Hypothesis Testing
A Motivating Example

- Between 1960 and 1980, there were many lawsuits in the South claiming racial bias in jury selection.
- Here’s some made up* (but similar) supporting data:
  - 50% of citizens in the local area are African American
  - On an 80 person panel, only 4 were African American
- Can this outcome be explained as the result of pure chance?
- Unlikely: If $X \sim B(80, .5)$, then $P[X \leq 4] = 1.8 \times 10^{-18}$
- \textit{N.B.:} Statistics can never \textit{prove} anything!

---

*This example was borrowed from \textit{The Cartoon Guide to Statistics}.\]
Hypothesis Testing

- A **hypothesis test** determines whether an observed distribution of data matches the expected distribution of data.
  - E.g. is a bag of M&Ms distributed as expected?
- A **test statistic** measures how well observed data match the data that would be expected under an assumed probability model (called the **null hypothesis**).
Basics of Hypothesis Testing

• Step 1: Formulate a null and alternative hypothesis
  ○ The null hypothesis is a claim that data are distributed in some way (e.g., $B(80, 0.5)$)
  ○ An alternative hypothesis is a claim that data are distributed in some other way (e.g., $p < 0.5$)
  ○ The null is so-called because it is usually a claim about no significant effect or difference, and it is often something we suspect the data will disprove

• Step 2: Compute a test statistic, assuming the null hypothesis
  ○ A test statistic measures the extent to which an observed sample of data agrees with an assumed probability model (i.e., the distribution under the null hypothesis).

• Step 3: Find the $p$-value corresponding to the value of the test statistic
  ○ What is the probability of observing this value of the test statistic?
A Worked Example

- Numerator: $p_{\text{hat}} - p = (4 / 80) - 0.5 = 0.05 - 0.5 = -0.45$
  - By subtracting $p$, we are assuming that the null hypothesis is true.

- Denominator: Standard Error
  - $\text{Var}[p_{\text{hat}}] = (0.5)(1 - 0.5) / 80 = 0.003125$
  - The standard error is the square root of this variance: $\sqrt{0.003125} = 0.056$

- $z$-statistic: $-0.45 / 0.056 \approx -8.0$
  - That’s a whole lot of standard deviations below the mean!

- If the null hypothesis were true, the probability of observing this value of our test statistic is essentially 0.
- We reject the null hypothesis and search for alternative explanations.
Who was expected to be the US president?

- YouGov surveyed 1300 people in the United States, asking them: “Who will you vote for in the election for President in November?”
- The poll was conducted on October 7th and 8th, 2016.
  - 44% of people planned to vote for Clinton.
  - 38% of people planned to vote for Trump.
Step 1: Formulate the Hypotheses

- Our null hypothesis is that the proportion of people that plan to vote for Trump is identical to the proportion of people that plan to vote for Clinton.
  - $p_H = p_T$: i.e., $p_{\text{NULL}} = 0$

- Our alternative hypothesis is that the proportion of people who plan to vote for Clinton is higher than the proportion who plan to vote for Trump.
  - $p_H > p_T$
Step 2: Calculate the Test Statistic

- **z-statistic for the difference between two sample proportions**
- **Numerator:** \( (p_H - p_T) - p_{\text{NULL}} \) = \( (.44 - .38) - 0 = .06 \)
  - By subtracting 0, we are assuming the null hypothesis (i.e., it is our baseline).
- **Denominator:** Standard Error
  - \( (.44)(1300) = 572 \) people preferred Clinton
  - \( (.38)(1300) = 494 \) people preferred Trump
  - \( \text{Var}[p_H - p_T] = (0.44)(1 - 0.44) / 572 + (0.38)(1 - 0.38) / 494 = 0.0009 \)
  - The standard error is the square root of this variance: \( \sqrt{0.0009} = 0.0301 \)
- **z-statistic:** \( 0.06 / 0.0301 = 1.99 \)
Step 3: Calculate the \( p \)-value

- What is the probability of observing this value of the test statistic?
  - \( \text{pnorm}(1.99, \ \text{lower.tail} = \text{FALSE}) = 0.023 \)

- Under the null hypothesis, there was only a 2.3\% chance we would see data that favor Clinton as much or more than what we saw in the YouGov poll.

- This means one of two things:
  - We witnessed something incredibly rare.
  - The assumption that the null hypothesis is true is incorrect.
Steps 0 and 4: Significance Testing

- Typically, we set a benchmark threshold (α-level) before we run the test.
- Often, the threshold is 5% (corresponding to a 95% confidence interval).
- If the p-value is below this threshold, then we reject the null hypothesis, and search for alternative explanations.
All the Steps in Hypothesis Testing

- Step 0: Set a significance level ($\alpha$)
- Step 1: Formulate a null and alternative hypothesis
- Step 2: Compute a test statistic, assuming the null hypothesis
- Step 3: Find the $p$-value corresponding to the value of the test statistic
- Step 4: Perform a significance test at the $\alpha$-level
Errors
“Innocent until proven guilty”

- Hypothesis testing is a statistical implementation of this maxim
- Null hypothesis: the defendant is innocent
- Alternative hypothesis: the defendant is guilty
  - A type 1 error (false positive) occurs when we put an innocent person in jail
  - A type 2 error (false negative) occurs when we do not jail a guilty person
- Another example:
  - Type 1 error: false alarm (fire alarm when there is no fire)
  - Type 2 error: fire but no fire alarm
- In sum:
  - Type 1 error: we reject the null when we should not
  - Type 2 error: we accept the null when we should not
Type I vs. Type II Errors

- **Cancer screening**
  - Null: no cancer
  - Type I: cancer suspected where there is none (not good, but not terrible)
  - Type II: cancer goes undetected (very very bad)

- **Err on the side of type I errors**
  - Make it easy to reject the null, even when we should not
  - Choose lower significance level
Type I vs. Type II Errors

● Spam Filters
  ○ Null: an email is legitimate
  ○ Type I: filter a legitimate email (could be very bad)
  ○ Type II: don’t filter spam (not so bad)

● Err on the side of type II errors
  ○ Make it hard to reject the null, even when we should
  ○ Choose higher significance level
Type I vs. Type II Errors

- **Suspected terrorists**
  - Null hypothesis: person is not a terrorist
  - Type I: send an innocent person to Guantánamo Bay
  - Type II: let a terrorist (who intends to commit mass murder) free

- **US is erring on the side of type I errors, which explains why people are often held at Guantánamo Bay without a fair trial**
Degrees of Freedom
Degrees of Freedom

- **Physics**: number of directions in which independent motion can occur
  - Elbows have one degree of freedom
  - Shoulders and wrists have three

- **Chemistry**: number of independent factors required to describe equilibrium

- **Statistics**: number of values in a calculation that are free to vary
  - Often, one less than the number of observations
Chi-Square Distribution
Chi-square Hypothesis Testing

- A total of 1453 people reported for jury duty in Alameda County in Northern California between 2009 and 2010.

<table>
<thead>
<tr>
<th></th>
<th>Asian</th>
<th>Black</th>
<th>Latinx</th>
<th>White</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>15%</td>
<td>18%</td>
<td>12%</td>
<td>54%</td>
<td>1%</td>
</tr>
<tr>
<td>Jurors</td>
<td>26%</td>
<td>8%</td>
<td>8%</td>
<td>54%</td>
<td>4%</td>
</tr>
</tbody>
</table>

- Were juries representative of the population from which they were drawn?

- **Null hypothesis**: Jurors are distributed according to a multinomial with probabilities consistent with the population.
# EDA

<table>
<thead>
<tr>
<th></th>
<th>Asian</th>
<th>Black</th>
<th>Latinx</th>
<th>White</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>15%</td>
<td>18%</td>
<td>12%</td>
<td>54%</td>
<td>1%</td>
</tr>
<tr>
<td>Jurors</td>
<td>26%</td>
<td>8%</td>
<td>8%</td>
<td>54%</td>
<td>4%</td>
</tr>
</tbody>
</table>

![Graph showing population and jurors by ethnicity](Image Source)
Chi-square Test Statistic

The value of the test-statistic is

\[ \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i / N - p_i)^2}{p_i} \]

where

- \( \chi^2 \) Pearson's chi-squared test statistic
- \( O_i \) the number of observations of type \( i \)
- \( N \) the total number of observations
- \( E_i = Np_i \) the expected (theoretical) frequency of type \( i \), asserted by the null hypothesis that the proportion of type \( i \) in the population is \( p_i \)
- \( n \) the number of cells in the table.
Chi-square Test Statistic, cont’d

- For Asians: \((26\% - 15\%)^2 / 15\% = 0.0807\)
- For Blacks: \((8\% - 18\%)^2 / 18\% = 0.0556\)
- For Latinx: \((8\% - 12\%)^2 / 12\% = 0.0133\)
- For Whites: \((54\% - 54\%)^2 / 54\% = 0\)
- For Others: \((1\% - 4\%)^2 / 4\% = 0.0225\)

The chi-square test statistic is thus:
\[1453(0.0807 + 0.0556 + 0.0133 + 0.0225) = 250\]
Chi-square Distribution

- The distribution of the sum of the squares of $k$ independent standard normal random variables
- The chi-square distribution is parameterized by degrees of freedom
Conclusion

- Choose $\alpha = 95\%$. Since there are 5 races, there are 4 degrees of freedom.
  - $qchisq(.95, \text{df} = 4)$
    
    \[
    [1] 9.487729
    \]
- Since $250 > 9.487729$, we reject the null hypothesis
- Likewise, the $p$-value is essentially 0:
  - $pchisq(250, \text{df} = 4, \text{lower.tail} = \text{FALSE}) = 6.50969e-53$

- We reject the null hypothesis: juries were not racially representative in Alameda County in 2009 and 2010.