Descriptive Statistics: Part 1/2 (Ch 3)

Will Landau

Iowa State University

January 24, 2013

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.
- Descriptive statistics seeks to find the following features of datasets:
  - Center: the point that the data are closest to on average
  - Spread: how wide the data look, how varied the points are
  - Shape (more on that when we get to plots)
  - Outliers: points that lie way beyond the rest of the data.

Outline

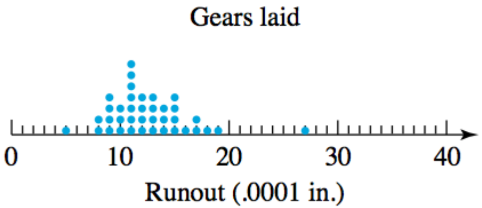
What is descriptive statistics?

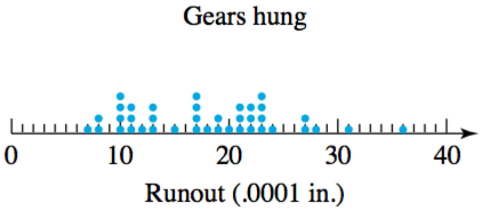
Graphical and Tabular Displays

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

Quantiles

Gear data


Gears laid
0
10
20
30
40
Runout (.0001 in.)


Gears hung
0
10
20
30
40
Runout (.0001 in.)

New example: bullet data

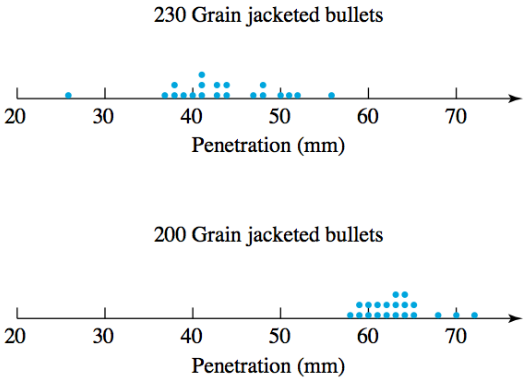
Portraying Bullet Penetration Depths

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

Gear 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

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 back 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

Frequency Table for Laid Gear Thrust Face Runouts

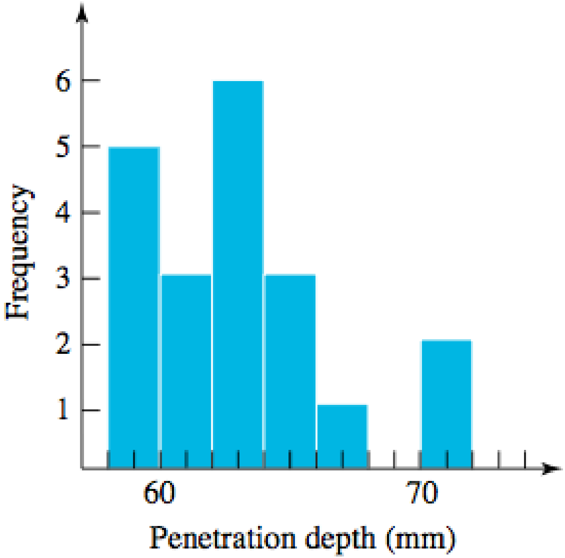
Runout (.0001 in.)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
5–8	|||	3	.079	.079
9–12	𝍸 𝍸 𝍸 |||	18	.474	.553
13–16	𝍸 𝍸 ||	12	.316	.868
17–20	||||	4	.105	.974
21–24		0	0	.974
25–28	|	1	.026	1.000
		38	1.000	

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	~~IIII~~	5	.25	.25
60.00–61.99	III	3	.15	.40
62.00–63.99	~~IIII~~ I	6	.30	.70
64.00–65.99	III	3	.15	.85
66.00–67.99	I	1	.05	.90
68.00–69.99		0	0	.90
70.00–71.99	II	2	.10	1.00
		20	1.00	

Histogram: bullet data, 200 grain



Histogram guidelines

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

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

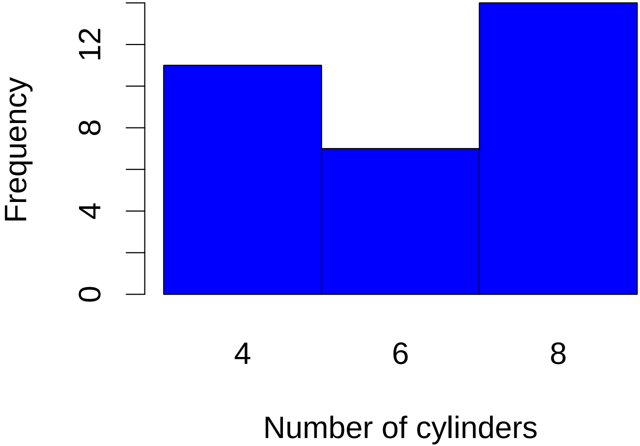
Discrete data: cars

```
## % latex table generated in R 2.15.1 by xtable 1.7-0 package
## % Mon Feb 25 23:40:38 2013
## \begin{table}[ht]
## \begin{center}
## \begin{tabular}{rll}
##   \hline
##  & mpg & cyl \\
##   \hline
## 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 \\
##   ... & ... & ... \\
##   \hline
```

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

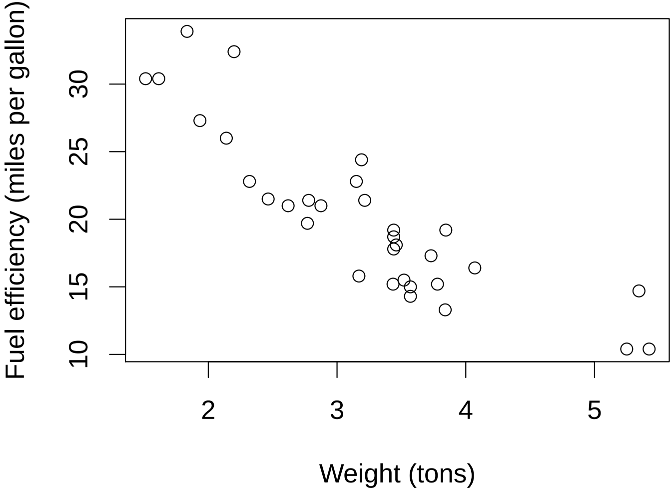
Bar plot (not a histogram)



Bivariate data: cars

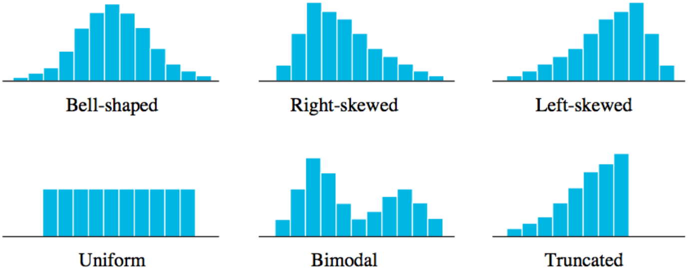
```
## % latex table generated in R 2.15.1 by xtable 1.7-0 packag
## % Mon Feb 25 23:40:38 2013
## \begin{table}[ht]
## \begin{center}
## \begin{tabular}{rll}
##   \hline
##  & mpg & wt \\
##   \hline
## Mazda RX4 & 21 & 2.62 \\
##   Mazda RX4 Wag & 21 & 2.875 \\
##   Datsun 710 & 22.8 & 2.32 \\
##   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 \\
```

Scatterplot: mpg vs wt, cats data



Distributional shapes

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



Outline

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."

Calculating quantiles of finite datasets: setup

- Given:
  - $x_1, \dots x_n$, an ordered list of numbers. This is the dataset.
  - 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**.

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 $Q(p) = (\lceil i' \rceil - i') x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor) x_{\lceil i' \rceil}$

Example: breaking strength (g) of towels

```
## % latex table generated in R 2.15.1 by xtable 1.7-0 packag
## % Mon Feb 25 23:40:38 2013
## \begin{table}[ht]
## \begin{center}
## \begin{tabular}{cc}
##   \hline
## test & strength \\
##   \hline
##   1 & 8577 \\
##     2 & 9471 \\
##     3 & 9011 \\
##     4 & 7583 \\
##     5 & 8572 \\
##     6 & 10688 \\
##     7 & 9614 \\
##     8 & 9614 \\
##     9 & 8527 \\
```

Example: breaking strength (g) of towels

```
## % latex table generated in R 2.15.1 by xtable 1.7-0 packag
## % Mon Feb 25 23:40:38 2013
## \begin{table}[ht]
## \begin{center}
## \begin{tabular}{ccc}
##  test & $\frac{i - .5}{10}$ & $i$'th smallest data point,
##   \hline
##   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 \\
```

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

```
## % latex table generated in R 2.15.1 by xtable 1.7-0 package
## % Mon Feb 25 23:40:38 2013
## \begin{table}[ht]
## \begin{center}
## \begin{tabular}{ccc}
##  test & $\frac{i - .5}{10}$ & $i$'th smallest data point, $x_i =
##   \hline
##   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 \\
##   \end{tabular}
## \end{center}
```

Q(0.5)

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

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

Q(0.18)

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

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

Q(0.94)

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

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

More on quantiles

- Special quantiles:
  - **Minimum**: $Q\left(\frac{1-.5}{n}\right)$
  - **Lower Quartile**: $Q(0.25)$
  - **Median**: $Q(0.5)$
  - **Upper Quartile**: $Q(0.75)$
  - **Maximum**: $Q\left(\frac{n-.5}{n}\right)$
- **Interquartile Range (IQR)**: $Q(0.75) - Q(0.25)$
  - Most points should be below $Q(0.75) + 1.5 \cdot \text{IQR}$ and above $Q(0.25) - 1.5 \cdot \text{IQR}$.
  - **Outlier**: a point above $Q(0.75) + 1.5 \cdot \text{IQR}$ or below $Q(0.25) - 1.5 \cdot \text{IQR}$.