Comparison of Two Proportions

A proportion is a summery measure for a categorical variable. When we want to compare two independent samples in relation to a categorical variable such as prevalence of smoking in boys and girls, or prevalence of TB in two cities, we use tests that compare proportions. For example a study on female genital mutilation in Kurdistan reported prevalence of 29% in Sulaimaniyah province and 37% in Erbil province. There is an observed difference of 8% (37%- 29%) in the prevalence of FGM between the two independent samples. This difference could be either due to chance (random variation) or it could be a true difference between the two regions. How can we make a decision on this? In this session we will discuss how to calculate 95% CI for a each of the two proportions, 95% CI for the difference between the two proportions and how to perform a significance test on the difference of the two proportions. The procedure used for comparing two independent proportions is somewhat similar to comparison of two independent means.

Example
A study was done to estimate prevalence of FGM in Kurdistan. 502 females in Erbil and 503 females in Sulaymaniyah were studied. Prevalence of FGM was 37% in Erbil and 29% in Sulaymaniyah.

Calculating standard deviation and standard error for a proportion
The standard deviation of a proportion is equal to the square root of p(1-p) using formula below:

SD(p)=√p(1-p)
Where p is the proportion

Example: find the SD of the proportion if FGM in Sulaymaniyah
SD(p)=√p(1-p)
SD(FGM)= √0.29(1-0.29)= √0.29X0,71)= √0.21=0.45

Note: if we use percentage instead of proportion we use the following formula

SD(p)=√p(100-p)
SD(FGM)= √29(100-29)= √29X71)= √210=45%

To calculate standard error, we divide the standard deviation by the square root of the sample size using the formula below

SE(p)= SD/√n
Example: The standard error of the FGM in Sulaymaniyah
SE(FGM)= SD/√n=0.45/√503=0.02 or 2%

We can also use the following formula directly without calculating the SD *(not required)*

$$\text{Standard error} = \sqrt{\frac{p(1-p)}{n}}$$

Example: SE(FGM)= √[29(1-29)/503]= √[29X71)/503]= √2059/503=√4=2%

Calculating 95% Confidence Interval for a proportion
The general formula for the 95% CI applies for the proportion

95% CI (p) =p $\pm$ 1.96 x SE (proportion)

Example: calculate 95% CI for the FGM prevalence in Sulymaniyah

95% CI (FGM) =p $\pm$ 1.96 x SE (proportion)
P=29%
SE=2%

Lower limit 95% CI:
p-1.96 SE= 29-1.96X2=29-4=25%

Upper limit 95% CI:
p +1.96 SE= 29 +1.96X2=29 +4=33%

What does this mean?

This means that we are 95% confident that the prevalence of FGM in Sulaymaniyah is somewhere between 25% and 33%.

Exercise: the same study found that prevalence of FGM in Kurdistan was 23% with a sample size of 1508, calculate the 95% CI for the prevalence of FGM in Kurdistan.

Comparison of two proportions
If want to know whether the difference in prevalence of FGM in Erbil and Sulaimaniyah is true, we need to do a hypothesis test for comparison of two proportion. To do a hypothesis test for this data we have to:

1. Calculate the difference in proportions (FGM prevalence)
2. Calculate the standard error of difference in proportions
3. Calculate the 95% confidence interval for this difference in prevalence of FGM.
4. Finally we perform the hypothesis test

1) The difference in proportions (p1-p2)
To calculate the difference in the two proportions we simply subtract one from the other

Difference in proportions=proportion in Erbil – proportion in Sulaimaniyah
=0.37-0.29=0.08

We can also use percentages

Difference in percentages= percentage in Erbil – percentage in Sulaymaniyah
=37%-29%=8%

2) The standard error of the difference , SE(p1-p2)

The standard error of the difference in two proportions is a combination of the two standard errors. We calculate the SE for the difference using the following formula *(not required)*

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

The standard error for the difference in the proportion of FGM in Sulaimaniyah and Erbil calculated with this formula is 0.03 or 3%.

3) The 95% CI for the difference

The 95% CI for the difference in two proportions is calculated using this formula

95% CI (p1-p2) =(p1-p2) ± 1.96 x SE (p1-p2)

Lower limit:
(p1-p2) – 1.96 SE(p1-p2)= (37%-29%) – 1.96X 3%=8%- 6%=2%

Upper limit
(p1-p2) + 1.96 SE(p1-p2)= (37%-29%) + 1.96X 3%=8% + 6%=14%

What does this mean? This means that we are 95% confident that there is a difference between prevalence of FGM in Erbil and Sulaimaniyah and this difference is somewhere between 2% and 14%. Since the range of the confidence interval does not include zero, the difference is probably true. But we have to do a hypothesis test to confirm this.

4) Hypothesis test for the difference between two proportions

To do a hypothesis test we first formulate the null and the alternative hypothesis. The null hypothesis says that there is no difference i.e. the difference between the two proportions is zero

H_0: (p1-p2)=0

The alternative hypothesis says that the two proportions are truly different i.e. the difference is not equal to zero:

H_1: (p1-p2)≠0

The hypothesis test calculates a z statistics using the following formula:

z= (p1-p2)/SE(p1-p2)

$$z = \frac{(p_1 - p_2)}{SE(p_1 - p_2)}$$

However the standard error is not the same standard error we calculated in point 2 above. This is another standard error for the common proportion, *p,* (a proportion between the two proportions) called pooled standard error which we calculate using the formula below *(not required):*

$$SE(p_1 - p_2) = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$ *Where* $$p = \frac{r_1 + r_2}{n_1 + n_2}$$

The standard error for the difference in the proportion of FGM in Sulaimaniyah and Erbil calculated with this formula is 0.03 or 3%.

Now we can calculate z statistics

z=(p1-p2)/SE(p1-p2)=8/3=2.7

In a table of normal distribution we look for z=2.7. This value of z corresponds to p=0.007 which is a very low probability of 7 per one thousand. Since this is smaller than 5%, we reject the null hypothesis (no difference) and accept the alternative hypothesis (there is difference). The difference which w have observed in our study could be due to chance only in 7 per 1000 cases which is a very low probability, therefore this difference is real difference. We conclude that there is a statistically significant difference in the prevalence of FGM between Sulaimaniyah and Erbil. This true difference is somewhere between 2% and 14% as shown by the 95% CI.

Table A1 Standard Normal distribution.

z	2-tailed *P*-value
0.0	1.000
0.1	0.920
0.2	0.841
0.3	0.764
0.4	0.689
0.5	0.617
0.6	0.549
0.7	0.484
0.8	0.424
0.9	0.368
1.0	0.317
1.1	0.271
1.2	0.230
1.3	0.194
1.4	0.162
1.5	0.134
1.6	0.110
1.7	0.089
1.8	0.072
1.9	0.057
2.0	0.046
2.1	0.036
2.2	0.028
2.3	0.021
2.4	0.016
2.5	0.012
2.6	0.009
2.7	0.007
2.8	0.005
2.9	0.004
3.0	0.003