Hypothesis testing for two populations

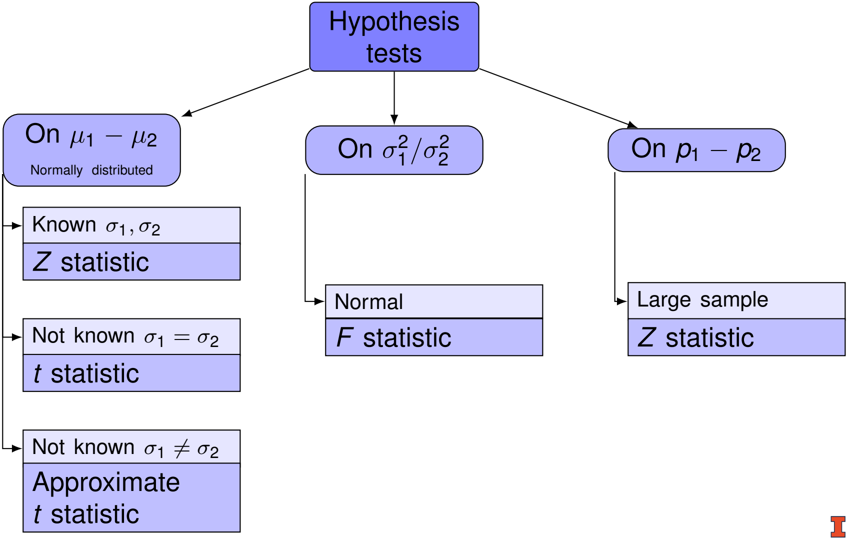
Chrysafis Vogiatzis

Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Lectures 28-29

ILLINOIS
ISE | Industrial & Enterprise Systems Engineering
GRAINGER COLLEGE OF ENGINEERING

Overview



Hypothesis tests
On $\mu_1 - \mu_2$
Normally distributed
Known σ_1, σ_2
Z statistic
Not known $\sigma_1 = \sigma_2$
t statistic
Not known $\sigma_1 \neq \sigma_2$
Approximate t statistic
On σ_1^2/σ_2^2
Normal
F statistic
On $p_1 - p_2$
Large sample
Z statistic

Hypothesis testing for $\mu_1 - \mu_2$

What if we are interested in testing what the difference between two means is?

- μ_1: unknown mean of population 1.
- μ_2: unknown mean of population 1.

We distinguish between:

1. normally distributed populations with known variances σ_1^2, σ_2^2.
2. normally distributed populations with unknown variances that are known to be equal, that is unknown $\sigma_1^2 = \sigma_2^2$.
3. normally distributed populations with unknown variances that are not known to be equal, that is unknown $\sigma_1^2 \neq \sigma_2^2$.

Hypothesis testing for $\mu_1 - \mu_2$

What if we are interested in testing what the difference between two means is?

- μ_1: unknown mean of population 1.
- μ_2: unknown mean of population 1.

We distinguish between:

1. normally distributed populations with known variances σ_1^2, σ_2^2.
2. normally distributed populations with unknown variances that are known to be equal, that is unknown $\sigma_1^2 = \sigma_2^2$.
3. normally distributed populations with unknown variances that are not known to be equal, that is unknown $\sigma_1^2 \neq \sigma_2^2$.

Hypothesis testing for $\mu_1 - \mu_2$

What if we are interested in testing what the difference between two means is?

- μ_1: unknown mean of population 1.
- μ_2: unknown mean of population 1.

We distinguish between:

1. normally distributed populations with known variances σ_1^2, σ_2^2.
2. normally distributed populations with unknown variances that are known to be equal, that is unknown $\sigma_1^2 = \sigma_2^2$.
3. normally distributed populations with unknown variances that are not known to be equal, that is unknown $\sigma_1^2 \neq \sigma_2^2$.

Hypothesis testing for $\mu_1 - \mu_2$

What if we are interested in testing what the difference between two means is?

- μ_1: unknown mean of population 1.
- μ_2: unknown mean of population 1.

We distinguish between:

1. normally distributed populations with known variances σ_1^2, σ_2^2.
2. normally distributed populations with unknown variances that are known to be equal, that is unknown $\sigma_1^2 = \sigma_2^2$.
3. normally distributed populations with unknown variances that are not known to be equal, that is unknown $\sigma_1^2 \neq \sigma_2^2$.

ILLINOIS

Hypothesis testing for $\mu_1 - \mu_2$

What if we are interested in testing what the difference between two means is?

- μ_1: unknown mean of population 1.
- μ_2: unknown mean of population 1.

We distinguish between:

1. normally distributed populations with known variances σ_1^2, σ_2^2.
2. normally distributed populations with unknown variances that are known to be equal, that is unknown $\sigma_1^2 = \sigma_2^2$.
3. normally distributed populations with unknown variances that are not known to be equal, that is unknown $\sigma_1^2 \neq \sigma_2^2$.

Means with known σ_1, σ_2

Null hypothesis: | Test statistic: | Distribution:

$$H_0 : \mu_1 - \mu_2 = \Delta_0.$$

$$Z_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

$$Z_0 \sim \mathcal{N}(0, 1).$$

H_1	Rejection region	P-value
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$	$2 \cdot (1 - \Phi(\lvert Z_0 \rvert))$
$\mu_1 - \mu_2 > \Delta_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$\mu_1 - \mu_2 < \Delta_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

ILLINOIS

Means with unknown $\sigma_1 = \sigma_2 = \sigma$

First, define a pooled estimator for the variance:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}.$$

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

Test statistic: $$T_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{s_p\sqrt{1/n_1 + 1/n_2}}.$$

Distribution: $T_0 \sim T_{n_1+n_2-2}.$

H_1	Rejection region	P-value
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert T_0 \rvert > t_{\alpha/2, n_1+n_2-2}$	$2 \cdot (1 - T_{n_1+n_2-2}(\lvert T_0 \rvert))$
$\mu_1 - \mu_2 > \Delta_0$	$T_0 > t_{\alpha, n_1+n_2-2}$	$1 - T_{n_1+n_2-2}(T_0)$
$\mu_1 - \mu_2 < \Delta_0$	$T_0 < -t_{\alpha, n_1+n_2-2}$	$T_{n_1+n_2-2}(T_0)$

ILLINOIS

Means with unknown $\sigma_1 = \sigma_2 = \sigma$

First, define a pooled estimator for the variance:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}.$$

Null hypothesis:

$$H_0 : \mu_1 - \mu_2 = \Delta_0.$$

Test statistic:

$$T_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{s_p \sqrt{1/n_1 + 1/n_2}}.$$

Distribution:

$$T_0 \sim T_{n_1+n_2-2}.$$

H_1	Rejection region	P-value
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert T_0 \rvert > t_{\alpha/2, n_1+n_2-2}$	$2 \cdot (1 - T_{n_1+n_2-2}(\lvert T_0 \rvert))$
$\mu_1 - \mu_2 > \Delta_0$	$T_0 > t_{\alpha, n_1+n_2-2}$	$1 - T_{n_1+n_2-2}(T_0)$
$\mu_1 - \mu_2 < \Delta_0$	$T_0 < -t_{\alpha, n_1+n_2-2}$	$T_{n_1+n_2-2}(T_0)$

ILLINOIS

Means with unknown $\sigma_1 \neq \sigma_2$

We can no longer calculate a pooled estimator for the variance: instead, we need to use the two sample standard deviations in the place of the actual ones.

We can no longer calculate the actual degrees of freedom: we estimate the approximate degrees of freedom as

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}}.$$

Means with unknown $\sigma_1 \neq \sigma_2$

We can no longer calculate a pooled estimator for the variance: instead, we need to use the two sample standard deviations in the place of the actual ones.

We can no longer calculate the actual degrees of freedom: we estimate the approximate degrees of freedom as

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

Means with unknown $\sigma_1 \neq \sigma_2$

Null hypothesis: | Test statistic: | Distribution:

$$H_0 : \mu_1 - \mu_2 = \Delta_0.$$

$$T_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

$$T_0 \sim T_v.$$

H_1	Rejection region	P-value
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert T_0 \rvert > t_{\alpha/2,v}$	$2 \cdot (1 - T_v(\lvert T_0 \rvert))$
$\mu_1 - \mu_2 > \Delta_0$	$T_0 > t_{\alpha,v}$	$1 - T_v(T_0)$
$\mu_1 - \mu_2 < \Delta_0$	$T_0 < -t_{\alpha,v}$	$T_v(T_0)$

ILLINOIS

Ratio of variances

For two normally distributed populations with unknown variances σ_1^2 and σ_2^2, we have:

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$.

Test statistic: $F_0 = \frac{s_1^2}{s_2^2}$.

Distribution: $F_0 \sim F_{n_1-1,n_2-1}$.

H_1	Rejection region
$\sigma_1^2 \neq \sigma_2^2$	$F_0 > f_{\alpha/2,n_1-1,n_2-1}$ or
	$F_0 < f_{1-\alpha/2,n_1-1,n_2-1}$
$\sigma_1^2 > \sigma_2^2$	$F_0 > f_{\alpha,n_1-1,n_2-1}$
$\sigma_1^2 < \sigma_2^2$	$F_0 < f_{1-\alpha,n_1-1,n_2-1}$

ILLINOIS

Difference in proportions

How about for the difference in the proportions of two populations $p_1 - p_2$?

- Let $n_i, \hat{p}_i$ be the sample sizes and observed proportions.
- Assume that $n_i p_i$ and $n_i(1 - p_i)$ are ≥ 30.

Define a pooled proportion estimator:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

Why do we need this? Well, we somehow need to quantify the variance of the combined sample...

Difference in proportions

How about for the difference in the proportions of two populations $p_1 - p_2$?

- Let $n_i, \hat{p}_i$ be the sample sizes and observed proportions.
- Assume that $n_i p_i$ and $n_i \left(1 - p_i\right)$ are ≥ 30.

Define a pooled proportion estimator:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

Why do we need this? Well, we somehow need to quantify the variance of the combined sample...

ILLINOIS

Difference in proportions

How about for the difference in the proportions of two populations $p_1 - p_2$?

- Let $n_i, \hat{p}_i$ be the sample sizes and observed proportions.
- Assume that $n_i p_i$ and $n_i (1 - p_i)$ are ≥ 30.

Define a pooled proportion estimator:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

Why do we need this? Well, we somehow need to quantify the variance of the combined sample...

Difference in proportions

Null hypothesis:	Test statistic:	Distribution:
$H_0 : p_1 - p_2 = \Delta_0.$	$Z_0 = \dfrac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$	$Z_0 \sim \mathcal{N}(0, 1).$

H_1	Rejection region	P-value
$p_1 - p_2 \neq \Delta_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$	$2 \cdot (1 - \Phi(\lvert Z_0 \rvert))$
$p_1 - p_2 > \Delta_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$p_1 - p_2 < \Delta_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

ILLINOIS