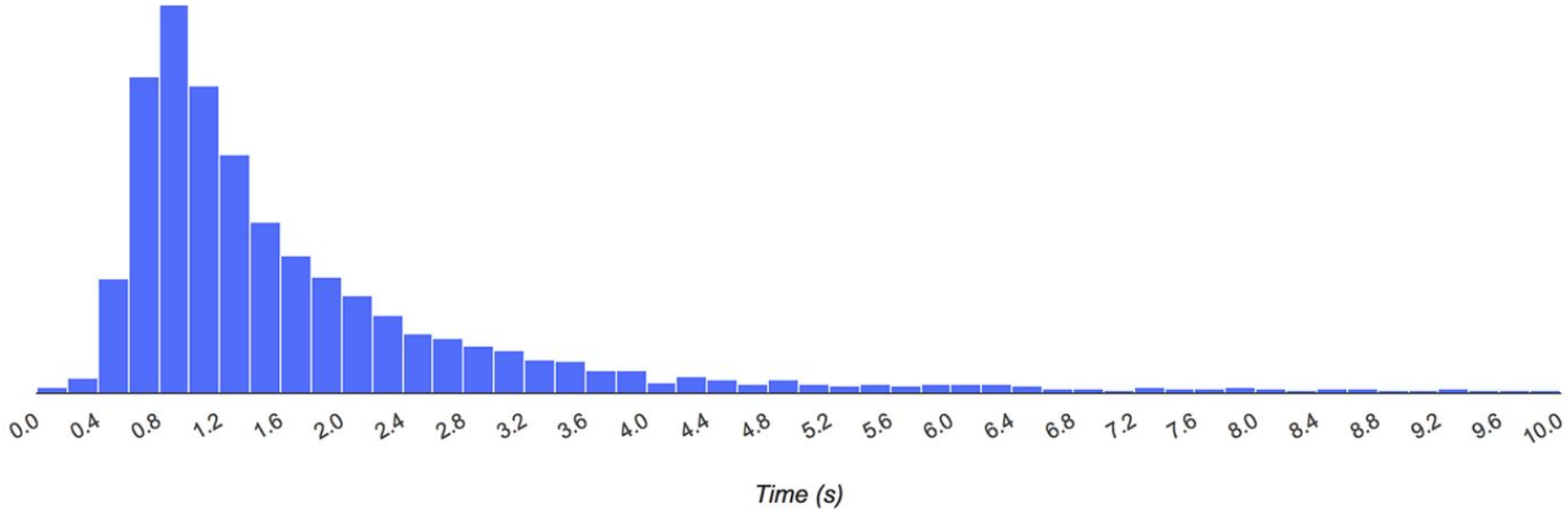
A map of protein dynamics during cell-cycle progression and cell-cycle exit



https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2003268

User-centric Performance Metrics (2018 Chrome Dev Summit)

In reality, your app's load time is the collection of all load times from every individual user, and the only way to fully represent that is with a distribution like in the histogram below:



https://developers.google.com/web/fundamentals/performance/user-centric-performance-metrics

How to draw a histogram?

How to build 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

Considerations for drawing histograms

How **many** class intervals?

Width of class intervals: (equal or not)?

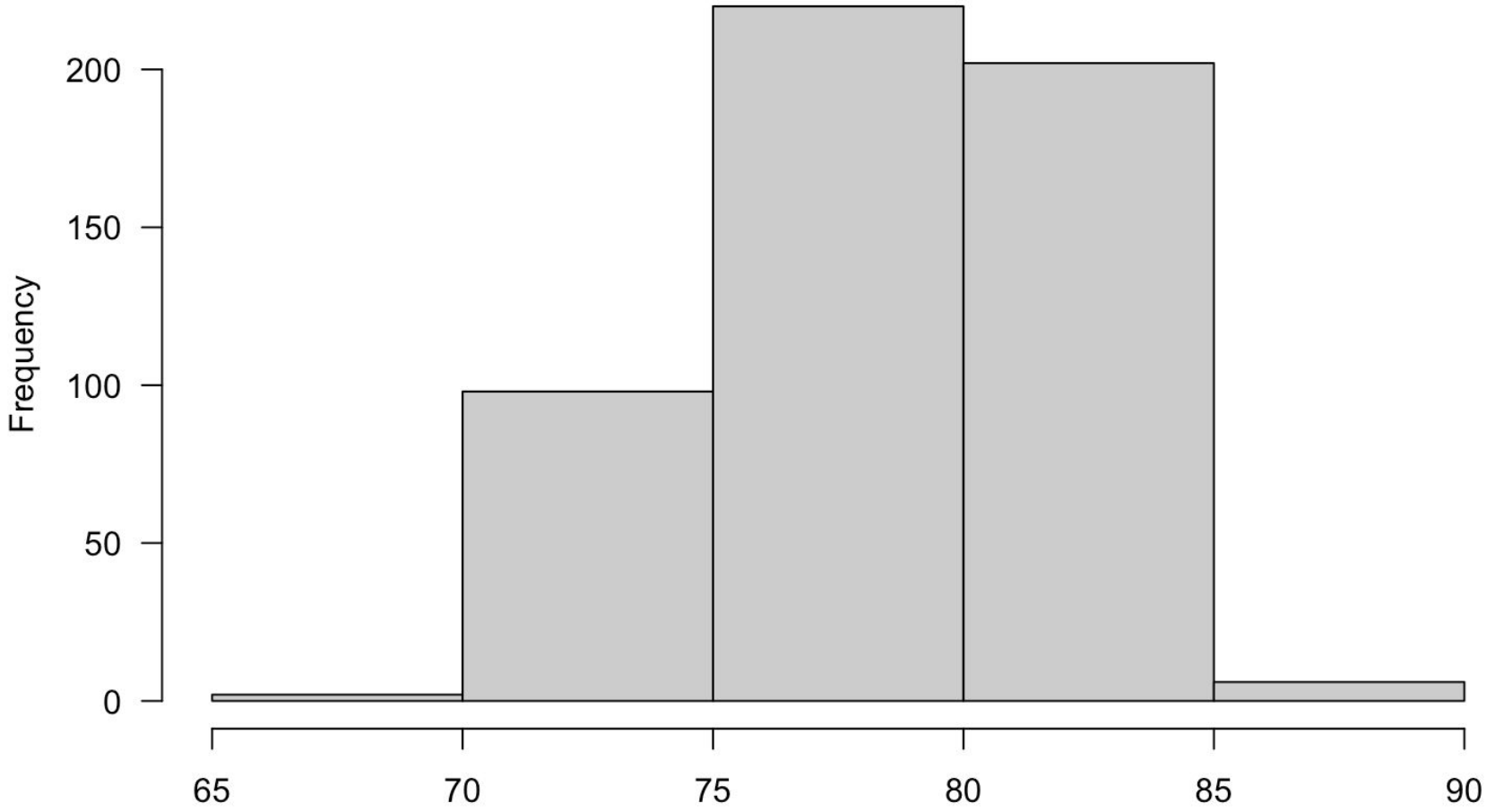
What **endpoint** should be included in a class interval?

What **scale** will be used?

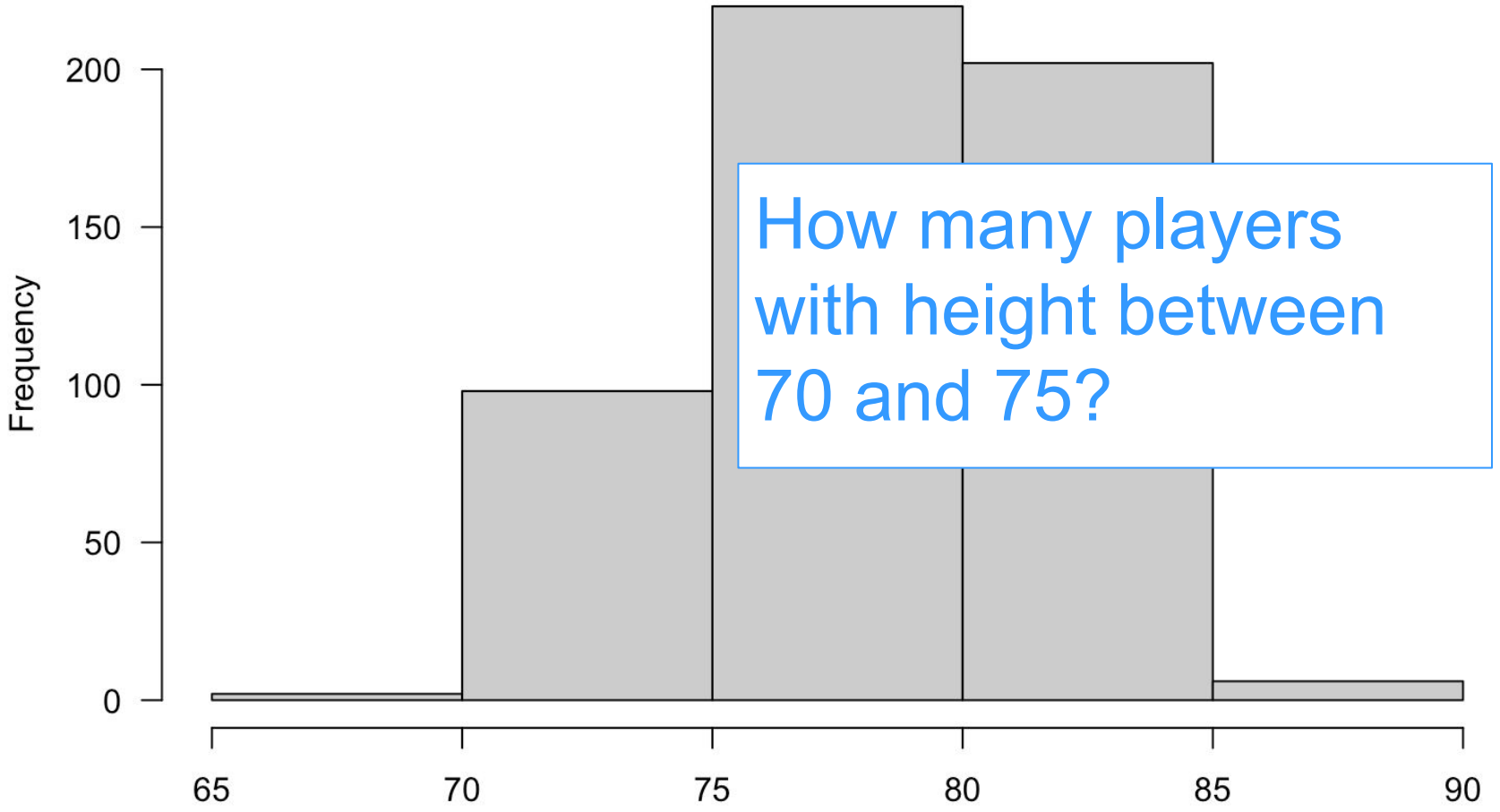
It's always good to try different # of bins

There’s a price to pay

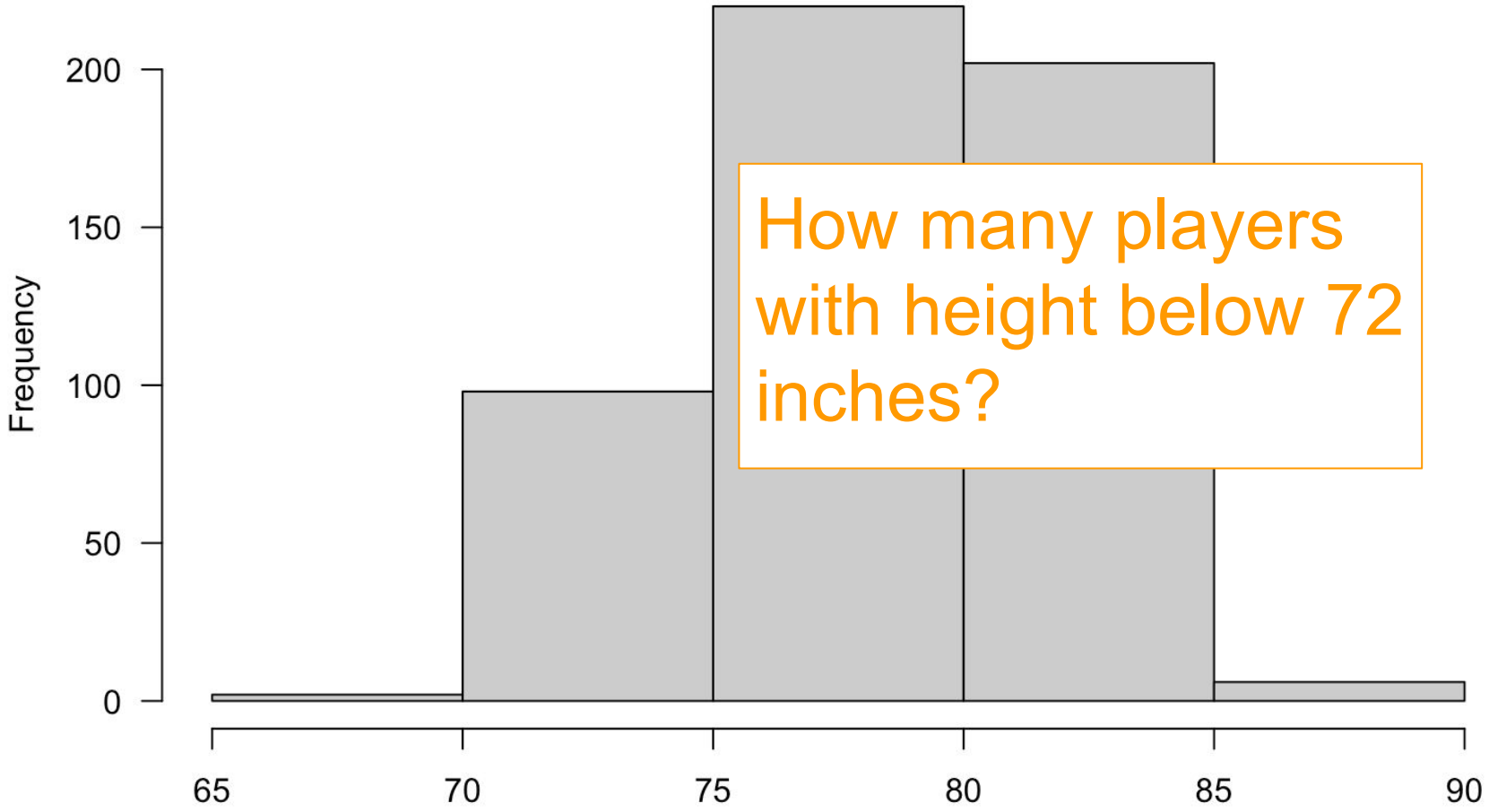
Histogram of NBA players height



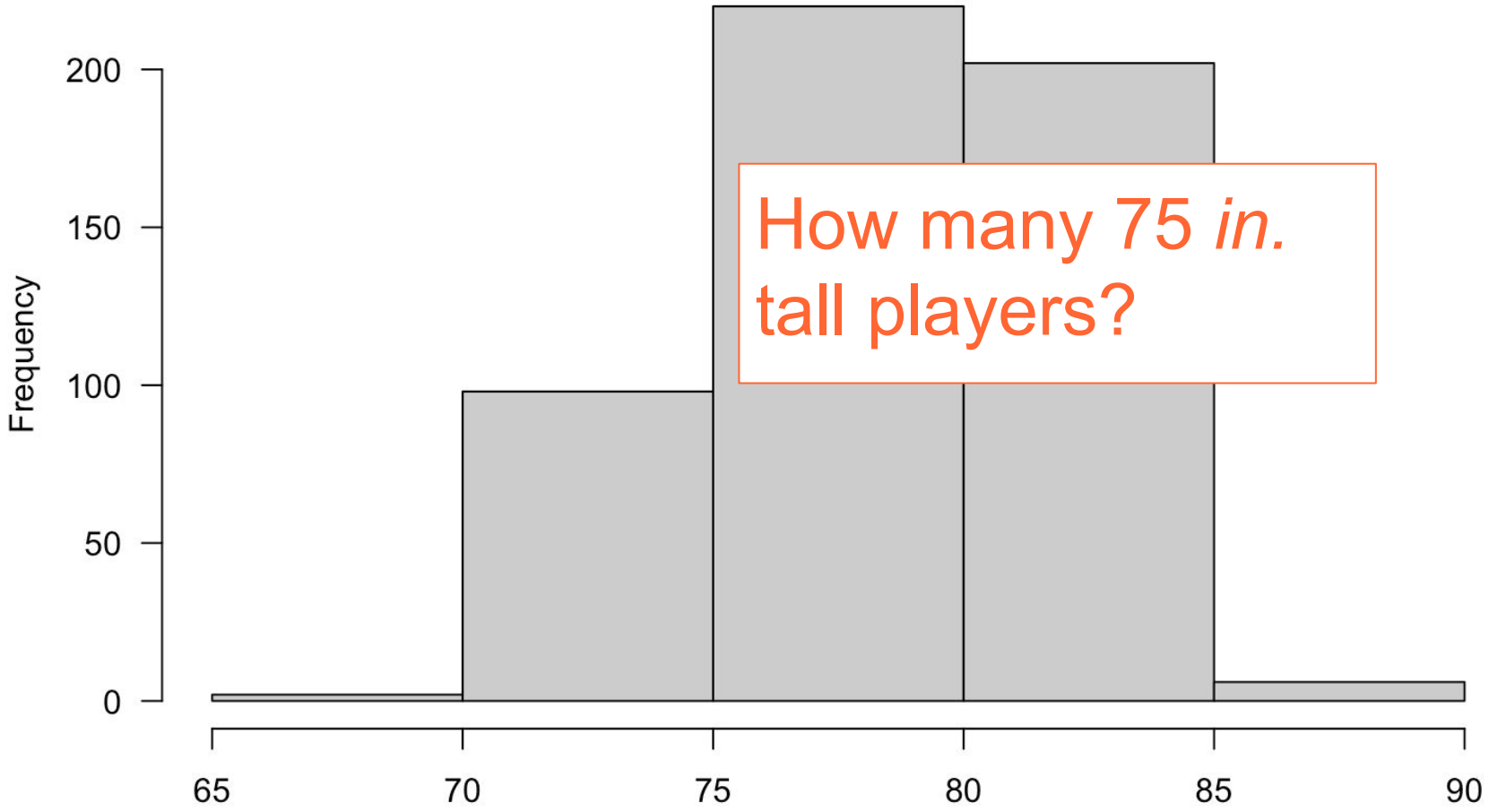
Histogram of NBA players height



Histogram of NBA players height



Histogram of NBA players height



Distribution Shapes

What sould we pay attention to?

Key Visual Characteristics of Distributions

- Spread
- Center
- Shape

Spread

Spread is a simple measure of dispersion

How spread out the values are

It is the easiest characteristic of a distribution to discern

Center

Center or central tendency

"Middle" of a set of values

Value that is "most typical"

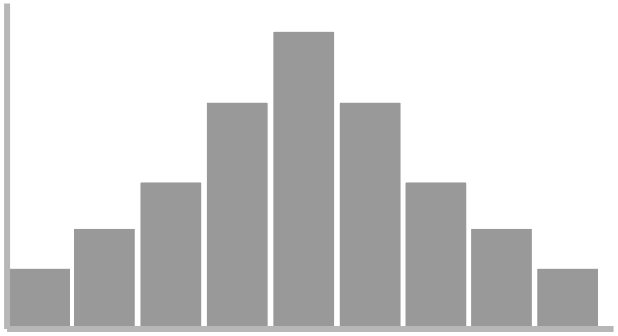
A "representative" value for the set of values as a whole

Shape

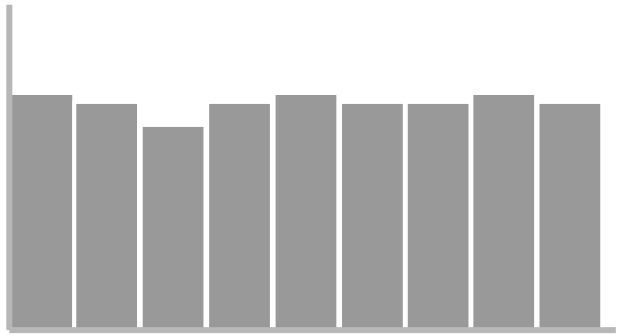
Shape or "profile"

Shows where the values are located throughout the spread

Distribution Shapes

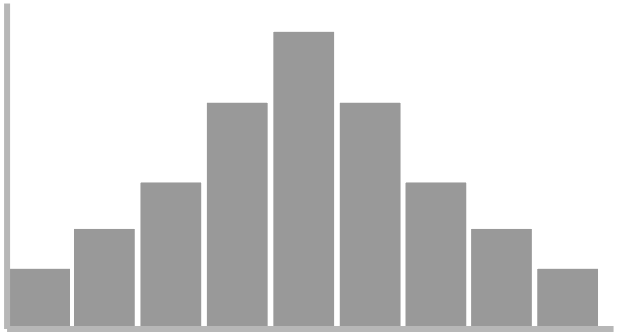


Curved

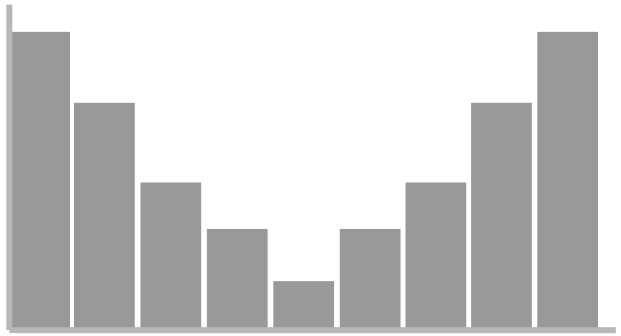


Flat or Uniform

Distribution Shapes

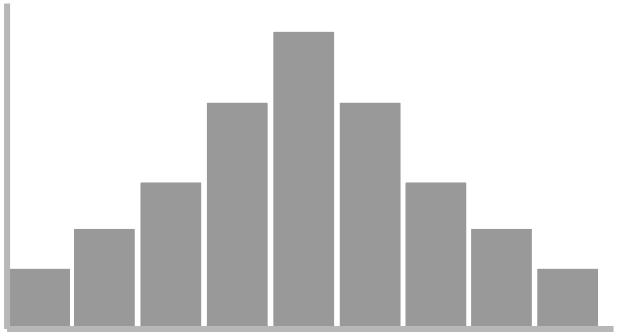


Curved Upward

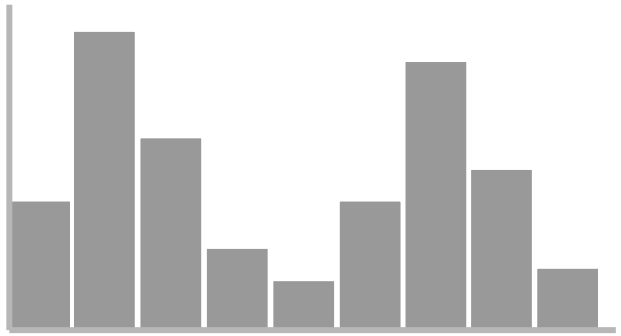


Curved Downward

Distribution Shapes

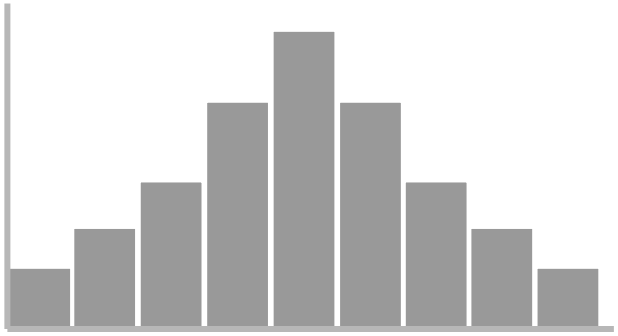


Single peak

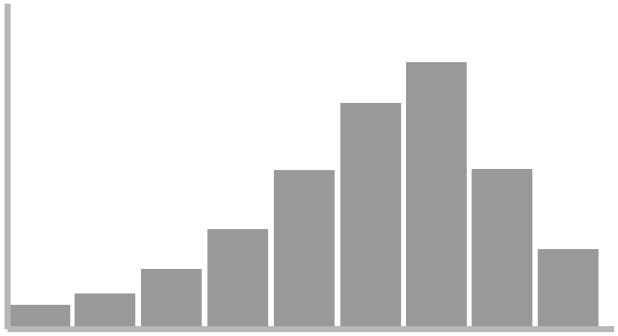


Multiple peaks
(e.g. bimodal, trimodal. etc)

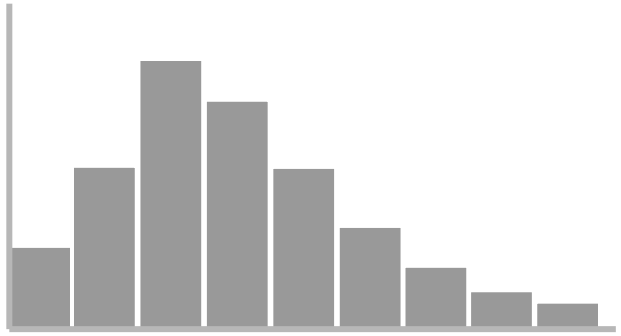
Distribution Shapes



Symmetrical



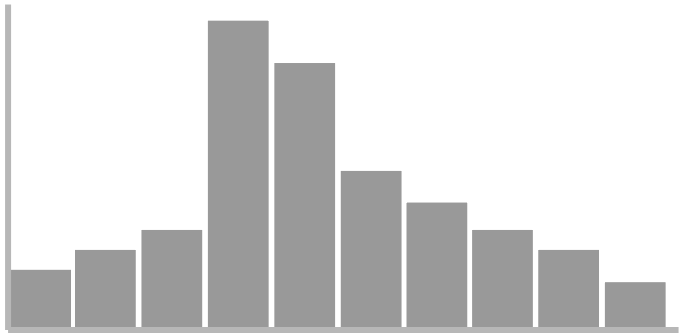
Skewed to the left

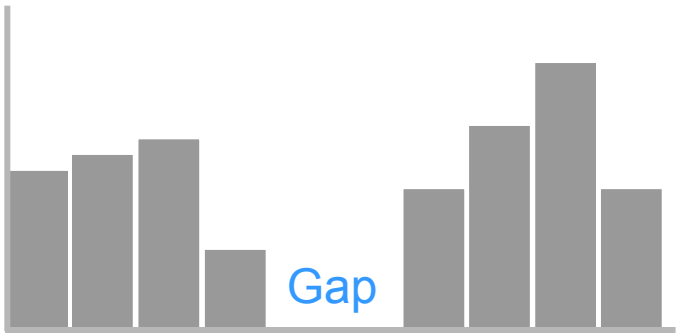


Skewed to the right

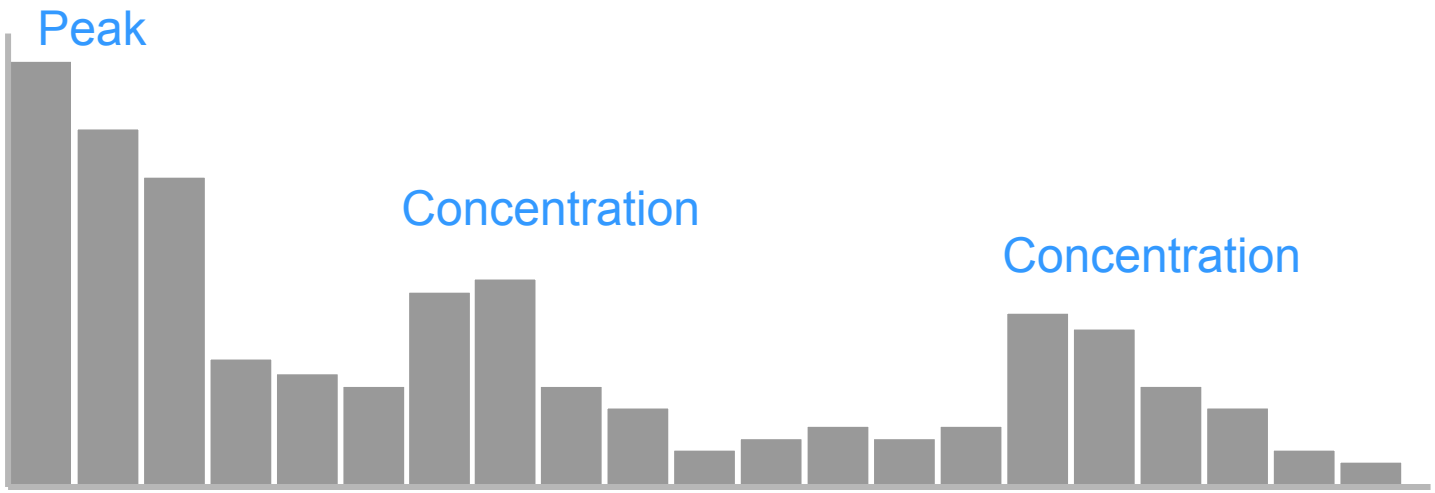
Distribution Shapes

Concentration

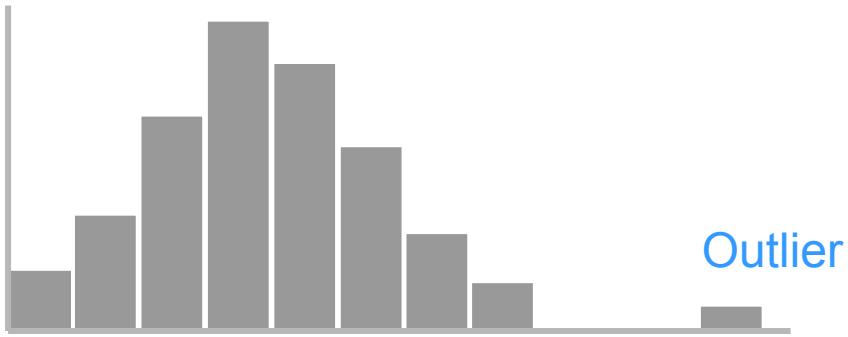




Distribution Shapes



Distribution Outliers



Final remarks

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.