# $\chi^2$ Tests

Pronounced "Khi squared"

$\chi^2$ 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 quantitive) if it has a discrete (or even finite) set of possible outcomes.
↑ Think to bins

Examples where $\chi^2$ 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 (♂/♀) and eye color (Brown, Blue, Green, ...).

(2)

Before carrying on, we need to model "binned variable" generically, when there are $t \geq 2$ bins.

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

10.2

**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 $\mathbb{P}(r_i) = p_i$. (The $i^{th}$ bin)

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

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

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

Sometimes denoted by $\binom{n}{k_1 \dots k_t}$: "Multinomial coefficient"

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

$$\mathbb{P}(X_1 = k_1, \dots, X_t = k_t) = 0$$

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 reed slot at least three times (in 7 spins)

Sol. 1) $n = 7$, $k_1 = 4$, $k_2 = 3$, $k_3 = n - k_1 - k_2 = 0$

$P_1 = \frac{18}{38} = P_2$, $P_3 = \frac{2}{38}$. We find 0.1873

2) This is modelled by a binomial ($t = 2$) with parameters $n = 7$ and $p = \frac{18}{38}$. We find 0.728

**Prop:** If $X = (X_1, \dots, X_t)$ has a multinomial distribution, then all the marginals of $X$ have binomial distribution and

$$\mathbb{E}(X) = \begin{pmatrix} np_1 \\ \vdots \\ np_t \end{pmatrix}$$

$$Var(X) = \begin{pmatrix} np_1(1-p_1) & -np_2p_1 & \cdots & -np_tp_1 \\ -np_1p_2 & \ddots & & \vdots \\ \vdots & & & -np_tp_{t-1} \\ -np_1p_t & \cdots & -np_{t-1}p_t & np_t(1-p_t) \end{pmatrix}$$

**Proof:** For all $1 \le i \le t$, $(X_i) = X$ can be seen a the sum $X^{(1)} + \dots + X^{(n)}$ of independent multinomial variable with parameters $n = 1$, and $p_1, \dots, p_t$. Hence, $\mathbb{E}(X) = n\mathbb{E}(X^{(1)}) = \begin{pmatrix} np_1 \\ \vdots \\ np_t \end{pmatrix}$ and

$Var(X) = n\,Var(X^{(1)})$. But for all $1 \le i, j \le t$,

$$\mathbb{E}(X_i^{(1)} X_j^{(1)}) = \begin{cases} 0 & \text{if } i \neq j \\ p_i & \text{if } i = j \end{cases}$$

so that $\begin{cases} Var(X_i) = p_i - p_i^2 = p_i(1-p_i) \\ Cov(X_i, X_j^{(1)}) = 0 - p_i p_j \end{cases}$, 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.

Hypotheses: $H_0: p_1 = p_{1_0}, p_2 = p_{2_0}, \dots, p_t = p_{t_0}$

$p_1$: the true probability of bin 1

$p_{1_0}$: the expected probability of bin 1.

$H_1: p_i \neq p_{i_0}$ for at least one $i$

Idea:
- If $p_i$ truly is equal to $p_{i_0}$ (i.e. under $H_0$), you expect to count approximately $n \times p_{i_0}$ elements, among $n$ samples, ending up in bin $i$
- If not, this count should deviate from $np_{i_0}$.

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

$$D = \sum_{\text{bins}} \frac{(\text{Expected}_{\text{counts}} - \text{Observed}_{\text{count}})^2}{\text{Expected}_{\text{counts}}}$$

To take bins into account altogether

For two bins not to compensate

Renormalization factor

(6) . Under $H_0$, $D$ should be small

. Under $H_1$, $D$ should be large (and go to infinity)

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

---

**Thm:** If $X = (X_1, \dots, X_t)$ has multinomial distribution with parameters $n$, and $p_1, \dots, p_t$, and that $np_i \geq 5$ $(1 \leq i \leq t)$,

(Observed counts; Expected counts)

then

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

($D$: $\chi^2$ statistic; $\approx$: "has approximate distribution")

---

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

$$\sqrt{n}\left\{\begin{pmatrix} X_1/n \\ \vdots \\ X_t/n \end{pmatrix} - \begin{pmatrix} p_1 \\ \vdots \\ p_t \end{pmatrix}\right\} \xrightarrow[n\to\infty]{D} N_t(0, \Sigma),$$

where $\Sigma = \begin{pmatrix} p_1(1-p_1) & \cdots & -p_j p_i \\ \vdots & \ddots & \vdots \\ -p_i p_j & \cdots & p_t(1-p_t) \end{pmatrix}$.

(7)

⚠ Notice that $\Sigma$ is not an invertible matrix (sum of columns is 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

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

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

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

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

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

And since $Z_n \xrightarrow[n\to\infty]{\mathcal{D}} N_{k-1}(0, I_{k-1})$, we get the result.

Rk: The proof only prove an asymptotic convergence. Condition "$np_i \geq 5$ 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:

Corr: ($\chi^2$ goodness-of-fit test)

Let $k_1, \dots, k_r$ be the observed counts for the outcomes $\pi_1, \dots, \pi_r$.

At the $\alpha$ level of confidence for the test

True proba. of bin $i$ — Tested proba. of bin $i$

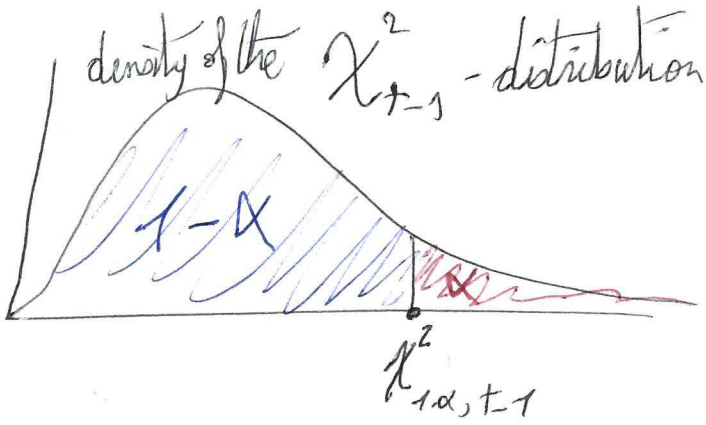
$$\begin{cases} H_0 : & p_i = p_{i_0} \quad \forall\, 1 \leq i \leq r \\ & \text{(VS)} \\ H_1 : & p_i \neq p_{i_0} \text{ for at least one } i. \end{cases}$$

with $np_i \geq 5$ for all $1 \leq i \leq r$, reject $H_0$ when

$$D = \sum_{i=1}^{r} \frac{(k_i - np_{i_0})^2}{np_{i_0}} \geq \chi^2_{1-\alpha,\, r-1}$$

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 $\begin{cases} 1-\alpha \text{ area to its left} \\ \alpha \quad \text{---} \quad \text{right} \end{cases}$




---

**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
- $\begin{cases} R \text{ and } Y \text{ are dominant} \\ r \text{ and } y \text{ are recessive} \end{cases}$

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

(10)

After 100 peas bred, you get:

| Phenotype | Observed counts |
|---|---|
| #1 | 62 |
| #2 | 24 |
| #3 | 9 |
| #4 | 5 |

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

⟶ This amounts to test if $H_0: (p_1, p_2, p_3, p_4) = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$ (= $p_{1_0}, p_{2_0}, p_{3_0}, p_{4_0}$) is true or false. (Here, $n = 100$)

We can display our data in the following way:

| Phenotype | Observed count | Expected count under $H_0$ |
|---|---|---|
| #1 | 62 | $np_{1_0} = 100 \times \frac{9}{16} = 56.5$ |
| #2 | 24 | $np_{2_0} = 100 \times \frac{3}{16} = 18.75$ |
| #3 | 9 | $np_{3_0} = 100 \times \frac{3}{16} = 18.75$ |
| #4 | 5 | $np_{4_0} = 100 \times \frac{1}{16} = 6.25$ |

⚠ DO NOT round the expected counts.

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

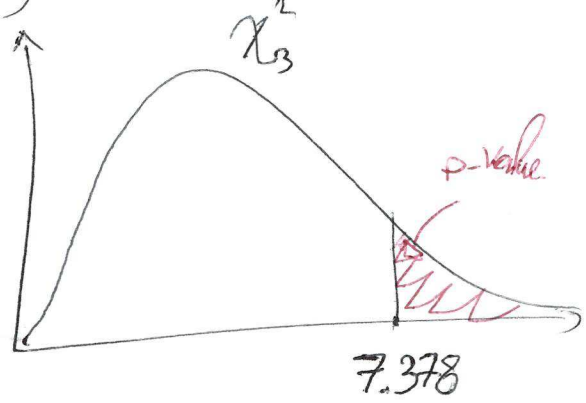
$\simeq 7.378.$

Since all the expected counts ($np_{i_0}$) are $\geq 5$ we can apply the $\chi^2$-test, with $df = 4 - 1 = 3$

(11)

On the table of the $\chi^2_3$ distribution, we see that the area to the right of 7.378 (= p-value) satisfies

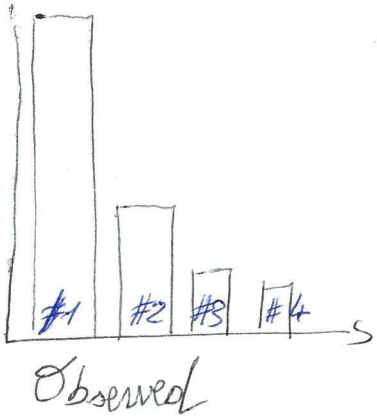
$$0.05 \leq \text{p-value} \leq 0.1 .$$




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 (barplots) 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 $(p_{1_0}, \dots, p_{t_0})$. Hence, the raw goodness-of-fit $\chi^2$ test cannot be applied.

A strategy to overcome this issue is to replace the expected counts $np_{i_0}$ (not available anymore) by estimated expected counts $n\hat{p}_i$, where $\hat{p}_i$ is the estimated proportion of outcome $r_i$.

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

---

**Thm:** If $X = (X_1, \dots, X_t)$ has multinomial distribution with $s \leq t$ unknown parameters ← (= parameters you need to estimate)
Let $\hat{p}_i$ denote the estimated probability of outcome $r_i$ based on $X$.

Then, if $n\hat{p}_i \geq 5$ for all $1 \leq i \leq t$,

$$\chi^2 = D = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i} \simeq \chi^2_{t-1-s}$$

---

**Proof:** Not covered here, it has the same flavor as for $\hat{p}_i$ replaced by $p_i$.

Rk: . 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 $\chi^2$ 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 in the US.

In general, say that: $X$ has possible outcomes $A_1, \dots, A_r$

$Y$ $\quad\quad$ $B_1, \dots, B_c$

($r$ and $c$ have no reason to be equal.)

We setup notation

. $\mathbb{P}(X = A_i \text{ and } Y = B_j) = p_{ij}$ , with $\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$

. $\mathbb{P}(X = A_i) = p_i$

. $\mathbb{P}(Y = B_j) = q_j$

(Clearly, $\begin{cases} p_i = \sum_{j=1}^{c} p_{ij} \\ q_j = \sum_{i=1}^{r} p_{ij} \end{cases}$)

(14)

Say we want to test if $X$ and $Y$ are independent ($H_0$). If $H_0$ is true, then $\mathbb{P}(X=A_i, Y=B_j) = \mathbb{P}(X=A_i) \times \mathbb{P}(Y=B_j)$, by definition of independence. In other words, $p_{ij} = p_i q_j$ for all $(i,j)$.

Say we estimate, from sample, $p_1, \dots, p_r$ and $q_1, \dots, q_c$ with $\hat{p}_1, \dots, \hat{p}_r$ and $\hat{q}_1, \dots, \hat{q}_c$, we can test if $p_{ij} = p_i q_j \simeq \hat{p}_i \hat{q}_j$.

Thm: ($\chi^2$ Independence test)

Suppose that $n$ observations are partitionned by the events $A_1, \dots, A_r$ and also by $B_1, \dots, B_c$. 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

vs

$H_1$: They're not independent

the null is rejected at the $\alpha$ level of significance if

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

Rk: This test is valid only if $n\hat{p}_i \hat{p}_j \geq 5$ for all $(i,j)$.

Proof: (Idea of the) We are considering the test of multinomial distributions for $r \times c$ possible outcomes. In the process, we estimated:

- $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_r \longrightarrow$ which makes only $\boxed{r-1}$ estimates, since we have $\sum_{i=1}^{r} \hat{p}_i = 1 \Rightarrow \hat{p}_r = 1 - \sum_{i=1}^{r-1} \hat{p}_i$
- $\hat{q}_1, \dots, \hat{q}_c \longrightarrow \boxed{c-1}$ estimates too.

In total, we estimated $s = (r-1) + (c-1)$ parameters, And at the end of the day, we get the degrees of freedom equal to

$$t - s - 1 = rc - (r+c-2) - 1$$
$$= (r-1)(c-1) \quad \square$$

---

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 developpement. In 230 rats injected, they got these data:

| | No tumor | One tumor | $\geq 2$ tumors |
|---|---|---|---|
| Control | 74 | 5 | 1 |
| Low dose | 63 | 12 | 5 |
| High dose | 45 | 15 | 10 |

$r = 3$

Observed counts

$c = 3$

Rk: Usually, it helps to display data in a two-way table, also called contingency table

(16) Do these data show evidence of an association between pyrolenzene exposure and tumor developpement for rats?

We compute the expected counts $n\hat{p}_{i\cdot}\hat{p}_{\cdot j}$.

Rk: Useful formula: $$n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{(\text{Row total}) \times (\text{Column total})}{n}$$

Here, we get the expected contingency table as follows

| | No tumor | One tumor | $\geq 2$ tumors |
|---|---|---|---|
| Control | 63.3 | 11.13 | 5.57 |
| low dose | 63.3 | 11.13 | 5.57 |
| High dose | 55.39 | 9.74 | 4.87 |

(Expected counts

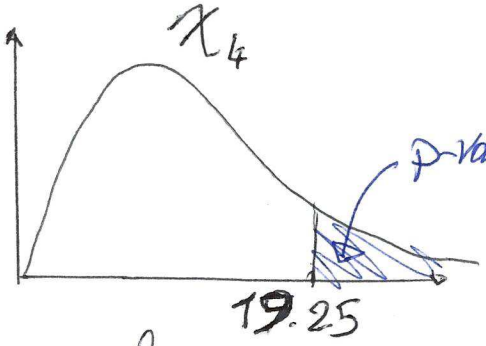
⚠ Do not round expected counts!

We get $$X^2 = D = \sum_{\text{cells}} \frac{(\text{Observed}_{\text{count}} - \text{Expected}_{\text{count}})^2}{\text{Expected}_{\text{count}}}$$

$$= 19.25$$

Looking at a statistical table of the $\chi^2$ distribution with $df = (r-1)\times(c-1) = (3-1)\times(3-1) = 4$, we get p-value $< 0.001$.

(17)




Hence, as p-value $< 0.001 < \alpha = 0.05$, we reject the null:

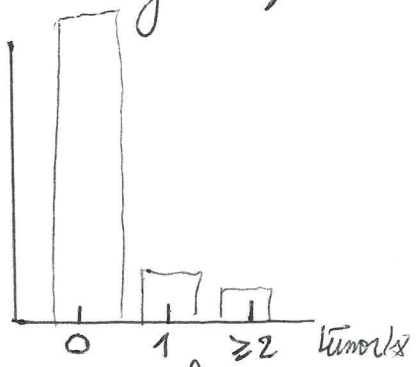
At the level of significance $\alpha = 5\%$, there is evidence of an association between pyrobenzene exposure and tumor development in rats.

⚠ 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!!

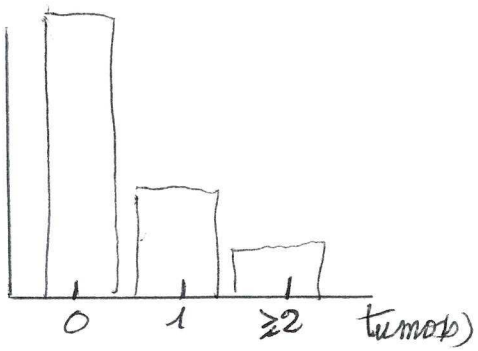
In the example above, pyrobenzene, the $\chi^2$ test might actually have detected that pyrobenzene cures cancer...!

Rk: What is this test really doing?

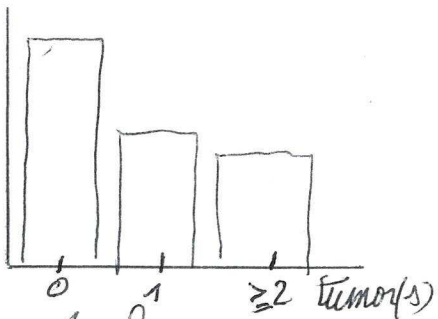
It basically compares how distributions look relative to each other




control




Low Dose




High Dose