Descriptive Statistics: Part 1/2 (Ch 3)

Yifan Zhu

Iowa State University

Outline

What is descriptive statistics?

Graphical and Tabular Displays

- Dot diagrams
- Stem and leaf plots
- Frequency tables
- Histograms
- Bar plots
- Scatterplots

Quantiles

What is descriptive statistics?

- **Descriptive statistics**: the use of plots and numerical summaries to describe data without drawing any formal conclusions. Exploratory data analysis (EDA)
- Descriptive statistics seeks to find the following features of datasets:
  mean, median . . .
  - Center: the point that the data are closest to on average
  - Spread: how wide the data look, how varied the points are
    variability
  - Shape (more on that when we get to plots)
  - Outliers: points that lie way beyond the rest of the data.

Unnormal observations /
sample units /
data points /

Outline

What is descriptive statistics?

Graphical and Tabular Displays

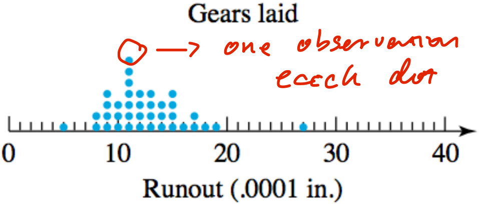
- Dot diagrams
- Stem and leaf plots
- Frequency tables
- Histograms
- Bar plots
- Scatterplots

Quantiles

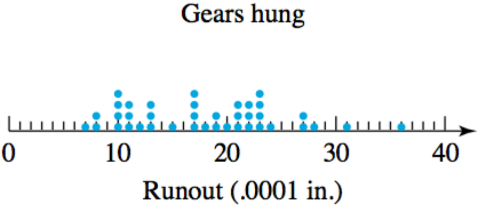
Gear data

dot diagram.

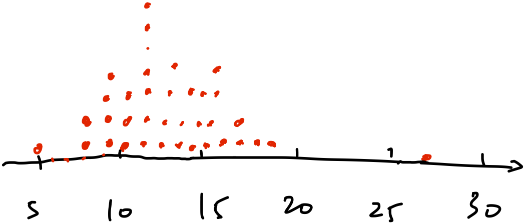
↓



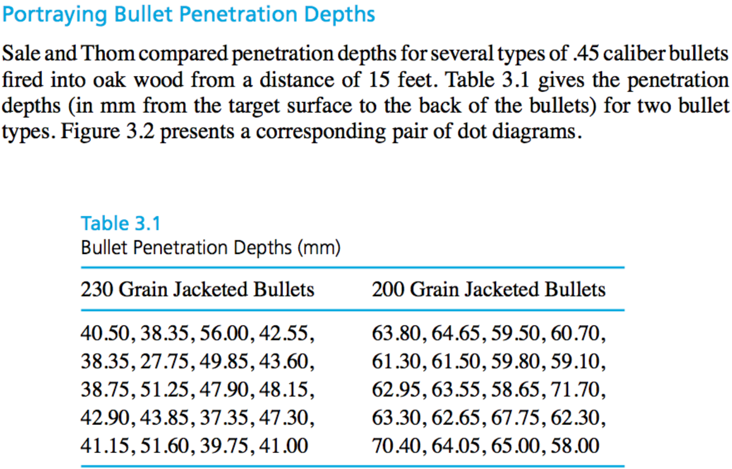
→ one observation for each dot



Gears Laid
5, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 13, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 16, 17, 17, 18, 19, 27



New example: bullet data



Portraying Bullet Penetration Depths

Sale and Thom compared penetration depths for several types of .45 caliber bullets fired into oak wood from a distance of 15 feet. Table 3.1 gives the penetration depths (in mm from the target surface to the back of the bullets) for two bullet types. Figure 3.2 presents a corresponding pair of dot diagrams.

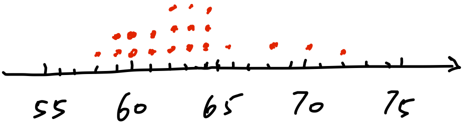
Table 3.1
Bullet Penetration Depths (mm)

230 Grain Jacketed Bullets	200 Grain Jacketed Bullets
40.50, 38.35, 56.00, 42.55,	63.80, 64.65, 59.50, 60.70,
38.35, 27.75, 49.85, 43.60,	61.30, 61.50, 59.80, 59.10,
38.75, 51.25, 47.90, 48.15,	62.95, 63.55, 58.65, 71.70,
42.90, 43.85, 37.35, 47.30,	63.30, 62.65, 67.75, 62.30,
41.15, 51.60, 39.75, 41.00	70.40, 64.05, 65.00, 58.00

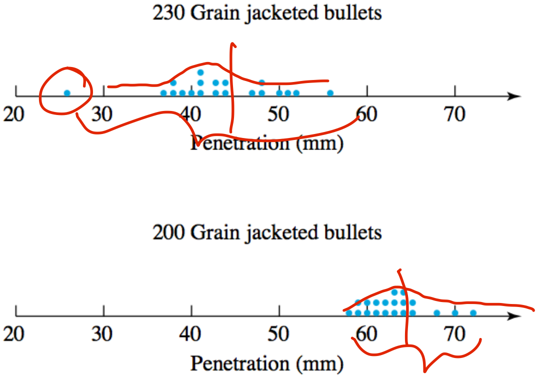
200 Grain Jacketed Bullets
63.80, 64.65, 59.50, 60.70, 61.30,
61.50, 59.80, 59.10, 62.95, 63.55,
58.65, 71.70, 63.30, 62.65, 67.75,
62.30, 70.40, 64.05, 65.00, 58.00

58 - 72 .

round to integer



Bullet data


230 Grain jacketed bullets
20
30
40
50
60
70
Penetration (mm)
200 Grain jacketed bullets
20
30
40
50
60
70
Penetration (mm)

Stem and leaf plots: laid gears

Gears Laid
5, 8, 8, 9, 9, 9, 9, 10, 10, 10,
11, 11, 11, 11, 11, 11, 11, 12, 12, 12,
12, 13, 13, 13, 13, 14, 14, 14, 15, 15,
15, 15, 16, 17, 17, 18, 19, 27

tens digit : stem

units digit : leaf .

stem	leaf
0	5 8 8 9 9 9 9
1	0 0 0 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 5 5 5 5 6 7 7 8 9
2	7

Stem and leaf plots: laid gears

```
0 | 5 8 9 9 9 9
1 | 0 0 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 5 5 5 5 6 7 7 8 9
2 | 7
3 |
```

```
0 |
0 | 5 8 9 9 9 9
1 | 0 0 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4
1 | 5 5 5 6 7 7 8 9
2 |
2 | 7
3 |
3 |
```

Back to stem and leaf plots

Laid runouts		Hung runouts
9 9 9 9 8 8 5	0	7 8 8
4 4 4 3 3 3 3 2 2 2 2 1 1 1 1 1 1 1 0 0 0	1	0 0 0 0 1 1 1 2 3 3 3
9 8 7 7 6 5 5 5 5	1	5 7 7 7 7 8 9 9
	2	0 1 1 1 2 2 2 3 3 3 3 4
7	2	7 7 8
	3	1
	3	6

Frequency Table: gear data

$$\frac{\text{frequency}}{\text{sample size}}$$

||

Frequency Table for Laid Gear Thrust Face Runouts

↓ intervals

Runout (.0001 in.)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
5–8	|||	3	.079	.079
9–12	𝍸 𝍸 𝍸 |||	18	.474	.553 = .079 + .474
13–16	𝍸 𝍸 ||	12	.316	.868
17–20	||||	4	.105	.974
21–24		0	0	.974
25–28	|	1	.026	1.000
		38	1.000	

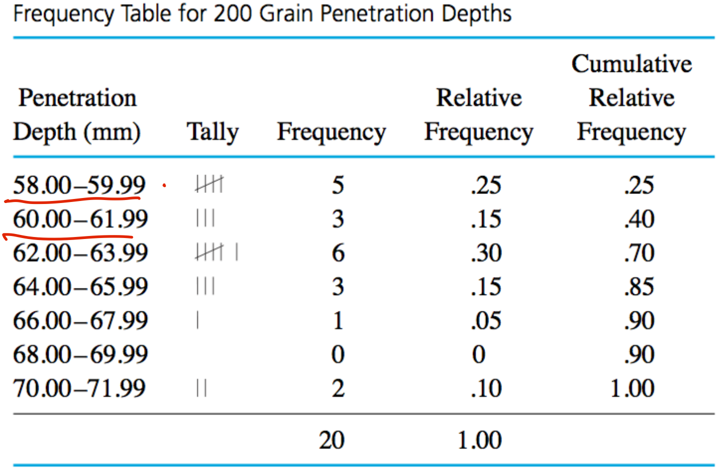
↓ sample size

5-8

: $\{x: 5 \leq x < 9\}$

$\{x: 5 \leq x \leq 8\} = [5, 9)$

Frequency Table: bullet data, 200 grain



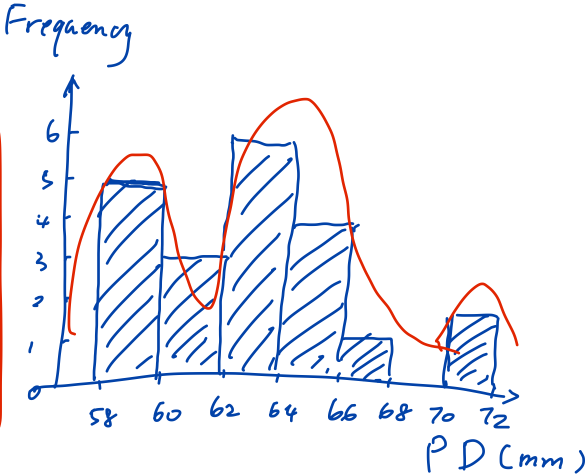
Frequency Table for 200 Grain Penetration Depths

Penetration Depth (mm)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
58.00–59.99	𝍸	5	.25	.25
60.00–61.99	|||	3	.15	.40
62.00–63.99	𝍸 |	6	.30	.70
64.00–65.99	|||	3	.15	.85
66.00–67.99	|	1	.05	.90
68.00–69.99		0	0	.90
70.00–71.99	||	2	.10	1.00
		20	1.00	

$[58, 60)$ $[60, 62)$

Histogram

Penetration Depth (mm)	Frequency
[58, 60)	5
[60, 62)	3
[62, 64)	6
[64, 66)	3
[66, 68)	1
[68, 70)	0
[70, 72)	2



h: width of the interval

n_i: frequency of the i-th interval

Histogram: bullet data, 200 grain

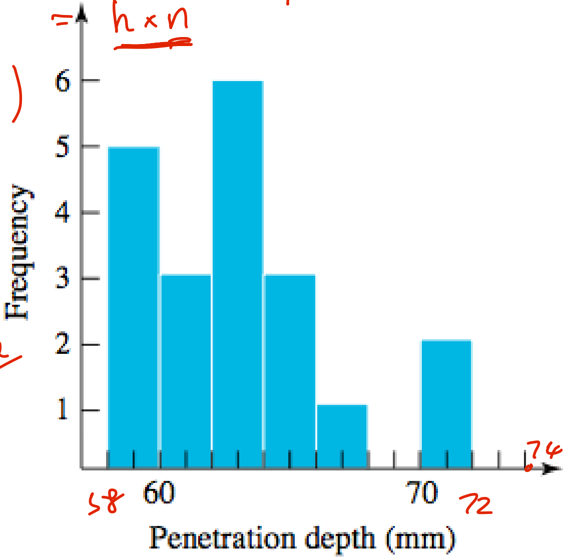
Q: what is the area of the boxes.

$\sum h \times n_i = \underline{h \times n}$

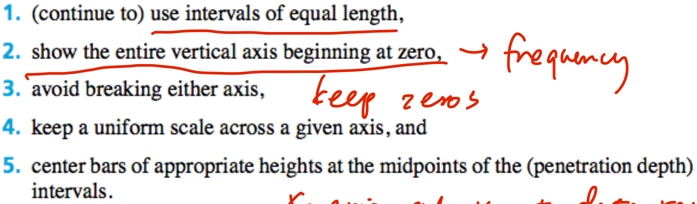
$\sum h \cdot \left(\frac{n_i}{n}\right)$

$= h$

Relative



Histogram guidelines



1. (continue to) use intervals of equal length,
2. show the entire vertical axis beginning at zero, → frequency
3. avoid breaking either axis, keep zeros
4. keep a uniform scale across a given axis, and
5. center bars of appropriate heights at the midpoints of the (penetration depth) intervals.

X- axis adaptive to data range

- Also: histograms are for continuous data only. The equivalent plot for discrete and categorical data is called a *bar plot*, featured next.

Discrete data: cars

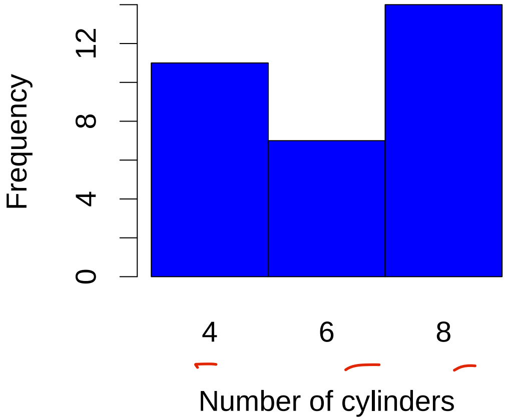
discrete variable

	mpg	cyl
Mazda RX4	21	6
Mazda RX4 Wag	21	6
Datsun 710	22.8	4
Hornet 4 Drive	21.4	6
Hornet Sportabout	18.7	8
Valiant	18.1	6
Duster 360	14.3	8
Merc 240D	24.4	4
Merc 230	22.8	4
Merc 280	19.2	6
Merc 280C	17.8	6
Merc 450SE	16.4	8
Merc 450SL	17.3	8
Merc 450SLC	15.2	8
Cadillac Fleetwood	10.4	8
...	...	...

Discrete data frequency table: cars data

Cylinders	Freq.	Relative Freq.	Cumulative Rel. Freq.
4	11	0.344	0.344
6	7	0.219	0.563
8	14	0.4375	1

Bar plot (not a histogram)



categories

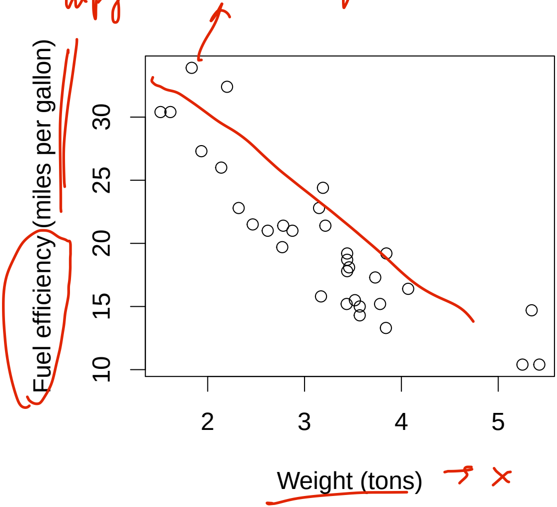
Bivariate data: cars

	mpg	wt
Mazda RX4	21	2.62
Mazda RX4 Wag	21	2.875
Datsun 710	22.8	2.32
Hornet 4 Drive	21.4	3.215
Hornet Sportabout	18.7	3.44
Valiant	18.1	3.46
Duster 360	14.3	3.57
Merc 240D	24.4	3.19
Merc 230	22.8	3.15
Merc 280	19.2	3.44
Merc 280C	17.8	3.44
Merc 450SE	16.4	4.07
Merc 450SL	17.3	3.73
Merc 450SLC	15.2	3.78
Cadillac Fleetwood	10.4	5.25
...	...	...

Scatterplot: mpg vs wt, cats data

mpg → y

(x, y)



→ x

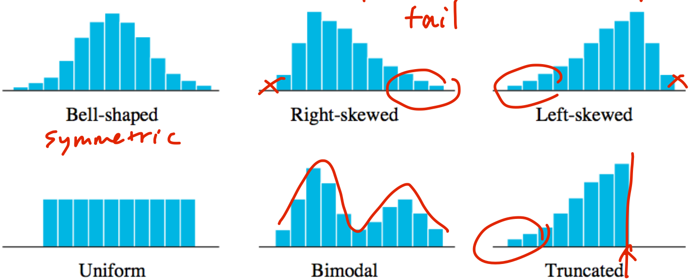
negative

linear

Distributional shapes

mean = median   mean > median   mean < median

Why do we plot data? To see the distributional shape.



$x_1 < \cdots < x_5$

mean : average $(x_1 + \cdots + x_5)/5$

median: 50% > . 50% < , x_3

Outline

What is descriptive statistics?

Graphical and Tabular Displays

- Dot diagrams
- Stem and leaf plots
- Frequency tables
- Histograms
- Bar plots
- Scatterplots

Quantiles

Percentiles and quantiles

- **The *p*'th percentile of a dataset**: a number greater than *p* % of the data and less than the rest.
  - "You scored at the 90'th percentile on the SAT" means that your score was higher than 90% of the students who took the test and lower than the other 10%
  - "Zorbit was positioned at the 80th percentile of the list of fastest growing companies compiled by INC magazine." means Zorbit was growing faster than 80% of the companies in the list and below the other 20%.
- **The *p* quantile of a dataset**: a percentile, except with *p* expressed as a decimal number, not a percentage.
  - "You scored at the 0.9 quantile on the SAT"
  - "Zorbit was positioned at the 0.8 quantile of the list compiled by INC magazine."

$0 \leq p \leq 1$

Calculating quantiles of finite datasets: setup

- Given:
  - $x_1, \ldots x_n$, an ordered list of numbers. This is the dataset. small to big
  - p, a number between 0 and 1.
- Goal: calculate $Q(p)$, the p quantile of the dataset.
- Notation:
  - $Q(p)$ is called the **quantile function**.
  - $\lfloor x \rfloor$ is called the **floor function**.
  - $\lceil x \rceil$ is called the **ceiling function**.

$\lfloor x \rfloor$: largest integer that is smaller than x

$\lfloor 5.2 \rfloor \rightarrow 5$, $\lfloor -1.2 \rfloor \rightarrow -2$

$\lceil x \rceil$: smallest integer that is bigger than x

$\lceil 5.2 \rceil \rightarrow 6$ $\lceil -1.2 \rceil = -1$

Calculating quantiles of finite datasets: procedure

1. Let $p_i = \frac{i-.5}{n}$, $i = 1, \ldots, n$
2. Define $Q(p_i) = x_i$ for $i = 1, \ldots n$.
   a. If $p = p_j$ for some index j, then $Q(p) = Q(p_j)$.
   b. Otherwise, linearly interpolate $Q(p)$:
      i. Let $i' = np + .5$ (Solve $p = \frac{i' - .5}{n}$ for i').
      ii. Take $Q(p) = (\lceil i' \rceil - i')x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor)x_{\lceil i' \rceil}$

$$i' = 5.2 . \quad \lceil i' \rceil = 6. \quad \lfloor i' \rfloor = 5$$

$$Q_p = (\underbrace{6 - 5.2}_{1-\text{decimal part}}) \cdot x_5 + (\underbrace{5.2 - 5}_{\text{decimal part}}) x_6$$

Example: breaking strength (g) of towels

test	strength
1	8577
2	9471
3	9011
4	7583
5	8572
6	10688
7	9614
8	9614
9	8527
10	9165

Example: breaking strength (g) of towels

p_i

$\parallel$

test	$\frac{i-.5}{10}$	i'th smallest data point, $x_i = Q(\frac{i-.5}{10})$
1	0.05	7583
2	0.15	8527
3	0.25	8572
4	0.35	8577
5	0.45	9011
6	0.55	9165
7	0.65	9471
8	0.75	9614
9	0.85	9614
10	0.95	10688

ordered

$n = 10$

Your turn: calculate $Q(0.5)$, $Q(0.18)$, and $Q(0.94)$.

pi

test	$\frac{i-.5}{10}$	i'th smallest data point, $x_i = Q(\frac{i-.5}{10})$
1	0.05	7583
2	0.15	8527
3	0.25	8572
4	0.35	8577
5	0.45	9011
6	0.55	9165
7	0.65	9471
8	0.75	9614
9	0.85	9614
10	0.95	10688

Q(0.5)

$$\begin{aligned} i' &= np + .5 \\ &= 10 \cdot 0.5 + 0.5 = 5.5 \end{aligned}$$

$$\begin{aligned} Q(0.5) &= (\lceil i' \rceil - i')x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor)x_{\lceil i' \rceil} \\ &= (\lceil 5.5 \rceil - 5.5)x_{\lfloor 5.5 \rfloor} + (5.5 - \lfloor 5.5 \rfloor)x_{\lceil 5.5 \rceil} \\ &= \underline{(6 - 5.5)x_5} + \underline{(5.5 - 5)x_6} \\ &= \underline{(0.5)9011} + \underline{(0.5)9165} \\ &= 9088 \end{aligned}$$

Q(0.18)

$$\begin{aligned} i' &= np + .5 \\ &= 10 \cdot 0.18 + 0.5 = \underline{\underline{2.3}} \end{aligned}$$

$$\begin{aligned} Q(0.18) &= (\lceil i' \rceil - i')x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor)x_{\lceil i' \rceil} \\ &= (\lceil 2.3 \rceil - 2.3)x_{\lfloor 2.3 \rfloor} + (2.3 - \lfloor 2.3 \rfloor)x_{\lceil 2.3 \rceil} \\ &= \underline{(3 - 2.3)}x_2 + \underline{(2.3 - 2)}x_3 \\ &= \underline{(0.7)8527 + (0.3)8572} \\ &= 8540.5 \end{aligned}$$

Q(0.94)

$$
\begin{aligned}
i' &= np + .5 \\
&= 10 \cdot 0.94 + 0.5 = 9.9
\end{aligned}
$$

$$
\begin{aligned}
Q(0.94) &= (\lceil i' \rceil - i')x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor)x_{\lceil i' \rceil} \\
&= (\lceil 9.9 \rceil - 9.9)x_{\lfloor 9.9 \rfloor} + (9.9 - \lfloor 9.9 \rfloor)x_{\lceil 9.9 \rceil} \\
&= (\underline{10 - 9.9})x_9 + (\underline{9.9 - 9})x_{10} \\
&= (0.1)9614 + (0.9)10688 \\
&= \underline{10580.6}
\end{aligned}
$$

More on quantiles

- Special quantiles:
  - **Minimum**: $Q\left(\frac{1-.5}{n}\right)$ $= X_1$
  - **Lower Quartile**: $Q(0.25)$
  - **Median**: $Q(0.5)$
  - **Upper Quartile**: $Q(0.75)$
  - **Maximum**: $Q\left(\frac{n-.5}{n}\right)$ $= X_n$

can be used to measure spread

- **Interquartile Range (IQR)**: $Q(0.75) - Q(0.25)$
  - Most points should be below $Q(0.75) + 1.5 \cdot \text{IQR}$ and above $Q(0.25) - 1.5 \cdot \text{IQR}$.
  - potential **Outlier**: a point above $Q(0.75) + 1.5 \cdot \text{IQR}$ or below $Q(0.25) - 1.5 \cdot \text{IQR}$.