# STP 226
# ELEMENTARY STATISTICS
### CHAPTER 10

### INFERENCES FOR TWO POPULATION MEANS

## 10.1 The Sampling Distribution of the difference between two sample means for independent samples

- **Independent samples** – samples taken at random from a population and that one sample selected from one population has no bearing on the sample selected from the other population.

**Sampling distribution of $\bar{x}_1 - \bar{x}_2$**

Suppose x is a normally distributed variable on each of two populations. Then, for independent samples of size $n_1$ and $n_2$ from two populations,

- $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$
- $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}$
- $\bar{x}_1 - \bar{x}_2$ is normally distributed.

**The two-sample z-procedures (two-sample z-test), $\sigma$ known**

Test statistic $$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$$

## 10.2 Inferences for two population means using independent samples: standard deviations assumed equal ($\sigma_1 = \sigma_2 = \sigma$)

We need to use sample information to estimate the common population standard deviation $\sigma$. We can regard the two sample variances $s_1^2, s_2^2$ as estimates for population variance $\sigma^2$. And pool those estimates by weighting them according to sample size. The resulting estimate called "pooled sample variance", $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$. By taking square root of it, we get the estimate of population standard deviation, $s_p$.

**Distribution of the pooled t-statistic**
Suppose x is a normally distributed variable on each of two populations and that the population standard deviations are equal. Then, for independent samples of sizes $n_1$ and $n_2$ from two populations, the variable

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{(1/n_1) + (1/n_2)}}$$ has a **t-distribution** with **df = $n_1 + n_2$ - 2**

Where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$ is the pooled sample standard deviation

**The pooled t-test for two population means**

ASSUMPTIONS

1. Independent samples
2. Normal populations or large samples
3. Equal population standard deviations ( OR )

STEPS

1. The null hypothesis is $H_0 : \mu_1 = \mu_2$ and the alternate hypothesis is one of the following:
   $H_a : \mu_1 \neq \mu_2$ (two-tailed)
   $H_a : \mu_1 < \mu_2$ (left-tailed)
   $H_a : \mu_1 > \mu_2$ (right-tailed)
2. Decide on the significance level, $\alpha$.
3. The critical value(s) are

   $\pm t_{\alpha/2}$ (two-tailed)
   $- t_{\alpha}$ (left-tailed)
   $z_{\alpha}$ (right-tailed)
   with df = $n_1 + n_2 - 2$ , use Table IV to find the critical value(s).
4. Compute the value of the test statistic $t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{(1/n_1) + (1/n_2)}}$

   where $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
5. If the value of the test statistic falls in the rejection region, reject $H_0$; otherwise, do not reject $H_0$.
6. State the conclusion in words.

The hypothesis test is exact for normal populations and is approximately correct for large samples from non-normal populations.

**The Pooled t-interval procedure for two population means**

ASSUMPTIONS

1. Independent samples
2. Normal populations or large samples
3. Equal population standard deviations

STEPS

1. For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with df = $n_1 + n_2 - 2$.

2. The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} . s_p \sqrt{(1/n_1) + (1/n_2)}$$

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

**What if the assumptions are not satisfied?**

1. Not independent samples: If paired, use the procedure in Section 10.4
2. Not normal and small samples: Use nonparametric method, *Mann-Whitney* test or *Mann-Whitney* confidence interval.
3. Unequal population standard deviations: If the problem satisfies assumptions 1 and 2, then use the procedure in Section 10.3.

## 10.3 Inferences for two population means using independent samples (standard deviations not assumed equal)

If the population standard deviations are known, use the z-interval procedure/test

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$$

**Distribution of the non-pooled t-statistic**

Suppose x is a normally distributed variable on each of two populations. Then, for independent samples of sizes $n_1$ and $n_2$ from the populations, the variable

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}}$$ has approximately a t-distribution.

The degrees of freedom used is obtained from the sample data; it is

$$\Delta = \frac{[(s_1^2/n_1)+(s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$$ rounded down to the nearest integer.

- **The Non-Pooled t-test for two population means**

ASSUMPTIONS

1. Independent samples
2. Normal populations or large samples

STEPS

1. The null hypothesis is $H_0 : \mu_1 = \mu_2$ and the alternate hypothesis is one of the following:
   $H_a : \mu_1 \neq \mu_2$ (two-tailed)
   $H_a : \mu_1 < \mu_2$ (left-tailed)
   $H_a : \mu_1 > \mu_2$ (right-tailed)
2. Decide on the significance level, $\alpha$.
3. The critical value(s) are
   $\pm t_{\alpha/2}$ (two-tailed)
   $- t_{\alpha}$ (left-tailed)
   $z_{\alpha}$ (right-tailed)
   with df = $\Delta$ , where

$$\Delta = \frac{[(s_1^2/n_1)+(s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$$ rounded down to the nearest integer

Use Table IV to find the critical value(s).

4. Compute the value of the test statistic $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1)+(s_2^2/n_2)}}$
5. If the value of the test statistic falls in the rejection region, reject $H_0$; otherwise, do not reject $H_0$.
6. State the conclusion in words.

- **The non-pooled t-interval procedure for two population means**

ASSUMPTIONS

1. Independent samples
2. Normal populations or large samples

STEPS

1. For a confidence level of 1 - $\alpha$, use Table IV to find $t_{\alpha/2}$ with df = $\Delta$, where

$$\Delta = \frac{[(s_1^2/n_1)+(s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$$ , rounded down to the nearest integer

2. The endpoints of the confidence interval for $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{(s_1^2/n_1)+(s_2^2/n_2)}$$

- **Choosing between a pooled and non-pooled procedure**

Suppose you want to compare the means of two populations using independent samples. When deciding between a pooled t-procedure and a non-pooled t-procedure, follow these guidelines: If you are reasonably sure the populations have nearly equal standard deviations, used a pooled t-procedure; otherwise, use a non-pooled t-procedure.

## 10.4 Inferences for two population means using paired samples

### The paired t-statistic

Paired difference variable, d of x is the difference between the two corresponding x's This implies that the difference between the population means is

$$\mu_d = \mu_1 - \mu_2$$

### Distribution of the paired t-statistic

Suppose x is a variable on each of two populations whose members can be paired. Further suppose that the paired-difference (normal differences) variable d is normally distributed. Then, for paired samples of size n, the variable

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{s_d / \sqrt{n}}$$ has the t-distribution with df = n - 1

- **The paired t-test for two population means**

ASSUMPTIONS

1. Paired sample
2. Normal differences of large sample

STEPS

1. The null hypothesis is $H_0 : \mu_1 = \mu_2$ and the alternate hypothesis is one of the following:

   $H_a : \mu_1 \neq \mu_2$ (two-tailed)

$H_a : \mu_1 < \mu_2$ (left-tailed)
$H_a : \mu_1 > \mu_2$ (right-tailed)

2. Decide on the significance level, $\alpha$.
3. The critical value(s) are
   $\pm t_{\alpha/2}$ (two-tailed)
   $-t_{\alpha}$ (left-tailed)
   $t_{\alpha}$ (right-tailed)

   with df = n - 1 , where Use Table IV to find the critical value(s).
4. Compute the value of the test statistic $t = \dfrac{\bar{d}}{s_d / \sqrt{n}}$
5. If the value of the test statistic falls in the rejection region, reject $H_0$; otherwise, do not reject $H_0$.
6. State the conclusion in words.

The hypothesis test is exact when the paired-difference variable is normally distributed (normal differences) and is approximately correct for large samples when the paired-difference variable is not normally distributed (non-normal differences).

- **The paired t-interval procedure for two population means**

ASSUMPTIONS

1. Paired sample
2. Normal differences or large samples

STEPS

1. For a confidence level of 1 - $\alpha$, use Table IV to find $t_{\alpha/2}$ with df = n – 1.
2. The endpoints of the confidence interval for $\mu_1 - \mu_2$ are

   $$\bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$$

The confidence interval is exact when the paired-difference variable is normally distributed (normal differences) and is approximately correct for large samples when the paired-difference variable is not normally distributed (non-normal differences).

**What if the assumptions are not satisfied?**

In case you have neither normality of the paired-difference variable nor a large sample, then a nonparametric method should be used.