CSCI 3022

intro to data science with probability & statistics

Lecture 3
January 22, 2018

1. InterQuartile Range
2. Histograms
3. Boxplots

① HW 1 posted ~ 5pm tonight
Due 1 week from Fri: 2/2

② Waitlist - 16.

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Last time on CSCI 3022:

- **Numerical summaries & summary statistics**: $x_1, x_2, \ldots x_n$ ← n = # data points
  - Mean: $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
  - Median: if n is odd, $\left(\frac{n+1}{2}\right)^{th}$ value if n even, average of $\frac{n}{2}$, $\frac{n+2}{2}$
  - Mode: most common value in dataset.
  - Variance: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (deg. of freedom)
  - Standard Deviation: $\sigma = \sqrt{var}$
  - Quartiles:

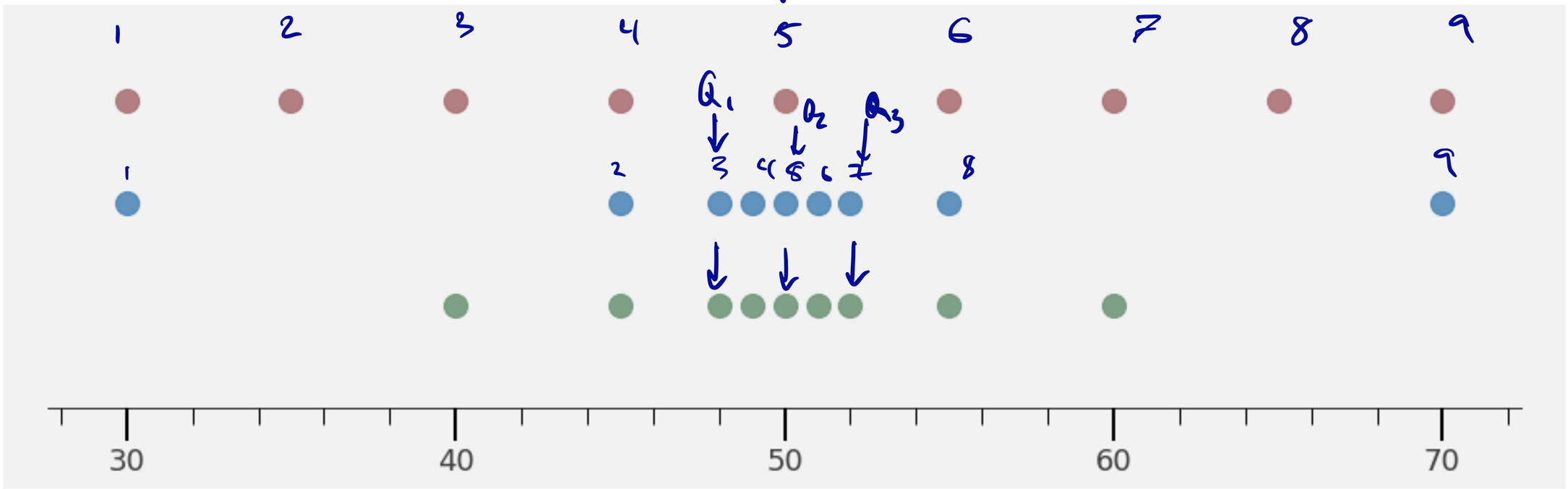
Q_2 = median

Q_1 = median lower half

Q_3 = median upper half.

InterQuartile Range (IQR)

Definition: IQR the difference between Q_3 and Q_1. It's the 'range' of 50% of the data.



Example: Compute the IQR of {6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49}

$\text{IQR} = Q_3 - Q_1 = 42.5 - 25.5$

$Q_1 = \frac{15 + 36}{2} = \frac{51}{2} = 25.5$

$Q_2 = 40$

$Q_3 = 42.5$

Tukey's 5 number summary.

John Tukey advocated that we summarize datasets with 5 values.

1. Minimum
2. Q1
3. Q2 (Median)
4. Q3
5. Max

John Wilder Tukey

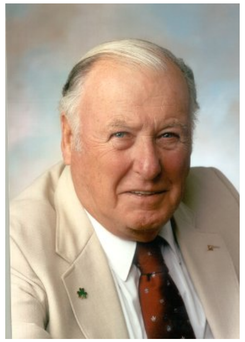
invented FFT, coined the term "bit" = "binary digit"
up there with Grace Hopper and the other demigods

Why we like this:

Gives the center of the data

Gives the spread through the easily computable IQR and range

Gives an idea of skewness



James Willian Cooley

What about Graphical Summaries of data?

Two key types that we'll dig into today:

Histogram: FYI, not the same as #throwbackthursday.
A histogram is a great way to visually understand a *single distribution*.

Boxplot: sometimes called *box-and-whisker-plot*.
A boxplot is a great way to visually compare *multiple distributions*.

Histograms. Why?

Yellowstone National Park

O.F. erupts every 44 to 125 mins

for ~~a~~ ~2 to 5 mins.

216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

272 eruption durations!

Histograms. Why?

And yet! Not particularly useful?
Let's dig in…

216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

272 eruption durations!

Histograms. Why?

And yet! Not particularly useful?
Let's dig in…

min: 96
max: 306
mean: 209.3
Q1: 129.5
Q2: 240
Q3: 267.5

$\frac{272}{2} = 136$

$\frac{136}{2} = 68$

$136 + 68 = 204$

1	96	100	102	104	105	105	105	105	105	105
2	107	107	108	108	108	108	109	109	109	110
3	110	110	110	110	110	110	111	111	112	112
4	112	112	112	112	112	112	113	113	113	113
5	115	115	116	116	117	118	118	118	119	119
6	119	120	120	120	120	121	121	121	122	122
7	124	125	125	126	126	126	128	129	130	130
8	131	132	132	132	133	134	134	135	135	136
9	137	138	139	140	141	142	143	144	144	145
10	145	149	157	158	168	173	174	184	199	200
11	200	202	205	207	210	210	214	214	216	216
12	216	216	221	223	224	225	226	226	229	230
13	230	230	230	230	231	231	233	235	235	235
14	237	237	238	238	240	240	240	240	240	240
15	242	242	243	244	244	245	245	245	245	245
16	246	246	247	247	248	248	249	249	249	249
17	250	250	250	250	251	252	254	254	254	255
18	255	255	255	256	256	257	257	258	258	259
19	260	260	260	260	260	261	261	261	261	262
20	262	262	262	263	264	265	265	265	265	266
21	266	267	267	267	268	268	269	270	270	270
22	270	270	270	270	270	271	272	272	272	272
23	272	273	274	274	274	275	275	275	275	276
24	276	276	276	277	278	278	278	279	280	280
25	282	282	282	282	282	282	283	284	285	286
26	287	288	288	288	288	288	288	289	289	290
27	290	291	293	294	294	296	296	296	300	302
28	304	306								

df.sort_values(…)

Histograms. Why?

A histogram shows us how the data are *distributed*.

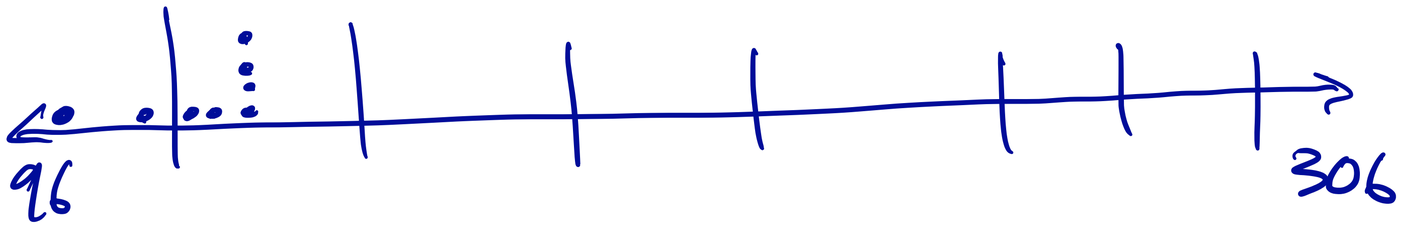
Easy to **compute**.
Easy to **understand**.

1	96	100	102	104	105	105	105	105	105	105
2	107	107	108	108	108	108	109	109	109	110
3	110	110	110	110	110	110	111	111	112	112
4	112	112	112	112	112	112	113	113	113	113
5	115	115	116	116	117	118	118	118	119	119
6	119	120	120	120	120	121	121	121	122	122
7	124	125	125	126	126	126	128	129	130	130
8	131	132	132	132	133	134	134	135	135	136
9	137	138	139	140	141	142	143	144	144	145
10	145	149	157	158	168	173	174	184	199	200
11	200	202	205	207	210	210	214	214	216	216
12	216	216	221	223	224	225	226	226	229	230
13	230	230	230	230	231	231	233	235	235	235
14	237	237	238	238	240	240	240	240	240	240
15	242	242	243	244	244	245	245	245	245	245
16	246	246	247	247	248	248	249	249	249	249
17	250	250	250	250	251	252	254	254	254	255
18	255	255	255	256	256	257	257	258	258	259
19	260	260	260	260	260	261	261	261	261	262
20	262	262	262	263	264	265	265	265	265	266
21	266	267	267	267	268	268	269	270	270	270
22	270	270	270	270	270	271	272	272	272	272
23	272	273	274	274	274	275	275	275	275	276
24	276	276	276	277	278	278	278	279	280	280
25	282	282	282	282	282	282	283	284	285	286
26	287	288	288	288	288	288	288	289	289	290
27	290	291	293	294	294	296	296	296	300	302
28	304	306								

df.sort_values(…)

Imagine all these data points on an axis…

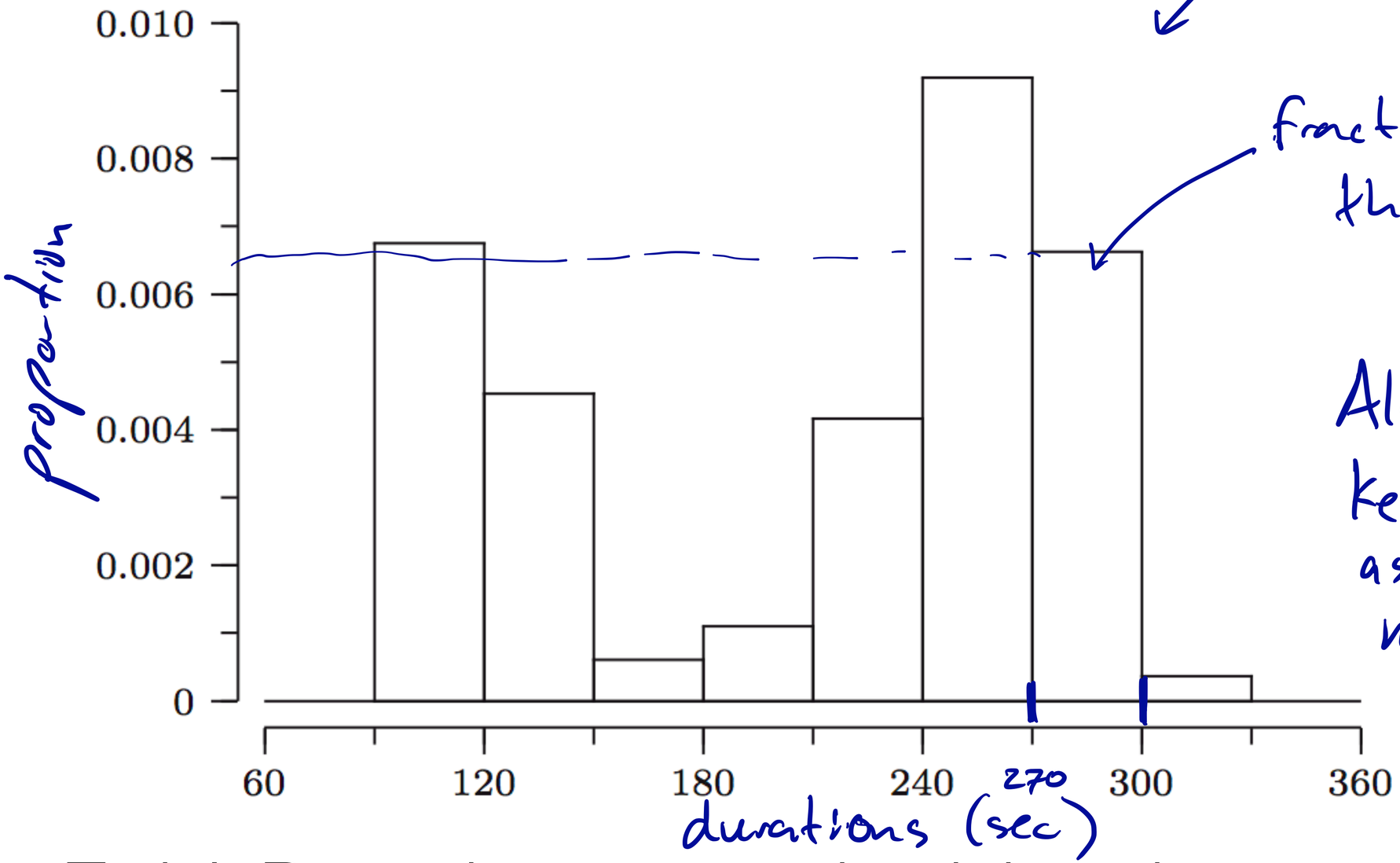
1	96	100	102	104	105	105	105	105	105	105
2	107	107	108	108	108	108	109	109	109	110
3	110	110	110	110	110	110	111	111	112	112
4	112	112	112	112	112	112	113	113	113	113
5	115	115	116	116	117	118	118	118	119	119
6	119	120	120	120	120	121	121	121	122	122
7	124	125	125	126	126	126	128	129	130	130
8	131	132	132	132	133	134	134	135	135	136
9	137	138	139	140	141	142	143	144	144	145
10	145	149	157	158	168	173	174	184	199	200
11	200	202	205	207	210	210	214	214	216	216
12	216	216	221	223	224	225	226	226	229	230
13	230	230	230	230	231	231	233	235	235	235
14	237	237	238	238	240	240	240	240	240	240
15	242	242	243	244	244	245	245	245	245	245
16	246	246	247	247	248	248	249	249	249	249
17	250	250	250	250	251	252	254	254	254	255
18	255	255	255	256	256	257	257	258	258	259
19	260	260	260	260	260	261	261	261	261	262
20	262	262	262	263	264	265	265	265	265	266
21	266	267	267	267	268	268	269	270	270	270
22	270	270	270	270	270	271	272	272	272	272
23	272	273	274	274	274	275	275	275	275	276
24	276	276	276	277	278	278	278	279	280	280
25	282	282	282	282	282	282	283	284	285	286
26	287	288	288	288	288	288	288	289	289	290
27	290	291	293	294	294	296	296	296	300	302
28	304	306								



Then, divide the line into “bins” and count.

df.sort_values(…)

Old Faithful Histogram



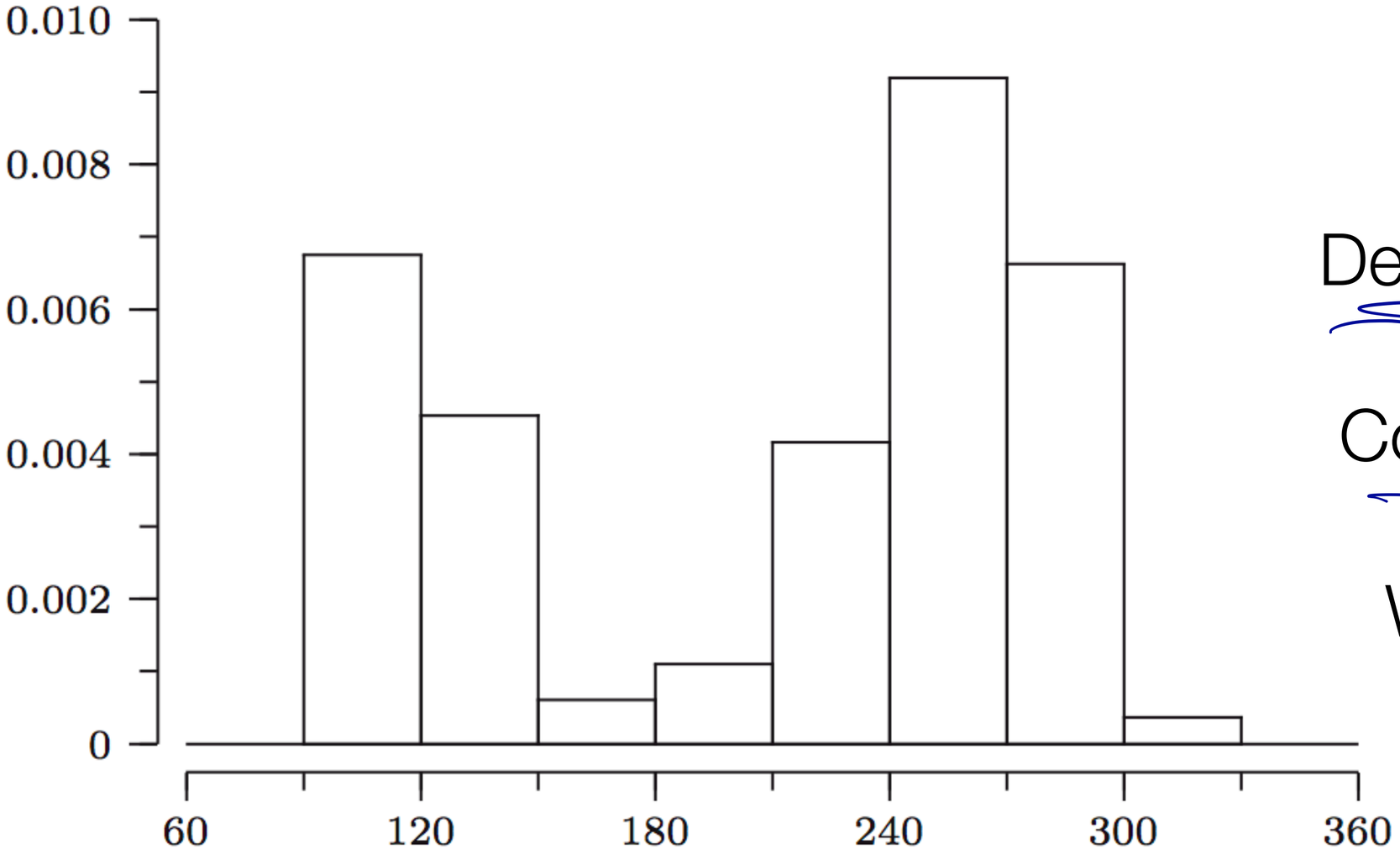
normalized ∑ bars = 1

fraction of eruptions in this bin

Also totally fine to keep vertical axis as counts, instead of normalizing.

Tada! But wait… your textbook has done something peculiar.
Can you spot it?

Old Faithful Histogram



Density histogram

vs

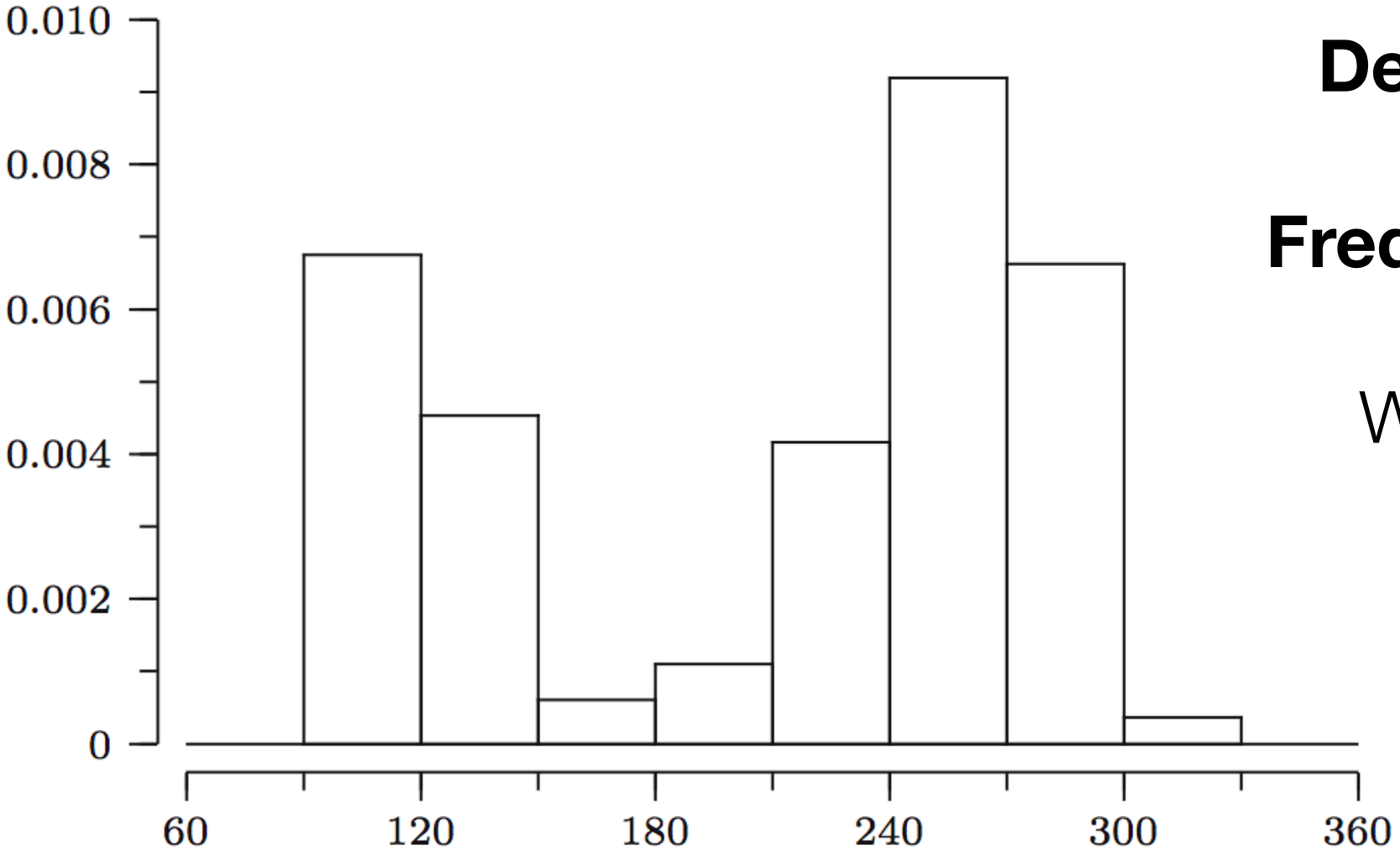
Count histogram

When/which?

Tada! But wait… your textbook has done something peculiar.

Can you spot it?

Old Faithful Histogram



Density histogram

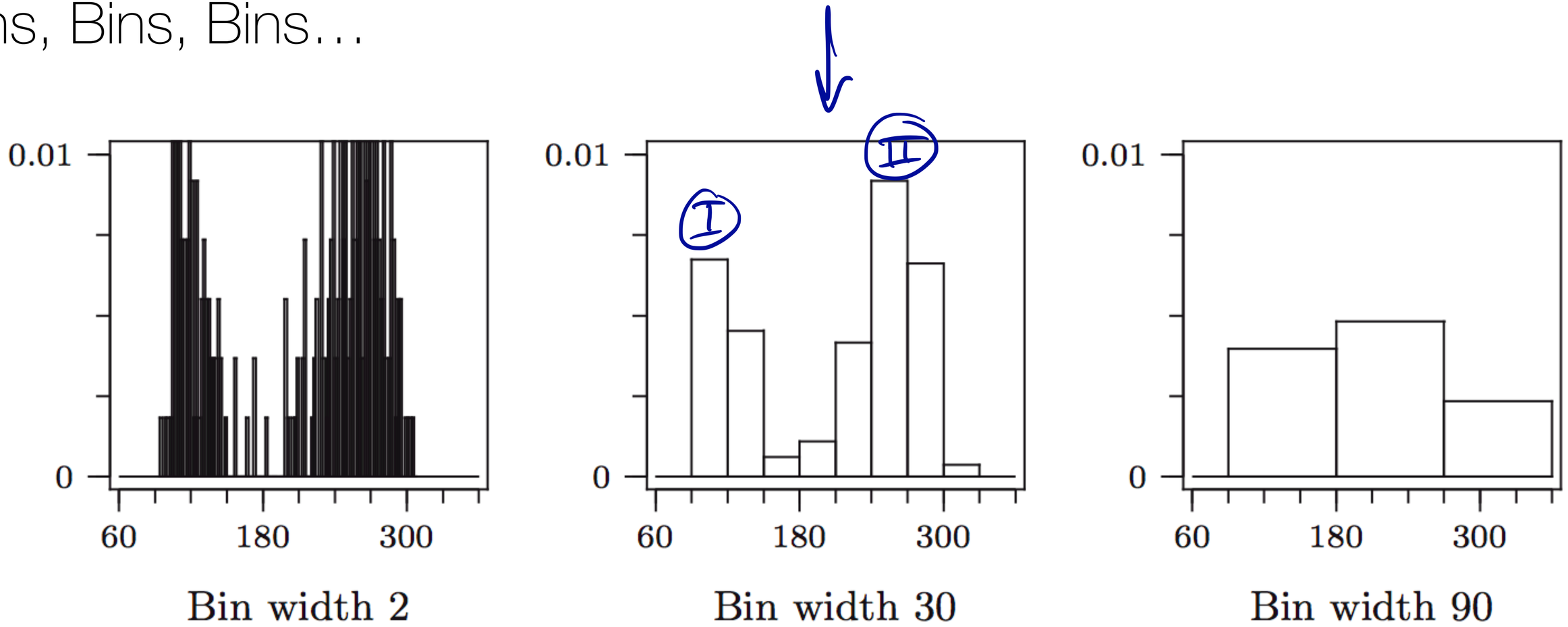
vs

Frequency histogram

Which is better?

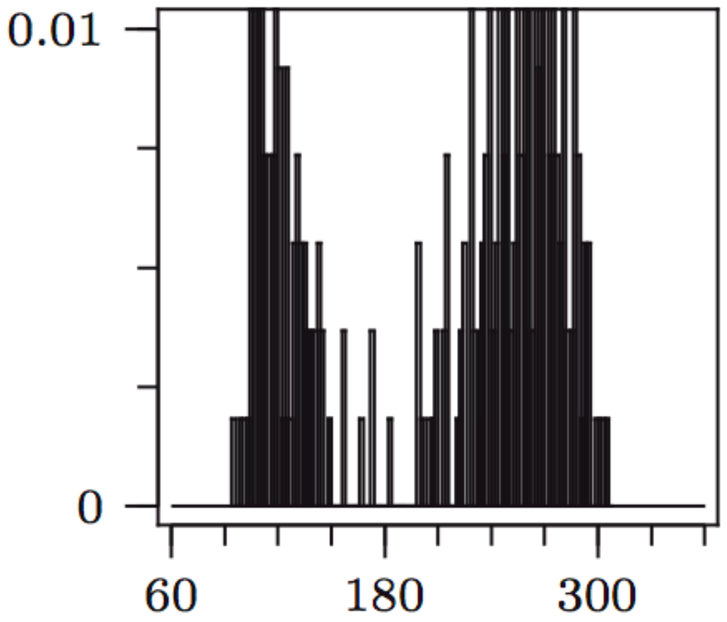
Tada! But wait… your textbook has done something peculiar.
Can you spot it?

Bins, Bins, Bins…

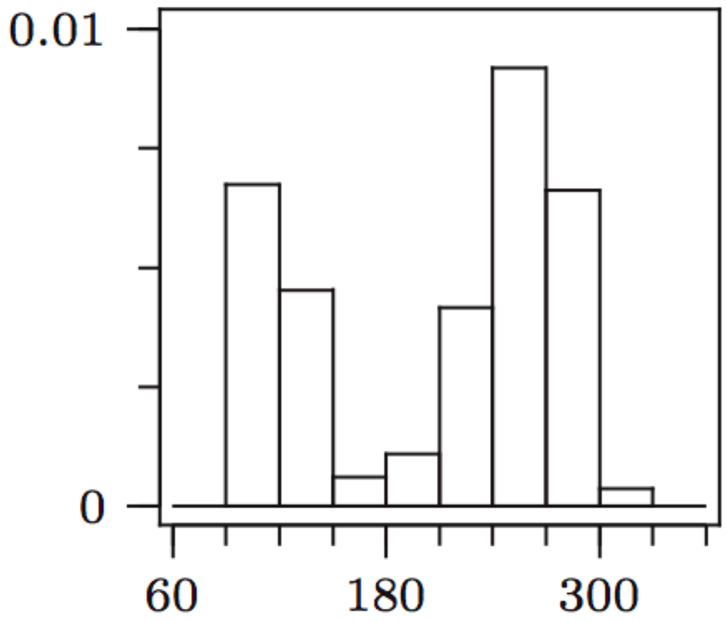


These are all histograms. They're all correct.
However, the one in the middle is more useful. Why?

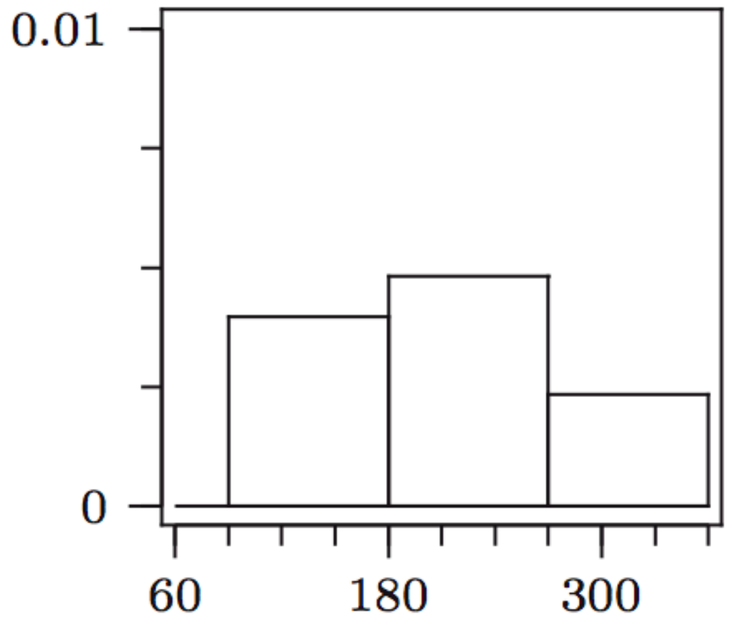
Bins, Bins, Bins…



Bin width 2



Bin width 30

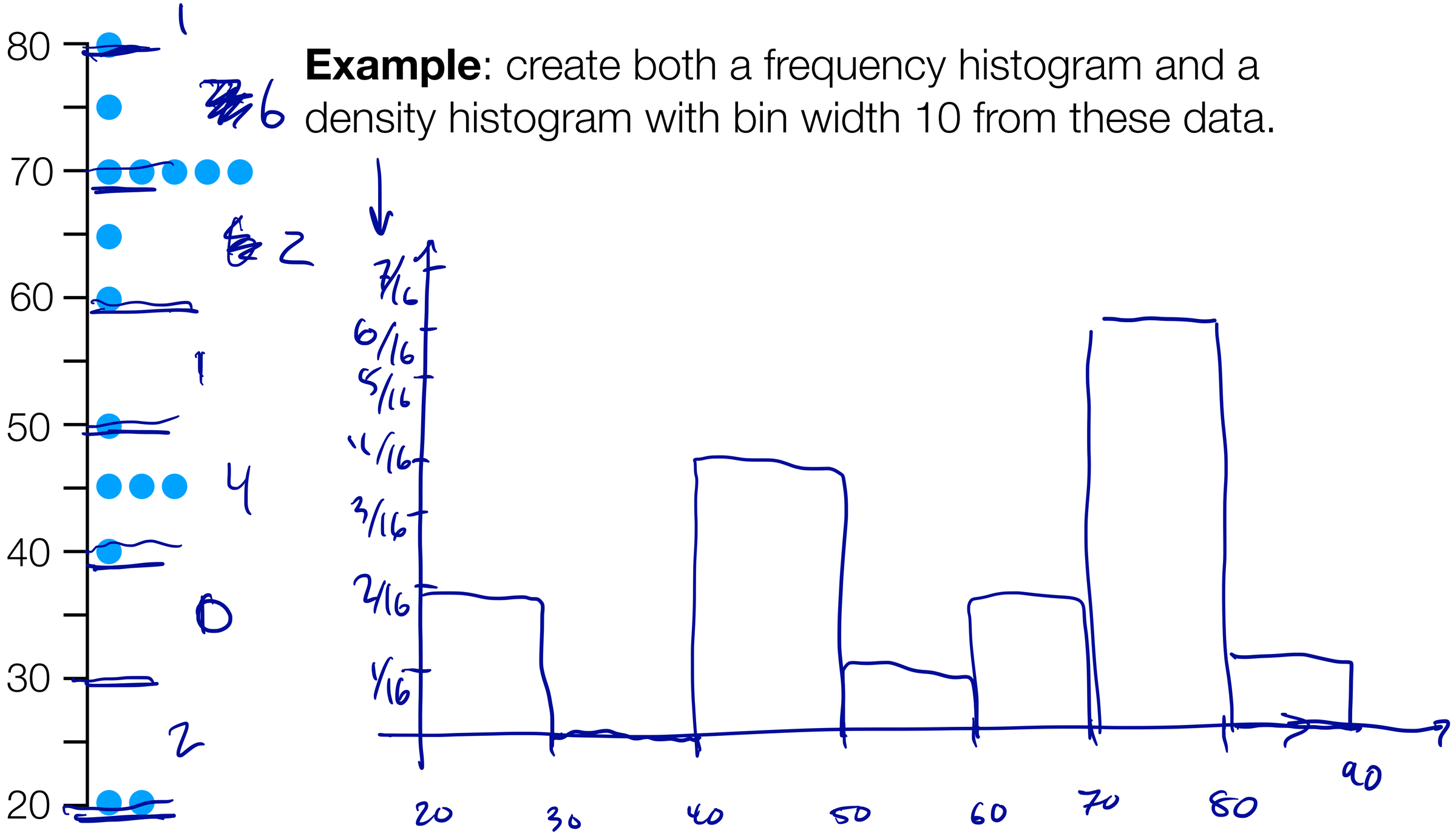


Bin width 90

These are all histograms. They're all correct.
However, the one in the middle is more useful. Why?

Goldilocks bin sizes, Freedman & Diaconis $= 2\,\frac{IQR}{n^{1/3}} = 2\,\frac{Q_3 - Q_1}{n^{1/3}}$

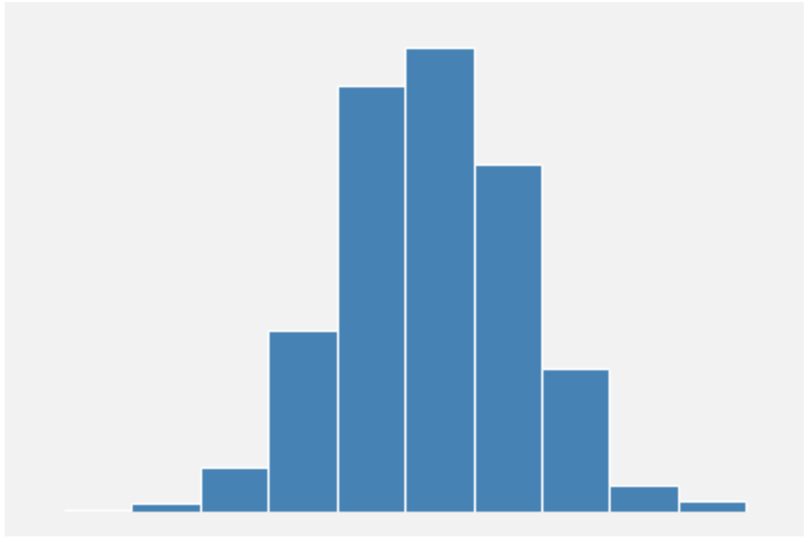
Example: create both a frequency histogram and a density histogram with bin width 10 from these data.



Modes…

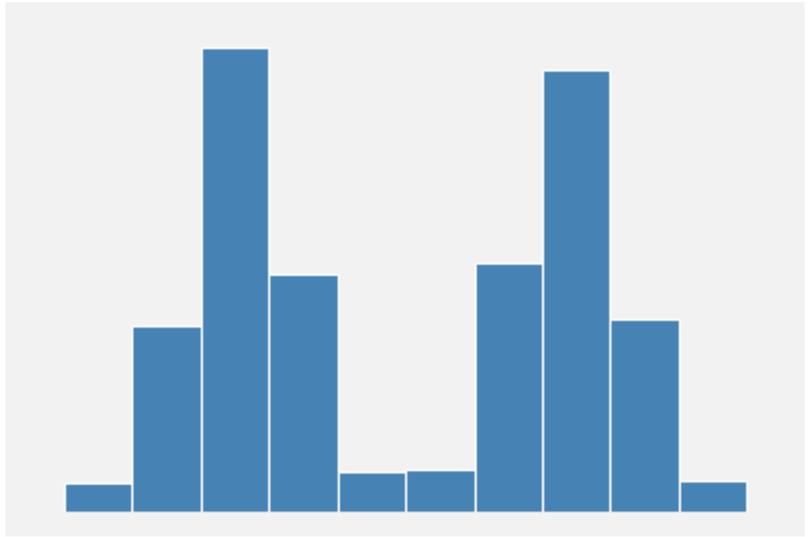
We talked before about “the” mode, i.e. the most frequent value.

Often, it’s useful to describe whether a distribution has multiple *separated* high frequency values. Histograms make this easy:



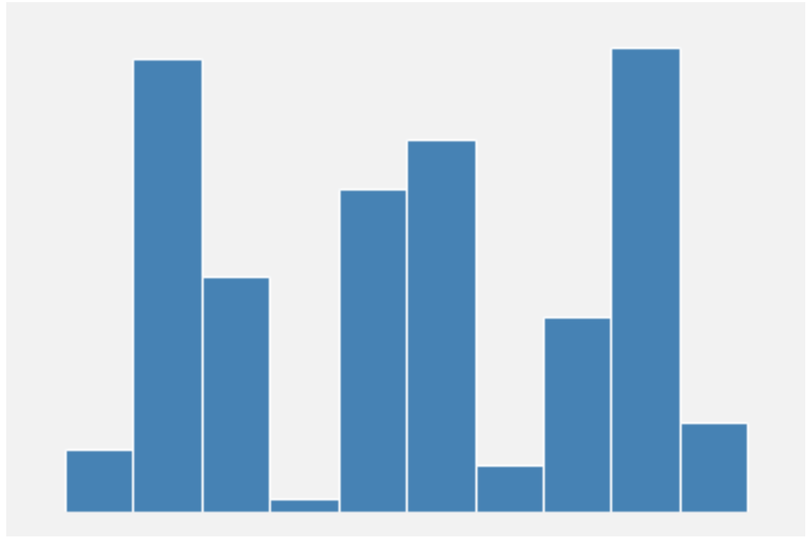
unimodal

1



bimodal

2

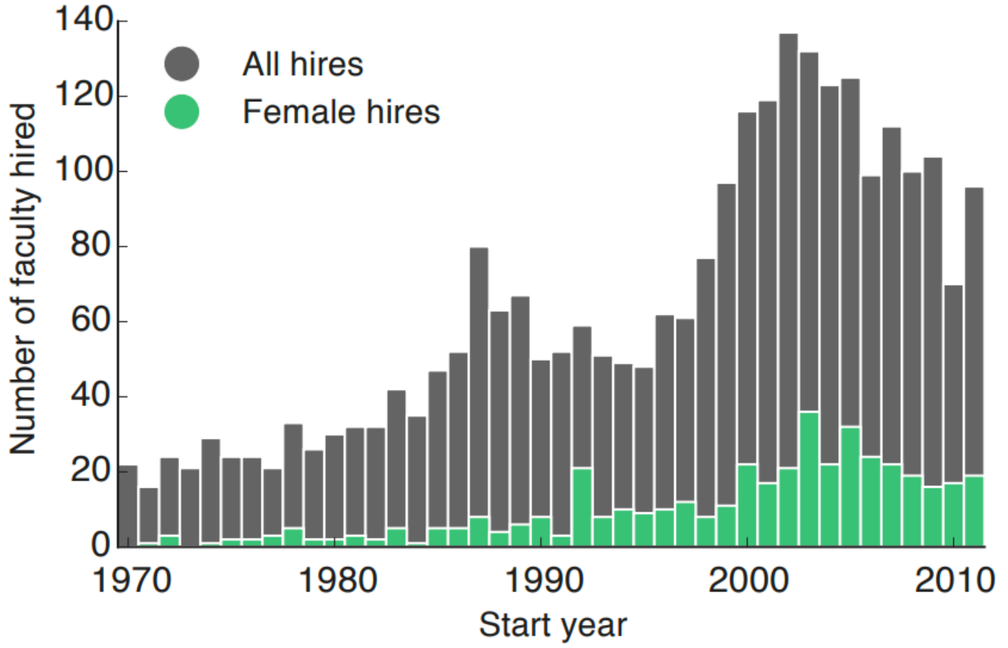


multimodal

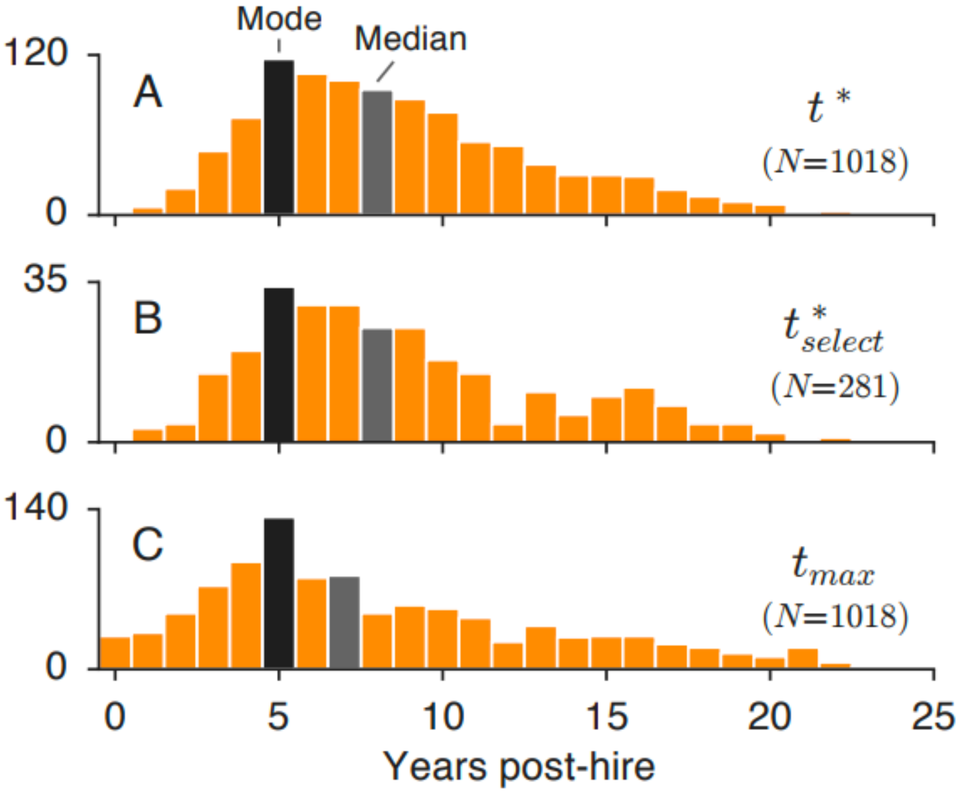
> 2

Review: what about skew?

Histograms in the wild!

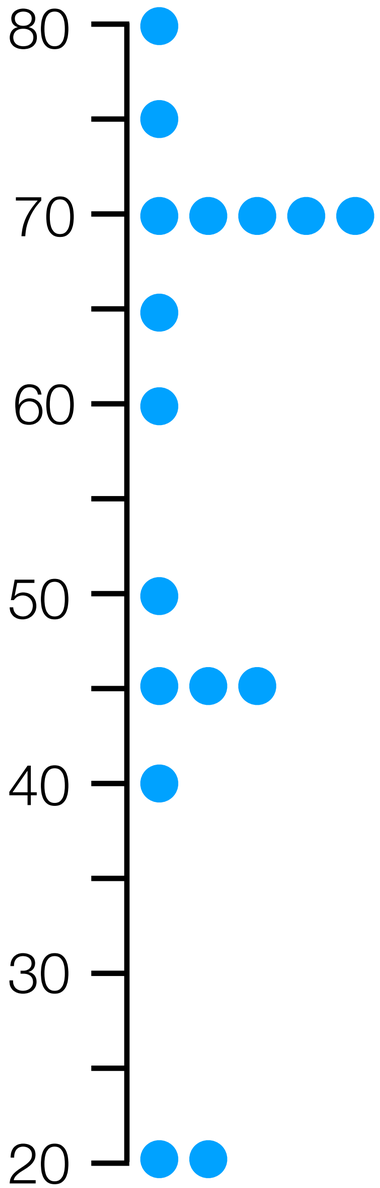


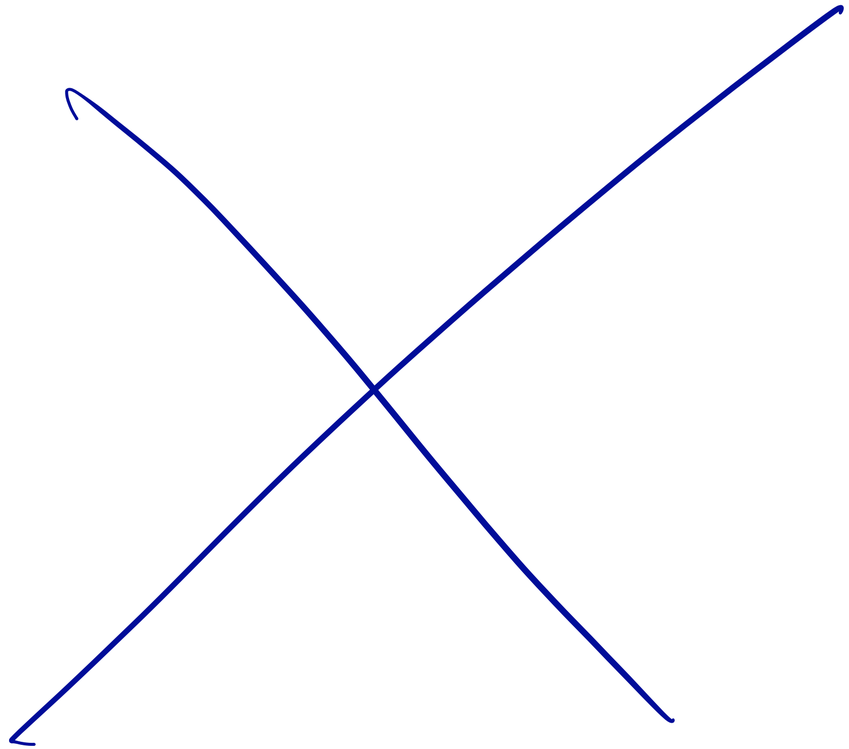
"Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks" Samuel F. Way, Daniel B. Larremore, and Aaron Clauset. Proc. 2016 World Wide Web Conference (WWW), 1169-1179 (2016).



"The misleading narrative of the canonical faculty productivity trajectory" Samuel F. Way, Allison C. Morgan, Aaron Clauset, and Daniel B. Larremore. (2016)

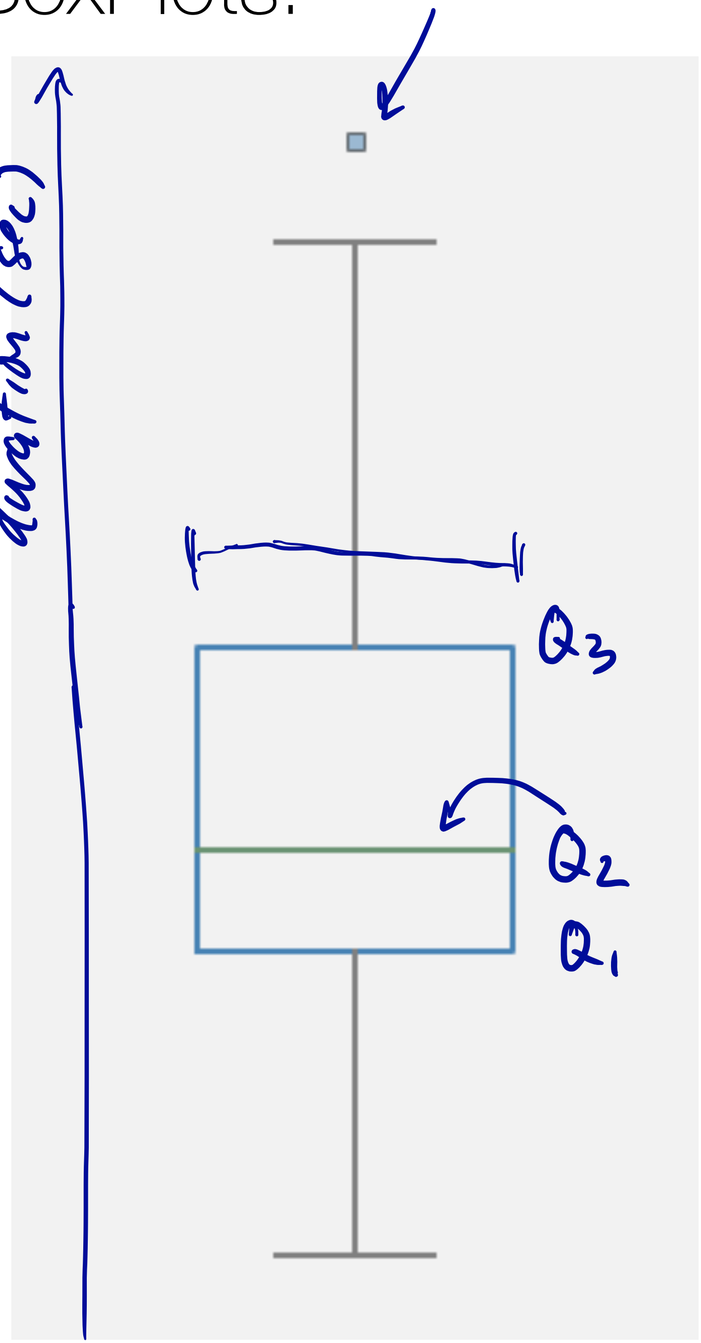
BoxPlots! But first… quartile review.





Example: what are the quartiles of this distribution?

BoxPlots!



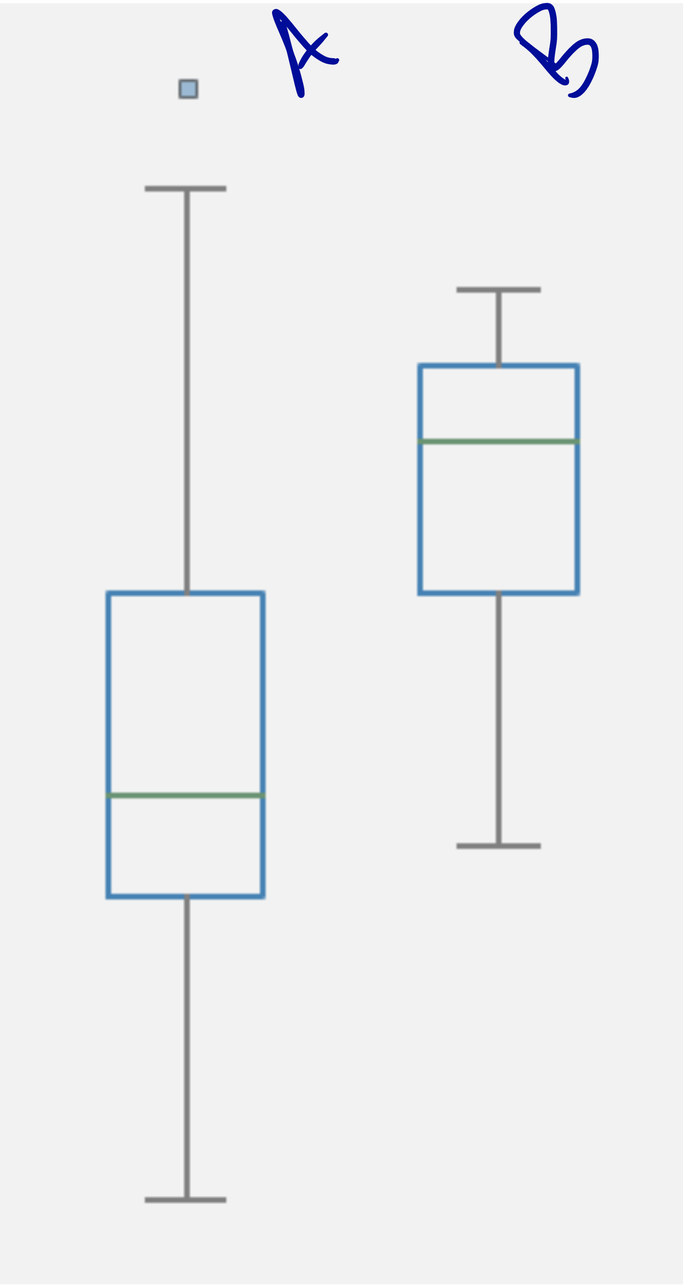
- The **Box** extends from Q_1 to Q_3.
- The **Median Line** goes through median
- The **Whiskers** extend to farthest point within 1.5 x IQR
- The Fliers or **outliers** are any points outside of whiskers
- The **width** of the box is unimportant.

Boxplots are a visualization of Tukey's 5# summary.

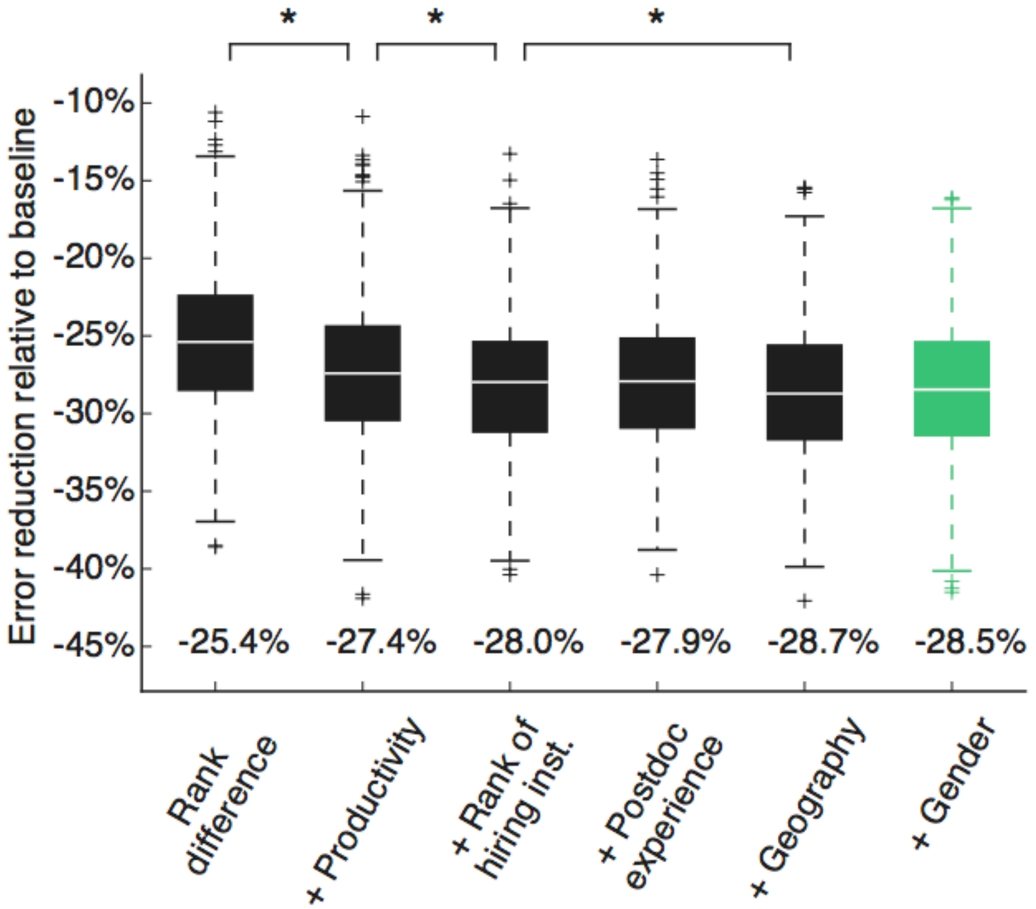
$1.5 \cdot IQR = (Q_3 - Q_1) \cdot 1.5$

$Q_2 \approx$ median

BoxPlots!



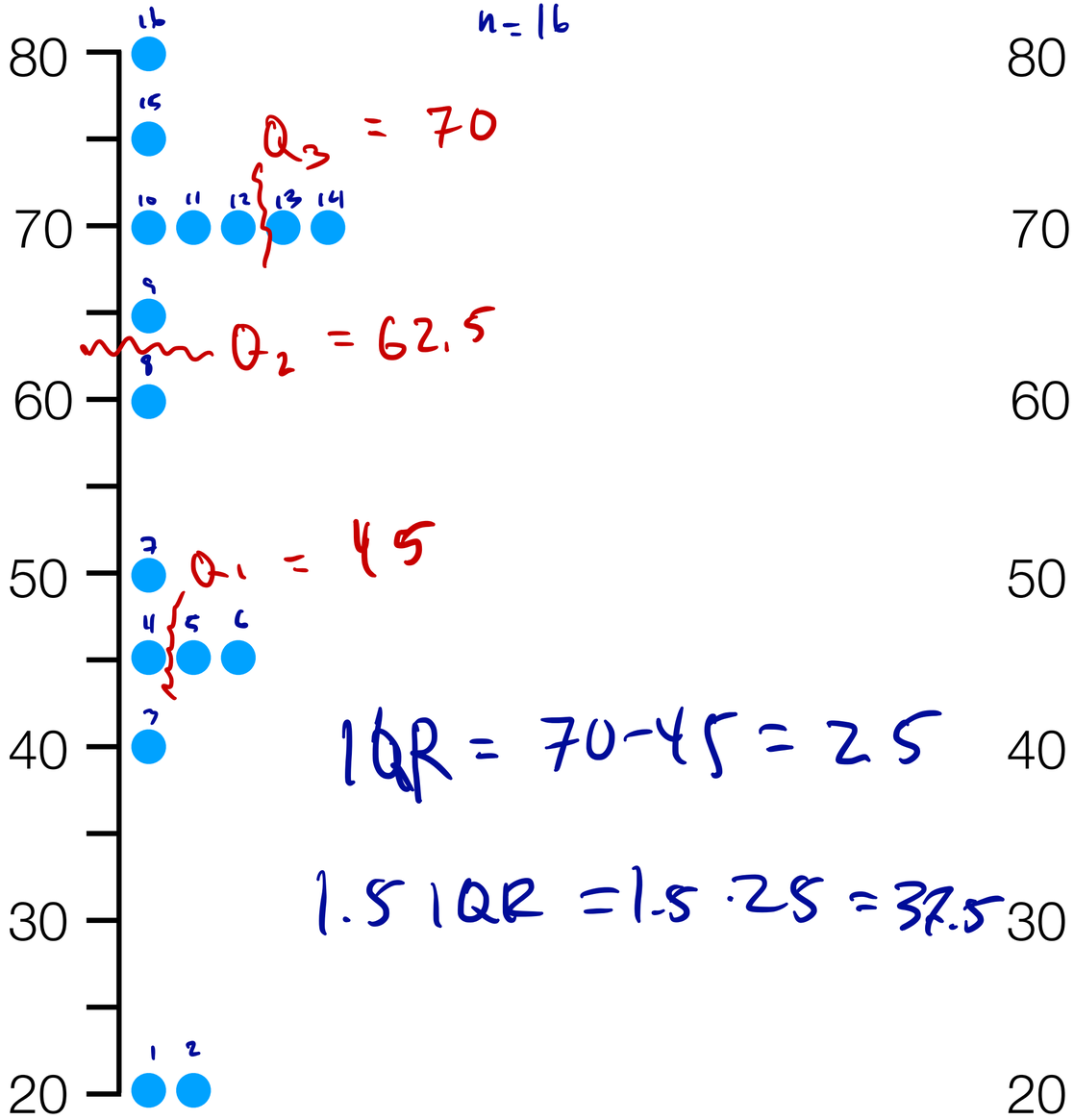
Boxplots are great for comparing multiple datasets or pieces of a dataset.

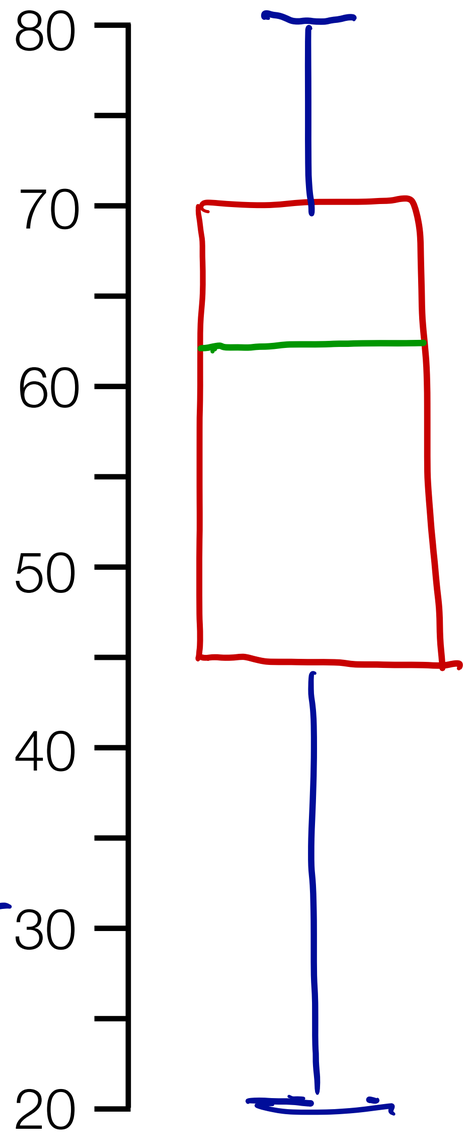


"Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks" Samuel F. Way, Daniel B. Larremore, and Aaron Clauset. Proc. 2016 World Wide Web Conference (WWW), 1169-1179 (2016).

BoxPlots practice

Example: draw a box & whisker plot from this dataset





70 + 37.5 = 107.5

45 − 37.5 = 7.5

Time to get cracking!

Now

1. Team up / laptops out.
2. Pull from the course github.
3. nb1 & nb2

Before next class

1. Complete nb1 notebook.
2. Start on HW1—even if you just glance at it. Come to office hours with questions!