STAT 2910: Statistics for the Sciences **InterSummer 2022**

Lecture Notes 2

Name: Edward Nafornita **Professor:** Sudhir Paul

Contents

1 Normal Probability Distribution

1.1 Continuous Random Variables

- Continuous Random Variables: can assume the infinitely many values corresponding to points on a line interval.
  - Examples: Heights, weights, length of life of a particular product, experimental laboratory error.
- A smooth curve describes the probability distribution of a continuous random variable.
- The depth or density of the probability, which varies with x, may be descirbed by a mathematical formula $f(x)$, called the **probability distribution** or **probability density function** for the random variable x.

1.2 Properties of Continuous Probability Distribution

1 The area under the curve is equal to 1.

2 $P(a \leq x \leq b) =$ **area under the curve** between a and b.

3 There is no probability attached to any single value of x. That is, $P(x = a) = 0$.

Continuous Proabability Distributions

- There are many different types of continuous random variables.
- We try to pick a model that
  - Fits the data well
  - Allows us to make the best possible inferences using the data.
- One important continuous random variable is the **normal random variable**.

1.3 Normal Distribution

- The formula that generates the normal probability distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1}$$

  - For $-\infty < x < \infty$
  - μ and σ are the population mean and standard deviation.
- The shape and location of the normal curve changes as the mean and standard deviation change.

1.4 Standard Normal Distribution

- To find $P(a < x < b)$, we need to find the area under the appropriate normal curve.
- To simplify the tabulation of these areas, we **standardize** each value of x by expressing it as a z-score, the number of standard deviations σ it lies from the mean μ.

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Properties of the Standard Normal Distribution

1 Mean = 0; Standard Deviation = 1

2 When $x = \mu$, $z = 0$

3 Symmetric about $z = 0$

4 Values of z to the left of center are neagtive

5 Values of z to the right of center are positive

6 Total area under the curve is 1.

1.4.1 Solve Problems Using Statistical Table 3

1 To find an area to the left of a z-value, find the area directly from the table.

2 To find an area to the right of a z-value, find the area in Table 3 and subtract from 1.

3 To find the area between two values of z, find the two areas in Table 3, and subtract one from the other.

Recall: Empirical Rule states that "approximately 95% of the measurements lie within 2 standard deviations of the mean."

1.4.2 Finding Probabilities for the General Normal Random Variable

1 To find an area for a normal random variable x with mean μ and standard deviation σ, standardize or rescale the interval in terms of z-values.

2 Find the appropriate area using Table 3.

1.5 Normal Approximation to the Binomial

- We can calculate binomial probabilities using:
  - The binomial formula
  - The cumulative binomial tables
- When n is large, and p is not too close to zero or one, areas under the normal curve with mean np and variance npq can be used to approximate binomial probabilities.

1.5.1 Approximating the Binomial

- Make sure to include the entire rectangle for the values of x in the interval of interest. This is called the **continuity correction**.
- Standardize the values of x using:

$$z = \frac{x - np}{\sqrt{npq}} \tag{3}$$

- Make sure that np and nq are both greater than 5 to avoid inaccurate approximations.

2 Sampling Distributions

- Parameters are numerical descriptive measures for populations. The values of the parameters are generally unknown.
  - Examples 1: Suppose height of the students of U. of W. is normally distributed with μ and σ.
    - The quantities μ and σ are parameters. These are fixed but unknown.
  - Example 2: Suppose a proportion p of people in Windsor have heart disease. Suppose you take a sample of n people from Windsor. Let x = number of people in the sample who have heart disease. Then, x has a binomial distribution with index n probability p.
    - The quantity p is the parameter which is fixed but unknown.

2.1 Sampling

Since it is time consuming and expensive to take all values of the population to obtain values of the population quantities such as μ, σ and p we can take a sample to estimate these quantities.

- Example: Suppose that the mean height of the students of U. of W. is $\mu = 5.6$ft. Now suppose that you take 10 students from a sample of 25 students and obtained their values is as follows;
  $\bar{x} : 5.2, 5.7, 5.4, 5.8, 5.6, 5.9, 5.1, 5.6, 5.55, 5.67.$
- This example shows that the sample mean is not fixed. It is known as a random variable and it depends on which 25 students are in the sample.
- Since $\bar{x}$ is a random variable, it has a probability distribution.
- The quantities $\bar{x}$ and s calculated from the sample to estimate the population quantities μ and σ are called statistics. Thus the sample mean $\bar{x}$ is a statistic obtained from the sample to estimate the population mean μ

Sample Variance: s^2 is a statistic obtained from the sample to estimate the population variance σ^2.
Sample Proportion: $\hat{p}$ is a statistic obtained from the sample to estimate the population proportion p.
Statistics vary from samle to sample and hence are random variables. The probability distributions for statistics are called **sampling distributions**. In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

2.1.1 Sampling Distributions for Statistics

1 Approximated with simulation techniques

2 Derived using mathematical theorems

- Example: The Central Limit Theorem is one such example.

> **Central Limit Theorem:** If random samples of n observations are drawn from a non-normal population with finite μ and standard deviation σ, then, when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The approximation becomes more accurate as n becomes large.

2.1.2 Importance of Sampling Distributions for Statistics

- The **Central Limit Theorem** also implies that the sum of n measurements is approximately normal with mean $n\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- Man statistics that are used for statistical inference are **sums** or **averages** of sample measurements.
- When n is large, these statistics will have approximately **normal** distributions.

2.1.3 Definition of Large

- If the sample is from a **normal population**, then the sampling distribution of $\bar{x}$ will also be normal, no matter the sample size.
- When the sample population is approximately **symmetric**, the distribution becomes approximately normal for relatively small values of n.
- When the sample population is **skewed**, the sample size must be **at least 30** before the sampling distribution of $\bar{x}$ becomes approximately normal.

2.2 Sampling Distribution of the Sample Mean

- A random sample of size n is selected from a population with mean μ and standard deviation σ.
- The sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- If the original population is **normal**, the sampling distribution will be normal for any sample size.
- If the original population is **non-normal**, the sampling distribution will be normal when n is large.

Note: Standard deviation of $\bar{x}$ is sometimes called the Standard Error (SE).

2.2.1 Finding Probabilities for the Sample Mean

- If the sampling distribution of $\bar{x}$ is normal or approximately normal, standardize or rescale the interval of interest in terms of

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Find the appropriate area using Table 3.

Example: A random sample of size $n = 16$ from a normal distribution with $\mu = 10$ and $\sigma = 8$. Then $P(\bar{x} > 12) = P(z > \frac{12-10}{\frac{8}{\sqrt{16}}}) = P(z > 1) = 1 - 0.8413 = 0.1587$.

Example 2

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of 0.2 oz. What is the probability that the average fill for a six-pack of soda is less than 12 oz?

$$P(\bar{x} < 12) = P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{12 - 12.1}{\frac{0.2}{\sqrt{6}}}\right) = P(z < 1.22) = 0.1112$$

Therefore, the probability that the average fill for a six-pack of soda is less than 12 oz is 0.1112 or 11.12%.

2.3 Sampling Distribution of the Sample Proportion

- The Central Limit Theorem can be used to conclude that the binomial random variable x is approximately normal when n is large, with mean np and standard deviation.
- The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a rescaling of the binomial random variable x, dividing it by n.
- From the Central Limit Theorem, the sampling distribution of $\hat{p}$ will also be approximately normal, with a rescaled mean and standard deviation.

- A random sample of size n is selected from a binomial population with parameter p.
- The sampling distribution of the sample proportion $\hat{p} = \frac{x}{n}$
- It will have mean p and standard deviation of $\sqrt{\frac{pq}{n}}$
- If n is large, and p is not too close to 0 or 1, the sampling distribution will be approximately normal.
- Note: The standard deviation of $\hat{p}$ is sometimes called the Standard Error (SE) of p-hat.

2.3.1 Finding Probabilities for the Sample Proportion

If the sampling distribution of $\hat{p}$ is normal or approximately normal, standardize or rescale the interval of interest in terms of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Then find the appropriate area using Table 3. **Example** A random sample of size $n = 100$ from a binomial population with $p = 0.4$

$$P(\hat{p} > 0.5) = P\left(z > \frac{0.5 - 0.4}{\sqrt{\frac{0.4(0.6)}{100}}}\right) = P(z > 2.04) = 1 - 0.9793 = 0.0207$$

The soda bottler in the previous example claims that only 5% of the soda cans are underfilled. A quality control technician randomly samples 200 cans of soda. What is the probability that more than 10% of the cans are underfilled?

$$\begin{aligned} n &= 200 \\ S &: \text{underfilled can} \\ p = P(S) &= 0.05 \\ q &= 0.95 \\ np &= 10 \\ nq &= 190 \end{aligned}$$

$$P(\hat{p} > 0.5) = P\left(z > \frac{0.10 - 0.05}{\sqrt{\frac{0.05(0.95)}{100}}}\right) = P(z > 3.24) = 1 - 0.9994 = 0.0006$$

Notice the very unusual probability, meaning that if $p = 0.05$ the probability would be unrealistic.

3 Large-Sample Estimation

- Populations are described by their probability distributions and parameters.
  - For quantitative populations, the location and shape are described by μ and σ.
  - For a binomial populations, the location and shape are determined by p.
- If the values of parameters are unknown, we make inferences about them using sample information.

Types of Inference

- A consumer wants to estimate the average price of similar homes in her city before putting her home on the market. (**Estimation:** Estimate μ, the average home price).
- A manufacturer wants to know if a new type of steel is more resistant to high temperatures than an old type was. (**Hypothesis Test:** Is the new average resistance μ_N equal to the old average resistance, μ_O)?

Since an estimator is calculated from sample values, it varies from sample to sample according to its sampling distribution.

Define

- A single value calculated from the sample is called a point estimate.
- $\bar{x}$ is a point estimate of the population parameter μ.
- $\hat{p} = \frac{x}{n}$ is the point estimate of the population parameter p.
- s^2 is the point estimate of σ^2.

3.1 Interval Estimation

- Create an interval (a, b) so that you are fairly sure that the parameter lies between these two values.
- "Fairly sure" means "with high probability" measured using the **confidence coefficient**, $1 - \alpha$.
- Suppose $1 - \alpha = 0.95$ and that the estimator has a normal distribution. Since we don't know the value of the parameter, consider the Estimator ± 1.96SE which has a variable center. Only if the estimator falls in the tail areas will the interval fail to enclose the parameter. This happens only 5% of the time.

Changing the Confidence Level

- To change to a general confidence level, $1 - \alpha$, pick a value of z that puts area $1 - \alpha$ in the center of the z-distribution.

$$100(1 - \alpha) = \text{Confidence Interval: Estimator} \pm z_{\frac{\alpha}{2}}\text{SE}$$

Margin of Error

- For estimators with normal sampling distributions, 95% of all point estimates will lie within 1.96 standard deviations of the parameter of interest.
- **Margin of Error:** the maximum error of estimation, calculated as 1.96(standard error of the estimator).

3.2 Estimating Means and Proportions

- For a quantitative population,
  Point Estimator of population mean $\mu : \bar{x}$
  Margin of Error $(n \geq 30) : \pm 1.96\frac{s}{\sqrt{n}}$

- For a binomial population,
  Point Estimator of population proportion $p : \hat{p} = \frac{x}{n}$
  Margin of Error $(n \geq 30) : \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Examples

1 A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is \$250,000 with a standard deviation of \$15,000. Estimate the average selling price for all similar homes in the city.

> Point Estimator of $\mu : \bar{x} = 250,000$
> Margin of Error: $\pm 1.96\frac{s}{\sqrt{n}} = \pm 1.96\frac{15,000}{\sqrt{64}} = \pm 3675$

2 A quality control technician wants to estimate the proportion of soda cans that are underfilled. He randomly samples 200 cans of soda and finds 10 underfilled cans.

> $n = 200$ $p =$ proportion of underfilled cans
> Point Estimator of $p : \hat{p} = \frac{x}{n} = \frac{10}{200} = 0.05$
> Margin of Error: $\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = \pm 1.96\sqrt{\frac{0.05(0.95)}{200}} = \pm 0.03$

3.3 Confidence Intervals for Means and Proportions

- For a quantitative population,
  Confidence interval for a population mean $\mu : \bar{x} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$
- For a binomial population,
  Confidence interval for a population proportion $p : \hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Examples

3 A random sample of $n = 50$ males showed a mean average daily intake of dairy products equal to 756 grams with a standard deviation of 35 grams. Find a 95% confidence interval for the population average μ.

- Note: The population mean of the daily intake of dairy products is μ. A 95% confidence interval for μ:

> $\bar{x} \pm 1.96\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 1.96\frac{35}{\sqrt{50}} \Rightarrow 750 \pm 9.70$ or $746.30 < \mu < 765.70$ grams.

Interpretation of the Confidence Interval in Example 3

- Intervals constructed in this manner will enclose the population mean μ 95% times in repeated sampling.
- Note: Interpretation is the same for confidence interval of any parameter we construct.

4 Find a 99% confidence interval for μ, the population average daily intake of dairy products for men.

$\bar{x} \pm 2.58\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 2.58\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 12.77$ or $743.23 < \mu < 768.77$ grams.
Note: The interval must be wider to provide for the increased confidence that is does indeed enclose the true value of μ.

- Intervals constructed in this manner will enclose the population mean 99% times in repeated sampling.

5 Of a random sample of $n = 150$ college students, 104 of the students said that they had played on a soccer team during their K-12 years. Estimate the proportions of college students who played soccer in their youth with a 98% confidence interval.

$\hat{p} \pm 2.33\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow \frac{104}{150} \pm 2.33\sqrt{\frac{0.69(0.31)}{150}} \Rightarrow 0.69 \pm 0.09$ or $0.60 < p < 0.78$

- Intervals constructed in this manner will enclose the population proportion 98% times in repeated sampling.

3.4 Estimating the Difference between Two Means

- Sometimes we are interested in comparing the means of two populations.
  - The average growth of plants fed using two different nutrients.
  - The average scores for students taught with two different teaching methods.
- We compare the two averages by making inferences about $\mu_1 - \mu_2$, the difference in the two population averages.
  - If the two population averages are the same, then $\mu_1 - \mu_2 = 0$.
  - The best estimate of $\mu_1 - \mu_2$ is the difference between the sample means, $\bar{x}_1 - \bar{x}_2$.

3.5 Sampling Distribution of the Sample Means

1 The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$, the difference in the population means.

2 The standard deviation of $\bar{x}_1 - \bar{x}_2$ is SE $= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

3 If the sample sizes are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and SE can be estimated as SE $= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

3.5.1 Estimating Population Means

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.

- Point Estimate for $\mu_1 - \mu_2 : \bar{x}_1 - \bar{x}_2$
  Margin of Error : $\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  Confidence Interval for $\mu_1 - \mu_2 : (\bar{x}_1 - \bar{x}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$

Example

Average Daily Intakes	Men	Women
Sample Size	50	50
Sample Mean	756	762
Standard Deviation	35	30

- Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

$$(\bar{x}_1 - \bar{x}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \Rightarrow (756 - 762) \pm 1.96\sqrt{\frac{35^2}{\sqrt{50}} + \frac{30^2}{\sqrt{50}}}$$
$$\Rightarrow -6 \pm 12.78 \text{ or } -18.78 < \mu_1 - \mu_2 < 6.78)$$

Notice

- Could you conclude, based on this confidence interval, that there is a difference in the average daily intake of dairy products for men and women?

- The confidence interval contains the value $\mu_1 - \mu_2 = 0$. Therefore, it is possible that $\mu_1 = \mu_2$. You would not want to conclude that there is a difference in average daily intake of dairy products for men and women.

3.5.2 Estimating the Difference between Two Proportions

- Sometimes we are interested in comparing the proportion of successes in two binomial populations.
  - The germination rates of untreated seeds and seeds treated with a fungicide.
  - The proportion of male and female voters who favour a particular candidate for governor.
- A random sample of size n_1 drawnfrom binomial population with parameter p_1 and a random sample size of n_2 drawn from binomial population with parameter p_2.
- We compare teh two proportions by making inferences about $p_1 - p_2$, the difference in the two population proportions.

- If the two population proportions are the same, then $p_1 - p_2 = 0$.
- The best estimate of $p_1 - p_2$ is the difference between the sample proportions,

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

3.6 Sampling Distribution of Sample Proportions

1 The mean of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$, the difference in the population proportions.

2 The standard deviation of $\hat{p}_1 - \hat{p}_2$ is SE $= \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$.

3 If the sample sizes are large, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, and SE can be estimated as SE $= \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$.

3.6.1 Estimating Population Proportions

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.

- Point Estimate for $p_1 - p_2 : \hat{p}_1 - \hat{p}_2$
  Margin of Error : $\pm 1.96\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$
  Confidence Interval for $p_1 - p_2 : (\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}})$

Example

Youth Soccer	Male	Female
Sample Size	80	70
Played Soccer	65	39

- Compare the proportions of male and female college students who said that they had played on a soccer team during their K-12 years using a 99% confidence interval.

$$(\hat{p}_1 - \hat{p}_2) \pm 2.58\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \Rightarrow \left(\frac{65}{80} - \frac{39}{70}\right) \pm 2.58\sqrt{\frac{0.89(0.19)}{80} + \frac{0.56(0.44)}{70}}$$
$$\Rightarrow 0.25 \pm 0.19 \text{ or } 0.06 < p_1 - p_2 < 0.44$$

Notice

- Could you conclude, based on this confidence interval, that there is a difference in the proportion of male and female college students who said that they had played on a soccer team during their K-12 years?
- The confidence interval does not contain the value $p_1 - p_2 = 0$. Therefore, it is not likely that $p_1 = p_2$. You would conclude that there is a difference in the proportions for males and femals.

IE: "A higher proportion of males than females played soccer in their youth."

3.6.2 Choosing the Sample Size

- The total amount of relevant information in a sample is controlled by two factors:
  - The **sampling plan** or **experimental design:** the procedure for collecting the information.
  - The **sample size** n**:** the amount of information you collect.
- In a statistical estimation problem, the accuracy of the estimation is measured by the margin of error or the "width of the confidence interval".
- To Choose the Sample Size:
  1. Determine the size of the margin of error, B, that you are willing to tolerate.
  2. Choose the sample size by solving for n or $n = n_1 = n_2$ in the inequality: 1.96 SE $= B$, where SE is a function of the sample size n.
  3. For quantitative populations, estimate the population standard deviation using a previously calculated value of s or the range approximation $\sigma \approx \frac{\text{Range}}{4}$.

4 For binomial populations, use the conservative approach and approximate p using the value $p = 0.5$.

Example

- A producer of PVC pipe wants to survey wholesalers who buy his product in order to estimate the proportion who plan to increase their purchases next year. What sample size is required if he wants his estimate to be within 0.04 of the actual proportion with probability equal to 0.95?

$$1.96\sqrt{\frac{pq}{n}} = 0.04 \Rightarrow 1.96\sqrt{\frac{0.5(0.5)}{n}} = 0.04 \Rightarrow \sqrt{n} = \frac{1.96\sqrt{0.5(0.5)}}{0.04} = 24.5$$
$$\Rightarrow n = 24.5^2 = 600.25$$

4 Large-Sample Tests of Hypotheses

4.1 Hypothesis, Null and Alternative Hypothesis

- Any statement regarding the value of a population parameter is called a hypothesis.
- The hypothesis to be tested is called null hypothesis.
- The complement of the null hypothesis is called an alternative hypothesis.

4.1.1 One and Two Sided Alternatives

- The null hypothesis is always a single value of the parameter, $H_0 : \mu = \mu_0$
- The alternative hypothesis can be two sided, $H_A : \mu \neq \mu_0$ or one sided, $H_A : \mu < \mu_0$ or $H_A : \mu > \mu_0$.

Example

1 A drug manufacturer claimed that the mean potency, μ, of one of its antibiotics was 80%. A random sample of $n = 100$ capsules were tested and produced a sample mean of $\bar{x} = 79.7\%$, with a standard deviation of $s = 0.8\%$. Does the data present sufficient evidence to refute the manufactuerer's claim?

- Here $H_0 : \mu = 80$ is the null hypothesis and $H_A : \mu \neq 80\%$ is the alternative hypothesis. Note: the alternative is two sided.

2 A drug manufactuerer claimed that the mean potency, μ, of one of its antibiotics was less than 80%. A random sample of $n = 100$ capsules were tested and produced a sample mean of $\bar{x} = 79.7\%$, with a standard deviation of $s = 0.8\%$. Does the data present sufficient evidence to refute the manufactuerer's claim?

- Here $H_0 : \mu = 80$ is the null hypothesis and $H_A : \mu < 80$ is the alternative hypothesis. Note: the alternative is one sided.

4.2 Testing a Null Hypothesis

- We need a quantity which is called a test statistic.
- We can then test using one of the three methods:

1 Critical value of the test statistic.

2 Using a P-value.

3 Confidence Interval Method.

4.2.1 Critical Value Method

- Consider a two sided alternative hypothesis. Notes that the null hypothesis always specifies a single value of the parameter. $H_0 : \mu = \mu_0$, $H_A : \mu \neq \mu_0$.
- Note: The Rejection Region - Reject H_0 if $z > 2.33$, if the test statistic falls in the rejection region, its p-value will be less than $\alpha = 0.01$.
  - Additionally, reject H_0 if $z < -1.645$if the test statistic falls in the rejection region, its p-value will be less than $\alpha = 0.05$.

Test Statistic

- Assume to begin with that H_0 is true. The sample mean $\bar{x}$ is our best estimate of μ, and we use it in a standardized form as the **test statistic:**

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \approx \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Since $\bar{x}$ has an approximate normal distribution with mean μ_0 and standard deviation $\frac{\sigma}{\sqrt{n}}$, the test statistic is approximately normal.
- If H_0 is true the value of $\bar{x}$ should be close to μ_0, and z will be close to zero. If H_0 is false, $\bar{x}$ will be much larger or smaller than μ_0, and z will be much larger or smaller than zero, indicating that we should reject H_0.

Large Sample Test of a Population Mean

- Take a random sample of size $n \geq 30$ from a population with mean μ and standard deviation σ.
- We assume that either: σ is known or $s \approx \sigma$ since n is large.

Parts of a Statistical Test

Conclusion:

- Either "Reject H_0" or "Do not reject "H_0", along with a statement about the reliability of your conclusion.

How do you decide when to reject H_0?

- Depends on the **significance level (α)**, the maximum tolerable risk you want to have of making a mistake, if you decide to reject H_0.
- Usually, the significance level is $\alpha = 0.01$ or $\alpha = 0.05$.

From Example [1], the value of the test statistic is,

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{79.7 - 80}{\frac{0.8}{\sqrt{100}}} = -3.75$$

Example 1 Revisited

- Critical Value Method:
  - Take $\alpha = 0.05$, then $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$.
  - Since $|Z| = 3.75 > Z_{0.025} = 1.96$, we reject the null hypothesis (H_0) at 5% level of significance. Therefore, there is insufficient evidence in the data to support the manufacturer's claim.

4.2.2 P-Value Method: Likely or Unlikely?

- Once you've calculated the observed value of the test statistic, calculate its **p-value** using the following formula:

p-value: The probability of observing, just by chance, a test statistic as extreme or even more extreme than what we've actually observed. If H_0 is rejected this is the actual probability that we have made an incorrect decision.

- If this probability is very small, less than some preassigned significance level (α), then we can reject H_0.
- If the alternative hypothesis is one-sided, $H_A : \mu < \mu_0$, then the $p-$value $= p(Z < -Z_0)$ and $Z_0 = \left|\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right|$. Note: Reverse both equality symbols if the null hypothesis had a inverted equality than the example.
- However if $H_A : \mu \neq \mu_0$, then the $p-$ value $= 2p(Z < -Z_0)$ and $Z_0 = \left|\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right|$.

Example 1 Revisited

- The null and alternative hypothesises are: $H_0 : \mu = 80$ and $H_A : \mu \neq 80$.
- Z = -3.75
- P-value for this test is: $p-$ value $= 2p(Z < -3.75) = 0.0$

- Since $p-$ value < 0.05, we reject the null hypothesis at 5% level of significance.

4.2.3 Statistical Significance of P-Values

- If the $p-$ value is less than 0.01, reject H_0. The results are **highly significant**.
- If the $p-$ value is between 0.01 and 0.05, reject H_0. The results are **statistically significant**.
- If the $p-$ value is between 0.05 and 0.10, do not reject H_0. The results are **tending towards significance**.
- If the $p-$value is greater than 0.10, do not reject H_0. The results are **not statistically significant**.

4.3 Confidence Interval Method

- Confidence interval method can only be applied to a two-sided test.
- Construct a $(1-\alpha)100\%$ confidence interval. Then, if the value of the parameter under teh nul hypothesis falls in this interval then we do not reject the null hypothesis. Otherwise we reject the null hypothesis in favor of the alternative.

Example

- A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is \$252,000 with a standard deviation of \$15,000. Is there sufficient evidence to conclude that the average selling price in her area is \$250,000? Use $\alpha = 0.01$.
- So: $H_0 : \mu = 250,000$ is the null hypothesis. $H_A : \mu \neq 250,000$ is the alternative hypothesis. Note: the alternative is two-sided.
- Additionally: A 99% confidence interval for population mean is:

$$\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = \left(252000 - 2.575\frac{15000}{\sqrt{64}}, 252000 + 2.575\frac{15000}{\sqrt{64}}\right)$$
$$\Rightarrow (247171.90, 256828.10)$$

- Since 250,000 is included in the 99% confidence interval we do not reject the null hypothesis. There is evidence in the data to suggest that the mean selling price is \$250,000.

4.3.1 Statistical Significance of the Confidence Interval Method

- The critical value approach and the $p-$value approach produce identical results.
- The $p-$value approach is often preferred because:
  - Computer printouts usually calculate $p-$value
  - You can evaluate the test results at any significance level you choose.

Q: What should you do if you are the experimenter and no one gives you a significance level to use?

- The General Test Statistic:

$$z = \frac{\text{statistic} - \text{hypothesized value}}{\text{standard error of statistic}}$$

4.4 Testing the Difference between Two Means

- A random sample of size n_1 is drawn from population 1 with mean μ_1 and variance σ_1^2.
- A random sample of size n_2 is drawn from population 2 with mean μ_2 and variance σ_2^2.
- The hypothesis of interest involves the difference, $\mu_1 - \mu_2$, in the form: $H_0 : \mu_1 - \mu_2 = D_0$ versus H_a : one of three where D_0 is the hypothesized difference, usually zero.

$H_0 : \mu_1 - \mu_2 = D_0$
H_a : one of three alternatives
Test Statistic: $z \approx \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
with rejection regions and/or p-values based on the standard normal z distribution.

4.5 Testing a Binomial Proportion

A random sample of size n from a binomial population is selected to test.
$H_0 : p = p_0$ versus H_a : one of three alternatives
Test Statistic: $z \approx \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$
with rejection regions and/or p-values based on the standard normal z distribution.

4.6 Testing the Difference between Two Proportions

$H_0 : p_1 - p_2 = 0$ versus H_a : one of three alternatives
Test Statistic: $z \approx \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
with $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ to estimate the common value of p and rejection regions or p-values based on the standard normal z distribution.

4.7 Steps in Hypothesis Testing

- Set up the null and the alternative hypothesises.
- Calculate the test statistic.
- Obtain the critical value or the p-value or and appropriate confidence interval and make a decision as to reject or not to reject the null hypothesis.
- Make a decision and write a conclusion with reason.