بسم الله الرحمن الرحيم

University of Khartoum

Faculty of Mathematical Sciences and Informatics

Statistics and Probability (S2013)

Assignment 4

(Normal Probability Distribution)

Content:

- ❖ Introduction.
- ❖ Curve, PDF, and MGF.
- ❖ Expectation and Variance.
- ❖ Standard Normal Distribution.
- ❖ Central Limit Theorem.
- ❖ References.

Written by:

Abobaker Ahmed Khidir Hassan 21-304

Computer Science

Normal Probability Distribution

Introduction

The normal distribution was introduced by the French mathematician Abraham DeMoivre in 1733, who used it to approximate probabilities associated with binomial random variables when the binomial parameter n is large. This result was later extended by Laplace and others and is now encompassed in a probability theorem known as the central limit theorem (It will be presented later at this assignment).

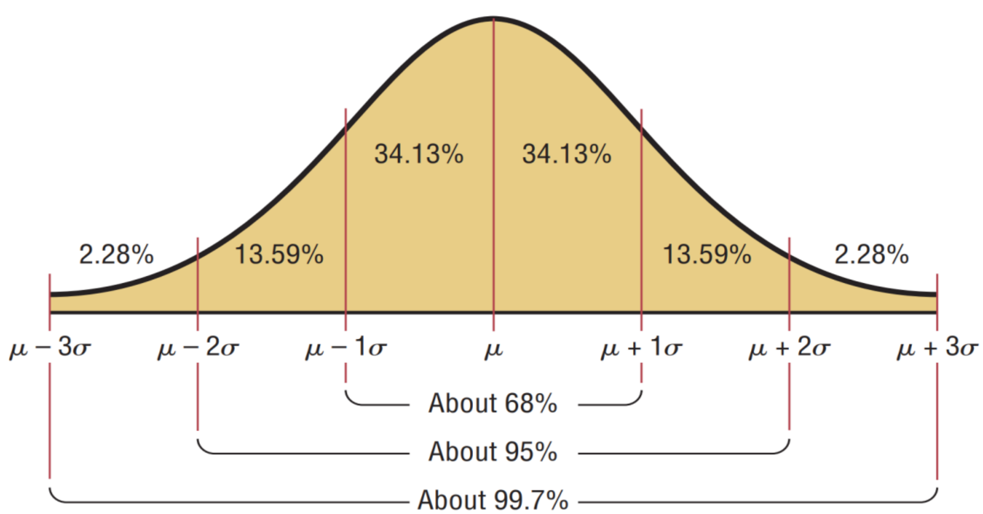
During the mid- to late 19th century, however, most statisticians started to believe that the majority of data sets would have histograms conforming to the Gaussian bell-shaped form. Indeed, it came to be accepted that it was "normal" for any well-behaved data set to follow this curve. As a result, following the lead of the British statistician Karl Pearson, people began referring to the Gaussian curve by calling it simply the normal curve.

Definition:

It is a continuous, symmetric, and bell-shaped distribution of a random variable.

Properties of the Curve of a Normal Distribution Variable

- **Continuous and Bell-shaped;** that is, it has the bell shape and there are no gaps or holes. For each value of X, there is a corresponding value of Y.
- **Symmetric about the mean,** its shape is the same on both sides of a vertical center line where located the mean, median and mode which are equal.
- **Unimodal** (i.e., it has only one mode).
- **Never touches the x-axis.** Theoretically, no matter how far in either direction the curve extends, it never meets the x-axis but it gets increasingly closer.
- **The total area under the curve is 1 or 100%, and it is classified as:**



Probability Density Function (PDF) of the Normal Distribution

We say a random variable X has a normal distribution if its pdf is:

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \quad for \quad -\infty < x < \infty}$$

Where:

$\boldsymbol{\mu} \equiv$ The mean of X.

$\boldsymbol{\sigma^2} \equiv$ The variance of X.

$\boldsymbol{e} \approx \mathbf{2.718}$ $\qquad\qquad$ $\boldsymbol{\pi} \approx \mathbf{3.14}$

We often write that X has a $\boldsymbol{N(\mu, \sigma^2)}$ distribution.

Moment Generating Function (MGF) of the Normal Distribution

$\boldsymbol{M_X(t) = E(e^{tX})}$

Assume $X = Z\sigma + \mu$ then $Z = \frac{X-\mu}{\sigma}$

$M_Z(t) = E(e^{tZ})$

$$= \int_{-\infty}^{\infty} e^{tZ}\left(\frac{1}{\sqrt{2\pi}}\right)\left(e^{\frac{-z^2}{2}}\right)dz = e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\right)\left(e^{\frac{-(z-t)^2}{2}}\right)dz$$

Assume $w = z - t \quad then \quad dw = dz$

$$= e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\right)\left(e^{\frac{-w^2}{2}}\right)dw = e^{\frac{t^2}{2}}$$

$\because M_Z(t) = e^{\frac{t^2}{2}} \;\; and \;\; X = Z\sigma + \mu$

$\therefore M_X(t) = E\left(e^{t(Z\sigma+\mu)}\right) = e^{\mu t} * E(e^{tZ\sigma}) = e^{\mu t} * e^{\frac{\sigma^2 t^2}{2}}$

$$\therefore \boldsymbol{M_X(t) = e^{\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}}$$

Expectation and Variance

Let X be a random variable normal distributed has a MGF $M_X(t) = e^{\left[\mu t+\frac{1}{2}\sigma^2 t^2\right]}$.

Then:

$$\boldsymbol{E(X) = M'_X(0)}$$

$$\frac{d}{dt}\left(M_X(t)\right) = \frac{d}{dt} e^{\mu t} * e^{\left[\frac{1}{2}\sigma^2 t^2\right]}\Big|_{t=0}$$

$$= e^{\left[\frac{1}{2}\sigma^2 t^2\right]} * \mu e^{\mu t} + e^{\mu t} * \sigma^2 t e^{\left[\frac{1}{2}\sigma^2 t^2\right]}\Big|_{t=0}$$

$$= \mu$$

$$\therefore \boldsymbol{E(x) = \mu}$$

$$\boldsymbol{var(X) = \ M''_X(0) - [M'_X(0)]^2}$$

$$\boldsymbol{M''_X(0) = \frac{d}{dt}\left(M'_X(t)\right)\Big|_{t=0}}$$

$$= \frac{d}{dt}(\mu + \sigma^2 t) e^{\left[\frac{1}{2}\sigma^2 t^2+\mu t\right]}\Big|_{t=0}$$

$$= \left[e^{\left[\frac{1}{2}\sigma^2 t^2+\mu t\right]} * \sigma^2 + (\mu + \sigma^2 t) * (\sigma^2 t + \mu) e^{\left[\frac{1}{2}\sigma^2 t^2+\mu t\right]}\right]\Big|_{t=0}$$

$$= \sigma^2 + \ \mu^2$$

$$\therefore \boldsymbol{Var(X) =} \sigma^2 + \ \mu^2 - \ \mu^2 \boldsymbol{= \sigma^2}$$

Standard Normal Distribution

It is a normal distribution with mean $(\mu = 0)$ and standard deviation $(\sigma = 1)$.

Standard Normal Distribution PDF:

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi}}\, e^{\frac{-z^2}{2}} \quad for \ -\infty < x < \infty}$$

Where: $\boldsymbol{z} = \frac{X-\mu}{\sigma}$

Standard Normal Distribution MGF:

$$\boldsymbol{M_Z\,(t) = E(e^{tZ})}$$

$$\int_{-\infty}^{\infty} e^{tZ}\left(\frac{1}{\sqrt{2\pi}}\right)\left(e^{\frac{-z^2}{2}}\right)dz = e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\right)\left(e^{\frac{-(z-t)^2}{2}}\right)dz$$

Assume $w = z - t \quad then \quad dw = dz$

$$= e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\right)\left(e^{\frac{-w^2}{2}}\right)dw$$

$$= e^{\frac{t^2}{2}}(1)$$

$$= e^{\frac{t^2}{2}}$$

$$\therefore\ \boldsymbol{M_Z\,(t) = e^{\frac{t^2}{2}}}$$

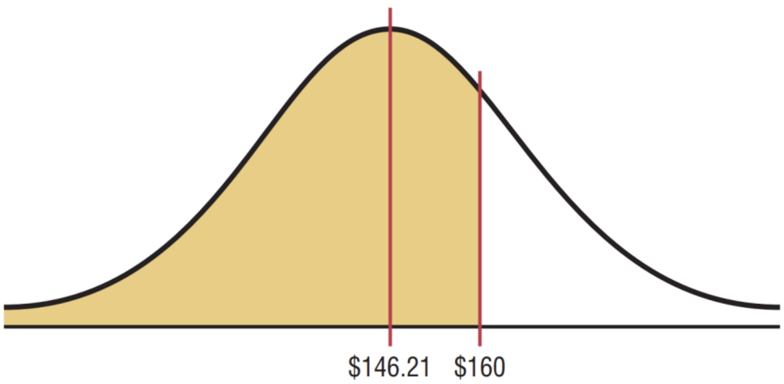
Examples

Example 1:

A survey found that women spend on average \$146.21 on beauty products during the summer months. Assume the standard deviation is \$29.44. Find the percentage of women who spend less than \$160.00. Assume the variable is normally distributed.

Solution:

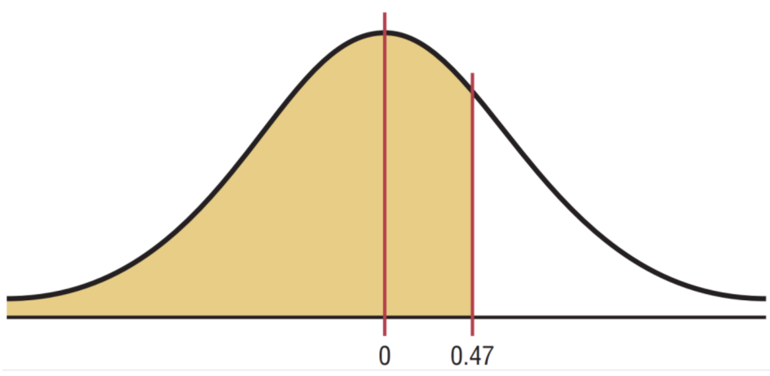
- Draw the figure and represent the area



- Find the z value corresponding to \$160.00

$$z = \frac{X - \mu}{\sigma} = \frac{160.00 - 146.21}{29.44} = 0.47$$

- Hence \$160.00 is 0.47 of a standard deviation above the mean of \$146.21,



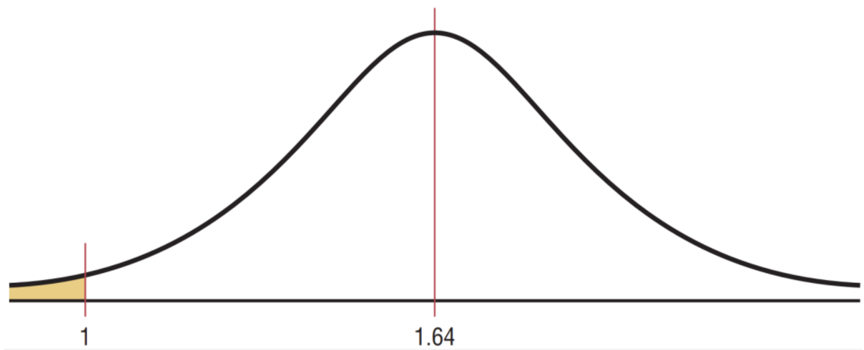
- Find the area, using Standard Normal Distribution Table. The area under the curve to the left of $z = 0.47$ is 0.6808.
- Therefore $\mathbf{0.6808}$, or $\mathbf{68.08}\%$, of the women spend less than \$160.00 on beauty products during the summer months.

Example 2:

Americans consume an average of 1.64 cups of coffee per day. Assume the variable is approximately normally distributed with a standard deviation of 0.24 cup. If 500 individuals are selected, approximately how many will drink less than 1 cup of coffee per day?

Solution:

- Draw the figure and represent the area



- Find the z value for 1

$$z = \frac{X - \mu}{\sigma} = \frac{1 - 1.64}{0.24} = -2.67$$

- Find the area to the left of z = 2.67. It is 0.0038.

- To find how many people drank less than 1 cup of coffee, multiply the sample size 500 by 0.0038 to get 1.9. Since we are asking about people, round the answer to 2 people.

  Hence, approximately 2 people will drink less than 1 cup of coffee a day.

Central Limit Theorem

The central limit theorem is one of the most remarkable results in probability theory. Loosely put, it states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped (that is, normal) curves.

Definition:

- As the sample size n increases without limit, 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. As previously shown, this distribution will have a mean $\mu_{\overline{X}}$ and a standard deviation $\sigma_{\overline{X}}$.
- In another words, if $X_1,\ X_2,\ X_3,\ \ldots..,\ X_n$ is a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2, then the distribution of their average $\overline{X}$ can be approximated by a normal distribution $N(\mu, \frac{\sigma^2}{n})$.

It's important to remember two things when you use the central limit theorem:

- When the original variable is normally distributed, the distribution of the sample means will be normally distributed, for any sample size n.
- When the distribution of the original variable might not be normal, a sample size of 30 or more is needed to use a normal distribution to approximate the distribution of the sample means. The larger the sample, the better the approximation will be.

Theoretical Formulas:

$$P\left\{\frac{X_1 + X_2 + X_3 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{\frac{-z^2}{2}}\, dx \quad as \quad n \to \infty$$

Where:

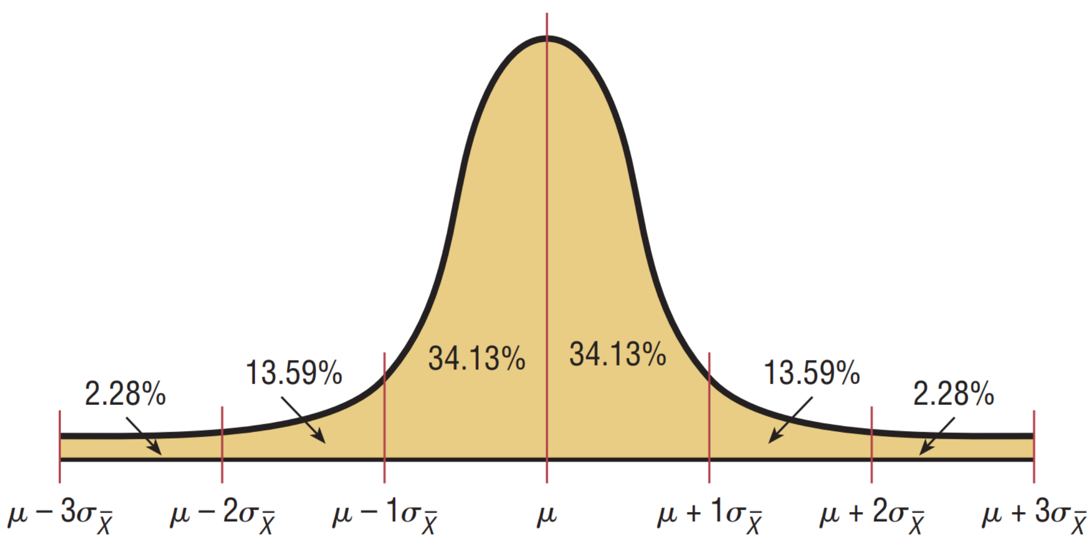
- $X_1,\ X_2,\ X_3,\ \ldots,\ X_n$ is a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2(standard deviation σ).
- n is the number variables in the sequence.
- $z = \frac{\overline{X}_i - \mu}{\sigma/\sqrt{n}}$ Where $\overline{X}_i$ are the sample means where $i = 1,2,3, \ldots, n$.
- $e \approx 2.718$ $\pi \approx 3.14$

$$\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}}^2 = \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}} = \sqrt{\sigma_{\overline{X}}^2} = \frac{\sigma}{\sqrt{n}}$$

Where:

- $\mu_{\overline{X}}$ is the mean of each variable in the sequence.
- $\sigma_{\overline{X}}^2$ is the variance of each variable in the sequence.
- $\sigma_{\overline{X}}$ is the standard deviation of each variable in the sequence.
- n is the number of variables in the sequence.

How the standard normal distribution is used to answer questions about sample means



Examples

Example 1:

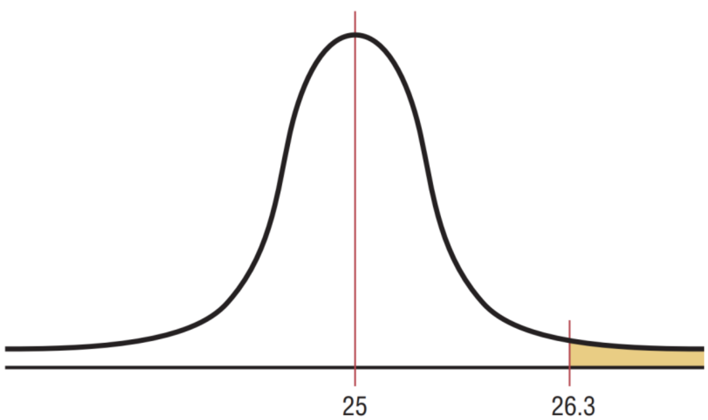
A. C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean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

Solution:

Since the variable is approximately normally distributed, the distribution of sample means will be approximately normal, with a mean of 25. The standard deviation of the sample means is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{20}} = 0.671$$

The distribution of the means is shown in Figure 6–32, with the appropriate area shaded.



The z value is

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{26.3 - 25}{3/\sqrt{20}} = \frac{1.3}{0.671} = 1.94$$

The area to the right of 1.94 is $1.000 - 0.9738 = 0.0262$, or 2.62%.
One can conclude that the probability of obtaining a sample mean larger than 26.3 hours is 2.62% [i.e., $P(\overline{X} > 26.3) = 2.62\%$].

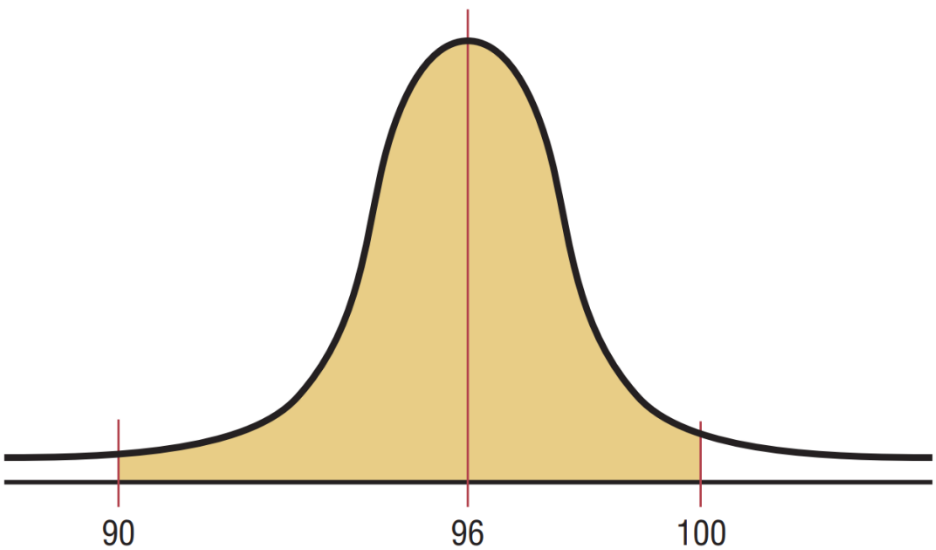
Example 2:

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.

Solution:

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

The desired area is



The two z values are

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$z_1 = \frac{90 - 96}{16/\sqrt{36}} = -2.25$$

$$z_2 = \frac{100 - 96}{16/\sqrt{36}} = 1.50$$

To find the area between the two z values of 2.25 and 1.50, look up the corresponding area in Standard Normal Distribution Table and subtract one from the other. The area for $z_1 = 2.25$ is 0.0122, and the area for $z_2 = 1.50$ is 0.9332. Hence the area between the two values is $0.9332 - 0.0122 = 0.9210$, or 92.1%.

Hence, the probability of obtaining a sample mean between 90 and 100 months is 92.1%; that is, $P(90 < \overline{X} < 100) = 92.1\%$.

References

- Bluman. Elementary Statistics A Step By Step Approach 8th Edition.
- Sheldon Ross. A First Course in Probability 9th Edition.
- Robert V. Hogg, Joseph W. McKean, Allen T. Craig. Introduction to Mathematical Statistics 8th Edition.
- Hogg Craig. Introduction to Mathematical Statistics 4th edition.
- Daniel Zwillinger. CRC Standard Probability and Statistics Tables and Formulae.
- An Introduction to Statistics and Probability Lectures Summary by Ms. Sara Faisal Ahmed.
- Old Researches.

The End

16th – July – 2024