Statistics and Estimation for Computer Science

İstanbul Teknik Üniversitesi

Mustafa Kamasak, PhD

Version: 2022.3.15

Two Sample Hypothesis Testing

Comparison of Two Populations

- Sometimes, hypothesis about two populations is tested
- For example,
  - Let μ_1 be the mean of population 1
  - Let μ_2 be the mean of population 2
- A sample hypothesis is
  $H_0 : \mu_1 = \mu_2$
  $H_1 : \mu_1 \neq \mu_2$

- Now, parameters of two populations are compared against eachother
- Sometimes two populations are totally different and independent
- Sometimes same population before/after a treatment is considered

Comparison of Two Populations – Procedure

- A sample from each population is collected
- A test statistics T is defined with two samples for the H_0 (called null distr)
- Sampling distribution of T, and significance level α is used to determine critical value(s) for T
- t_{obs} is computed
- If $t_{obs} \in \mathcal{R}$ H_0 is rejected, otherwise it is retained

Two Sample Hypothesis Testing For Population Mean

- Sample-1: $\mathbf{x}_1 = [x_{11}, 12, \cdots, x_{1N_1}] \leftarrow$ population 1, size N_1
- Sample-2: $\mathbf{x}_2 = [x_{21}, x_{22}, \cdots, x_{2N_2}] \leftarrow$ population 2, size N_2
- Both populations are normal distributed:

$$\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ and } \mathbf{x}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

- $x_{11}, x_{12}, \cdots, x_{1N_1}$ are iid
- $x_{21}, x_{22}, \cdots, x_{2N_2}$ are iid
- $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent
- Then

$$\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2})$$

$$T = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N1} + \frac{\sigma_2^2}{N2}}} \sim \mathcal{N}(0, 1)$$

- This sampling distr can be used to form confidence interval for $\mu_1 - \mu_2$

Two Sample Hypothesis Testing For Population Mean

- Hypothesis can be stated in terms of $\mu_1 - \mu_2$

$$T = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim \mathcal{N}(0,1)$$

- Let $\Delta = \mu_1 - \mu_2$
  $\mathrm{H}_0 : \mu_1 - \mu_2 = \Delta$
  $\mathrm{H}_1 : \mu_1 - \mu_2 \neq \Delta$
- Then, test statistic T becomes

$$T = \frac{\overline{x}_1 - \overline{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{N1} + \frac{\sigma_2^2}{N2}}} \sim \mathcal{N}(0,1)$$

if population variances are known.

Two Sample Hypothesis Testing For Population Mean

- From sample 1 ($\mathbf{x_1}$) and sample 2 ($\mathbf{x_2}$), t_{obs} is computed

$$t_{obs} = \frac{\overline{x}_1 - \overline{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim \mathcal{N}(0, 1)$$

- Use significance level α to determine the critical values for T
- For this two-tailed test $z_c = z_{\alpha/2}$
- If $|t_{obs}| \geq z_c \implies$ reject H_0
- If $|t_{obs}| < z_c \implies$ retain H_0
- p-value is

$$\text{p-value} = 2(1 - F_z(t_{obs}))$$

  where F_z is the cdf of standard normal distr

Two Sample Hypothesis Testing For Population Mean

- A common value for $\Delta = 0$
- Hypothesis becomes
  $H_0 : \mu_1 = \mu_2$
  $H_1 : \mu_1 \neq \mu_2$
- Test statistic T becomes

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim \mathcal{N}(0, 1)$$

- Rest is the same

- Population variances are typically unknown
- Sample variances are used instead

Two Sample Hyp Test for Population Mean – Exercise

- Assume a Calculus lecture is offered in two sections, in-class and online, to the students of same department.
- Final exam scores are

	In-class	Online
# of students	20	25
mean score	62	67

- Variance is known to be $\sigma^2 = 25$ for both sections

(a) Test following hypothesis for significance level of $\alpha = 0.01$
$H_0 : \mu_{in-class} = \mu_{online}$
$H_1 : \mu_{in-class} \neq \mu_{online}$

(b) Find p-value

Confidence Interval for Population Mean Difference

- $1 - \alpha$ confidence interval for population difference $\mu_1 - \mu_2$ is
- When population variances are known

$$\left[\, \overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}},\ \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}\, \right]$$

One Sided Confidence Bound on Population Mean Difference

- Consider the case, where an upper or lower limit is required with confidence level $1 - \alpha$
- When population variances are known
- For upper limit

$$\mu_1 - \mu_2 \leq \overline{x}_1 - \overline{x}_2 + z_\alpha \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

- For lower limit

$$\mu_1 - \mu_2 \geq \overline{x}_1 - \overline{x}_2 - z_\alpha \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

Two Sample Hypothesis Testing For Population Mean

- Use sample variances for population variances

$$T = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N1} + \frac{s_2^2}{N2}}}$$

- Let $\Delta = \mu_1 - \mu_2$
  $\mathrm{H}_0 : \mu_1 - \mu_2 = \Delta$
  $\mathrm{H}_1 : \mu_1 - \mu_2 \neq \Delta$
- Then, test statistic T becomes

$$T = \frac{\overline{x}_1 - \overline{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{N1} + \frac{s_2^2}{N2}}}$$

- Test statistics T is not normal distributed anymore

Two Sample Hypothesis Testing For Population Mean

- Let's assume population variances are the same $\sigma^2 = \sigma_1^2 = \sigma_2^2$
- Use all available data to estimate σ^2
- Pooled estimator of σ^2 is

$$S_p^2 = \frac{(N_1)s_1^2 + (N_2)s_2^2}{N_1 + N_2 - 2}$$

- Then

$$T = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim t_{N_1+N_2-2}$$

  has t distribution with $N_1 + N_2 - 2$ dof

Two Sample Hypothesis Test for Population Mean – Exercise

- Assume a Calculus lecture is offered in two sections, in-class and online, to the students of same department.
- Final exam scores are

	In-class	Online
# of students	20	25
mean score	62	67
sample var	15	25

- Variance is unknown but it is expected to be same for both sections as the students come from same department

(a) Test following hypothesis for significance level of $\alpha = 0.01$
$H_0 : \mu_{in-class} = \mu_{online}$
$H_1 : \mu_{in-class} \neq \mu_{online}$

(b) Find p-value

Two Sample Hypothesis Testing For Population Mean

- If population variances are different $\sigma_1^2 \neq \sigma_2^2$
- Use all available data to estimate σ^2
- Then

$$T = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \sim t_\nu$$

  has t distribution with ν dof
- ν is the effective dof

$$\nu = \left\lfloor \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{(s_1^2/N_1)^2}{N_1 - 1} + \frac{(s_2^2/N_2)^2}{N_2 - 1}} \right\rfloor$$

Two Sample Hypothesis Test for Population Mean – Exercise

- Assume a Calculus lecture is offered in two sections, in-class and online, to the students of different departments.
- Final exam scores are

	In-class	Online
# of students	20	25
mean score	62	67
sample var	15	25

- Variance is unknown and it is expected to be different for two sections as the students come from different departments

(a) Test following hypothesis for significance level of $\alpha = 0.01$
$H_0 : \mu_{in-class} \geq \mu_{online}$
$H_1 : \mu_{in-class} < \mu_{online}$

(b) Find p-value

Confidence Interval for Population Mean Difference

- When population variances are not known
- If population variances are same

$$\left[\ \overline{x}_1 - \overline{x}_2 - t_c S_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}},\ \overline{x}_1 - \overline{x}_2 + t_c S_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}\ \right]$$

where $t_c = t_{N_1+N_2-2,\alpha/2}$

- If population variance are different

$$\left[\ \overline{x}_1 - \overline{x}_2 - t_c\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}},\ \overline{x}_1 - \overline{x}_2 + t_c\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}\ \right]$$

where $t_c = t_{\nu,\alpha/2}$

Paired t-test

- Sometimes an effect of a treatment is investigated on the population
- A sample is taken and measurements/observations are performed before and after treatment

$$[(x_{1b}, x_{1a}), (x_{2b}, x_{2a}), \cdots, (x_{Nb}, x_{Na})]$$

- Hence data from each case is taken twice (before/after treatment)
  x_{ib}: Data taken from case i before treatment
  x_{ia}: Data taken from case i after treatment

- For example, effects of a cholesterol pill on population
- A random sample is taken
- Blood cholesterol is measured before and after pill

Paired t-test

- Define sample difference d as

$$d = [x_{1b} - x_{1a}, x_{2b} - x_{2a}, \cdots, x_{Nb} - x_{Na}]$$

- Hypothesis is
  $\mathrm{H}_0 : d = \Delta$
  $\mathrm{H}_1 : d \neq \Delta$
- Test statistics is

$$T = \frac{\overline{d} - \Delta}{s_d/\sqrt{N}} \sim t_{N-1}$$

- Using significance level, find critical value for T: $t_c = t_{N-1,\alpha/2}$
- From data compute t_{obs}
- If $|t_{obs}| \geq t_c$ reject H_0
- If $|t_{obs}| < t_c$ retain H_0

Paired vs Unpaired t-test

- What happens if unpaired t-test is used for paired data?
- Unpaired t-test

$$T_1 = \frac{\overline{x}_1 - \overline{x}_2 - \Delta}{S_p\sqrt{\frac{1}{N} + \frac{1}{N}}} = \frac{\overline{x}_1 - \overline{x}_2 - \Delta}{\sqrt{\frac{s_1^2+s_2^2}{N-1}}} \sim t_{2N-2}$$

- Paired t-test

$$T_2 = \frac{\overline{d} - \Delta}{s_d/\sqrt{N}} = \frac{\overline{d} - \Delta}{\sqrt{\frac{s_1^2-2cov(x_1,x_2)+s_2^2}{N-1}}} \sim t_{N-1}$$

- Numerators of the test statistics are the same as $\overline{d} = \overline{x}_1 - \overline{x}_2$
- Denominators are same if $cov(x_1, x_2)$ is zero $\implies$ unlikely

Paired vs Unpaired t-test

- For positive covariance, denominator of paired t-test will be lower $\implies T_1 < T_2$
- Unpaired t-test statistics have higher dof compared to paired t-test statistics

- For high covariance, unpaired t-test will overestimate p-value
- For low covariance, unpaired t-test will have lower power

- If covariance between pairs is low $\implies$ use unpaired t-test
- If covariance between pairs is high $\implies$ use paired t-test

Two Sample Testing for Population Variance

- Consider two population and the following simple hypothesis that compares their variances
  $H_0 : \sigma_1^2 = \sigma_2^2$
  $H_1 : \sigma_1^2 \neq \sigma_2^2$
- Equivalently
  $H_0 : \sigma_1^2/\sigma_2^2 = 1$
  $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$

Two Sample Testing for Population Variance

- Let sample 1 from population 1 with size N_1 and sample 2 from population 2 with size N_2
- Recall
$$\frac{(N_1 - 1)s_1^2}{\sigma_1^2} \sim \chi^2_{N_1-1} \qquad \frac{(N_2 - 1)s_2^2}{\sigma_2^2} \sim \chi^2_{N_2-1}$$
- Division of two independent normalized (by dof) χ^2 distributed random variables has F distribution
- Define following test statistics

$$F = \frac{\frac{(N_1-1)s_1^2}{\sigma_1^2}\frac{1}{(N_1-1)}}{\frac{(N_2-1)s_2^2}{\sigma_2^2}\frac{1}{(N_2-1)}} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

F Distribution

- Sampling distribution of F is F distribution with $N_1 - 1$ dof for the numerator and $N_2 - 1$ dof for the denominator

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{N_1-1, N_2-1}$$

- The pdf of F distr with u dof for numerator and v dof for denominator ($F_{u,v}$) is

$$F_{u,v}(x) = \frac{\Gamma(\frac{u+v}{2})(\frac{u}{v})^{u/2} x^{u/2-1}}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})[\frac{ux}{v}+1]^{(u+v)/2}}$$

for $0 < x < \infty$

- Mean and variance of $F_{u,v}$ is

$$\mu = v/(v-2) \text{ for } v > 2$$

$$\sigma^2 = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)} \text{ for } v > 4$$

- Need numeric integration and look-up table for cdf

F Distribution

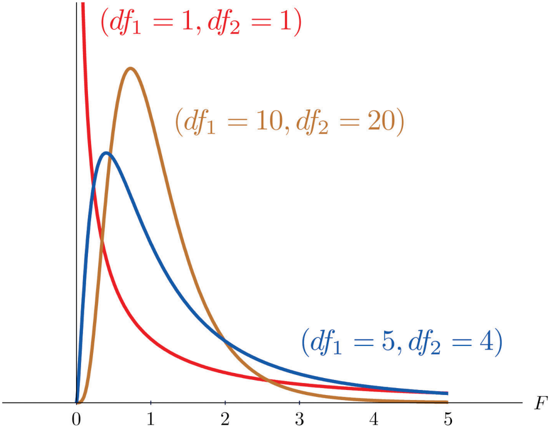
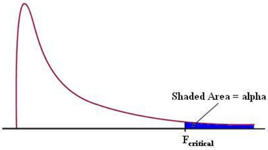


Table for F Distribution

- Table for F distr has 3 dimensions
  - Numerator dof
  - Denominator dof
  - Significance level α
- Not possible to show all 3 dimensions in a single table
- For each significance level F distr is given by numerator dof ν_1 (on columns) and denominator dof ν_2 (on rows)



F Distribution: Critical Values of F (5% significance level)

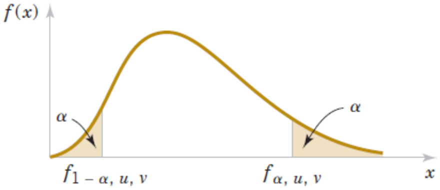
ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	14
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86

Two Sample Testing for Population Variance

- Test
  $H_0 : \sigma_1^2/\sigma_2^2 = 1$
  $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$
- Test statistics under H_0

$$F = \frac{s_1^2}{s_2^2} \sim F_{N_1-1, N_2-1}$$

- If F statistics is far away from 1, H_0 can be rejected
- Using significance level α and F tables find two critical values



Two Sample Testing for Population Variance

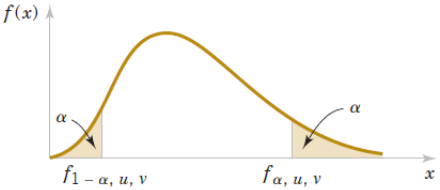
- Lower critical value can be estimated from upper critical value

$$F_{(u,v),1-\alpha/2} = \frac{1}{F_{(v,u),\alpha/2}}$$

note the reversed dof for numerator and denominator.

- For example

$$F_{(5,10),0.975} = \frac{1}{F_{(10,5),0.025}}$$



Two Sample Testing for Population Variance

- If $F_{(u,v),1-\alpha/2} < F_{obs} < F_{(u,v),\alpha/2}$ then retain H_0 , otherwise reject H_0

- To test composite hypothesis
  $H_0 : \sigma_1^2 \geq \sigma_2^2$
  $H_1 : \sigma_1^2 < \sigma_2^2$
  use the lower tail only
- If $F_{(u,v),1-\alpha} < F_{obs}$ then retain H_0 , otherwise reject H_0

- To test composite hypothesis
  $H_0 : \sigma_1^2 \leq \sigma_2^2$
  $H_1 : \sigma_1^2 > \sigma_2^2$ use the upper tail only
- If $F_{obs} < F_{(u,v),\alpha}$ then retain H_0 , otherwise reject H_0

Two Sample Testing for Population Proportions

- Consider a hypothesis that compares proportions in two populations
  $\mathrm{H}_0 : p_1 = p_2$
  $\mathrm{H}_1 : p_1 \neq p_2$
- Let $\hat{p}_1 = x_1/N_1$ and $\hat{p}_2 = x_2/N_2$ are the sample proportions
- Test statistics

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}} \sim \mathcal{N}(0,1)$$

- Test statistics has normal distribution
- To test H_0 $p = p_1 = p_2$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \sim \mathcal{N}(0,1)$$

Two Sample Testing for Population Proportions

- To test $\mathrm{H}_0 : p = p_1 = p_2$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \sim \mathcal{N}(0,1)$$

- Use pooled estimation for $\hat{p}$

$$\hat{p} = \frac{x_1 + x_2}{N_1 + N_2}$$

- For significance level of α
  - If $|z_{obs}| < z_{\alpha/2}$, retain H_0
  - If $|z_{obs}| \geq z_{\alpha/2}$, reject H_0
- p-value is

$$\text{p-value} = 2(1 - F_z(z_{obs}))$$

where $F_z()$ is the cdf of standard normal distr

Two Sample Testing for Population Proportions

- Use same test statistics for composite hypothesis

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \sim \mathcal{N}(0, 1)$$

- For significance level of α

- To test
  $\mathrm{H}_0 : p_1 \geq p_2$
  $\mathrm{H}_1 : p_1 < p_2$
  - If $z_{obs} > -z_\alpha$, retain H_0
  - If $z_{obs} \leq -z_\alpha$, reject H_0

- p-value is

$$\text{p-value} = F_z(z_{obs})$$

  where $F_z()$ is the cdf of standard normal distr

Two Sample Testing for Population Proportions

- Use same test statistics for composite hypothesis

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \sim \mathcal{N}(0, 1)$$

- For significance level of α

- To test
  $\mathrm{H}_0 :\ p_1 \leq p_2$
  $\mathrm{H}_1 :\ p_1 > p_2$
  - If $z_{obs} < z_\alpha$, retain H_0
  - If $z_{obs} \geq z_\alpha$, reject H_0

- p-value is

$$\text{p-value} = 1 - F_z(z_{obs})$$

where $F_z()$ is the cdf of standard normal distr

Two Sample Testing for Population Proportions – Exercise

- At the end of the semester, number of students passing in-class/online sections is

	In-class	Online
N	20	22
# passed	16	14

- Let p_1 and p_2 be the proportion of students passing the section for in-class and online sections respectively

(a) Test
$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$
with $\alpha = 0.05$

(b) Find the significance level (p value)