03-Inferential Statistics
Notes and Examples

Inferential Statistics

- Suppose you **randomly sampled 10 people** from the population of men in Mumbai, between the ages of 21 and 35 years and computed the **mean height** of your sample.
- You would **not expect** your sample mean to be equal to the mean of all men in Mumbai. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly.
- Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.
- **Inferential statistics** concerns generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter.
- In this example, the **sample statistics** are the sample means and the **population parameter** is the population mean

Sampling Distribution

- A **sampling distribution of sample means** is a distribution obtained by using the means computed from random samples of a specific size taken from a population
- Knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean.
- The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean
- **Sampling error** is the difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population

Central Limit Theorem

- *Given a population with a finite mean μ and a finite nonzero variance $\sigma 2$, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of $\sigma 2/N$ as N, the sample size, increases*
- What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases.

Sampling Distribution

Mean and Standard Deviation of Sample Means

- As the sample size n increases, the shape of the distribution of the sample means taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution

- The mean of the sample means equals the population mean

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

- The standard deviation of the sample means is called the standard error of the mean

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

z - score

- The central limit theorem can be used to answer questions about sample means in the same manner that the normal distribution can be used to answer questions about individual values.
- A new formula must be used for the z values:

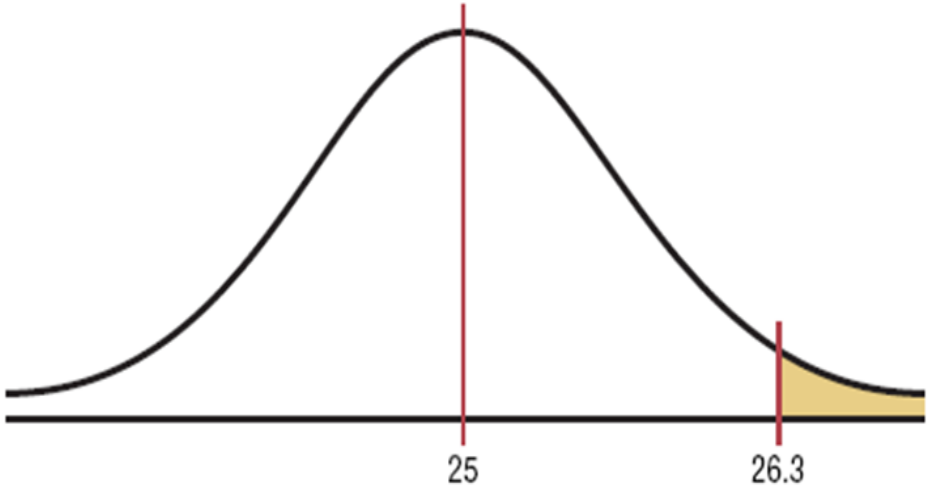
$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Example – Hours of Television

- A. C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week
- Assume the variable is normally distributed and the standard deviation is 3 hours
- If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours

- Since we are calculating probability for a sample mean

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{26.3 - 25}{3/\sqrt{20}} = 1.94$$



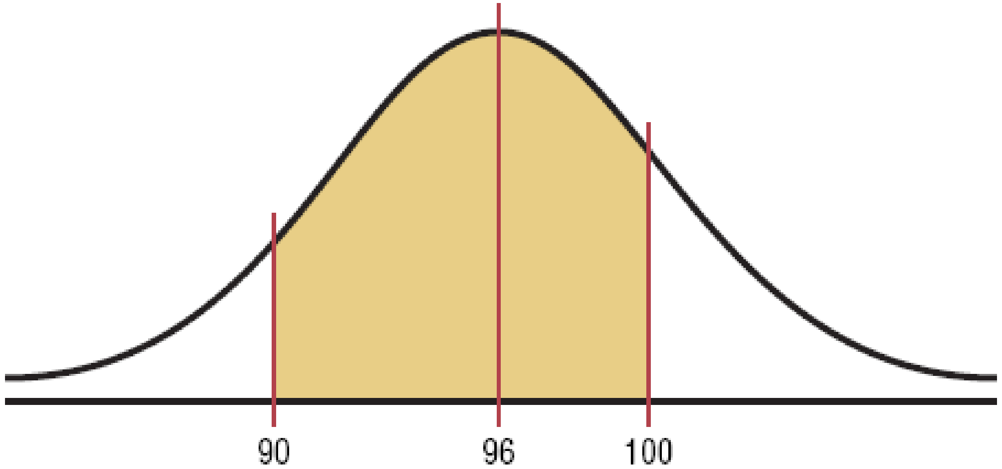
- For z=1.94, from the table, area under the curve=0.9738
- Thus the area is 1.0000 – 0.9738 = 0.0262.
- The probability of obtaining a sample mean larger than 26.3 hours is 2.62%.

Example – Vehicle Age

- The average age of a vehicle registered in the United States is 8 years, or 96 months
- Assume the standard deviation is 16 months
- If a random sample of 36 vehicles is selected, find the probability that the mean of their age is between 90 and 100 months.

- Since the sample is 30 or larger, the normality assumption is not necessary

$$z = \frac{90 - 96}{16/\sqrt{36}} = -2.25 \qquad z = \frac{100 - 96}{16/\sqrt{36}} = 1.50$$



- Table gives us areas 0.9332 and 0.0122, respectively.
- The desired area is 0.9332 - 0.0122 = 0.9210.
- Thus, the probability of obtaining a sample mean between 90 and 100 months is 92.1%.

Statistical Inference and Estimation

- Statistical inference is the act of generalizing from the data (“sample”) to a larger phenomenon (“population”) with calculated degree of certainty.
- The act of generalizing and deriving statistical judgments is the process of inference
- The two common forms of statistical inference are:
  - Estimation
  - Null hypothesis tests of significance (NHTS) – Hypothesis Testing
- There are two forms of estimation
  - Point estimation (maximally likely value for parameter)
  - Interval estimation (also called confidence interval for parameter)

Parameters and Point Estimates

- Both Estimation and Hypothesis Testing are used to infer parameters.
- A parameter is a statistical constant that describes a feature about a phenomena/population
- Examples of parameters include:
  - Expected value μ (also called “the population mean”)
  - Standard deviation σ (also called the “population standard deviation”)
  - Binomial probability of “success” p (also called “the population proportion”)
- Point estimates are single points that are used to infer parameters directly. For example
  - Sample mean x (“x bar”) is the point estimator of μ
  - Sample standard deviation s is the point estimator of σ
  - Sample proportion pˆ (“p hat”) is the point estimator of p

Population Parameters – Point Estimates

- One of the major applications of statistics is estimating population parameters from sample statistics
- For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium
- Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition.
- This value of 0.53 is called a **point estimate** of the population proportion. It is called a point estimate because the estimate consists of a single value or point

Interval Estimates

- An interval estimate of a parameter is an interval or a range of values used to estimate the parameter.
- Interval estimates are called **Confidence Intervals**
- A Confidence Interval will have a lower limit and an upper limit – and are called the confidence limits.

Confidence Intervals

- For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \pi < 0.60$.
- The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal.
- The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%.

Confidence Interval of the mean

- Confidence Interval = Observed mean ± Margin of error
- Margin of error = (Confidence Coefficient) × Standard Error of the mean

$$\boxed{\overline{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}$$

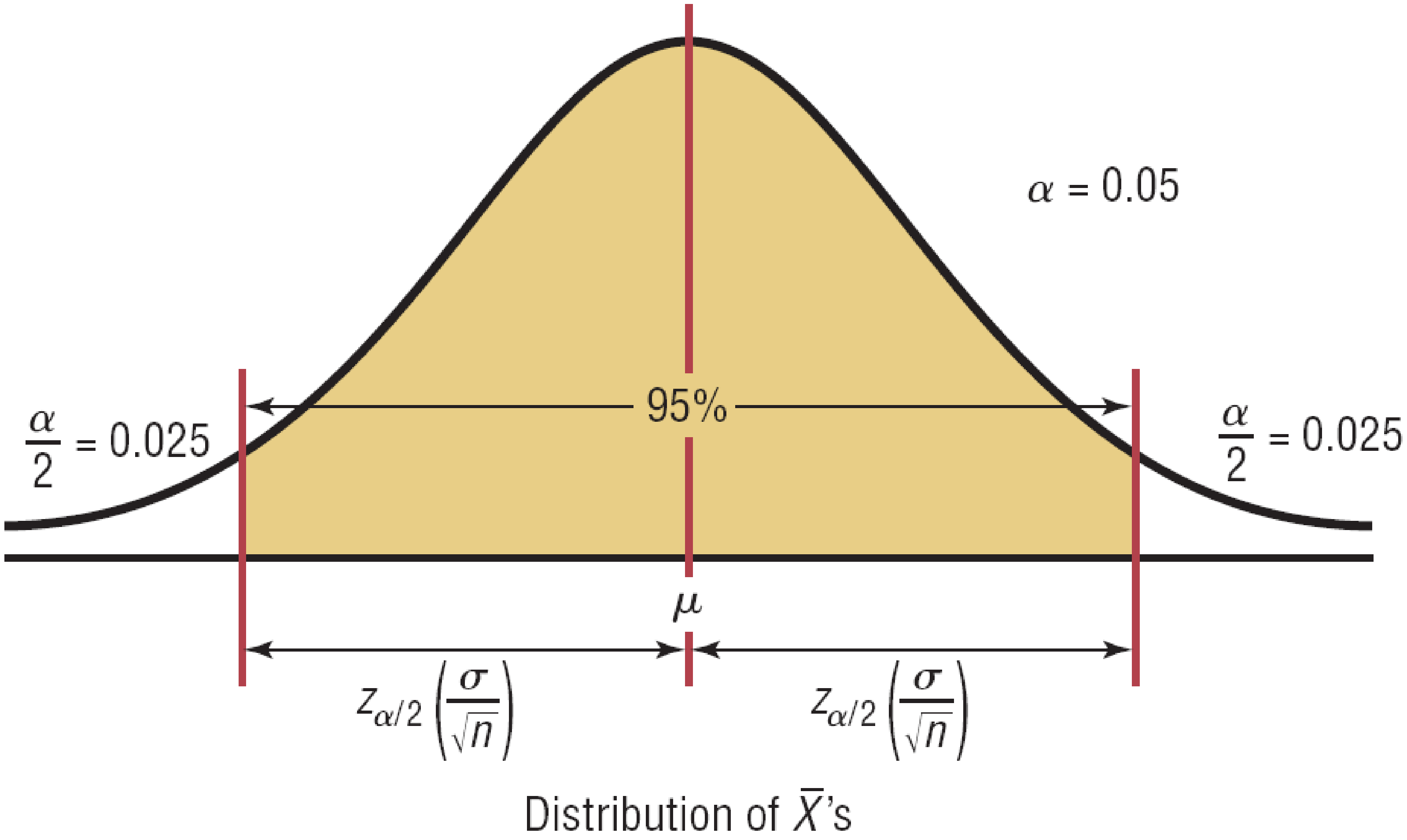
For a 90% confidence interval: $z_{\alpha/2} = 1.65$

For a 95% confidence interval: $z_{\alpha/2} = 1.96$

For a 99% confidence interval: $z_{\alpha/2} = 2.58$

- Relationship between α and the confidence level
  - α(alpha) represents the total area in both tails of the standard normal distribution curve.
  - Confidence level is the percentage equivalent of $1 - \alpha$
  - When 95% confidence interval is to be found , α = 0.05 since $1 - \alpha = 1 - 0.05 = 0.95$ (95%)
  - α is called the Significance level

95% confidence interval of the mean



$\alpha = 0.05$
$\frac{\alpha}{2} = 0.025$
95%
$\frac{\alpha}{2} = 0.025$
μ
$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$
$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$
Distribution of $\overline{X}$'s

Example: Age of Automobiles

- A survey of 50 adults found that the mean age of a person's primary vehicle is 5.6 years. Assuming the standard deviation of the population is 0.8 year, find the 99% confidence interval of the population mean.

  - $\overline{X} = 5.6$
  - $\sigma = 0.8$
  - For 99% Confidence Level, α= 0.005 on each side of the bell curve
  - z = -2.58 on left side and +2.58 on right side (from the table)
  - $Confidence\ Interval = \ 5.6 \pm 2.58\left(\frac{0.8}{\sqrt{50}}\right) = 5.6 \pm 0.4 = (5.2{,}6.0)$
  - One can be 99% confident that the mean age of all primary vehicles is between 5.2 and 6.0 years, based on a sample of 50 vehicles.

z-value in python

- import scipy.stats as st
- area=(1+conf_level)/2
- z=st.norm.ppf(area)