DATA VISUALIZATION

UNIT – IV (PART – II)

Variation and Uncertainty

- It’s good to feel confident when armed with a bit of data, but it's crucial to stay humble.
- The world is complex and ever-changing.
- If our data is unreliable or our conclusions are questionable, we should be cautious.
- When sharing data, honesty is key.
- Clearly communicate what we know, what we don't, and represent reality as accurately as possible.
- If data has high variation or is from a limited sample, be transparent to avoid misleading our audience.

Variation and Uncertainty

- Variation refers to how much individual observations differ within a group.
- Example: Students in a class may have different heights because of factors like genetics and nutrition.

- Uncertainty is the lack of confidence in making inferences about a population from data collected in samples.
- Example: It's hard to be completely sure about the average income of a city when relying on a small survey.

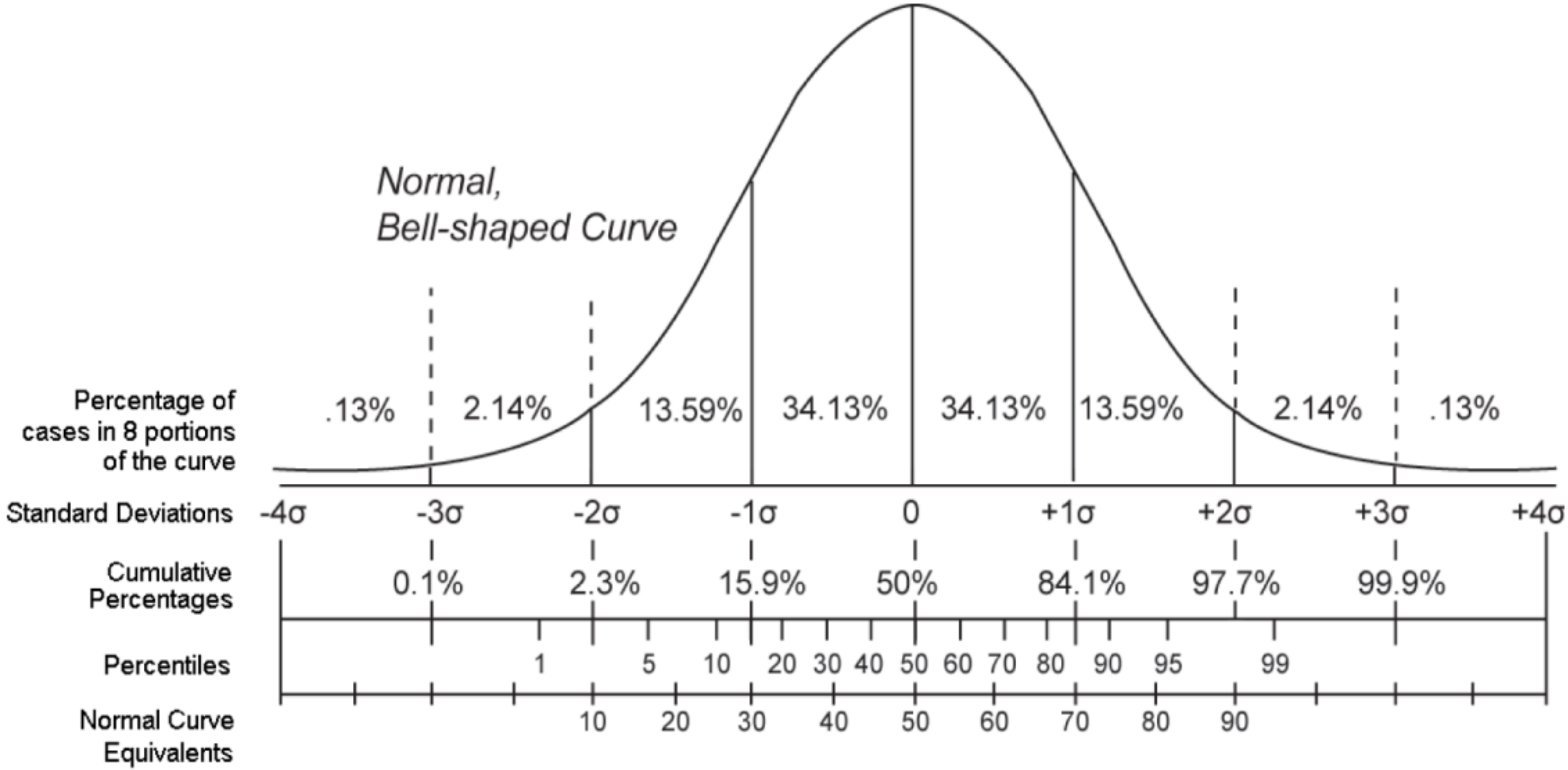
Variation and Uncertainty

1. Respecting variation
2. Variation over time-Control charts
3. Understanding uncertainty

Respecting variation

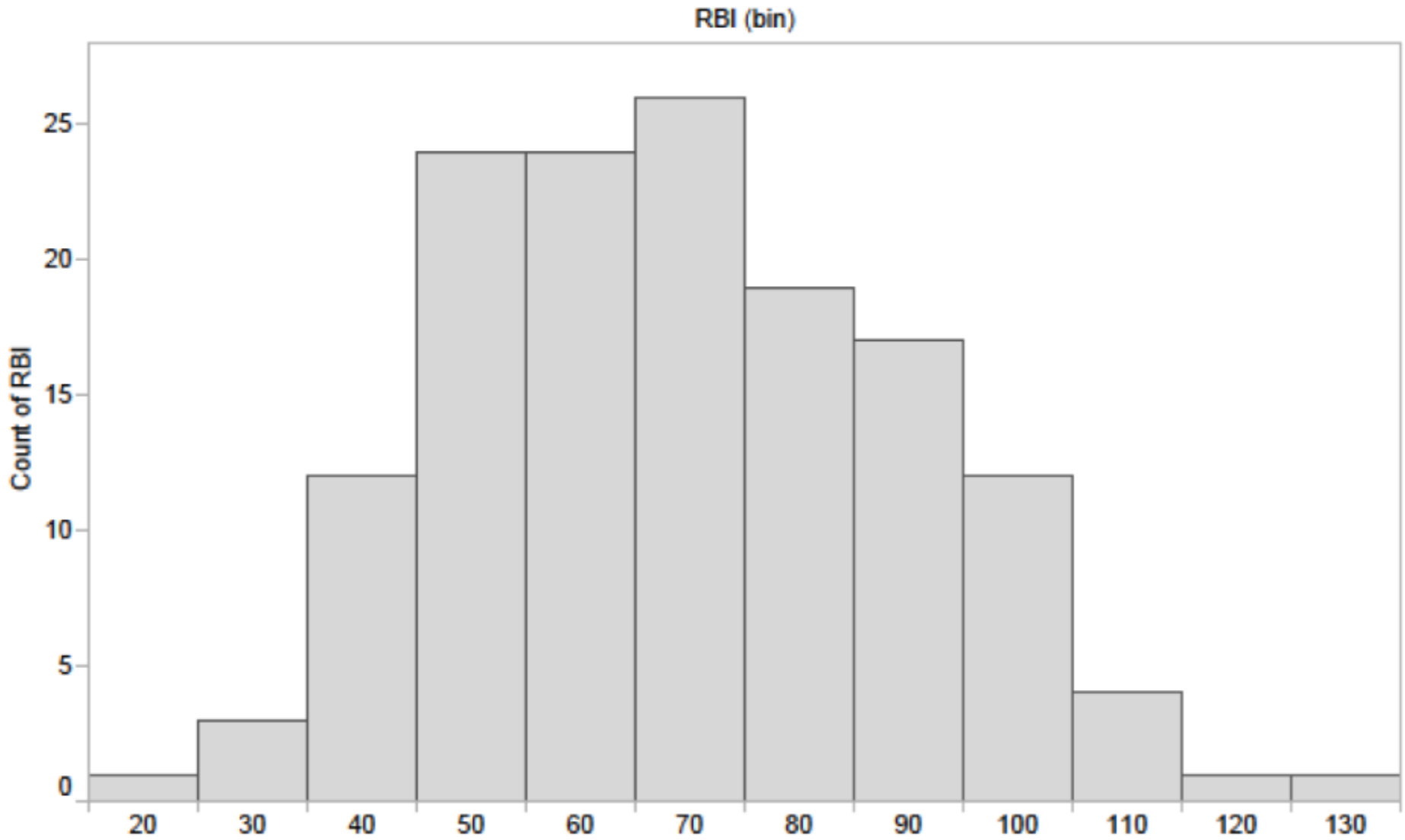
- In the previous chapter, we explored central tendency measures, such as the mean and median.

- This discussion also included fundamental measures of variation, like standard deviation and the interquartile range, as illustrated in Figure 7-1.

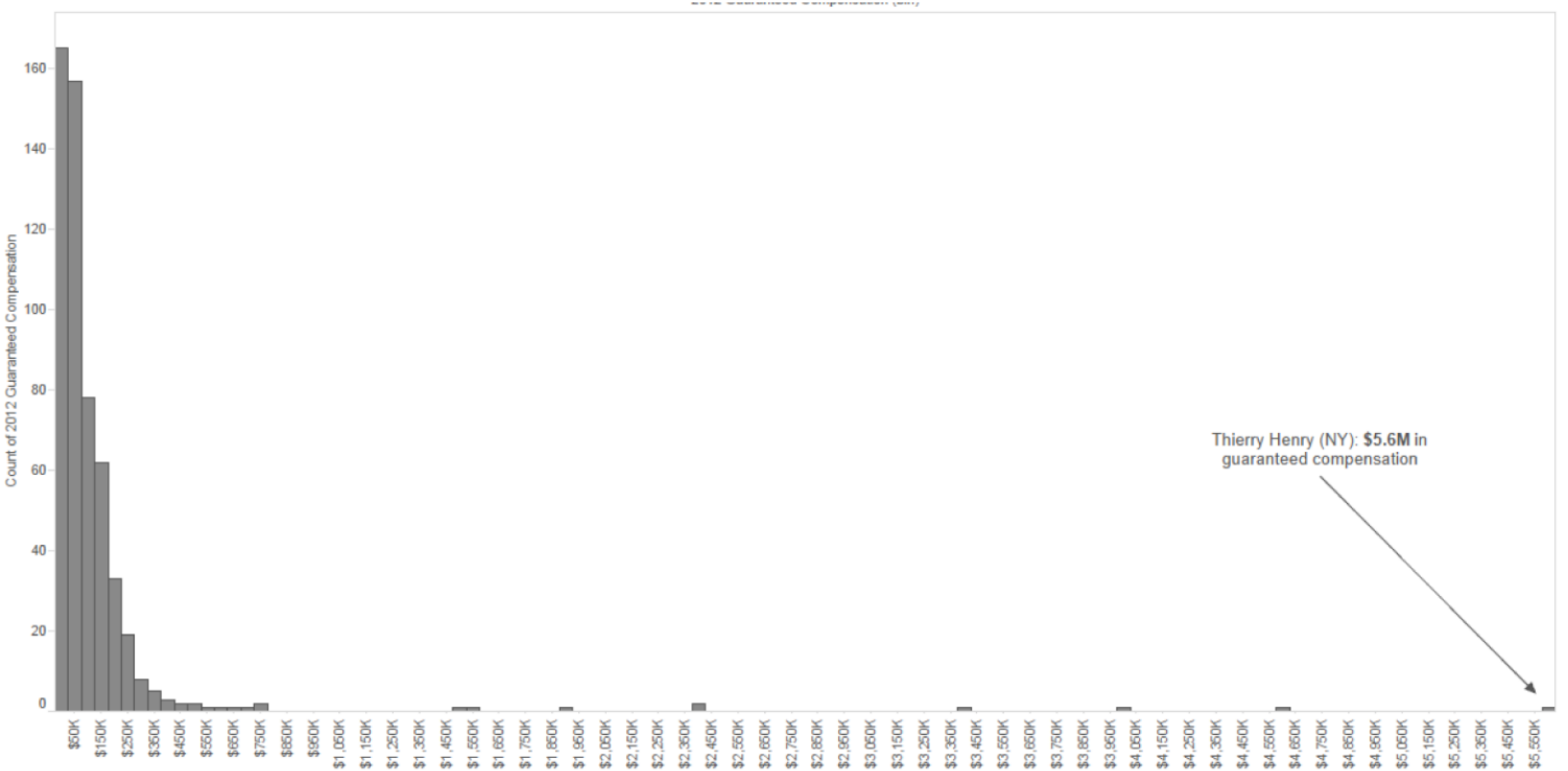

Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Normal Curve Equivalents
10
20
30
40
50
60
70
80
90

Respecting variation

- In the previous chapter, we examined two distinct types of variables in the realm of sports: baseball batting statistics (RBI) and soccer players' salaries, depicted in Figure 7-2.



2012 Guaranteed Compensation (bin)


Count of 2012 Guaranteed Compensation
160
140
120
100
80
60
40
20
0
Thierry Henry (NY): $5.6M in guaranteed compensation
$50K
$150K
$250K
$350K
$450K
$550K
$650K
$750K
$850K
$950K
$1,050K
$1,150K
$1,250K
$1,350K
$1,450K
$1,550K
$1,650K
$1,750K
$1,850K
$1,950K
$2,050K
$2,150K
$2,250K
$2,350K
$2,450K
$2,550K
$2,650K
$2,750K
$2,850K
$2,950K
$3,050K
$3,150K
$3,250K
$3,350K
$3,450K
$3,550K
$3,650K
$3,750K
$3,850K
$3,950K
$4,050K
$4,150K
$4,250K
$4,350K
$4,450K
$4,550K
$4,650K
$4,750K
$4,850K
$4,950K
$5,050K
$5,150K
$5,250K
$5,350K
$5,450K
$5,550K

Respecting variation

Visualizing Variation:

- To honor the inherent variation in our data, it's essential to display it.
- Presenting only averages creates a too simple view of the world.
- Just as not every person in a country shares the most common physical traits, not every data point aligns with the mean, median, or mode.
- If we consider once again the number of strikeouts per nine innings in professional baseball over the past 100 years, we can show a simple line plot of average strikeouts per nine innings, as shown in Figure 7-3

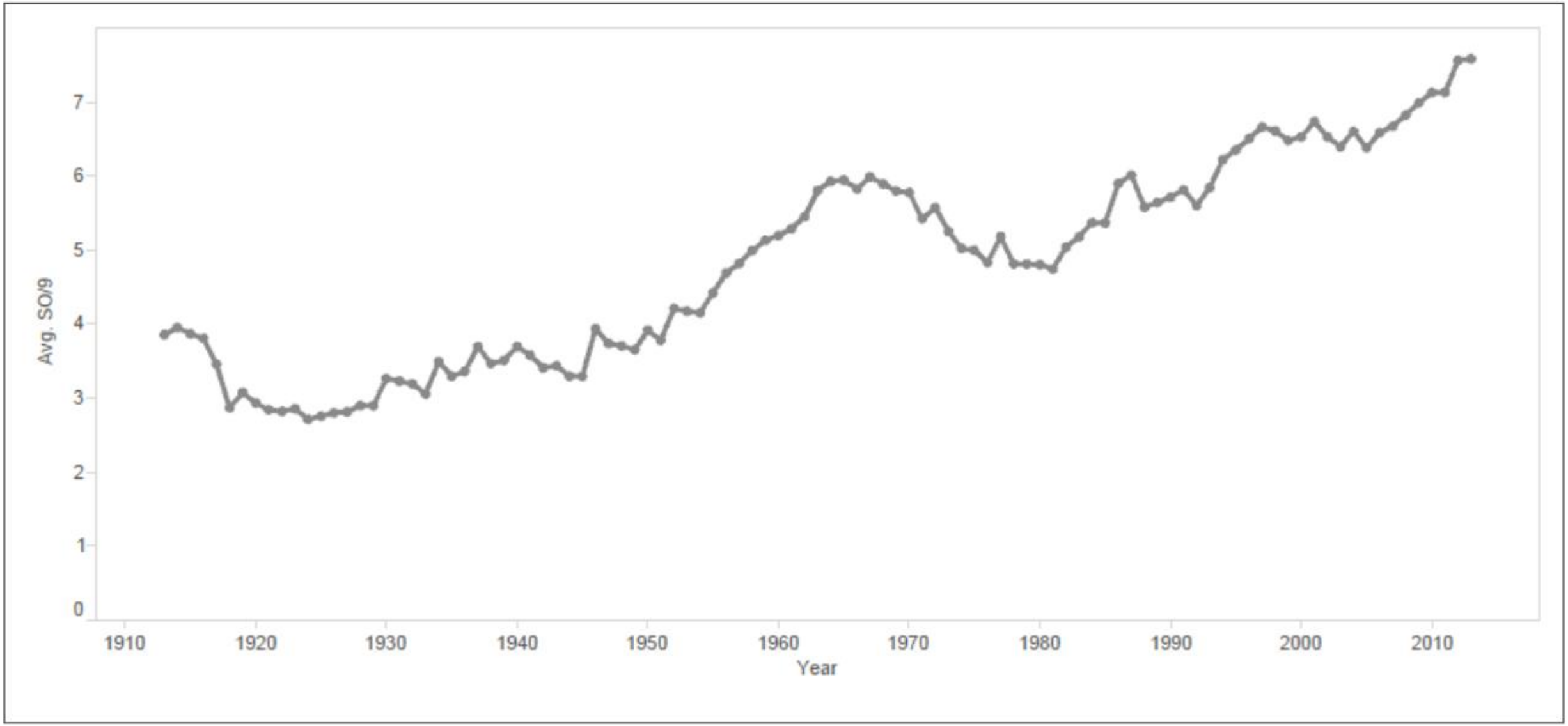


Figure 7-3. Average number of strikeouts per nine innings

Respecting variation

Visualizing Variation:

- However, this chart doesn't reveal how the strikeout rates varied among different teams in the league each year.
- We're left in the dark about the contrast between the team with the highest strikeout rate and the one with the lowest rate annually.

- To capture the inherent variation in the data, we can represent it in various ways, as depicted in Figure 7-4

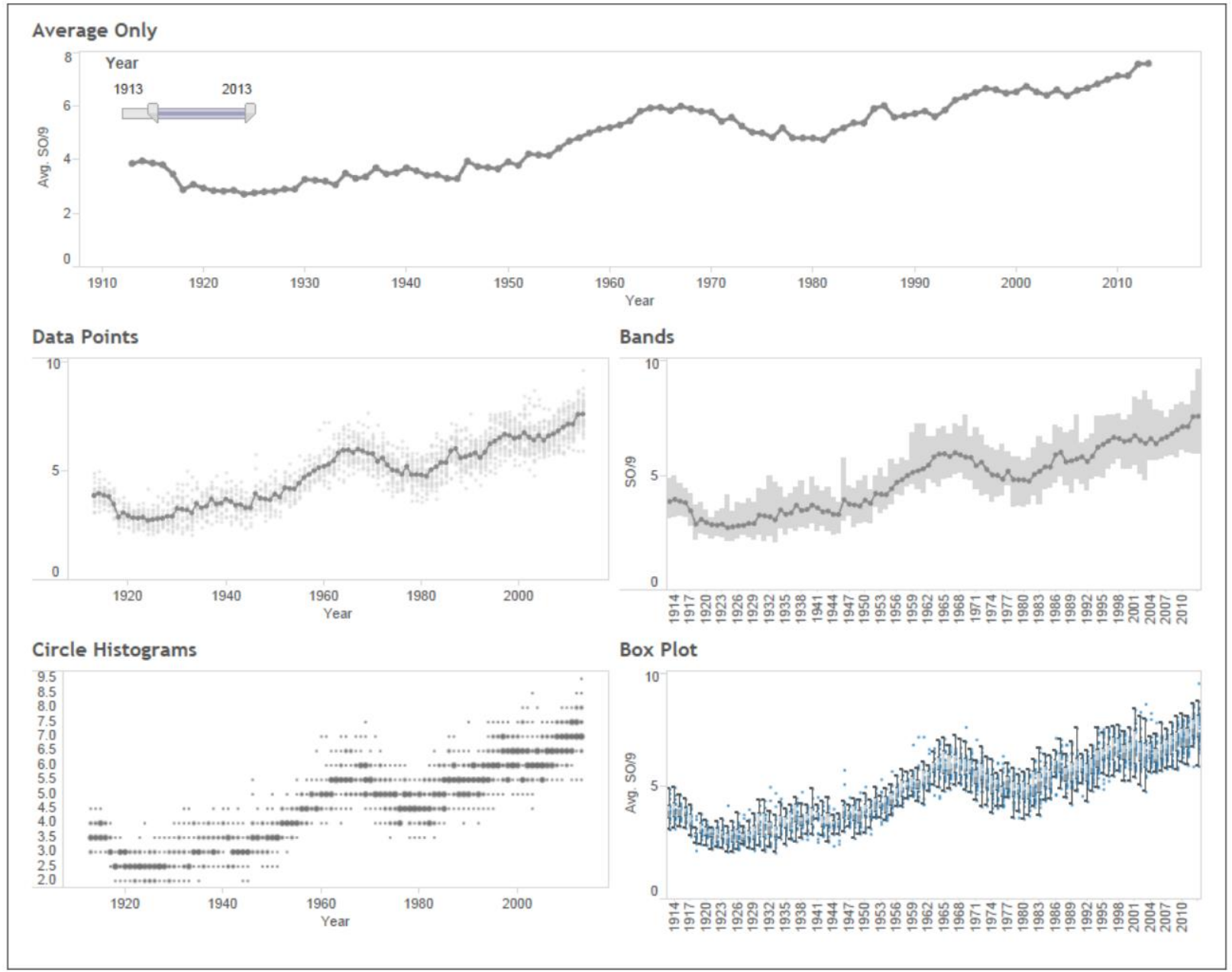


Figure 7-4. Four different ways of showing variation in a time series

Respecting variation

Data points:

- Each team is represented by its own circle in each year.

Bands:

- Includes reference bands from the minimum to the maximum for each year.

Circle Histograms:

- Consists of circle histograms, where the area of each circle is proportional to the number of teams in each bin.

Box-Plots

- Displays a series of box plots for each year.

Variation and Uncertainty

1. Respecting variation
2. Variation over time-Control charts
3. Understanding uncertainty

Variation over time-Control charts

- Control charts help determine if data collected over time contains statistically significant signals or if the variation is just noise.
- Walter Shewhart developed them in the 1920s at the Western Electric Company for industrial quality control.

- The Six Sigma movement has popularized these charts, with "black belts" using them to measure process behavior and reduce variation to enhance quality.

Variation over time-Control charts

- The idea is that reducing variation leads to fewer defects.
- This concept is particularly applicable in manufacturing and any scenario where a consistent output is essential.
- For instance, when ordering a burger from a fast-food chain or starting a new car, we expect a standardized product.
- In such cases, variation would likely be undesirable.

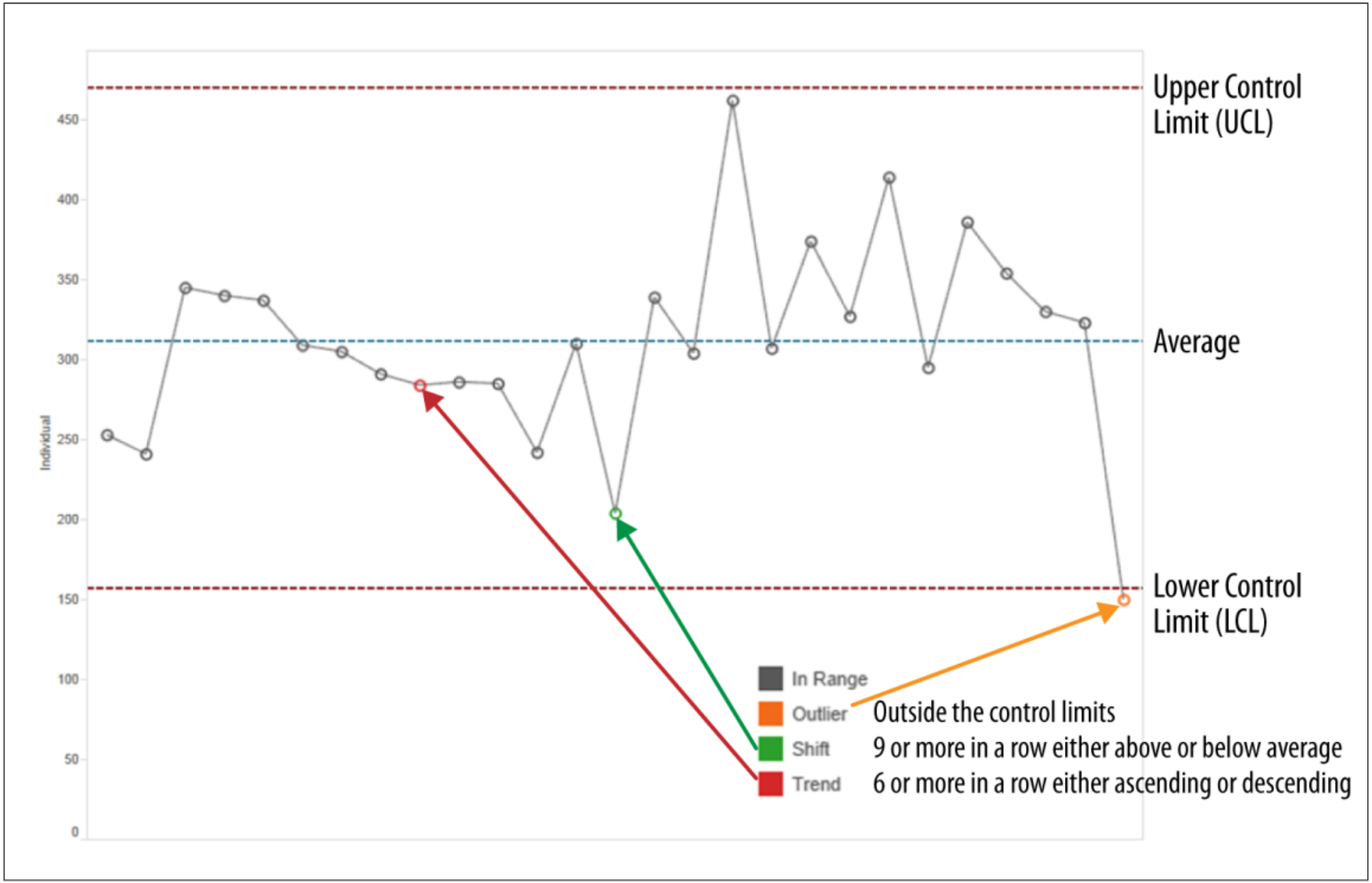


Figure 7-5. The elements of a Shewhart Control Chart

Variation over time-Control charts

Anatomy of a Control Chart:

- A control chart contains the following basic elements:
  1. The time series data itself
  2. The average line
  3. The control limits: UCL (the upper control limit) LCL (the lower control limit)
  4. Signals:
     - Outliers (data points either above the UCL or below the LCL)
     - Trends (six or more points either all ascending or all descending)
     - Shifts (nine or more points either all above or all below the average line)

Variation over time-Control charts

How to Create a Control Chart in Tableau

- Let's explore two methods for control chart analysis: the quick method and the rigorous method.
- The main distinction lies in how the control limits are determined.

- The quick method employs a global measure of dispersion, specifically the standard deviation of all data points.
- The rigorous method utilizes a local measure of dispersion known as Sigma(x), derived from the differences between successive data points.

Variation over time-Control charts

Example:

- Consider the total number of earthquakes recorded worldwide that registered magnitude 6.0 or higher on the Richter scale from 1983 through 2013.

- The source for the data is the USGS Earthquake Archive Search website.
- There were 4,136 such events recorded, and Figure 7-6 gives a view of the most recent records in the data set

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Date & Time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	type
2	2013-12-17 23:38:06	20.7727	146.7903	9	6.2	mww		14	4.162	0.81	us	usc000lmmc	2014-02-27T21:49:52.353Z	198km E of Farallon de Pajaros, Northern Mariana Islands	earthquake
3	2013-12-08 17:24:54	44.4438	149.1667	28	6	mww		26	4.726	0.99	us	usb000lds9	2014-02-15T02:35:45.869Z	134km SE of Kuril'sk, Russia	earthquake
4	2013-12-01 06:29:57	2.044	96.8261	20	6	mww		27	1.049	0.89	us	usb000l8pb	2014-02-12T02:20:23.821Z	69km SE of Sinabang, Indonesia	earthquake
5	2013-12-01 01:24:13	-7.0269	128.3791	9.87	6.4	mww		11	3.046	0.7	us	usb000l8mb	2014-02-12T02:20:47.278Z	Kepulauan Barat Daya, Indonesia	earthquake
6	2013-11-25 07:21:18	-53.8708	-53.9107	14.83	6	mwc		58	3.346	0.76	us	usb000l633	2014-02-11T02:22:33.206Z	South Atlantic Ocean	earthquake
7	2013-11-25 06:27:33	-53.9451	-55.0033	11.78	7	mww		31	2.935	1.08	us	usb000l5zn	2014-02-11T02:25:27.101Z	Falkland Islands region	earthquake
8	2013-11-25 05:56:50	45.5613	151.0047	34	6	mww		26	5.885	0.66	us	usb000l5z1	2014-02-11T02:32:38.383Z	247km E of Kuril'sk, Russia	earthquake
9	2013-11-23 07:48:32	-17.1171	-176.5449	371	6.5	mww		22	5.194	0.83	us	usb000l51g	2014-02-11T02:22:07.739Z	Fiji region	earthquake
10	2013-11-19 17:00:44	18.4753	145.2041	511	6	mww		10	1.848	1.05	us	usb000l25i	2014-02-11T02:29:01.431Z	58km WSW of Agrihan, Northern Mariana Islands	earthquake
11	2013-11-19 13:32:51	2.6403	128.4339	38	6	mww		19	2.14	1.01	us	usb000l219	2014-02-11T02:36:43.624Z	111km NNE of Tobelo, Indonesia	earthquake
12	2013-11-17 09:04:55	-60.2738	-46.4011	10	7.7	mww		23	8.05	1.33	us	usb000l0gq	2014-01-31T21:29:01.439Z	Scotia Sea	earthquake
13	2013-11-16 03:34:31	-60.2627	-47.0621	9.97	6.9	mww		17	8.284	0.84	us	usb000kznc	2014-01-31T21:32:02.803Z	Scotia Sea	earthquake
14	2013-11-13 23:45:47	-60.2814	-47.1233	11.07	6.1	mww		23	8.319	1.19	us	usb000kxhr	2014-01-31T21:26:05.362Z	Scotia Sea	earthquake
15	2013-11-12 07:03:51	54.6859	162.3024	43	6.4	mww		20	2.73	0.87	us	usb000kw1x	2014-01-31T21:35:40.466Z	172km S of Ust'-Kamchatsk Staryy, Russia	earthquake
16	2013-11-02 18:53:46	-19.1711	-172.6411	10.05	6.2	mww		21	5.297	0.72	us	usb000krlz	2014-01-10T13:04:16.196Z	152km ESE of Neiafu, Tonga	earthquake
17	2013-11-02 15:52:46	-23.6357	-112.5956	9.98	6	mww		35	4.558	0.81	us	usb000krjt	2014-01-10T13:03:30.639Z	Easter Island region	earthquake
18	2013-10-31 23:03:59	-30.2921	-71.5215	27	6.6	mww		31	0.636	1.28	us	usb000kqnc	2014-01-10T13:06:12.899Z	41km SSW of Coquimbo, Chile	earthquake
19	2013-10-31 12:02:08	23.5904	121.4366	10	6.3	mww		15	0.234	1.29	us	usc000ksdy	2014-01-10T13:05:25.588Z	46km SSW of Hualian, Taiwan	earthquake
20	2013-10-30 02:51:47	-35.314	-73.395	41.5	6.2	mww				1.68	us	usc000kr9k	2014-01-10T13:04:41.053Z	88km W of Constitucion, Chile	earthquake
21	2013-10-25 17:10:19	37.1557	144.6611	35	7.1	mww		10	3.968	1.01	us	usc000kn4n	2014-01-03T00:48:15.801Z	Off the east coast of Honshu, Japan	earthquake
22	2013-10-24 19:25:10	-58.153	-12.7964	22.87	6.7	mww		53	13.711	0.99	us	usc000kmfw	2014-01-03T00:40:15.133Z	East of the South Sandwich Islands	earthquake
23	2013-10-23 08:23:30	-23.0067	-177.1425	160	6	mwb		19	6.252	0.84	us	usb000kj1z	2014-02-21T19:59:38.000Z	283km SW of Vaini, Tonga	earthquake

Figure 7-6. Sample of global earthquakes data set, registering magnitude 6.0 or greater

Variation over time-Control charts

Quick Method:

- Create a simple timeline with a YEAR (Date & Time) on the Columns shelf, and SUM (Number of Records) on the Rows shelf, fit to width as shown in Figure 7-7.

- Right-click on the y-axis, select Add Reference Line, and add an average line by filling out the resulting dialog box.
- Then right-click on the y-axis, select Add Reference Line again, and this time add a distribution of +3 and −3 times the standard deviation, with dotted red lines and no fill.
- Both reference line dialog boxes are shown in Figure 7-8

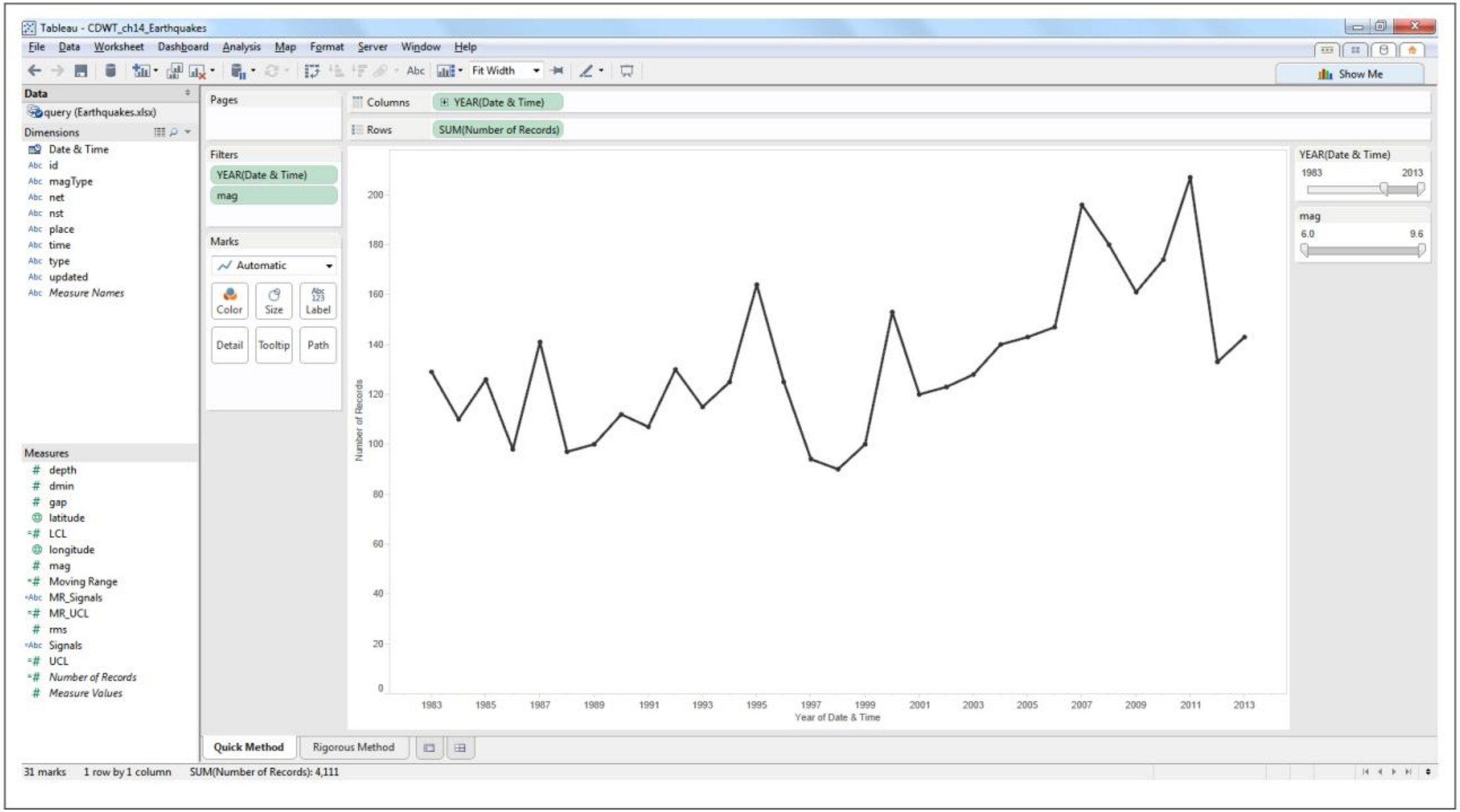


Figure 7-7. A simple timeline of the number of annual earthquakes

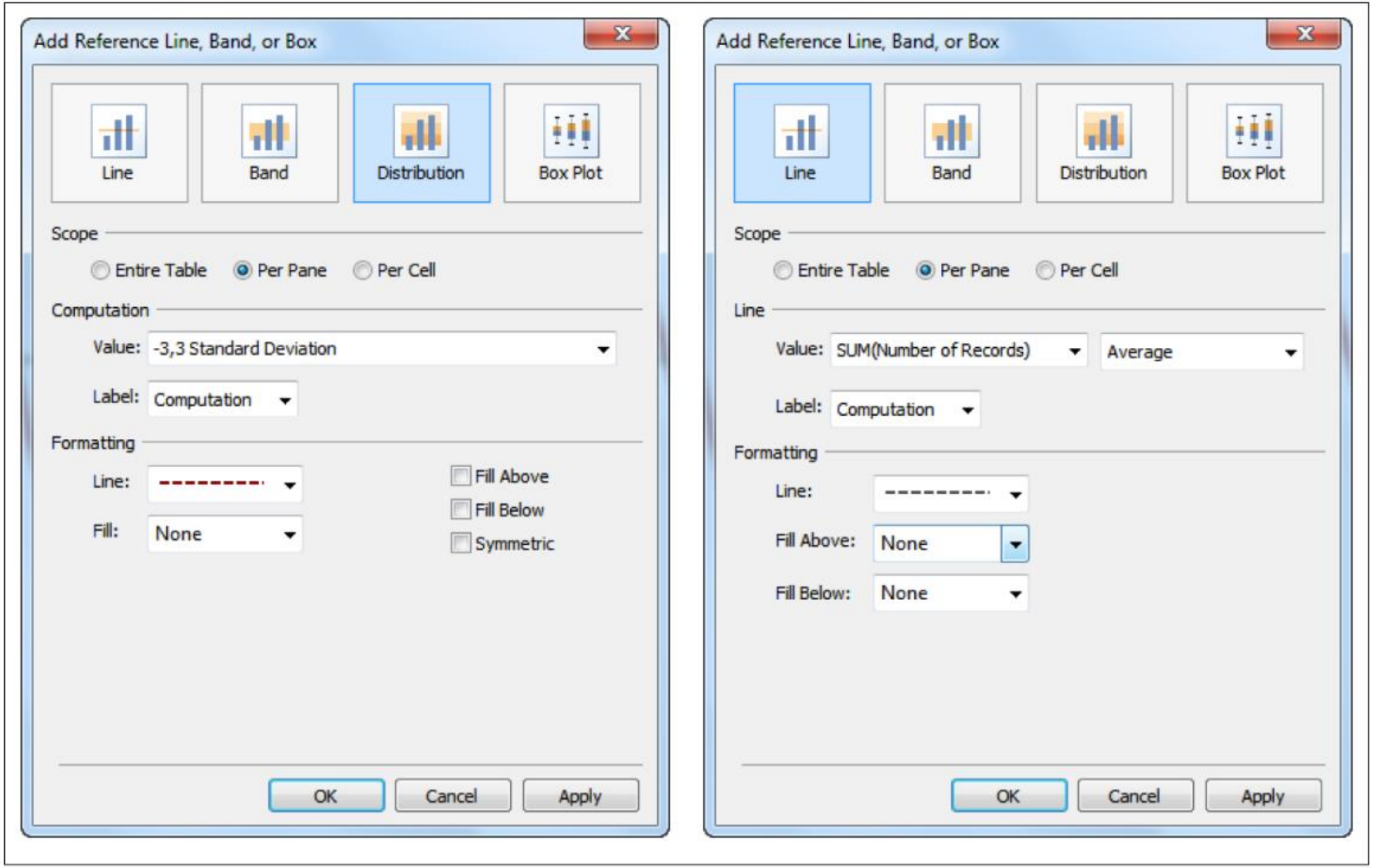


Figure 7-8. Adding reference lines to the line chart

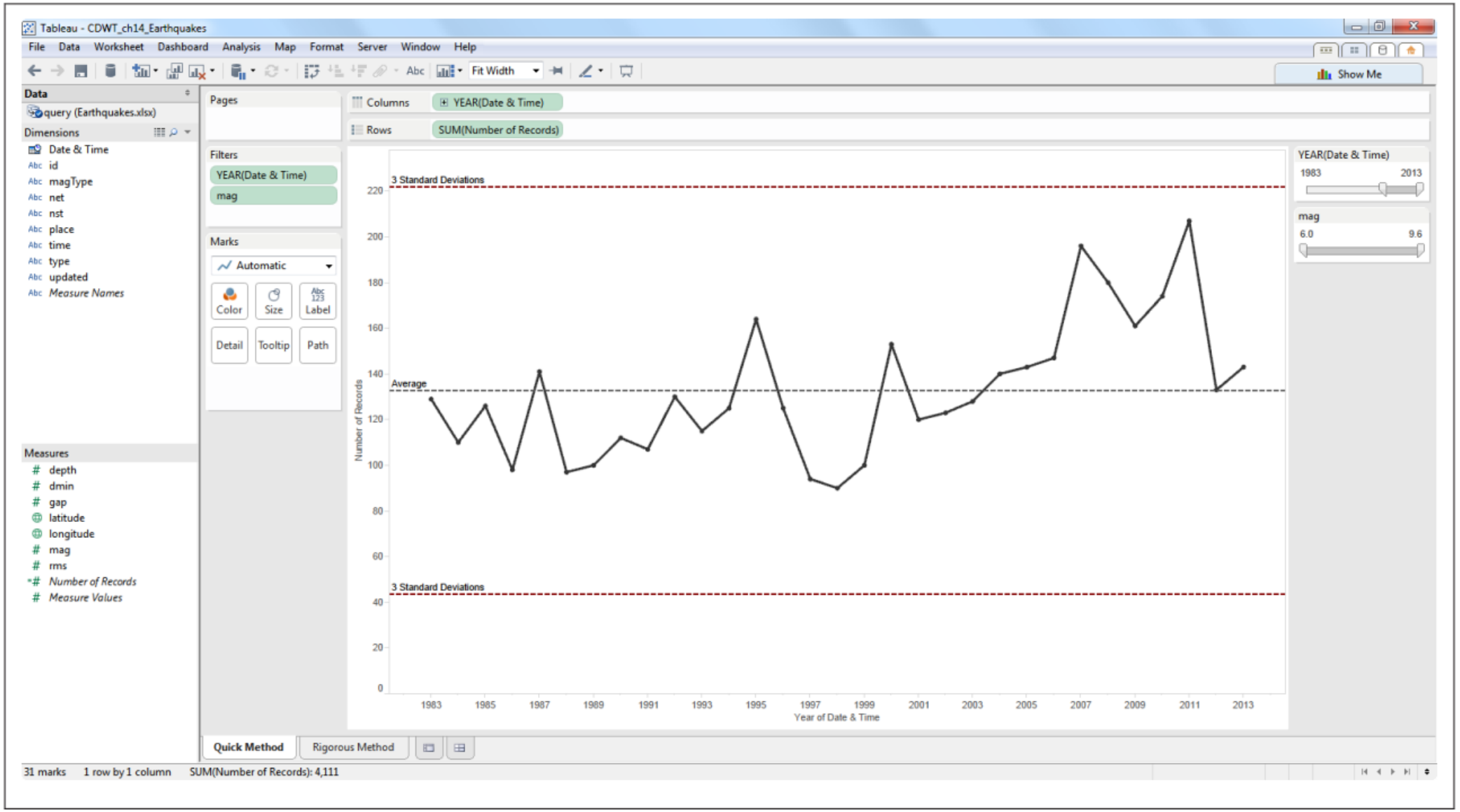


Figure 7-9. Simple control chart of annual earthquakes of magnitude 6.0 or greater

Variation over time-Control charts

- If we change from YEAR to MONTH, then the control chart changes to show several points above the 3-sigma line, including a sharp outlier in March 2011 corresponding to the Great East Japan earthquake, as shown in Figure 7-10

- Also note that the lower limit is not real. It's below 0, and it's not possible to have a negative number of earthquakes recorded

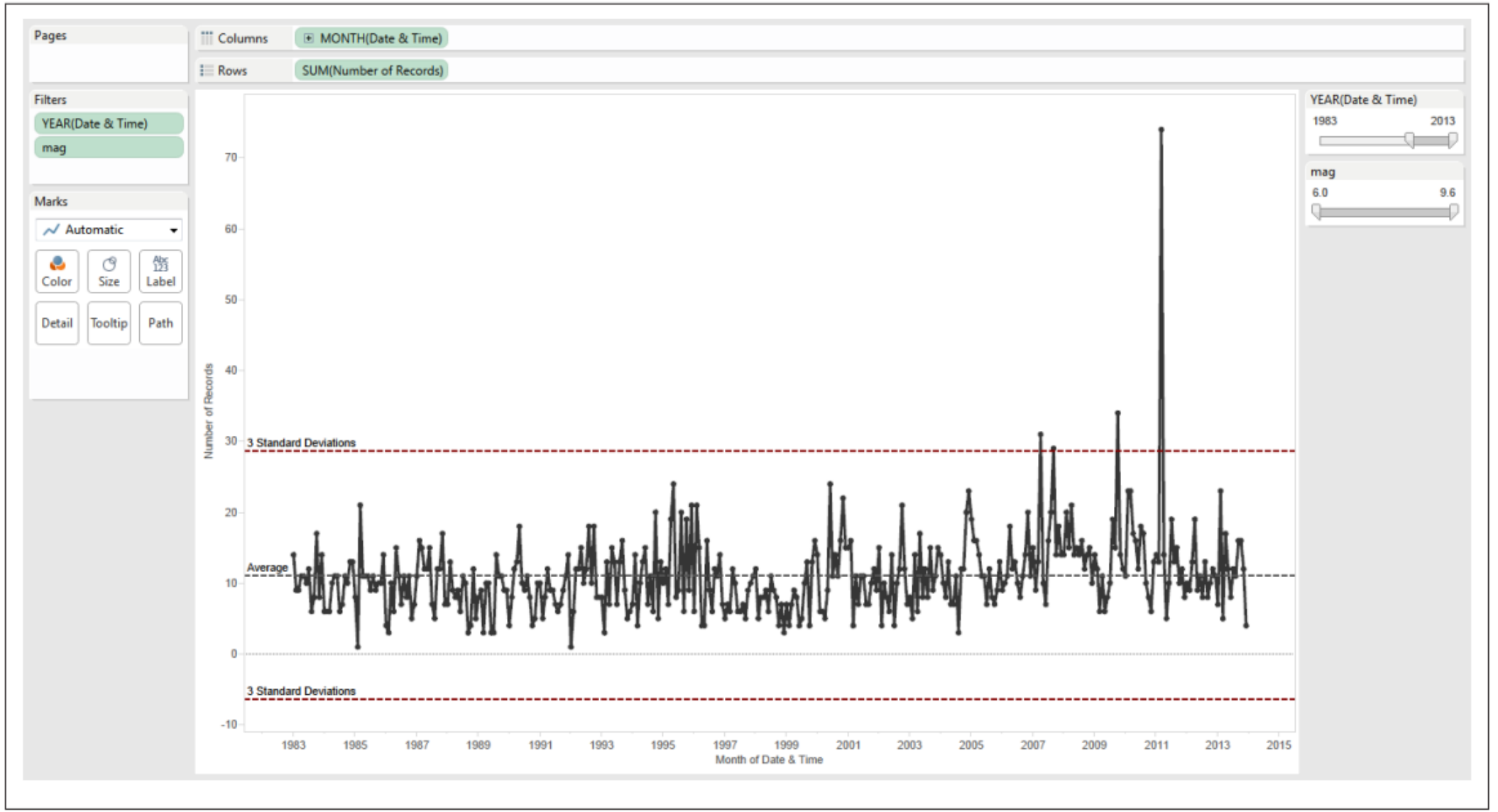


Figure 7-10. The simple control chart showing monthly counts of worldwide earthquakes

Variation over time-Control charts

The rigorous method:

- Create a new sheet and begin with Step 1 of the quick method outlined in the previous section to establish a basic timeline.
- Duplicate the SUM(Number of Records) and generate a dual-axis plot with synchronized axes.
- Represent the first set of marks as a line and the second set as circles, as illustrated in Figure 7-11.
- Additionally, introduce extra elements like a "Moving Range" timeline, which displays the absolute value of the change from one quake to another.

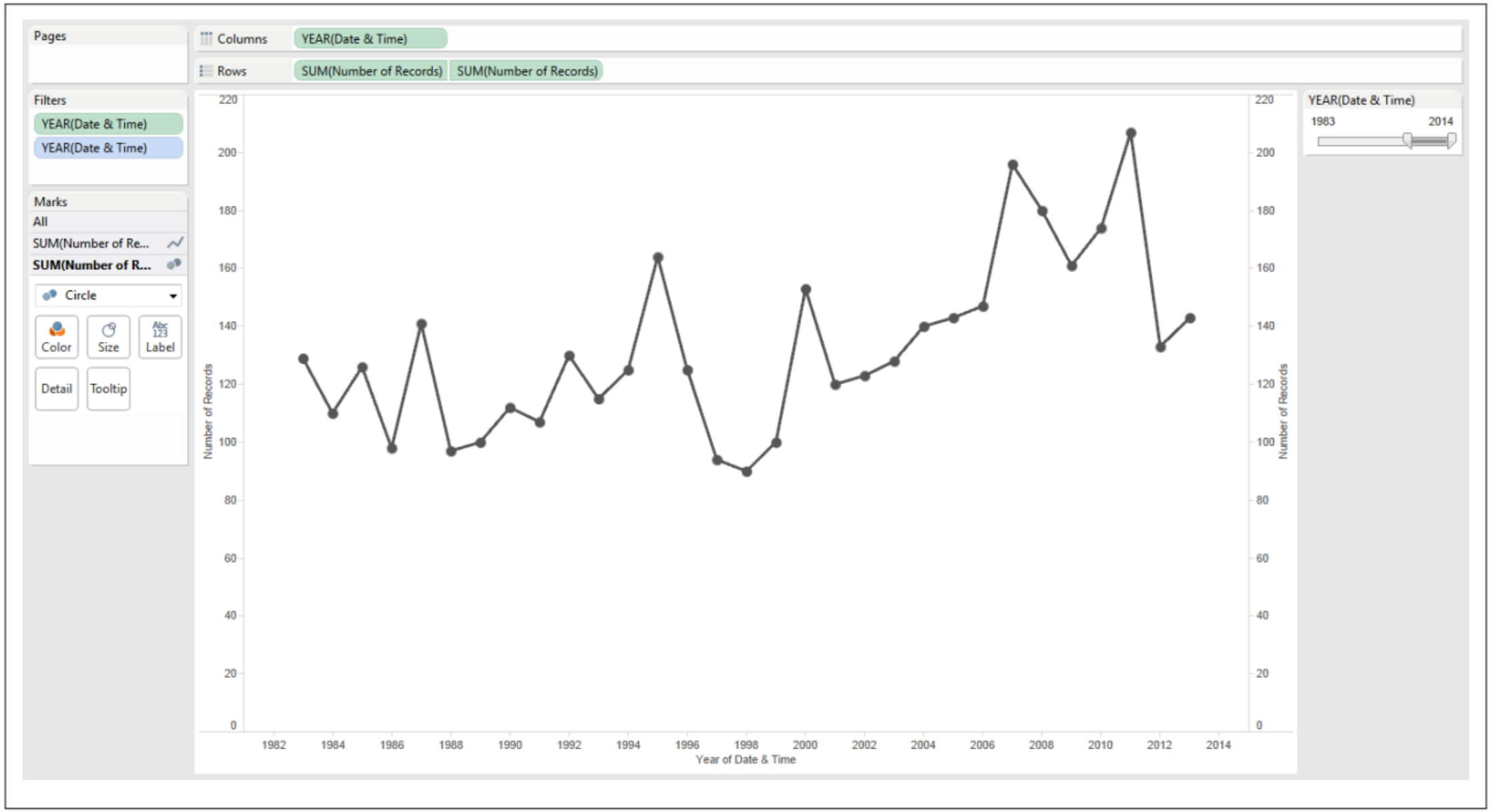


Figure 7-11. Dual-axis timeline of annual earthquake count

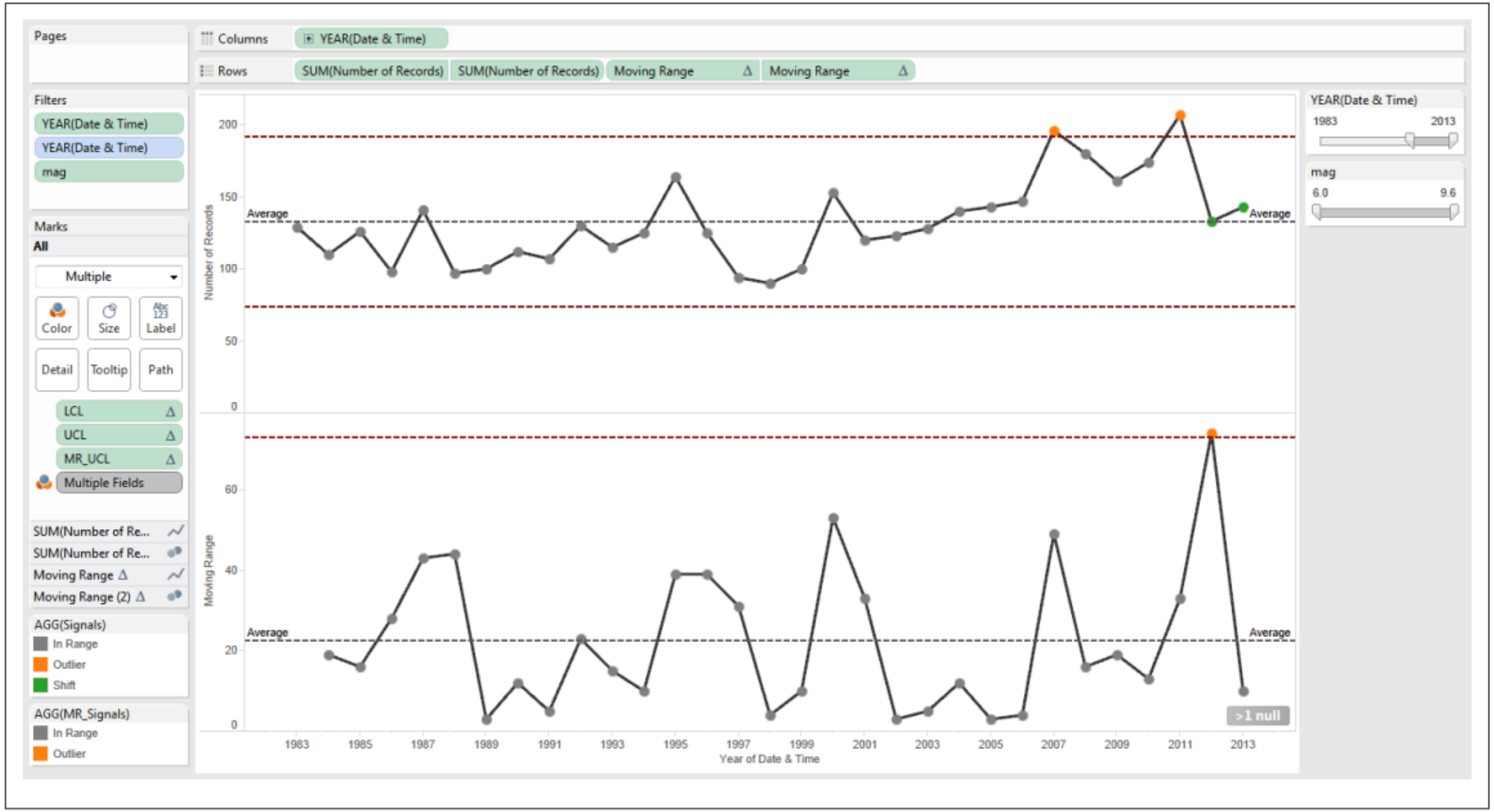


Figure 7-15. The rigorous control chart showing worldwide earthquake count by year

Variation and Uncertainty

1. Respecting variation
2. Variation over time-Control charts
3. Understanding uncertainty

Understanding uncertainty

- When we present data to an audience, it's important to make clear whether we are presenting data from the entire population, or from just a sample of the population.

- To illustrate the point, let's consider a fictional case of chess club participation rates of students in ten different cities in the peaceful country of Chesslandia.
- The President of Chesslandia, Garry Fischer, wanted to know whether students in his country were involved in chess club or not, and which cities were more successful in getting students to participate.

	A	B	C	D
1	**Grade**	**Town**	**In Chess Club**	**Not in Chess Club**
2	5th-6th grade	Kingston	29	40
3	5th-6th grade	Queensville	39	30
4	5th-6th grade	Rooktown	39	31
5	5th-6th grade	Bishop Village	18	32
6	5th-6th grade	Knightfield	35	33
7	5th-6th grade	Pawnford	36	18
8	5th-6th grade	Castleborough	38	22
9	5th-6th grade	Checkshire	33	31
10	5th-6th grade	Fianchettoberg	32	33
11	5th-6th grade	Gambitopolis	31	30
12	3rd-4th grade	Kingston	14	28
13	3rd-4th grade	Queensville	16	22
14	3rd-4th grade	Rooktown	13	21
15	3rd-4th grade	Bishop Village	7	26
16	3rd-4th grade	Knightfield	15	26
17	3rd-4th grade	Pawnford	14	21
18	3rd-4th grade	Castleborough	22	14
19	3rd-4th grade	Checkshire	15	27
20	3rd-4th grade	Fianchettoberg	14	27
21	3rd-4th grade	Gambitopolis	11	27
22	1st-2nd grade	Kingston	7	20
23	1st-2nd grade	Queensville	6	12
24	1st-2nd grade	Rooktown	8	20
25	1st-2nd grade	Bishop Village	4	17
26	1st-2nd grade	Knightfield	6	20
27	1st-2nd grade	Pawnford	6	12
28	1st-2nd grade	Castleborough	10	14
29	1st-2nd grade	Checkshire	7	20
30	1st-2nd grade	Fianchettoberg	8	15
31	1st-2nd grade	Gambitopolis	6	16

Figure 7-17. Spreadsheet of results of chess club student survey

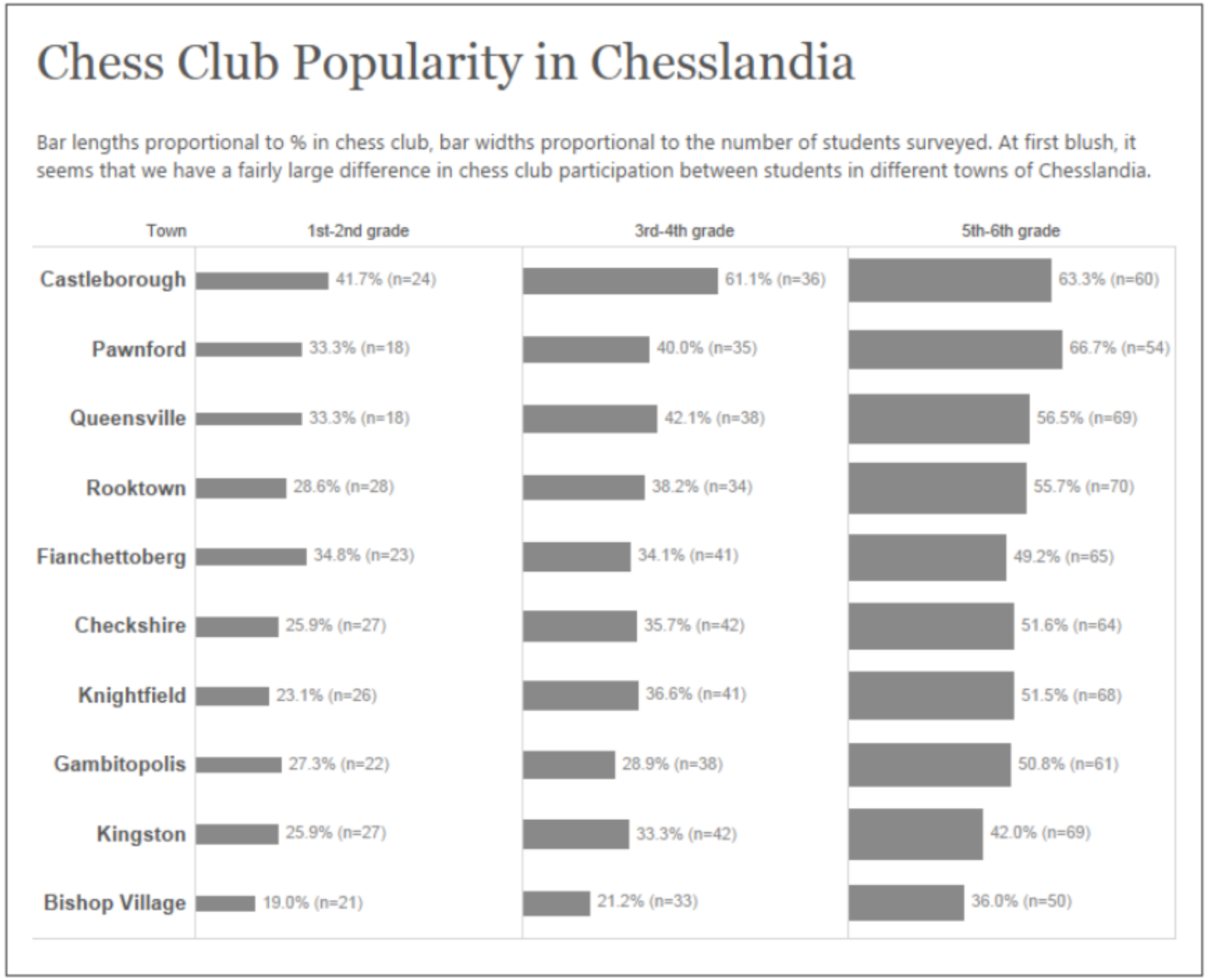


Figure 7-18. Visualization of the survey results

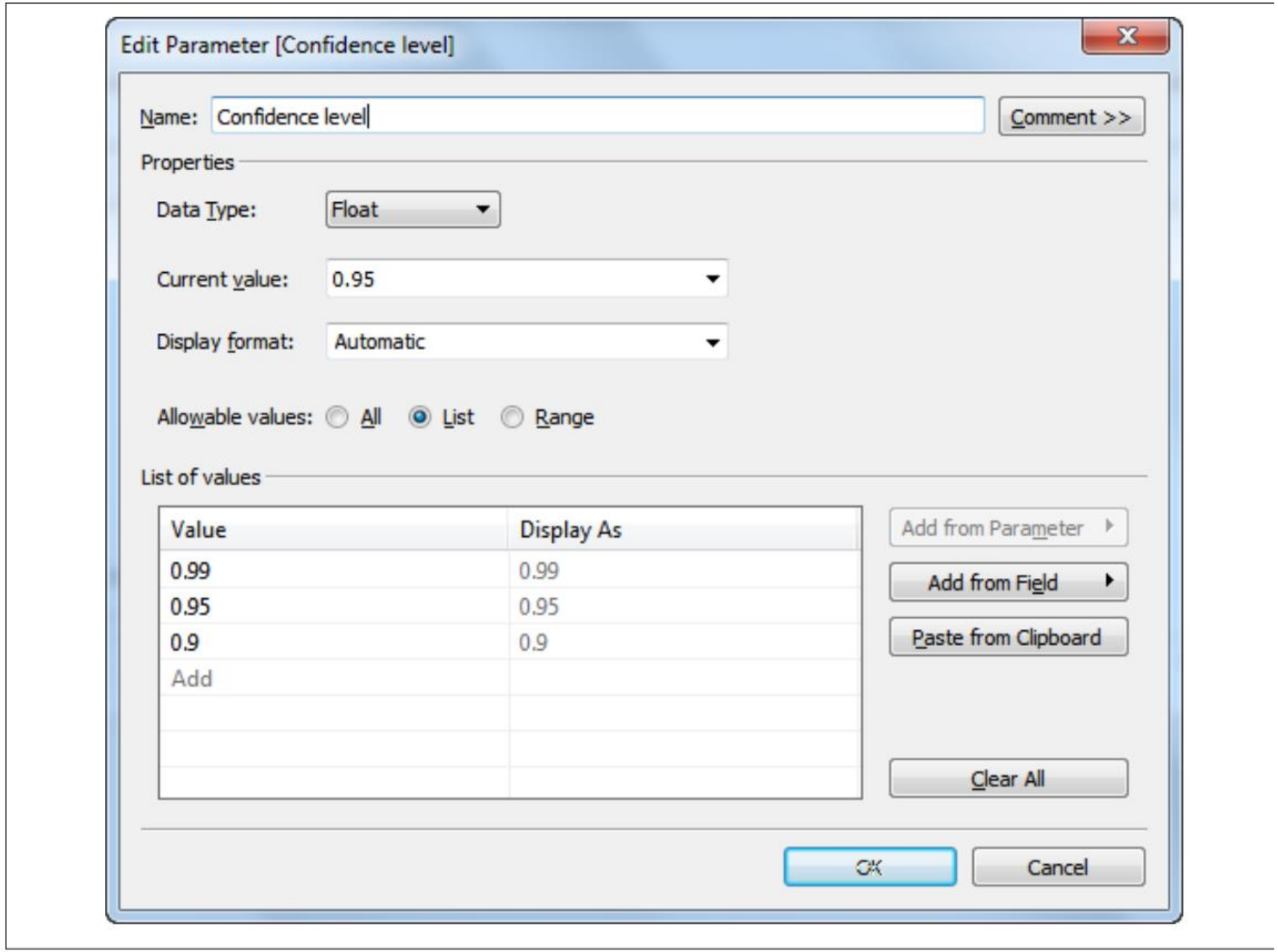


Figure 7-19. A new parameter to select desired confidence level

Understanding uncertainty

Next, he needed to create several calculated fields to generate the error bars:

- % in Chess Club, p = [In Chess Club]/([In Chess Club]+[Not in Chess Club])
- Sample Size, n = [In Chess Club]+[Not in Chess Club]
- Standard Error = SQRT(([% in Chess Club]*(1–[% in Chess Club]))/[n])
- z upper = CASE [Confidence level]
- — WHEN 0.99 THEN 2.575829
- — WHEN 0.95 THEN 1.959964
- — WHEN 0.90 THEN 1.644854

Understanding uncertainty

- Margin of Error = [Standard Error]*[z upper]
- Lower limit = [% in Chess Club]–[Margin of Error]
- Upper limit = [% in Chess Club]+[Margin of Error]
- Error Bar Line = [Upper limit]–[Lower limit]
- np = [n]*[% in Chess Club]

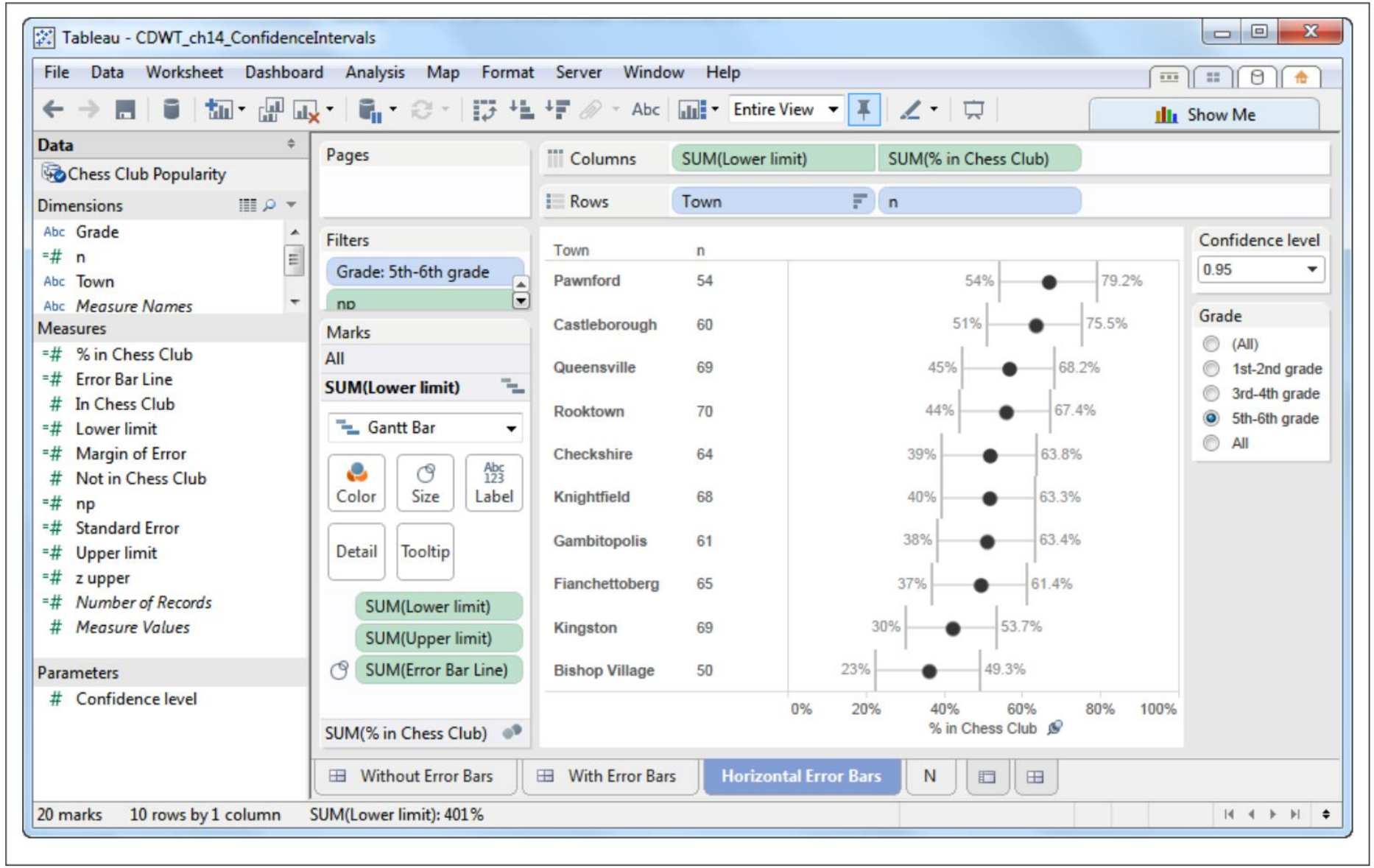


Figure 7-20. A binomial probability distribution for the chess club survey results

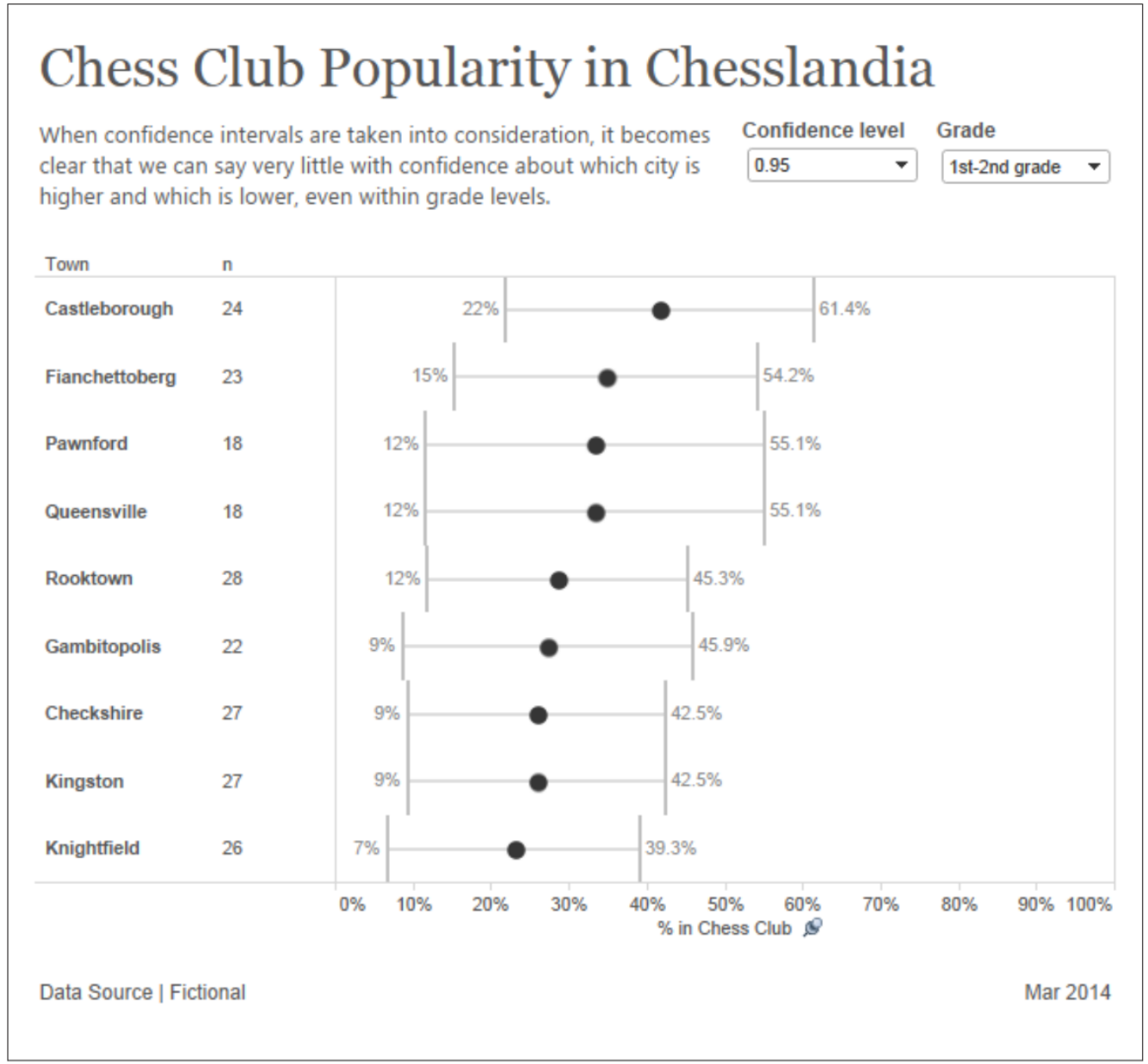


Figure 7-21. 95% confidence intervals for first- and second-grade chess club participation

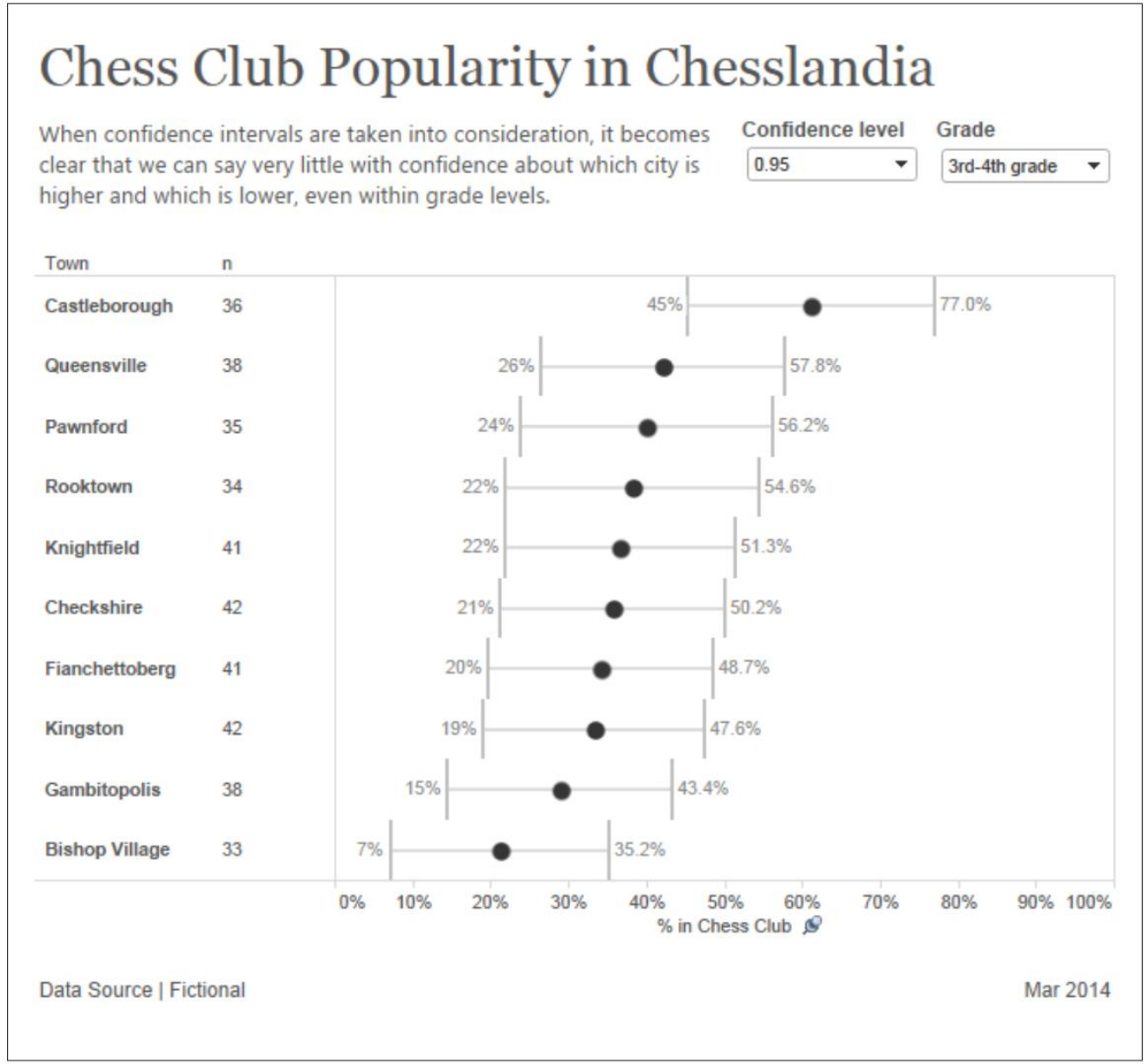


Figure 7-22. 95% confidence intervals for third- and fourth-grade chess club participation

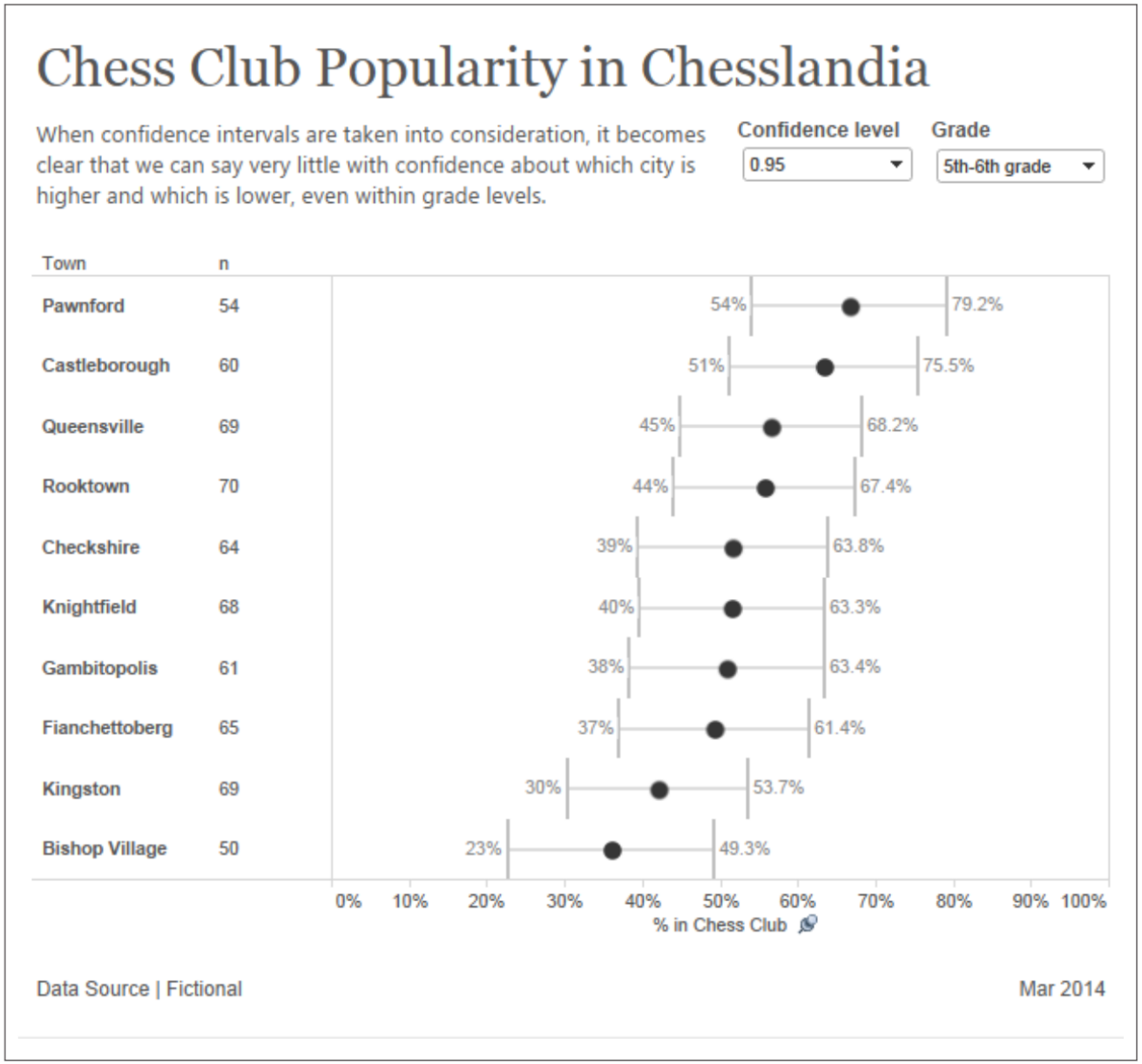


Figure 7-23. 95% confidence intervals for fifth- and sixth-grade chess club participation