Descriptive statistics

Chrysafis Vogiatzis

Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Lecture 14

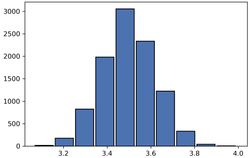
Probability vs. statistics

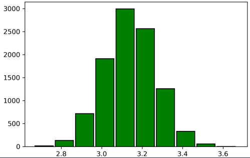
What is probability?

- An estimate of how likely an outcome is.
  "What are my chances of rolling a 6 and a 1?" $\frac{2}{36} = \frac{1}{18}$

What is statistics?

- All the methods involved with collecting, describing, analyzing, interpreting data.
  "Are two dice fair?"





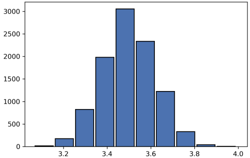
Probability vs. statistics

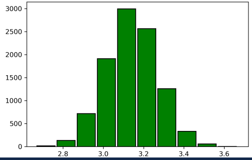
What is probability?

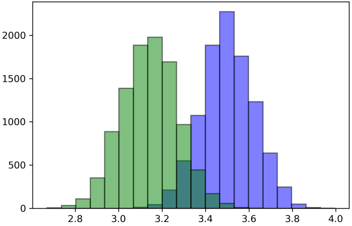
- An estimate of how likely an outcome is.
  “What are my chances of rolling a 6 and a 1?” $\frac{2}{36} = \frac{1}{18}$

What is statistics?

- All the methods involved with collecting, describing, analyzing, interpreting data.
  “Are two dice fair?”







Data science *is* statistics

The use of **statistics** has two facades:

1. *Data*: the presentation of
   - interesting numerical facts.
   - representative numbers, specific to the data.
2. *Information*: or the communication of
   - Knowledge and predictions for a specific aspect.

Statistical methods are categorized in:

1. Descriptive statistics: methods to *describe* and *present* data.
2. Inferential statistics: methods to use observations in a smaller **sample** to *draw conclusions* for the larger **population**.
3. Model building: methods to build models to *predict* future data based on past observations.

Data science *is* statistics

The use of **statistics** has two facades:

1. *Data*: the presentation of
   - interesting numerical facts.
   - representative numbers, specific to the data.
2. *Information*: or the communication of
   - Knowledge and predictions for a specific aspect.

Statistical methods are categorized in:

1. Descriptive statistics: methods to *describe* and *present* data.
2. Inferential statistics: methods to use observations in a smaller **sample** to *draw conclusions* for the larger **population**.
3. Model building: methods to build models to *predict* future data based on past observations.

ILLINOIS

Data science *is* statistics

The use of **statistics** has two facades:

1. *Data*: the presentation of
   - interesting numerical facts.
   - representative numbers, specific to the data.
2. *Information*: or the communication of
   - Knowledge and predictions for a specific aspect.

Statistical methods are categorized in:

1. Descriptive statistics: methods to *describe* and *present* data.
2. Inferential statistics: methods to use observations in a smaller **sample** to *draw conclusions* for the larger **population**.
3. Model building: methods to build models to *predict* future data based on past observations.

Data science *is* statistics

The use of **statistics** has two facades:

1. *Data*: the presentation of
   - interesting numerical facts.
   - representative numbers, specific to the data.
2. *Information*: or the communication of
   - Knowledge and predictions for a specific aspect.

Statistical methods are categorized in:

1. Descriptive statistics: methods to *describe* and *present* data.
2. Inferential statistics: methods to use observations in a smaller **sample** to *draw conclusions* for the larger **population**.
3. Model building: methods to build models to *predict* future data based on past observations.

Data science *is* statistics

The use of **statistics** has two facades:

1. *Data*: the presentation of
   - interesting numerical facts.
   - representative numbers, specific to the data.
2. *Information*: or the communication of
   - Knowledge and predictions for a specific aspect.

Statistical methods are categorized in:

1. Descriptive statistics: methods to *describe* and *present* data.
2. Inferential statistics: methods to use observations in a smaller **sample** to *draw conclusions* for the larger **population**.
3. Model building: methods to build models to *predict* future data based on past observations.

Statistical methods: descriptive statistics

What we will focus on in this lecture and in the worksheet is **descriptive statistics**. More specifically:

1. Numerical summaries of data.
    - sample mean, mode, median.
    - sample variance, standard deviation.
    - percentiles, quartiles, ranges.

this video lecture

2. Graphical displays of data.
    - Dot diagrams.
    - Histograms.
    - Stem-and-leaf diagrams.
    - Box plots.
    - Scatter diagrams.
    - Time series plots.
    - Q-Q plots.

in-class worksheet

ILLINOIS

Populations vs. samples

A **population** implies *all of the observations* with which we are concerned:

- The height of every person in the world.
- The SAT scores of every person that took the SATs in 2018.
- The delays in all of the flights of a specific company.

A **sample** implies *subset of the observations* selected from a population:

- The height of every person in Chicago.
- The SAT scores of 20 randomly selected people that took the SATs in 2018.
- The delays in all of the flights of a specific company at the airport of Atlanta.

In most cases, our data is just a sample. We need to remember this and consider it in our analyses.

Populations vs. samples

A **population** implies *all of the observations* with which we are concerned:

- The height of every person in the world.
- The SAT scores of every person that took the SATs in 2018.
- The delays in all of the flights of a specific company.

A **sample** implies *subset of the observations* selected from a population:

- The height of every person in Chicago.
- The SAT scores of 20 randomly selected people that took the SATs in 2018.
- The delays in all of the flights of a specific company at the airport of Atlanta.

In most cases, our data is just a sample. We need to remember this and consider it in our analyses.

Populations vs. samples

A **population** implies *all of the observations* with which we are concerned:

- The height of every person in the world.
- The SAT scores of every person that took the SATs in 2018.
- The delays in all of the flights of a specific company.

A **sample** implies *subset of the observations* selected from a population:

- The height of every person in Chicago.
- The SAT scores of 20 randomly selected people that took the SATs in 2018.
- The delays in all of the flights of a specific company at the airport of Atlanta.

In most cases, our data is just a sample. We need to remember this and consider it in our analyses.

Data.. summarized

Presenting all of the data (raw or processed) is rarely ever an effective way to communicate its pattern. Instead, we present measures to reveal two important characteristics.

1. The center of the data.
   - Average/mean.
   - Median (the middle value of the ordered data).
   - Mode (the most frequent value or values).
2. The variation in the data.
   - Variance and standard deviation.
   - Range.
   - Interquartile range.

We'll see all of these in the subsequent slides.

Data.. summarized

Presenting all of the data (raw or processed) is rarely ever an effective way to communicate its pattern. Instead, we present measures to reveal two important characteristics.

1. The center of the data.
   - Average/mean.
   - Median (the middle value of the ordered data).
   - Mode (the most frequent value or values).
2. The variation in the data.
   - Variance and standard deviation.
   - Range.
   - Interquartile range.

We'll see all of these in the subsequent slides.

Data.. summarized

Presenting all of the data (raw or processed) is rarely ever an effective way to communicate its pattern. Instead, we present measures to reveal two important characteristics.

1. The center of the data.
   - Average/mean.
   - Median (the middle value of the ordered data).
   - Mode (the most frequent value or values).
2. The variation in the data.
   - Variance and standard deviation.
   - Range.
   - Interquartile range.

We'll see all of these in the subsequent slides.

ILLINOIS

Data.. summarized

Presenting all of the data (raw or processed) is rarely ever an effective way to communicate its pattern. Instead, we present measures to reveal two important characteristics.

1. The center of the data.
   - Average/mean.
   - Median (the middle value of the ordered data).
   - Mode (the most frequent value or values).
2. The variation in the data.
   - Variance and standard deviation.
   - Range.
   - Interquartile range.

We'll see all of these in the subsequent slides.

Data.. summarized

Presenting all of the data (raw or processed) is rarely ever an effective way to communicate its pattern. Instead, we present measures to reveal two important characteristics.

1. The center of the data.
   - Average/mean.
   - Median (the middle value of the ordered data).
   - Mode (the most frequent value or values).
2. The variation in the data.
   - Variance and standard deviation.
   - Range.
   - Interquartile range.

We'll see all of these in the subsequent slides.

Sample mode

Definition

Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample mode** is the value(s) x_i that appears most times.

Example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the sample mode is 60 as it appears twice.

Sample average/mean

Definition

Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample mean** is calculated as

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the sample mean is

$$\frac{1}{5}(60 + 67 + 72 + 63 + 60) = 64.4.$$

Population means

Definition

When a population is finite and has N observations $x_1, x_2, \ldots, x_N$, then the **population mean** is calculated as

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

When a population is infinite and the observations are represented by a continuous random variable with pdf $f(x)$, then the population mean is calculated as

$$\mu = \int_{x} x f(x) dx.$$

Usually, the actual population mean is unknown.

Sample variance

Definition

Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample variance** is calculated as

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} =$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}.$$

The sample standard deviation is denoted by $s = \sqrt{s^2}$. Furthermore, $n-1$ is also called the **degrees of freedom** of the sample.

Example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60 with $\overline{x} = 64.4$. Then, the sample variance is

$$\frac{1}{4}\left(4.4^2 + 2.7^2 + 7.8^2 + 1.4^2 + 4.4^2\right) = \frac{108.81}{4} = 27.2025.$$

Sample variance

Definition

Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample variance** is calculated as

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} =$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}.$$

The sample standard deviation is denoted by $s = \sqrt{s^2}$. Furthermore, $n-1$ is also called the **degrees of freedom** of the sample.

Example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60 with $\overline{x} = 64.4$. Then, the sample variance is

$$\frac{1}{4}\left(4.4^2 + 2.7^2 + 7.8^2 + 1.4^2 + 4.4^2\right) = \frac{108.81}{4} = 27.2025.$$

Population variance

Definition

When a population is finite and has N observations $x_1, x_2, \ldots, x_N$ with mean μ, then the **population variance** is calculated as

$$\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_N-\mu)^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 .$$

When a population is infinite and the observations are represented by a continuous random variable with pdf $f(x)$ with mean μ, then the population variance is calculated as

$$\sigma^2 = \int_x (x-\mu)^2 f(x) dx.$$

Usually, the actual population variance is unknown.

- Population variance: calculated by dividing by N.
- Sample variance: calculated by dividing by $n-1$.

ILLINOIS

Percentiles

We refer to the value of $p\%$ percentile as the number below which we find approximately $p\%$ of the data.

Assume we sorted the data in increasing order: the $p\%$ percentile value can be found by finding the $(n+1)\,p/100$-st value.

If the calculation of $(n+1)p/100$ is fractional (i.e., the rank falls between two values), then we interpolate.

Example

Assume the heights of 9 people are 62, 64, 67, 58, 70, 61, 67, 65, 64. What is the 30% *and the* 67% *percentile?*

Answer: The ordered heights are 58, 61, 62, 64, 64, 65, 67, 67, 70.

- **30% percentile**: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10\cdot 30}{100} = 3$. The 3rd value is 62.
- **67% percentile**: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10\cdot 67}{100} = 6.7$. The 6th value is 65 and the 7th is 67: interpolating, we get: $0.3\cdot 65 + 0.7\cdot 67 = 66.4$.

Percentiles

We refer to the value of $p\%$ percentile as the number below which we find approximately $p\%$ of the data.

Assume we sorted the data in increasing order: the $p\%$ percentile value can be found by finding the $(n+1)\,p/100$-st value.

If the calculation of $(n+1)p/100$ is fractional (i.e., the rank falls between two values), then we interpolate.

Example

Assume the heights of 9 people are 62, 64, 67, 58, 70, 61, 67, 65, 64. What is the 30% and the 67% percentile?

Answer: The ordered heights are 58, 61, 62, 64, 64, 65, 67, 67, 70.

- **30% percentile**: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 30}{100} = 3$. The 3rd value is 62.
- **67% percentile**: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 67}{100} = 6.7$. The 6th value is 65 and the 7th is 67: interpolating, we get: $0.3 \cdot 65 + 0.7 \cdot 67 = 66.4$.

Percentiles

We refer to the value of $p\%$ percentile as the number below which we find approximately $p\%$ of the data.

Assume we sorted the data in increasing order: the $p\%$ percentile value can be found by finding the $(n+1)\,p/100$-st value.

If the calculation of $(n+1)p/100$ is fractional (i.e., the rank falls between two values), then we interpolate.

Example

Assume the heights of 9 people are 62, 64, 67, 58, 70, 61, 67, 65, 64. What is the 30% *and the* 67% *percentile?*

Answer: The ordered heights are 58, 61, 62, 64, 64, 65, 67, 67, 70.

30% percentile: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 30}{100} = 3$. The 3rd value is 62.

67% percentile: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 67}{100} = 6.7$. The 6th value is 65 and the 7th is 67: interpolating, we get: $0.3 \cdot 65 + 0.7 \cdot 67 = 66.4$.

ILLINOIS

Quartiles

A special percentile for presenting purposes is called a quartile. There are three quartiles: Q1, Q2, Q3.

- Q1: Splits the lower 25% from the rest of the data.
- Q2: Splits the lower 50% from the rest of the data.
- Q3: Splits the lower 75% from the rest of the data.

Q2 is also called the **median**.

Example

Earlier, we got the ordered 9 heights to be 58, 61, 62, 64, 64, 65, 67, 67, 70.

Answer:

Q1: $\frac{(n+1)p}{100} = \frac{10 \cdot 25}{100} = 2.5$. So $Q1 = 61.5$.

Q2: $\frac{(n+1)p}{100} = \frac{10 \cdot 50}{100} = 5 \implies Q2 = 64$.

Q3: $\frac{(n+1)p}{100} = \frac{10 \cdot 75}{100} = 7.5 \implies Q3 = 67$.

Quartiles

A special percentile for presenting purposes is called a quartile. There are three quartiles: Q1, Q2, Q3.

- Q1: Splits the lower 25% from the rest of the data.
- Q2: Splits the lower 50% from the rest of the data.
- Q3: Splits the lower 75% from the rest of the data.

Q2 is also called the **median**.

Example

Earlier, we got the ordered 9 heights to be 58, 61, 62, 64, 64, 65, 67, 67, 70.

Answer:

Q1: $\frac{(n+1)p}{100} = \frac{10 \cdot 25}{100} = 2.5$. So $Q1 = 61.5$.

Q2: $\frac{(n+1)p}{100} = \frac{10 \cdot 50}{100} = 5 \implies Q2 = 64$.

Q3: $\frac{(n+1)p}{100} = \frac{10 \cdot 75}{100} = 7.5 \implies Q3 = 67$.

Quartiles

A special percentile for presenting purposes is called a quartile.
There are three quartiles: Q1, Q2, Q3.

- Q1: Splits the lower 25% from the rest of the data.
- Q2: Splits the lower 50% from the rest of the data.
- Q3: Splits the lower 75% from the rest of the data.

Q2 is also called the **median**.

Example

Earlier, we got the ordered 9 heights to be 58, 61, 62, 64, 64, 65, 67, 67, 70.

Answer:

Q1: $\frac{(n+1)p}{100} = \frac{10 \cdot 25}{100} = 2.5$. So $Q1 = 61.5$.

Q2: $\frac{(n+1)p}{100} = \frac{10 \cdot 50}{100} = 5 \implies Q2 = 64$.

Q3: $\frac{(n+1)p}{100} = \frac{10 \cdot 75}{100} = 7.5 \implies Q3 = 67$.

ILLINOIS

Ranges and outliers

Range:

- The range is simply the difference of the maximum and the minimum value: $R = \max\{x_i\} - \min\{x_i\}$.
- The range of a population will always be greater than or equal to the range of a sample.

Interquartile range:

- Calculated by $IQR = Q3 - Q1$.
- Essentially provides the range of the "middle" part of our data.

Outliers:

- An outlier is a value that affects the range of our data but leaves the "middle" part unaffected.
- A data point is considered an outlier if it lies outside $[Q1 - 1.5IQR, Q3 + 1.5IQR]$.

Ranges and outliers

Range:

- The range is simply the difference of the maximum and the minimum value: $R = \max\{x_i\} - \min\{x_i\}$.
- The range of a population will always be greater than or equal to the range of a sample.

Interquartile range:

- Calculated by $IQR = Q3 - Q1$.
- Essentially provides the range of the "middle" part of our data.

Outliers:

- An outlier is a value that affects the range of our data but leaves the "middle" part unaffected.
- A data point is considered an outlier if it lies outside $[Q1 - 1.5IQR, Q3 + 1.5IQR]$.

Ranges and outliers

Range:

- The range is simply the difference of the maximum and the minimum value: $R = \max\{x_i\} - \min\{x_i\}$.
- The range of a population will always be greater than or equal to the range of a sample.

Interquartile range:

- Calculated by $IQR = Q3 - Q1$.
- Essentially provides the range of the "middle" part of our data.

Outliers:

- An outlier is a value that affects the range of our data but leaves the "middle" part unaffected.
- A data point is considered an outlier if it lies outside $[Q1 - 1.5IQR, Q3 + 1.5IQR]$.

ILLINOIS

Summary statistics: a quick review

Given a random sample of size n containing observations $x_1, x_2, \ldots, x_n$, then:

- Sample mode: the most frequent value in the sample.
- Sample mean: $\overline{x} = \sum_{i=1}^{n} x_i / n$.
- Sample variance: $s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 / n - 1$.
- Sample degrees of freedom: $n - 1$.
- Sample range: $R = \max\{x_i\} - \min\{x_i\}$.
- Quartiles: Q1, Q2 (also called the median), Q3.
- Interquartile range: $IQR = Q3 - Q1$

Where do we go from here?

- Well, providing summary statistics is *great*.
- But, many of us better understand relationships in pictorial form...

Summary statistics: a quick review

Given a random sample of size n containing observations $x_1, x_2, \ldots, x_n$, then:

- Sample mode: the most frequent value in the sample.
- Sample mean: $\overline{x} = \sum_{i=1}^{n} x_i/n$.
- Sample variance: $s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2/n - 1$.
- Sample degrees of freedom: $n - 1$.
- Sample range: $R = \max\{x_i\} - \min\{x_i\}$.
- Quartiles: Q1, Q2 (also called the median), Q3.
- Interquartile range: $IQR = Q3 - Q1$

Where do we go from here?

- Well, providing summary statistics is *great*.
- But, many of us better understand relationships in pictorial form...