Lecture 8
Difference in Population Mean

by Hiro Kasahara

Vancouver School of Economics
University of British Columbia

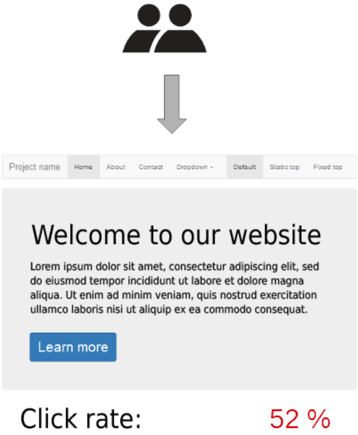
Examples

We are often interested in knowing if the mean of two populations is different or not.

- A/B testing
- A Study of Lung Cancer
- Homework and Final Grades

A/B testing

A/B testing



By Maxime Lorant - `https://commons.wikimedia.org/wiki/File:A-B_testing_simple_example.png`

A/B testing

Randomly assign 200 visitors into two versions of web designs.

	Click	No Click	Total visits
Design A	52	48	100
Design B	72	28	100

- 52 out of 100 visitors clicked for design A: $\hat{p}_x = 0.52$.
- 72 out of 100 visitors clicked for design B: $\hat{p}_y = 0.72$.

Clicker Question 8-1

Which of the following is true in population?

A). With certainty, design B has higher click rates than design A.

B). It is likely that design B has higher click rates than design A.

C). There is not enough information to tell how likely design A has higher click rates than B.

A/B testing

Data from Design A: $\{X_1, X_2, ..., X_{100}\}$, where $X_i \in \{0,1\}$

Data from Design B: $\{Y_1, Y_2, ..., Y_{100}\}$, where $Y_j \in \{0,1\}$

$$\textit{Population}: \quad \Pr(X_i = 1) = p_x \quad \text{and} \quad \Pr(Y_j = 1) = p_y.$$

$$\textit{Sample}: \quad \hat{p}_x = \frac{1}{100}\sum_{i=1}^{n} X_i = 0.52$$

$$\hat{p}_y = \frac{1}{100}\sum_{j=1}^{n} Y_j = 0.72$$

How to construct 95 percent Confidence Interval for $p_y - p_x$?

A point estimator for $p_y - p_x$

- A point estimator for $p_y - p_x$ is given by

$$\hat{p}_y - \hat{p}_x.$$

- $\hat{p}_y - \hat{p}_x$ is an unbiased estimator of $p_y - p_x$ because

$$E(\hat{p}_y - \hat{p}_x) = p_y - p_x.$$

- $\hat{p}_y - \hat{p}_x$ is a consistent estimator of $p_y - p_x$ because

$$\hat{p}_y - \hat{p}_x \xrightarrow{p} p_y - p_x \quad \text{as } n \to \infty.$$

95 percent Confidence Interval for $p_y - p_x$

Because

$$E(\hat{p}_y - \hat{p}_x) = p_y - p_x$$

$$Var(\hat{p}_y - \hat{p}_x) = Var(\hat{p}_y) + Var(\hat{p}_x) - 2\underbrace{Cov(\hat{p}_y, \hat{p}_x)}_{=0}$$

$$= \frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y},$$

we have

$$\frac{(\hat{p}_y - \hat{p}_x) - (p_y - p_x)}{\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}} \xrightarrow{d} N(0,1)$$

95 percent Confidence Interval for μ

Because $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$ as $n \to \infty$:

$$\Pr\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$\Leftrightarrow \Pr\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Leftrightarrow \Pr\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

95 percent Confidence Interval for p

Because $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0,1)$ as $n \to \infty$:

$$\Pr\left(-1.96 < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < 1.96\right) = 0.95$$

$$\Leftrightarrow \Pr\left(-1.96\sqrt{\frac{p(1-p)}{n}} < \hat{p}-p < 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

$$\Leftrightarrow \Pr\left(\hat{p}-1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p}+1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

95 percent Confidence Interval for $p_y - p_x$

Because $\frac{(\hat{p}_y - \hat{p}_x) - (p_y - p_x)}{\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}} \xrightarrow{d} N(0,1)$ as $n \to \infty$:

$$\Pr\left(-1.96 < \frac{(\hat{p}_y - \hat{p}_x) - (p_y - p_x)}{\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}} < 1.96\right) = 0.95$$

$\Leftrightarrow$

$$\Pr\left((\hat{p}_y - \hat{p}_x) - 1.96\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}} < p_y - p_x < (\hat{p}_y - \hat{p}_x) + 1.96\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}\right) = 0.95$$

95 percent Confidence Interval for $p_y - p_x$

Because $\hat{p}_x$ and $\hat{p}_y$ converge in probability to p_y and p_x, we replace $\hat{p}_x$ and $\hat{p}_y$ with p_y and p_x.

$$\Pr\left((\hat{p}_y - \hat{p}_x) - 1.96\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} < p_y - p_x < (\hat{p}_y - \hat{p}_x) + 1.96\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \right) \approx 0.95$$

when n is large.

A/B testing

Randomly assign 200 visitors into two versions of web designs.

	Click	No Click	Total visits
Design A	52	48	100
Design B	72	28	100

- 52 out of 100 visitors clicked for design A: $\hat{p}_x = 0.52$.
- 72 out of 100 visitors clicked for design B: $\hat{p}_y = 0.72$.

95 percent Confidence Interval for $p_y - p_x$

In this example, $\hat{p}_y = 0.72$, $\hat{p}_x = 0.50$, and $n_y = n_x = 100$.

95 percent confidence interval is given by

$$(0.72 - 0.5) \pm 1.96\sqrt{\frac{0.72(1-0.72)}{100} + \frac{0.5(1-0.5)}{100}}$$

or

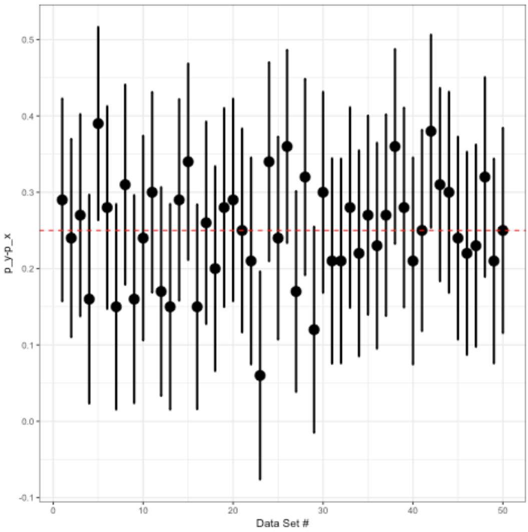
$$0.22 \pm 0.1317 = [0.088, 0.352]$$

$\Rightarrow$ Evidence that Design B's click rate is higher than Design A's in population.

Simulating 95 % Confidence Interval

- Suppose that $p_y = 0.7$ and $p_x = 0.45$ in population.
- $p_y - p_x = 0.25$ in population.
- Generate 50 data sets, where each data set contains 100 $\times 2$ observations.
- For each of 50 data sets, compute the 95 percent confidence interval for $p_y - p_x$.
- The 95 percent confidence interval may or may not contain the true value of $p_y - p_x = 0.25$ but it will contain 0.25 approximately 95 percent of times.

Simulating 95 % Confidence Interval



Study of Lung Cancer

A Study of Lung Cancer

- Doll and Hill (1952) interviewed 1357 men with lung cancer in hospitals.
- Doll and Hill also interviewed another set of 1357 men without lung cancer but with other diseases including other types of cancer ("control group").
- In the interview, each individual was asked about smoking frequency per day.

A Study of Lung Cancer

Disease Group	No. of Non-Smokers	No. of Smokers
1357 lung-cancer patients	7 (0.5%)	1350 (99.5%)
1357 patients with other diseases	61 (4.5%)	1296 (95.5%)

- p_y = population fraction of smokers among lung-cancer patients.
- p_x = population fraction of smokers among patients with other diseases

95 percent confidence interval for $p_y - p_x$?

95 percent Confidence Interval for $p_y - p_x$

$$\Pr\left(-1.96 < \frac{(\hat{p}_y - \hat{p}_x) - (p_y - p_x)}{\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}} < 1.96\right) = 0.95$$

$\Leftrightarrow$

$$\Pr\left((\hat{p}_y - \hat{p}_x) - 1.96\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}} < (p_y - p_x) < (\hat{p}_y - \hat{p}_x) + 1.96\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}\right) = 0.95$$

95 percent Confidence Interval for $p_y - p_x$

Because $\hat{p}_x$ and $\hat{p}_y$ converge in probability to p_y and p_x, we replace $\hat{p}_x$ and $\hat{p}_y$ with p_y and p_x.

$$\Pr\left((\hat{p}_y - \hat{p}_x) - 1.96\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \right.$$

$$\left. < p_y - p_x < (\hat{p}_y - \hat{p}_x) + 1.96\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \right) \approx 0.95$$

when n is large.

95 percent Confidence Interval for $p_y - p_x$

In this example, $\hat{p}_y = 0.995$, $\hat{p}_x = 0.955$, and $n_y = n_x = 1357$.

95 percent confidence interval is given by

$$(0.995 - 0.955) \pm 1.96\sqrt{\frac{0.995(1-0.995)}{1357} + \frac{0.955(1-0.955)}{1357}}$$

or

$$0.04 \pm 0.012 = [0.028, 0.052]$$

$\Rightarrow$ Evidence that lunger cancer patients are more likely to be smokers than patients with other diseases in population.

Worksheet Question

Table: Two-way table of results of tests on 10,000 patients with Tumors

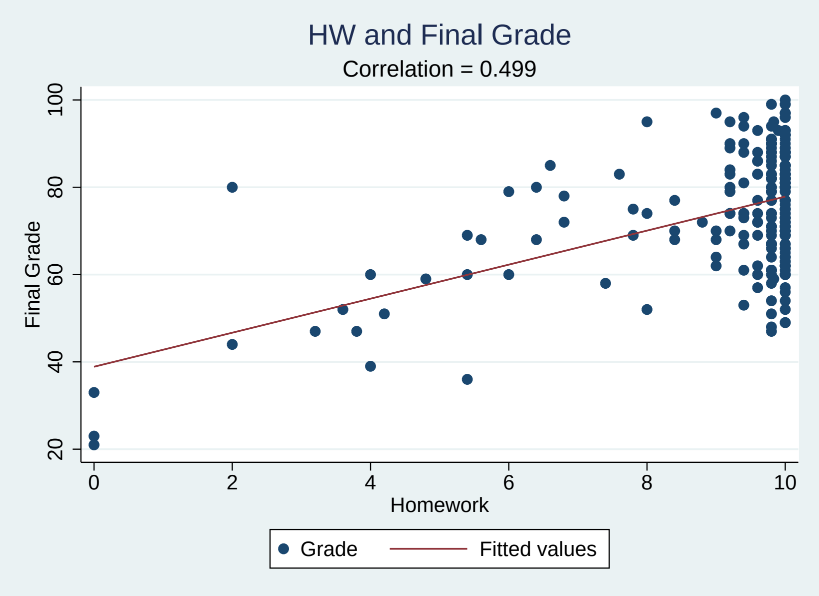
	Cancer	No Cancer	Total
Test Positive	85	1485	1570
Test Negative	15	8415	8430
Total	100	9900	10000

- p_y = the probability of having cancer if test is positive.
- p_x = the probability of having cancer if test is negative.

Construct 95 percent confidence interval for $p_y - p_x$.

Homework and Final Grades

Scatter Plot of HW Grade and Final Grade



Summary Statistics by Stata

Define Low HW group as students with HW grade less than 6 out of 10.

```
. gen Low_HW = 0

. replace Low_HW = 1 if hw<6

. sum grade if Low_HW==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       grade |       224      76.625    11.96154         47        100


. sum grade if Low_HW==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       grade |        16     49.3125    16.51956         21         80
```

95 confidence interval for $\mu_y - \mu_x$

- μ_y = population mean of final grade among students who receive HW grade more than 6 out of 10.
- μ_x = population mean of final grade among students who receive HW grade less than 6 out of 10.
- Sample:

$$\bar{Y} = 76.63, \quad \bar{X} = 49.31, \quad s_y = 11.96, \quad s_x = 16.51.$$

$$n_y = 224, \quad n_x = 16.$$

95 percent confidence interval for $\mu_y - \mu_x$?

95 percent Confidence Interval for $\mu_y - \mu_x$

Because

$$E(\bar{Y} - \bar{X}) = \mu_y - \mu_x,$$

$$Var(\bar{Y} - \bar{X}) = \frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x},$$

we have

$$\frac{(\bar{Y} - \bar{X}) - (\mu_y - \mu_x)}{\sqrt{\frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}}} \xrightarrow{d} N(0,1)$$

Because s_x^2 and s_y^2 converge in probability to σ_x^2 and σ_y^2, we replace s_x^2 and s_y^2 with σ_x^2 and σ_y^2.

95 percent Confidence Interval for $\mu_y - \mu_x$

Applying the Central Limit Theorem,

$$\Pr\left(-1.96 < \frac{(\bar{Y} - \bar{X}) - (\mu_y - \mu_x)}{\sqrt{\frac{s_y^2}{n_y} + \frac{s_x^2}{n_x}}} < 1.96\right) \approx 0.95$$

$\Leftrightarrow$

$$\Pr\left((\bar{Y} - \bar{X}) - 1.96\sqrt{\frac{s_y^2}{n_y} + \frac{s_x^2}{n_x}} < (\mu_y - \mu_x) < (\bar{Y} - \bar{X}) + 1.96\sqrt{\frac{s_y^2}{n_y} + \frac{s_x^2}{n_x}}\right) \approx 0.95$$

95 percent Confidence Interval for $\mu_y - \mu_x$

$$\bar{Y} = 76.63, \quad \bar{X} = 49.31, \quad s_y = 11.96, \quad s_x = 16.51.$$
$$n_y = 224, \quad n_x = 16.$$

95 percent confidence interval is given by

$$(76.63 - 49.31) \pm 1.96\sqrt{\frac{(11.96)^2}{224} + \frac{(16.51)^2}{16}}$$

or

$$27.32 \pm 8.24 = [19.08, 35.56]$$

$\Rightarrow$ Evidence that students who did well in HW do better in final grades than those who did not do well in HW.

Difference-in-differences (DID)

Figure: Does installing blue lights at train stations prevent suicides?

Blue lights and suicides at train stations

- Railway and metro suicides constitute a major problem in Japan.
- *Matsubayashi et al. (2014)* examines the effect of blue lights on the number of suicides by using panel data from 71 train stations between 2000 and 2013.
- Compare the number of suicides before and after the intervention of blue lights at 14 stations, using other stations without the intervention as a control group.
- **The effect of installing blue LED lamps on a decrease in the number of suicides is estimated at 74% (with 95% Confidence Interval given by 48–87%).**

Blue lights and suicides at train stations

Table 1
The average number of suicides before and after the installation of blue lights.

	(1) Station with blue lights Installed	(2) One station away	(3) Two stations away	(4) Three stations away	(5) Four stations away	(6) Five stations away	(7) Six and more stations away
Before	0.435 (115)	0.269 (182)	0.234 (201)	0.275 (189)	0.245 (200)	0.259 (220)	0.090 (546)
After	0.189 (53)	0.274 (84)	0.269 (93)	0.275 (91)	0.266 (94)	0.245 (102)	

Note: Table entries are the average number of suicides per year before and after the installation of blue lights with the number of station-year in parentheses. Data represent the number of suicides at 71 stations between 2000 and 2013. The total number of observations is 994.

Blue lights and suicides at train stations

Suppose that there is at most one suicide per station within one year.

- p_{y0} = a population fraction of stations with suicides before installation in Treatment group
- p_{y1} = a population fraction of stations with suicides after installation in Treatment group
- p_{x0} = a population fraction of stations with suicides before installation in Control group
- p_{x1} = a population fraction of stations with suicides after installation in Control group

Blue lights and suicides at train stations

Treatment group: $\hat{p}_{y0} = 0.435 \Rightarrow \hat{p}_{y1} = 0.189$

Control group: $\hat{p}_{x0} = 0.269 \Rightarrow \hat{p}_{x1} = 0.274$

We would like to construct the 95 percent confidence interval for

$$(p_{y1} - p_{y0}) - (p_{x1} - p_{x0}),$$

i.e., the difference in the changes in suicide rates after installing blue lights between the treatment group (stations that installed blue lights) and the control group (one station away).

Confidence Interval for $(p_{y1} - p_{y0}) - (p_{x1} - p_{x0})$

We assume that suicide events are independent across stations:

$$E[(\hat{p}_{y1} - \hat{p}_{y0}) - (\hat{p}_{x1} - \hat{p}_{x0})] = (p_{y1} - p_{y0}) - (p_{x1} - p_{x0}),$$

$$\begin{aligned} & Var((\hat{p}_{y1} - \hat{p}_{y0}) - (\hat{p}_{x1} - \hat{p}_{x0})) \\ &= Var(\hat{p}_{y1}) + Var(\hat{p}_{y0}) + Var(\hat{p}_{x1}) + Var(\hat{p}_{x0}) \\ &= \frac{p_{y1}(1 - p_{y1})}{n_{y1}} + \frac{p_{y0}(1 - p_{y0})}{n_{y0}} + \frac{p_{x1}(1 - p_{x1})}{n_{x1}} + \frac{p_{x0}(1 - p_{x0})}{n_{x0}}. \end{aligned}$$

we have

$$\frac{[(\hat{p}_{y1} - \hat{p}_{y0}) - (\hat{p}_{x1} - \hat{p}_{x0})] - [(p_{y1} - p_{y0}) - (p_{x1} - p_{x0})]}{\sqrt{\frac{p_{y1}(1-p_{y1})}{n_{y1}} + \frac{p_{y0}(1-p_{y0})}{n_{y0}} + \frac{p_{x1}(1-p_{x1})}{n_{x1}} + \frac{p_{x0}(1-p_{x0})}{n_{x0}}}} \xrightarrow{d} N(0,1)$$

95% Confidence Interval

By the Central Limit Theorem and the Law of Large Numbers, with probability 95 percent, $[(p_{y1} - p_{y0}) - (p_{x1} - p_{x0})]$ is within

$$(\hat{p}_{y1} - \hat{p}_{y0}) - (\hat{p}_{x1} - \hat{p}_{x0})$$
$$\pm 1.96 \times \sqrt{\frac{\hat{p}_{y1}(1-\hat{p}_{y1})}{n_{y1}} + \frac{\hat{p}_{y0}(1-\hat{p}_{y0})}{n_{y0}} + \frac{\hat{p}_{x1}(1-\hat{p}_{x1})}{n_{x1}} + \frac{\hat{p}_{x0}(1-\hat{p}_{x0})}{n_{x0}}}$$

The 95 percent Ci is given by

$$[-0.4314 - 0.0705]$$

$\Rightarrow$ installation of blue lights have likely reduced suicides.