The YB for Medical Specializations

Lectures in Advanced Biostatistics

For YBPH program Students

2024-2025

2024-2025

Prof. Dr. Salem Saleh

2024-2025

Normal Distribution, Standard N.D, Z-Score, Critical value, Sig. level, CI

The normal distribution: is one of the most important and widely used probability distributions in medical sciences and biostatistics. It has unique properties that make it suitable for describing many natural and medical phenomena. In this lecture, we will explain the normal distribution, its key characteristics, the critical value, and the standard score (Z-score).

Definition.

The normal distribution is a continuous probability distribution that is symmetric and bell-shaped. It is mathematically defined by the **mean** μ and the **standard deviation** σ.

Normal Distribution Formula

$$\boxed{f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

- μ: The mean of the distribution
- σ: The standard deviation
- x: The variable for which we are calculating the probability.

Characteristics of the Normal Distribution:

1. Symmetry:
   - The normal distribution curve is symmetric around the mean, meaning that half of the data lies to the right of the mean and the other half lies to the left.
2. Bell-shaped Curve:
   - The curve resembles the shape of a bell, with most of the data clustered around the mean, and probabilities decrease gradually as we move further away from the mean.
3. Equal Mean, Median, and Mode:
   - In a normal distribution, the mean (μ), median, and mode are equal and located at the center of the distribution.

Proportion of Data Around the Mean:

- About 68% of the values lie within one standard deviation from the mean ($\mu \pm \sigma$).
- About 95% of the values lie within two standard deviations from the mean ($\mu \pm 2\sigma$).
- About 99.7% of the values lie within three standard deviations from the mean ($\mu \pm 3\sigma$).
- This is known as the 68-95-99.7 Rule.

The Standard Normal Distribution

The **standard normal distribution** is a special case of the normal distribution where the **mean** $\mu = 0$ and the **standard deviation** $\sigma = 1$.

The Standard Score (Z-Score)

The **Z-score** represents the number of standard deviations a value is from the mean. It allows for the comparison of different values across different distributions by standardizing the data.

The formula for calculating the Z-score is:

$$\boxed{Z = \frac{x - \mu}{\sigma}}$$

- Z: The Z-score
- x: The original value
- μ: The mean
- σ: The standard deviation

Importance of Z-Score

- Z-scores allow us to compare values from different distributions and determine how far a value is from the mean.
- They help in calculating probabilities using the standard normal distribution, where Z-scores are used in **Z-tables**.

Example of Z-Score:

If the mean height of a group of students is 170 cm with a standard deviation of 10 cm, what is the Z-score for a student who is 185 cm tall?

$$\boxed{Z = \frac{185 - 170}{10} = \frac{15}{10} = 1.5}$$

- A Z-score of 1.5 means that this student's height is 1.5 standard deviations above the mean.

The following is a graphical representation of the normal distribution curve:

[redacted] [redacted]

```
-3σ         -2σ         -1σ          0          +1σ         +2σ         +3σ
|-----------|-----------|-----------|-----------|-----------|-----------|
Z = -3      Z = -2      Z = -1      Z = 0      Z = +1      Z = +2      Z = +3
```

- **Z = 0** represents the mean of the standard normal distribution.
- Each Z value corresponds to the correct position on the horizontal axis.
- Each unit on the axis represents one standard deviation (σ).

This diagram clearly shows the distribution of Z-scores in the standard normal distribution, centered around the mean $Z = 0$.

What is a Critical Value?

A **critical value** is a point on the normal distribution curve used to determine the rejection region in hypothesis testing. Critical values depend on the **confidence level** or **significance level** (α) of the analysis.

- **Confidence level** ($1 - \alpha$): Indicates how confident you are that the result is not due to chance.
- **Significance level** (α): Represents the probability of making a Type I error (i.e., rejecting the null hypothesis when it is true).

Examples of Critical Values:

- At a significance level of $\alpha = 0.05$, the critical value for the standard normal distribution is $Z = \pm 1.96$.
- At a significance level of $\alpha = 0.01$, the critical value is $Z = \pm 2.58$.

The diagram below represents the critical values for a two-tailed test with $\alpha = 0.05$ (95% confidence level):

[redacted] [redacted]

```
Rejection      Acceptance      Rejection
 region          region          region
α/2 = 0.025                    α/2 = 0.025
|--------|--------|--------|--------|
       -1.96      0       +1.96
```

- The **acceptance region** is between -1.96 and $+1.96$.
- The **rejection regions** are the tails beyond -1.96 and $+1.96$, each with a probability of $\alpha/2 = 0.025$.
- If the Z-score falls in the rejection region (less than -1.96 or greater than $+1.96$), we reject the null hypothesis.

Practical Examples of Normal Distribution in Medicine

Example 1: Blood Pressure Distribution Among Patients

Suppose the distribution of blood pressure for a group of patients follows a normal distribution with a mean of 120 mmHg and a standard deviation of 15 mmHg. If a patient has a blood pressure of 145 mmHg, what is the Z-score for this patient?

Since,

$$\boxed{Z = \frac{145 - 120}{15} = \frac{25}{15} = 1.67}$$

- A Z-score of 1.67 means that this patient's blood pressure is 1.67 standard deviations above the mean.

Example 2: Blood Sugar Levels

If blood sugar levels follow a normal distribution with a mean of 100 mg/dL and a standard deviation of 20 mg/dL, what percentage of people have blood sugar levels below 80 mg/dL?

First, we calculate the Z-score:

$$\boxed{Z = \frac{80 - 100}{20} = \frac{-20}{20} = -1}$$

- Using the standard normal distribution table, the probability corresponding to $Z = -1$ is about 0.1587.
- This means that about 15.87% of people have blood sugar levels below 80 mg/dL.

Confidence Intervals

Confidence intervals (CIs) are a fundamental concept in statistics and are widely used in biostatistics to estimate the range within which a population parameter (such as a mean or proportion, odds ratio, or risk ratio) is likely to fall. Understanding and interpreting confidence intervals is crucial for making inferences from sample data to the broader population.

What is a Confidence Interval?

A confidence interval is a range of values, derived from sample data, that is likely to contain the true value of an unknown population parameter. Instead of giving a single point estimate (like a sample mean), a confidence interval provides a range that, with a certain level of confidence, includes the true parameter value.

Confidence Level

The confidence level is the percentage of all possible samples that can be expected to include the true population parameter. Common confidence levels are 90%, 95%, and 99%.

- **95% Confidence Interval:** This means that if we were to take 100 different samples and compute a confidence interval for each sample, we would expect 95 of the intervals to contain the true population parameter.

Formula

- For a population mean μ when the population standard deviation σ is known:

$$\boxed{\text{CI} = \bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}}$$

Medical Example

Suppose a sample of 30 patients has an average systolic blood pressure of 140 mmHg, with a sample standard deviation of 15 mmHg. Construct a 95% confidence interval for the population mean systolic blood pressure.

- Sample mean ($\bar{x}$) = 140 mmHg
- Sample standard deviation (s) = 15 mmHg
- Sample size (n) = 30
- Degrees of freedom (df) = 29
- Critical value ($t_{0.025,29}$) $\approx$ 2.045 (from t-distribution table)

$$\boxed{\mathrm{CI} = 140 \pm 2.045 \times \frac{15}{\sqrt{30}} = 140 \pm 5.6 = [134.4, 145.6] \text{ mmHg}}$$

This means we are 95% confident that the true mean systolic blood pressure in the population is between 134.4 and 145.6 mmHg.

Risks and Odds

In this section we introduce some important statistical measurements such as:

- Absolute risk reduction,
- number needed to treat,
- Relative risk (or risk ratio),
- Odds ratios

Which are measures helpful for comparing probability values and measuring risk.

- If we Summarize the Results of a Prospective Study as follows:

	Disease	No Disease
Treatment	a	b
Placebo	c	d

We can define the following important concepts:

- **Absolute risk reduction** $= \left| \dfrac{a}{a+b} - \dfrac{c}{c+d} \right|$
- **Number Needed to Treat (NNT)** $= \dfrac{1}{\text{Absolute Risk Reduction}}$

$$\boxed{\text{Relative risk (RR)}} = \frac{p_t}{p_c} = \frac{\dfrac{a}{a+b}}{\dfrac{c}{c+d}}$$

(for prospective only)

$$\boxed{\text{Odds ratio (OR)}} = \frac{ad}{bc}$$

* **Odds Ratio (OR)**

The odds ratio is a measure of association between an exposure and an outcome. It compares the odds of the outcome occurring in the exposed group to the odds of it occurring in the non-exposed group

Confidence Interval for Odds Ratio

The confidence interval for the odds ratio can be calculated using the following formula:

$$\boxed{\mathrm{CI} = \exp\left[\ln(\mathrm{OR}) \pm Z\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right]}$$

Where:

- $\ln(\mathrm{OR})$ is the natural logarithm of the odds ratio.
- Z is the Z-value corresponding to the desired confidence level (e.g., 1.96 for 95% confidence).

Example: Confidence Interval for Odds Ratio

Consider a study examining the association between smoking and lung cancer. The data are summarized in the following 2x2 table:

	Lung Cancer	No Lung Cancer
Smokers	30	20
Non-Smokers	10	40

Table 2: Association between Smoking and Lung Cancer

Here,

$$a = 30, \quad b = 20, \quad c = 10, \quad d = 40$$

The odds ratio is:

$$\mathrm{OR} = \frac{30 \times 40}{20 \times 10} = 6.0$$

Now, we calculate the 95% confidence interval:

$$\ln(\mathrm{OR}) = \ln(6.0) = 1.79176$$

$$\mathrm{SE} = \sqrt{\frac{1}{30} + \frac{1}{20} + \frac{1}{10} + \frac{1}{40}} = 0.45644$$

$$\mathrm{CI} = \exp\left[1.79176 \pm 1.96 \times 0.55902\right] = \exp\left[1.79176 \pm 1.09567\right]$$

$$\mathrm{CI} = \exp\left[0.89714, 2.6864\right] = \left[2.4526, 14.678\right]$$

Thus, the odds ratio's 95% confidence interval is [2.4526, 14.678].

Then, find for the above example the measures:

- ARR,
- RR,
- NNT,
- OR

Exercise

In Exercises 13-16, use the data in the accompanying table that summarizes results from a clinical trial of atorvastatin

	Infection	No Infection
Atorvastatin (10 mg)	89	774
Placebo	27	243

13. Absolute Risk Reduction

a. What is the probability of infection in the atorvastatin treatment group?

b. What is the probability of infection in the placebo group?

c. Find the value of the absolute risk reduction for infection in the placebo group and the atorvastatin treatment group. Write a brief statement interpreting the result.

14. Number Needed to Treat Calculate the number needed to treat and interpret the result.

15. Odds For those who were treated with atorvastatin, find the odds in favor of an infection. Also find the odds in favor of an infection for those given a placebo. Is there much of a difference between these two results?

16. Odds Ratio and Relative Risk Find the odds ratio and relative risk for an infection in the group treated with atorvastatin compared to the placebo group. Based on this result, does atorvastatin appear to increase the risk of an infection? Why or why not?

17. Odds Ratio and Relative Risk In a clinical trial of 2103 subjects treated with Nasonex (mometasone), 26 reported headaches. In a control group of 1671 subjects given a placebo, 22 reported headaches. Find the relative risk and odds ratio for the headache data. What do the results suggest about the risk of a headache from the Nasonex treatment?

18. Design of Experiments You would like to conduct a study to determine the effectiveness of seat belts in saving lives in car crashes.

a. What would be wrong with randomly selecting 2000 drivers, then randomly assigning half of them to a group that uses seat belts and another group that does not wear seat belts?

b. If 2000 drivers are randomly selected and separated into two groups according to whether they use seat belts, what is a practical obstacle in conducting a prospective study of the effectiveness of seat belts in car crashes?