American University of Beirut

STAT 210: Elementary Statistics for Sciences

2022–2023 Fall

Siamak Taati

Chapter 7

Sampling distributions

Part 1

Distribution of sample statistics

Samples in the language of probabilities

We earlier talked about samples as typical sources of data sets. We use samples as statistical evidence to learn about populations.

The sample statistics (e.g., sample mean, sample variance, ...) vary from sample to sample, and typically deviate from the corresponding population values (e.g., population mean, population variance, ...).

If the sample is taken at random, then the sample statistics will also be random. Armed with the language of probabilities, we can now view sample statistics as random variables and wonder about their distributions.

Samples in the language of probabilities

Random samples

Recall the two types of random samples:

- With replacement
- Without replacement

Samples in the language of probabilities

Random samples

Recall the two types of random samples:

- With replacement

  (n elements are picked independently, uniformly at random.)

- Without replacement

  (n distinct elements are picked at random, uniformly from all possibilities.)

Samples in the language of probabilities

Random samples

Recall the two types of random samples:

- With replacement

  (n elements are picked independently, uniformly at random.)
- Without replacement

  (n distinct elements are picked at random, uniformly from all possibilities.)

Sample statistics

We will focus on the following:

- Sample mean
- Sample proportion
- Sample variance
- ...

Samples in the language of probabilities

Random samples

Recall the two types of random samples:

- With replacement

  (n elements are picked independently, uniformly at random.)
- Without replacement

  (n distinct elements are picked at random, uniformly from all possibilities.)

Sample statistics

We will focus on the following:

- Sample mean
- Sample proportion
- Sample variance (discussed later in the course)
- ...

Samples in the language of probabilities

Random samples

Recall the two types of random samples:

- With replacement

  (n elements are picked independently, uniformly at random.)

- Without replacement

  (n distinct elements are picked at random, uniformly from all possibilities.)

Sample statistics

We will focus on the following:

- Sample mean
- Sample proportion (a special case of sample mean!)
- Sample variance (discussed later in the course)
- ...

Samples in the language of probabilities

Random samples

Recall the two types of random samples:

- With replacement

  (n elements are picked independently, uniformly at random.)
- Without replacement

  (n distinct elements are picked at random, uniformly from all possibilities.)

Sample statistics

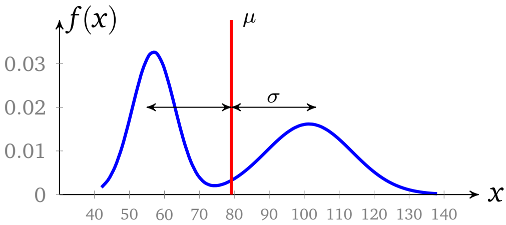
We will focus on the following:

- Sample mean
- Sample proportion (a special case of sample mean!)
- Sample variance (discussed later in the course)
- ...

Sample mean

Example (Weight of orangutans)

The weights of adult orangutans are approximately distributed according to following probability density function, with mean $\mu = 79.15\,\text{Kg}$ and standard deviation $\sigma = 24.28\,\text{Kg}$.

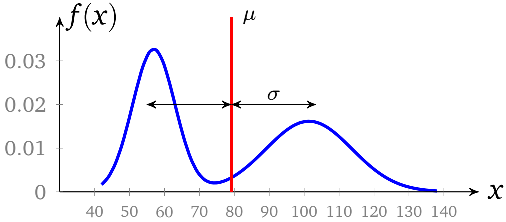


Sample mean

Example (Weight of orangutans)

The weights of adult orangutans are approximately distributed according to following probability density function, with mean $\mu = 79.15\,\text{Kg}$ and standard deviation $\sigma = 24.28\,\text{Kg}$.

Q Why do you think the distribution is bimodal?



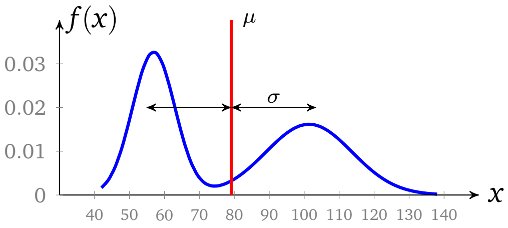
Sample mean

Example (Weight of orangutans)

The weights of adult orangutans are approximately distributed according to following probability density function, with mean $\mu = 79.15\,\text{Kg}$ and standard deviation $\sigma = 24.28\,\text{Kg}$.

Q Why do you think the distribution is bimodal?

A The two peaks correspond to the typical weights of males and females.

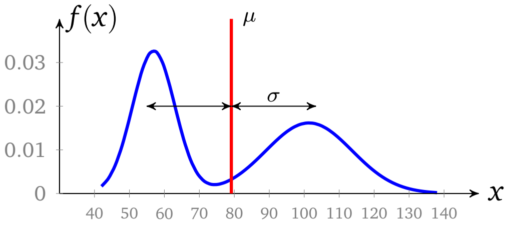


Sample mean

Example (Weight of orangutans)

The weights of adult orangutans are approximately distributed according to following probability density function, with mean $\mu = 79.15\,\text{Kg}$ and standard deviation $\sigma = 24.28\,\text{Kg}$.

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.



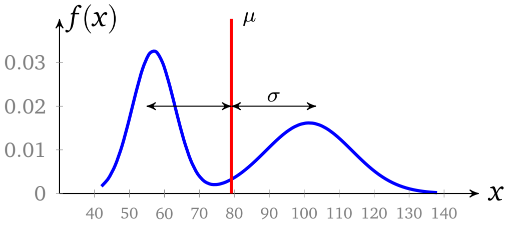
Sample mean

Example (Weight of orangutans)

The weights of adult orangutans are approximately distributed according to following probability density function, with mean $\mu = 79.15\,\text{Kg}$ and standard deviation $\sigma = 24.28\,\text{Kg}$.

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.

Q What can we say about the sample mean?



Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.

Q What can we say about the sample mean?

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.

- Q What can we say about the sample mean?
- A Taking a random sample of 50 orangutans is a random experiment.

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.

Q What can we say about the sample mean?

A Taking a random sample of 50 orangutans is a random experiment. The weights of the picked orangutans

$$W_1, W_2, W_3, \ldots, W_{50}$$

are random variables in this experiment.

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.

Q What can we say about the sample mean?

A Taking a random sample of 50 orangutans is a random experiment. The weights of the picked orangutans

$$W_1, W_2, W_3, \ldots, W_{50}$$

are random variables in this experiment. Hence, the sample mean

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

is again a random variable.

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.
Let

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

be the sample mean. This is a random variable.

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.
Let

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

be the sample mean. This is a random variable.

Q Should we expect $\overline{W}$ to be close to the population mean μ?

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.

Let

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

be the sample mean. This is a random variable.

- Q Should we expect $\overline{W}$ to be close to the population mean μ?
- Q What is the chance that $\overline{W}$ deviates from μ by a large amount?

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.
Let

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

be the sample mean. This is a random variable.

- Q Should we expect $\overline{W}$ to be close to the population mean μ?
- Q What is the chance that $\overline{W}$ deviates from μ by a large amount?

In order to answer these questions, we first need to address the following questions:

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.
Let

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

be the sample mean. This is a random variable.

Q What is the mean of $\overline{W}$?

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.
Let

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

be the sample mean. This is a random variable.

- Q What is the mean of $\overline{W}$?
- Q What is the standard deviation of $\overline{W}$?

Sample mean

Example (Weight of orangutans)

Researchers in Borneo plan to take a random sample of 50 orangutans and measure their weights.

Let

$$\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}$$

be the sample mean. This is a random variable.

- Q What is the mean of $\overline{W}$?
- Q What is the standard deviation of $\overline{W}$?
- Q Can we determine the distribution of $\overline{W}$?

Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What do we know about $W_1, W_2, W_3, \ldots, W_{50}$?

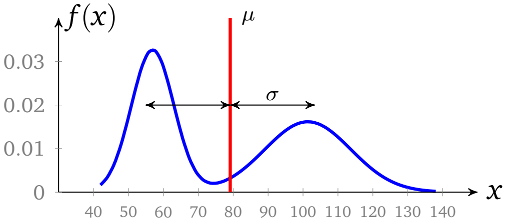
Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What do we know about $W_1, W_2, W_3, \ldots, W_{50}$?

A ① Each of $W_1, W_2, W_3, \ldots, W_{50}$ is distributed according to the population distribution.



Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What do we know about $W_1, W_2, W_3, \ldots, W_{50}$?

$\boxed{\text{A}}$

1. Each of $W_1, W_2, W_3, \ldots, W_{50}$ is distributed according to the population distribution.
2. If the sample is taken with replacement, then $W_1, W_2, W_3, \ldots, W_{50}$ are independent.

Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What do we know about $W_1, W_2, W_3, \ldots, W_{50}$?

A
1. Each of $W_1, W_2, W_3, \ldots, W_{50}$ is distributed according to the population distribution.
2. If the sample is taken with replacement, then $W_1, W_2, W_3, \ldots, W_{50}$ are independent.
3. If the sample is taken without replacement, then $W_1, W_2, W_3, \ldots, W_{50}$ are distinct (and not independent).

Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What do we know about $W_1, W_2, W_3, \ldots, W_{50}$?

A
1. Each of $W_1, W_2, W_3, \ldots, W_{50}$ is distributed according to the population distribution.
2. If the sample is taken with replacement, then $W_1, W_2, W_3, \ldots, W_{50}$ are independent.
3. If the sample is taken without replacement, then $W_1, W_2, W_3, \ldots, W_{50}$ are distinct (and not independent).

In general:

If the sample size is "small" compared to the population size, then the distinction between sampling with or without replacement becomes negligible! (Rule of thumb: $\frac{\text{sample size}}{\text{population size}} \leq 5\%$)

Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What is the mean of $\overline{W}$?

Q What is the standard deviation of $\overline{W}$?

Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What is the mean of $\overline{W}$?

A The mean of each W_i is the population mean μ. Hence,

$$\mathbb{E}[\overline{W}] = \mu = \boxed{79.15\,\mathrm{Kg}}$$

Q What is the standard deviation of $\overline{W}$?

Sample mean

Example (Weight of orangutans)

$$\boxed{\overline{W} := \frac{W_1 + W_2 + W_3 + \cdots + W_{50}}{50}}$$

Q What is the mean of $\overline{W}$?

A The mean of each W_i is the population mean μ. Hence,

$$\mathbb{E}[\overline{W}] = \mu = \boxed{79.15\,\text{Kg}}$$

Q What is the standard deviation of $\overline{W}$?

A The standard deviation of each W_i is the population standard deviation σ. Assuming the sample is with replacement,

$$\mathbb{SD}[\overline{W}] = \frac{\sigma}{\sqrt{50}} = \frac{24.28}{\sqrt{50}} \approx \boxed{3.43\,\text{Kg}}$$

Sample mean

Mean and standard deviation of sample mean

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a population with mean μ and standard deviation σ, then the sample mean

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is a random variable with

- $\boxed{\mathbb{E}[\overline{X}] = \mu}$ (valid for both types of sample)
- $\boxed{\mathbb{SD}[\overline{X}] = \sigma/\sqrt{n}}$ (assuming the sample is with replacement)

Sample mean

Mean and standard deviation of sample mean

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a population with mean μ and standard deviation σ, then the sample mean

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is a random variable with

- $\boxed{\mathbb{E}[\overline{X}] = \mu}$ (valid for both types of sample)
- $\boxed{\mathbb{SD}[\overline{X}] = \sigma/\sqrt{n}}$ (assuming the sample is with replacement)

These are consequences of the following general principles:

If X and Y are arbitrary random variables, then

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- $\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]$ provided that X and Y are independent.

Sample mean

Mean and standard deviation of sample mean: interpretation

The sample mean $\overline{X}$ can be thought of as an estimator for the population mean.

Sample mean

Mean and standard deviation of sample mean: interpretation

The sample mean $\overline{X}$ can be thought of as an estimator for the population mean.

- The fact that $\mathbb{E}[\overline{X}] = \mu$ means that

  $\overline{X}$ is an unbiased estimator for μ.

Sample mean

Mean and standard deviation of sample mean: interpretation

The sample mean $\overline{X}$ can be thought of as an estimator for the population mean.

- The fact that $\mathbb{E}[\overline{X}] = \mu$ means that

  $\overline{X}$ is an unbiased estimator for μ.

- The fact that $\mathbb{SD}[\overline{X}] = \sigma/\sqrt{n}$ reflects out expectation that

  The larger the sample, the more accurate the estimator.

  (The standard deviation of the sample mean is inversely proportional to the square root of the sample size.)