MBA - 2^{nd} year

Business Statistics - Course - MS6107E - Statistical Inference

Arjun Anil Kumar

SOMS NIT Calicut

September 5, 2023

Contents

Contents

Statistical Inference

- Statistical methods are used to make **decisions** and draw **conclusions** about **populations**. This aspect of statistics is generally called statistical inference.
- In general, a statement would be made about the **population**. You have to derive **conclusions** on the **veracity** of the statement based on the **random** samples of size n obtained from **population** of size N.
  - Eg : Proportion of students who feel they get placed in two companies is 20% or p = .2
  - Eg : The difference of means between salary of employees with 5 years and employees with 7 years experience post MBA is 45000 Rs a month or $\mu_{G1} - \mu_{G2}$ =45000
  - Eg : The average marks scored in Statistics exam is 54 or μ=54
  - Eg : The variance of salary of men is 250000 or $\sigma^2 = 250000$
  - Eg : The variance of salary of men and variance of salary of women are equal or ${\sigma_{G1}}^2 = {\sigma_{G2}}^2$
  - Observations : Some statements about one group and some are about two groups.
  - Statistical Inference is divided into Two Major Areas : **Parameter Estimation** & **Hypothesis Testing**

Contents

Parameter Estimation

- Parameter Estimation means estimating the parameters of the population distribution.
- Parameter Estimation has two methods
  - Point Estimation - Estimates the parameter as a single point or value. Examples are
    Single point estimate for population mean is sample mean
    Single point estimate for population variance is sample variance
    Single point estimate for population proportion is sample proportion
  - Interval Estimation - Estimates the interval with lower and upper limits for a parameter. Examples are
    Interval Estimates for population mean with a upper and lower limit
    Interval Estimates for population variance with a upper and lower limit
    Interval Estimates for population proportion with a upper and lower limit

Point Estimation of Population Mean

Problem : Estimate the population mean μ using n samples $[x_1, x_2, x_n]$ randomly drawn from population?

- We need a single point estimate of the population mean.
- For estimating the population mean from the samples, we need a estimator.
- The sample mean $\hat{\mu}$ is the point estimator of unknown population mean μ.

$$\hat{\mu} = \frac{x_1 + x_2 ... + x_n}{n} \tag{1}$$

- **A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic T . The statistic T is called the point estimator.**
- In the above example population parameter $\theta = \mu$, Statistic T $= \sum_{i=1}^{n} \frac{x_i}{n}$ and the sample mean $\hat{\theta}$ or $\hat{\mu}$ computed at the random sample $[x_1, x_2, x_n]$ is the single point estimate of population parameter μ.

Point Estimation of Population Variance

Problem : Estimate the population variance[1] σ^2 using n samples $[x_1, x_2, x_n]$ randomly drawn from population?

- We need a single point estimate of the population variance.
- For estimating the population variance from the samples, we need a estimator.
- The sample variance[2] is the point estimator of unknown population mean σ^2.

$$\hat{\sigma}^2 = \frac{\sum_1^n [x_i - \hat{\mu}]^2}{n-1} \tag{2}$$

- **A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic T . The statistic T is called the point estimator.**
- In the above example, population variance $\theta = \sigma^2$, Statistic T $= \frac{\sum_1^n [x_i - \hat{\mu}]^2}{n-1}$ and the sample variance $\hat{\theta}$ or $\hat{\sigma}^2$ computed at the random sample $[x_1, x_2, x_n]$ is the single point estimate of population parameter $\theta = \sigma^2$.

[1] Population Variance is the square of the Population Standard Deviation

[2] n-1 is used to ensure that the estimator is unbiased or $E[\hat{\sigma}^2] = \sigma^2$

Point Estimation of Population Proportion

Problem : Estimate the population proportion p in favour of something using n samples $[x_1, x_2, x_n]$ randomly drawn from population, where x_i are binary variables (0 or 1) corresponding to a condition?

- We need a single point estimate of the population proportion.
- For estimating the population proportion from the samples, we need a estimator.
- The sample proportion $\hat{p}$ is the point estimator of unknown population population proportion p.

$$\hat{p} = \frac{\sum_{i=1}^{n}[x_i]}{n} \tag{3}$$

- **A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic T . The statistic T is called the point estimator.**
- In the above example, population proportion θ = p, Statistic T = $\frac{\sum_{1}^{n}[x_i]}{n}$ and the sample proportion $\hat{\theta}$ or $\hat{p}$ computed at the random sample $[x_1, x_2, x_n]$ is the single point estimate of population parameter $\theta = p$.

Point Estimation of Difference in Means and Proportions

We want to estimate the difference in means of two populations (μ_1 - μ_2) or difference in proportion of two population (p_1 - p_2) using n randomly drawn samples from first group $[x_1, x_2, x_n]$ and n randomly drawn samples from second group $[y_1, y_2, y_n]$

- We need a single point estimate of the difference of population means and population proportions.
- For estimating the the difference of population means $\mu_1 - \mu_2$ and population proportions $p_1 - p_2$ from the samples, we develop single point estimators such as ($\hat{\mu}_1$ - $\hat{\mu}_2$) and $\hat{p_1} - \hat{p_2}$ respectively.

$$\hat{\mu}_1 = \frac{x_1 + x_2 ... + x_n}{n} \tag{4}$$

$$\hat{\mu}_2 = \frac{y_1 + y_2 ... + y_n}{n} \tag{5}$$

$$\hat{p_1} = \frac{\sum_{i=1}^{n}[x_i]}{n} \tag{6}$$

$$\hat{p_2} = \frac{\sum_{1}^{n}[y_i]}{n} \tag{7}$$

Confidence Intervals

- Assume the sample mean is $\hat{\mu} = 1000$
- The sample mean does not say anything about how close the sample mean $\hat{\mu}$ is to the population mean μ
- Is the population mean likely to be in the range 900 and 1000 or $900 \leq \mu \leq 1000$
- Is the population mean likely to be in the range 990 to 1010 or $990 \leq \mu \leq 1010$?
- An interval estimate for a population parameter is called a **confidence interval**.
- Shorter the length of the interval, more precise is the estimation.

CONFIDENCE INTERVAL ON THE MEAN OF A NORMAL DISTRIBUTION, VARIANCE KNOWN

Development of Confidence Interval

- Suppose that we have a normal population with unknown mean μ and known variance $\sigma^2 => $ Unrealistic.
- Suppose that $X_1, X_2, .., X_n$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2.
- CLT says

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{8}$$

  As $n- > \infty$, is the standard normal distribution.
- A confidence interval estimate for μ is an interval of the form $l \leq \mu \leq u$, where the end points l and u are computed from the sample data.
- Because different samples will produce different values of l and u, these end-points are values of random variables L and U, respectively.

CONFIDENCE INTERVAL ON THE MEAN OF A NORMAL DISTRIBUTION, VARIANCE KNOWN

Development of Confidence Interval

- Is it possible to determine the values of L and U such that the following probability condition is satisfied. $P\left[L \leq \mu \leq U\right] = 1 - \alpha$, where 1 - α is called the confidence.
- The answer is yes! $P\left[- z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$
  where $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- For a sample of size n, the confidence interval for population mean μ and population standard deviation σ with a confidence of 1 - α, where α is the level of significance, is given by

$$P\left[\bar{X} - K \leq \mu \leq \bar{X} + K\right] = 1 - \alpha \tag{9}$$

$$K = z_{\frac{\alpha}{2}} * SE \tag{10}$$

$$SE = \frac{\sigma}{\sqrt{n}} \tag{11}$$

Confidence Interval

- If the population mean[3] is between 60 and 80 with a confidence of 95%

$$P(60 < \mu < 80) = 95\% \tag{12}$$

Does it mean that around 95% of the times μ falls between 60 and 80? Is this right?

[3] Although population mean is unknown, it is a constant

Confidence Interval - Frequentist Interpretation

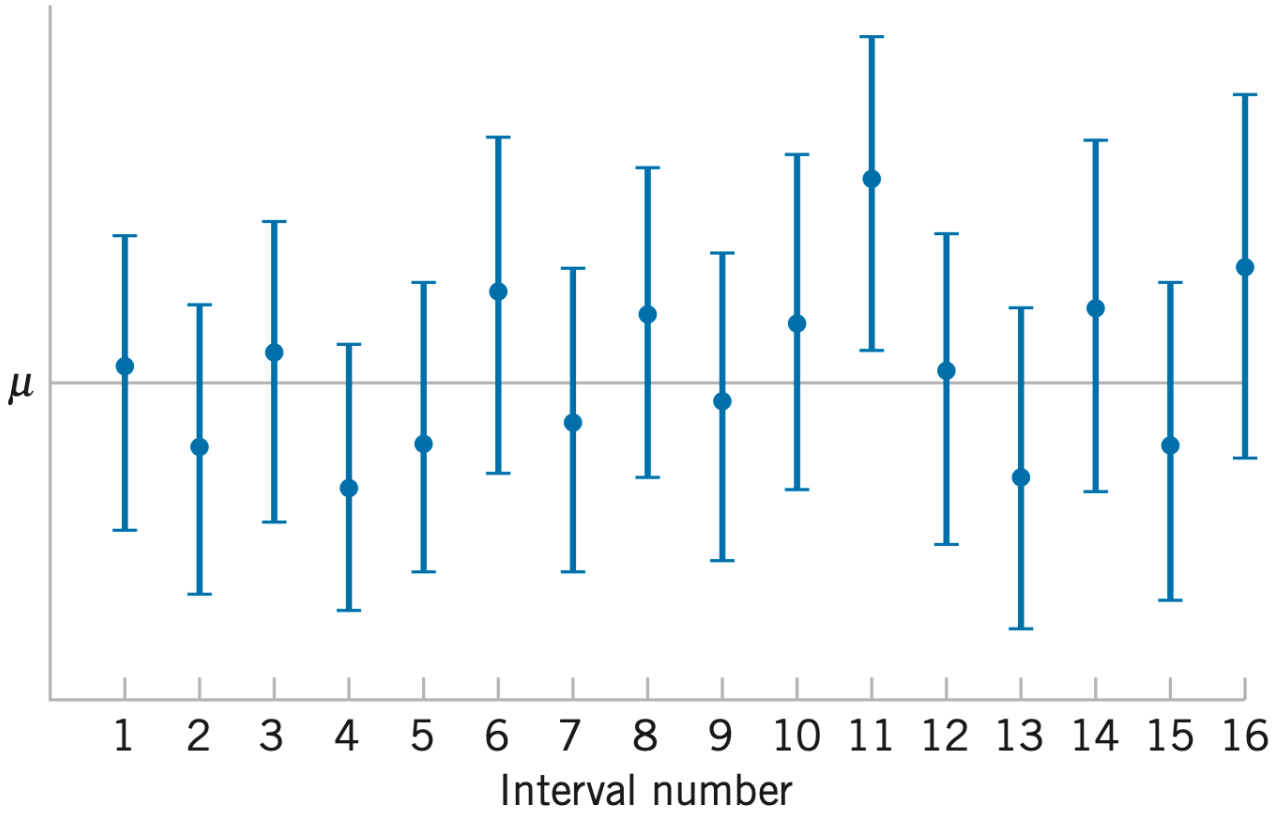


Figure: Confidence Interval - Frequentist Interpretation

Confidence Interval and Precision - Inverse Relation

- The confidence interval is given by 2K or $2.z_{\frac{\alpha}{2}} * SE$ or $2.z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$
- When confidence (1-α) is increased, α is decreased. For example : For 95% confidence or α =.05, we have $Z_{.025} = 1.96$ and 99% confidence or $\alpha = .1$ or $Z_{.005} = 2.58$.
- Confidence level $\uparrow$ Confidence Interval $\uparrow$ Precision $\downarrow$

Precision and Confidence Interval

- Ideally you would want a high precision and high confidence interval.
- But when we increase the confidence level or the confidence interval, the precision is actually decreasing.
- **But for a given confidence interval, we can improve the precision by increasing the sample size n!**

Choice of Sample Size and Error

$$P\left[\hat{\mu} - z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha \tag{13}$$

$$P\left[- z_{\frac{\alpha}{2}} \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha \tag{14}$$

$$P\left[|E| < z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha \tag{15}$$

$$P\left[|E| < E_{max}\right] = 1 - \alpha \tag{16}$$

If $\hat{\mu}$ is the sample mean and μ is the population mean, we can be $100(1-\alpha)$, that the error $|\hat{\mu} - \mu|$ will not exceed an error E_{max} if the sample size n $= \left[\frac{z_{\frac{\alpha}{2}} * \sigma}{E_{max}}\right]^2$

Error and Sample Size

- Error $E_{max} \downarrow$ implies increase in the sample size n if the confidence level $(1-\alpha)$ and population standard deviation σ is fixed.
- As population standard deviation $\sigma \downarrow$ the sample size n also reduces when the error E and the confidence level $(1-\alpha)$ is fixed.
- As the level of confidence increases or $z_{\frac{\alpha}{2}}$ increases, the sample size n also increases when the error E and population standard deviation σ is fixed.

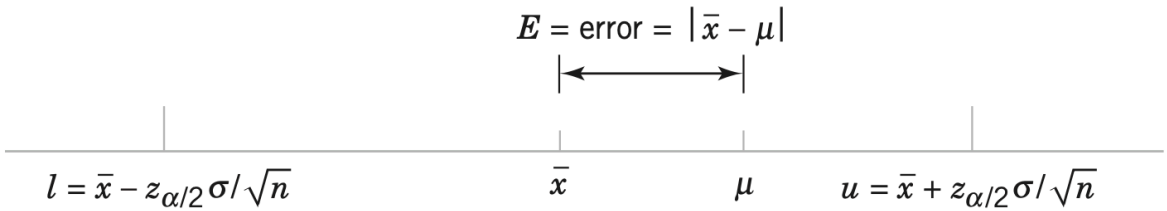


Figure: Error in estimating μ with $\hat{\mu}$

Large Sample ($n > 40$) Confidence Interval for Mean

CONFIDENCE INTERVAL ON THE MEAN OF A NORMAL DISTRIBUTION, VARIANCE UNKNOWN

- Suppose that $X_1, X_2, .., X_n$ is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2.
- CLT says

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{17}$$

As $n- > \infty$, is the standard normal distribution.

$$P\left[\hat{\mu} - z_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}}\right] = 1 - \alpha \tag{18}$$

Small Sample ($n <= 40$) Confidence Interval for Mean

CONFIDENCE INTERVAL ON THE MEAN OF A NORMAL DISTRIBUTION, VARIANCE UNKNOWN

- Suppose that $X_1, X_2, .., X_n$ is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2.
- CLT says

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{19}$$

As $n- > \infty$, is the T distribution with n-1 degrees of freedom

$$P\left[\hat{\mu} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \hat{\mu} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}\right] = 1 - \alpha \tag{20}$$

Standard T and Standard Normal Distribution Z

- The general appearance of the T-distribution is similar to the standard normal distribution in that both distributions are symmetric and unimodal, and the maximum ordinate value is reached when the μ is zero.
- However, the t distribution has heavier tails than the normal; that is, it has more probability in the tails than the normal distribution.
- As the number of degrees of freedom k increases, the limiting form of the t distribution is the standard normal distribution.
- Generally, the number of degrees of freedom for t is the number of degrees of freedom associated with the estimated standard deviation.

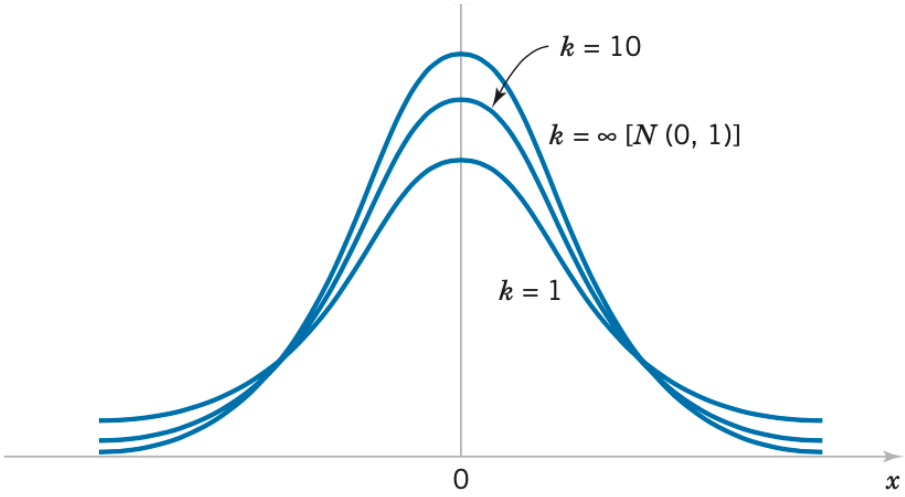


Figure 8-4 Probability density functions of several t distributions.

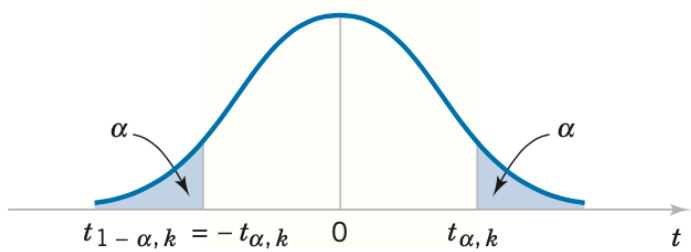


Figure 8-5 Percentage points of the t distribution.

Confidence Interval of Population Mean

Suppose scores on exams in statistics are normally distributed with an unknown population mean and unknown population standard deviation. A random sample of 36 scores is taken to compute the sample mean score of 68 marks and sample standard deviation of 3 marks.Find a confidence interval estimate for the population mean exam score? Compute the CI with 90% confidence for the true (population) mean of statistics exam scores.

- Sample mean is t distributed[4] with unknown mean μ and unknown standard deviation of $\frac{s}{\sqrt{n}}$
- s is the sample standard deviation
- s = 3
- population standard deviation $\sigma \approx$ s
- Confidence interval equals to [$\bar{x}$ - E,$\bar{x}$+E]
- Sample mean $\bar{x}$ is 68
- Error E is $t_{\frac{\alpha}{2}}$*SE($\bar{x}$)
- SE($\bar{x}$) = $\frac{s}{\sqrt{n}}$
- n is sample size = 36

[4] n-1 degrees of freedom, n is the sample size

Confidence Interval of Population Mean

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three marks. A random sample of 36 scores is taken to compute the sample mean score of 68 marks. Find a confidence interval estimate for the population mean exam score? Compute the CI with 90% confidence for the true (population) mean of statistics exam scores.

- Sample mean is normally distributed with unknown mean μ and known standard deviation of $\frac{\sigma}{\sqrt{n}}$
- Confidence interval equals to [$\bar{x}$ - E,$\bar{x}$+E]
- Sample mean $\bar{x}$ is 68
- Error E is $z_{\frac{\alpha}{2}}$*SE($\bar{x}$)
- SE($\bar{x}$) = $\frac{\sigma}{\sqrt{n}}$
- $z_{\frac{\alpha}{2}}$ = 1.645 for 90% confidence.
- n is sample size = 36
- Population standard deviation $\sigma = 3$

Large Sample Confidence Interval for a Population Proportion

Construction of population proportion

- We need to construct confidence interval for population proportion.
- Assume we randomly sample x_1,x_2,x_3,..x_n
- Note that the samples x_1,x_2,x_3,..x_n are drawn from a bernouli distribution of parameter p.
- The sample proportion is the point estimator of unknown population population proportion p.

$$\hat{p} = \frac{\sum_{i=1}^{n}[x_i]}{n} \tag{21}$$

- $\hat{p}$ is a random variable with mean $\frac{1}{n} * n * p = p$ and variance $\frac{1}{n^2}.n.p.(1-p)$ $= \frac{p(1-p)}{n}$.
- As per CLT, if n is large, the distribution of

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p.(1-p)}{n}}} \tag{22}$$

Large Sample Confidence Interval for a Population Proportion

- $$P\big[-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\big] = 1-\alpha \tag{23}$$
- $$P\big[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{p.(1-p)}{n}} \leq p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{p.(1-p)}{n}}\big] = 1-\alpha \tag{24}$$
- Confidence Interval for Proportion

$$P\big[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}\big] = 1-\alpha \tag{25}$$

- Choice of sample size n based on max error E_{max} or $max|(\hat{p}-p)|$

$$P\Big[|E| < E_{max}\Big] = 1-\alpha \tag{26}$$

If $\hat{p}$ is the sample proportion and p is the population proportion, we can be $100(1-\alpha)$, that the error $|\hat{p}-p|$ will not exceed an error E_{max} if the sample size n $= \left[\frac{z_{\frac{\alpha}{2}} * \sigma}{E_{max}}\right]^2$ or $\left[\frac{z_{\frac{\alpha}{2}} * \sqrt{p.(1-p)}}{E_{max}}\right]^2$

Sample Problem : Confidence Interval for Proportion

Q1 : Out of random sample of 85 students, 10 students voted for party A and 75 voted for the rest. Compute the confidence interval for the population proportion voting for party A

Sample proportion $\hat{p} = \frac{10}{85}$ For a two sided 95% interval, $\alpha = .05$, $z_{.025} = 1.96$

The confidence interval of population proportion is given by

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}\right]$$

.05<=p<=.19

Q2 : How many samples do we pick for taking a decision assuming the max error $max|(\hat{p} - p)|$ is .05?

$$n = \left[\frac{z_{\frac{\alpha}{2}} * \sqrt{p.(1-p)}}{E_{max}}\right]^2$$

For observed sample, sample size n $= \left[\frac{z_{\frac{\alpha}{2}} * \sqrt{\hat{p}.(1-\hat{p})}}{E_{max}}\right]^2 = 163$

If no observed sample is there, maximum sample size n $= \left[\frac{z_{\frac{\alpha}{2}} * \sqrt{.5*(1-.5)}}{E_{max}}\right]^2 = 385$

Confidence Interval of Population Proportion

In the Framingham Offspring study 1,219 subjects were on anti-hypertensive medication out of 3,532 total subjects. Construct the confidence interval on the population proportion with 95% confidence?

- $\hat{p} = \frac{1219}{3532}$
- Confidence interval equals to [$\hat{p}$ - E,$\hat{p}$+E]
- Error E is $z_{\frac{\alpha}{2}}$*SE($\hat{p}$)
- SE($\hat{p}$) = $\sqrt{\frac{p.(1-p)}{n}}$
- Since p is unknown, SE($\hat{p}$) = $\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$
- n is sample size = 3532
- Confidence Interval is [.329,.361]

Confidence Interval for Population Variance

- If X_1,X_2,X_3...X_n are standard normal variables with mean 0 and variance 1, the variable χ^2 expressed as the sum of squares of standard normal variable would have a χ^2 distribution.

$$\chi^2 = X_1^2 + X_2^2 + X_3^2 + X_n^2 \tag{27}$$

- $\chi^2 = \frac{n-1.s^2}{\sigma^2}$ is a chi-square distribution with n-1 degrees of freedom.
- $\chi^2_{-\frac{\alpha}{2},n-1} \leq \chi^2 \leq \chi^2_{\frac{\alpha}{2},n-1}$
- $\chi^2_{-\frac{\alpha}{2},n-1} \leq \frac{n-1.s^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-1}$
- The confidence interval of σ^2 is $[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{-\alpha}{2},n-1}}]$

Prediction Interval

- Suppose X_1, X_2 .. X_n are n random samples
- We wish to predict X_{n+1} and we want to construct an interval for X_{n+1}.
- A point prediction of X_{n+1} is $\bar{X}$
- Prediction Error E is $X_{n+1} - \bar{X}$
- Prediction Error E is a random variable with mean 0 and variance $\sigma^2 + \frac{\sigma^2}{n}$.
- $$Z = \frac{X_{n+1} - \bar{X}}{\sqrt{\sigma^2 + \frac{\sigma^2}{n}}} \tag{28}$$
- $$T = \frac{X_{n+1} - \bar{X}}{\sqrt{s^2 + \frac{s^2}{n}}} \tag{29}$$
- The prediction interval is given by

$$\bar{x} - t_{\frac{\alpha}{2}} * \sqrt{s^2 + \frac{s^2}{n}} \leq X_{n+1} \leq \bar{x} + t_{\frac{\alpha}{2}} * \sqrt{s^2 + \frac{s^2}{n}} \tag{30}$$

Tolerance Interval

- Suppose the mean annual sales of cars during the 50 years is 600 hundred thousands and the standard deviation is 30 hundred thousands, the interval 600 - 1.96*30 to 600 + 1.96*30 would capture the sales of 95% (.95*50 years) of the data. This is the idea behind tolerance interval. This was possible because we knew the population mean and population standard deviation => Not really practical!
- A tolerance interval for capturing at least γ of the values in a normal distribution with confidence level 100(1-α)% is $\bar{x} - ks, \bar{x} + ks$, where $\bar{x}$ is the sample mean and s is the sample standard deviance. k is the tolerance factor that needs to be obtained from standard normal table.

Confidence	Critical Value
90	1.645
95	1.96
99	2.58

Table: Caption

Contents

Central Limit Theorem

If X_1, X_2, .., X_n is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, the limiting form of the distribution of random variable Z

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{31}$$

As $n- > \infty$, is the standard normal distribution Z $\sim \mathcal{N}(0, 1)$.

- Sample Mean $\bar{X}$ is a random variable and follows a normal distribution with mean equal to population mean and standard deviation equal to population standard deviance divided by square root of n or $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
- If n $>=$ 30, the distribution of the sample mean would be always normal irrespective of the distribution of the parent distribution from which X_1, X_2, .., X_n are sampled.
- If n $<$ 30, the central limit theorem will work only if the distribution of the parent distribution from which X_1, X_2, .., X_n are sampled are not severly non-normal. Ideally, they should have been normal.

Distribution of average scores of throwing n dies m times

Throw n dies m times, the distribution of mean score is close to normal distribution

Each Experiment E_i refers to throwing 'i' dies m times and taking the average of the dies $\frac{\sum_1^m X_{i1}+X_{i2}+,,X_{im}}{N}$, where i is 1,2,5 and 10.

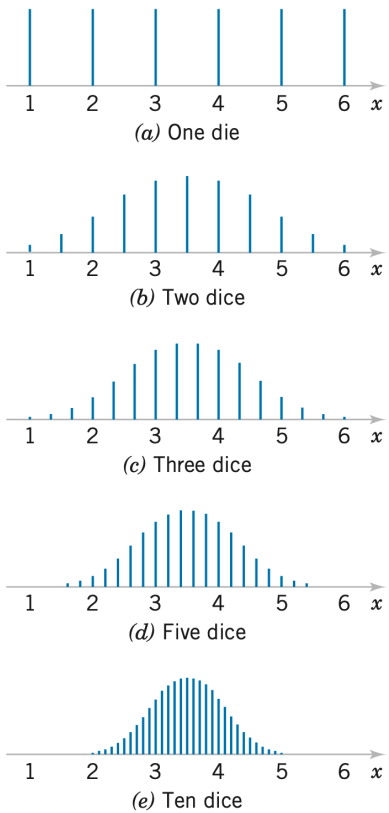


Figure: Distribution of average scores of throwing n dies m times

CLT Theorem

The Central Limit Theorem (CLT) tells us that the sampling distribution of the sample mean is, at least approximately, normally distributed, regardless of the distribution of the underlying random sample.

- If the distribution of the X_i, i from 1 to n, is symmetric, unimodal or continuous, then a sample size as small as 4 or 5 yields an adequate approximation.
- If the distribution of the X_i, i from 1 to n, is skewed, then a sample size of at least 25 or 30 yields an adequate approximation.
- If the distribution of the X_i, i from 1 to n, is extremely skewed, then you may need an even larger .

Sampling Mean $\bar{X}$ and CLT

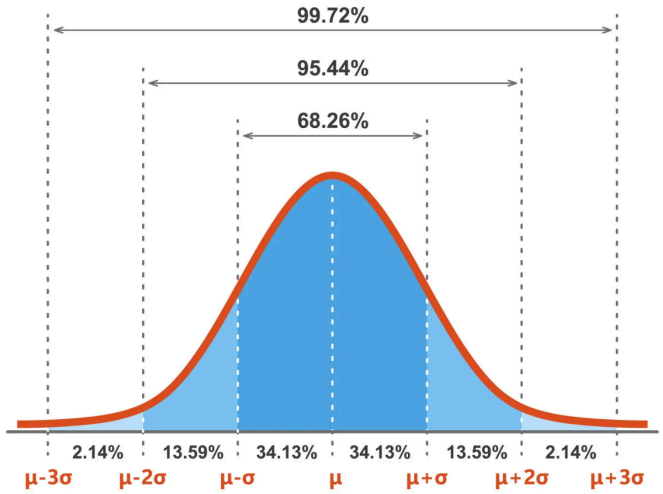
- $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
- Construct standard normal variable

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{32}$$

- $Z \sim \mathcal{N}(0, 1)$
- Hypothesis Testing for Population Mean can be done using the above statistic Z

  $H_0 : \mu = \mu_0$

  $H_a : \mu \neq \mu_0$



Difference in Sample Means $\bar{X}_1$ - $\bar{X}_2$ and CLT

- $\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$.
- If we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, and if $\bar{X}_1$ and $\bar{X}_2$ are the sample means of two independent random samples of size n_1 and size n_2, then the sampling distribution of

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{33}$$

  is approximately standard normal.
- $Z \sim \mathcal{N}(0, 1)$
- If the two populations are normal, the sampling distribution Z is exactly standard normal.
- Hypothesis Testing for Population Mean can be done using the above statistic Z
  $H_0 : \mu_1 = \mu_2$
  $H_a : \mu_1 \neq \mu_2$

Contents

Hypothesis Testing - A motivation

Suppose that an engineer is designing an air crew escape system that consists of an ejection seat and a rocket motor that powers the seat. The rocket motor contains a propellant, and in order for the ejection seat to function properly, the propellant should have a mean burning rate of 50 cm/sec. If the burning rate is too low, the ejection seat may not function properly, leading to an unsafe ejection and possible injury of the pilot. Higher burning rates may imply instability in the propellant or an ejection seat that is too powerful, again leading to possible pilot injury.
The important question is **"Is the mean burning rate equal to 50 cm/sec"**.
The solution is **Hypothesis Testing**

Hypothesis Testing

Definitions

- The statements made about the parameters of the general population are called as hypotheses.
- **No hypothesis is made about samples.**
- The decision making procedure is called hypothesis testing
- A **statistical hypothesis** is a statement about the parameters of one or more populations.
- Since we use probability distributions to represent populations, a statistical hypothesis may also be thought of as a statement about the probability distribution of a random variable. The hypothesis will usually involve one or more parameters of this distribution.
- $H_0 : \mu = 50$
  $H_1 : \mu \neq 50$
- H_1 is double sided. $H_1 : \mu \neq 50$ means $\mu<50$ or $\mu>50$

Test of Statistical Hypothesis

- Hypotheses -
  Null Hypothesis H_0 and Alternate Hypothesis H_1
  $H_0 : \mu = 50$ cm/sec
  $H_1 : \mu \neq 50$ cm/sec
- Take n samples $X_1, X_2 ... X_n$ of the fuel and perform the fuel burn test and find the average burning rate $\bar{X}$
- If the average burning rate $\bar{X}$ is close to the real unknown population mean μ, accept the Null Hypothesis. Else fail to accept the Null Hypothesis.

Rejection & Acceptance Region

- Suppose if the sample mean $\bar{X}$ is between 48.5 cm/sec and 51.5 cm/sec or $48.5 \leq \bar{X} \leq 51.5$ and we decide to accept the null hypothesis $H_0 : \mu = 50$ cm/sec.
- Suppose if the sample mean $\bar{X}$ is lesser than 48.5 or greater than 51.5, we decide to not accept the null hypothesis $H_0 : \mu = 50$ cm/sec or accept the alternate hypothesis $H_1 : \mu \neq 50$ cm/sec

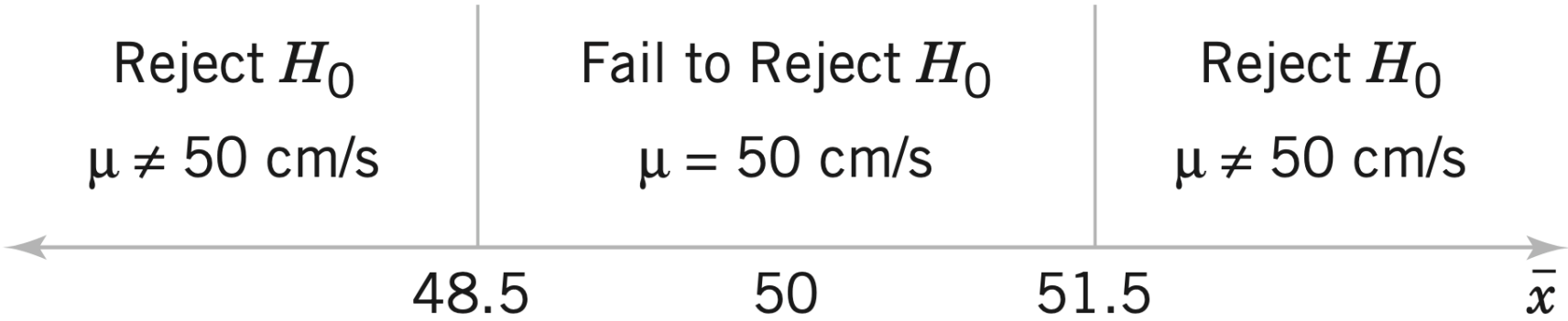


Figure: Critical Regions & Acceptance Region, Critical Values

Hypothesis Testing & Wrong Conclusions : Type I and Type II Errors

- Scenario 1 : There is a chance that the true mean μ was indeed 50 cm/sec. However, due to some sampling issue, the sample mean $\bar{X}$ was less than 48.5 or greater than 51.5. The chance or the probability that we fail to accept the null hypothesis given that the null hypothesis is true is called as the significance level α.
  α = P(Type I Error) = P(Reject H_0 when H_0 is true)
- Rejecting the null hypothesis H_0 when it is true is defined as a type I error.
- Scenario 2 : The true mean is not 50 cm/sec. However, due to some sampling issue, the sample mean $\bar{X}$ was found to be between 48.5 and 51.5. The chance or the probability that we accept the null hypothesis given that the null hypothesis is false is called as β error.
- Failing to reject the null hypothesis when it is false is defined as a type II error. β = P(Type II Error)= P(Fail to Reject H_0 when H_0 is false)

Computing Type 1 Error Probability or Significance Level α

- $\alpha = P(\bar{X} < 48.5/\mu = 50) + P(\bar{X} > 51.5/\mu = 50)$
- Assume population mean $\mu = 50$ and standard deviation of burning rate $\sigma = 2.5$ and number of samples $= 10$, the sample mean $\bar{X}$ would have a mean $\mu = 50$ and standard deviation of $\frac{\sigma}{\sqrt{n}}$ or .79

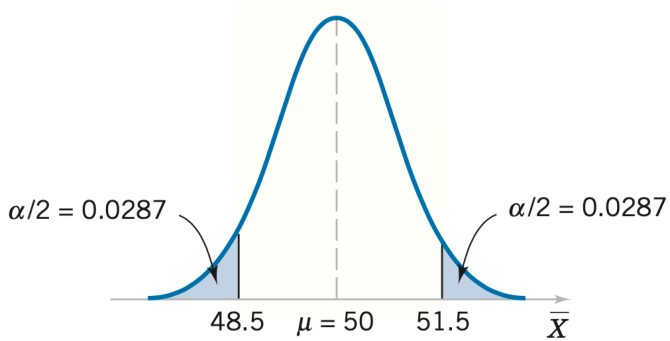


Figure: Type 1 Error

- $\alpha = P(Z < -1.9) + P(Z > 1.9) = .0287 + .0287 = .0574$
- This means that that 5.74% of all random samples would lead to rejection of the null hypothesis H_0 even if the null hypothesis is true
- Assume you repeat the experiment of taking n samples $X_1, X_2 ... X_n$ 100 times, around 5 times, you might get a rejection of null hypothesis even if the null hypothesis was really true.

Type 1 Error and Sample Size n

- If you increase the sample size, it is reasonable to expect that the type 1 error is reduced.
- Assume the sample size is increased from 10 to 16, standard deviation of $\frac{\sigma}{\sqrt{n}}$ would be .625.
- $\alpha = P(Z < -2.4) + P(Z > 2.4) = .0082 + .0082 = .0164$
- From 5.74%, we have reduced the Type 1 Error to 1.64%.

Computing Type 2 Error Probability

- The null hypothesis is false for Type 2 Error or $\mu \neq 50$. It means μ can be 51 or 52 or 49 or 48 or anything other than 50.
- The type 2 error probability can be computed only for some specific values of population mean μ
- We can compute the type 2 error probability of $\mu = 48$ and $\mu = 52$.
- In short, we are trying to compute the probability of accepting the null hypothesis H_O : $\mu = 50$ when the true mean μ is 52.
- $\beta = P(48.5 <= \bar{X} <= 51.5)$ when true mean μ is 52.
- For sample mean $\bar{X}$ with mean 52 and standard deviation $\frac{\sigma}{\sqrt{n}}$, where $\sigma = 2.5$
- $\beta = P(-4.43 <= Z <= -.63) = .2643 - .00000 = .2643$
- Thus, if we are testing H_0: $\mu = 50$ against H_1: $\mu \neq 50$ with n samples n = 10, and the true value of the mean $\mu = 52$, the probability that we will fail to reject the false null hypothesis ($\mu = 50$) is 0.2643.
- Assume you repeat the experiment of taking n samples $X_1, X_2 ... X_n$ 100 times, around 26 times, you might get a acceptance of null hypothesis even if the null hypothesis was wrong.

Computing Type 2 Error Probability

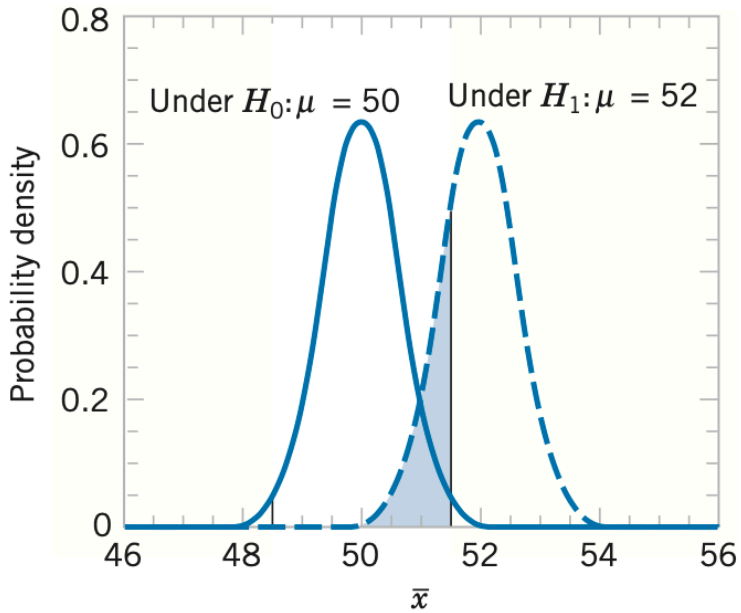


Figure: Type 2 Error

Decisions in Hypothesis Testing - Summary

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	no error	type II error
Reject H_0	type I error	no error

Figure: Decisions in Hypothesis Testing

Hypothesis Testing and Power of Statistical Test

- Type 1 and Type 2 errors are related. Both cannot be simultaneously reduced.
- The general idea is to fix the significance level α and increase the sample size n to reduce the beta error β, provided that the α is a constant.
- When the null hypothesis is false, it has to be rejected.
- The probability of rejecting a false null hypothesis is called the power of a test
- The power is the probability of rejecting a false null hypothesis
- Power of a test $=$ 1 - β, where β is the beta error or the probability of failing to reject H_0 when H_0 is false.
- When μ is 52, the power of the statistical test is 1 - .2643 $=$ .7357.

Hypothesis Testing for Population Mean with known Population Standard Deviation

Lower Tail Test	Upper Tail Test	Lower and Upper Tail Test
$H_O : \mu >= \mu_0$	$H_O: \mu <= \mu_0$	$H_O : \mu = \mu_0$
$H_a : \mu < \mu_0$	$H_O: \mu > \mu_0$	$H_O : \mu \neq \mu_0$

Table: Hypothesis Testing for Population Mean

The test statistic Z is given by

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{34}$$

σ is the population standard deviation.
Decision Making based on Z-Distribution and confidence level 100*α:
If T falls in the rejection region, reject the null hypothesis.
If T falls in the acceptance region, fail to reject the null hypothesis.

Hypothesis Testing for Population Mean with unknown Population Standard Deviation

Lower Tail Test	Upper Tail Test	Lower and Upper Tail Test
$H_O : \mu >= \mu_0$	$H_O: \mu <= \mu_0$	$H_O : \mu = \mu_0$
$H_a : \mu < \mu_0$	$H_a: \mu > \mu_0$	$H_a : \mu \neq \mu_0$

Table: Hypothesis Testing for Population Mean

The test statistic T is given by

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{35}$$

s is the sample standard deviation Decision Making based on T-Distribution with n-1 degrees of freedom and confidence level 100*α::
If T falls in the rejection region, reject the null hypothesis.
If T falls in the acceptance region, fail to reject the null hypothesis.

Hypothesis Testing for Population Proportion

Lower Tail Test	Upper Tail Test	Lower and Upper Tail Test
$H_O : p >= p_0$	$H_O: p <= p_0$	$H_O : p = p_0$
$H_a : p < p_0$	$H_a: p > p_0$	$H_a : p \neq p_0$

Table: Hypothesis Testing for Population Proportion

The test statistic T is given by

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0.(1-p_0)}{n}}} \tag{36}$$

s is the sample standard deviation Decision Making based on Z-Distribution and confidence level 100*α::
If T falls in the rejection region, reject the null hypothesis.
If T falls in the acceptance region, fail to reject the null hypothesis.

Hypothesis Testing for Variance

- Suppose that we wish to test the hypothesis that the variance of a normal population σ^2 equals to ${\sigma_0}^2$.
  $H_0 : \sigma^2 = {\sigma_0}^2$
  $H_a : \sigma^2 \neq {\sigma_0}^2$
-

$$\chi^2 = \frac{(n-1)s^2}{{\sigma_0}^2} \tag{37}$$

- The variable χ^2 has a chi-square distribution with n-1 degrees of freedom.

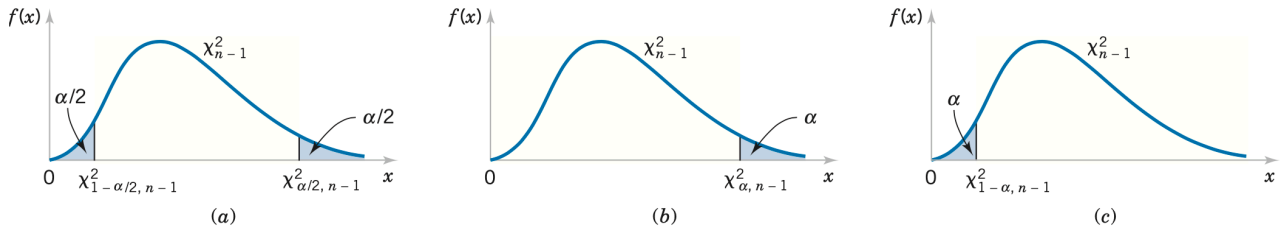


Figure 9-14 Reference distribution for the test of H_0: $\sigma^2 = \sigma_0^2$ with critical region values for (a) H_1: $\sigma^2 \neq \sigma_0^2$, (b) H_1: $\sigma^2 > \sigma_0^2$, and (c) H_1: $\sigma^2 < \sigma_0^2$.

Figure: Chi-Square Test for Hypothesis Testing of Variance

Hypothesis Testing for Variance

Tests on the Variance of a Normal Distribution

Null hypothesis: $H_0: \sigma^2 = \sigma_0^2$

Test statistic: $$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

Alternative hypothesis	Rejection criteria
$H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi_{\alpha/2,n-1}^2$ or $\chi_0^2 < -\chi_{\alpha/2,n-1}^2$
$H_1: \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi_{\alpha,n-1}^2$
$H_1: \sigma^2 < \sigma_0^2$	$\chi_0^2 < -\chi_{\alpha,n-1}^2$

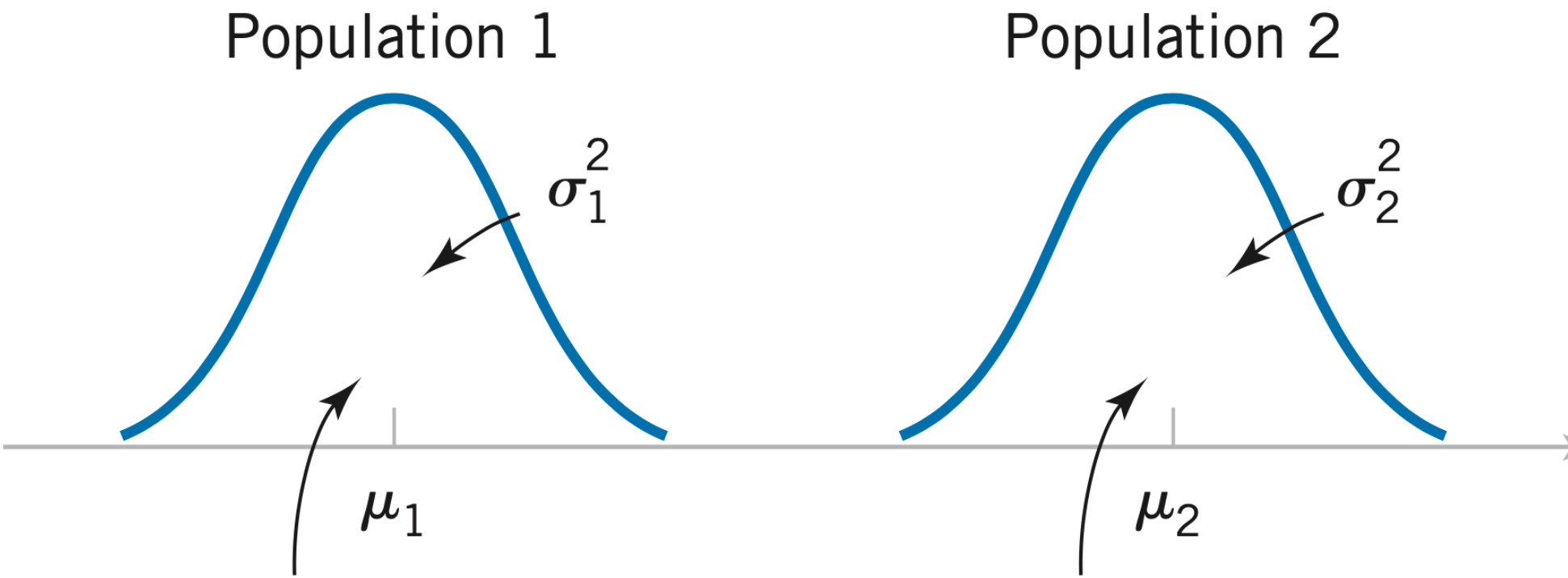
Figure: Chi-Square Test for Hypothesis Testing for Variance

Hypothesis Testing of Difference of Means, Variances Known

- $X_{11}, X_{12}, , X_{1,n_1}$ is a random sample of size n1 from population 1.
- $X_{21}, X_{22}, , X_{2,n_2}$ is a random sample of size n2 from population 2.
- The two populations represented by $X_1$and X_2 are independent.
- Both populations are normal.
- We want to draw inferences on $\mu_1 - \mu_2$, where μ_1 and μ_2 are the unknown population means of the two populations.
- The variable of interest is $\bar{X}_1 - \bar{X}_2$
- $\bar{X}_1 - \bar{X}_2$ has a mean of $\mu_1 - \mu_2$ and variance $\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}$
- Statistic Z has a normal distribution with mean 0 and variance 1.

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{38}$$

Hypothesis Testing of Difference of Means, Variances Known



Sample 1:

$x_{11}, x_{12}, \ldots, x_{1n_1}$

Sample 2:

$x_{21}, x_{22}, \ldots, x_{2n_2}$

Hypothesis Testing of Difference of Means, Variances Known

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (10\text{-}2)$$

Alternative Hypotheses	*P*-Value	Rejection Criterion For for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Figure: Two Sample Hypothesis Testing

Hypothesis Testing of Difference of Means, Variances Unknown

- $X_{11}, X_{12}, , X_{1,n_1}$ is a random sample of size n1 from population 1.
- $X_{21}, X_{22}, , X_{2,n_2}$ is a random sample of size n2 from population 2.
- Let $\bar{X_1}$,$\bar{X_2}$,${S_1}^2$,${S_2}^2$ be the sample mean and sample variances of the two groups
- $\bar{X_1} - \bar{X_2}$ has a mean of $\mu_1 - \mu_2$ and variance $\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}$
- Assume the unknown variances are equal or ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$
- We define a pooled estimator of σ^2, ${S_p}^2$

$$ {S_p}^2 = \frac{A}{B} \tag{39} $$

$$ A = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \tag{40} $$

$$ B = n_1 + n_2 - 2 \tag{41} $$

Hypothesis Testing of Difference of Means, Variances Unknown

- The pooled variance is a weighted sum of the estimated variances of the two groups
- Since pooled variance is used, this test is also called as a Pooled T Test.
- $$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{42}$$
- Test Statistic follows a T distribution with n1 + n2 - 2 degrees of freedom.

Hypothesis Testing of Difference of Means, Variances Unknown

Pooled T-Test

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (10\text{-}14)$$

Alternative Hypothesis	*P-Value*	Rejection Criterion for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert t_0\rvert$ and probability below $-\lvert t_0\rvert$	$t_0 > t_{\alpha/2,n_1+n_2-2}$ or $t_0 < -t_{\alpha/2,n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha,n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha,n_1+n_2-2}$

Figure: Pooled T-Test

Hypothesis Testing of Difference of Means, Variances Unknown

- $X_{11}, X_{12}, , X_{1,n_1}$ is a random sample of size n1 from population 1.
- $X_{21}, X_{22}, , X_{2,n_2}$ is a random sample of size n2 from population 2.
- Let $\bar{X_1}$,$\bar{X_2}$,${S_1}^2$,${S_2}^2$ be the sample mean and sample variances of the two groups
- $\bar{X_1} - \bar{X_2}$ has a mean of $\mu_1 - \mu_2$ and variance $\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}$
- Assume the unknown variances are unequal or ${\sigma_1}^2 \neq {\sigma_2}^2$

$$T = \frac{\bar{X_1} - \bar{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{43}$$

The statistic would be T-distributed with v degrees of freedom

Unpooled T-Test

Degrees of Freedom

- $$v = \frac{A}{B} \tag{44}$$

$$A = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \tag{45}$$

$$B = \frac{\left[\frac{s1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{s2^2}{n_2}\right]^2}{n_2 - 1} \tag{46}$$

- If v is not an integer, round down to the nearest integer.

Thank you!

References I