

MAT1372, Classwork20, Fall2025

6.2 Difference of Two Proportions

1. A 30-year study was conducted with nearly 90,000 female participants. During a 5-year screening period, each woman was randomized to one of two groups: in the first group, women received regular mammograms to screen for breast cancer, and in the second group, women received regular non-mammogram breast cancer exams. We'll consider death resulting from breast cancer over the full 30-year period and the results are in the figure.

	Death from breast cancer?	
	Yes	No
Mammogram	500	44,425
Control	505	44,405

(a) What is the death rate in the treatment group? $p_t \approx \hat{p}_t = \frac{500}{500 + 44425} = \frac{500}{44925}$

(b) What is the death rate in the control group? $p_c \approx \hat{p}_c = \frac{505}{505 + 44405} = \frac{505}{44910}$

(c) Can we model the difference in sample proportions $p_t - p_c$ using the normal distribution? **Yes.**

2. Conditions for the Sampling Distribution of $\hat{p}_1 - \hat{p}_2$ to be Normal.

The difference $\hat{p}_1 - \hat{p}_2$ can be modeled using a normal distribution when

- *Independence, extended.* The data are independent within and between the 2 groups. Generally this is satisfied if the data come from 2 independent random samples

- *Success-failure condition.* The success-failure condition holds for both groups, where we check the condition separately. When it's satisfied, standard error of $\hat{p}_1 - \hat{p}_2$ is

$$SE = \sqrt{(SE_{p_1})^2 + (SE_{p_2})^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where p_1, p_2 represent the population proportions, and n_1, n_2 represent the sample sizes

3. Check whether we can model the difference in sample proportions using the normal distribution in 1.(c)?

Independent: this is a randomized experiment, this condition is satisfied

success-failure condition: since 500, 44425, 505, 44405 are all more than 10, so this condition is also satisfied.

With both conditions satisfied, " $p_1 - p_2$ " can be reasonably modeled using a normal distribution.

4. Confidence Intervals for $p_1 - p_2$.

$$\text{point estimate} \pm z^* \times SE \Rightarrow (p_1 - p_2) \pm \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \cdot z^*$$

5. Create and interpret a 95% confidence interval of the difference for the death rates in the breast cancer study.

Let p_t, p_c be the death rate in treatment and control group respectively.

$$p_t - p_c = \hat{p}_t - \hat{p}_c = 0.01113 - 0.01125 = -0.00012$$

$$SE = \sqrt{\frac{p_t(1-p_t)}{n_t} + \frac{p_c(1-p_c)}{n_c}} \approx \sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{n_t} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_c}} = \sqrt{\frac{(0.01113)(1-0.01113)}{44925} + \frac{(0.01125)(1-0.01125)}{44910}} = 0.00070191$$

Then 95% C.I. is $-0.00012 \pm 1.96 \cdot 0.00070191$
 $\Rightarrow (-0.001495, 0.001255)$

6. Set up hypotheses to test whether there was a difference in breast cancer deaths in the mammogram and control groups

H_0 : _____

H_A : _____

In this case, _____

7. Based on the Confidence interval in 5., can we reject the null hypothesis?

8. Definition of Pooled Proportion \hat{p}_{pooled} .

$$\hat{p}_{pooled} = \frac{\hat{p}_t n_t + \hat{p}_c n_c}{n_t + n_c}$$

In CPR case, we have $\hat{p}_{pooled} = \frac{0.95 \times 100 + 0.95 \times 100}{100 + 100} = 0.95$

This proportion is an average of the survival rate across the entire study, and it's our average of the proportions p_t and p_c if _____

9. Is it reasonable to model the difference in proportions using a normal distribution with \hat{p}_{pooled} in this study?

$$\hat{p}_{pooled} \times n_t =$$

$$(1 - \hat{p}_{pooled}) \times n_t =$$

$$\hat{p}_{pooled} \times n_c =$$

$$(1 - \hat{p}_{pooled}) \times n_c =$$

10. Testing Hypotheses for $p_t - p_c$ Using Significance Level.

Assume we choose significance level _____.

First, we calculate SE by \hat{p}_{pooled} : (why?)

$$SE = \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_t} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_c}} = \sqrt{\frac{0.95(1-0.95)}{100} + \frac{0.95(1-0.95)}{100}} = 0.0014$$

Second, we use the _____ and _____, calculate a p-value for the hypothesis test and write a conclusion:

$$Z = \frac{\hat{p}_t - \hat{p}_c}{SE} = \frac{0.95 - 0.95}{0.0014} = 0$$

In conclusion, _____

