$X^2$ Tests

$X^2$ tests aim at testing values of parameters / independence of categorical variables (as opposed to continuous variables and T-tests you saw in Math 181A for quantitative ideas.)

Roughly speaking, a variable is "categorical" (or quantitative) if it has a discrete (or even finite) set of possible outcomes.

Examples where $X^2$ tests are useful:

- With a sample of 100 San Diegans, you wonder whether the distribution of ages (grouped by categories) in S.D. is the same as in the U.S. overall.
- With a sample of 100 San Diegans and 100 New Yorkers, you wonder if the age distribution in S.D. is the same as in N.Y.
- With a sample of 100 San Diegans, you try to figure out if there is a link between sex (0/M, 1/F) and eye color (Brown, Blue, Green, ...).
Before carrying on, we need to model "binomial variable" generically, when there are $t > 2$ bins.

This is a generalisation of the binomial distribution, that can be seen as $t=2$ bins "0" and "1".

**Def. (Multinomial Distribution)**

Let $X_i$ denote the number of times the outcome $r_i$ occurs ($1 \leq i \leq t$) in a series of $n$ independent Bernoulli trials with $P(r_i) = p_i$.

Then the vector $(X_1, ..., X_t)$ has a multinomial distribution, and

$$P(X_1 = k_1, ..., X_t = k_t) = \frac{n!}{k_1! \times ... \times k_t!} p_1^{k_1} \ldots p_t^{k_t},$$

for all $0 \leq k_1, ..., k_t \leq n$ with $\sum_{i=1}^{t} k_i = n$.

Sometimes denoted by $(m \choose k_1, ..., k_t)$, "multinomial coefficient".

**Rk:** If $\sum_{i=1}^{t} k_i \neq n$, then

$$P(X_1 = k_1, ..., X_t = k_t) = 0$$

condition required because we put $n$ elements (exactly) in the $t$ bins.
Example: The positions on a roulette wheel are divided into 3 colors:

red, black and green: $\begin{cases} 18 \text{ red} \\ 18 \text{ black} \\ 2 \text{ green} \end{cases}$

If the wheel is fair, determine

1) The probability that in 7 independent spins of the wheel, the ball lands in a red slot 4 times and in a black slot 3 times.

2) The chance the ball in a red slot at least three times (in 7 spins)

\[ \text{Sol. } \begin{align*}
1) & \quad m = 7, \quad k_1 = 4, \quad k_2 = 3, \quad k_3 = m - k_1 - k_2 = 0 \\
& \quad P_1 = \frac{18}{38} \quad P_2, \quad P_3 = \frac{2}{38}. \quad \text{We find } 0.1873
\end{align*} \]

2) This is modelled by a binomial $(k=2)$ with parameters

$m = 7$ and $p = \frac{18}{38}$. We find 0.7281
Prop. If \( X = (X_1, \ldots, X_t) \) has a multinomial distribution, then all the marginals of \( X \) have binomial distribution and
\[
E(X) = \begin{pmatrix}
mp_1 \\
\vdots \\
mp_t
\end{pmatrix}
\]
\[
\text{Var}(X) = 
\begin{pmatrix}
mp_1(1-p_1) & -mp_2p_1 & \cdots & -mp_tp_1 \\
-mp_1p_2 & mp_1p_1 & \cdots & -mp_2p_t \\
\vdots & \vdots & \ddots & \vdots \\
mP_{t-1}p_t & \cdots & -mp_tp_{t-1} & mp_t(1-p_t)
\end{pmatrix}
\]

Proof: For all \( 1 \leq i \leq t \), \((X_i)_i = X \) can be seen as the sum of independent multinomial variables with parameters \( n = 1 \) and \( p_1, \ldots, p_t \). Hence, \( E(X) = mp \bar{E}(X)^{(n)} = \left( \begin{pmatrix} mP_1 \\ \vdots \\ mPt \end{pmatrix} \right) \) and
\[
\text{Var}(X) = m \text{Var}(X)^{(n)}, \text{ But for all } 1 \leq i \neq j \leq t,
\]
\[
\text{Cov}(X_i, X_j) = \begin{cases}
0 & \text{if } i \neq j \\
p_i & \text{if } i = j
\end{cases}
\]
\[
\text{so that } \begin{cases}
\text{Var}(X_i) = p_i - p_i^2 = p_i(1-p_i) \\
\text{Cor}(X_i, X_j) = 0 - p_i p_j
\end{cases}
\]
which is the result.
10.3 Goodness of Fit \( \chi^2 \) Test

To check if a sample made of counts (categorical/binned) "fits" some multinomial you think it should fit.

**Setting:** Independent samples, sample size large enough.

**Hypothesis:**

\[ H_0: \ p_1 = p_{1o}, p_2 = p_{2o}, \ldots, p_r = p_{ro}. \]

- The true probability of bin \( i \).
- The expected probability of bin \( i \).

\[ H_1: \ p_i \neq p_{io} \text{ for at least one } i \]

**Idea:** If \( p_i \) truly is equal to \( p_{io} \) (i.e. under \( H_0 \)), you expect to count approximately \( n \times p_{io} \) elements among \( n \) samples, ending up in bin \( i \).

If not, this count should deviate from \( n p_{io} \).

Take away formula to enclose this (i.e test statistic):

\[ D = \sum \left( \frac{\text{Expected counts}}{\text{Observed counts}} \right) \]

To take bins into account altogether, Renormalization factor.
Under $H_0$, $D$ should be small.

Under $H_a$, $D$ should be large (and go to infinity).

We formalize this idea through the following result, that states how multinomials behave:

**Theorem:** If $X = (X_1, \ldots, X_t)$ has multinomial distribution with parameters $m$, and $p_1, \ldots, p_t$, and that $mp_i \geq 5 \ (1 \leq i \leq t)$, then

$$D = \sum_{i=1}^{t} \frac{(X_i - mp_i)}{mp_i} \chi^2 \sim \chi^2_{t-1}$$

has approximate distribution.

**Proof:** From the central limit theorem,

$$\sqrt{m} \left\{ \frac{(X_i - \bar{X})}{\bar{X}} \right\} \xrightarrow{p \to \infty} \mathcal{N}_t(0, \Sigma)$$

where

$$\Sigma = \begin{pmatrix}
    p_1(1-p_1) & \cdots & -p_1p_t \\
    \vdots & \ddots & \vdots \\
    -p_t p_1 & \cdots & p_t(1-p_t)
\end{pmatrix}.$$
Notice that $\Sigma$ is not an invertible matrix (sum of columns equal to 0), so that we cannot renormalize the left-hand side properly.

Let us consider $Y = \begin{pmatrix} X_1 \\ \vdots \\ X_{k-1} \end{pmatrix}$ (subvector of $X$). Then we have

$$\sqrt{m} \left( Y - \begin{pmatrix} p_1 \\ \vdots \\ p_{k-1} \end{pmatrix} \right) \xrightarrow{m \to \infty} N_{k-1}(0, \Sigma^*)$$

where $\Sigma^*$ is the upper-left $(k-1) \times (k-1)$ submatrix of $\Sigma$.

One may check that $\Sigma^*$ is invertible, and that

$$(\Sigma^*)^{-1} = \begin{pmatrix}
\frac{1}{p_1} & \frac{1}{p_{k-1}} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}} \\
\frac{1}{p_{k-1}} & \frac{1}{p_1} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}} \\
\frac{1}{p_k} & \frac{1}{p_k} & \frac{1}{p_1} & \cdots & \frac{1}{p_{k-1}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{p_{k-1}} & \frac{1}{p_{k-1}} & \frac{1}{p_{k-1}} & \cdots & \frac{1}{p_1} 
\end{pmatrix}$$

Furthermore,

$$D = \left\| \sqrt{m} \left( Y - \begin{pmatrix} p_1 \\ \vdots \\ p_{k-1} \end{pmatrix} \right) \right\|^2 = m \left( \frac{Y - \begin{pmatrix} p_1 \\ \vdots \\ p_{k-1} \end{pmatrix}^T (\Sigma^*)^{-1} (Y - \begin{pmatrix} p_1 \\ \vdots \\ p_{k-1} \end{pmatrix}) \right)$$

And since $Z_m \xrightarrow{m \to \infty} N_{k-1}(0, \Sigma^*)$, we get the result.
The proof only prove an asymptotic convergence. Condition "\( m_i \geq 5 \) for all \( 1 \leq i \leq r \) is here to account for this. Always have in mind that this is an approximate statement.

From this, we derive a goodness-of-fit test:

**Corollary (\( \chi^2 \) goodness-of-fit test)**

Let \( k_1, \ldots, k_r \) be the observed counts for the outcomes \( \pi_1, \ldots, \pi_r \).

At the \( \alpha \) level of confidence for the test

\[
H_0 : p_i = \pi_i \quad \forall 1 \leq i \leq r
\]

VS

\[
H_1 : p_i \neq \pi_i \quad \text{for at least one } i.
\]

with \( m_i \geq 5 \) for all \( 1 \leq i \leq r \), reject \( H_0 \) when

\[
D = \sum_{i=1}^{r} \frac{(k_i - m_i \pi_i)^2}{m_i \pi_i} \geq \chi^2_{r-1 - (1-\alpha)}
\]

Quantile of order \( 1-\alpha \) of the \( \chi^2_{r-1} \) distribution
Rk: The quantile of order \( 1 - \alpha \) of a distribution is the value that has \( \alpha \) area to its right.

Example: Testing Mendel's theory on alleles for peas.

Mendel's theory for peas:
- An allele \( R/r \) codes roundness
- An allele \( Y/y \) codes yellowness
- \( R \) and \( Y \) are dominant
- \( r \) and \( y \) are recessive

In a cross, this theory predicts that observed phenotypes come with frequencies:

- \( \frac{9}{16} \): Yellow + Round → Phenotype #1
- \( \frac{3}{16} \): non-yellow + Round → #2
- \( \frac{3}{16} \): Yellow + non-round → #3
- \( \frac{1}{16} \): non-yellow + non-round → #4
After 100 peas bred, you get:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Observed Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>62</td>
</tr>
<tr>
<td>#2</td>
<td>24</td>
</tr>
<tr>
<td>#3</td>
<td>9</td>
</tr>
<tr>
<td>#4</td>
<td>5</td>
</tr>
</tbody>
</table>

Do these data provide evidence that Mendel's theory is wrong?

This amounts to test if $H_0: (p_1, p_2, p_3, p_4) = \left( \frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right)$ is true or false. (Here, $n = 100$)

We can display our data in the following way:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Observed Count</th>
<th>Expected count under $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>62</td>
<td>$mP_{10} = 100 \times \frac{9}{16} = 56.5$</td>
</tr>
<tr>
<td>#2</td>
<td>24</td>
<td>$mP_{20} = 100 \times \frac{3}{16} = 18.75$</td>
</tr>
<tr>
<td>#3</td>
<td>9</td>
<td>$mP_{30} = 100 \times \frac{3}{16} = 18.75$</td>
</tr>
<tr>
<td>#4</td>
<td>5</td>
<td>$mP_{40} = 100 \times \frac{1}{16} = 6.25$</td>
</tr>
</tbody>
</table>

We get $\chi^2 = D = \frac{(62 - 56.5)^2}{56.5} + \frac{(24 - 18.75)^2}{18.75} + \frac{(9 - 18.75)^2}{18.75} + \frac{(5 - 6.25)^2}{6.25}$

$\approx 7.378$.

Since all the expected counts ($mP_{ij}$) are $\geq 5$, we can apply the $\chi^2$-test, with df $= 4 - 1 = 3$
On the table of the $\chi^2$ distribution, we see that the area to the right of 7.378 (= p-value) satisfies

$0.05 \leq p\text{-value} \leq 0.1$.

Hence, at the level $\alpha = 5\% = 0.05$, we do not reject $H_0$.

Rk: What is this test really doing?

In essence, a $\chi^2$ goodness-of-fit test tells you if the observed data differ from the data expected on some theory by an amount that exceeds the type of variation we expect from random effect.

Said differently, $\chi^2 = D$ tells you if the two histograms (boxplots) are "close enough" to each other (null) or not (alternative).
10.4 Goodness-of-fit $\chi^2$ test: Parameters unknown

In some situations one has reason to believe that some observed counts should be distributed as a multinomial (and wants to test this) but has no a-priori parameter values ($p_1, \ldots, p_k$). Hence, the raw goodness-of-fit $\chi^2$ test cannot be applied.

A strategy to overcome this issue is to replace the expected counts $m \cdot p_i$ (not available anymore) by estimated expected counts $m \cdot \hat{p}_i$, where $\hat{p}_i$ is the estimated proportion of outcome $r_i$.

Done so, we'll have to take into account the extra randomness of $m \cdot \hat{p}_i$ into the number of degrees of freedom of the $\chi^2$ distribution.

**Theorem:**

If $X = (X_1, \ldots, X_k)$ has multinomial distribution with $\Delta \leq t$ unknown parameters (= parameters you need to estimate), let $\hat{p}_i$ denote the estimated probability of outcome $r_i$ based on $X$.

Then, if $m \cdot \hat{p}_i \geq 5$ for all $i \leq t$, $X^2 = \sum_{i=1}^{t} \frac{(X_i - m \cdot \hat{p}_i)^2}{m \cdot \hat{p}_i} \sim \chi^2_{t-1-\Delta}$

**Proof:** Not covered here, it has the same flavor as for $\hat{p}_i$ replaced by $p_i$. 
As a corollary, we get a test procedure to test if a sample comes from some (partially unspecified) multinomial distribution.

Next section gives an important application of this result.

10.5 Contingency Tables and the $X^2$ Independence Test

Generalizing the above, where we studied a single variable $X$ with $t$ possible outcomes, we now move to the study two categorical variables $(X, Y)$. The main idea here is to determine (test) whether or not $X$ and $Y$ are independent: $(X, Y)$ describes two traits of a single individual.

Example: $X$: Gender, $Y$: Eye color

$X$: Sexual orientation, $Y$: State of birth is the US.

In general, say that $X$ has possible outcomes $A_1, \ldots, A_n$,

$Y$ has possible outcomes $B_1, \ldots, B_c$.

We setup notation

$P(X = A_i \text{ and } Y = B_j) = P_{ij}$, with

$\sum_{i=1}^{n} \sum_{j=1}^{c} P_{ij} = 1$

$P(X = A_i) = p_i$

$P(Y = B_j) = q_j$

Clearly,

$p_i = \sum_{j=1}^{c} P_{ij}$

$q_j = \sum_{i=1}^{n} P_{ij}$
Say we want to test if \( X \) and \( Y \) are independent (\( H_0 \)). If \( H_0 \) is true, then 
\[
P(X = A_i, Y = B_j) = P(X = A_i) \times P(Y = B_j),
\]
by definition of independence. In other words, \( \hat{P}_{ij} = \hat{p}_i \hat{q}_j \) for all \((i,j)\).

Say we estimate, from sample, \( \hat{p}_1, \ldots, \hat{p}_n \) and \( \hat{q}_1, \ldots, \hat{q}_c \) with \( \hat{P}_{11}, \ldots, \hat{P}_{np} \) and \( \hat{Q}_{11}, \ldots, \hat{Q}_{cp} \), we can test if \( \hat{P}_{ij} = \hat{p}_i \hat{q}_j \approx \hat{P}_{ij} \).

**Thm: (\( \chi^2 \) Independence Test)**

Suppose that \( n \) observations are partitioned by the events \( A_1, \ldots, A_n \) and also by \( B_1, \ldots, B_c \). Let \( k_{ij} \) denote the number of observations in the sample that belong to \( A_i \cap B_j \).

To test \( H_0: \) The \( A_i \)'s are independent from the \( B_j \)'s

\( \text{vs} \)

\( H_a: \) They're not independent

the null is rejected at the \( \alpha \) level of significance if

\[
\chi^2 = D = \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{\left( k_{ij} - m \hat{p}_i \hat{q}_j \right)^2}{m \hat{p}_i \hat{q}_j} \geq \chi^2_{1-n, (n-1)(c-1)}
\]

**P.S:** This test is valid only if \( m \hat{p}_i \hat{q}_j \geq 5 \) for all \((i,j)\).
Proof: We are considering the test of multinomial distributions for \( n \times c \) possible outcomes. In the process, we estimated:

\[
\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n \rightarrow \text{which makes only } (n-1) \text{ estimates},
\]

since we have \( \sum_{i=1}^{n} \hat{p}_i = 1 \Rightarrow \hat{p}_n = 1 - \sum_{i=1}^{n-1} \hat{p}_i \).

\[
\hat{q}_1, \ldots, \hat{q}_c \rightarrow (c-1) \text{ estimates too.}
\]

In total, we estimated \( \Delta = (n-1) + (c-1) \) parameters. And at the end of the day, we get the degrees of freedom equal to:

\[
1 - \Delta - 1 = n \times c - (n+c-2)-1
\]

\[
= (n-1)(c-1)
\]

Example: Testing association between smoking and cancer.

Pyrobenzene is a major component of cigarette smoke. Researchers injected rats with different levels of pyrobenzene, and looked for tumor development. In 230 rats injected, they got these data:

<table>
<thead>
<tr>
<th></th>
<th>No tumor</th>
<th>One tumor</th>
<th>( \geq 2 ) tumors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>74</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Low dose</td>
<td>63</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>High dose</td>
<td>45</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

\( n = 3 \)  \( c = 3 \)

\([Rk]: \text{Usually, it helps to display data in a two-way table, also called contingency table.} \)
Do these data show evidence of an association between pyroligneous exposure and tumor development for rats?

We compute the expected counts \( \hat{\lambda}_i \hat{\lambda}_j \).

**Useful formula:**
\[
\hat{\lambda}_i \hat{\lambda}_j = \frac{(\text{Row total}) \times (\text{Column total})}{m}
\]

Here, we get the expected contingency table as follows:

<table>
<thead>
<tr>
<th></th>
<th>No tumor</th>
<th>One tumor</th>
<th>( \geq 2 ) tumors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>63.3</td>
<td>11.13</td>
<td>5.57</td>
</tr>
<tr>
<td>Low dose</td>
<td>63.3</td>
<td>11.13</td>
<td>5.57</td>
</tr>
<tr>
<td>High dose</td>
<td>55.39</td>
<td>9.74</td>
<td>4.87</td>
</tr>
</tbody>
</table>

We get \( X^2 = 19.25 \).

Looking at a statistical table of the \( X^2 \) distribution with
\[
df = (r-1) \times (c-1) = (3-1) \times (3-1) = 4,
\]
we get \( p \)-value < 0.001.
Hence, as $p$-value $< 0.001 < \alpha = 0.05$, we reject the null:

At the level of significance $\alpha = 5\%$, there is evidence of an association between pyrobenzene exposure and tumor development in rats.

⚠️ A $\chi^2$ test of independence tells you if there is an association or not, but it doesn't inform you on the sense of this association!!

In the example above, pyrobenzene, the $\chi^2$ test might actually have detected that pyrobenzene cures cancer...!

Rk: What is this test really doing?
It basically compares how distributions look relative to each other.