5 Sampling Distributions of Statistics

MTH 3240 Environmental Statistics

Spring 2020

Notes

Objectives

Objectives:

- Explain the term *sampling distribution* of a statistic.
- Explain the term *standard error* of a statistic.
- Identify the mean and standard error of the sampling distribution of the sample mean, and state the two situations in which the distribution will be a normal distribution.

Notes

Introduction

- A ***statistic*** is a numerical value computed from **random** sample data.

  Examples:
  - Sample mean $\bar{X}$
  - Sample median $\tilde{X}$
  - Sample standard deviation S
  - Sample proportion $\hat{P}$
- Statistics are **random variables** because their values are determined by chance.

Notes

- The sample-to-sample variation in the value of a statistic is called ***sampling variation***.
- The **probability distribution** of a statistic is called its ***sampling distribution***.

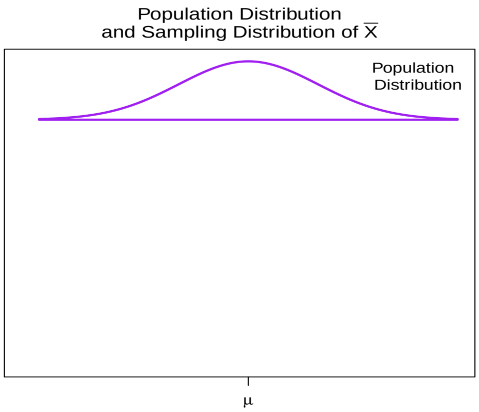
  The **sampling distribution** of a statistic specifies the values that the statistic might take and the probabilities with which it takes those values.

Notes

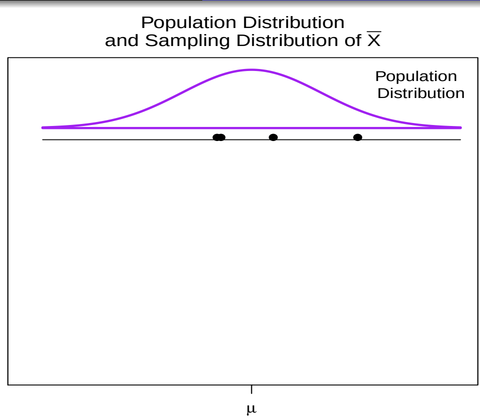
The Sampling Distribution of $\bar{X}$

- For now, we'll focus on the **sampling distribution** of the **sample mean** $\bar{X}$.
- The next slide shows a **population distribution** (top), ten samples of size $n = 4$ from the population, their ten sample means, and the **sampling distribution** of the sample mean (bottom).

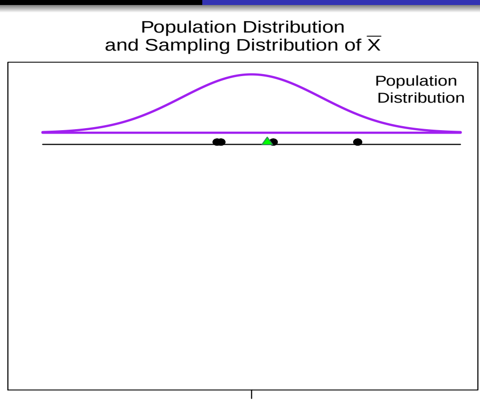
Notes



Notes

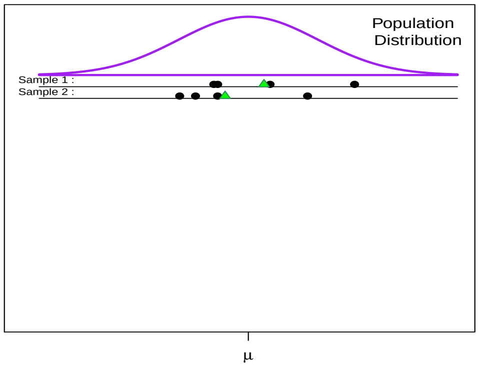


Notes



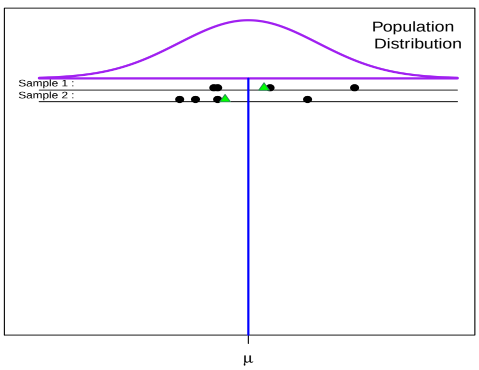
Notes

Population Distribution and Sampling Distribution of $\overline{X}$



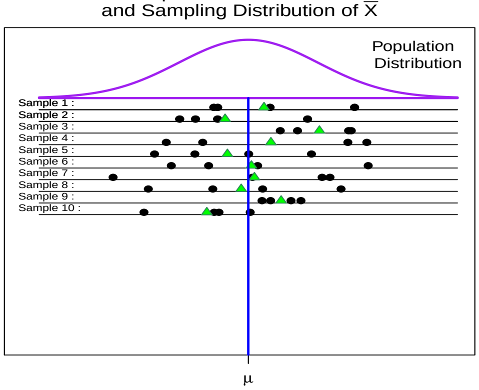
Notes

Population Distribution and Sampling Distribution of $\overline{X}$



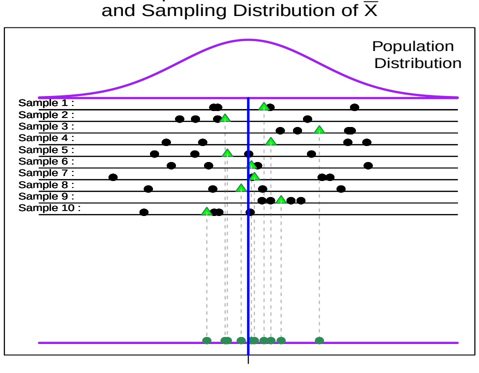
Notes

Population Distribution and Sampling Distribution of $\overline{X}$



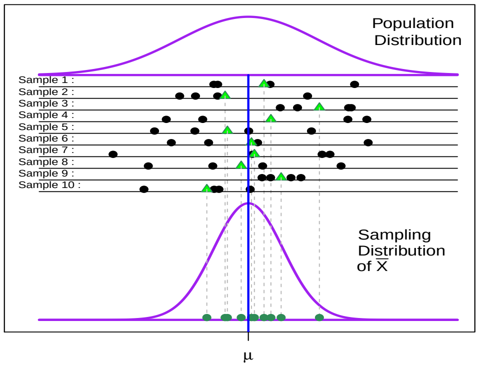
Notes

Population Distribution and Sampling Distribution of $\overline{X}$



Notes

Population Distribution
and Sampling Distribution of $\overline{X}$



Notes

- The **shape**, **center**, and **spread** (variation) of the **sampling distribution of $\bar{X}$** will depend on:
  - The shape of the population from which the sample is drawn.
  - The center of the population, μ.
  - The spread of the population, σ.
  - The sample size n.

Notes

- Under either of two important scenarios, $\bar{X}$ **follows** (at least approximately) a **normal distribution**:
  1. When the sample is from a **normal population**, or
  2. When the **sample size is n large**.

  More details are given in the next two facts.

Notes

Normality of the $\bar{X}$ Distribution

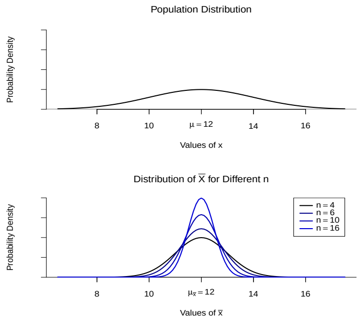
Fact: Suppose we have a random sample from a **N(μ, σ) population**. Then

$$\bar{X} \sim \text{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

As a consequence, if we **standardize** $\bar{X}$, the resulting random variable Z will follow a **standard normal** distribution, i.e.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{N}(0, 1).$$

Notes



Notes

Fact: **The Central Limit Theorem**: Suppose we have a random sample from **any population** (*not* necessarily normal) whose mean is μ and whose standard deviation is σ. Then **if n is large**,

$$\bar{X} \sim \text{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right),$$

at least **approximately**. The larger n is, the more closely the $\bar{X}$ distribution resembles the **normal distribution**.

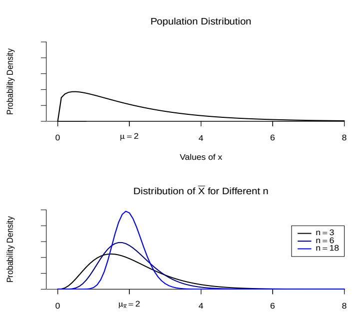
Notes

As a consequence, if we **standardize** $\bar{X}$, the resulting random variable Z follows a **standard normal** distribution, i.e.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{N}(0, 1)$$

(at least **approximately**).

Notes



Notes

- Sometimes we use $\mu_{\bar{X}}$ to denote the **mean** of the $\bar{X}$ **distribution**, and $\sigma_{\bar{X}}$ to denote its **standard deviation**, i.e.

$$\mu_{\bar{X}} = \mu$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

- The **standard deviation** of the $\bar{X}$ **distribution** is sometimes called the ***standard error*** of $\bar{X}$ (more on this later).

Notes

- The next example shows how to obtain **probabilities** from the sampling distribution of $\bar{X}$.

Notes

Example

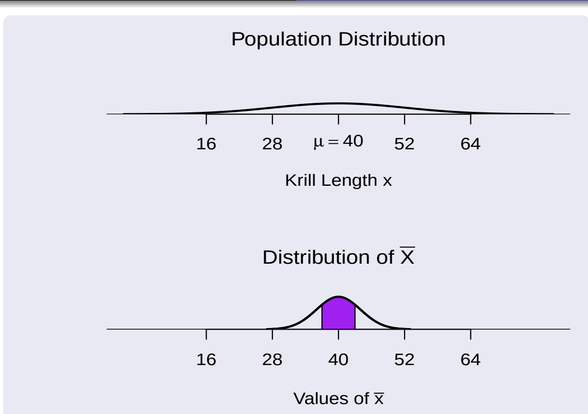
Body lengths of the Southern ocean krill species *Euphausia supeba* are normally distributed with **mean** $\mu = 40$ mm and **standard deviation** $\sigma = 12$ mm.

A random sample of $n = 9$ krill is to be taken.

The **sampling distribution** of $\bar{X}$ is a **normal** distribution with **mean** and **standard error**

$$\mu_{\bar{X}} = 40 \quad \text{and} \quad \sigma_{\bar{X}} = \frac{12}{\sqrt{9}} = 4.0.$$

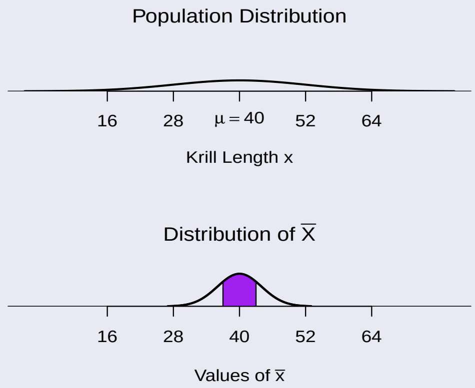
Notes



Notes

The **probability** that $\bar{X}$ will fall between **37** and **43** mm when a random sample of size $n = 9$ is selected is the shaded area shown on the next slide.

Notes



Notes

The **probability** (obtained using software) is **0.5468**, i.e. there's a **54.68%** chance that the **mean** of the nine krill lengths will be between **37** and **43** mm.

Notes

- The next example shows how to obtain **percentiles** from the sampling distribution of $\bar{X}$.

Notes

Example (Cont'd)

Recall that lengths of krill follow a normal distribution with **mean** $\mu = 40$ mm and **standard deviation** $\sigma = 12$ mm.

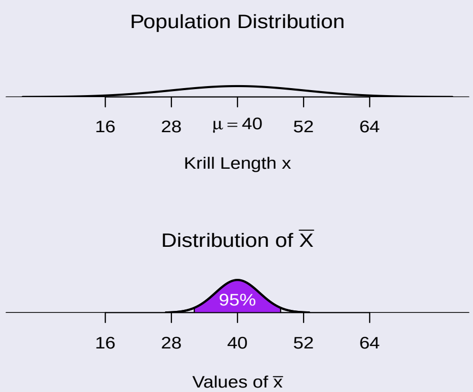
The **sampling distribution** of $\bar{X}$, for samples of size $n = 9$, is a **normal** distribution with **mean** and **standard error**

$$\mu_{\bar{X}} = 40 \quad \text{and} \quad \sigma_{\bar{X}} = \frac{12}{\sqrt{9}} = 4.0.$$

Notes

The **2.5th** and **97.5th percentiles** of the **sampling distribution** of $\bar{X}$ when a random sample of size $n = 9$ is selected are the two lengths (mm) that delimit the shaded area shown on the next slide.

Notes



Notes

These **percentiles** are the two values that capture the **middle 95%** of the $\bar{X}$ **distribution**.

They're obtained by "unstandardizing" the corresponding N$(0, 1)$ percentiles ($\pm$1.96):

$$\mu_{\bar{X}} + 1.96\,\sigma_{\bar{X}} = 40 + 1.96 \times 4.0 = \mathbf{47.8}$$

and

$$\mu_{\bar{X}} + (-1.96)\,\sigma_{\bar{X}} = 40 + (-1.96) \times 4.0 = \mathbf{32.2}.$$

Thus there's a **95% chance** that $\bar{X}$ will be between **47.8** and **32.2** mm.

Notes

- The fact that the **mean** of the $\bar{X}$ **distribution** equals the **mean** of the **population**, i.e.

$$\mu_{\bar{X}} = \mu,$$

says that *on average*, the **sample mean** $\bar{X}$ will equal the **population mean** μ.

- But any *particular* $\bar{X}$ almost certainly won't equal μ exactly.

The discrepancy between a *particular* **estimate** $\bar{X}$ and the **true value** μ is called the ***sampling error***.

Sampling Error of $\bar{X}$:

$$\text{Sampling Error} = \bar{X} - \mu$$

Notes

- The standard deviation of the $\bar{X}$ **distribution**,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

represents the size of a **typical sampling error** of $\bar{X}$ **away from** μ.

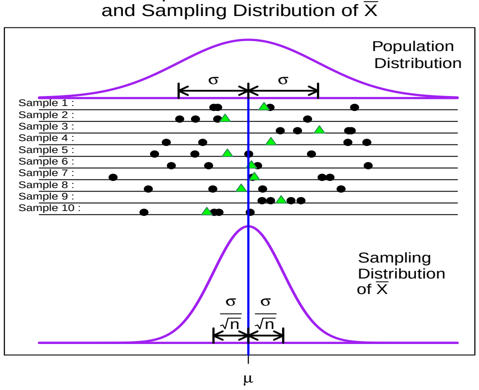
Because of this, $\sigma/\sqrt{n}$ is called the ***standard error*** of $\bar{X}$.

- The **standard error** will be **small** when either:
  - The sample size n **is large**,
    or
  - The population standard deviation σ **is small**.

We *can't* control σ, but we *can* choose a value for n.

Notes

Population Distribution
and Sampling Distribution of $\bar{X}$

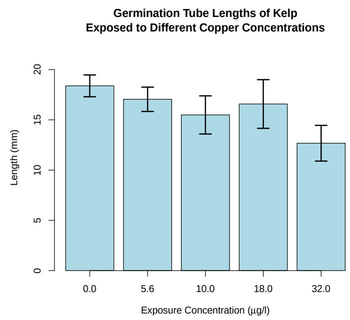


Notes

- It's useful, when graphing **sample means**, to include ***error bars*** that extend one or two **standard errors** above and below the means.

In practice, we use the ***estimated standard error***, $S/\sqrt{n}$, where S is the sample standard deviation, because σ isn't known.

Notes



Notes

Sampling Distributions of Other Statistics

- **Every statistic** (e.g the sample median $\tilde{X}$, sample standard deviation S, sample proportion $\hat{P}$, etc.), follows **some *sampling distribution***, but it *might not* be a normal distribution.

  The standard deviation of that sampling distribution is always called the ***standard error*** of the statistic.

  The ***standard error*** indicates the size of a **typical sampling error** when the statistic is used to estimate the corresponding population parameter.

Notes

Notes

Notes