0623 slides

MA 116

June 2025

Attendance

Attendance for this lecture is for record purpose only and will not affect your attendance credit.

Categorical data

In the chapter of regression, we discussed how to use dummy variables to do linear regression when our variable u is categorical. A data point in this setting looks like (u, y) for y a numerical value of the response variable; for example, when investigating ages of people living in different countries, a data point may be (UK, 32).

If my data point has only an individual entry that's categorical/qualitative, i.e. no numerical value at all, how can I do inference about this kind of sample?

Example.

Setting 1

Experiment: Randomly choose a candy from a bag of M&M candies and observe its color.

The sample space of this experiment is {brown, yellow, red, orange, blue, green}.

Suppose a manufacturer claims that the distribution of a bag of M&M candies is 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green. How can we test this claim?

We can associate a variable u to the outcomes of this experiment.

Goodness of fit test

This kind of test is called a **goodness of fit test** because it's testing whether our proposed/hypothetical distribution of u fits well to out sample data.

Another setting. I might want to test if a die is fair, i.e. the possibility of getting each number is the same $p_1 = \frac{1}{6} = p_2 = \cdots = p_6$.

For each category i , there associates a proposed/hypothetical probability p_i such that the sum of p_i is 1.

How to test a hypothesis?

Suppose I want to test if a die is fair. I roll this die n times for a sample size n large enough. For each $1 \leq i \leq 6$, let O_i denote the number of trials I get number i. Then $\sum n_i = n$.

Population parameters μ_i

The **expected count of the i-th outcome** is a population parameter $E_i = \mu_i = np_i$ for a fixed n. This is the population parameter that represents "the average number of the i-th outcome I get from n independent trials of an experiment with 6 mutually exclusive possible outcomes".

Idea: Instead of categorical data, we consider the quantitative quantities observed counts O_i vs. the expected counts E_i.

Consider a new variable

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where k is the number of categories, or number of mutually exclusive possible outcomes. Look at this definition, this variable is a **quantitative variable** that measures how a group of observed counts O_i differ from the expected counts E_i.

Distribution of this variable

When

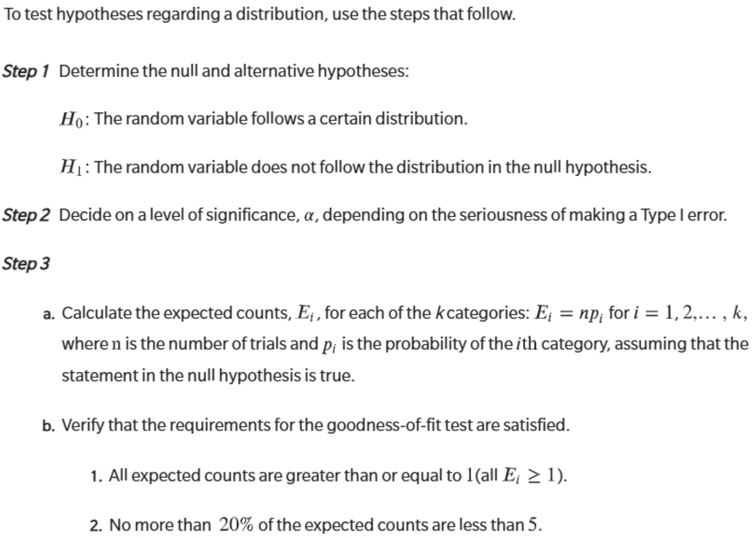
1. $E_i \geq 1$ for all $1 \leq i \leq k$,
2. No more than 20% of E_i are less than 5,

the distributino of χ^2 approximately follows the **chi-square distribution** with $k - 1$ degrees of freedom.

Chi-square distribution

- Not symmetrix
- The shape of a chi-square distribution depends on the degrees of freedom (just like Student's t-distirbution)
- As df increases, chi-square distributions become more and more symmetric
- this distribution is a good probability density function (values ≥ 0, total area $=1$)
- Read critical values from a table: Table VIII in our textbook.

The Goodness-of-Fit Test



To test hypotheses regarding a distribution, use the steps that follow.

Step 1 Determine the null and alternative hypotheses:

H_0: The random variable follows a certain distribution.

H_1: The random variable does not follow the distribution in the null hypothesis.

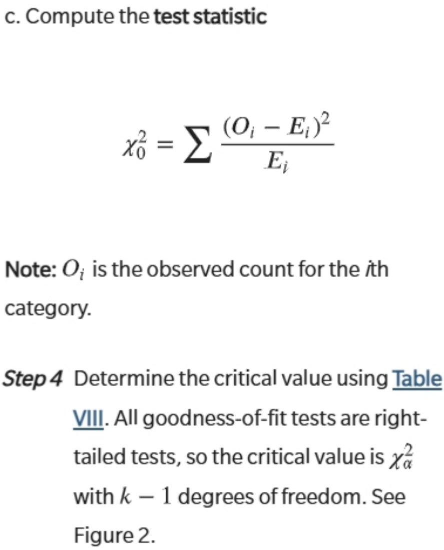
Step 2 Decide on a level of significance, α, depending on the seriousness of making a Type I error.

Step 3

a. Calculate the expected counts, E_i, for each of the k categories: $E_i = np_i$ for $i = 1, 2, \ldots, k$, where n is the number of trials and p_i is the probability of the $i\mathrm{th}$ category, assuming that the statement in the null hypothesis is true.

b. Verify that the requirements for the goodness-of-fit test are satisfied.

1. All expected counts are greater than or equal to 1 (all $E_i \geq 1$).
2. No more than 20% of the expected counts are less than 5.

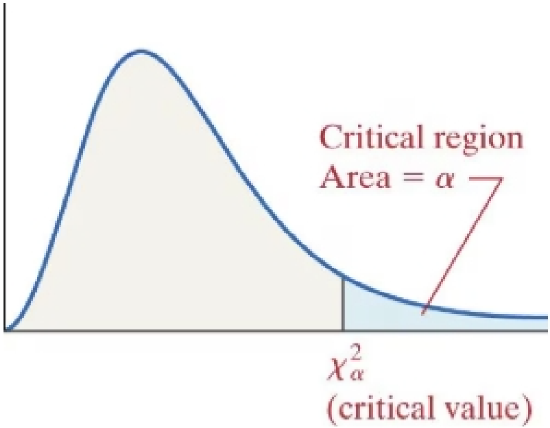


c. Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed count for the ith category.

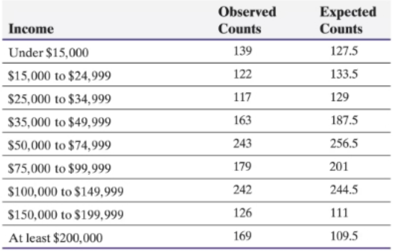
Step 4 Determine the critical value using Table VIII. All goodness-of-fit tests are right-tailed tests, so the critical value is χ_α^2 with $k - 1$ degrees of freedom. See Figure 2.



Compare the critical value to the statistic. If $\chi_0^2 > \chi_\alpha^2$, reject the null hypothesis.

One growing concern regarding the U.S. economy is the inequality in the distribution of income. An economist wants to know if the distribution of income 1500 is changing, so they randomly select households and obtain the household income shown in Table 2. Table 2 also contains the expected counts under the assumption the distribution has not changed since 2000 (obtained in Example 1). Does the evidence suggest that the distribution of income has changed since 2000 at the $\alpha = 0.05$ level of significance? **Note:** The data in Table 2 are based on the 2021 Current Population Survey and have been adjusted for inflation.

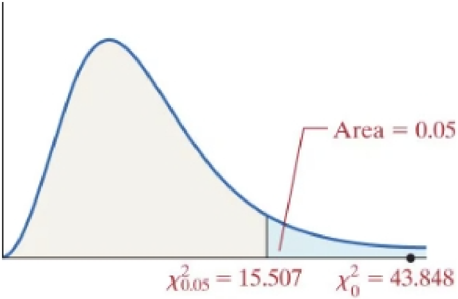
TABLE 2



Income	Observed Counts	Expected Counts
Under $15,000	139	127.5
$15,000 to $24,999	122	133.5
$25,000 to $34,999	117	129
$35,000 to $49,999	163	187.5
$50,000 to $74,999	243	256.5
$75,000 to $99,999	179	201
$100,000 to $149,999	242	244.5
$150,000 to $199,999	126	111
At least $200,000	169	109.5

$$
\begin{aligned}
\chi_0^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(139 - 127.5)^2}{127.5} + \frac{(122 - 133.5)^2}{133.5} \\
&\quad + \frac{(163 - 187.5)^2}{187.5} + \frac{(243 - 256.5)^2}{256.5} \\
&\quad + \frac{(242 - 244.5)^2}{244.5} + \frac{(126 - 111)^2}{111} \\
&= 43.848
\end{aligned}
$$

Figure 4



Because the test statistic, 43.848, is greater than the critical value, 15.507,we reject the null hypothesis.

14.2

Given two categorical variable, can we tell if they are related?

Let u be a categorical variable that represents "maritual statues". I.e. u may assume value in {married, widowed, divorced/separated, never married}. Let v be a categorical variable that represents "happiness". We let v assume value in {very happy, pretty happy, not too happy}.

A question we may ask is **is there a relationship between marital status and happness?**

Is there a relation between level of education and employment status?

Is there a relation between level of education and job satisfaction?

Is there a relation between political affiliation and job type?

Is there a relation between phone brand and age group? (10-20 years old, 20-30, 30-40, etc.)

A hypothesis test for answering this type of question is called a **chi-square test of independence**

Chi-square test of independence

H_0 : The two categorical variables are independent.
H_1 : The two categoriables are dependent.

Example.

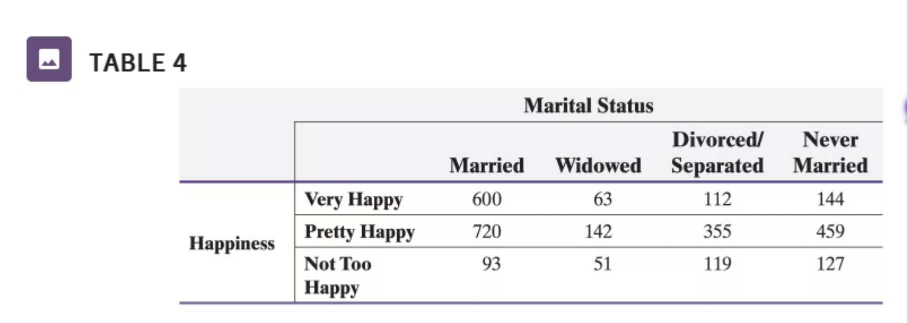
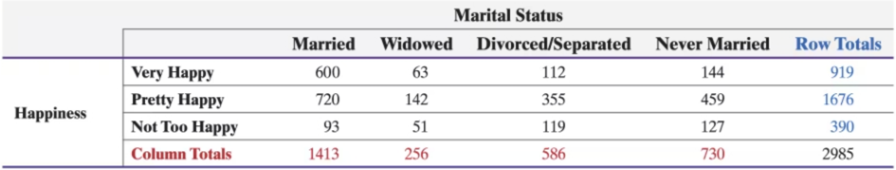
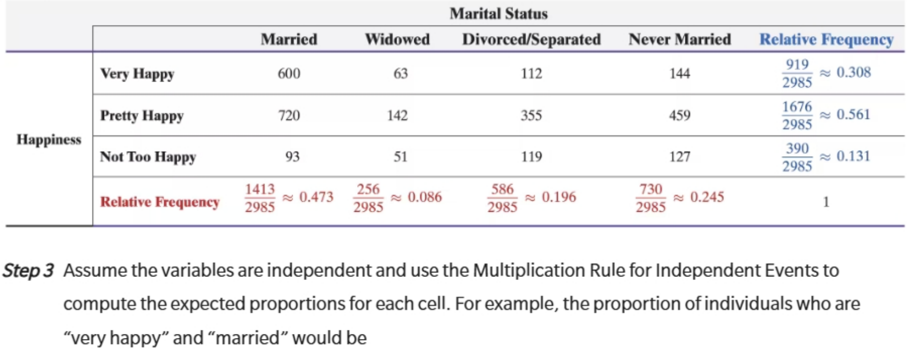


TABLE 4

		Marital Status			
		Married	**Widowed**	**Divorced/ Separated**	**Never Married**
Happiness	**Very Happy**	600	63	112	144
	Pretty Happy	720	142	355	459
	Not Too Happy	93	51	119	127



		Marital Status				
		Married	Widowed	Divorced/Separated	Never Married	Row Totals
Happiness	Very Happy	600	63	112	144	919
	Pretty Happy	720	142	355	459	1676
	Not Too Happy	93	51	119	127	390
	Column Totals	1413	256	586	730	2985



		Marital Status				
		Married	Widowed	Divorced/Separated	Never Married	Relative Frequency
Happiness	Very Happy	600	63	112	144	$\frac{919}{2985} \approx 0.308$
	Pretty Happy	720	142	355	459	$\frac{1676}{2985} \approx 0.561$
	Not Too Happy	93	51	119	127	$\frac{390}{2985} \approx 0.131$
	Relative Frequency	$\frac{1413}{2985} \approx 0.473$	$\frac{256}{2985} \approx 0.086$	$\frac{586}{2985} \approx 0.196$	$\frac{730}{2985} \approx 0.245$	1

Step 3 Assume the variables are independent and use the Multiplication Rule for Independent Events to compute the expected proportions for each cell. For example, the proportion of individuals who are "very happy" and "married" would be

Indepdendent events -> Multiplication of probability!

$$\begin{aligned}\begin{pmatrix}\text{Proportion “very happy”}\\ \text{and “married”}\end{pmatrix} &= (\text{proportion “very happy”}) \cdot (\text{proportion “married”})\\ &= \left(\frac{919}{2985}\right)\left(\frac{1413}{2985}\right)\\ &= 0.145737\end{aligned}$$

Table 7 shows the expected proportion in each cell, assuming independence.

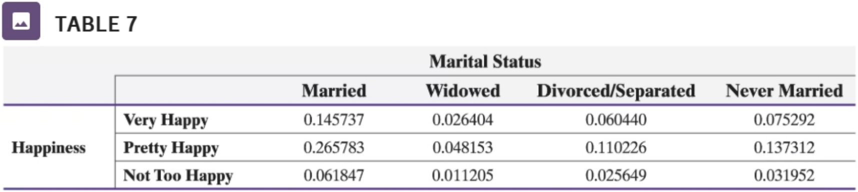
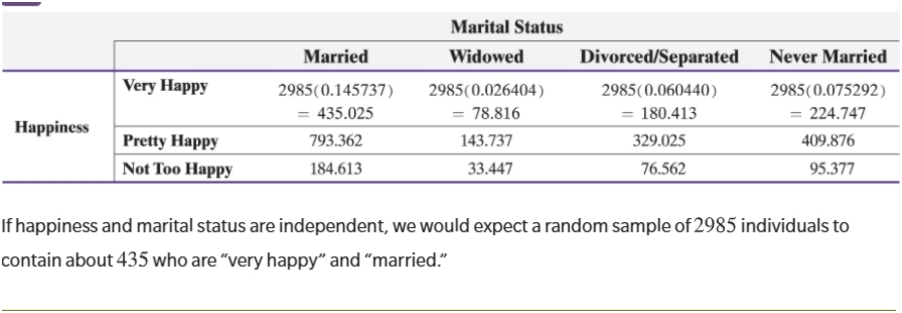


TABLE 7

		Marital Status			
		Married	Widowed	Divorced/Separated	Never Married
Happiness	Very Happy	0.145737	0.026404	0.060440	0.075292
	Pretty Happy	0.265783	0.048153	0.110226	0.137312
	Not Too Happy	0.061847	0.011205	0.025649	0.031952

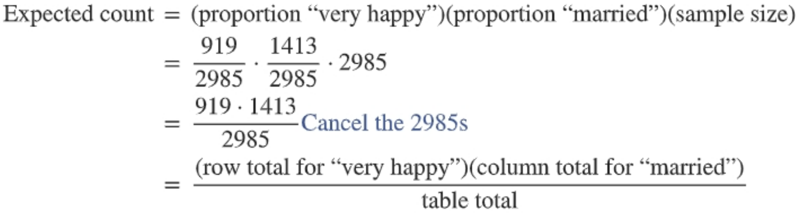
Step 4 Multiply the expected proportions in Table 7 by 2985, the sample size, to obtain the expected counts. See Table 8.

Calculate expected counts



		Marital Status			
		Married	**Widowed**	**Divorced/Separated**	**Never Married**
Happiness	**Very Happy**	2985(0.145737) = 435.025	2985(0.026404) = 78.816	2985(0.060440) = 180.413	2985(0.075292) = 224.747
	Pretty Happy	793.362	143.737	329.025	409.876
	Not Too Happy	184.613	33.447	76.562	95.377

If happiness and marital status are independent, we would expect a random sample of 2985 individuals to contain about 435 who are “very happy” and “married.”



$$\begin{aligned}\text{Expected count} &= (\text{proportion ``very happy''})(\text{proportion ``married''})(\text{sample size}) \\ &= \frac{919}{2985} \cdot \frac{1413}{2985} \cdot 2985 \\ &= \frac{919 \cdot 1413}{2985} \text{Cancel the 2985s} \\ &= \frac{(\text{row total for ``very happy''})(\text{column total for ``married''})}{\text{table total}}\end{aligned}$$

Once we have obtained the expected counts for each cell, we may let

$$\chi^2 = \sum_{1 \le i \le r,\, 1 \le j \le c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where E_{ij} is the expected count for the ij-th cell. Under certain conditions, this variable follows a chi-squares distribution of $(r-1)(c-1)$ degrees of freedom.

This variable allows us to do hypothesis test in the same way as in the previous section!

RMK. Unlike other hypothesis techniques we've learnt in this course, the chi-square test for independence involves uses of probability theory $P(A \cap B) = P(A)P(B)$ if A and B are independent events!

Applications of hypothesis tests

1. Personal decision making, such as personal financial decisions
2. Understanding of how statistical claims are used in news and ads, e.g. in a political setting
3. Understanding of how statistical claims are used in experimental subjects like biology, chemistry, and physics
4. Understanding of how statistical claims are used in industry, e.g. manufactural process quality control
5. Evidence-based reasoning