Descriptive Statistics: Part 2/2 (Ch 3)

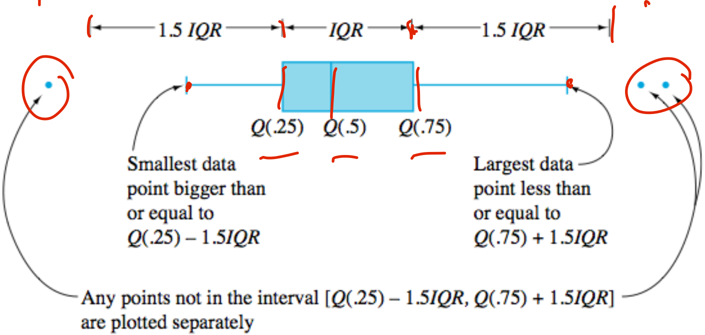
Yifan Zhu

Iowa State University

Outline

Generic Boxplot

$[Q(.25) - 1.5\,IQR \;,\; Q(.75) + 1.5\,IQR]$



Example: bullet data

$i = 20 \times 0.25 + 0.5 = 5.5$

Quantiles of the Bullet Penetration Depth Distributions

i	p_i $\frac{i-.5}{20}$	ith Smallest 230 Grain Data Point $= Q(\frac{i-.5}{20})$	ith Smallest 200 Grain Data Point $= Q(\frac{i-.5}{20})$
1	.025	27.75	58.00
2	.075	37.35	58.65
3	.125	38.35	59.10
4	.175	38.35	59.50
5	.225	38.75	59.80
6	.275	39.75	60.70
7	.325	40.50	61.30
8	.375	41.00	61.50
9	.425	41.15	62.30
10	.475	42.55	62.65
11	.525	42.90	62.95
12	.575	43.60	63.30
13	.625	43.85	63.55
14	.675	47.30	63.80
15	.725	47.90	64.05
16	.775	48.15	64.65
17	.825	49.85	65.00
18	.875	51.25	67.75
19	.925	51.60	70.40
20	.975	56.00	71.70

Example: bullet data (230-grain bullets)

Step 1:

$$\boxed{Q(.25)} = .5Q(.225) + .5Q(.275) = .5(38.75) + .5(39.75) = 39.25 \text{ mm}$$

$$\boxed{Q(.5)} = .5Q(.475) + .5Q(.525) = .5(42.55) + .5(42.90) = 42.725 \text{ mm}$$

$$\boxed{Q(.75)} = .5Q(.725) + .5Q(.775) = .5(47.90) + .5(48.15) = 48.025 \text{ mm}$$

So Step 2:

$$IQR = 48.025 - 39.25 = 8.775 \text{ mm}$$

$$1.5IQR = 13.163 \text{ mm}$$

$$\boxed{Q(.75) + 1.5IQR} = 61.188 \text{ mm}$$

$$\boxed{Q(.25) - 1.5IQR} = 26.087 \text{ mm}$$

Descriptive Statistics: Part 2/2 (Ch 3)

Yifan Zhu

Boxplots

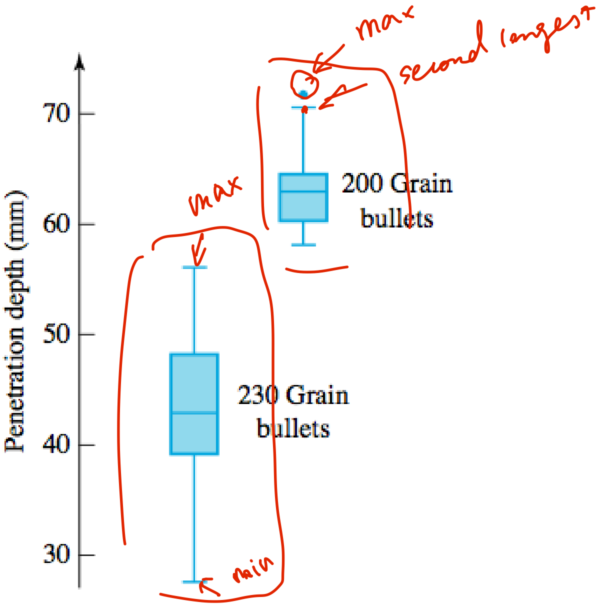
Quantile-Quantile (QQ) Plots

Theoretical Quantile-Quantile Plots

Numerical Summaries

Parameters

Example: bullet data



Descriptive Statistics: Part 2/2 (Ch 3)

Yifan Zhu

Boxplots

Quantile-Quantile (QQ) Plots

Theoretical Quantile-Quantile Plots

Numerical Summaries

Parameters

Outline

quantiles

- **Quantile-quantile (QQ) plot**: a scatterplot of the sorted values of one dataset on the sorted values of another dataset.
  - This plot is used to tell if the distributional shapes of the datasets are the same or different.
    - If the points in the plot lie in a straight line, the distributional shapes are the same.
    - Otherwise, the shapes are different.
  - The datasets must be univariate, numerical, and of the same size.

- **Quantile-quantile (QQ) plot**: a scatterplot of the sorted values of one dataset on the sorted values of another dataset.
  - This plot is used to tell if the distributional shapes of the datasets are the same or different.
    - If the points in the plot lie in a straight line, the distributional shapes are the same.
    - Otherwise, the shapes are different.
  - The datasets must be univariate, numerical, and of the same size.

- **Quantile-quantile (QQ) plot**: a scatterplot of the sorted values of one dataset on the sorted values of another dataset.
  - This plot is used to tell if the distributional shapes of the datasets are the same or different.
    - If the points in the plot lie in a straight line, the distributional shapes are the same.
    - Otherwise, the shapes are different.
  - The datasets must be univariate, numerical, and of the same size.

- **Quantile-quantile (QQ) plot**: a scatterplot of the sorted values of one dataset on the sorted values of another dataset.
  - This plot is used to tell if the distributional shapes of the datasets are the same or different.
    - If the points in the plot lie in a straight line, the distributional shapes are the same.
    - Otherwise, the shapes are different.
  - The datasets must be univariate, numerical, and of the same size.

- **Quantile-quantile (QQ) plot**: a scatterplot of the sorted values of one dataset on the sorted values of another dataset.
  - This plot is used to tell if the distributional shapes of the datasets are the same or different.
    - If the points in the plot lie in a straight line, the distributional shapes are the same.
    - Otherwise, the shapes are different.
  - The datasets must be univariate, numerical, and of the same size.

$x_1, \ldots, x_n$ dataset 1
$y_1, \ldots, y_n$ dataset 2

(x_i, y_i)

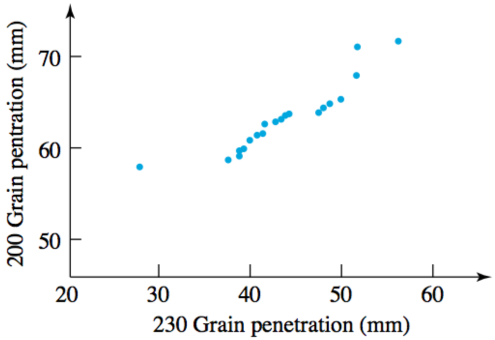
Example: bullet data

Quantiles of the Bullet Penetration Depth Distributions

i	$\frac{i-.5}{20}$	ith Smallest 230 Grain Data Point $= Q(\frac{i-.5}{20})$	ith Smallest 200 Grain Data Point $= Q(\frac{i-.5}{20})$
1	.025	27.75	58.00
2	.075	37.35	58.65
3	.125	38.35	59.10
4	.175	38.35	59.50
5	.225	38.75	59.80
6	.275	39.75	60.70
7	.325	40.50	61.30
8	.375	41.00	61.50
9	.425	41.15	62.30
10	.475	42.55	62.65
11	.525	42.90	62.95
12	.575	43.60	63.30
13	.625	43.85	63.55
14	.675	47.30	63.80
15	.725	47.90	64.05
16	.775	48.15	64.65
17	.825	49.85	65.00
18	.875	51.25	67.75
19	.925	51.60	70.40
20	.975	56.00	71.70

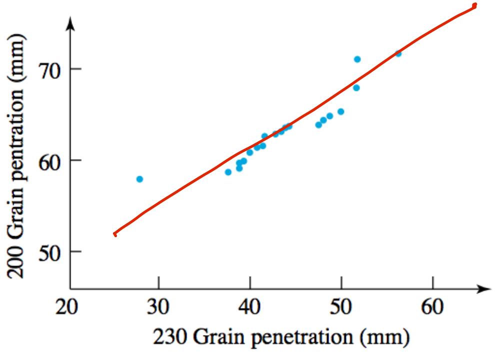
Example: bullet data

- I can make a QQ plot of the bullet data by plotting the sorted 200-grain depths against the sorted 230-grain depths.
- The points lie in approximately a straight line, so the 200-grain depths are similarly shaped in distribution to the 230-grain depths.



Example: bullet data

- I can make a QQ plot of the bullet data by plotting the sorted 200-grain depths against the sorted 230-grain depths.
- The points lie in approximately a straight line, so the 200-grain depths are similarly shaped in distribution to the 230-grain depths.



Outline

Boxplots

Quantile-Quantile (QQ) Plots

Theoretical Quantile-Quantile Plots

Numerical Summaries

Parameters

Theoretical quantile-quantile (QQ) plots

- **Theoretical quantile-quantile (QQ) plot**: a scatterplot with:
  - The sorted values $x_1, x_2, \ldots x_n$ of some real data set on the x axis.
  - $Q(\frac{1-.5}{n}), Q(\frac{2-.5}{n}), \ldots, Q(\frac{n-.5}{n})$ on the y axis.
    - Q is some **theoretical quantile function**: the quantile function we would *expect* from a dataset if that dataset had a certain shape.
- Example theoretical quantile functions:
  - "Standard" bell-shaped (or normally-distributed) data should have:

$$Q(p) \approx 4.9(p^{0.14} - (1-p)^{0.14})$$

  - "Exponentially distributed" data (a kind of highly right-skewed data) should have:

$$Q(p) \approx -\lambda^{-1} \log(1-p)$$

where λ is some constant.

Theoretical quantile-quantile (QQ) plots

- **Theoretical quantile-quantile (QQ) plot**: a scatterplot with:
  - The sorted values $x_1, x_2, \ldots x_n$ of some real data set on the x axis.
  - $Q(\frac{1-.5}{n}), Q(\frac{2-.5}{n}), \ldots, Q(\frac{n-.5}{n})$ on the y axis.
    - Q is some **theoretical quantile function**: the quantile function we would *expect* from a dataset if that dataset had a certain shape.
- Example theoretical quantile functions:
  - "Standard" bell-shaped (or normally-distributed) data should have:

$$Q(p) \approx 4.9(p^{0.14} - (1-p)^{0.14})$$

  - "Exponentially distributed" data (a kind of highly right-skewed data) should have:

$$Q(p) \approx -\lambda^{-1} \log(1-p)$$

where λ is some constant.

Theoretical quantile-quantile (QQ) plots

- **Theoretical quantile-quantile (QQ) plot**: a scatterplot with:
  - The sorted values $x_1, x_2, \ldots x_n$ of some real data set on the x axis.
  - $Q(\frac{1-.5}{n}), Q(\frac{2-.5}{n}), \ldots, Q(\frac{n-.5}{n})$ on the y axis.
    - Q is some **theoretical quantile function**: the quantile function we would *expect* from a dataset if that dataset had a certain shape.
- Example theoretical quantile functions:
  - "Standard" bell-shaped (or normally-distributed) data should have:

$$Q(p) \approx 4.9(p^{0.14} - (1-p)^{0.14})$$

  - "Exponentially distributed" data (a kind of highly right-skewed data) should have:

$$Q(p) \approx -\lambda^{-1} \log(1-p)$$

    where λ is some constant.

Theoretical quantile-quantile (QQ) plots

- **Theoretical quantile-quantile (QQ) plot**: a scatterplot with:
  - The sorted values $x_1, x_2, \ldots x_n$ of some real data set on the x axis.
  - $Q(\frac{1-.5}{n}), Q(\frac{2-.5}{n}), \ldots, Q(\frac{n-.5}{n})$ on the y axis.
    - Q is some **theoretical quantile function**: the quantile function we would *expect* from a dataset if that dataset had a certain shape.
- Example theoretical quantile functions:
  - "Standard" bell-shaped (or normally-distributed) data should have:

$$Q(p) \approx 4.9(p^{0.14} - (1-p)^{0.14})$$

  - "Exponentially distributed" data (a kind of highly right-skewed data) should have:

$$Q(p) \approx -\lambda^{-1} \log(1-p)$$

where λ is some constant.

Theoretical quantile-quantile (QQ) plots

- **Theoretical quantile-quantile (QQ) plot**: a scatterplot with:
  - The sorted values $x_1, x_2, \dots x_n$ of some real data set on the x axis.
  - $Q(\frac{1-.5}{n}), Q(\frac{2-.5}{n}), \dots, Q(\frac{n-.5}{n})$ on the y axis.
    - Q is some **theoretical quantile function**: the quantile function we would *expect* from a dataset if that dataset had a certain shape.
- Example theoretical quantile functions:
  - "Standard" bell-shaped (or normally-distributed) data should have:

$$Q(p) \approx 4.9(p^{0.14} - (1-p)^{0.14})$$

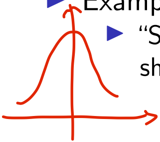
  - "Exponentially distributed" data (a kind of highly right-skewed data) should have:

$$Q(p) \approx -\lambda^{-1} \log(1-p)$$

where λ is some constant.

Theoretical quantile-quantile (QQ) plots

- **Theoretical quantile-quantile (QQ) plot**: a scatterplot with:
  - The sorted values $x_1, x_2, \ldots x_n$ of some real data set on the x axis.
  - $Q(\frac{1-.5}{n}), Q(\frac{2-.5}{n}), \ldots, Q(\frac{n-.5}{n})$ on the y axis.
    - Q is some **theoretical quantile function**: the quantile function we would *expect* from a dataset if that dataset had a certain shape.
- Example theoretical quantile functions:



  - "Standard" bell-shaped (or normally-distributed) data should have:

$$Q(p) \approx 4.9(p^{0.14} - (1-p)^{0.14})$$

N(0,1)

  - "Exponentially distributed" data (a kind of highly right-skewed data) should have:

$$Q(p) \approx -\lambda^{-1} \log(1-p)$$

where λ is some constant.

Theoretical quantile-quantile (QQ) plots

- **Theoretical quantile-quantile (QQ) plot**: a scatterplot with:
  - The sorted values $x_1, x_2, \ldots x_n$ of some real data set on the x axis.
  - $Q(\frac{1-.5}{n}), Q(\frac{2-.5}{n}), \ldots, Q(\frac{n-.5}{n})$ on the y axis.
    - Q is some **theoretical quantile function**: the quantile function we would *expect* from a dataset if that dataset had a certain shape.
- Example theoretical quantile functions:
  - "Standard" bell-shaped (or normally-distributed) data should have:

$$Q(p) \approx 4.9(p^{0.14} - (1-p)^{0.14})$$

  - "Exponentially distributed" data (a kind of highly right-skewed data) should have:

Exp(λ)

$$Q(p) \approx -\lambda^{-1} \log(1-p)$$

where λ is some constant.

λ = 1

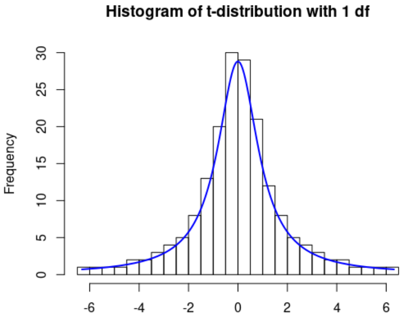
Normal quantile-quantile (QQ) Plots

- **Normal quantile-quantile (QQ) plot**: a theoretical QQ plot where the quantile function, Q, is the quantile function for "standard" bell-shaped (normally-distributed) data. $N(0,1)$
- If the points in a normal QQ plot are in a straight line, the dataset in question is normally-distributed and bell-shaped. Otherwise, the data is not normally-distributed (but still could be bell-shaped).

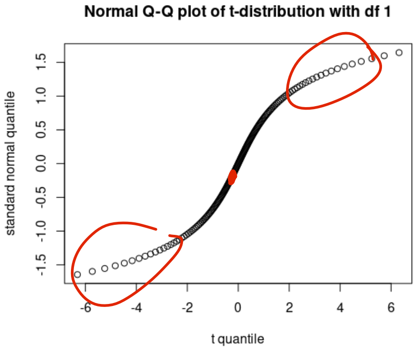
Normal quantile-quantile (QQ) Plots

- **Normal quantile-quantile (QQ) plot**: a theoretical QQ plot where the quantile function, Q, is the quantile function for "standard" bell-shaped (normally-distributed) data.
- If the points in a normal QQ plot are in a straight line, the dataset in question is normally-distributed and bell-shaped. Otherwise, the data is not normally-distributed (but still could be bell-shaped).

bell-shaped.



Histogram of t-distribution with 1 df
Frequency
0
5
10
15
20
25
30
-6
-4
-2
0
2
4
6



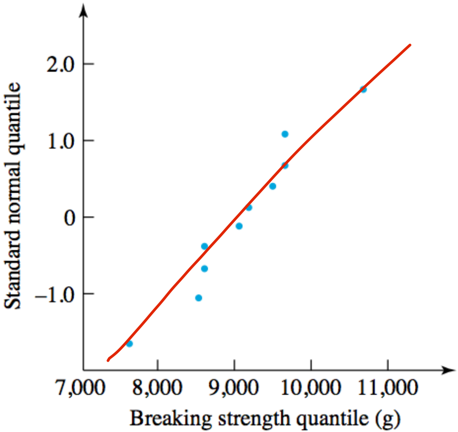
Normal Q-Q plot of t-distribution with df 1
standard normal quantile
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
-6
-4
-2
0
2
4
6
t quantile

Example: towel breaking strength data

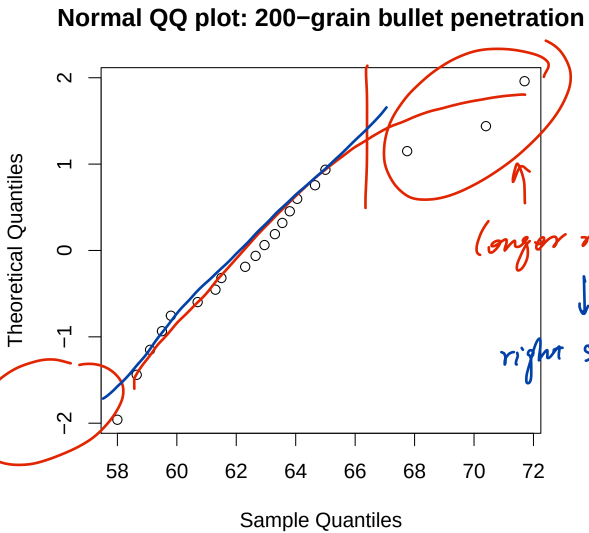
Breaking Strength and Standard Normal Quantiles

i	$\frac{i-.5}{10}$ p_i	$\frac{i-.5}{10}$ Breaking Strength Quantile x_i	$\frac{i-.5}{10}$ Standard Normal Quantile $Q(p_i)$
1	.05	7,583	−1.65
2	.15	8,527	−1.04
3	.25	8,572	−.67
4	.35	8,577	−.39
5	.45	9,011	−.13
6	.55	9,165	.13
7	.65	9,471	.39
8	.75	9,614	.67
9	.85	9,614	1.04
10	.95	10,688	1.65

Example: towel breaking strength data



- The points are roughly straight-line-shaped, so the breaking strength data is roughly bell-shaped and normally-distributed.



longer right tail

↓ ||

right skewed

Observations

- Since the points in the normal QQ plot are not quite arranged in a straight line, the 200-grain penetration depths are not quite normally-distributed. However, the departure from normality is not severe.
- The QQ plot of the bullet data from before revealed that the 200-grain depths had the same distributional shape as the 200-grain bullet depths. Thus, the 230-grain bullet data is not quite normally-distributed either.

Observations

- Since the points in the normal QQ plot are not quite arranged in a straight line, the 200-grain penetration depths are not quite normally-distributed. However, the departure from normality is not severe.
- The QQ plot of the bullet data from before revealed that the 200-grain depths had the same distributional shape as the 200-grain bullet depths. Thus, the 230-grain bullet data is not quite normally-distributed either.

↑ also make QQ plot.

Outline

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean ← Sample mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Numerical summaries

- **Numerical summary (statistic)**
  - A number or list of numbers calculated using the data (and only the data).
  - Numerical summaries highlight important features of the data (shape, center, spread, outliers).
- Examples:
  - Measures of center:
    - Arithmetic mean
    - Median
    - Mode
  - Measures of spread:
    - Sample variance
    - Sample standard deviation
    - Range
    - IQR
  - Measures of shape:
    - All the quantiles together
    - Skew (beyond the scope of the class)
    - Kurtosis (beyond the scope of the class)

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0 + 1 + 1 + 2 + 3 + 5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1 + 2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode = 1.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode = 1.

Measures of center

n = 6 (sample size).

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

Sample mean

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0 + 1 + 1 + 2 + 3 + 5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1 + 2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode = 1.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

$n = 5$, $\lceil 5/2 \rceil$
$= \lceil 2.5 \rceil$
$= 3$
$n = 6$. $n/2 = 3$
$(x_3 + x_4)/2$

Descriptive Statistics: Part 2/2 (Ch 3)

Yifan Zhu

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0 + 1 + 1 + 2 + 3 + 5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1 + 2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of center

x_1	x_2	x_3	x_4	x_5	x_6
0	1	1	2	3	5

- Arithmetic mean:
  - $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Here, $\overline{x} = \frac{1}{6}(0+1+1+2+3+5) = 2$
- Median: $Q(0.5)$.
  - A shortcut to calculating $Q(0.5)$ is:
    - $Q(0.5) = x_{\lceil n/2 \rceil}$ if n is odd
    - $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$ if n is even.
  - Here, $Q(0.5) = (1+2)/2 = 1.5$
- Mode (of a discrete or categorical dataset)
  - the most frequently-occurring value
  - Here, mode $= 1$.

Measures of spread

	x_1	x_2	x_3	x_4	x_5	x_6
x_i	0	1	1	2	3	5
$\frac{i-.5}{n}$	.083	0.25	0.417	0.583	0.75	0.917

- Sample variance
  - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ ← unbiased estimator of variance
  - Here, $s^2 = \frac{1}{6-1}[(0-2)^2 + (1-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (5-2)^2] = 3.2$
- Sample standard deviation
  - $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
  - Here, $s = \sqrt{3.2} = 1.78885438199983$
- Range
  - Range = Maximum - Minimum
  - Here, Range = 5 - 0 = 5
- Interquartile range
  - $\text{IQR} = Q(0.75) - Q(0.25)$
  - Here, $\text{IQR} = 3 - 1 = 2$.

Your turn: sensitivity to outliers

Compare:

	x_1	x_2	x_3	x_4	x_5	x_6
x_i	0	1	1	2	3	5
$\frac{i-.5}{n}$	.083	0.25	0.417	0.583	0.75	0.917

to:

outlier

	y_1	y_2	y_3	y_4	y_5	y_6
x_i	0	1	1	2	3	817263489
$\frac{i-.5}{n}$	.083	0.25	0.417	0.583	0.75	0.917

extreme case of right skewed

which measures of center and spread differ drastically between the x_i's and the y_i's? Which ones are about the same?

Descriptive Statistics: Part 2/2 (Ch 3)

Yifan Zhu

Answers: sensitivity to outliers

Data	x_i	y_i
Mean	2	136210582.666667
Median	1.5	1.5
Mode	1	1
Sample Variance	3.2	111319934694019232
Sample Std. Dev.	1.78885438199983	333646421.671235
Range	5	817263489
IQR	2	2

Sensitivity of numerical summaries

- Numerical summaries sensitive to outliers *and* skewness:
  - Mean Sample mean
  - Sample variance
  - Sample standard deviation
  - Range
- Less sensitive numerical summaries:
  - Median
  - Mode
  - IQR

Outline

Boxplots

Quantile-Quantile (QQ) Plots

Theoretical Quantile-Quantile Plots

Numerical Summaries

Parameters

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

approx

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical characteristic summary of a theoretical distribution or data on an entire *population*. unknown fixed
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean): $N = \infty$, $E(X)$
    - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\bar{x} \approx \mu$. $n \ll N$, $\bar{x} \to \mu$, $n \to \infty$
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{\mathbf{N}} \sum_{i=1}^{N} (x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.

      Var(X)
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{\mathbf{N}} \sum_{i=1}^{N} (x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Descriptive Statistics: Part 2/2 (Ch 3)

Yifan Zhu

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{N}(x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{\mathbf{N}} \sum_{i=1}^{N} (x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{\mathbf{N}} \sum_{i=1}^{N} (x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.

Statistics and parameters

- **Statistic**: numerical summary of data on the *sample*
- **Parameter**: numerical summary of a theoretical distribution or data on an entire *population*.
  - Population mean ("true" mean):
    - $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ if N the *finite* population size.
    - $\overline{x} \approx \mu$.
  - Population variance ("true" variance):
    - $\sigma^2 = \frac{1}{\mathbf{N}} \sum_{i=1}^{N} (x_i - \mu)^2$ if N the *finite* population size.
    - $s^2 \approx \sigma^2$.
  - Population standard deviation ("true" standard deviation):
    - $\sigma = \sqrt{\frac{1}{\mathbf{N}} \sum_{i=1}^{N} (x_i - \mu)^2}$ if N is the *finite* population size.
    - $s \approx \sigma$.