IEDA 2540 Statistics for Engineers
Comparing Multiple Samples

Wei YOU

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Spring, 2025

Introduction

In the previous topic, we considered hypothesis tests for one population. For example:

- One-sample Z-test and one-sample T-test for the mean of one normal population.
- One-sample χ^2-test for the variance of one normal population.

In many applications, we often need to compare multiple populations in terms of their means or variances.

Example: In medical trials, we need to determine if a certain treatment makes a difference. A common approach is to assign patients to two groups—the control group and the treatment group—and compare the mean responses of the two groups. This practice is sometimes referred to as A/B testing.

Methods for Comparison Among Multiple Populations

In this topic, we discuss the methods for comparing multiple populations.

- **Two Populations:**
  - **Two-sample Z-test:** Compare the means of two normal populations (known variance).
  - **Two-sample T-test:** Compare the means of two normal populations (unknown variance).
  - **Two-sample F-test:** Compare the variances of two normal populations.
- **Categorical Data:** Pearson's chi-squared test.
- **Multiple (≥ 3) Populations:** Analysis of variance (ANOVA).

Two-Sample Tests

Suppose that we observe

- $X_1, \ldots, X_{n_x} \sim N(\mu_x, \sigma_x^2)$, i.i.d. random sample;
- $Y_1, \ldots, Y_{n_y} \sim N(\mu_y, \sigma_y^2)$, i.i.d. random sample.

Example: Comparing one-sample and two-sample hypotheses.

	H_0	H_1
One-sample test	$\mu_x = \mu_0$	$\mu_x \neq \mu_0$
	$\sigma_x = \sigma_0$	$\sigma_x \neq \sigma_0$
	$\mu_y = \mu_0$	$\mu_y \neq \mu_0$
	$\sigma_y = \sigma_0$	$\sigma_y \neq \sigma_0$
Two-sample test	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$
	$\sigma_x = \sigma_y$	$\sigma_x \neq \sigma_y$

Two-Sample Z-Test – Normal Means with Known Variances

Assumptions

- The two samples $\{X_1, \ldots, X_{n_x}\}$ and $\{Y_1, \ldots, Y_{n_y}\}$ are independent.
- The two samples are normal random samples, or n_x and n_y large enough;
- The variances σ_x^2 and σ_y^2 are known, but can be unequal.

Hypotheses

$$H_0 : \mu_x = \mu_y. \quad vs. \quad H_1 : \mu_x \neq \mu_y.$$

Test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}} \sim N(0,1) \text{ under } H_0$$

p**-value**

$$p = \mathbb{P}(|Z| > |z| \mid H_0).$$

Reject H_0 is rejected at significance level α if $p \leq \alpha$, or if $\{|z| \geq z_{\alpha/2}\}$.

One- and two-sided tests

H_1	Rejection region	p-value
$\mu_x \neq \mu_y$	$\lvert z\rvert > z_{\alpha/2}$	$P(\lvert Z\rvert > \lvert z\rvert \mid H_0)$
$\mu_x > \mu_y$	$z > z_\alpha$	$P(Z > z \mid H_0)$
$\mu_x < \mu_y$	$z < -z_\alpha$	$P(Z < z \mid H_0)$

*$Z \sim N(0,1)$, $\mathbb{P}(Z > z_\alpha) = \alpha$.

- The test procedure is exactly the same as that of the one-sample Z-test, except that the definition of Z is different.

Two-Sample Z-Test – Normal Means with Known Variances

Now, suppose that we want to compare the difference in means to a non-zero value δ.

Hypotheses

$$H_0 : \mu_x - \mu_y = \delta. \quad vs. \quad H_1 : \mu_x - \mu_y \neq \delta.$$

Test statistic

$$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}} \sim N(0,1) \text{ under } H_0$$

The rest of the test procedure is exactly the same:

- p-**value**: $P(|Z| > |z| \mid H_0)$; or
- **Rejection region**: $|z| > z_{\alpha/2}$.

Two-sample T-Test – Normal Means with Unknown Equal Variance

Hypothesis

$$H_0 : \mu_x - \mu_y = 0.$$

Assumptions

- The two samples $\{X_1, \dots, X_{n_x}\}$ and $\{Y_1, \dots, Y_{n_y}\}$ are independent.
- The two samples are normal random samples, or n_x and n_y large enough;
- The variances σ_x^2 and σ_y^2 are unknown, but must be equal.

The pooled estimator for variance. Let S_x^2, S_y^2 be the sample variances

$$S_{xy}^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2} = \frac{\sum_{i=1}^{n_x}(X_i - \bar{X})^2 + \sum_{i=1}^{n_y}(Y_i - \bar{Y})^2}{n_x + n_y - 2}.$$

- $(n_x + n_y - 2)S_{xy}^2/\sigma^2 \sim \chi^2_{n_x + n_y - 2}$.

Test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_{xy}\sqrt{1/n_x + 1/n_y}} = \frac{\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma^2/n_x+\sigma^2/n_y}}}{S_{xy}/\sigma}$$ (What happens if the variance is unequal?)

$$= \frac{N(0,1)}{\sqrt{\chi^2_{n_x+n_y-2}/(n_x+n_y-2)}} \sim t_{n_x+n_y-2} \text{ under } H_0.$$

*** We omit the proof that S_{xy} is independent of $\bar{X} - \bar{Y}$, which is needed for us to have t distribution. (To be covered in advanced statistic course.)**

H_1	Rejection region	p-value
$\mu_x \neq \mu_y$	$\lvert t\rvert > t_{n_x+n_y-2,\alpha/2}$	$P(\lvert T\rvert > \lvert t\rvert \mid H_0)$
$\mu_x > \mu_y$	$t > t_{n_x+n_y-2,\alpha}$	$P(T > t \mid H_0)$
$\mu_x < \mu_y$	$t < -t_{n_x+n_y-2,\alpha}$	$P(T < t \mid H_0)$

*$T \sim t_{n_x+n_y-2}$, $\mathbb{P}(T > t_{n_x+n_y-2,\alpha}) = \alpha$.

Two-sample T-Test – Normal Means with Unknown Unequal Variance

Hypothesis:

$$H_0 : \mu_x - \mu_y = 0.$$

Assumptions:

- The two samples $\{X_1, \ldots, X_{n_x}\}$ and $\{Y_1, \ldots, Y_{n_y}\}$ are independent.
- The two samples are normal random samples (or n_x and n_y are sufficiently large).
- The variances σ_x^2 and σ_y^2 are unknown and not assumed to be equal.

Test Statistic (Welch's t-test):

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_x^2/n_x + S_y^2/n_y}} \approx t_{df} \quad \text{under } H_0,$$

with approximate degrees of freedom given by $df \approx \dfrac{\left(S_x^2/n_x + S_y^2/n_y\right)^2}{\frac{(S_x^2/n_x)^2}{n_x - 1} + \frac{(S_y^2/n_y)^2}{n_y - 1}}$.

Two-sample T-Test – Normal Means with Unknown Unequal Variance

Rejection Regions and p-values:

H_1	Rejection Region	p-value
$\mu_x \neq \mu_y$	$\lvert t \rvert > t_{df,\alpha/2}$	$p = P(\lvert T \rvert > \lvert t \rvert \mid H_0)$
$\mu_x > \mu_y$	$t > t_{df,\alpha}$	$p = P(T > t \mid H_0)$
$\mu_x < \mu_y$	$t < -t_{df,\alpha}$	$p = P(T < t \mid H_0)$

*The approximation is valid under moderate departures from normality or with large sample sizes.

Paired Observations

Recall that in the two-sample T-test, the two samples are assumed to be independent.

What if we observe pairs (X_1, Y_1), ..., (X_n, Y_n), where X_i and Y_i are correlated?

- If measurements are made on the same subject rather than on two different (independent) individuals.
  Example: Measurements before and after treatment of the same subject.
- Paired observations are not necessarily of the same subject, but still can be correlated.
  Example: Pair of siblings or family members.

Two-sample Paired T-Test

Hypothesis

$$H_0 : \mu_x - \mu_y = 0.$$

Assumptions

- $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are i.i.d. bivariate normal random vectors.

Test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_{x-y}/\sqrt{n}} \sim t_{n-1} \text{ under } H_0$$

where S_{x-y} is the sample standard deviation of the difference $\{X_1 - Y_1, \dots, X_n - Y_n\}$.

- Essentially, we are performing the one-sample T-test on the difference $X_i - Y_i$ of the observations.

Example: The following are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation:

Before (X):	45	73	46	124	33	57	83	34	26	17
After (Y):	36	60	44	119	35	51	77	29	24	11

Use the 0.05 level to test whether the safety program is effective.

- **Hypotheses**

$$H_0 : \mu_x = \mu_y. \quad vs. \quad H_1 : \mu_x > \mu_y.$$

- **Test statistics**

$$T = (\bar{X} - \bar{Y})/(S/\sqrt{10}) \sim t_9 \text{ under } H_0$$
$$t = (\bar{x} - \bar{y})/(s/\sqrt{10}) = 4.03$$

- p-**value**

$$p = P(T > 4.03 \mid H_0) = 0.0015 < 0.05 = \alpha.$$

- **Conclusion**: Reject the null hypothesis at 0.05 significance level. There is significant evidence that the safety program improves operation efficiency.

Comparing Two-Sample T-Test and Paired T-Test

	Two-Sample T-test	Paired T-test
Number of samples	n_x and n_y can be different	paired $n_x = n_y = n$
Variance	unknown and $\sigma_x^2 = \sigma_y^2 = \sigma^2$	unknown, σ_x^2 and σ_y^2 can be different
Independence	X_i and Y_i are independent	X_i and Y_i can be correlated
Order of observations	does not matter	matters, must be paired

Two-sample F-Test – Normal Variance

In the two-sample T-test, **we need the assumption that the variance of the two population are the same**. How do we test this hypothesis?

Suppose we observe $X_1, \ldots, X_{n_x}$, and $Y_1, \ldots, Y_{n_y}$.

Hypotheses

$$H_0 : \sigma_x^2 = \sigma_y^2 \Leftrightarrow \frac{\sigma_x^2}{\sigma_y^2} = 1 \quad vs. \quad H_1 : \sigma_x^2 \neq \sigma_y^2 \Leftrightarrow \frac{\sigma_x^2}{\sigma_y^2} \neq 1$$

Assumptions

- Normal random samples or large samples

Test statistic: Naturally, one would expect that the ratio $F = S_x^2/S_y^2$ can tell something about H_0.

F distribution

The F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled independent chi-squared random variables $\chi^2_{d_1}$ and $\chi^2_{d_2}$ with degrees of freedom of d_1 and d_2, respectively.

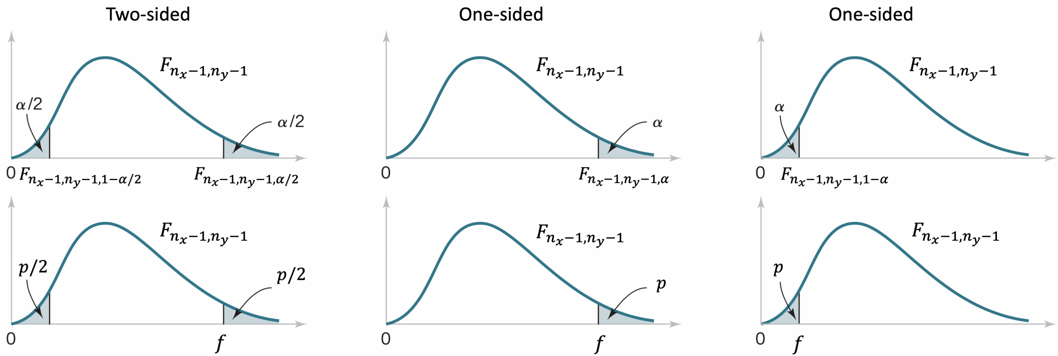
$$F = \frac{\chi^2_{d_1}/d_1}{\chi^2_{d_2}/d_2} \sim F_{d_1,d_2}.$$

- It becomes relevant when we try to calculate the ratio of sample variances of normally distributed statistics.

  Test statistic:

$$F = \frac{S_x^2}{S_y^2} = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} = \frac{\frac{(n_x-1)S_x^2}{\sigma_x^2}/(n_x-1)}{\frac{(n_y-1)S_y^2}{\sigma_y^2}/(n_y-1)} \sim F_{n_x-1,n_y-1} \text{ under } H_0$$

- The distribution is skewed to the right, and the F-values can only be positive.
- Useful facts: $F_{1,d_2} = t^2_{d_2}$, $F_{d_1,\infty} = \chi^2_{d_1}/d_1$



H_1	Rejection region	p-value
$\sigma_x^2 \neq \sigma_y^2$	$f > F_{n_x-1,n_y-1,\alpha/2}$ or $f < F_{n_x-1,n_y-1,1-\alpha/2}$	$2 \times \min\{P(F > f \mid H_0), P(F < f \mid H_0)$
$\sigma_x^2 > \sigma_y^2$	$f > F_{n_x-1,n_y-1,\alpha}$	$P(F > f \mid H_0)$
$\sigma_x^2 < \sigma_y^2$	$f < F_{n_x-1,n_y-1,1-\alpha}$	$P(F < f \mid H_0)$

*$F \sim F_{n_x-1,n_y-1}$, $\mathbb{P}(F > F_{n_x-1,n_y-1,\alpha}) = \alpha$.

Example: Want to check whether there is less variability in the silver plating done by company 1 than company 2. If independent random samples of size $n_1 = n_2 = 12$ of the two companies' work yield $s_1 = 0.035$ and $s_2 = 0.062$, test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 < \sigma_2^2$ at the 0.05 level.

- **Test statistics**

$$F = S_1^2/S_2^2 \sim F_{n_1-1,n_2-1} \text{ under } H_0.$$

$$f = s_1^2/s_2^2 = 0.318.$$

- p-**value**

$$p = P(F < f \mid H_0) = P(F < 0.318 \mid H_0) = 0.0352,$$

  where $F \sim F_{11,11}$ under H_0.

- **Conclusion**: Reject H_0 because $p < \alpha = 0.05$. There is significant statistical evidence in favor of the claim that company 1 has better silver plating quality.
- Alternatively, reject because $0.318 < f_{11,11,0.95} = 0.355$.

Categorical Data

Recall that an important type of variable is the **categorial variable**, where the values it takes are discrete and unordered.

- These variables cannot be considered as normally distributed and hence the previous tests such as two-sample t-test and ANOVA cannot be applied.
- New method is needed.

Example: Let us study how smoking affects cardiovascular health. Suppose we sampled a group of people and collect information on whether they smoke or not and whether they suffer from heart disease or not. The data classified by two different variables, each of which has only two possible outcomes.

- Is the subject smoking: "yes" and "no".
- Does the subject have heart disease: "yes" and "no".

Categorical Data

Example: Let us study how smoking affects cardiovascular health. Suppose we collected the following data

- The data classified by two different variables, each of which has only two possible outcomes.
  - Is the subject smoking: "yes" and "no".
  - Does the subject have heart disease: "yes" and "no".

Smoking	Heart disease		**Total**
	Yes	No	
Yes	O_{11}	O_{12}	$n_{1\cdot}$
No	O_{21}	O_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

* "O" stands for "observed"

Question: Does smoking change the population proportions of subjects with heart disease? In other words, is smoking independent of the risk of getting heart disease?

2×2 Contingency Table

In general, this table is called a 2×2 contingency table.

X	Y		**Total**
	Yes	No	
Yes	O_{11}	O_{12}	$n_{1\cdot}$
No	O_{21}	O_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

* "O" stands for "observed"

Hypotheses

$$H_0 : \;\; X \text{ and } Y \text{ are independent.} \quad vs \quad H_1 : \;\; X \text{ and } Y \text{ are correlated.}$$

Example:

Say, we observe the following data.

X	Y		**Total**
	Yes	No	
Yes	20	80	100
No	15	135	150
Total	35	215	250

- Would you believe there is a significant difference in the proportion?
    - 20% in the smoking group versus 10% in the non-smoking group.
    - Is it purely by randomness? Is it reliable difference?
- We'll start, as always, by formulating the null hypothesis H_0.

$$H_0 : \text{Smoking and heart disease are independent}$$

- What would we expect to see under H_0?

Example:

What is the expected number of smoking with heart disease?

- The overall proportion of heart disease is $\frac{35}{250} = 0.14$.
- If the proportion are the same for the smoking and non-smoking groups, then we should expect the number of smoking people with heart disease to be around $\frac{35}{250} \times 100 = 14$.
- Similarly, we the expected number of
  - smoking people without heart disease is $\frac{215}{250} \times 100 = 86$.
  - non-smoking people with heart disease is $\frac{35}{250} \times 150 = 21$.
  - non-smoking people without heart disease is $\frac{215}{250} \times 150 = 129$.

Example:

Now, we can compare what we expect (under the null hypothesis)with what we actually observe, and see how much they differ—If the difference is large, then we should not believe in H_0 : the proportion is the same.

Expected:

X	Y		**Total**
	Yes	No	
Yes	14	86	100
No	21	129	150
Total	35	215	250

Observed:

X	Y		**Total**
	Yes	No	
Yes	20	80	100
No	15	135	150
Total	35	215	250

How to formulate this comparison with a mathematical expression?

χ^2-Test for 2×2 Contingency Table

Under H_0, i.e. the independence of X and Y, we have

$$\mathbb{P}(\{X = \text{“Y”}\} \cap \{Y = \text{“Y”}\}) = \mathbb{P}(\{X = \text{“Y”}\}) \times \mathbb{P}\left(\{Y = \text{“Y”}\}\right).$$

It is natural to consider the estimations

$$\mathbb{P}(\{X = \text{“Y”}\}) \approx \frac{n_{1\cdot}}{n}$$

$$\mathbb{P}(\{Y = \text{“Y”}\}) \approx \frac{n_{\cdot 1}}{n}$$

$$\frac{O_{11}}{n} \approx \mathbb{P}(\{X = \text{“Y”}\} \cap \{Y = \text{“Y”}\}) \approx \frac{n_{1\cdot}}{n} \times \frac{n_{\cdot 1}}{n}$$

Note that $n\mathbb{P}(\{X = \text{“Y”}\} \cap \{Y = \text{“Y”}\})$ is the expected number observation of $E_{11} \triangleq \{X = \text{“Y”}\} \cap \{Y = \text{“Y”}\} \Rightarrow O_{11} \approx E_{11}$.

Example:

We call O_{11} the *observed number of observations* for smokers with heart disease and $E_{11} = \frac{n_{1\cdot}}{n} \times \frac{n_{\cdot 1}}{n}$ the *expected number of observations* for smokers with heart disease **under the null hypothesis** H_0.

- The expected number should match the observed number if H_0 were to be true.
- We can repeat this for the other three groups of people:
  - $O_{10} \approx E_{10}$: smoker without heart disease
  - $O_{01} \approx E_{01}$: non-smoker with heart disease
  - $O_{00} \approx E_{00}$: non-smoker without heart disease

In summary, we should expect $O_{i,j} \approx E_{i,j}$ for all $i, j = 0, 1$.

- How do we measure the discrepancy between $O_{i,j}$ and $E_{i,j}$?
- We can use the **sum of squares** of the difference between $O_{i,j}$ and $E_{i,j}$.
- The larger the difference, the more evidence against H_0.

Pearson's Chi-Squared Test of Independence

- **Test statistic**

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi_1^2 \text{ under } H_0$$

  - The approximation is good if every entry has at least 5 observations.
  - The proof for it is beyond the scope of this course.

  If H_0 is not true, then O_{ij} is not likely to be close to E_{ij}, and the test statistic X^2 tends to be large.

- p-**value** for the observed value x^2

$$p = \mathbb{P}(X^2 > x^2 \mid H_0), \text{ where } X^2 \sim \chi_1^2.$$

- Rejection region $[\chi^2_{1,\alpha}, \infty)$.

$R \times C$ Contingency Table

More generally, each of the two categorical variable may have more than 2 possible values (0,1 in the previous case).

Example: Let X denote the previous working experience, with possible value of "¡5 years", "5-10 years" or "¿10 years." Let Y denote the job rank offered to the candidate, e.g., "Analyst", "Head of Business", "Partner."

- We aim to check if the previous working experience is correlated to the rank offered.

$$H_0: \; X \text{ and } Y \text{ are independent.} \quad vs \quad H_1: \; X \text{ and } Y \text{ are correlated.}$$

- In general, X may take R possible values and Y may take C possible values. How should we test this hypothesis?

$R \times C$ Contingency Table

In general, we can consider the $R \times C$ contingency table.

X	Y				**Total**
	1	2	$\cdots$	C	
1	O_{11}	O_{12}	$\cdots$	O_{1C}	$n_{1\cdot}$
2	O_{21}	O_{22}	$\cdots$	O_{1C}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
R	O_{R1}	O_{R2}	$\cdots$	O_{RC}	$n_{R\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot C}$	n

Hypotheses

$H_0 :$ X and Y are independent. vs $H_1 :$ X and Y are correlated.

Similarly, E_{ij} estimates the expected number of observations in the ij-th entry, where

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}, \quad i = 1, 2, \ldots, R \text{ and } j = 1, 2, \ldots, C.$$

- **Test statistic**

$$X^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2_{(R-1)(C-1)} \text{ under } H_0$$

  The approximation is good if

  - (1) no more than $1/5$ of the cells have expected values < 5; and
  - (2) no cell has expected value < 1.

- If H_0 is not true, then O_{ij} is not likely to be close to E_{ij}, and the test statistic X^2 tends to be large.
- p-**value** for the observed value x^2

$$p = \mathbb{P}(X^2 > x^2 \mid H_0), \text{ where } X^2 \sim \chi^2_{(R-1)(C-1)}.$$

- Rejection region $[\chi^2_{(R-1)(C-1),\alpha}, \infty)$.

Example: Samples of three kinds of materials, subjected to extreme temperature changes, produced the results shown below:

Material	Under extreme heat		**Total**
	Crumbled	Intact	
A	41	79	120
B	27	53	80
C	22	78	100
Total	90	210	300

Use the 0.05 level to test whether the probability of crumbling is the same for the three kinds of materials.

Hypothesis

$$H_0: \text{ Crumbling prob. is independent of the matrials.}$$

Test statistic

$$X^2 = \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(3-1)(2-1)} \text{ under } H_0$$

$$x^2 = \frac{(41 - 120 \times 90/300)^2}{120 \times 90/300} + \frac{(79 - 120 \times 210/300)^2}{120 \times 210/300} + \cdots = 4.575.$$

p-**value**

$$p = \mathbb{P}(X^2 > x^2 \mid H_0) = \mathbb{P}(X^2 > 4.575 \mid H_0) = 0.1015 > 0.05 = \alpha,$$

where $X^2 \sim \chi^2_2$.

Conclusion: Fail to reject H_0, there is no evidence showing that these materials show different crumbling probabilities.

Alternatively, since $\chi^2 = 4.575 < 5.991 = \chi^2_{2,0.05}$, fail to reject reject the null.

Comparing more than two populations

Previously, we saw several ways to compare samples from multiple populations.

- Two groups of normal populations ⇒ Two-sample t-test
    - Variances are known
    - Variances are unknown but equal
    - Variances are unknown but unequal[1]
- Two or more groups of categorical data
    - 2×2 contingency table
    - $R \times C$ contingency table

[1] Use the F-test to check whether variances are equal.

Comparing more than two populations

Question: How do we compare multiple normal populations?

Example: How does three brands of fertilizer affects the growth of plants (e.g. weight of the potatos)?

- The weight of the potato can be regarded as a continuous variable.
- The brands of the fertilizer can be regarded as a categorical variable (Brand A, Brand B, Brand C).

Question: Is there significant difference in the effectiveness of the fertilizers?

$$H_0 : \mu_A = \mu_B = \mu_C . \quad \text{v.s.} \quad H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j.$$

Essentially, we are asking: how does a continuous variable depend on a categorical variable?

Example: Suppose that we are comparing the mean weight of the potatos, applied with 10 brands of fertilizers.

- A straight forward choice is to apply two-sample T test to very pair of fertilizers.
- If H_0 in at least one test is rejected, we say that there is a significant difference.
- There are $\binom{10}{2} = \frac{10\times 9}{2} = 45$ different pairs $\Rightarrow$ 45 tests to be done!

Problem with performing multiple two-sample T tests

- **Inefficient**: number of tests needed grows very fast as number of brands grows.
- **Large Type-I error**:
  - If $\alpha = 0.05$ for each comparison, there is a 5% chance of making a Type-I error: a pair of samples falsely regarded as significantly different.
  - For 10 independent pairs, the probability of making a Type I error at least once is $1 - 0.95^{10} = 0.40$!

Is there a smarter way to compare multiple populations?

Overview

Y	–	X
Response	–	Explanatory
Dependent	–	Independent
Outcome	–	Predictor

Y	X	Test	Lecture
Categorical	Categorical	Pearson's χ^2-test	Previous lecture
Continuous	Binary	Two-sample t-test	
	Categorical	(One-way) ANOVA	This lecture
Continuous	Continuous	Simple linear regression	Next lecture(s)
Binary	Continuous	Logistic regression	(Probably) not covered

ANOVA

ANalysis **O**f **VA**riance (**ANOVA**) is a collection of statistical models used to compare means for multiple (≥ 3) independent populations.

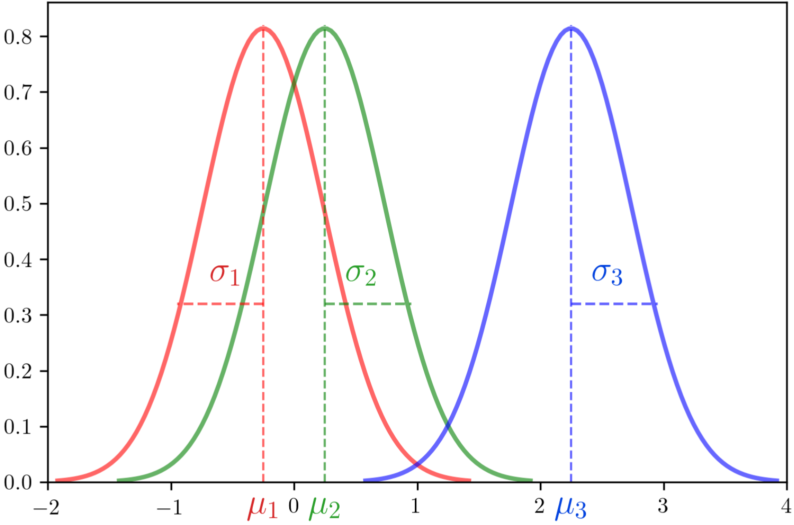
We only focus on the simplest case of ANOVA

- Assume that we have k independent, normally distributed groups with equal variance

$$X_1 \sim N(\mu_1, \sigma^2),\ X_2 \sim N(\mu_2, \sigma^2), \ldots, X_k \sim N(\mu_k, \sigma^2)$$

- We wish to determine if there is a significant difference among the population means of these groups, i.e.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k. \quad \text{v.s.} \quad H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j.$$



$H_0 : \mu_1 = \mu_2 = \mu_3.$ v.s. $H_1 :$ At least one mean is significantly different from others.

Assumptions

$$\begin{gathered}
X_{11}, X_{12}, \ldots, X_{1n_1} \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2) \\
X_{21}, X_{22}, \ldots, X_{2n_2} \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2) \\
\vdots \\
X_{k1}, X_{k2}, \ldots, X_{kn_k} \overset{i.i.d.}{\sim} N(\mu_k, \sigma^2)
\end{gathered}$$

- The k samples are mutually independent.
- Each sample $\{X_{i1}, X_{i2}, \ldots, X_{in_i}\}$ is a normal random sample - we check this condition with histograms and Q-Q plots.
- The population variance is the same across all k groups - we check this condition by comparing sample standard deviations.

Checking the ANOVA Assumptions

To check whether a sample comes from a normally distributed population, we have two main choice:

- Histogram
- Q-Q plot

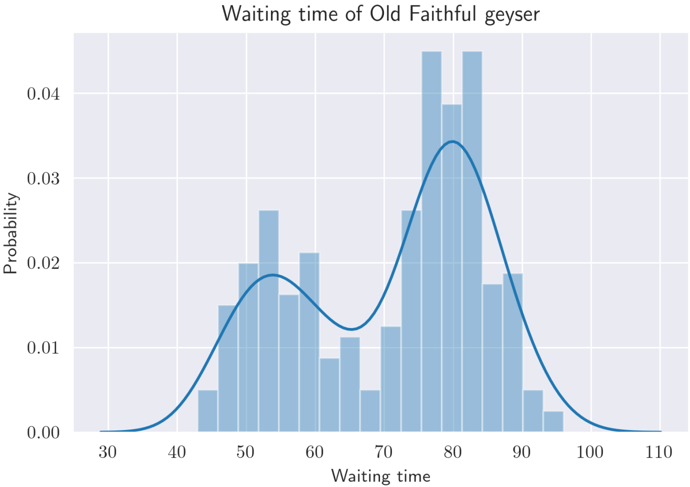
We now demonstrate both approaches using the `old-faithful` data.

Checking the ANOVA Assumptions

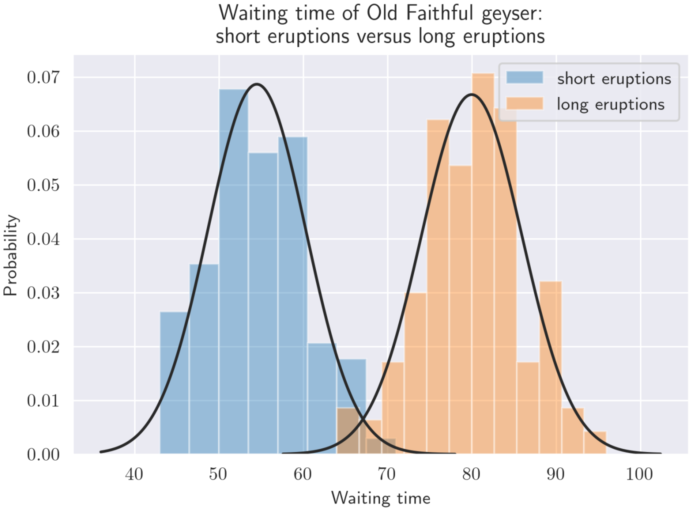
Example: Old Faithful is a cone geyser in Yellowstone National Park. It erupts every 40-100 mins, and each eruption lasts for 1-5 mins.
We observe the eruptions and record the following two variables:

- The duration in minutes of the eruption.
- The duration in minutes until the next eruption.

Checking the Normality Assumptions – Histogram



Checking the Normality Assumptions – Histogram



Checking the Normality Assumptions – Q-Q Plot

Quantile-quantile plot

In statistics, a Q-Q (quantile-quantile) plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

We use the Q-Q plot to check if the data can be regarded as a normal random sample.

- Step 1: Given a sample $\{x_1, \ldots, x_n\}$, normalized the data by subtracting the sample mean $\bar{x}$ and dividing the sample standard deviation s. So $\tilde{x}_i = (x_i - \bar{x})/s$.
- Step 2: Sort the (normalized) sample in increasing order $\{\tilde{x}_{(1)}, \ldots, \tilde{x}_{(n)}\}$, then $\tilde{x}_{(i)}$ is (approximately) the $\frac{i}{n}$-th sample quantile.
- Step 3: For each $\tilde{x}_{(i)}$, calculate z_i, the $\frac{i}{n}$-th population quantile of the standard normal distribution. Then plot the point $(z_i, \tilde{x}_{(i)})$.

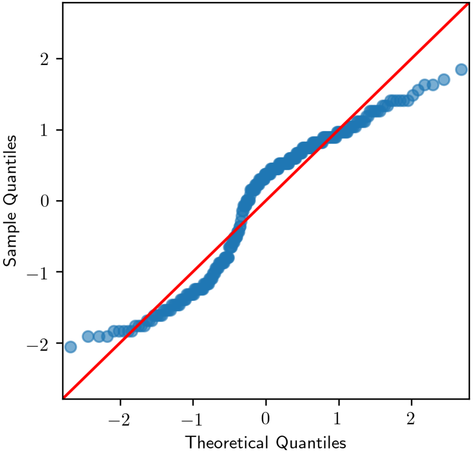
Decision: If $z_i \approx \tilde{x}_{(i)}$ for all i, then the sample is approximately normally distributed.

Checking the Normality Assumptions – Q-Q Plot

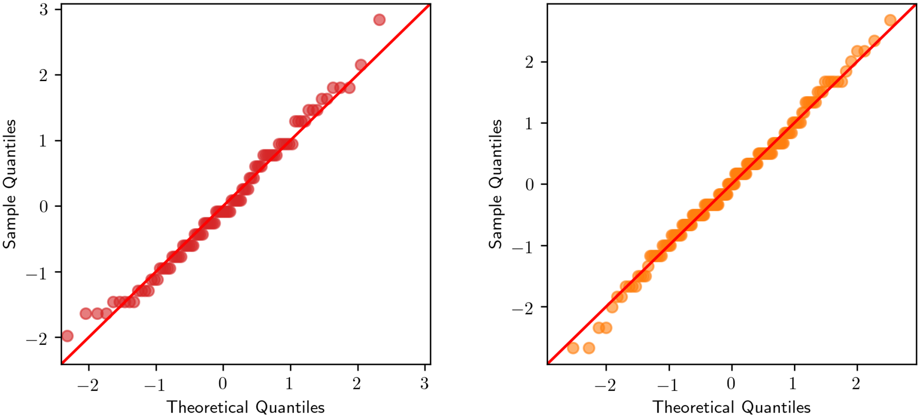
The sample is approximately normally distributed, if

- $z_i \approx \tilde{x}_{(i)}$ for all i; or equivalently
- if the points lines-up in a 45° line (the red line).

For the Old Faithful waiting time, we cannot assume normality.

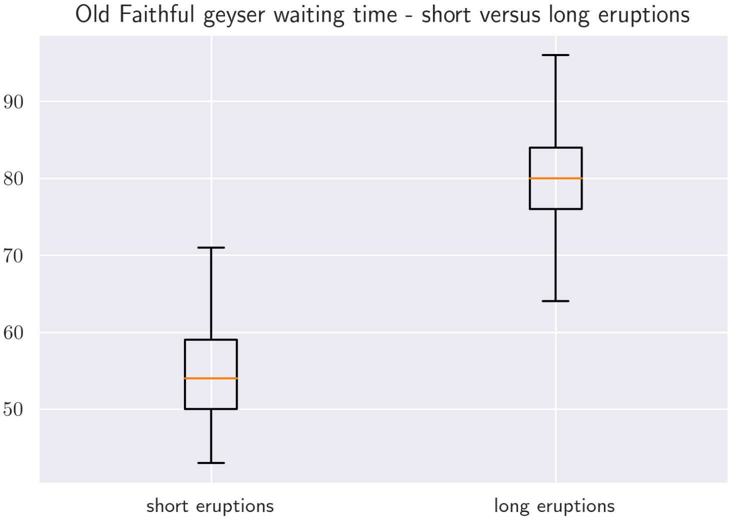


Checking the Normality Assumptions – Q-Q plot



Can assume normality for long/short eruptions considered separately.

Visual check of the Homogeneity of Variance – Boxplot



*Checking the Homogeneity of Variance – Bartlett's Test

Recall that we also need to assume equal variance across different samples.

- To test equal variance for $k = 2$ populations $\Rightarrow$ two-sample F-test.

What about $k \geq 3$ populations?

- Pair-wise comparison.
- A smarter way: **Bartlett's Test**. Here is the `Python` code:

```
from scipy.stats import bartlett
T, p = bartlett(sample1, sample2, sample3,...)
```

Here `T` is the value of the test statistic and `p` is the p-value.

Example: Old Faithful – short vs. long eruptions. We have $p = 0.77$, so we fail to reject H_0 and can assume equal variance at any significance level $\alpha < 0.77$.

One-Way ANOVA: Comparing k Normal Means

Model: For $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$,

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

where μ_i is the mean of the i-th group and all ε_{ij} are independent.

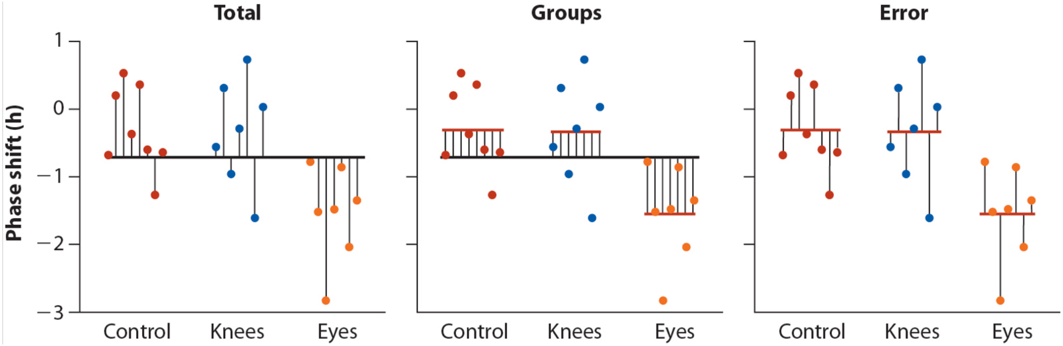
Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \text{not all } \mu_i \text{ are equal.}$$

Assumptions:

- Each group is a random sample from a $\mathcal{N}(\mu_i, \sigma^2)$ population.
- All groups share the same variance σ^2.
- Observations are independent both within and between groups.

Intuition – Variation Between/Within Groups



Consider $H_0 : \mu_1 = \cdots = \mu_k \equiv \mu$ versus $H_1 : \mu_i \neq \mu_j,$ for some $i \neq j$.

If H_0 is true, **(variation in the entire data)** $\approx$ **(variation within each group)**.

Intuition – Quantifying the Variation

To quantify the variation between/within groups

- Let $\bar{X}_{i\cdot}$ be the sample mean of the ith group.
- Let $\bar{X}$ be the sample mean of all observations.

For variation in a single observation, we write

$$\underbrace{X_{ij} - \bar{X}}_{\text{Deviation from mean}} \quad = \quad \underbrace{(\bar{X}_{i\cdot} - \bar{X})}_{\text{between groups}} \quad + \quad \underbrace{(X_{ij} - \bar{X}_{i\cdot})}_{\text{within group}}$$

We then aggregate the above to obtain variation in all the samples:

$$SS_{\mathrm{T}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}((\bar{X}_{i\cdot} - \bar{X}) + (X_{ij} - \bar{X}_{i\cdot}))^2$$

Measuring Variation in One-Way ANOVA

Within-Group Variation: Each group's variability is measured by its sum of squared deviations from its own mean:

$$SS_{\mathrm{W}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \big(X_{ij} - \bar{X}_{i\cdot}\big)^2, \quad \bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}.$$

Between-Group Variation: The variability of the group means around the overall mean is

$$SS_{\mathrm{B}} = \sum_{i=1}^{k} n_i \big(\bar{X}_{i\cdot} - \bar{X}\big)^2, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}, \quad n = \sum_{i=1}^{k} n_i.$$

Sum of Squares and ANOVA Intuition

Sums of Squares

- $SS_{\mathrm{T}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$ is called the **total sum of squares**.
- $SS_{\mathrm{W}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\cdot}\right)^2$ is called the **sum of squares within groups**.
- $SS_{\mathrm{B}} = SS_{\mathrm{B}} = \sum_{i=1}^{k} n_i \left(\bar{X}_{i\cdot} - \bar{X}\right)^2$ is called the **sum of squares between groups**.

Intuition:

- If the population means $\mu_1, \ldots, \mu_k$ are all equal (H_0 true), then the group means $\bar{X}_{i\cdot}$ will vary only due to within-group noise.
- Thus, we expect

$$SS_{\mathrm{T}} \approx SS_{\mathrm{W}} \iff SS_{\mathrm{B}} \ll SS_{\mathrm{W}}.$$

- A very small ratio $SS_{\mathrm{B}}/SS_{\mathrm{W}}$ suggests H_0 is likely true.

ANOVA F-Test

F-Statistic:

$$F = \frac{SS_{\mathrm{B}}/(k-1)}{SS_{\mathrm{W}}/(N-k)} = \frac{MS_{\mathrm{B}}}{MS_{\mathrm{W}}} \sim F_{k-1,N-k} \quad \text{under } H_0.$$

- $N = \sum_{i=1}^{k} n_i$ is the total sample size.
- $MS_{\mathrm{B}} = SS_{\mathrm{B}}/(k-1)$ is the **mean square between groups**.
- $MS_{\mathrm{W}} = SS_{\mathrm{W}}/(N-k)$ is the **mean square within groups**.
- Large values of F indicate that MS_{B} is large relative to MS_{W}, suggesting significant differences among group means.

Decision Rule

Reject H_0 at level α if

$$F > F_{k-1,\,N-k,\alpha}.$$

One-Way ANOVA

Hypotheses

$$H_0 : \mu_1 = \cdots = \mu_k \equiv \mu \text{ versus } H_1 : \mu_i \neq \mu_j, \text{ for some } i \neq j.$$

ANOVA table

Source	df	Sum of squares	Mean squares	F-statistic
Between	$k-1$	$SS_B = \sum_{i=1}^{k} n_i(\bar{X}_{i\cdot} - \bar{X})^2$	$MS_B = SS_B/(k-1)$	$F = \frac{MS_B}{MS_W}$
Within	$N-k$	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2$	$MS_W = SS_W/(N-k)$	
Total	$N-1$	$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$		

Test Statistics:

$$F = \frac{MS_B}{MS_W} \sim F_{k-1,N-k} \text{ under } H_0,$$

$$f = \frac{ms_B}{ms_W}.$$

ANOVA p-Value

How to we determin the p-**value**?

- The p-value is the probability that the test statistic is more "extreme" than the observed value.
- Recall that
$$F = \frac{MS_B}{MS_W} \sim F_{k-1,N-k} \text{ under } H_0,$$
If the variation between groups (MS_B) is sufficiently large, in compare with the variation within groups (MS_W), then H_0 is not likely to be true.
- Hence, $p = \mathbb{P}(F > f \mid H_0)$ where $F \sim F_{k-1,N-k}$.

Conclusion: reject if $p < \alpha$. Alternatively, reject if $f > F_{k-1,N-k,\alpha}$.

*Remark: Decomposition of Total Sum of Squares

Partitioning the Total Sum of Squares

$$SS_T = SS_B + SS_W.$$

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}((\bar{X}_{i\cdot}-\bar{X})+(X_{ij}-\bar{X}_{i\cdot}))^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot}-\bar{X})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot}-\bar{X})(X_{ij}-\bar{X}_{i\cdot}) + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot}-\bar{X})^2 + 2\sum_{i=1}^{k}(\bar{X}_{i\cdot}-\bar{X})\underbrace{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})}_{=0} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2 \\
&= \sum_{i=1}^{k}n_i(\bar{X}_{i\cdot}-\bar{X})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2
\end{aligned}$$

Example

A group of researchers conducted a clinical trial to determine the effectiveness of 3 pain relievers.

- 3 experimental groups (Advil, Tylenol, Aleve) and 1 placebo group
- 4 people in each group
- Each participant completed a pain self-assessment on a scale of 0 (no pain) - 10 (extreme pain).

A: Advil	B: Tylenol	C: Alevel	D: placebo
3	2	1	6
4	1	2	9
2	0	1	7
1	9	2	10

Example

Hypotheses:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

Calculation:

$$\bar{x}_{1\cdot} = 2.5, \bar{x}_{2\cdot} = 3, \bar{x}_{3\cdot} = 1.5, \bar{x}_{4\cdot} = 8, \bar{x} = 3.75$$

$$s_1 = 1.29, s_2 = 4.08, s_3 = 0.58, s_4 = 1.83$$

$$ss_B = \sum_{i=1}^{4} (x_{i\cdot} - \bar{x})^2 = 4((2.5 - 3.75)^2 + (3 - 3.75)^2 + (1.5 - 3.75)^2 + (8 - 3.75)^2) = 101$$

$$ss_W = \sum_{i=1}^{4} \sum_{j=1}^{4} (x_{ij} - \bar{x}_{i\cdot})^2 = \sum_{i=1}^{4} (n_1 - 1)s_i^2 = \sum_{i=1}^{4} 3s_i^2 = 3(1.29^2 + 4.08^2 + 0.58^2 + 1.83^2) = 65.9874$$

$$ms_B = 101/(4 - 1) = 33.67$$

$$ms_W = 65.9874/(16 - 4) = 5.50$$

Test Statistics

$$F \sim F_{3,12} \text{ under } H_0$$

$$f = \frac{ms_B}{ms_W} = \frac{33.67}{5.5} \approx 6.12$$

ANOVA table

Source	Degree of freedom	Sum of squares	Mean squares	F-statistic
Between	$4-1=3$	101	33.67	6.12
Within	$16-4=12$	65.9874	5.50	
Total	$16-1=15$	166.9874		

Conclusion: Set the significant level $\alpha = 0.05$. Reject H_0 since $f = 6.12 > 3.49 = F_{3,12,0.05}$. We can conclude that at least one pair of the groups has significantly different mean pain scores.

ANOVA: Possible Outcomes and Next Steps

We tested

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \text{not all } \mu_i \text{ are equal.}$$

- **Fail to reject H_0:** No significant difference among the group means.
  Example: For example, if vital-sign measurements across different dosage groups are statistically the same, then the drug shows no effect.
- **Reject H_0:** At least one group mean differs significantly from the others.
  Example: This indicates that dosage has an effect on the vital signs.

Next Question: "*Where do the differences lie?*"

Tukey's HSD Post-Hoc Test

We will now introduce post-hoc multiple comparison procedures to pinpoint which group(s) differ.

- Tukey's Honestly Significant Difference (HSD) test performs all pairwise comparisons among group means while controlling the family-wise error rate.
- For each pair (i, j), it:
  - Tests $H_0 : \mu_i = \mu_j$ versus $H_1 : \mu_i \neq \mu_j$.
  - Constructs a simultaneous $(1 - \alpha) \times 100\%$ confidence interval for $\mu_i - \mu_j$.
- **Equal sample sizes required:** The standard Tukey HSD assumes $n_1 = n_2 = \cdots = n_k$.
- The test statistic is based on the studentized range distribution; derivation is beyond the scope of this course.