Statistics I – Chapter 8 Estimation for One Population (Part 1)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

November 28, 2012

Introduction

- We have studied Descriptive Statistics (Chapters 2 and 3) and Probability (Chapters 4 to 7).
- Now we are ready to study inferential Statistics.
- In particular, we want to:
  - Estimate population parameters (Chapters 8 and 10).
  - Test hypotheses about parameters (Chapter 9 to 11).
  - And more.
- The concepts introduced in Chapters 8 and 9 are the heart of Inferential Statistics!

Introduction

- Consider the quality control problem again.
- For all LED lamps of brand IM, we are interested in μ, the average number of hours of luminance.
- Let's select a random sample of 40 lamps. A test shows that the sample mean is $\bar{x} = 28000$ hours.
  - What's the probability that $\mu = \bar{x}$?
  - What's the probability that $\mu \in [27000, 29000]$?
  - What's the probability that $\mu \in [26000, 30000]$?
  - Why don't we use the median?
- Now we are able to answer these questions.

Road map

- **Point estimation**.
- Interval estimation.
- Estimating the population mean.
  - When the population variance is known.

Estimators

- From a population, we may collect a subset as a sample.
- From a sample, we may calculate **statistics**.
- A statistic is a **function** of values in a sample.
  - E.g., the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  - E.g., the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
- When a statistic is used to **estimate** a population **parameter**, it is called an **estimator** of that parameter.
  - E.g., $\overline{X}$ can be used as an estimator of μ.
  - E.g., S^2 can be used as an estimator of σ^2.

Estimators

- A statistic is an estimator **of a parameter**.
  - It is meaningless to say "The sample mean is an estimator." An estimator of what?
- An estimator is nothing but a statistic of a particular use.
  - It is still a function of values in a sample.
  - It is a **random variable**.
  - It has a specific **target**: the parameter.
- The **realized value** of an estimator is called an **estimate**.

Estimators

- For a parameter, there are **multiple** estimators.
- Suppose we want to estimate the population mean μ.
  - One (intuitive) estimator is the sample mean $\overline{X}$.
  - One may also use the sample median as an estimator.
  - One may even use the sample maximum $X_{\max} \equiv \max\limits_{i=1,\ldots,n}\{X_i\}$, sample minimum $X_{\min} \equiv \min\limits_{i=1,\ldots,n}\{X_i\}$, or something creative such as $\frac{1}{2}(X_{\max} + X_{\min})$, $\frac{1}{3}(X_1 + 2X_2)$, etc.
- Which estimator is **good**?

Point estimation

- One way to estimate a parameter is as follows:
  - Define an estimator.
  - Conduct sampling and generate a sample.
  - Calculate the **realized value**, the **estimate**, of the estimator.
  - Claim that "I think the parameter is close to the estimate."
- In short, we "guess" that the parameter is close to the realized value of an estimator, the estimate.
- The above process is called **point estimation**.

Point estimation: An example

- Suppose we want to estimate the average number of hours one spend in homework per week in this class. Let it be μ.
- Suppose we ask 10 students and get

$$6 \quad 2 \quad 4 \quad 2 \quad 5 \quad 3 \quad 12 \quad 4 \quad 2 \quad 1.$$

  - If we have defined the sample mean as our estimator, the estimate will be 4.1. We will guess that μ is close to 4.1.
  - If we have defined the sample maximum as our estimator (which is obviously bad), the estimate will be 12.
  - If we have define the sample median as our estimator, the estimate will be 3.5.

Point estimation

- Probably it is obvious that in estimating the population mean, the best idea is to use the sample mean.
- But some things are not so obvious.
- Consider the population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$:
  - We define the sample variance as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
  - Why $n - 1$?
  - Why don't we define it as $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$?
  - Is S^2 a good estimator of σ^2?

Properties of a point estimator

- To answer all these questions, we need to first define "good".
- Among many properties, three of them are:
  - Unbiasedness,
  - Relative efficiency, and
  - Consistency.
- An estimator is "good" if it is unbiased, relatively efficient, and consistent.

Unbiasedness

- Believed by most statisticians, the first thing is for an estimator to be **unbiased**.

Definition 1

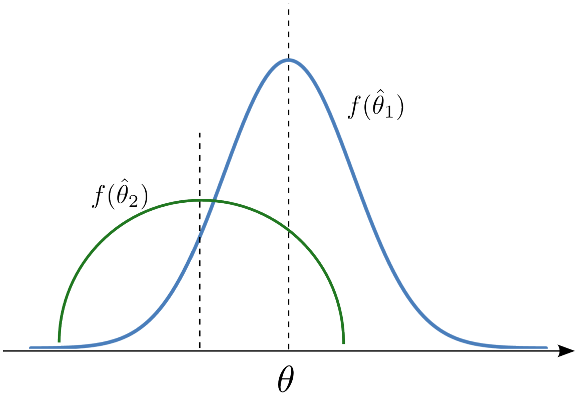
Let θ be a parameter and $\hat{\theta}$ be an estimator of θ. $\hat{\theta}$ is unbiased if

$$\mathbb{E}\left[\hat{\theta}\right] = \theta.$$

  - The parameter θ is a constant.
  - The estimator $\hat{\theta}$ is a random variable.
  - $\hat{\theta}$ may take different values, but **in expectation** it is θ.

Unbiasedness

- $\hat{\theta}_1$ is unbiased while $\hat{\theta}_2$ is biased.



Unbiasedness of the sample variance

- Now we may answer why the denominator of the sample variance is $n-1$ instead of n.

Proposition 1

The sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is unbiased for the population variance σ^2, *i.e.,* $\mathbb{E}[S^2] = \sigma^2$.

Proof. Because $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$, we have $\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) - n\mathbb{E}\left(\overline{X}^2\right)$.

Unbiasedness of the sample variance

- *Proof (cont'd).* Because $\mathbb{E}[X_i^2] = \text{Var}(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2$ and $\mathbb{E}\big[\overline{X}^2\big] = \text{Var}(\overline{X}) + \mathbb{E}[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2$, we have,

$$\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

It follows that

$$\mathbb{E}(S^2) = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2,$$

so we see that S^2 is an unbiased estimator for σ^2. □

Unbiasedness

- For the population mean μ:
  - The sample mean $\overline{X}$ is unbiased:

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}(n\mu) = \mu.$$

  - The sample median is biased.
  - The sample maximum is biased as long as $n > 1$. E.g., suppose $X_i \sim \text{Uni}(0,2)$, we have

$$\mathbb{E}[X_{\max}] = \int_0^2 x\left(\frac{nx^{n-1}}{2^n}\right)dx = \frac{2n}{n+1} > 1.$$

  - How about this statistic: $\frac{1}{3}(X_1 + 2X_2)$?

Relative efficiency

- Between two unbiased estimators, we prefer the one that is **relatively efficient**, i.e., with **smaller variance**.

Definition 2

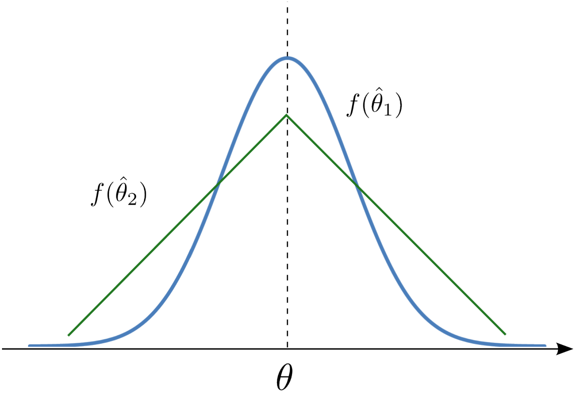
Let θ be a parameter and $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ. The efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is the ratio

$$\frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_1)}.$$

- The smaller the variance, the larger the relative efficiency (if they are unbiased).

Relative efficiency

- $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.



Relative efficiency

- For the population mean μ:
  - $\frac{1}{2}(X_1 + X_2)$ and $\frac{1}{3}(X_1 + 2X_2)$ are both unbiased.
  - Which one is more efficient?
  - We have

$$\text{Var}\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}(1+1) = \frac{1}{2} \text{ and}$$
$$\text{Var}\left(\frac{X_1 + 2X_2}{3}\right) = \frac{1}{9}(1+4) = \frac{5}{9},$$

  so $\frac{1}{2}(X_1 + X_2)$ is more efficient.
- In general, the sample mean is more efficient than any weighted average with various weights (why?).

Consistency

- An estimator should be **consistent**, i.e., get closer to the parameter (probabilistically) as the **sample size** n goes up.
  - In particular, it should **converge** to the parameter as $n \to \infty$.

Definition 3

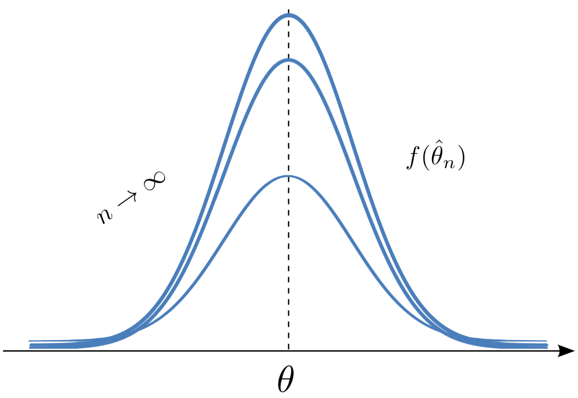
Let θ be a parameter and $\hat{\theta}_n$ be an estimator of θ whose sample size is n. $\hat{\theta}_n$ is consistent if for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} \Pr\left(\left|\hat{\theta}_n - \theta\right| \leq \epsilon\right) = 1.$$

- In other words, the "guess" will be "correct" when the sample size goes to infinity.

Consistency

- $\hat{\theta}_n$ converges to θ as $n \to \infty$.



Consistency

- Is a sample mean consistent? Do we have

$$\lim_{n\to\infty} \Pr\left(|\overline{X} - \mu| > \epsilon\right) = 0 \quad \forall \epsilon > 0.$$

- You have proved this in Problem 5 of Homework 6!
- This important result is called the **law of large numbers**.

Proposition 2 (Law of large numbers)

The sample mean converges to the population mean as the sample size goes to infinity.

Summary

- We use statistics to estimate parameters.
- When we use a single number as an estimate, we are doing point estimation.
- For a single parameter, there are multiple point estimators.
- Some estimators are better than others.
- A good estimator should be:
  - Unbiased,
  - Relatively effective, and
  - Consistent.

Road map

- Point estimation.
- **Interval estimation**.
- Estimating the population mean.
  - When the population variance is known.

Drawbacks of point estimation

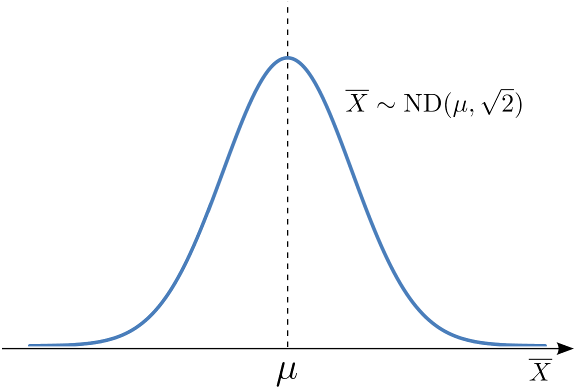
- Indeed some point estimators are good.
  - E.g., the sample mean is a good for the population mean.
- However, there are some drawbacks of point estimation:
  - We know the population mean is close to the sample mean. But **how close** it is?
  - No matter how good an estimator is, if we use just one value, the probability of making a correct guess is typically **zero**!
- Therefore, instead of suggesting a number, it will be better to suggest an **interval**.
- We need to measure how good an interval is.

Interval estimation: the first illustration

- Let's illustrate the idea with population and sample means.
- Let's assume the population is normal with known variance $\sigma^2 = 16$. The population mean, μ, is unknown.
- Let the sample mean $\overline{X}$ be the estimator.
- The sample size $n = 8$.
- We have observed the value of sample mean, $\bar{x} = 10$.
  - $\overline{X}$ is a statistic and $\bar{x}$ is a realized value.
- Intuitively, the interval should **center at $\bar{x}$**.
- We want to find the smallest $b > 0$ such that the interval $I(b) = [\bar{x} - b, \bar{x} + b]$ **covers μ** with a 95% probability.
- How?

The sampling distribution

- This is possible because we know the distribution of $\overline{X}$.
- As the population is normal, $\overline{X} \sim \text{ND}(\mu, \frac{\sigma}{\sqrt{n}} = \sqrt{2})$.



The sampling distribution

- Suppose someone randomly says: How about

$$I(\sigma) = [\bar{x} - \sigma, \bar{x} + \sigma] = \left[10 - \sqrt{2}, 10 + \sqrt{2}\right]?$$

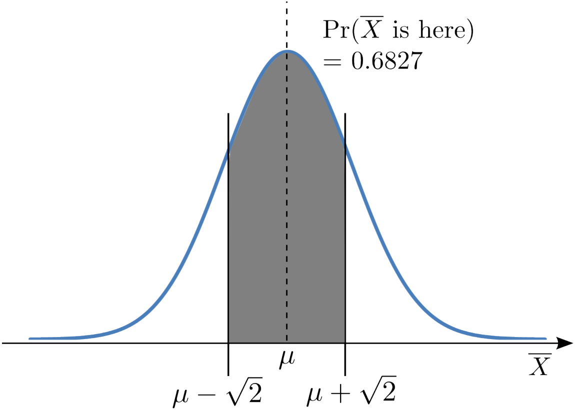
- How to measure the quality of this interval?
- Consider an "unknown" interval centered at μ: $U(\sqrt{2}) = [\mu - \sqrt{2}, \mu + \sqrt{2}]$. Let $Z \sim \text{ND}(0, 1)$, we have

$$\Pr\left(\overline{X} \in U\right) = \Pr\left(\mu - \sqrt{2} \leq \overline{X} \leq \mu + \sqrt{2}\right)$$
$$= \Pr(-1 \leq Z \leq 1) = 0.6827.$$

- The **location** of the interval $U(\sqrt{2})$ is unknown because μ is unknown. But its **size** is known: $2\sqrt{2}$.

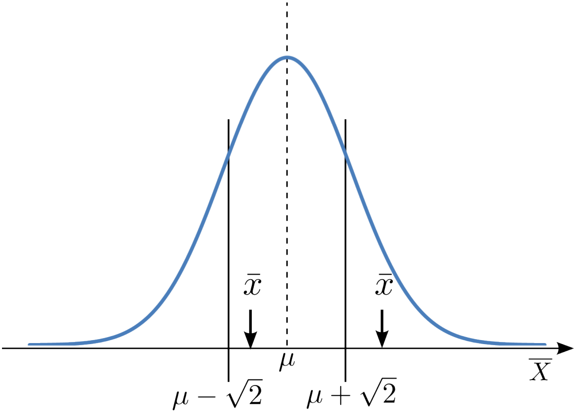
The sampling distribution

- We do not know where μ is, but we know the probability for $\overline{X}$ to deviate from μ by less than $\frac{\sigma}{\sqrt{n}} = \sqrt{2}$.



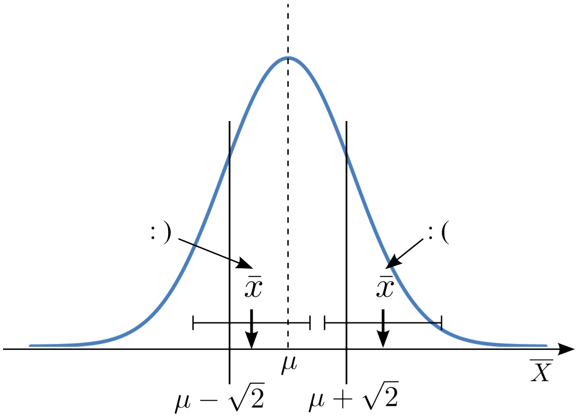
How good an interval is?

- Now, let's consider $I(\sqrt{2}) = \left[10 - \sqrt{2}, 10 + \sqrt{2}\right]$ again.
- $\bar{x} = 10$ can be close to or far from μ.



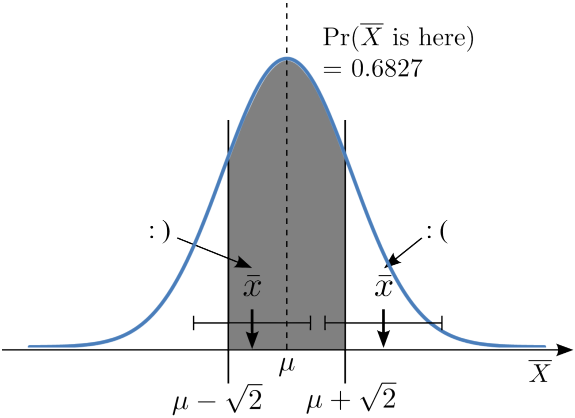
How good an interval is?

- If, luckily, $\bar{x} = 10$ is close enough to μ, $I(\sqrt{2})$ covers μ.
- If, unluckily, $\bar{x} = 10$ is far from μ, $I(\sqrt{2})$ does not cover μ.



How good an interval is?

- The probability that "we are lucky" is exactly 0.6827!
  - $\Pr\left(\left|\overline{X} - \mu\right| \leq \sqrt{2}\right) = 0.6827.$

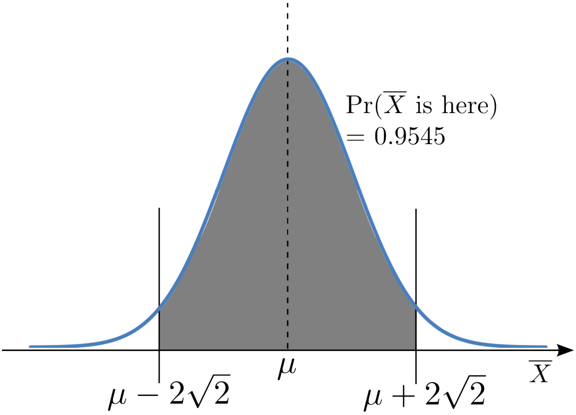


How good an interval is?

- In conclusion, given **any** realization $\bar{x}$, $\left[\bar{x} - \sqrt{2}, \bar{x} + \sqrt{2}\right]$ **covers** μ with probability 0.6827.
  - We can reach this conclusion as we know $\overline{X} \sim \text{ND}(\mu, \sqrt{2})$.
- But 0.6827 is not enough: We want 0.95.
- So instead of having $\sqrt{2}$ as the leg length, let's try $2\sqrt{2}$.

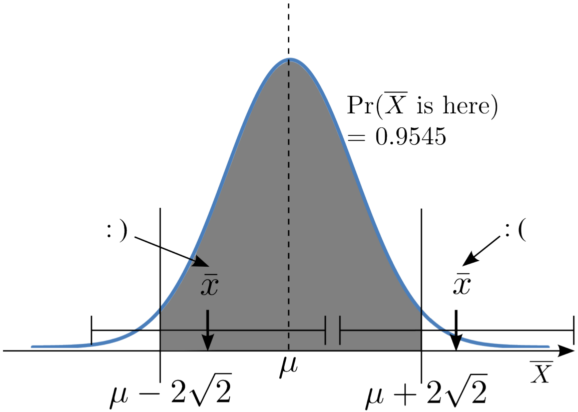
A larger interval

- We do not know where μ is, but we know the probability for $\overline{X}$ to deviate from μ by less than $2\frac{\sigma}{\sqrt{n}} = 2\sqrt{2}$.



A larger interval

- The probability that "we are lucky" now becomes 0.9545!
  - $\Pr\left(\left|\overline{X} - \mu\right| \leq 2\sqrt{2}\right) = 0.9545.$



What should be the leg size?

- We made two attempts:
  - $\left[10-\sqrt{2}, 10+\sqrt{2}\right]$ is too small: The covering probability is 0.6827, which is $\Pr(-1 <= Z <= 1)$.
  - $\left[10-2\sqrt{2}, 10+2\sqrt{2}\right]$ is too large: The covering probability is 0.9545, which is $\Pr(-2 <= Z <= 2)$.
- To get exactly 0.95, we need to solve

$$\Pr(-z \leq Z \leq z) = 0.95$$

  The answer is $z = 1.96$.
- So the desired interval is

$$\left[10-1.96\sqrt{2}, 10+1.96\sqrt{2}\right] = [7.228, 12.772].$$

  It covers μ with probability 0.95.

Summary

- We want to construct an interval that will cover the population mean with a predetermined probability.
- As we have the value of the sample mean, it is natural to make the interval centering at the sample mean.
- We may measure the quality (the probability of covering the population mean) of each interval because:
  - $\left[\overline{X} - b, \bar{X} + b\right]$ covers $\mu \Leftrightarrow \left|\overline{X} - \mu\right| \leq b$.
  - The probability of the latter can be calculated if we know the distribution of $\overline{X}$.

Summary

- The interval is called a **confidence interval** (CI).
- The probability of covering the desired parameter is called the **confidence level**.
- The typical way to state a conclusion is

  "With a $1 - \alpha$ confidence level, the population parameter will be covered by the confidence interval."

  - In practice, $1 - \alpha$ is typically chosen to be 90%, 95%, or 99%.

Road map

- Point estimation.
- Interval estimation.
- **Estimating the population mean**.
  - When the population variance is known.

Estimating population mean

- Let's consider the task again: to suggest an interval that covers the **population mean** μ with a certain probability.
- While we do this based on the **sample mean** $\overline{X}$, the key is to know the **sampling distribution** of $\overline{X}$.
- We need to study many different cases:
  - Known or unknown population variance.
  - Normal or nonnormal population.
  - Large or small sample size.
  - Infinite or finite population (or sampling without or with replacement).

Known population variance

- In this section, we will assume that the population variance σ^2 is **known**.
- Is it possible that the population mean is unknown but the population variance is known?
  - Certainly this is not so common.
- Consider the following example:
  - A machine produces an item.
  - Once the desired length is set manually, the variance of the lengths of items is known to be 0.04cm^2.
  - However, after you fire a bad employee, he modified the setting without telling anyone...

Known population variance

- In practice, if you do not know the population variance, try to use the methods introduced in the next section.
- If only methods which assumes known population variance are available, you will need to estimate or test the population variance first.
  - To be introduced later in this semester.

General setting

- The unknown population mean is μ.
- The known population variance is σ^2.
- The sample mean is $\overline{X}$.
- The realized value of sample mean is $\bar{x}$.
- The sample size is n.
- The desired confidence level is $1 - \alpha$.
  - α is the allowed probability for not covering μ.

General setting

- In general, we are looking for a smallest $b > 0$ such that the interval $[\bar{x} - b, \bar{x} + b]$ covers μ with probability $1 - \alpha$.
- For simplicity, b is (almost always) measured as the **number of standard deviations** of $\overline{X}$:

$$b = z\sigma_{\overline{X}},$$

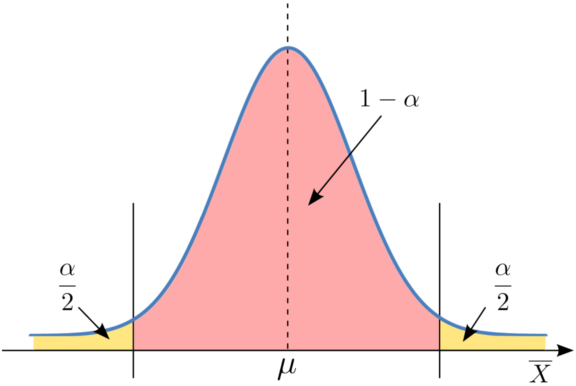
  where z is the z-score of b and $\sigma_{\overline{X}} \equiv \sqrt{\mathrm{Var}(\overline{X})}$ is the **standard error**.
- The standard error of an estimator is just a special name of standard deviations particularly for estimators.

Normal populations

- If the population is normal, the sample mean $\overline{X}$ is **normal** regardless of the sample size.
  - With sampling with replacement or infinite population ($n < 0.05N$), the standard error $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.
  - With sampling without replacement and finite population ($n > 0.05N$), the standard error $\sigma_{\overline{X}} = \left(\frac{\sigma}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$.
- We say we use **the z distribution** to construct the interval.
- Suppose we have obtained the value of $\sigma_{\overline{X}}$. How to construct the interval for the desired confidence level?

Normal populations

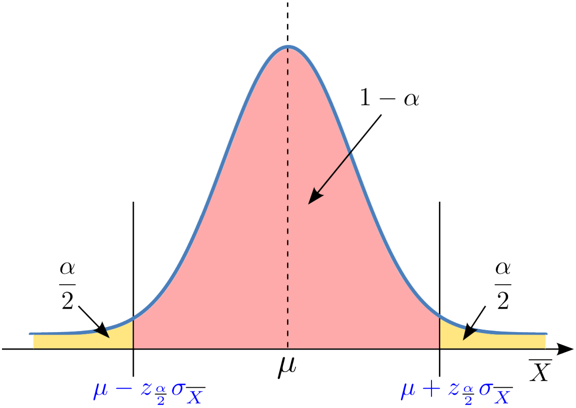
- The distribution of $\overline{X}$ can be divided into three regions based on μ and α.



- Our mission is to find the **two cutoffs**.

Normal populations

- The two cutoffs depends on $\sigma_{\overline{X}}$ and α.
  - z_t denotes the critical value such that $\Pr(Z > z_t) = t$.



Normal populations

- The confidence interval can be found in the following way:
  - Given the sample, calculate the **sample mean** $\bar{x}$.
  - Given the population variance σ^2 and sample size n (and population size N if $n > 0.05N$), calculate the **standard error** $\sigma_{\overline{X}}$.
  - Given the confidence level $1 - \alpha$, use software or table to calculate the **critical value** $z_{\frac{\alpha}{2}}$ such that $\Pr(Z > z_{\alpha/2}) = \frac{\alpha}{2}$.
  - The confidence interval is

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\sigma_{\overline{X}}, \bar{x} + z_{\frac{\alpha}{2}}\sigma_{\overline{X}}\right].$$

Normal populations

- What if the population is nonnormal?
- If the sample size is large ($n \geq 30$), we may apply the **central limit theorem** and conclude that the sample mean is still normal. Everything then follows.
- If the sample size is small ($n < 30$), we can do nothing at this moment. We need to study **Nonparametric Statistics** (in Chapter 17).

Example 1

- Recall that someone messed up our machine.
- While the variance of items produced is 0.04cm^2, the mean is unknown and must be found.
- 100 items are produced and the lengths are recorded:

$$6.01 \quad 6.12 \quad 6.03 \quad 5.96 \quad 5.51 \quad 6.31 \quad 5.79 \quad \ldots \quad 6.25$$

  The sample mean is 6.09 cm.
- Estimate the population mean with a 95% confidence interval.

Example 1

- What is the population? What is the parameter?
- Is the population normal? Is it finite or infinite?
- Answer:
  - (**Important!**) Because the population variance is known and the sample size 100 is large enough, we may use the z distribution to construct the confidence interval.
  - The sample mean is 6.09. The standard error is $\frac{0.2}{\sqrt{100}} = 0.02$.
  - The critical values are $z_{0.025} = 1.96$.
  - The confidence interval is

$$[6.09 - 1.96 \times 0.02, 6.09 + 1.96 \times 0.02] \approx [6.051, 6.129].$$

  - Conclusion: With a 95% confidence interval, the population mean is between 6.051 and 6.129.

Example 1: remarks

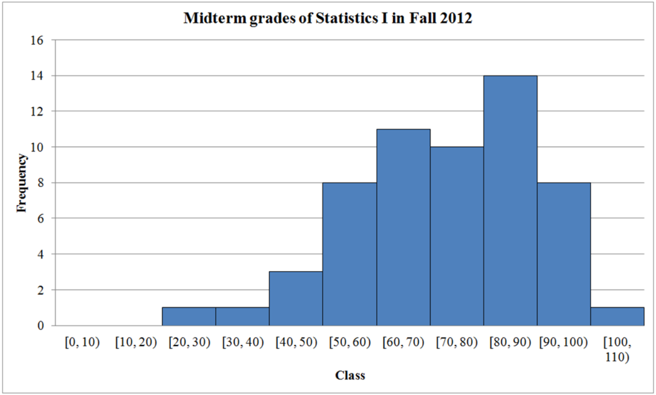
- Suppose now only 10 items are produced.
  - If according to past experience we know the population is normal, we may still construct the confidence interval.
  - If the population is nonnormal (or if we do not know whether it is normal), we can do nothing.
- If you want to see whether the population is normal:
  - At least you should draw a histogram.
  - A rigorous way (which has the chi-square distribution involved) will be introduced in Chapter 16.

Example 2

- Let's assume I didn't announce the average of the midterm.
  - But I announced the standard deviation as 16.72.
- You want to know the average of the 57 scores.
- Because some classmates refuse to tell you their scores, you cannot conduct a census.
- Among your friends, you randomly asked three persons. Their grades are 69, 72, and 92. You got 78.

Example 2

- The population distribution looks like normal:



- With a 90% confidence level, what is the confidence interval?

Example 2

- Answer:
  - Because the population variance is known and the population is normal, we use the z distribution to construct the interval.
  - The sample mean is 77.75.
  - The standard error is $\left(\frac{16.72}{\sqrt{4}}\right)\sqrt{\frac{57-4}{57-1}} = 8.13$.
  - The critical values are $z_{0.05} = 1.645$.
  - The confidence interval is

$$\begin{aligned} & [77.75 - 1.645 \times 8.13, 77.75 + 1.645 \times 8.13] \\ \approx & [64.371, 91.129]. \end{aligned}$$

  - Conclusion: With a 90% confidence interval, the population mean is between 64.371 and 91.129.
- Obviously a larger sample size will help.