

통계분석

Statistical Analysis

Topics in Hypothesis Testing

Two-Sample Tests

- So far we have discussed hypothesis testing on a single population.
- We can also apply hypothesis testing on comparison between two distinct populations.

Basic Assumption: Two-Sample Tests

X_1, X_2, \dots, X_m : m random sample from a distribution of (μ_x, σ_x^2)

Y_1, Y_2, \dots, Y_n : n random sample from a distribution of (μ_y, σ_y^2)

Samples $\{X_i, Y_j\}$ are independent.

We want to know $\mu_x = \mu_y$.

- Null hypothesis: $\mu_x - \mu_y = 0$
- Alternative hypothesis: $\mu_x - \mu_y \neq 0$

Test Statistic: Two-Sample Test

X_1, X_2, \dots, X_m : m random sample from a distribution of (μ_x, σ_x^2)

Y_1, Y_2, \dots, Y_n : n random sample from a distribution of (μ_y, σ_y^2)

Samples $\{X_i, Y_j\}$ are independent.

Test Statistic = $\bar{X} - \bar{Y}$

- unbiased estimator of $\mu_x - \mu_y$
- $E(\bar{X} - \bar{Y}) = \mu_x - \mu_y$ $\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$

Two-Sample Test: Case (1)

Two independent normal populations

X_1, X_2, \dots, X_m : m random sample from $N(\mu_x, \sigma_x^2)$

Y_1, Y_2, \dots, Y_n : n random sample from $N(\mu_y, \sigma_y^2)$

Variances σ_x^2, σ_y^2 are known, but means are unknown.

Test Statistic = $\bar{X} - \bar{Y}$

- unbiased estimator of $\mu_x - \mu_y$
- $E(\bar{X} - \bar{Y}) = \mu_x - \mu_y$ $\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$
- $\bar{X} - \bar{Y}$ is normally distributed

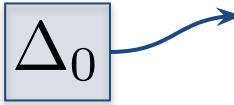
Standardizing test statistic $Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}$ 0 (Null Hypothesis)

Two-Sample Test (1)

Two independent normal populations

Standardizing test statistic
$$Z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}$$

Applying z-test to the standardized test statistic.

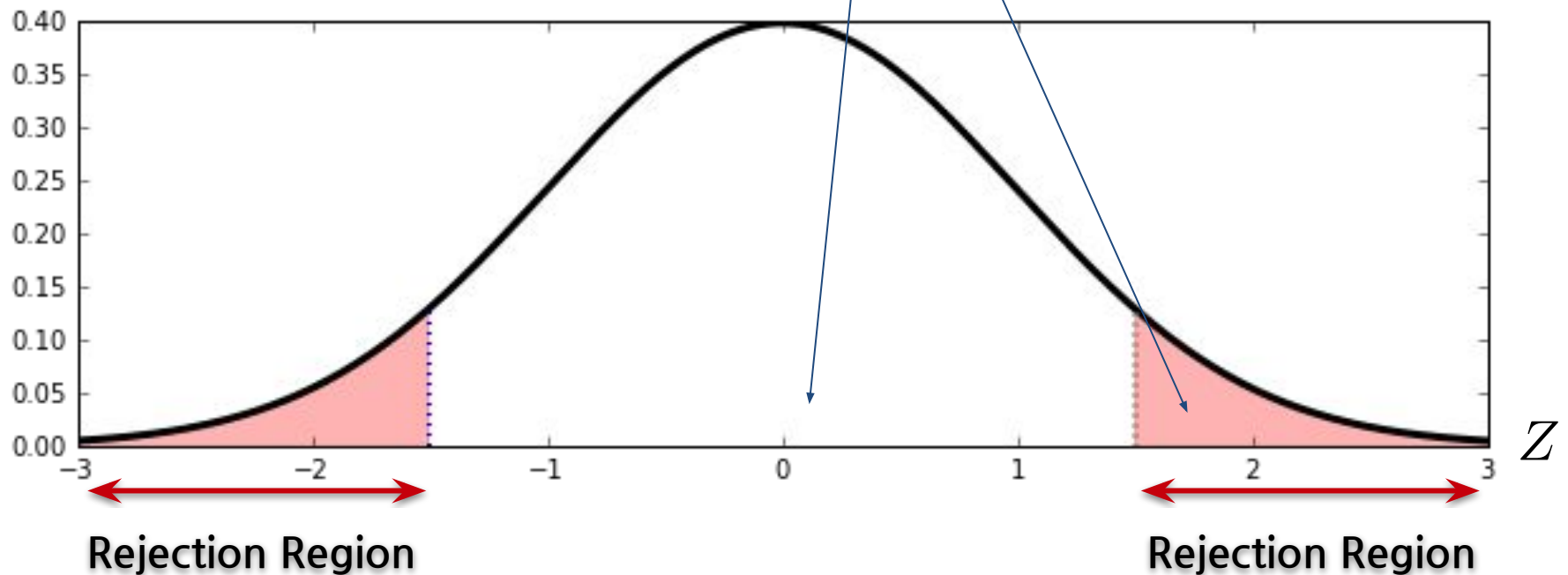
- Null hypothesis $H_0 : \mu_x - \mu_y = \Delta_0$  We can generalize the previous case
- Alternative hypothesis $H_a : \mu_x - \mu_y \neq \Delta_0$

Two-Sample Test (1)

Two independent normal populations

Standardizing test statistic $Z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}$

Applying z-test to the standardized test statistic.



Two-Sample Test (1)

Analysis of a random sample consisting of $m=20$ specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of **29.8 ksi**. A second random sample of $n=25$ two-sided galvanized steel specimens gave a sample average strength of **34.7 ksi**. Assuming that the two yield-strength distributions are **normal** with standard deviations **4.0 and 5.0** (suggested by a graph in the article “Zinc-Coated Sheet Steel: An Overview,” Automotive Engr., Dec. 1984: 39–43), does the data indicate that the corresponding true average yield strengths are different? Let’s carry out a test at significance level 0.01.

Two-Sample Test (2): Large Sample Case

Two independent populations

X_1, X_2, \dots, X_m : m random sample from population with μ_x, σ_x

Y_1, Y_2, \dots, Y_n : n random sample from population with μ_y, σ_y

- Here, we do **NOT** assume that they are **normally distributed**.
- We do **NOT** assume that the population variances are known, too.
- Both m and n are **large enough** for Central limit theorem.

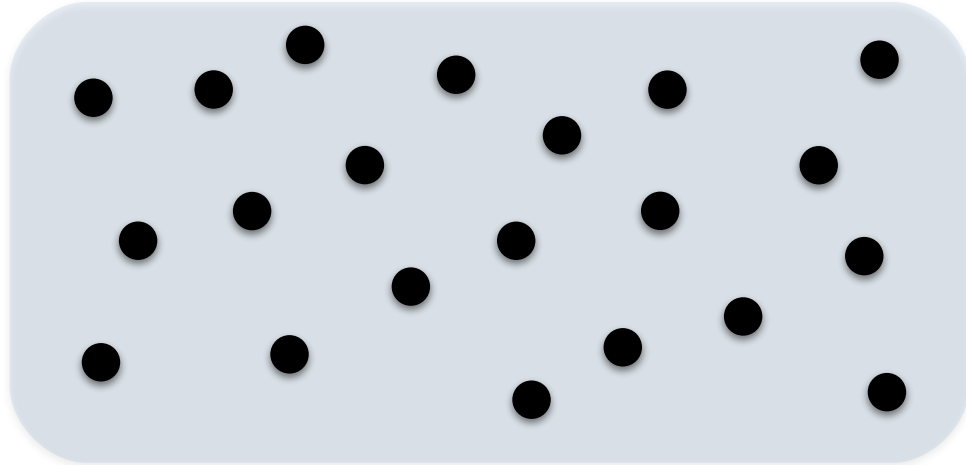
→ $\bar{X} - \bar{Y}$ is approximately normal

→ Standardizing test statistic
$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}}$$

Unknown σ_x^2, σ_y^2 is replaced by S_x^2, S_y^2 .

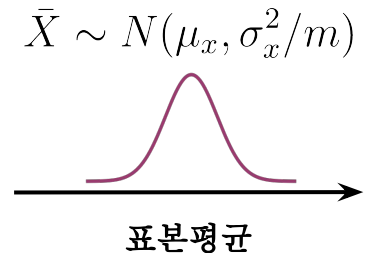
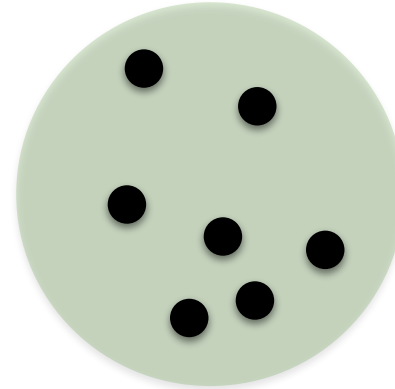
Two-Sample Test (2): Large Sample Case

Population #1 (모집단 #1)



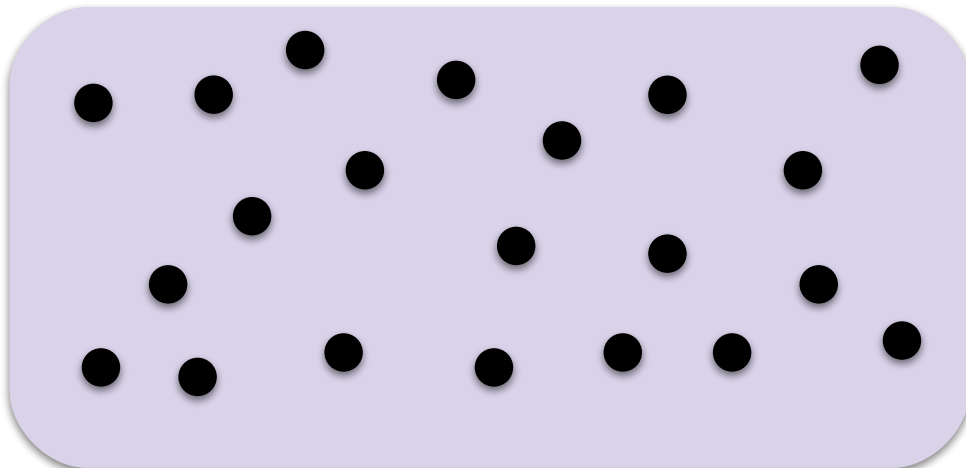
Distribution (Population) is not necessarily normal.

Large Sample (표본) X



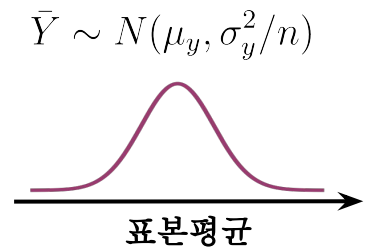
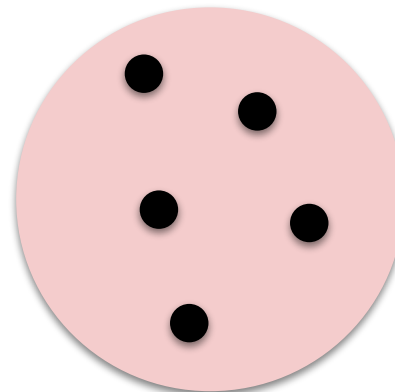
Due to Central Limit Theorem

Population #2 (모집단 #2)



Distribution (Population) is not necessarily normal.

Large Sample (표본) Y



중심극한정리에 의하여
두 표본평균은 모두 정규분포를 만족한다.

Two-Sample Test (2): Large Sample Case

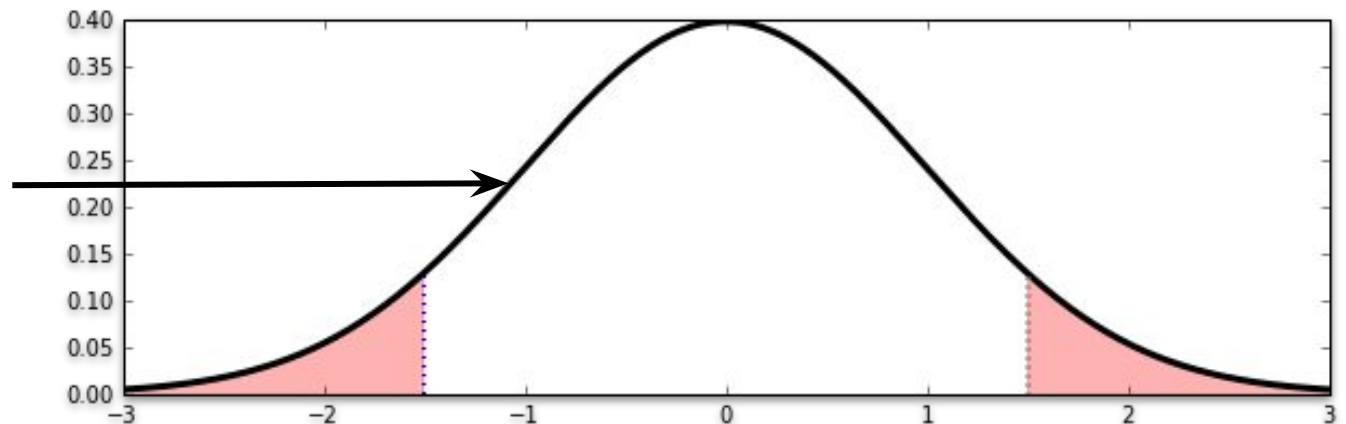
→ $\bar{X} - \bar{Y}$ is approximately normal

→ Standardizing test statistic $Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}}$

Applying z-test to the standardized test statistic.

- Null hypothesis $H_0 : \mu_x - \mu_y = \Delta_0$
- Alternative hypothesis $H_a : \mu_x - \mu_y \neq \Delta_0$ (Two-sided)

$$z = \frac{(\bar{x} - \bar{y}) - \Delta_0}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}}$$

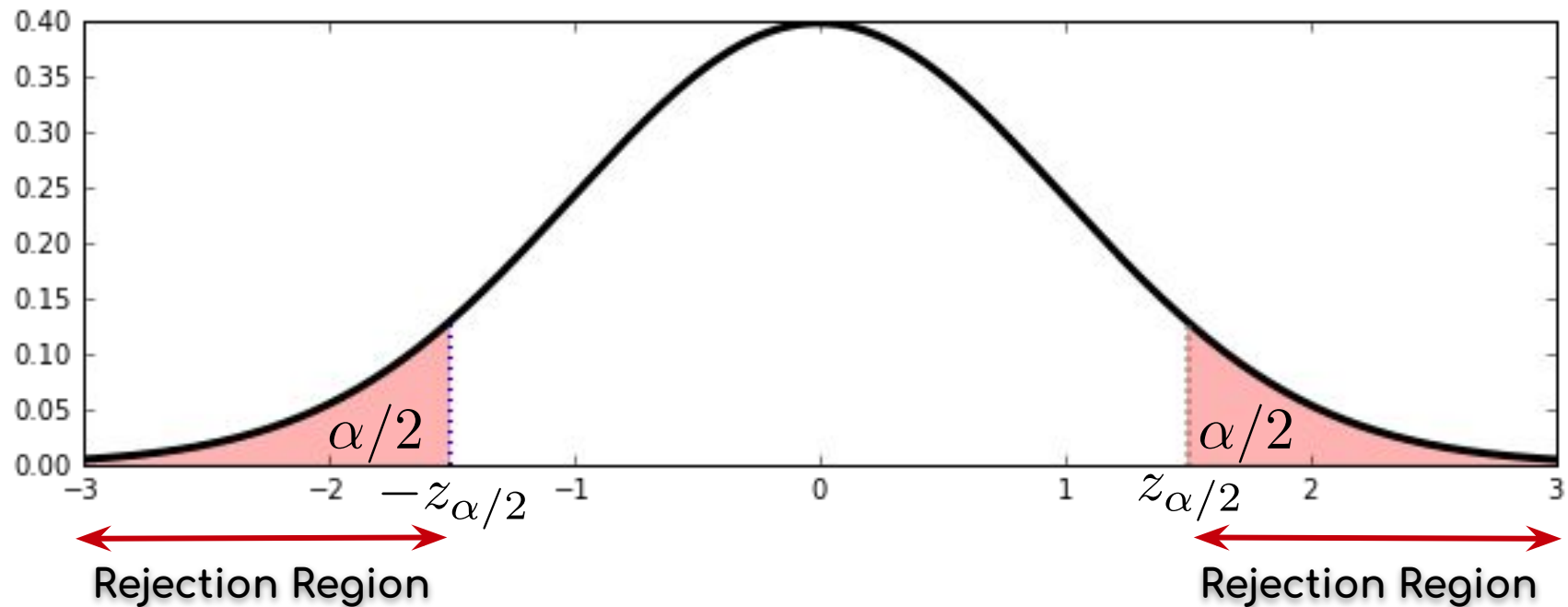


Two-Sample Test (2): Large Sample Case

$$z = \frac{(\bar{x} - \bar{y}) - \Delta_0}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}}$$

sample means: \bar{x} , \bar{y}

sample variances: s_x^2 , s_y^2



significance level = α

Two-Sample Test (2): Large Sample Case

How large are m and n ?

→ When does the Central Limit Theorem hold?

→ $m > 40$ and $n > 40$

Two-Sample Test (2): Example

What impact does fast-food consumption have on various dietary and health characteristics? The article “Effects of Fast-Food Consumption on Energy Intake and Diet Quality Among Children in a National Household Study” (Pediatrics, 2004: 112–118) reported the accompanying summary data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat Fast Food	Sample Size	Sample Mean	Sample Std
No (Y)	663	2258	1519
Yes (X)	413	2637	1138

Does this data provide strong evidence for concluding that true (population) average calorie intake for teens who typically eat fast food exceeds by more than 200 calories per day the true (population) average intake for those who don't typically eat fast food? Let's investigate by carrying out a test of hypotheses at a significance level of approximately 0.05.

$$\mu_x - \mu_y = 200 \text{ (Null)}$$

$$\mu_x - \mu_y > 200 \text{ (Alternative) (Upper-sided)}$$

$$\alpha = 0.05$$

Two-Sample Test (2): Example

Eat Fast Food	Sample Size	Sample Mean	Sample Std
No (X)	663	2258	1519
Yes (Y)	413	2637	1138

μ_1 = true average calorie intake for teens who don't eat fast food

μ_2 = true average calorie intake for teens who eat fast food

$$\begin{aligned} H_0 : \mu_1 - \mu_2 &= -200 \\ H_a : \mu_1 - \mu_2 &< -200 \end{aligned} \quad \Rightarrow \quad z = \frac{\bar{x} - \bar{y} - (-200)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

Two-Sample Test (2): Example

Eat Fast Food	Sample Size	Sample Mean	Sample Std
No	663	2258	1519
Yes	413	2637	1138

- **Test Statistic** $\implies z = \frac{\bar{x} - \bar{y} - (-200)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{2637 - 2258 - (-200)}{\sqrt{\frac{1519^2}{663} + \frac{1138^2}{413}}} \approx -2.20$
- **Significance level** $\alpha = 0.05 \longrightarrow z_{\alpha=0.05} = 1.645$
- **Rejection region (lower-tailed)** $z < -z_{0.05} = -1.645$
 $-2.20 < -1.645 \implies H_0$ is rejected.

Two-Sample Test (2): Example

Eat Fast Food	Sample Size	Sample Mean	Sample Std
No	663	2258	1519
Yes	413	2637	1138

- **Test Statistic** $\implies z = \frac{\bar{x} - \bar{y} - (-200)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{2637 - 2258 - (-200)}{\sqrt{\frac{1519^2}{663} + \frac{1138^2}{413}}} \approx -2.20$
- **Significance level** $\alpha = 0.05 \longrightarrow z_{\alpha=0.05} = 1.645$
- **P-value (lower-tailed)** $p = \Phi(-2.20) = 0.0139$

$0.0139 < 0.05 \implies H_0$ is rejected.

Two-Sample Test: Case (3)

Two independent populations

X_1, X_2, \dots, X_m : m random sample from population with μ_x, σ_x

Y_1, Y_2, \dots, Y_n : n random sample from population with μ_y, σ_y

- Both m and n are ***NOT large***, so we cannot use *Central limit theorem*.
- We assume that they are ***normally distributed (nearly normal)***.
- Normality can be justified by drawing a normal probability plot of x and y.
- We do not know population variances here.

→ $\bar{X} - \bar{Y}$ is no longer normally distributed.

→ Instead, $\bar{X} - \bar{Y}$ follows the t-distribution.

Two-Sample Test: Case (3)

Assuming that X_i and Y_j are normal,

$\bar{X} - \bar{Y}$ (approximately) follows the t -distribution as follows.

Standardizing T test statistic
$$T = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}}$$

Degree of Freedom ν

$$\nu \approx \frac{\left[\frac{s_x^2}{m} + \frac{s_y^2}{n} \right]^2}{\frac{(s_x^2/m)^2}{m-1} + \frac{(s_y^2/n)^2}{n-1}}$$

rounded down to the nearest integer (ex. $\nu = 29.4 \rightarrow \nu = 29$)

Two-Sample t-Test: Example

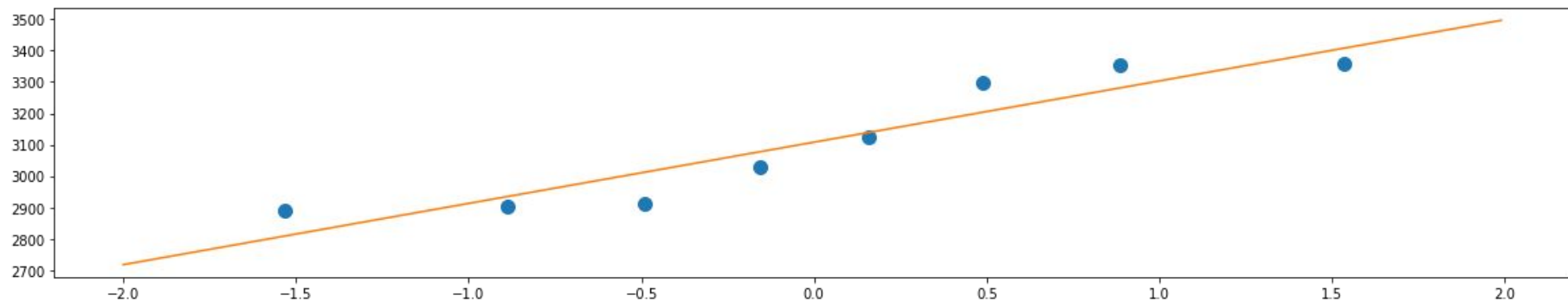
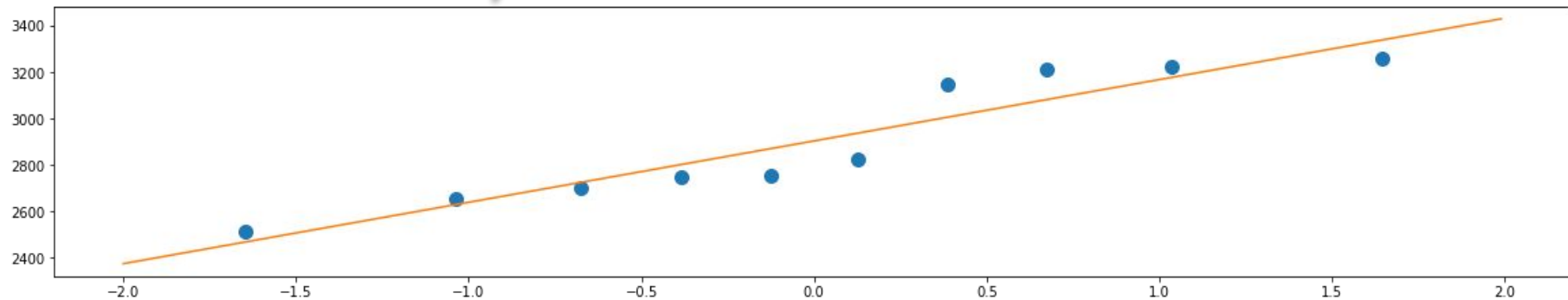
The deterioration of many municipal pipeline networks across the country is a growing concern. One technology proposed for pipeline rehabilitation uses a flexible liner threaded through existing pipe. The article “Effect of Welding on a High-Density Polyethylene Liner” (J. of Materials in Civil Engr., 1996: 94–100) reported the following data on tensile strength (psi) of liner specimens both when a certain fusion process was used and when this process was not used.

No Fusion	2748	2700	2655	2822	2511	3149	3257	3213	3220	2753	m = 10
	Sample Size		10	Sample Mean			Sample Std				
Fusion	3027	3356	3359	3297	3125	2910	2889	2902			n = 8
	Sample Size		8	Sample Mean			Sample Std				

Two-Sample t-Test: Example

No Fusion	2748	2700	2655	2822	2511	3149	3257	3213	3220	2753
	Sample Size			Sample Mean			Sample Std			
Fusion	3027	3356	3359	3297	3125	2910	2889	2902		
	Sample Size			Sample Mean			Sample Std			

● Normal Probability Plot



Two-Sample Test (4) : Pooled t Procedure

Two independent populations

X_1, X_2, \dots, X_m : m random sample from population with μ_x, σ_x

Y_1, Y_2, \dots, Y_n : n random sample from population with μ_y, σ_y

We assume that they are (1) *normally distributed (nearly normal)*, and that

(2) *their variances are equal*. $\sigma_x^2 = \sigma_y^2 = \sigma^2$

But, *we do not know the value of the variance here.*

Two-Sample Test: Pooled t -Procedure

Assuming that X_i and Y_j are normal with the same variance,

$\bar{X} - \bar{Y}$ (approximately) follows the t -distribution as follows.

- Standardizing T test statistic
$$T = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2 \left(\frac{1}{m} + \frac{1}{n} \right)}}$$

- Pooled (Combined) estimator of σ^2

$$S_p^2 = \frac{m-1}{m+n-2} S_x^2 + \frac{n-1}{m+n-2} S_y^2$$

- Degree of Freedom ν

$$\nu = m + n - 2$$