Degrees of Freedom
Degrees of Freedom

- Physics: number of directions in which independent motion can occur
  - Example: elbows have one degree of freedom
- 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
Student $t$-Distribution
When does the CLT kick in?

Sampling Distribution of the Mean with $n = 10$

Sampling Distribution of the Mean with $n = 30$

Sampling Distribution of the Mean with $n = 100$

Sampling Distribution of the Mean with $n = 1000$
What is a large enough sample size?

- The distribution looks roughly normal with a sample size of 30.
- So, as a rule of thumb, people often say that 30 is a large enough sample size for the central limit theorem to apply.
- However, the histogram becomes more and more bell-shaped as the sample size increases.
- So all things being equal, larger sample sizes are always better than smaller ones!
What if the sample size is not large enough?

- We can extend our methodologies by introducing a new distribution, called the *(Student)* $t$-distribution, which approximates the normal.
  - The student in question was one William S. Gosset.
  - He discovered this (family of) distribution(s) in 1908, while employed as a statistician by the Guinness brewing company, who forbid him from publishing under his own name.
  - He wrote under the pen name “Student” instead.

- The $t$-distribution allows us to perform statistical inference even when the sample size is not large enough to apply the central limit theorem.
The \( t \)-Distribution

- The \( t \)-distribution has 1 parameter: the **degrees of freedom**.
- As \( n \) goes to infinity, the \( t \)-distribution converges to the normal.
- Thus, using the \( t \)-distribution when the sample size is small is consistent with using the normal distribution when the sample size is large.
Building Confidence Intervals

- The only difference between building confidence intervals using the
  $t$-distribution and building them using the normal is: the critical values for
  the normal distribution are familiar numbers to most statisticians. Recall:
  
  - Find $z_{lo}$ and $z_{hi}$ s.t. $P[z_{lo} \leq Z \leq z_{hi}] = 1 - \alpha$
    - If $1 - \alpha = 90\%$, then $|z_{a/2}| = z_{1-\alpha/2} = 1.645$
    - If $1 - \alpha = 95\%$, then $|z_{a/2}| = z_{1-\alpha/2} = 1.96$
    - If $1 - \alpha = 98\%$, then $|z_{a/2}| = z_{1-\alpha/2} = 2.33$
    - If $1 - \alpha = 99\%$, then $|z_{a/2}| = z_{1-\alpha/2} = 2.58$
  - $z_{lo} = z_{a/2}$ & $z_{hi} = z_{1-\alpha/2}$ are called critical values.

- For the $t$-distribution, we have to look these values up.
  
  - The magnitude of these values for $\alpha = .05$, for example, is higher than 1.96.
  - There is more uncertainty because of the smaller sample size.
An Example

- Let’s say we poll 30 people to see who they want to win the presidency, and 60% of people say Clinton.
- We build a 95% confidence interval as follows:
  - The estimate is 0.6.
  - The $SE = \sqrt{(.6)(.4)/30} = 0.09$.
  - The critical $t$-values are found using R.
    - $qt(0.025, 29) = -2.05$ and $qt(0.975, 29) = 2.05$
- The 95% CI is $(0.6 - (2.05)(0.09), 0.6 + (2.05)(0.09)) = (0.42, 0.78)$
Chi-Square Distribution
Chi-square Hypothesis Testing

- This test tries to determine whether the observed distribution of categorical data matches the expected distribution of data.
- Example: 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>1%</td>
</tr>
</tbody>
</table>

- Were juries representative of the populations from which they were drawn?
### 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>1%</td>
</tr>
</tbody>
</table>

![Bar chart showing ethnicity distribution](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: \((1453) (26\% - 15\%)^2 / 15\% = 117\)
- For Blacks: \((1453) (8\% - 18\%)^2 / 18\% = 81\)
- For Latinx: \((1453) (8\% - 12\%)^2 / 12\% = 19\)
- For Whites: \((1453) (54\% - 54\%)^2 / 54\% = 0\)
- For Others: \((1453) (1\% - 4\%)^2 / 1\% = 131\)

- The chi-square test statistic is thus \(117 + 81 + 19 + 0 + 131 = 348\)
Chi-square Distribution

- The distribution of the sum of the squares of $k$ independent standard normal random variables
- Like the student $t$-distribution, 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.
  - $\text{qchisq}(.95, \ df = 4)$
    
    \[
    [1] \ 9.487729
    \]

- Since $348 > 9.487729$, we reject the null hypothesis

- Likewise, the $p$-value is essentially 0:
  - $\text{pchisq}(348, \ df = 4, \ lower\_tail = \text{FALSE}) = 4.740217\times10^{-74}$

- We can reject the null hypothesis in favor of the alternative.
- Juries were not selected to be racially representative in Alameda County.