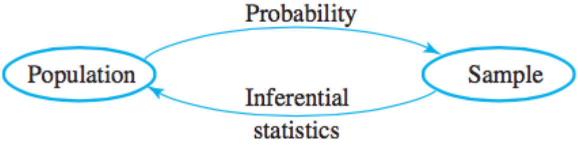
# MATH 205: Statistical methods

Lecture 24: Confidence intervals of the population mean

The story so far...

# Random sample




## Definition

The random variables $X_1, X_2, ..., X_n$ are said to form a (simple) random sample of size $n$ if

1. the $X_i$'s are independent random variables
2. every $X_i$ has the same probability distribution

# Linear combination of random variables

**Theorem**

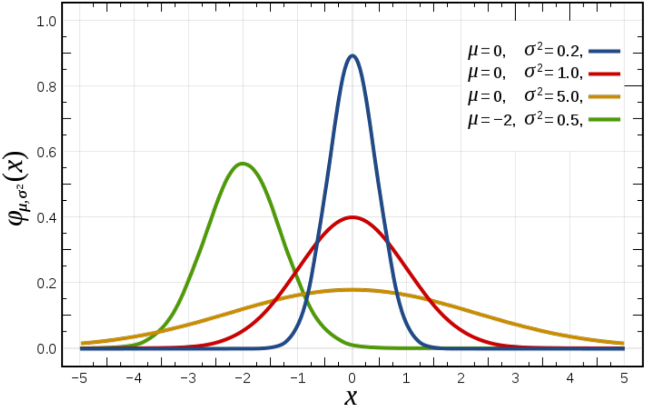
*Let $X_1, X_2, \ldots, X_n$ be independent random variables (with possibly different means and/or variances). Define*

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

*then the mean and the standard deviation of $T$ can be computed by*

- $E(T) = a_1 E(X_1) + a_2 E(X_2) + \ldots + a_n E(X_n)$
- $Var(T) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + \ldots + a_n^2 Var(X_n)$

# $\mathcal{N}(\mu, \sigma^2)$




$$E(X) = \mu, Var(X) = \sigma^2$$

# Linear combination of normal random variables

**Theorem**

*Let $X_1, X_2, \ldots, X_n$ be independent normal random variables (with possibly different means and/or variances). Then*

$$T = a_1 X_1 + a_2 X_2 + \ldots a_n X_n$$

*also follows the normal distribution with*

- $E(T) = a_1 E(X_1) + a_2 E(X_2) + \ldots + a_n E(X_n)$
- $Var(T) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + \ldots + a_n^2 Var(X_n)$

# Mean and variance of the sample mean

**Theorem**

*Given independent random samples* $X_1, X_2, ..., X_n$ *from a distribution with mean* $\mu$ *and standard deviation* $\sigma$*, the mean is modeled by a random variable* $\bar{X}$*,*

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

*Then*

$$E[\bar{X}] = \mu$$

*and*

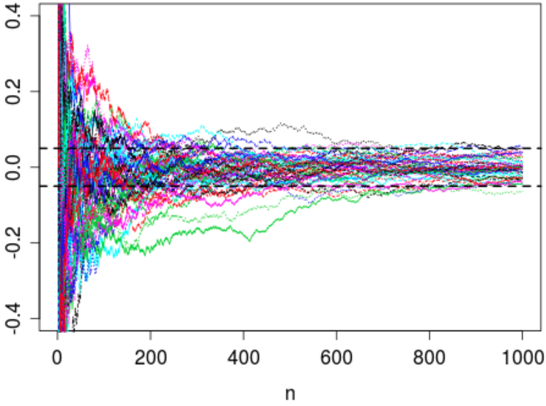
$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

# Law of large numbers

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$. Then

$$\bar{X} \to \mu$$

as $n$ approaches infinity



# The Central Limit Theorem

**Theorem**

*Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$. Then, in the limit when $n \to \infty$, the $\bar{X}$ follows normal distribution.*

*Recall that*

$$E[\bar{X}] = \mu, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

*this means we have*

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

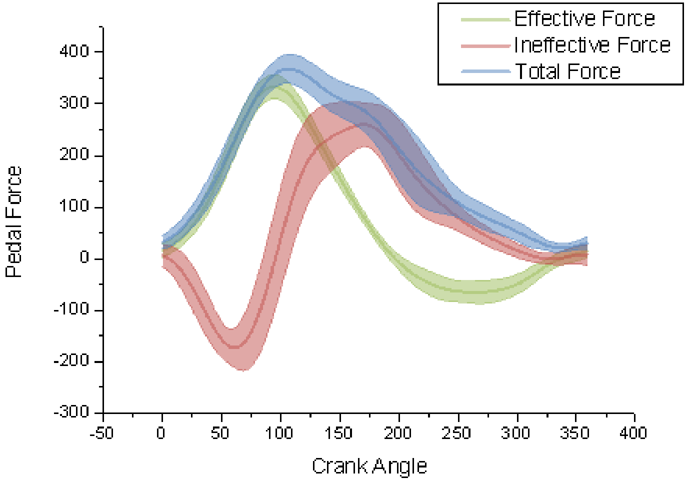
*follows the standard normal distribution.*

Rule of Thumb:

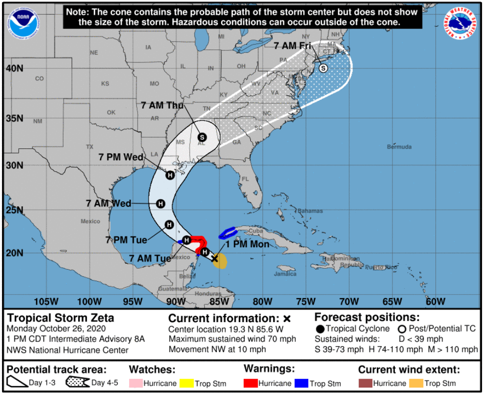
If $n > 30$, the Central Limit Theorem can be used for computation.

Confidence intervals of the population mean

# A good prediction comes with a range



# A good prediction comes with a range




A 70% confidence region of the path of a hurricane.

# Confidence

- Assume that you have been using an AI to predict the stock price of Microsoft every day in the last few years
- The prediction comes as a range, e.g., [295, 305]
- The algorithm, on average, is correct 95 out of 100 days
- Then we say that a prediction from this AI has a confidence of 95%

# Confidence interval: Example 1

## Problem

*Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean* $\mu$ *(unknown) and standard deviation* $\sigma = 0.85$*. A random sample of* $n = 25$ *specimens is selected with sample average* $\bar{X}$*.*
*Find a number c such that*

$$P\left[-c < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < c\right] = 0.95$$

# Confidence interval

- We have

$$P\left[-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

- Rearranging the inequalities gave

$$P\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

- This means that if you use

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

as a range to estimate $\mu$, then you are correct 95% of the time.

## Normal distribution with know $\sigma$

- Using

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

as a range to estimate $\mu$ is correct 95% of the time.

- If after observing $X_1 = x_1$, $X_2 = x_2, \ldots$, $X_n = x_n$, we compute the observed sample mean $\bar{x}$. Then

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a 95% confidence interval of $\mu$

# z-critical value

**NOTATION** $z_\alpha$ will denote the value on the measurement axis for which $\alpha$ of the area under the $z$ curve lies to the right of $z_\alpha$. (See Figure 4.19.)

For example, $z_{.10}$ captures upper-tail area .10 and $z_{.01}$ captures upper-tail area .01.

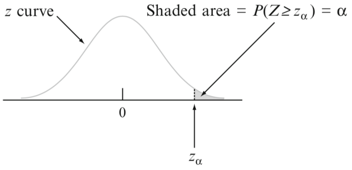



**Figure 4.19** $z_\alpha$ notation illustrated

Since $\alpha$ of the area under the standard normal curve lies to the right of $z_\alpha$, $1 - \alpha$ of the area lies to the left of $z_\alpha$. Thus $z_\alpha$ *is the* $100(1 - \alpha)$*th percentile of the standard normal distribution*. By symmetry the area under the standard normal curve to the left of $-z_\alpha$ is also $\alpha$. The $z_\alpha$'s are usually referred to as **$z$ critical values**. Table 4.1 lists the most useful standard normal percentiles and $z_\alpha$ values.

# $100(1-\alpha)\%$ confidence interval

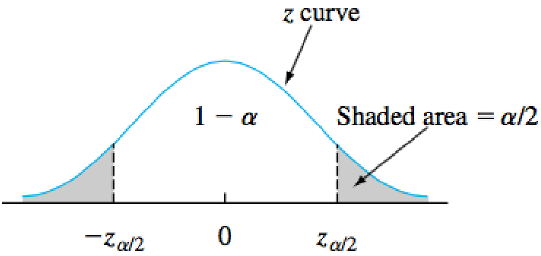



**Figure 8.4** $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$

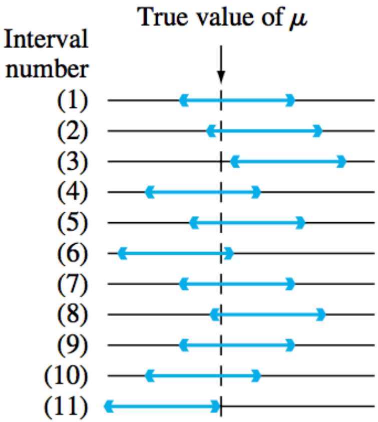
# $100(1-\alpha)\%$ confidence interval

A **$100(1-\alpha)\%$ confidence interval** for the mean $\mu$ of a normal population when the value of $\sigma$ is known is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{8.5}$$

or, equivalently, by $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

# Interpreting confidence intervals




95% confidence interval: If we repeat the experiment many times, the interval contains $\mu$ about 95% of the time

# Interpreting confidence intervals

- Writing
$$P[\mu \in (\bar{X} - 1.7, \bar{X} + 1.7)] = 95\%$$
is okay.
- If $\bar{x} = 2.7$, writing
$$P[\mu \in (1, 4.4)] = 95\%$$
is NOT correct.
- Saying $\mu \in (1, 4.4)$ with confidence level 95% is good.
- Saying "if we repeat the experiment many times, the interval contains $\mu$ about 95% of the time" is perfect.

# Example

## Example

Assume that the helium porosity (in percentage) of coal samples taken from any particular seam is normally distributed with true standard deviation $\sigma = .75$.

- Compute a 95% CI for the true average porosity of a certain seam if the average porosity for 20 specimens from the seam was 4.85.
- How large a sample size is necessary if the width of the 95% interval is to be .40?

# One-sided CIs (Confidence bounds)

# Example 1b: One-sided confidence interval

## Problem

*Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean* $\mu$ *(unknown) and standard deviation* $\sigma = 0.85$*. A random sample of* $n = 25$ *specimens is selected with sample average* $\bar{X}$*.*
*Find a number b such that*

$$P\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < b\right] = 0.95$$

# CIs vs. one-sided CIs

CIs:

- $100(1-\alpha)\%$ confidence

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

- 95% confidence

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

One-sided CIs:

- $100(1-\alpha)\%$ confidence

$$\left(-\infty, \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

- 95% confidence

$$\left(-\infty, \bar{x} + 1.64\frac{\sigma}{\sqrt{n}}\right)$$