# Section 5.3
# Hypothesis testing for a proportion

# Gender discrimination

- In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as "routine".
- The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.
- It was randomly determined which supervisors got "male" applications and which got "female" applications.
- Of the 48 files reviewed, 35 were promoted.
- The study is testing whether females are unfairly discriminated against.

**Q**: *Is this an observational study or an experiment?* *Experiment*

*B.Rosen and T. Jerdee (1974), "Influence of sex role stereotypes on personnel decisions", J.Applied Psychology, 59:9-14.*

# Data

**Q**: *At a first glance, does there appear to be a relationship between promotion and gender?*

| | | *Promotion* | | |
|---|---|---|---|---|
| | | *Promoted* | *Not Promoted* | *Total* |
| *Gender* | *Male* | *21* | *3* | *24* |
| | *Female* | *14* | *10* | *24* |
| | *Total* | *35* | *13* | *48* |

**% of males promoted:** $21/24 = 0.875$
**% of females promoted:** $14/24 = 0.583$

# Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

(a) If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.

(b) Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions. *Maybe*

(c) The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions. *Maybe*

# Two competing claims

1. "There is nothing going on."
   Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance. $\rightarrow$ *Null hypothesis*, denoted $H_o$.
2. "There is something going on."
   Promotion and gender are *dependent*, there is gender discrimination, observed difference in proportions is not due to chance. $\rightarrow$ *Alternative hypothesis*, denoted $H_a$.

# A court trial as a hypothesis test

- Hypothesis testing is very much like a court trial.
- $H_0$: Defendant is innocent
  $H_A$: Defendant is guilty
- We then present the evidence - collect data.
- Then we judge the evidence - "Could these data plausibly have happened by chance if the null hypothesis were true?"
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
- Ultimately we must make a decision. How unlikely is unlikely?

# A court trial as a hypothesis test (cont.)

- If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of "not guilty".
  - The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  - The defendant may, in fact, be innocent, but the jury has no way of being sure.
- Said statistically, we fail to reject the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - Therefore we never "accept the null hypothesis".

# A trial as a hypothesis test (cont.)

- In a trial, the burden of proof is on the prosecution.
- In a hypothesis test, the burden of proof is on the unusual claim.
- The null hypothesis is the ordinary state of affairs (the status quo), so it's the alternative hypothesis that we consider unusual and for which we must gather evidence.

# Simulating the experiment...

... under the assumption of independence, i.e. leave things up to chance: 35 of the 48 managers would have promoted the candidate regardless of gender.

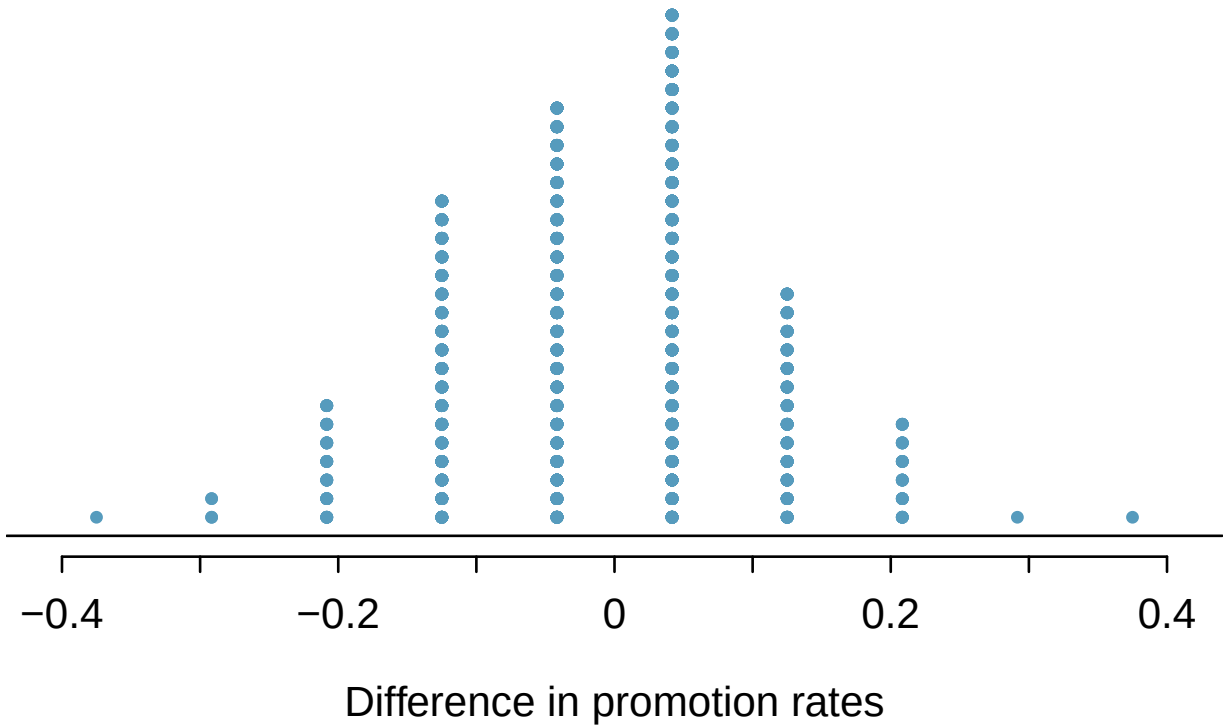
If results from the simulations based on the *chance model* look like the data, then we can determine that the difference between the proportions of promoted files between males and females was simply *due to chance* (promotion and gender are independent).

If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but *due to an actual effect of gender* (promotion and gender are dependent).

# Simulating the experiment

1. Build a deck with 48 cards, 35 of which say "promote" and 13 say "do not promote".
2. Shuffle the cards and deal them intro two groups of size 24, representing males and females.
3. Count and record how many cards in each group say "promoted."
4. Calculate the proportion of promoted cards in each group and take the difference (male - female), and record this value.
5. Repeat steps 2 - 4 many times.

# Simulation Results and Conclusion




Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (male promotions being 30% or more higher than female promotions), we have good reason to reject the null hypothesis in favor of the alternative.

# Recap: hypothesis testing framework

- We start with a *null hypothesis* ($H_0$) that represents the status quo.
- We also have an *alternative hypothesis* ($H_A$) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods (which is the focus of this section).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

# Hypothesis Tests via Theoretical Medthods

In this section we learn methods based on the central limit theorem for conducting an hypothesis test. The key elements:

- State *hypotheses* in terms of the parameter(s) of interest
- Gather good data
- Calculate *test statistic*
- Determine *p-value*
- State conclusion, often in relation to a *significance level.*

# Facebook interest categories

**Q**: *A survey asked 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?*

*https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/*

# Setting the hypotheses

- The *parameter of interest* is the proportion of all American Facebook users who are comfortable with Facebook creating categories of interests for them.
- There may be two explanations why our sample proportion is lower than 0.50 (minority).
  - The true population proportion is different than 0.50.
  - The true population proportion is 0.50, and the difference between the true population proportion and the sample proportion is simply due to natural sampling variability.

# Setting the hypotheses

- We start with the assumption that 50% of American Facebook users are comfortable with Facebook creating categories of interests for them

$$H_0 : p = 0.50$$

- We test the claim that the proportion of American Facebook users who are comfortable with Facebook creating categories of interests for them is different than 50%

$$H_A : p \neq 0.50$$

# Gathering good data

- Respondents in the sample should be independent of each other with respect to whether or not they feel comfortable with their interests being categorized by Facebook.
- Sampling should have been done randomly.
- The sample size should be less than 10% of the population of all American Facebook users.
- There should be at least 10 expected successes and 10 expected failure.

# Test statistic

In order to evaluate if the observed sample proportion is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.

$$\hat{p} \sim N\left(\mu = 0.50, SE = \sqrt{\frac{0.50 \times 0.50}{850}}\right)$$

$$Z = \frac{0.41 - 0.50}{0.0171} = -5.26$$

**Q**: *The sample proportion is 5.26 standard errors away from the hypothesized value. Is this considered unusually low? That is, is the result* statistically significant*?*
*Yes, and we can quantify how unusual it is using a p-value.*

# p-value

- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *low* (lower than the significance level, $\alpha$, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* $H_0$.
- If the p-value is *high* (higher than $\alpha$) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* $H_0$.

# p-value

*p-value:* probability of observing data at least as favorable to $H_A$ as our current data set (a sample proportion lower than 0.41), if in fact $H_0$ were true (the true population proportion was 0.50).

$$\begin{aligned} P(\hat{p} < 0.41 \text{ or } \hat{p} > 0.59) &= P(z < -5.26 \text{ or } z > 5.26) \\ &= 2 * \texttt{pnorm}(-5.26) \\ &< 0.0001 \end{aligned}$$

# Making a decision

- p-value $< 0.0001$
  - If 50% of all American FB users are comfortable with FB creating these interest categories, there is less than a 0.01% chance of observing a random sample of 850 American Facebook users where 41% or fewer or 59% of higher feel comfortable with it.
- Since p-value is low (lower than 5%) we reject $H_0$ (at the 5% significance level).
- The data provide convincing evidence that the proportion of American FB users who are comfortable with FB creating a list of interest categories for them is different than 50%.
- The difference between the null value of 0.50 and observed sample proportion of 0.41 is not due to chance or sampling variability.

# Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

# Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

| | | **Decision** | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | ✓ | Type 1 Error |
| | $H_A$ true | Type 2 Error | ✓ |

- A *Type 1 Error* is rejecting the null hypothesis when $H_0$ is true.
- A *Type 2 Error* is failing to reject the null hypothesis when $H_A$ is true.
- We (almost) never know if $H_0$ or $H_A$ is true, but we need to consider all possibilities.

# Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error
- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"

– William Blackstone

# Type 1 error rate

- As a general rule we reject $H_0$ when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where $H_0$ is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of $\alpha$ – increasing $\alpha$ increases the Type 1 error rate.

# Choosing a significance level

- While the the traditional level is 0.05, it is helpful to adjust the significance level based on the application.
- Select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring $H_A$ before we would reject $H_0$.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject $H_0$ when the null is actually false.

# Testing hypotheses using confidence intervals

**Q**: *Earlier we calculated a 95% confidence interval for the proporton of American Facebook users who think Facebook categorizes their interests accurately as 64% to 67%. Based on this confidence interval, do the data support the hypothesis that a majority of American Facebook users think Facebook categorizes their interests accurately.*

- The associated hypotheses are:
  - $H_0$: $p = 0.50$: 50% of American Facebook users think Facebook categorizes their interests accurately
  - $H_A$: $p > 0.50$: More than 50% of American Facebook users think Facebook categorizes their interests accurately
- Null value is not included in the interval $\rightarrow$ reject the null hypothesis.
- This is a quick-and-dirty approach for hypothesis testing, but it doesn't tell us the likelihood of certain outcomes under the null hypothesis (p-value).

# One vs. two sided hypothesis tests

- In two sided hypothesis tests we are interested in whether $p$ is either above or below some null value $p_0$: $H_A : p \neq p_0$.
- In one sided hypothesis tests we are interested in $p$ differing from the null value $p_0$ in one direction (and not the other):
  - If there is only value in detecting if population parameter is less than $p_0$, then $H_A : p < p_0$.
  - If there is only value in detecting if population parameter is greater than $p_0$, then $H_A : p > p_0$.
- Two-sided tests are often more appropriate as we often want to detect if the data goes clearly in the opposite direction of our alternative hypothesis as well.

# Hypothesis testing for a single proportion

Once you've determined a one-proportion hypothesis test is the correct procedure, there are four steps to completing the test:

**Prepare.** Identify the parameter of interest, list hypotheses, identify the significance level, and identify $\hat{p}$ and $n$.

**Check.** Verify conditions to ensure $\hat{p}$ is nearly normal under $H_0$. For one-proportion hypothesis tests, use the null value to check the success-failure condition.

**Calculate.** If the conditions hold, compute the standard error, again using $p_0$, compute the Z-score, and identify the p-value.

**Conclude.** Evaluate the hypothesis test by comparing the p-value to $\alpha$, and provide a conclusion in the context of the problem.