Comparing population means
  - t-test for two samples
  - ANOVA (F-test)
Comparing population variances
  - F-test
  - Levene test (F-test)
Today’s lecture

Categorical data analysis

- Contingency Tables
- Chi-square test
  - Goodness-of-fit test
  - Test of homogeneity
  - Test of independence
- Fisher’s exact test
Chi-squared theoretical justification I

Let \( X_1 \sim \text{Binomial}(n, p_1) \).

As \( n \to \infty \), \( \frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}} \to Z \), where \( Z \sim N(0, 1) \).

\[
Q = \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} \to Z^2, \text{ where } Z^2 \sim \chi^2_1.
\]

Let \( X_2 = n - X_1 \) and \( p_2 = 1 - p_1 \).

\[
Q = \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1 - p_1)}
\]
\[
= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}
\]

\( Q \) has an approximate \( \chi^2_1 \) with 1 d.f.

- \( X_1, X_2 \): observed values
- \( np_1, np_2 \): expected values
Let $(X_1, \ldots, X_k) \sim Multinomial(n, p_1, \ldots, p_k)$, where $\sum_{i=1}^{k} X_i = n$ and $\sum_{i=1}^{k} p_i = 1$. Then, $Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ has an approximate $\chi^2$ with d.f. $k - 1$

Note: a binomial distribution is a special case of a multinomial distribution with $k = 2$. 
Goodness-of-fit test

The goodness-of-fit test is a way of determining whether a set of categorical data came from a claimed discrete distribution or not. It answers the question: are the frequencies I observe for my categorical variable consistent with my theory? The goodness-of-fit test is used if you have two or more categories.

<table>
<thead>
<tr>
<th>Yellow</th>
<th>Green</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>416</td>
<td>140</td>
<td>556</td>
</tr>
</tbody>
</table>

- **$H_0$:** The ratio of Yellow:Green is 3:1.
- **$H_1$:** The ratio of Yellow:Green is NOT 3:1.
Goodness-of-fit test

Let \((X_1, \ldots, X_k) \sim Multinomial(n, p_1, \ldots, p_k)\). We want to test for

\[ H_0 : p_1 = p_1^0, \ldots, p_k = p_k^0 \text{ vs. } H_1 : \text{at least one } p_i \neq p_i^0 \]

- Contingency table: \((O_i = X_i \text{ and } E_i = np_i^0)\)

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>\cdots</th>
<th>k</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>(O_1)</td>
<td>\cdots</td>
<td>(O_k)</td>
<td>(n)</td>
</tr>
<tr>
<td>Expected</td>
<td>(E_1)</td>
<td>\cdots</td>
<td>(E_k)</td>
<td>(n)</td>
</tr>
</tbody>
</table>

- Test statistic: \(Q = \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}\) under \(H_0\)

- Rejection region: \(Q > \chi^2_{k-1,1-\alpha}\)

- p-value: \(p = \Pr(Q > Q(\text{obs}))\)
Goodness-of-fit test

- $H_0: p_1 = 3/4, p_2 = 1/4$
- $H_1: p_1 \neq 3/4, p_2 \neq 1/4$

<table>
<thead>
<tr>
<th>Cate.</th>
<th>Yellow</th>
<th>Green</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs.</td>
<td>416</td>
<td>140</td>
<td>556</td>
</tr>
<tr>
<td>Exp.</td>
<td>417(= 556 × 3/4)</td>
<td>139</td>
<td>556</td>
</tr>
</tbody>
</table>

Test statistic:

$$\frac{(416 - 417)^2}{417} + \frac{(140 - 139)^2}{139}$$

Chi-squared test for given probabilities data: obs

$X$-squared = 0.0095923, df = 1, p-value = 0.922
Test of homogeneity: multiple samples

Example: A group of 10,000 current smoking women ages 50-69, 40,000 never smoking women, and 50,000 ex-smoking women who previously smoked but do not currently smoke are followed for 5 years. Fifty of the current smokers, 10 of the never smokers and 25 of ex-smokers developed lung cancer over 5 years. We want to test whether the probabilities of having lung cancer (prevalence) of current smoker, never smoker, and ex-smoker are the same or not.

\[ H_0: \Pr(\text{lung cancer}|\text{current smoker}) = \Pr(\text{lung cancer}|\text{never smoker}) = \Pr(\text{lung cancer}|\text{ex-smoker}) \]

vs. \[ H_1: \text{at least one of such probability is different.} \]
Test of homogeneity

The **test of homogeneity** is a way of determining whether two or more sub-groups of a population share the same distribution of a single categorical variable.

Consider $r$ samples of $c$-categorical variable:

\[
X_{1,1}, \ldots, X_{1,c} \sim Multinomial(n_1, p_{1,1}, \ldots, p_{1,c}) \\
\vdots \\
X_{r,1}, \ldots, X_{r,c} \sim Multinomial(n_r, p_{r,1}, \ldots, p_{r,c})
\]

We want to test

\[
p_{1,1} = \ldots = p_{r,1} = p_{1}^0 \\
H_0: \quad \ldots \quad vs. \quad H_1: \quad not \quad H_0 \\
p_{1,c} = \ldots = p_{r,c} = p_{c}^0
\]
Test of homogeneity

Observation table ($r \times c$ contingency table)

<table>
<thead>
<tr>
<th>Cate.</th>
<th>1</th>
<th>\cdots</th>
<th>c</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grp 1</td>
<td>$O_{1,1}$</td>
<td>\cdots</td>
<td>$O_{1,c}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td></td>
<td>\vdots</td>
<td></td>
<td>\vdots</td>
<td></td>
</tr>
<tr>
<td>Grp r</td>
<td>$O_{r,1}$</td>
<td>\cdots</td>
<td>$O_{r,c}$</td>
<td>$n_r$</td>
</tr>
<tr>
<td>total</td>
<td>$m_1$</td>
<td>\cdots</td>
<td>$m_c$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Under $H_0: p_{1,j} = \ldots, p_{r,j} = p_j^0$ for all $j$, the estimate $\hat{p}_j^0 = \frac{m_j}{n}$

Under $H_0$, the expected value $E_{i,j} = n_i \hat{p}_i^0 = n_i \frac{m_j}{n}$ and the expected table:

<table>
<thead>
<tr>
<th>Cate.</th>
<th>1</th>
<th>\cdots</th>
<th>c</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grp 1</td>
<td>$E_{1,1}$</td>
<td>\cdots</td>
<td>$E_{1,c}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td></td>
<td>\vdots</td>
<td></td>
<td>\vdots</td>
<td></td>
</tr>
<tr>
<td>Grp r</td>
<td>$E_{r,1}$</td>
<td>\cdots</td>
<td>$E_{r,c}$</td>
<td>$n_r$</td>
</tr>
</tbody>
</table>
Test of homogeneity

\[ H_0 : p_{1,j} = \ldots, p_{r,j} = p_j^0 \text{ for all } j \]

- Test statistic: \[ Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \]

- Under \( H_0 \): \( Q \sim \chi^2_{(r-1)(c-1)} \)
  - Each group: \( \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \) has \((c - 1)\) d.f.
  - The last row of the observed contingency table is fully determined when \( \hat{p}_j^0 \) is given.

- Rejection region: \( Q > \chi^2_{(r-1)(c-1),1-\alpha} \)
- \( p \)-value: \( p = \Pr(Q > Q(\text{obs})) \)
Test of homogeneity: lung cancer

<table>
<thead>
<tr>
<th></th>
<th>lc</th>
<th>nc</th>
<th>lc</th>
<th>nc</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>50</td>
<td>9950</td>
<td>cs</td>
<td>8.5</td>
</tr>
<tr>
<td>ns</td>
<td>10</td>
<td>39990</td>
<td>ns</td>
<td>34.0</td>
</tr>
<tr>
<td>xs</td>
<td>25</td>
<td>49975</td>
<td>xs</td>
<td>42.5</td>
</tr>
</tbody>
</table>

Proportion of women with lung case in each group

<table>
<thead>
<tr>
<th></th>
<th>cs</th>
<th>ns</th>
<th>xs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00500</td>
<td>0.00025</td>
<td>0.00050</td>
</tr>
</tbody>
</table>

Pearson’s Chi-squared test

$X^2 = 226.96$, $df = 2$, $p$-value $< 2.2e-16$
Example. We are interested in the relationship between lung cancer incidence and heavy drinking (defined as $\geq 2$ drinks per day). We followed 30,000 nonsmokers for 10 years to determine cancer endpoints. The table below shows the observed results to examine the relationship between lung cancer and initial drinking habit:

<table>
<thead>
<tr>
<th></th>
<th>Lung caner</th>
<th>no lung caner</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy drinker</td>
<td>9</td>
<td>891</td>
<td>900</td>
</tr>
<tr>
<td>Nondrinker</td>
<td>30</td>
<td>2079</td>
<td>2100</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>2970</td>
<td>30000</td>
</tr>
</tbody>
</table>

We want to test whether two categorical variables (lung cancer incidence, drinking habit) are related or not.

$H_0$ : lung cancer incidence and drinking habit are independent vs. $H_0$ : lung cancer incidence and drinking habit are dependent
The **test of independence** is a way of determining whether two categorical variables are associated with one another in the population.

Let \( X_{1,1}, \ldots, X_{1,c}, \ldots, X_{r,1}, \ldots, X_{r,c} \sim \textit{Multinomial}(n, p_{1,1}, \ldots, p_{r,c}) \) and be classified by two attributes (e.g., lung cancer incidence and drinking habit).

### Observation table

<table>
<thead>
<tr>
<th>var.</th>
<th>( A_1 )</th>
<th>( \ldots )</th>
<th>( A_c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_1 )</td>
<td>( O_{1,1} )</td>
<td>( \ldots )</td>
<td>( O_{1,c} )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>( B_r )</td>
<td>( O_{r,1} )</td>
<td>( \ldots )</td>
<td>( O_{r,c} )</td>
<td>( n_r )</td>
</tr>
<tr>
<td>total</td>
<td>( m_1 )</td>
<td>( \ldots )</td>
<td>( m_c )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

### Probabilities

<table>
<thead>
<tr>
<th>var.</th>
<th>( A_1 )</th>
<th>( \ldots )</th>
<th>( A_c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_1 )</td>
<td>( p_{1,1} )</td>
<td>( \ldots )</td>
<td>( p_{1,c} )</td>
<td>( p_1 )</td>
</tr>
<tr>
<td>( B_r )</td>
<td>( p_{r,1} )</td>
<td>( \ldots )</td>
<td>( p_{r,c} )</td>
<td>( p_r )</td>
</tr>
<tr>
<td>total</td>
<td>( p_1 )</td>
<td>( \ldots )</td>
<td>( p_c )</td>
<td>1</td>
</tr>
</tbody>
</table>

**\( H_0 \):** Two categorical variables are independent.

That is, \( p_{i,j} = p_i.p_j \) for all \( i, j \).
Test of independence

\( H_0: \ p_{i,j} = p_i \cdot p_j \) for all \( i, j \) vs. \( H_1: \) at least one \( p_{i,j} \neq p_i \cdot p_j \)

Under \( H_0 \), \( \hat{p}_i = \frac{m_i}{n} \), \( \hat{p}_j = \frac{n_j}{n} \), and \( \hat{p}_{i,j} = \hat{p}_i \cdot \hat{p}_j = \frac{m_i n_j}{n^2} \).

Therefore, the expected value \( E_{i,j} = n \frac{m_i n_j}{n^2} = \frac{m_i n_j}{n} \)

Observation table

<table>
<thead>
<tr>
<th>var.</th>
<th>( A_1 )</th>
<th>( \cdots )</th>
<th>( A_c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_1 )</td>
<td>( O_{1,1} )</td>
<td>( \cdots )</td>
<td>( O_{1,c} )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>( B_r )</td>
<td>( O_{r,1} )</td>
<td>( \cdots )</td>
<td>( O_{r,c} )</td>
<td>( n_r )</td>
</tr>
<tr>
<td>total</td>
<td>( m_1 )</td>
<td>( \cdots )</td>
<td>( m_c )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

Expected table

<table>
<thead>
<tr>
<th>var.</th>
<th>( A_1 )</th>
<th>( \cdots )</th>
<th>( A_c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_1 )</td>
<td>( E_{1,1} )</td>
<td>( \cdots )</td>
<td>( E_{1,c} )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>( B_r )</td>
<td>( E_{r,1} )</td>
<td>( \cdots )</td>
<td>( E_{r,c} )</td>
<td>( n_r )</td>
</tr>
<tr>
<td>total</td>
<td>( m_1 )</td>
<td>( \cdots )</td>
<td>( m_c )</td>
<td>( n )</td>
</tr>
</tbody>
</table>
Test of independence

$H_0$: $p_{i,j} = p_i p_j$ for all $i, j$ vs. $H_1$: at least one $p_{i,j} \neq p_i p_j$

- Test statistic: $Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

- Under $H_0$: $Q \sim \chi^2_{(r-1)(c-1)}$
  - d.f.: $rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1)$

- Rejection region: $Q > \chi^2_{(r-1)(c-1), 1-\alpha}$

- p-value: $p = \Pr(Q > Q_{\text{obs}})$
Test of independence: lung cancer and drinking habit

observed
  cancer nc
hd  9   891
nd 30  2079

Expected table
  cancer nc
hd 11.665  888.335
nd 27.335  2081.665

Pearson’s Chi-squared test
X-squared = 0.88008, df = 1, p-value = 0.3482
Test of independence vs. Test of homogeneity

- Same test statistic, same null distribution
- However, they have different hypotheses and designed for different data sets!
  - In the test of independence, observational units are collected at random from a population and two categorical variables are observed for each unit. (e.g., 30,000 non-smoker were studied, but the frequency of each category of two variables is random.)
  - In the test of homogeneity, the data are collected by randomly sampling from each sub-group separately. (e.g., 10,000 smoker, 40,000 never-smoker and 50,000 ex-smoker)
  - The difference between these two tests is subtle yet important.
Yates-Corrected Chi-Square Test for a $2 \times 2$ Contingency Table

$\chi^2$-test:
- normal approximation
- introducing some errors

Yates-Corrected Chi-Square test:
- To reduce the error in approximation
- continuity correction by subtracting 0.5 from the difference between each observed value and its expected value

Test statistic: $Q = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|O_{i,j} - E_{i,j}| - 0.5)^2}{E_{i,j}}$

Generally, p-values obtained using the continuity correction are slightly larger and thus are slightly less significant than comparable results.
Chi-square tests

Use chi-square tests only if both of the following two conditions are satisfied

- No more than 20% of the cells have expected values < 5
- No cell has an expected value < 1

For a $2 \times 2$ contingency table, use chi-square tests when

- the expected values of all cells are $\geq 5$
Fisher’s Exact Test: small-sample test of independence

<table>
<thead>
<tr>
<th></th>
<th>$A_1$</th>
<th>$A_2$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>$O_{1,1}$</td>
<td>$O_{1,2}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$O_{2,1}$</td>
<td>$O_{2,2}$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>total</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Under $H_0$: $p_{i,j} = p_i p_j$ for all $i, j$,

$$
\Pr(O_{1,1}|n, n_1, m_1) = \frac{\binom{n_1}{O_{1,1}} \binom{n - n_1}{m_1 - O_{1,1}}}{\binom{n}{m_1}} = \frac{\binom{n_1}{O_{1,1}} \binom{n_2}{O_{2,1}}}{\binom{n}{m_1}}
$$

That is, $O_{1,1} \sim Hypergeometric(n, m_1, n_1)$

**Recap: Hypergeometric distribution**

- a finite population of $n$ size, containing $n_1$ successes
- $m_1$ draws without replacement
- the probability of $O_{1,1}$ successes out of $m_1$ draws
Fisher’s Exact Test: example

For example, we observed: \( \Pr(O_{1,1} = 3 \mid 8, 4, 4) = 0.2285 \)

<table>
<thead>
<tr>
<th></th>
<th>( A_1 )</th>
<th>( A_2 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_1 )</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>( B_2 )</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>total</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

\[
P = 0.0142 \quad \begin{array}{ccc} \hline \( B_1 \) & \( A_1 \) & \( A_2 \) & \text{Total} \\
0 & 4 & 4 & 4 \\
4 & 0 & 4 & 4 \\
\hline \text{total} & 4 & 4 & 8 \\
\end{array}
\]

\[
P = 0.5142 \quad \begin{array}{ccc} \hline \( B_1 \) & \( A_1 \) & \( A_2 \) & \text{Total} \\
2 & 2 & 4 & 4 \\
2 & 2 & 4 & 4 \\
\hline \text{total} & 4 & 4 & 8 \\
\end{array}
\]

\[
P = 0.2285 \quad \begin{array}{ccc} \hline \( B_1 \) & \( A_1 \) & \( A_2 \) & \text{Total} \\
1 & 3 & 4 & 4 \\
3 & 1 & 4 & 4 \\
\hline \text{total} & 4 & 4 & 8 \\
\end{array}
\]

\[
P = 0.0142 \quad \begin{array}{ccc} \hline \( B_1 \) & \( A_1 \) & \( A_2 \) & \text{Total} \\
4 & 0 & 4 & 4 \\
0 & 4 & 4 & 4 \\
\hline \text{total} & 4 & 4 & 8 \\
\end{array}
\]
Fisher’s Exact Test

<table>
<thead>
<tr>
<th></th>
<th>$A_1$</th>
<th>$A_2$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>$O_{1,1}$</td>
<td>$O_{1,2}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$O_{2,1}$</td>
<td>$O_{2,2}$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>total</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

- $H_0$: $p_{i,j} = p_{i}.p_{.j}$ for all $i, j$ vs. $H_1$: at least one $p_{i,j} \neq p_{i}.p_{.j}$
- Test statistic: $O_{1,1}$
- $O_{1,1}|n, m_1, n_1 \sim \text{Hypergeometric}(n, m_1, n_1)$
- p-value: $2 \times \min\{\Pr(O_{1,1} \geq O(\text{obs})), \Pr(O_{1,1} \leq O(\text{obs}))\}$

- For $r \times c$ contingency table, Fisher’s exact test can be applied based on multiple hypergeometric distribution.
Fisher’s Exact Test: example

For example, we observed: \( \Pr(O_{1,1} = 3 \mid 8, 4, 4) = 0.2285 \)

<table>
<thead>
<tr>
<th></th>
<th>( A_1 )</th>
<th>( A_2 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_1 )</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>( B_2 )</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>total</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

p-value: \( 2 \times (\Pr(O_{1,1} = 3) + \Pr(O_{1,1} = 4)) = 0.4857 \)
Summary and Future lectures

Compare population means of group
- to study the change in population mean according to the change in the level of groups/categories

Compare the distribution of categorical data of group/another categorical variable
- to study the change in the distribution of the categorical data according to the change in the level of groups or another categorical variable

Association of a categorical data
- with another categorical data

Study the change in population mean of a variable with the change in the value of another continuous variable
- Regression

Association of a continuous variable with another continuous data
- Correlation