1 Inferential Statistics for Proportions

1.1 Hypotheses test about a Population Proportion p

Instead of answering questions concerning a population mean, μ, we now want to answer questions about a population proportion, p.
Is the proportion of voters who support Mr Harper greater than 0.5?
Does the percentage of patients who survive at least 5 years after treatment exceed 60%?
Is this coin fair? (Is the probability to flip head 0.5?)

In general we have a feature in the population we are interested in: e.g. "support Mr Harper" or "survive at least 5 years". And we have questions about the proportion of members in the population who are described by this feature (we call theses successes).
The proportion of successes in the population, p, is estimated through the sample proportion

$$\hat{p} = \frac{\text{number of successes}}{\text{sample size}}$$

From the Central Limit Theorem for the sample proportion we learn:

1. $mu_{\hat{p}} = p$
2. $\sigma_{\hat{p}} = \sigma/\sqrt{n}$
3. If the sample size is large, then $\hat{p}$ is approximately normally distributed.

we know how we can standardize the sample proportion.
Therefore, if p is the true proportion of success in the population and the sample size is large, then:

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately normally distributed.
We can use this result to obtain a test about a proportion p based on this z-score.

z test about a population proportion p

1. Hypotheses:

   Choose p_0 (Given through the question).

Test type	**Hypotheses**
Upper tail	$H_0 : p \leq p_0$ versus $H_a : p > p_0$
Lower tail	$H_0 : p \geq p_0$ versus $H_a : p < p_0$
Two tail	$H_0 : p = p_0$ versus $H_a : p \neq p_0$

   Choose α.

2. Assumptions:
   The sample size is large, i.e. $pn > 5$, and $(1-p)n > 5$

3. Test statistic:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

4. P-value:

Test type	**P-value**
Upper tail	$P(z > z_0)$
Lower tail	$P(z < z_0)$
Two tail	$2 \cdot P(z > abs(z_0))$

5. Decision:
   $p - value \leq \alpha$ then we reject H_0 and accept H_a.
   $p - value > \alpha$ then we do not reject H_0 (fail to reject H_0).

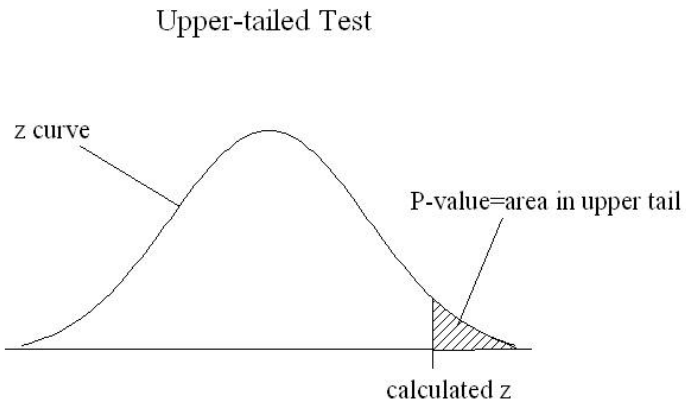
1. **Upper-tailed test:**

   $H_0 : p \leq p_0$ vs. $H_a : p > p_0$

   A large z will contradict the null hypothesis, so compute

   $$\text{P} - \text{value} = P(z > z_0) = 1 - P(z \leq z_0).$$

   Look up $P(z \leq z_0)$ in Table A.



2. **Lower-tailed test:**

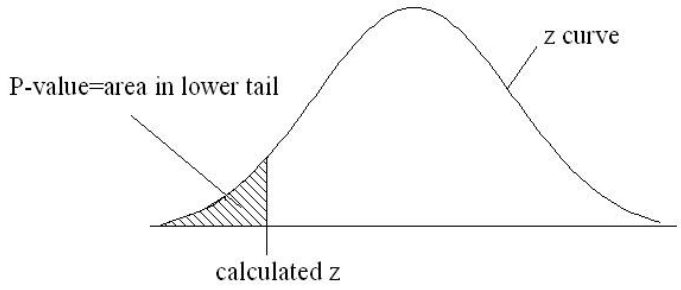
   $H_0 : p \geq p_0$ vs. $H_a : p < p_0$

   A small z will contradict the null hypothesis, so compute

   $$\text{P} - \text{value} = P(z < z_0).$$

   Look up $P(z < z_0)$ in Table A.

Lower-tailed Test



3. **Two-tailed test:**

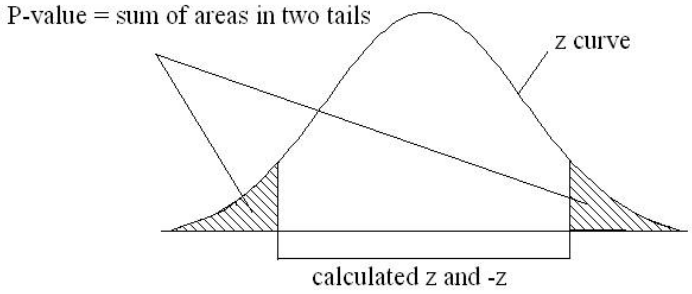
   $H_0 : p = p_0$ vs. $H_a : p \neq p_0$

   A small or large z will contradict the null hypothesis, so (with $abs(z_0)$ being the absolute value)

   $$\text{P} - \text{value} = P(z > abs(z_0) \text{ or } z < -abs(z_0)) = 2 \cdot P(z \leq -abs(z_0)).$$

   Look up P

Two-tailed Test



Example 1
A company wants to know if 30% of their employees are for the motion of weekly staff meetings. They ask a random sample of 69 people for their opinion.

1. Hypotheses: $H_0 : p = 0.3$ vs. $H_a : p \neq 0.3$. Choose significance level of $\alpha = 0.05$.

2. Assumptions: Random sample? Is the sample size large enough?

   $n = 69$ $p_0 = 0.3$, $n(0.3) = 20.7$ and $n(1 - 0.3) = 48.3$, both are greater than 5, so the sample size is large enough.

3. test statistic

   The number of people (out of 69) who find the meeting beneficial is 19, so that $\hat{p} = 0.275$.

   Then
   $$z_0 = \frac{\hat{p} - 0.3}{\sqrt{\frac{0.3 \cdot 0.7}{69}}} = -0.453$$

4. P-value:

   This is a two tailed test, so

   $$\text{P} - \text{value} = P(z > 0.453 \text{ or } z < -0.453) = 2 \cdot P(z \leq -0.453) = 2(.3264) = 0.6528.$$

   From the z-table (A).

5. Decision: Since 0.6528=P-value$>$ 0.05 $= \alpha$, H_0 is not rejected. There is not sufficient evidence to reject that $p = 0.3$.

6. Conclusion: The sample did not provide sufficient evidence at significance level 0.05, that the percentage of employees who agree with weekly staff meeting is different from 30

   Note: This does not mean that $p = 0.3$, only that THIS sample did not provide sufficient evidence against it.

1.2 A large sample confidence interval for a proportion p

For finding a confidence interval we use the results from the central limit theorem on the distribution of $\hat{p}$ and get

A large sample $(1 - \alpha) \times 100\%$ **confidence interval for a population proportion** p.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where z^* is the $(1 - \alpha/2)$ percentile of the standard normal distribution.

Example 2
Let's find a 93% confidence interval for the proportion p of employees who support the motion of weekly staff meetings. $\hat{p} = 0.275$, $\alpha = 0.07, \alpha/2 = 0.035, 1-\alpha/2 = 0.965$, therefore $z^* = 1.81$ (table A).

$$0.275 \pm 1.81 \sqrt{\frac{0.275(1 - 0.275)}{69}} \leftrightarrow 0.275 \pm 0.0973$$

We are 93% confident that the proportion of employees who support the motion falls between 0.1777 and .3723.

1.3 Plus 4 Method for Estimating population proportions

For small samples the large sample confidence interval is very inaccurate, it actually states a higher confidence level than it actually has. That is bad.
The following method leads to a more appropriate interval for small sample sizes.

A quick fix for this situation can be obtained by adding four imaginary observations (2 successes + 2 failures). Simulations have shown that the resulting confidence interval is better than the large sample confidence interval.

Plus 4 Confidence Interval for p
n= sample size, X = number of successes.
The plus 4 estimate of p:

$$\tilde{p} = \frac{X+2}{n+4}$$

with mean and standard deviation

$$\mu_{\tilde{p}} \neq p \quad \sigma_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

We get an approximate $(1-\alpha)100\%$ CI for p

$$\tilde{p} \pm z_{(1-\alpha/2)}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

This confidence interval is appropriate when the sample size is at least 10!

Example 3 A chess program is tested against a very good human player (Grandmaster). We want to estimate the winning probability of the program against the human player. They played 12 games, 4 were tied, 5 were won by the program, and 3 by the human.
We will find a 95% confidence interval for the winning probability of the program. The plus four estimate is

$$\tilde{p} = \frac{X+2}{n+4} = \frac{5+2}{12+4} = 0.4375$$

and the 95% ci is

$$0.4375 \pm 1.96\sqrt{\frac{0.4375(1-0.4375)}{12}} \quad \leftrightarrow \quad 0.4375 \pm 0.2431 \quad \leftrightarrow \quad [0.1944, 0.6806]$$

We are 95% confident that the winning probability of this program against the human Grandmaster falls between about 20% and 68%.

1.4 Choosing the sample size

The margin of error for the large sample confidence interval is

$$ME = z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If we desire to design a study which will result in a confidence interval no wider than $\pm$ a certain amount m, the sample size should be at least

$$n \geq \left(\frac{z^*}{m}\right) p^*(1-p^*)$$

For p^* we either choose a value from a former study, or preliminary study, or we can use the most conservative value, 0.5, which will result in highest sample size possible.

Example 4 Assume a group is planning a poll before an election. They are interested in estimating the proportion of voters who support the incumbent candidate. They plan on finding a 95% confidence interval which should not be wider than ±3%. Therefore $z^* = 1.96$, $m = 0.03$, and use $p^* = 0.5$, then

$$n \geq \left(\frac{1.96}{0.03}\right) 0.5(1 - 0.5) = 1067.1$$

They should use a sample of at least 1068 voters.
Check for polls in papers. They usually involve a little above 1000 participants, this is because they do the same calculation we just did.

1.5 Statistical Inference for Two Population Proportions p_1 and p_2

Instead of comparing population means, like the mean survival time for different treatment, or the mean amount spend by visitors in different malls, we now want to compare proportions that describe two populations.
For example: According to the proportion of wins, who is the better chess player, Jack or Jill? Or, is the proportion of international students different at Grant MacEwan and Mount Royal Universities?
Or, Is the percentage of male and female students who work while attending school the same? In order to make decisions, we will have to rely on samples again.

Notation:

	population	sample		
	proportion	size	successes	proportion
population 1	p_1	n_1	x_1	$\hat{p}_1 = x_1/n_1$
population 2	p_2	n_2	x_2	$\hat{p}_2 = x_2/n_2$

If we want to estimate the difference in the population proportions, $p_1 - p_2$ based on sample data, the first point estimator that comes to our mind (I hope) is the difference in the sample proportions, $\hat{p}_1 - \hat{p}_2$.
In order to be able to develop a test and to find a confidence interval using this estimator, we have to discuss its distribution.

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$ from two independent samples.

- For the mean: $\mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$, so that $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator for $p_1 - p_2$.
- For the variance:
$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$
- For the standard deviation:
$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Shape: If n_1 and n_2 are both large (that is, if $n_1p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2p_2 \geq 5$, and $n_2(1-p_2) \geq 5$), then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

Conclusion:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

is approximately standard normally distributed.

Before we can use this z-score for testing and constructing confidence intervals, we have to find an estimator for the standard deviation of $\hat{p}_1 - \hat{p}_2$, since to calculate z, p_1 and p_2 (the population proportions we are seeking) have to be known.

Definition:
If the population proportions are equal, that is $p = p_1 = p_2$, then

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

is the combined (or pooled) estimator of the common population proportion p.
If $p_1 = p_2$ we can combine the two samples to estimate the common proportion.

Large-Sample z Test for comparing p_1 and p_2

1. Assumption: Both sample sizes are large:

$$n_1p_1 \geq 5,\ n_1(1-p_1) \geq 5,\ n_2p_2 \geq 5,\ n_2(1-p_2) \geq$$

2. Hypotheses:

Test type	hypotheses
Upper tail	$H_0 : p_1 - p_2 \leq 0$ versus $H_a : p_1 - p_2 > 0$
Lower tail	$H_0 : p_1 - p_2 \geq 0$ versus $H_a : p_1 - p_2 < 0$
Two tail	$H_0 : p_1 - p_2 = 0$ versus $H_a : p_1 - p_2 \neq 0$

3. Test statistic:

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$$

4. P-value/Rejection Region:

Test type		**P-value**
Upper tail	$P(z > z_0)$	$z_0 > z_\alpha$
Lower tail	$P(z < z_0)$	$z_0 < z_\alpha$
Two tail	$2 \cdot P(z < -abs(z_0))$	$abs(z_0) > z_{\alpha/2}$

5. Decision: If P-value$\leq \alpha$ or z_0 falls into the rejection region, then reject H_0.
   If P-value$> \alpha$ or z_0 does not fall into the rejection region, then do not reject H_0.

6. Interpretation: Put decision into context.

Large-Sample z Confidence Interval for $p_1 - p_2$

Assumption: Both sample size are large:

$$n_1 p_1 \geq 10,\ n_1(1-p_1) \geq 10,\ n_2 p_2 \geq 10,\ n_2(1-p_2) \geq 10$$

The $(1-\alpha)$ Confidence Interval for μ_d:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ percentile of the standard normal distribution (Table A).

Example 5 Suppose you want to compare the infestation by a specific pest of two forests. Let p_1 be the proportion of trees in forest 1 that are affected and p_2 be the proportion of trees in forest 2 that are affected.
Two samples are drawn
sample 1: $n_1 = 100$ $x_1 = 10$ $\hat{p}_1 = 10/100 = 0.1$
sample 2 $n_2 = 100$ $x_2 = 15$ $\hat{p}_2 = 15/100 = 0.15$ and $p_c = (10+15)/(100+100) = 25/200 = 0.125$

Test is the infestation is different in those two forests.

1. Assumption is met, since $n_1\hat{p}_1 \geq 5,\ n_1(1-\hat{p}_1) \geq 5,\ n_2\hat{p}_2 \geq 5,\ n_2(1-\hat{p}_2) \geq 5$.

2. Hypotheses: $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$ at a significance level of $\alpha = 0.05$.

3. Test statistic:

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}} = \frac{(0.1 - 0.15)}{\sqrt{\frac{0.125(1-0.875)}{100} + \frac{0.125(1-0.875)}{100}}} = -0.05/0.046 = -1.087$$

4. p-value: This is a two tailed test, so we get p-value$=2 \cdot P(z < -abs(z_0))=2 \cdot 0.1379 =$ 0.2748 (use table A).

5. Decision: Since the p-value is greater than α, H_0 is not rejected.

6. Interpretation: The data does not provide enough evidence at significance level 0.05 to show that there is a difference in infestation between the two forests.

We give a 95% confidence interval for the difference in infestation between the two forests.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{(1-\alpha/2)}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

this becomes for the these data

$$(0.1 - 0.15) \pm 1.96\sqrt{\frac{0.1(0.9)}{100} + \frac{0.15(0.85)}{100}}$$

that is

$$(-0.05) \pm 0.0914 \leftrightarrow [-0.1414, 0.0414]$$

Since zero is captured within the confidence interval, zero is one of the possible values for $p_1 - p_2$. The data does not exclude the the possibility that in fact $p_1 = p_2$. This does not mean that the proportions are the same. The interval includes infinite may other values for the difference.

1.6 Plus 4 Confidence Interval for $\mathbf{p_1 - p_2}$

Let n_1= sample size from the sample from population 1,
X_1 = number of successes in sample 1,
n_2= sample size from the sample from population 2,
X_2 = number of successes in sample 2.

The plus 4 estimate of $p_1 - p_2$:

$$\tilde{p_1} = \frac{X_1 + 1}{n_1 + 2}, \quad \tilde{p_2} = \frac{X_2 + 1}{n_2 + 2}$$

standard deviation

$$\sigma_{\tilde{p_1} - \tilde{p_2}} = \sqrt{\frac{\tilde{p_1}(1-\tilde{p_1})}{n_1 + 2} + \frac{\tilde{p_2}(1-\tilde{p_2})}{n_2 + 2}}$$

We get an approximate $(1-\alpha)100\%$ CI for $p_1 - p_2$

$$(\tilde{p_1} - \tilde{p_2}) \pm z_{(1-\alpha/2)}\sqrt{\frac{\tilde{p_1}(1-\tilde{p_1})}{n_1 + 2} + \frac{\tilde{p_2}(1-\tilde{p_2})}{n_2 + 2}}$$

This confidence interval is appropriate when both sample sizes are at least 10!

Example 6 Is it true that young adults (20-30) play more computer games more often than older adults (30-40)?
The following are results for two random samples. x is the number of people who said they played computer games at least once a week.

young adults:	$n_1 = 10$	$x_1 = 4$	$\tilde{p}_1 = 4/12 = 0.3333$
older adults	$n_2 = 11$	$x_2 = 2$	$\hat{p}_2 = 3/13 = 0.2308$

A 95% confidence interval for $p_1 - p_2$, the difference in the proportion of people who play at least once a week a computer game for young and older adults, is given by

$$(0.3333 - 0.2308) \pm 1.96\sqrt{\frac{0.3333(0.6666)}{12} + \frac{0.2308(0.7692)}{13}}$$

$$0.0025 \pm 0.3517 \rightarrow [-0.3492, 0.3542]$$

We are 95% confident that the difference in the proportions of people who play at least once a week a computer game for young and older adults falls between -0.3492 and 0.3542. Since 0 falls within the interval, we do not have sufficient evidence that the proportion differs for young and older adults at 95% confident.