CS555 Data Analysis and Visualization

Lecture 11

One and Two-Sample Tests for Proportions

Kia Teymourian

One-Sample Tests for Proportions

- We are interested in the **proportion of the population, p**, that has a particular outcome.

- The **population parameter, p**, is unknown and we aim to estimate it.

- As in previous sections, we seek to make conclusions about the population parameter (p in this case) by using information from a sample. The **sample proportion, denoted as pˆ**, is an estimate of the **population parameter, p**.

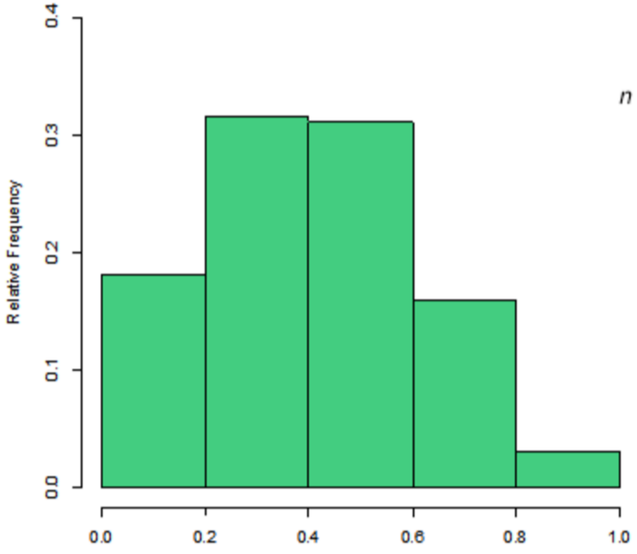
- We can look at the sampling distribution of the sample proportion to see how well it estimates the population proportion.

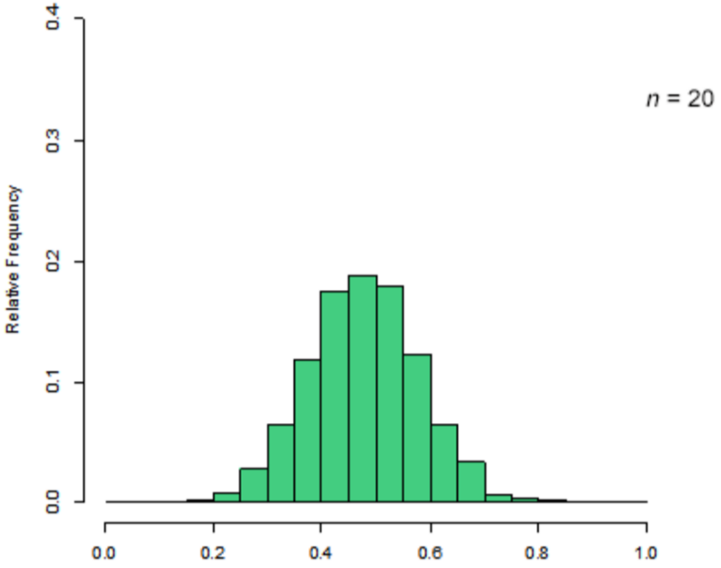
An example: One-Sample Tests for Proportions

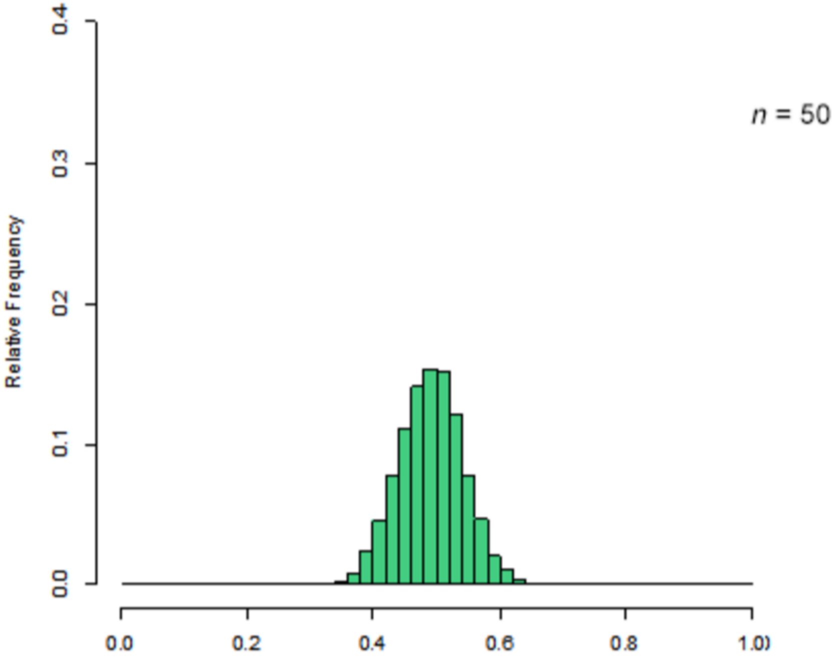
We are trying to estimate the proportion of all **dice rolls** that result in a **value >3**. That is, we’d like to estimate how often the die will roll a 4, 5, or 6. For simplicity, we will call rolling a 4, 5, or 6 a “**success**.”

We know that the probability of “success” using a fair die is 3/6 or 1/2 or 50%. Thus the underlying population parameter, p, is 0.50.
Let’s assume we roll a die 5 times. In the 5 times, we get a 2, 2, 5, 6, and 3. In this sample, we get 2/5=40% as our sample proportion of “successes”.

As n increases, the distribution of the **sample proportion looks more and more like a normal distribution**.
Also, the variability of the **sample proportion (and the width of the distribution) decreases as n increases**.
In the graphs, the axes are the same so as to facilitate easier comparisons across graphs.







One-Sample Tests for Proportions

We look at the sampling distribution of $\hat{p}$ as sample size increases:

1. As the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.
2. The mean of the sampling distribution is p.
3. The standard deviation of the sampling distribution decreases as the sample size increases. In fact, the variability (as measured by the standard deviation) of $\hat{p}$ is given by

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

These properties mimic the properties of the sampling distribution of the sample mean as discussed in relation to the central limit theorem.

The Central Limit Theorem:
When the number of samples taken from a population is sufficiently large, the sampling distribution of the sample mean, $\bar{x}$, will be approximately normally distributed with an expected value of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$ where μ and σ are the mean and the standard deviation from the population.

One-Sample Tests for Proportions

Given the nature of the **data being dichotomous** (taking on one of two values “Success” and “failures”) in this setting, there are a few caveats.

- In general the normal distribution only holds when the **sample size is sufficiently large.**
- The formula noted above for the standard deviation of **p^ is only accurate when the population is much larger than the sample** (at least 10 times larger).
- Further, the normal approximation of the sampling distribution of p^ **works poorly when the underlying population parameter p is close to 0 or close to 1.**
- When p is close to 0 or 1, much larger samples are required.

Significance Tests for a Proportion

Tests of hypotheses about a population proportion p are based on the sampling distribution of the sample proportion $\hat{p}$ and use data from a sample to evaluate evidence against the null hypothesis.

We are interested testing the null hypothesis H0:p=p_0. To do this, we use the following z-statistic

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

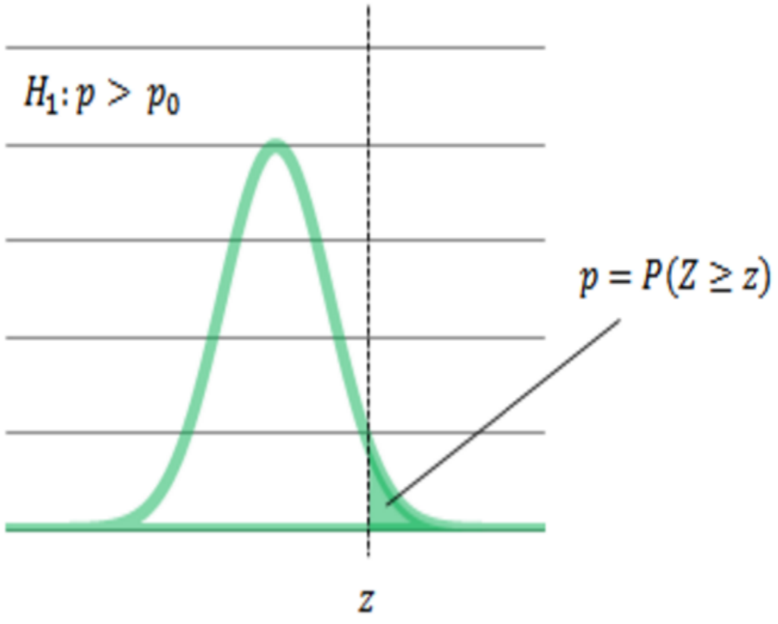
where $\hat{p}$ is the sample proportion, p_0 is the value of the population proportion under the null hypothesis, and n is the number of observations in the sample.

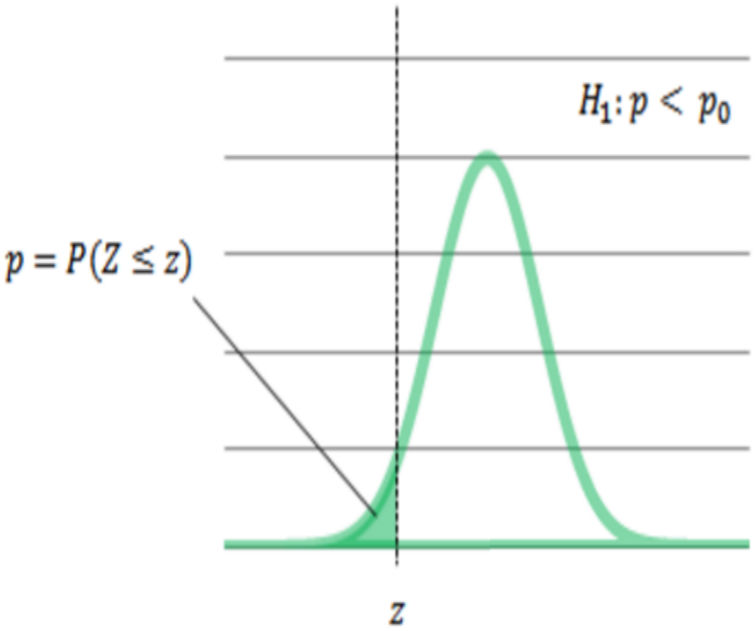
This test is valid if the sample size n is sufficiently large.

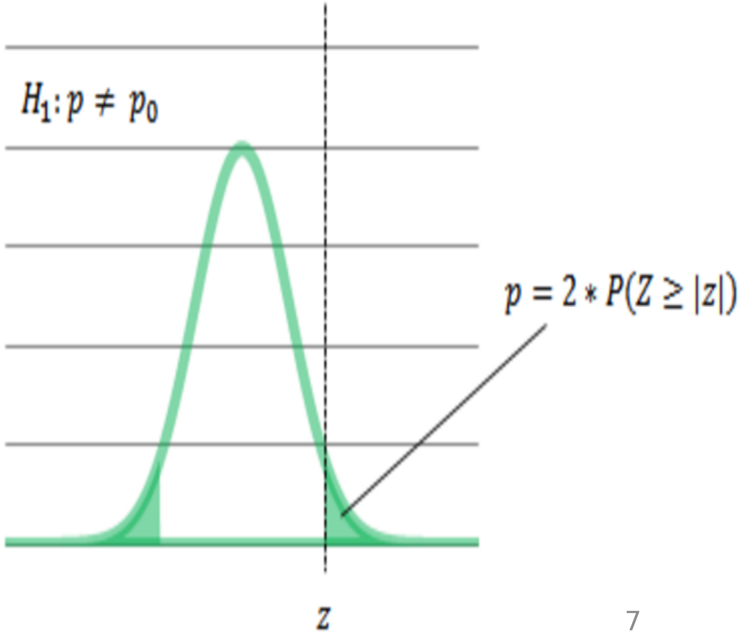
In general, we use the condition that inference is valid if $n \cdot p_0$ and $n \cdot (1 - p_0)$ are both greater than 10.

Significance Tests for a Proportion

- The test statistic for tests relating to the population proportion measure how far p^ is from the value of p (under the null hypothesis) **in standard deviation units**.
- The **z-statistic is approximately normally distributed** with a mean of 0 and a standard deviation of 1 (the standard normal distribution).
- This result allows us to quantify **how far the point estimate is from the expected value** of the population parameter under the null hypothesis and to make inference about the population proportion.
- This is generally done by **calculating the corresponding p-value**. The **p-value for the z statistic** described above is calculated using the **standard normal distribution**.







An Example: Significance Tests for a Proportion

We are interested in estimating the proportion of children in the county that are vaccinated for measles. We suspect that it may be as low as 80%. A random sample is taken of 100 children from the county. Of those sampled, only 70 were vaccinated. Formally test if the proportion of vaccinated children is different than 80%.

1. Set up the hypotheses and select the alpha level
H_0:p=0.80 (the underlying population proportion is 80%)
H_1:p≠0.80 (the underlying population proportion is different than 80%)
α=0.05

2. Select the appropriate test statistic $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

3. State the decision rule
Determine the appropriate critical value from the standard normal distribution table associated with a right hand tail probability of α=0.05/2=0.025. Using the table, the appropriate critical value is 1.960.
Decision Rule: Reject H_0 if |z|>1.96
Otherwise, do not reject H_0

An Example: Significance Tests for a Proportion

4. Compute the test statistic and the associated p-value

$$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.7-0.8}{\sqrt{\frac{0.8(1-0.8)}{100}}} = \frac{-0.1}{\sqrt{0.0016}} = -2.5$$

5. Conclusion

Reject H_0 since |−2.50|≥1.960. We have significant evidence at the α=0.05 level that p≠0.80 (p=2*0.0062=0.0124). That is, we reject the null hypothesis that the proportion of children in the county that is vaccinated is 80%. In our sample, just 70% of children were vaccinated.

Confidence Intervals for a Proportion

To calculate a confidence interval with a confidence level of C for the population proportion, p, we'd like to use the following formula:

$$\hat{p} \pm z \cdot \sigma_{\hat{p}}$$

where $\hat{p}$ is the sample proportion, z is the appropriate critical value corresponding to the confidence level, and $\sigma_{\hat{p}}$ is the standard deviation of $\hat{p}$.

However, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ and p (the population proportion) is not known (and if it was, we wouldn't need to calculate a confidence interval!).

We must instead estimate the quantity.

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

And the formula for the confidence interval for the population proportion, p, is given by

$$\hat{p} \pm z \cdot \mathrm{SE}_{\hat{p}}=\hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

An example: Confidence Intervals for a Proportion

Continue with the measles vaccination example.

The 95% confidence interval is calculated as follows:

$$\hat{p} \pm z \cdot \mathrm{SE}_{\hat{p}} = \hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7 \pm 1.96 \cdot \sqrt{\frac{0.7(1-0.7)}{100}} = 0.7 \pm 0.90 \approx (0.61, 0.79)$$

We are 95% confident that the true proportion of vaccinated children in the county is between 61% and 79%.

One Sample Tests for a Proportion: R commands

Use the prop.test() funciton

- **prop.test**([s], [n], p=[p0], alternative=[alternative], conf.level=[confidence level], correct=FALSE)

- [s]: number of successes
- [n]: sample size
- [p0]: population proportion p
- [alternative]: “two.sided” (default), “greater”, “less”
- [confidence level]: default 0.95
- Correct = FALSE specifies not to use a continuity correction; Default is to apply correction

Continuity correction corrects for the fact that the sampling distribution of a proportion is **not a continuous distribution**. The correction should only be used when a single sample proportion is compared with the population proportion.

It is a very slight alteration of the formula for the test statistics and the confidence interval that makes the test a **little more conservative** (and thus the confidence interval a little wider). Maybe used anytime, but is especially helpful when the sample size is small.

An Example: One Sample Tests for a Proportion

```
> prop.test(70, 100, p=0.8, conf.level=0.95, correct=FALSE)
1-sample proportions test without continuity correction

data:  70 out of 100, null probability 0.8
X-squared = 6.25, df = 1, p-value = 0.01242
alternative hypothesis: true p is not equal to 0.8
95 percent confidence interval:
 0.6041515 0.7810511
sample estimates:
  p
0.7

> prop.test(70, 100, p=0.8, conf.level=0.95, correct=TRUE)
1-sample proportions test with continuity correction

data:  70 out of 100, null probability 0.8
X-squared = 5.6406, df = 1, p-value = 0.01755
alternative hypothesis: true p is not equal to 0.8
95 percent confidence interval:
 0.5989396 0.7854574
sample estimates:
  p
0.7
```

Conditions for Inference

In order for inference based on the test statistic and confidence interval formulas shown above, the following conditions need to be met:

- The sample needs to be randomly drawn from the population
- The population should be at least 10 times the size of the sample. This condition helps to ensure that the standard error of $\hat{p}$ is approximately equal to $\sqrt{\frac{p(1-p)}{n}}$.
- The sample size n is large enough to ensure approximate normality. For significance testing, as a general rule, we use the condition that inference is valid if $n \cdot p_0$ and $n \cdot (1 - p_0)$ are both greater than 10.
- For confidence intervals, we use the general rule that the counts of both successes and failures are both greater than or equal to 15. That is, $n \cdot \hat{p}$ and $n \cdot (1 - \hat{p})$ are both at least 15.

These conditions should always be checked before calculating the confidence interval or before formally testing a proportion using the formulas above.

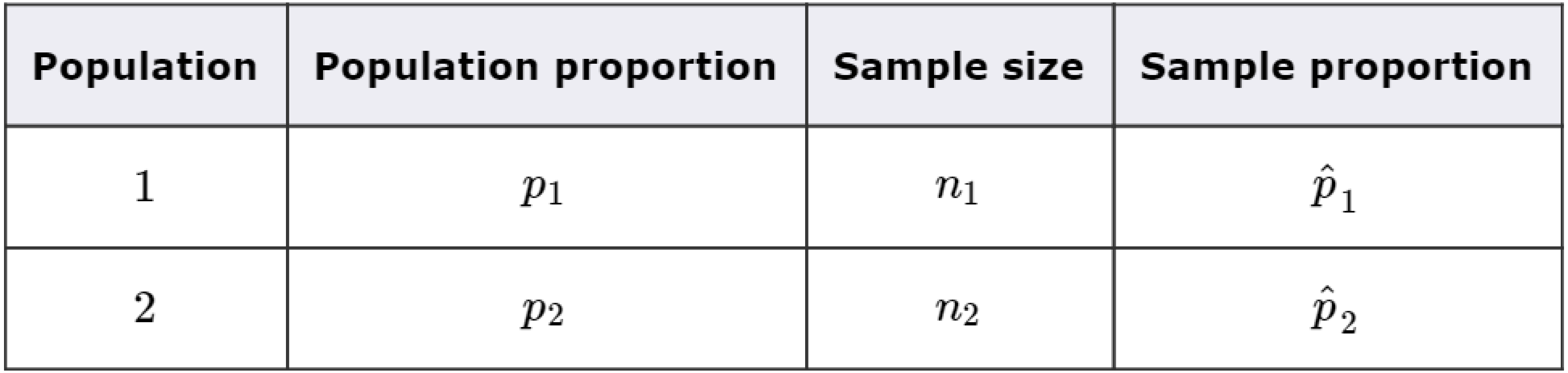
If the sample size is not large enough, then there are some approximations that can be used to estimate the confidence interval and adjust the z-statistic appropriately.

Two-Sample Tests for Proportions

Let us consider situations where we are interested in comparing the **sample proportions from two populations**. We have separate samples from each population and we want to make conclusions about the **characteristics of the two populations**.

Inference in this setting is referred to as two-sample procedures.

The table summarizes the commonly used notation for the population parameters and sample statistics. The populations proportions, p_1 and p_2, are often unknown. We are interested in estimating the quantity p_1-p_2 or testing the null hypothesis of no difference between proportions **$p_1-p_2=0$ or, equivalently, $p_1=p_2$**. Since the population proportions are unknown, they are estimated by the sample proportions. Thus, inference about p_1-p_2 is based on the difference between sample proportions, **p^$_1$−p^$_2$**.



Population	Population proportion	Sample size	Sample proportion
1	p_1	n_1	$\hat{p}_1$
2	p_2	n_2	$\hat{p}_2$

Confidence Intervals for Differences in Proportions

The standard deviation of the difference in the sample proportions, $\hat{p}_1 - \hat{p}_2$, is given by

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

This quantity must be estimated as it is based on the unknown parameters p_1 and p_2. We use the following formula for inference which is referred to as the standard error of the statistic $\hat{p}_1 - \hat{p}_2$

$$\mathrm{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\widehat{p_1}\,(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

The level C confidence interval for the difference in population proportions is given by

$$(\widehat{p_1} - \widehat{p_2}) \pm z \cdot \sqrt{\frac{\widehat{p_1}\,(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

where $\widehat{p_i}$ is the sample proportion from population i and n_i is the number of observations in the sample from population i, and z is the appropriate critical value from the standard normal distribution with area C between −z and z.

This confidence interval should only be used when the populations are at least 10 times as large as the samples and the counts of successes and failures are 10 or more in both samples.

An Example: Two-Sample Tests for Proportions

An investigator is interested in the **long-term effects of preschool programs on low-income children**.

A study was conducted where by two groups of children were followed over time.

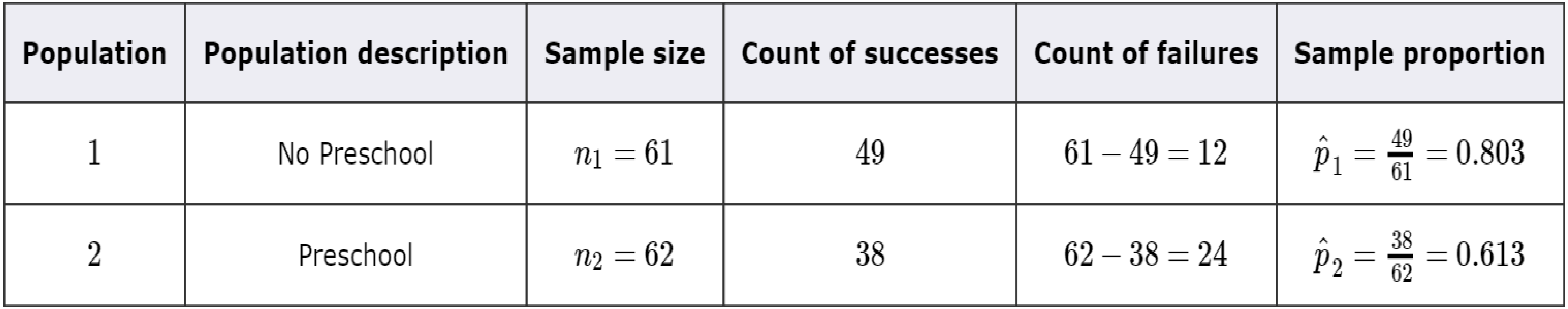
- The first group of 61 children **did not attend preschool**.
- The second group of 62 children (from similar areas and with similar backgrounds of those in the first sample) **attended preschool** as 3- and 4-year-olds.

The need for social programs as adults was the outcome of interest. Of the group who did not attend preschool, 49 of them needed social services (mainly welfare) between the ages of **18 and 30**. In the preschool group, 38 required social services in the same age range.

Calculate the 95% confidence interval for the difference in **proportions of adults requiring social services between those who did not attend preschool versus those who did attend preschool**.

Lets define a “**success**” in this case as requiring social services.

An Example: Two-Sample Tests for Proportions



Population	Population description	Sample size	Count of successes	Count of failures	Sample proportion
1	No Preschool	$n_1 = 61$	49	$61 - 49 = 12$	$\hat{p}_1 = \frac{49}{61} = 0.803$
2	Preschool	$n_2 = 62$	38	$62 - 38 = 24$	$\hat{p}_2 = \frac{38}{62} = 0.613$

In the table, the smallest of these quantities is 12. Since >10, we can proceed with using the formulas specified above.

The level C confidence interval for the difference in population proportions is given by

$$(\widehat{p_1} - \widehat{p_2}) \pm z \cdot \sqrt{\frac{\widehat{p_1}\,(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} = (0.803 - 0.613) \pm 1.960 \cdot \sqrt{\frac{0.803(1-0.803)}{61} + \frac{0.613(1-0.613)}{62}} \approx (0.033, 0.347)$$

Two Sample Tests for Proportions: R commands

Use the prop.test() funciton

- **prop.test([s], [n], alternative=[alternative], conf.level=[confidence level], correct=FALSE)**

- **[s]**: number of successes in each group: **c(s_1, s_2)**
- **[n]**: sample size in each group: **c(n_1, n_2)**
- [alternative]: “two.sided” (default), “greater”, “less”
- [confidence level]: default 0.95
- Correct = FALSE specifies not to use a continuity correction

Continuity correction corrects for the fact that the sampling distribution of a proportion **is not a continuous distribution**. The correction should only be used when a single sample proportion is compared with the population proportion.

It is a very slight alteration of the formula for the test statistics and the confidence interval that makes the test a little more conservative (and thus the confidence interval a little wider). Maybe used anytime, but is especially helpful when the sample size is small.

Two Sample Tests for Proportions: an example

```
> prop.test(c(49, 38), c(61, 62), conf.level=0.95, correct=FALSE)
 2-sample test for equality of proportions without continuity correction

data:  c(49, 38) out of c(61, 62)
X-squared = 5.383, df = 1, p-value = 0.02033
alternative hypothesis: two.sided
95 percent confidence interval:
 0.03336798 0.34738294
sample estimates:
   prop 1    prop 2
0.8032787 0.6129032

> prop.test(c(49, 38), c(61, 62), conf.level=0.95, correct=TRUE)
 2-sample test for equality of proportions with continuity correction

data:  c(49, 38) out of c(61, 62)
X-squared = 4.5027, df = 1, p-value = 0.03384
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01710674 0.36364418
sample estimates:
   prop 1    prop 2
0.8032787 0.6129032
```

Significance Tests for Differences in Proportions

We are interested testing the null hypothesis H_0: $p_1=p_2$. The two-sample z-statistic used for hypothesis testing is calculated by dividing the difference in the sample proportions by the standard error of the difference in sample proportions where the standard error is calculated under the assumption of the null hypothesis:

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\widehat{p_i}$ is the sample proportion from population i and n_i is the number of observations in the sample from population i, and $\hat{p}$ is the pooled sample proportion.

This quantity represents how far the difference in sample proportions is from 0 in standard deviations units under the null hypothesis where the underlying population proportion is the same between the two populations.

This test statistic is approximately normally distributed and the standard normal distribution can be used to calculate the p-values associated with this test.

Inference using this test is valid when the populations are at least 10 times as large as the samples and the counts of successes and failures are 5 or more in both samples.

An Example: Significance Tests for Differences in Proportions

1. Set up the hypotheses and select the alpha level

H_0: $p_1=p_2$ (the underlying population proportion needing adult social services among those who attended preschool is the same as that of those who did not attended preschool)

H_1: $p_1 \neq p_2$ (the underlying population proportion needing adult social services among those who attended preschool is different than that of those who did not attended preschool)

ɑ=0.05

2. Select the appropriate test statistic $z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}$

3. State the decision rule

Determine the appropriate critical value from the standard normal distribution associated with a right hand tail probability of ɑ/2=0.05/2=0.025.

Decision Rule: Reject H_0 if $|z| \geq 1.960$

Otherwise, do not reject H_0

An Example: Significance Tests for Differences in Proportions

4. Compute the test statistic and the associated p-value

$$z = \frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p})\cdot(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0.803-0.613}{\sqrt{\frac{49+38}{61+62}\,(1-\frac{49+38}{61+62})\cdot(\frac{1}{61}+\frac{1}{62})}} = \frac{0.190}{\sqrt{0.707\,(1-0.707)\cdot(\frac{1}{61}+\frac{1}{62})}} = 2.320$$

5. Conclusion

Reject H_0 since |2.320|≥1.960. We have significant evidence at the α=0.05 level that $p_1 \neq p_2$ (p=0.0204). The percentage of adults needing social services was 20% lower among those who attended preschool versus those who did not. We reject the null hypothesis that the underlying population proportions are the same between those attending preschool versus those who did not.

Effect Measures

We focused on the **differences in proportions** (also known as the **risk difference**)

Measures of the differences between risks are called effect measures. Effect measures include the risk difference, the relative risk, and the odds ratio.
The risk difference is simply the absolute difference in risk between two populations.

Estimate of risk difference: $\hat{RD} = \hat{p}_1 - \hat{p}_2$

- A risk difference of 0 indicates that there is no difference in risk between the populations
- A positive risk difference indicates a higher risk in group 1
- A negative risk difference indicates a lower risk in group 1.
- Generally, group 2 as defined here is thought of as the reference group

The relative risk is simply the ratio of the risks in each of the two populations.

Estimate of risk ratio: $\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$

- A relative **risk of 1** indicates that there is no difference in risk between the populations.
- A relative **risk greater than 1** indicates a higher risk in group 1
- A relative **risk less than 1** indicates a lower risk in group 1

Effect Measures

The odds ratio is the ratio of the odds of the outcome in each of the two populations.

The odds of an event is the ratio of the events to the non-events (or it is the ratio of the number of successes to the number of failures).

If among n subjects in a particular group, x had the event, then odds of the event in that group would be $\frac{x}{n-x}$ which is equal to $\frac{\widehat{p_1}}{1-\widehat{p_1}}$. The odds ratio, then is estimated as:

Estimate of odds ratio: $\widehat{OR} = \frac{\frac{\widehat{p_1}}{1-\widehat{p_1}}}{\frac{\widehat{p_2}}{1-\widehat{p_2}}}$

- An odds ratio of 1 indicates that there is no difference in odds between the populations
- An odds ratio greater than 1 indicates a higher odds in group 1
- An odds ratio less than 1 indicates a lower odds in group 1

- Odds ratios are less intuitive to interpret, but have some favorable mathematical properties and as such are used quite often in statistics.
- When the event is rare (when $\widehat{p_1}\ and\ \widehat{p_2}$ are small), the risk ratio and the odds ratios are similar in magnitude.

An example: Estimate of risk difference

Suppose we are interested in the association between gender and risk of having a coronary event in a high-risk patient population (who have had an event in the past). We have data for 25 females and 25 males. We followed each subject for a year to see if they had another coronary event (defined here as a “success”). Calculate the estimated risk difference.

Population	Population description	Sample size	Count of “Successes”	Count of “Failures”	Sample proportion
1	Males	$n_1 = 25$	18	$25 - 18 = 7$	$\hat{p}_1 = \frac{18}{25} = 0.72$
2	Females	$n_2 = 25$	8	$25 - 8 = 17$	$\hat{p}_2 = \frac{8}{25} = 0.32$

Estimate of risk difference: $\widehat{RD} = \widehat{p_1} - \widehat{p_2}$=0.72−0.32=0.40
The risk of a coronary event is 40% higher among males as among females.

An example: Estimate of risk ratio

Estimate of risk ratio: $\widehat{RR} = \frac{\widehat{p_1}}{\widehat{p_2}} = \frac{0.72}{0.32} = 2.25$

The risk of a coronary event is 2.25 times higher among males as among females.

If the reference group had been males,

Estimate of risk ratio: $\widehat{RR} = \frac{\widehat{p_1}}{\widehat{p_2}} = \frac{0.32}{0.72} = 0.44$

The risk of a coronary event among females is less than one half the risk among males.

Sometimes the actual result and the wording of the interpretation help dictate the choice of the reference group. Here, “the risk of a coronary event is 2.25 times higher among males as among females” is easier to understand than “the risk of a coronary event among females is less than one half (RRˆ=0.44) the risk among males.”

An example: Estimate of odds ratio

Estimate of odds ratio:$\widehat{OR} = \frac{\frac{\widehat{p_1}}{1-\widehat{p_1}}}{\frac{\widehat{p_2}}{1-\widehat{p_2}}} = \frac{\frac{0.72}{1-0.72}}{\frac{0.32}{1-0.32}} = 5.46$

The odds of a coronary event is 5.46 times higher among males as among females.

If the reference group had been males,

Estimate of odds ratio: $\widehat{OR} = \frac{\frac{\widehat{p_1}}{1-\widehat{p_1}}}{\frac{\widehat{p_2}}{1-\widehat{p_2}}} = \frac{\frac{0.32}{1-0.32}}{\frac{0.72}{1-0.72}} = 0.18$

The risk of a coronary event among females is less than one half the risk among males.

As noted above, sometimes the actual result and the wording of the interpretation help dictate the choice of the reference group.

“The odds of a coronary event is 5.46 times higher among males as among females” is easier to understand than “The odds of a coronary event among females is less than one fifth the odds of the same event among males.”