# Confidence intervals for two populations

Chrysafis Vogiatzis

Lecture 23

## Learning objectives

After lectures 20-23, we will be able to:

- Build confidence intervals for:
  - unknown means:
  - unknown variances;
  - unknown proportions.
- Build confidence intervals for:
  - the difference between two unknown means;
  - the ratio between two unknown variances;
  - the difference between two unknown proportions.
- Understand the effect of Type I error, or probability  $\alpha$ .
- Calculate errors and interval margins.
- Select appropriate sample sizes to keep errors below a limit.

Motivation: Do masks work?

There has been an ongoing discussion about whether mandating universal mask wearing curbs COVID-19. All politics aside, there was a very interesting study coming from Kansas: apparently, counties that mandated masks saw smaller increases (or decreases) in the onset of new COVID-19 cases, than counties that did not mandate masks. Could we prove that (within a given specified level of confidence)?

Motivation: Does IE 300 have more variable grades than IE 310?

When deciding a technical elective, we also look for how *variable* the grading is; not only what the average is! A class that has an A- average is not necessarily "easier" or "more straightforward" than a class that has a B average. Instead, we also want to see what the variances are. The question we would like to answer then: how much more variable is class *A* compared to class *B*? Or, to put it in confidence interval terms, what is the ratio of variances between two populations with 95% confidence?

# Two population confidence intervals

In this set of notes, we turn our focus to two different populations and how they compare. More specifically, in this lecture we see confidence intervals on:

- 1. the difference of two means,  $\mu_1 \mu_2$ .
- 2. the difference between two proportions,  $p_1 p_2$ .
- 3. the ratio of two variances,  $\frac{\sigma_1^2}{\sigma^2}$ .

Why would we look at two populations? Well, in many practical applications, we are given more than one populations to compare. For example, we may want to compare the performance of a drug in two groups of patients. At a similar vein, we may want to check the differences in driving on ice between more (> 10 years) and less experienced (0 - 10 years) drivers. Finally, at a problem that we can relate to in 2020, we may want to see how people living in two different states vote?

One thing is for sure: in all these cases, it is imperative to create confidence intervals for more than just one population.

## Difference in means

Consider two normally distributed populations with unknown means  $\mu_1, \mu_2$ . We are interested in quantifying the difference in their means:

$$\mu_1 - \mu_2$$
.

How about we try again what we did before? That is:

- Take a sample of size  $n_1$  from the first population and calculate the sample average  $\overline{X}_1$ .
- Take a sample of size  $n_2$  from the second population and calculate the sample average  $X_2$ .
- Estimate  $\mu_1 \mu_2$  by  $\overline{X}_1 \overline{X}_2$ .

This will be a good point estimate, for sure <sup>1</sup>. But what about the confidence interval?

1 Why? Can you prove that?

Difference in means of two populations with known variances

Before we get to the confidence intervals, let us define a new "kind" of standard deviation.

**Definition 1 (Pooled standard deviation)** Given two samples of sizes  $n_1, n_2$ , with known respective standard deviations  $\sigma_1, \sigma^2$ , we define their pooled standard deviation as:

$$\sigma_P = \sqrt{\frac{(n_1 - 1)\,\sigma_1^2 + (n_2 - 1)\,\sigma_2^2}{n_1 + n_2 - 2}}.$$

With this definition, and keeping the same logic as in Lecture 20, we get our first confidence interval:

$$\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sigma_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sigma_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Note how similar the setup is as in the case of single population means with known standard deviation! The only differences are in the point estimate used  $(\overline{X} \text{ vs. } \overline{X}_1 - \overline{X}_2)$ , in the standard deviation used ( $\sigma$  vs. the pooled standard deviation  $\sigma_P$ ), and in the population size (we multiplied by  $1/\sqrt{n}$  vs. multiplying by  $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ ).

#### **BMI**

The body mass index of 1545 people was found to be on average 28.8. The same index for another population of 1781 people was calculated as (on average) 27.6. The body mass index has variance of 9 (in both populations). Build a 99%confidence interval for the difference in the BMI between the two populations.

First, we will need  $z_{0.005} = 2.576$ . Then, we have:

- $\overline{X}_1 \overline{X}_2 = 1.2$ .
- $\sigma_P = \sqrt{\frac{1544 \cdot 9 + 1780 \cdot 9}{3324}} = 3$ . (unsurprising as both populations had the same  $\sigma$  to begin with.)
- $L = 1.2 2.576 \cdot 3 \cdot \sqrt{\frac{1}{1545} + \frac{1}{1781}} = 0.931.$
- $U = 1.2 + 2.576 \cdot 3 \cdot \sqrt{\frac{1}{1545} + \frac{1}{1781}} = 1.469.$

Hence, the difference in the mean BMI between these two populations is

$$[0.931, 1.469]$$
.

Difference in means of two populations with unknown variances

Now, assume we do not know the population variances! If we do not know them, then we can take a page from the single population confidence intervals book: we can estimate these unknown variances using the sample variances,  $s_1^2$  and  $s_2^2$ . As soon as we estimate

the variances though (rather than having the true ones), we get two changes:

Change 1: we no longer have a z-value (from a normal distribution), but we have a t-value (from a Student's T distribution) with  $n_1 + n_2 - 2$  degrees of freedom.

Change 2: as we do not know  $\sigma_1, \sigma_2$ , we **estimate** the pooled standard deviation as

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

$$\overline{X_1 - \overline{X}_2 - t_{\alpha/2, n_1 + n_2 - 2} s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + t_{\alpha/2, n_1 + n_2 - 2} s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

#### BMI: take 2

Assume that we run a smaller experiment on the body mass index of two populations. We now have collected a sample of 4 from some population with  $\overline{X}_1 = 26.2$  and  $s_1 = 2$ , and a sample of 6 from a different population with  $\overline{X}_2 = 28$  and  $s_2 = 3.6$ . Build a 99% confidence interval for  $X_1 - X_2$ .

Now, instead of a *z*-value, we will need  $t_{0.005,9} = 4.297$ . We have:

- $\overline{X}_1 \overline{X}_2 = -1.8$ .
- $s_P = \sqrt{\frac{3.4 + 5.12.96}{8}} = \sqrt{9.6} = 3.1.$
- $L = -1.8 4.297 \cdot 3.1 \cdot \sqrt{\frac{1}{4} + \frac{1}{6}} = -10.4.$
- $U = -1.8 + 4.297 \cdot 3.1 \cdot \sqrt{\frac{1}{4} + \frac{1}{6}} = 6.8.$

Hence, the difference in the mean BMI between these two populations has now become

$$[-10.4, 6.8]$$
.

A bit of critical analysis in these two results. Take a look at the first confidence interval from the larger experiment with the known standard deviations. We got (with 99% confidence) that the first population has bigger BMI values by at least 0.931 points and up to 1.469 points. Hence, we could say that the first population has more BMI with 99% confidence! On the other hand, looking at the second smaller experiment with unknown standard deviations, we got a

much bigger and much less clear confidence interval: specifically, we got that the BMI difference is between -10.4 and 6.8. This translates to possibly the first population having smaller BMI by a whole 10.4 points or bigger BMI by 6.8 points! Hence, we could not claim that the first population has more nor less BMI with 99% confidence!

This type of critical thinking will be invaluable when we move to hypothesis testing.

Confidence intervals for the ratio of the variances of two normally distributed populations

Before we calculate confidence intervals on the variances of two normally distributed populations, we define the ratio of two sample variances as:

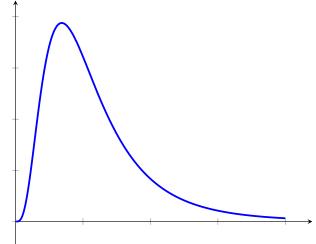
$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}.$$

Recall that variances have degrees of freedom: hence, assuming we have a sample of  $n_1$  observations from the first and  $n_2$  observations from the second population, then we say that *F* is distributed as an *F* distribution with  $n_1 - 1$  degrees of freedom in the numerator and  $n_2 - 1$  degrees of freedom in the denominator, or, mathematically, we write that:

$$F \sim F_{n_1-1,n_2-1}$$
.

We show this visually in Figure ??. Note that much like the  $\chi^2$  distribution, the *F* distribution is also not symmetric.

Figure 1: The *F* distribution visually.



Much like with the other distributions we have seen, we also have a table for the F distribution (see the last two pages here)! It is a little different to read: here the columns and rows represent the degrees of freedom of the numerator  $(v_1)$  and the denominator  $(v_2)$ . Then, we find the value of  $\alpha$  we are interested in to get the value.

## Finding *F* distribution values

- $n_1 = 9$ ,  $n_2 = 5$ ,  $\alpha = 10\%$ : we then look for  $\nu_1 = n_1 1 = 8$ ,  $\nu_2 = n_2 - 1 = 4$ ,  $\alpha = 0.1$  and find  $f_{8,4,0,1} = 3.95$ .
- $n_1 = 5$ ,  $n_2 = 16$ ,  $\alpha = 5\%$ : we then look for  $\nu_1 = n_1 1 = 4$ ,  $\nu_2 = n_2 - 1 = 15$ ,  $\alpha = 0.05$  and find  $f_{4,15,0.05} = 3.06$ .

Wait! What do I do if I am looking at other values, such as the ones used for  $\alpha = 90\%$ ? Those are clearly not available in the table, right? Well, in that case we have:

$$f_{u,v,\alpha} = \frac{1}{f_{v,u,1-\alpha}}$$

In English: the f value for u degrees of freedom in the numerator, v degrees of freedom in the denominator and  $\alpha$  is equal to 1 over the f value for v degrees of freedom in the numerator, u degrees of freedom in the denominator and  $1 - \alpha$ . Nifty, no? Let's put it to the use.

#### Finding *F* distribution values

In general, for two-sided confidence intervals we need values for  $\alpha/2$  and 1 -  $\alpha/2$ . So, assume we are building 80% confidence intervals and 95% confidence intervals for:

•  $n_1 = 9$ ,  $n_2 = 5$ ,  $\alpha = 20\%$ : we then look for  $\nu_1 = n_1 - 1 = 8$ ,  $\nu_2 = n_2 - 1 = 4$ ,  $\alpha/2 = 0.10$  and find  $f_{8.4,0.1} = 3.95$ . We also need the same value but for  $1 - \alpha/2$ ,  $f_{8,4,0,9}$ . We know it can be found as:

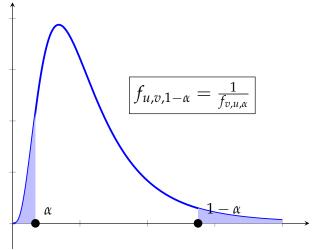
$$f_{8,4,0.9} = \frac{1}{f_{4,8,0.1}} = \frac{1}{2.81} = 0.356.$$

•  $n_1 = 5$ ,  $n_2 = 15$ ,  $\alpha = 10\%$ : we then look for  $\nu_1 = n_1 - 1 = 4$ ,  $\nu_2 = n_2 - 1 = 15$ ,  $p = 1 - \alpha = 95\%$  and find  $f_{4,15,0.05} = 3.06$ . We also need  $f_{4,15,0.95}$ . We know it can be found as:

$$f_{4,15,0.95} = \frac{1}{f_{15,4,0.05}} = \frac{1}{5.86} = 0.171.$$

Please check this awesome online tool to help you do these calculations: https://stattrek.com/online-calculator/f-distribution. aspx. Additionally, see Figure ?? for an example of how  $\alpha$  and  $1 - \alpha$ are located in the *F* distribution.

Figure 2: The *F* distribution and the marks for  $1 - \alpha$  and  $\alpha$ .



Following a similar logic to the single population variance case, we finally have:

$$f_{n_2-1,n_1-1,1-\alpha/2} \frac{s_1^2}{s_2^2} \le \frac{\sigma_1^2}{\sigma_2^2} \le f_{n_2-1,n_1-1,\alpha/2} \frac{s_1^2}{s_2^2}$$

Before we put this to the test, let us remember what a critical value is! In essence, for a critical value  $f_{u,v,\alpha}$  we need:

$$P(F \ge f_{u,v,\alpha}) = \alpha.$$

This implies the following: for a given value for  $\alpha$ , we look at p = $1 - \alpha$  in the *F*-table!

Let's put this to use right away!

#### Semiconductor wafers and their oxide layers

The variability in the thickness of oxide layers in semiconductor wafers is a critical characteristic, where low variability is desirable. A company is investigating two different ways to mix gases so as to reduce the variability of the oxide thickness. We produce 16 wafers with each gas mixture and our results indicate that the standard deviation is  $s_1 = 1.96\text{Å}$  and  $s_2 = 2.13\text{Å}$  for the two mixtures. What is the 95% confidence intervals for the ratio between the two variances?

#### Semiconductor wafers and their oxide layers

We have been given a series of information:

- size of population 1:  $n_1 = 16$ ;
- sample standard deviation for sample from population 1:  $s_1 = 1.96$ ;
- size of population 2:  $n_2 = 16$ ;
- sample standard deviation for sample from population 2:  $s_2 = 2.13$ .

Since we are looking for a 95% confidence interval we need two f values:

- $f_{n_2-1,n_1-1,\alpha/2} = f_{15,15,0.025} = 2.86.$
- $f_{n_2-1,n_1-1,1-\alpha/2} = \frac{1}{f_{n_1-1,n_2-1,\alpha/2}} = \frac{1}{2.86} = 0.35.$

Finally, the confidence interval for  $\sigma_1^2/\sigma_2^2$  is found as:

$$\left[ f_{n_2-1,n_1-1,1-\alpha/2} \frac{s_1^2}{s_2^2}, f_{n_2-1,n_1-1,\alpha/2} \frac{s_1^2}{s_2^2} \right] =$$

$$= [0.35 \cdot 0.847, 2.86 \cdot 0.847] = [0.296, 2.422].$$

Confidence intervals for the difference of the proportions of two populations

As a reminder, if we had a single population with proportion p, then our confidence intervals, are given by:

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}.$$

Recall that  $\hat{p}$  is the observed (estimated) proportion based on the sample collected. Similarly,  $\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$  is the estimated square error (standard deviation).

Now, assume we have two populations: one with true proportion  $p_1$  and the other with true proportion  $p_2$ . If we do not know what  $p_1$ and  $p_2$  are, how can we estimate  $p_1 - p_2$ ?

Well, let us follow a similar process:

1. Collect a sample from the first population of size  $n_1$  and calculate the observed proportion  $\hat{p}_1$ .

- 2. Collect a sample from the second population of size  $n_2$  and calculate the observed proportion  $\hat{p}_2$ .
- 3. Estimate  $p_1 p_2$  as  $\hat{p}_1 \hat{p}_2$ .

Great! That will do! But, what about the confidence interval around it? Following the theory from the single population proportions, we get...

If  $n_1p_1, n_2p_2, n_1(1-p_1), n_2(1-p_2)$  are all greater than or equal to 5, then  $\hat{p}_1 - \hat{p}_2$  is **normally distributed** with mean  $p_1 - p_2$  and variance  $\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$ .

Using that, we finally obtain our confidence interval as:

$$\begin{aligned} \hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} &\leq p_1 - p_2 \leq \\ &\leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}. \end{aligned}$$

#### Environmentally conscious

Residents of major metropolitan areas in the US were asked whether they agree with the following statement:

"I consider my self environmentally conscious."

The answers they could give were either a "Yes" or a "No". Specifically, we focus on two cities: Portland and Philadelphia.

- Out of n = 91 respondents in Portland, 61 answered Yes.
- Out of n = 100 respondents in Philadelphia, 45 said Yes.

Build a 95% confidence interval for the true proportion difference  $p_1 - p_2$ , where  $p_1$  is the proportion of people agreeing with the statement in Portland and  $p_2$  the proportion of the same people in Philadelphia.

We have:

• 
$$n_1 = 91$$
,  $\hat{p}_1 = \frac{61}{91} = 0.67$ .

• 
$$n_2 = 100, \hat{p}_2 = \frac{45}{100} = 0.45.$$

We also have  $z_{\alpha/2} = z_{0.025} = 1.96$ . Plugging everything to-

$$p_1 - p_2 \in [0.22 - 1.96 \cdot 0.07, 0.22 + 1.96 \cdot 0.07] = [0.08, 0.36].$$

Let us dwell on this last result for one minute. What do we learn from this confidence interval? Well, we learn that residents of Portland are (with 95% confidence) more environmentally conscious than residents of Philadelphia! We could never make the same assertion simply by looking at the individual observed proportions: that is, we cannot make the claim simply through the argument that  $\hat{p}_1 > \hat{p}_2$ . But now? We most definitely can make it! Always with the sidenote that "with 95% confidence".

More on that, in the coming lectures.

## CRITICAL VALUES OF THE F DISTRIBUTION

$\nu_2 \backslash \nu_1$		2	3	4	5	6	7	8	10	12	15	20	30	50	$\infty$
	α		(	0		-0 -			<i>(</i>	<i>(</i>	<i>(</i>	<i>(</i>	(- (	(	(
1	0.100												62.6		
	0.050												250	252	254
	0.025									977					
2	0.100												9.46		
	0.050												19.5		
•	0.025												39.5		
3	0.100												5.17		
	0.050												8.62		
	0.025												14.1		
4	0.100												3.82		
	0.050												5.75 8.46		
5	0.025 0.100												3.17		
3	0.050												4.50		
	0.025												6.23		
6	0.100												2.80		
Ü	0.050												3.81		
	0.025												5.07		
7	0.100												2.56		
,	0.050												3.38		
	0.025												4.36		
8	0.100												2.38		
	0.050												3.08		
	0.025	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.29	4.20	4.10	4.00	3.89	3.81	3.67
9	0.100	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.42	2.38	2.34	2.30	2.25	2.22	2.16
	0.050	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	3.01	2.94	2.86	2.80	2.71
	0.025	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.96	3.87	3.77	3.67	3.56	3.47	3.33
10	0.100	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.32	2.28	2.24	2.20	2.16	2.12	2.06
	0.050												2.70		
	0.025	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.72	3.62	3.52	3.42	3.31	3.22	3.08
11	0.100												2.08		
	0.050												2.57		
	0.025												3.12		
12	0.100												2.01		
	0.050												2.47		
	0.025												2.96		
13	0.100												1.96		
	0.050												2.38		
	0.025	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.25	3.15	3.05	2.95	2.84	2.74	2.60

## CRITICAL VALUES OF THE F DISTRIBUTION

$\nu_2 \backslash \nu_1$		2	3	4	5	6	7	8	10	12	15	20	30	50	$\infty$
	α														
14	0.100	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.10	2.05	2.01	1.96	1.91	1.87	1.80
	0.050	3.743	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.46	2.39	2.31	2.24	2.13
	0.025	4.86	<b>1.2</b> 4	3.89	3.66	3.50	3.38	3.29	3.15	3.05	2.95	2.84	2.73	2.64	2.49
15	0.100	2.702	2.49	2.36	2.27	2.21	2.16	2.12	2.06	2.02	1.97	1.92	1.87	1.83	1.76
	0.050	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.40	2.33	2.25	2.18	2.07
	0.025	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.06	2.96	2.86	2.76	2.64	2.55	2.40
16	0.100	2.672	2.46	2.33	2.24	2.18	2.13	2.09	2.03	1.99	1.94	1.89	1.84	1.79	1.72
	0.050												2.19		
	0.025	4.69	4.08	3.73	3.50	3.34	3.22	3.12	2.99	2.89	2.79	2.68	2.57	2.47	2.32
17	0.100												1.81		
	0.050												2.15		
	0.025	4.62													
18	0.100												1.78		
	0.050	3.553													
	0.025	4.56													
19	0.100												1.76		
	0.050	3.523													
	0.025	4.51													
20	0.100												1.74		
	0.050	3.493													
	0.025	4.46													
25	0.100												1.66		
	0.050												1.92		
•	0.025	4.293													
30	0.100												1.61		
	0.050	3.322 4.183													
60	0.025 0.100												1.48		
00	0.050	3.152													
	0.025												1.82		
80	0.100												1.44		
	0.050	3.112													
	0.025	3.86													
100	0.100	2.362													
	0.050	3.092													
	0.025	3.83													
$\infty$	0.100	2.302													
	0.050	3.002													
	0.025	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.05	1.94	1.83	1.71	1.57	1.43	1.00