Grade 12 Math of Data Management

MDM4U

Qinghao Hu

December 10, 2025

Contents

Chapter 2

Normal Distribution

2.1 Normal Distributions

2.1.1 Definitions

Definition 2.1.1 (Normal Distribution)
The properties of normal distributions can be described as this:

1. *Graphical displays are mound-shaped histograms that are also Bell-Shaped*
2. *The MEAN, MEDIAN and MODE fall on the line of symmetry*
3. *Denoted by the notation $X \sim N(\mu, \sigma^2)$ where μ is the mean and σ is the standard deviation*
4. *Modelled by a NORMAL CURVE where the curve passes through the midpoint of the top of each bar of the histogram of the data.*
5. *The stadnard normal curve has a mean of 0 and a standard deviation of 1*

Empirical Rule

$$P(\mu - \sigma < X < \mu + \sigma) = 0.68$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

2.2 Normal Modelling

In the test, you definitely have no graphing calculator. How can hack can we solve for the probablity of Normal Distribution between two region. At this time, we can use Z-score

Theorem 2.2.1
Assume we have a normal distribution, $N \sim X(\mu, \sigma^2)$, we want to solve for

$$P(x \leq X) \tag{2.1}$$

First we need to solve for z-score of X

$$Z = \frac{X - \mu}{\sigma}$$

After we have the z-score, we can rewrite 2.1:

$$P(z \leq Z)$$

Then use your z-score table to find the value correspond to this z-score.

Theorem 2.2.2
How to use z-score to solve for some special P:

Case 1:

$$P(X_1 < x < X_2) = P(x < X_2) - P(x < X_1)$$
$$P(X_1 > x) = 1 - P(x < X_1)$$

Case 2:

$$P(x > X) = 1 - P(x < X)$$

For discrete data, always perform $+/-0.5$

2.3 Approximate Binomial

At certain situation, we can use a normal distribution to approximate a Binomial distribution

Theorem 2.3.1
If X is a binomial random distribution of n independent trials, each with probability of success p, and if

$$np > 5$$
$$n(1-p) > 5$$

then the binomial random variable can be approximated by a normal distribution with

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

Theorem 2.3.2
The distribution of a discrete random variable X is a binomial probability distribution if:

1. *The probability experiment is repeated a FIXED number of times (In the example, $n = 3$)*
2. *The outcome of each trial can be CATEGORIZED into successful or failed outcomes*
3. *The random varialbe X represents the number of successes*
4. *Each trial of the experiment is Independent*

2.4 Approximate Hypergeometric

If the population size N is fairly large, and the sample size is relatively small, we can use a Normal Distribution to simulate Hypergeometric

Theorem 2.4.1
The distribution of a discrete random variable X is a hypergeometric probability distribution if:

1. *the probability experiment is repeated a **fixed** number of times*
2. *the outcome of each trial can be categorized into **success** or failed outcomes*
3. *the random varaible X represents the number of **successes***
4. *each trial of experiment is **dependant***

Theorem 2.4.2
Given that N is the population size, n is the sample size, and a is the number of elements in the population with the particular characteristic, if

$$\frac{n}{N} < \tfrac{1}{10}$$
$$n\frac{a}{N} > 5$$
$$n(\frac{N-a}{N}) > 5$$

then the hypergeometric random variable can be approximated by a normal distribution with

$$\mu = \frac{an}{N}$$
$$\sigma = \sqrt{n(\frac{a}{N})(\frac{N-a}{N})(\frac{N-n}{N-1})}$$

2.5 Sampling Distributions

Theorem 2.5.1
If a population has a mean of μ and a standard deviation of σ, then the means obtained from repeatedly sampling the population are normally distributed with a mean of

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

and a standard deviation of

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where n is the size of each repeated sample $n \geq 30$ for population data that is not normally distributed

Definition 2.5.2 (Point Estimation)
The process of using a sample mean $\bar{x}$ to estimate the population mean μ. The downside to point estimates is that we have no way of knowing if the sample statistics are actually close to the population summary value. It could be that, because of variability, the sample mean is 'way off' from the population mean. For that reason, interval estimation is more preferred.

Definition 2.5.3 (Confidence Interval)
An interval estimate accompanied by a confidence level. The confidence level indicates the percent of confidence that the given interval size will, in the "long run", capture the population summary value.

2.6 Confidence Intervals

Theorem 2.6.1
Repeated sampling from a normally distributed population produces a normally distributed sample means. Hence, the probability of observing a single sample mean, $\bar{x}$, within $z\sigma_{\bar{x}}$ of $\mu_{\bar{x}}$ is $1-\alpha$

As a result, we can get:

$$P(\mu_{\bar{x}} - z\sigma_{\bar{x}} < x < \mu_{\bar{x}} + z\sigma_{\bar{x}}) = 1 - \alpha$$

After rearrange, for a significance level of α, we can get this:

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu_{\bar{x}} < \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

Therefore, the boundaries for the interval estimate is $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$

Definition 2.6.2 (Porportion)
A ratio of the population that shares the same characteristic.

Theorem 2.6.3
If one wants to find the proportion of a population that have a particular characteristic, then

$$\mu_{\bar{p}} = \mu \qquad \sigma_{\bar{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

p and $\hat{p}$ are the proportion of the population and a sample, respectively that have the characteristic.
$\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ are the mean and the standard deviation of the distribution of the sampling proportions.

As well,

$$P(\mu_{\hat{p}} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \mu_{\hat{p}} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

The boundaries for the interval estimate is defined as:

$$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$