Data Science 1
Statistical Inference for Two Population Proportions

Ann Maharaj

Statistical Inference for two Population Proportions

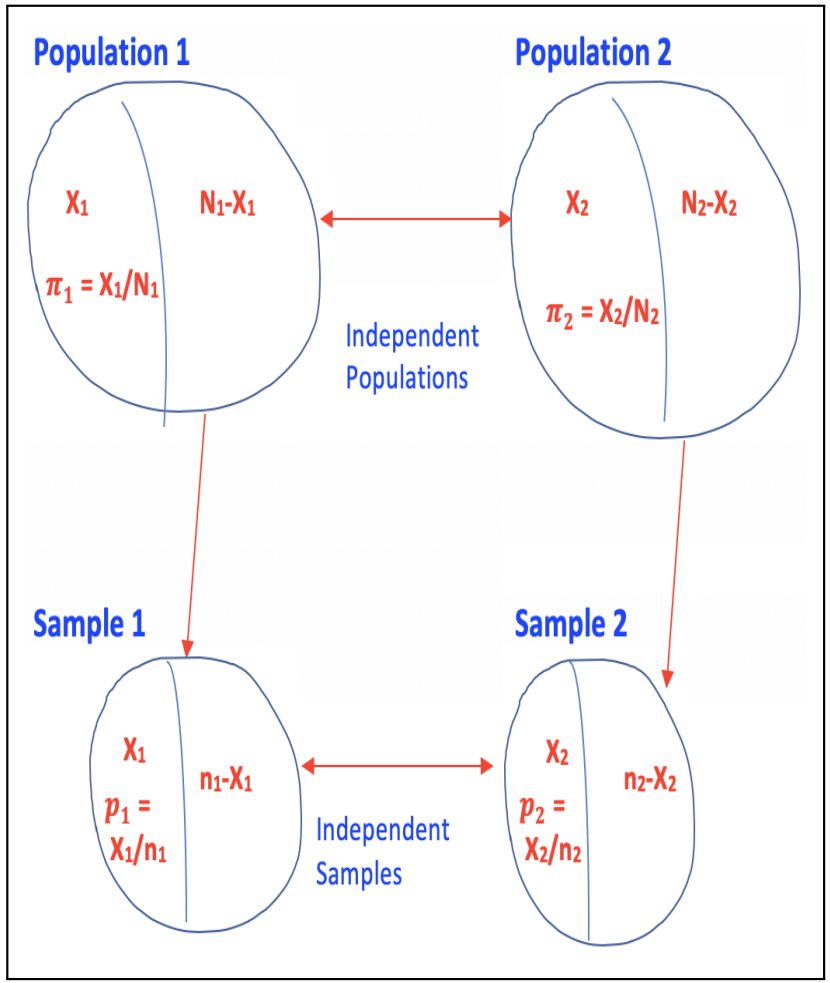
Introduction

There are many situations where we may want to compare proportions between two populations. For example:

- Comparing unemployment rates in rural and urban populations.
- Comparing the proportion of defective items produced by two competing manufacturing processes.

Introduction

- We are interested in a particular characteristic in each of two independent populations.
- Let X_1 and X_2 be the number of cases with this characteristic in the populations which are of size N_1 and N_2, respectively.
- X_1 and X_2 each follows a binomial distribution.
- Let π_1 and π_2 represent the population proportions, where $\pi_1 = X_1/N_1$ and $\pi_2 = X_2/N_2$.
- We take random samples of size n_1 and n_2 from the respective populations and obtain the sample proportions, $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$.
- The point estimator of $\pi_1 - \pi_2$ is $p_1 - p_2$.



Sampling Distribution

- Since the populations are independent:
  - The mean of $p_1 - p_2$ is

$$E(p_1 - p_2) = \pi_1 - \pi_2,$$

    implying that $p_1 - p_2$ is an unbiased estimator of $\pi_1 - \pi_2$.
  - The standard deviation (standard error) of $p_1 - p_2$ is

$$SE(p_1 - p_2) = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}.$$

- For large sample sizes, it can be shown that the sampling distribution of $p_1 - p_2$ is approximately normal.
- Then the standardised difference $p_1 - p_2$,

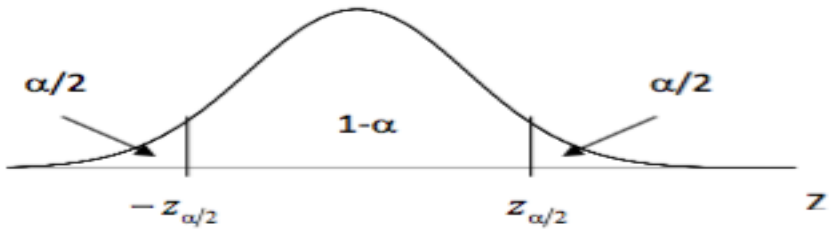
$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)} \sim N(0, 1).$$

Confidence Interval Estimation

Since the standardised difference between the sample proportions, $p_1 - p_2$ follows a $N(0,1)$ distribution, and

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-z_{\alpha/2} < \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)} < z_{\alpha/2}\right) = 1 - \alpha,$$



It can be shown that on making $\pi_1 - \pi_2$ the subject of the formula:
A $100(1-\alpha)\%$ Confidence Interval estimator of $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{\alpha/2} \times SE(p_1 - p_2).$$

Replacing π_1 and π_2 with p_1 and p_2, respectively, the standard error term becomes

$$SE(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Table: 1

Null Hypothesis	$H_0 : \pi_1 - \pi_2 = D_0$	$H_0 : \pi_1 - \pi_2 = D_0$ or $H_0 : \pi_1 - \pi_2 \leq D_0$	$H_0 : \pi_1 - \pi_2 = D_0$ or $H_0 : \pi_1 - \pi_2 \geq D_0$
Alternative Hypothesis	$H_1 : \pi_1 - \pi_2 \neq D_0$	$H_1 : \pi_1 - \pi_2 > D_0$	$H_1 : \pi_1 - \pi_2 < D_0$
Level of Significance	α	α	α
Test Statistic	$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)}$	$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)}$	$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)}$
P-Value Approach Reject H_0 if	$p - value < \alpha$	$p - value < \alpha$	$p - value < \alpha$

If $D_0 \neq 0$

$$SE(p_1 - p_2) = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}.$$

If $D_0 = 0$ the common proportion is estimated by the pooled proportion

$$\overline{p} = \frac{n_1}{n_1 + n_2} p_1 + \frac{n_2}{n_1 + n_2} p_2.$$

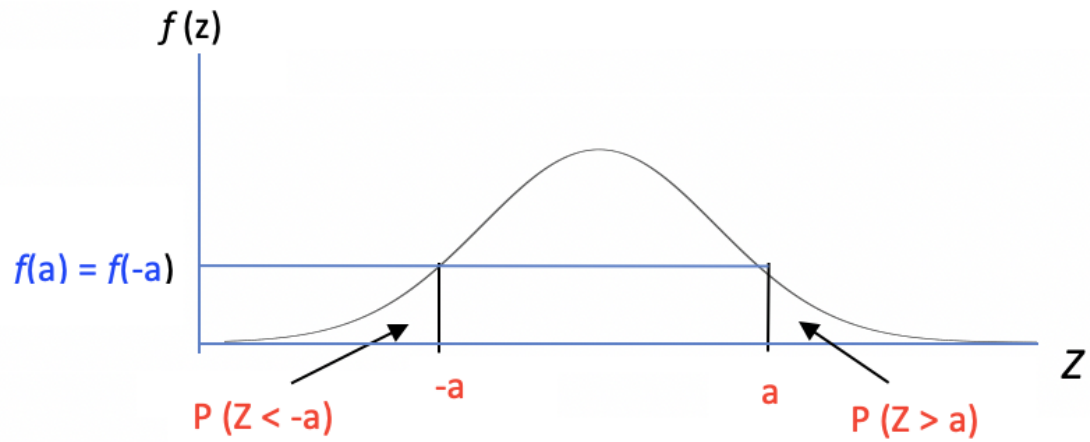
Hence

$$SE(p_1 - p_2) = \sqrt{\overline{p}(1 - \overline{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

P-value approach

$H_0 : \pi_1 - \pi_2 = D_0$
$H_1 : \pi_1 - \pi_2 \neq D_0$

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)} \sim N(0, 1)$$



a = |numerical value of the test statistic|

$f(z)$ is the density function of the standard normal distribution and $f(-a) = f(a)$

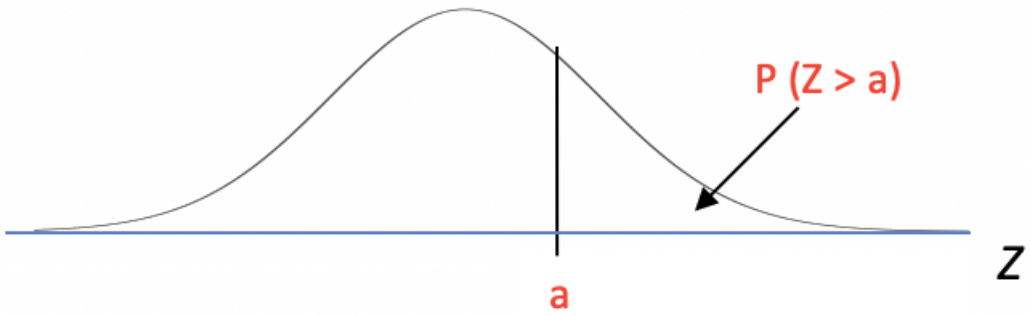
p-value = P(Z > a) + P(Z < -a)

Reject H_0 if $p - value < \alpha$

P-value approach

$H_0 : \pi_1 - \pi_2 = D_0$ or $H_0 : \pi_1 - \pi_2 \leq D_0$
$H_1 : \pi_1 - \pi_2 > D_0$

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)} \sim N(0, 1)$$



a = numerical value of the test statistic

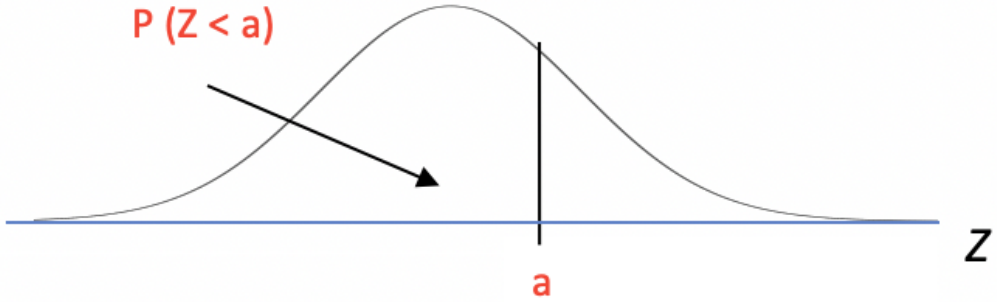
p-value = P(Z > a)

Reject H_0 if $p - value < \alpha$

P-value approach

$H_0 : \pi_1 - \pi_2 = D_0$ or $H_0 : \pi_1 - \pi_2 \geq D_0$
$H_1 : \pi_1 - \pi_2 < D_0$

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(p_1 - p_2)} \sim N(0, 1)$$



a = numerical value of the test statistic

p-value = P(Z < a)

Reject H_0 if $p - value < \alpha$

A lecturer who prefers innovative teaching methods wishes to compare the effectiveness of teaching Statistics by the traditional classroom lecture method and by the extensive use of audio-visual aids. To do so, 100 students were selected at random from a class of 250 and were assigned to audio-visual instruction. The remaining 150 students are taught Statistics in classroom lectures. At the end of the semester all 250 students are given a test. The number of students from each group who passed the test is given in the following table.

Table: 2

	Audio-Visual Instruction	Classroom Lecture
Pass	63	107
Fail	37	43
Total	100	150

1. Obtain and report a 95% confidence interval estimate of the difference between success rates for the two methods of instruction.
2. Do the data support the hypothesis that a better pass rate is achieved using the classroom lecture method, than is achieved using the audio-visual method?

Example - Part 1

π_1 and π_2: population proportions associated with the classroom lecture and audio-visual instruction methods, respectively.

```
> #Classroom Lecture
> n1 <- 150
> pass1 <- 107
> fail1 <- 43
> #sample proportion
> p1 <- pass1/n1
> p1
[1] 0.7133333
> #Audio-Visual Instruction
> n2 <- 100
> pass2 <- 63
> fail2 <- 37
> #sample proportion
> p2 <- pass2/n2
> p2
[1] 0.63
> #difference bertween sample proportions
> p1-p2
[1] 0.08333333
> #standard error
> se <- sqrt((p1*(1-p1)/n1) +(p1*(1-p1)/n1))
> se
[1] 0.05221608
```

```
> #97.5th percentile of the standard normal distribution
> z025 <- qnorm(.975)
> z025
[1] 1.959964
>
> #Lower confidence limit
> LCL <- (p1-p2) - (z025 *se)
> LCL
[1] -0.01900829
>
> #Upper confidence limit
> UCL <- (p1-p2) + (z025 *se)
> UCL
[1] 0.185675
>
> #95% confidence interval estimate of pi1-pi2
> cbind(LCL, UCL)
              LCL      UCL
[1,] -0.01900829 0.185675
```

- The 95% confidence interval estimate of the difference in population proportions $\pi_1 - \pi_2$ is $[-0.0190, 0.1857]$
- We conclude with 95% confidence that the difference between success rates for the two methods of instruction is between -2% and 19%
- Because the confidence interval contains zero, the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$ is not rejected in favour of the alternative $H_1 : \pi_1 - \pi_2 \neq 0$.

Example: Part 2

π_1 and π_2: population proportions associated with the classroom lecture and audio-visual instruction methods, respectively.

1. Hypotheses:
   $H_0 : \pi_1 - \pi_2 = 0 \quad H_1 : \pi_1 - \pi_2 > 0$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $Z = 1.384$
4. P-value: $p - value = 0.083$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level of significance. There is insufficient evidence to conclude that the classroom lecture method is superior to the audio-visual instruction method.

```
> #H0: pi1-pi2 = 0; H1: pi1-pi2 > 0
> #pooled proportion
> pbar <- (n1/(n1+n2)*p1) + (n2/(n1+n2)*p2)
> pbar
[1] 0.68
> sep <- sqrt(pbar*(1-pbar)*(1/n1 + 1/n2))
> sep
[1] 0.06022181
> #Compute test statistic
> ts <- (p1-p2)/sep
> ts
[1] 1.383773
> #p-value = P(Z > ts)
> pvalue <-pnorm(ts,lower.tail = FALSE )
> pvalue
[1] 0.08321395
```

However, if we test the hypothesis at the 10% level of significance, we will reject H_0 and conclude that the classroom lecture method could be superior to the audio-visual instruction method.