Preview: Where Are We Going?

Descriptive Statistics

Numerical

- Average
- median
- percentiles
- standard deviation
- correlation coefficient

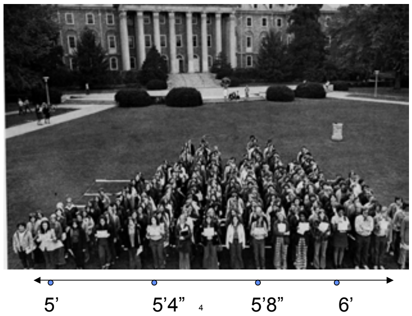
Graphical

- Histograms
- bar charts
- scatter diagrams

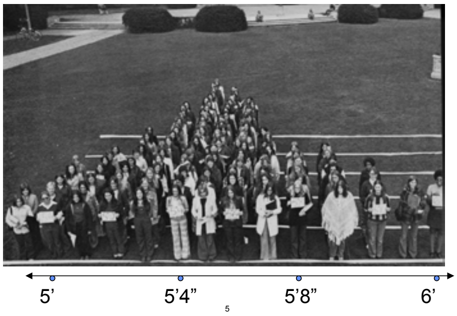
Why Descriptive Statistics?

Human beings cannot cope with more than a few numbers at once. Descriptive statistics are concise summaries.

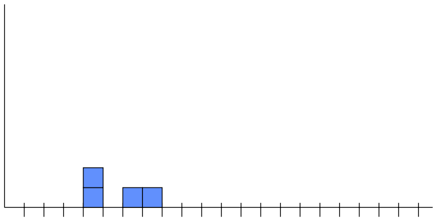
Histograms



Histogram of Just the Women



Histograms with Equal Bin Widths

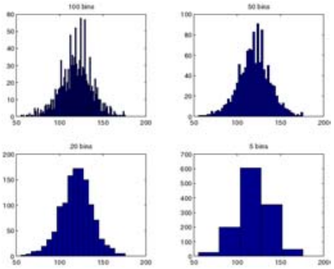


Bins
(need endpoint convention)

The Effects of Bin-Width

Durations (minutes) of eruptions of Old Faithful Geyser: a histogram

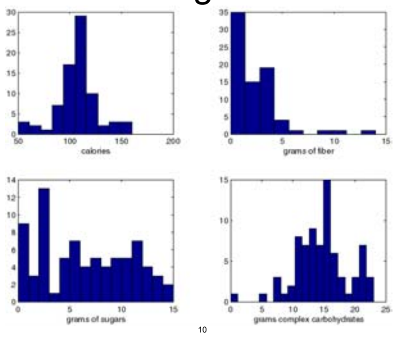
Birthweights of 1230 Male Babies



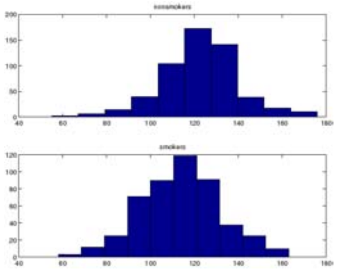
Histograms for Data Summary: Contents of 77 Breakfast Cereals

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamin	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.402973
100%_Natural_Br	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.983679
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.425505
All-Bran_with_Ex	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.704912
Almond_Delight	R	C	110	2	2	200	1	14	8	-1	25	3	1	0.75	34.384843
Apple_Cinnamon	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.509541
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.174094
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.038562
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.120253
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.313813
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.042851
Cheerios	G	C	110	6	2	290	2	17	1	105	25	1	1	1.25	50.764999
Cinnamon_Toast_	G	C	120	1	3	210	0	13	9	45	25	2	1	0.75	19.823573
Clusters	G	C	110	3	2	140	2	13	7	105	25	3	1	0.5	40.400208
Cocoa_Puffs	G	C	110	1	1	180	0	12	13	55	25	2	1	1	22.736446
Corn_Chex	R	C	110	2	0	280	0	22	3	25	25	1	1	1	41.445019
Corn_Flakes	K	C	100	2	0	290	1	21	2	35	25	1	1	1	45.863324
Corn_Pops	K	C	110	1	0	90	1	13	12	20	25	2	1	1	35.782791
Count_Chocula	G	C	110	1	1	180	0	12	13	65	25	2	1	1	22.396513
Cracklin'_Oat_Bra	K	C	110	3	3	140	4	10	7	160	25	3	1	0.5	40.448772
Cream_of_Wheat_	N	H	100	3	0	80	1	21	0	-1	0	2	1	1	64.533816
Crispix	K	C	110	2	0	220	1	21	3	30	25	3	1	1	46.895644
Crispy_Wheat_&	G	C	100	2	1	140	2	11	10	120	25	3	1	0.75	36.176196

Histograms

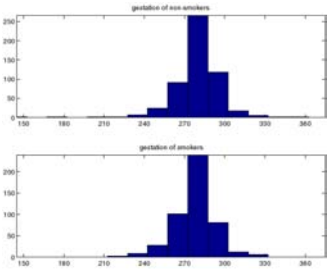


Histograms for Comparisons: Babies of Smokers and Nonsmokers: Histograms of Weights

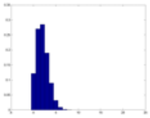


Could this difference be explained by a confounder?

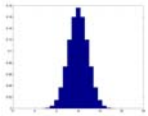
Comparison of Gestation Ages



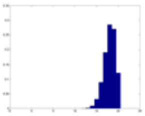
Shapes of Histograms: Symmetry and Skewness



skewed right



symmetric

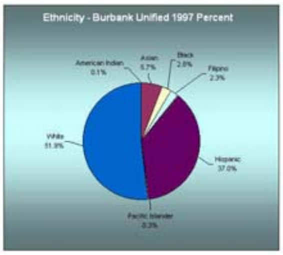


skewed left

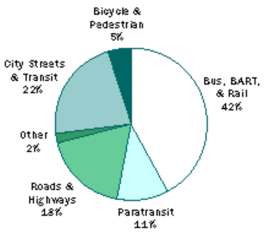
The Bin Height of a Histogram

- Previous examples used counts in each bin, which is common. Comparisons of different histograms can then be difficult.
- Problems arise when bins are different widths.
- Book: area under histogram = 100%
- Another alternative: area = 1

A density histogram is like a pie chart: it represents percentages by areas



Ethnicity breakdown

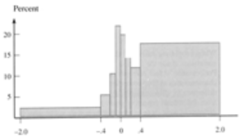


Breakdown of transportation $

Reported minus Actual GPA

Class Interval	Percent
-2.0 -- -.4	2.3
-.4 -- -.2	5.5
-.2 -- -.1	9.7
-.1 -- 0.0	21.0
0.0 -- .1	18.9
.1 -- .2	13.9
.2 -- .4	11.6
.4 – 2.0	17.1

What's misleading with this histogram?

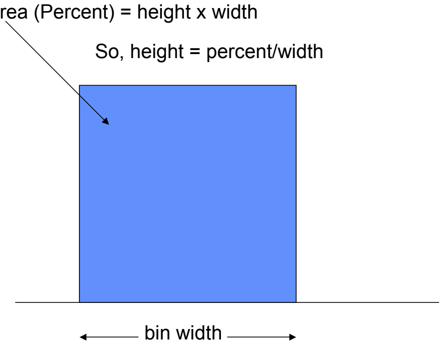


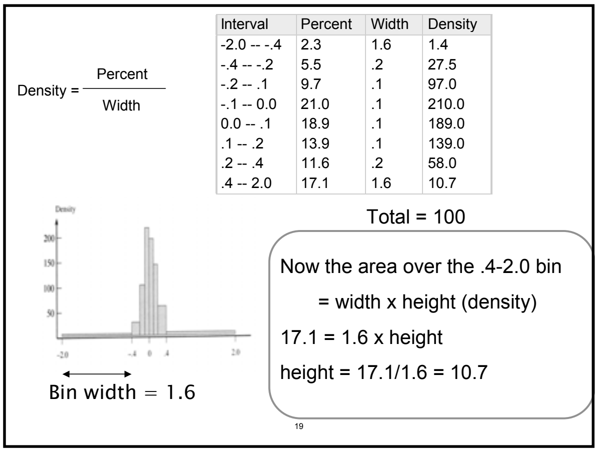
Constructing a 100% Area Histogram

- Calculate percentage in each bin ("class interval")
- The area should equal that percentage, and *area=height x width*
- So, divide each percentage by the bin width, giving the height of the bar ("block") over that bin. This is called the *density scale*.

Area (Percent) = height x width

So, height = percent/width





$$\text{Density} = \frac{\text{Percent}}{\text{Width}}$$

Interval	Percent	Width	Density
-2.0 -- -.4	2.3	1.6	1.4
-.4 -- -.2	5.5	.2	27.5
-.2 -- .1	9.7	.1	97.0
-.1 -- 0.0	21.0	.1	210.0
0.0 -- .1	18.9	.1	189.0
.1 -- .2	13.9	.1	139.0
.2 -- .4	11.6	.2	58.0
.4 -- 2.0	17.1	1.6	10.7

Total = 100

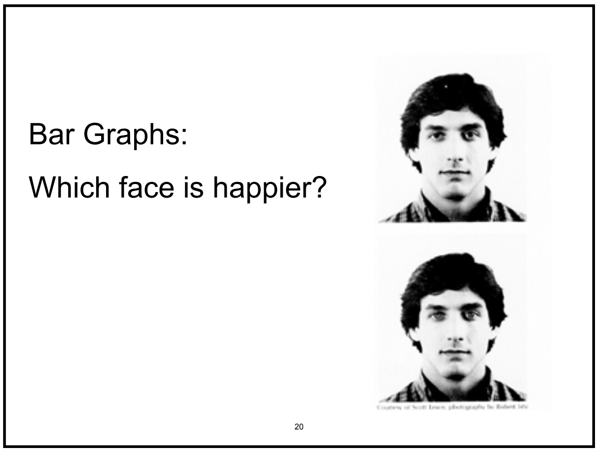
Bin width = 1.6

Now the area over the .4-2.0 bin

= width x height (density)

17.1 = 1.6 x height

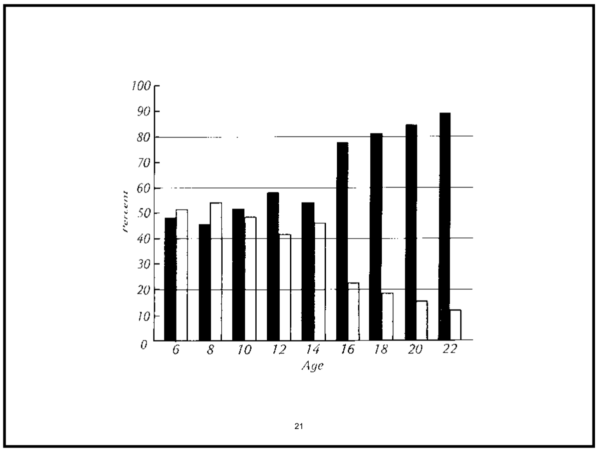
height = 17.1/1.6 = 10.7



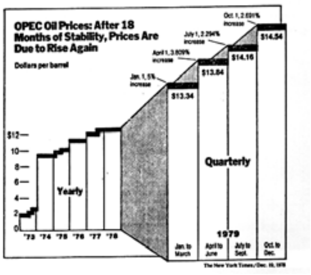
Bar Graphs:

Which face is happier?

Courtesy of Scott Lewis; photographs by Robert Ishi



How to Lie with a Bargraph



New York Times

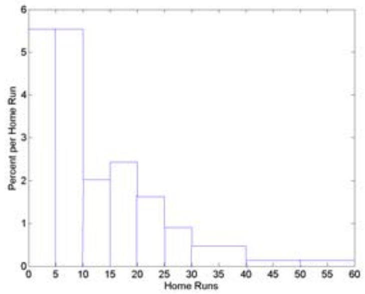
Dec 19, 1978

Review Exercise

Number of home runs in 2002 by American League players with at least 100 plate appearances.

# HR*	count	frequency	block width	block height
0-5	41	27.7 %		
5-10	41	27.7%		
10-15	15	10.1%		
15-20	18	12.2%		
20-25	12	8.1%		
25-30	10	6.8%		
30-40	7	4.7%		
40-50	2	1.4%		
50-60	2	1.4%		

*class interval contains left endpoint but not right endpoint



Is the histogram symmetric or skewed?

Key Concepts

A *histogram* represents percentages by areas.

Density scale: the height of each block equals the percentage in that block divided by the width of the block. The total area = 100%

When the bin widths are equal, it is common for a histogram to just show the counts in each bin.

A histogram shows the shape of the “distribution” of a batch of numbers. The shape may be symmetric, skewed left, or skewed right