Chisquare Goodness-of-fit Test

Cheng Peng

STA200 Statistics II

Contents

1 Introduction

We have discussed the relationship between two numerical variables using the linear correlation coefficient and linear regression. We now turn to the relationship between two categorical variables. The idea is to make an assumption (hypothesis) about the variable(s) and use this assumption to construct a frequency table, known as the expected table. At the same time, we tabulate the data to obtain an observed table. The discrepancy between the expected and observed tables can be used to make an inference about the relationship between the categorical variables.

This module introduces a new positive distribution—the chi-square (χ^2) distribution—and its application in **goodness-of-fit** testing. **Goodness-of-fit tests for categorical distribution** assess the relationship between the **observed frequency distribution** and the **hypothetical distribution** under the null hypothesis (i.e., the expected distribution under H_0).

2 Chi-square (χ^2) Distribution

As we learned from MAT121/125, the **Chi-square (χ^2) distribution** is a continuous probability distribution commonly used in inferential statistics, especially for testing hypotheses about variances and categorical data (e.g., contingency tables). We will review this distribution and related problems.

2.1 Chi-Square Distribution

The key characteristics of χ^2 distribution are

- **Shape**: Skewed right, but becomes more symmetric as degrees of freedom (df) increase.
- **Support**: Defined only for non-negative values ($x \geq 0$).
- **Parameter**: Degrees of freedom (df), typically a positive integer.

The following YouTube video gives a nice introduction to the χ^2 distribution.

2.2 Two Types of Questions

As we learned from normal and t-distributions, we need to calculate tail probabilities, in particular, the right-tail probabilities and quantiles for given probabilities. We used chi^2 table in MAT121/125. You are encouraged to use R commands to find probabilities and quantiles.

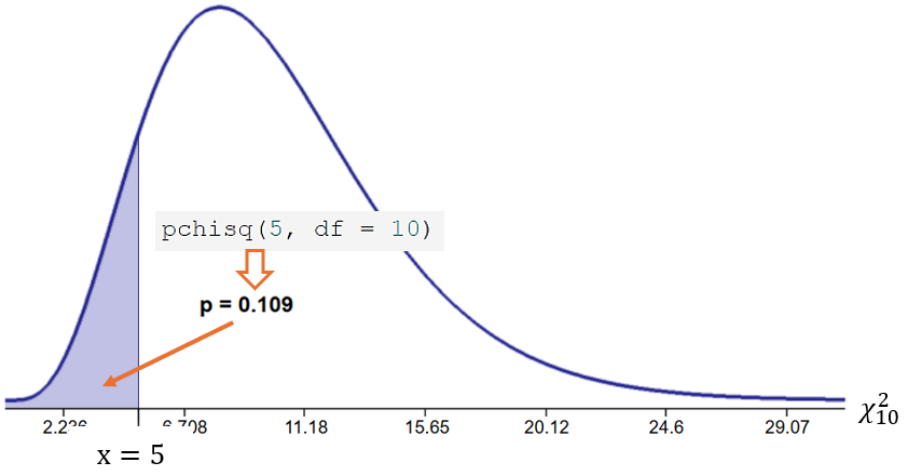
Finding Probability (Area under the curve)

- **Goal**: Calculate the probability that a χ^2 random variable falls below, above, or between certain values.
- **Typical Question**: What is $P(X \leq x)$ or $P(X \geq x)$ for a χ^2-distributed variable with $df = k$?
- **R Command**:

```
pchisq(x, df)                          # P(X <= x)
pchisq(x, df, lower.tail = FALSE)      # P(X >= x), x = chi-square test statistic
```

Note: **All χ^2 tests are right-tailed since the test statistic measures the discrepancy between two distributions.** The p-value of a χ^2 test is always the area of the tail region to the right-hand side of the test statistic, which can be found using `pchisq(x, df, lower.tail = FALSE)`. The argument `lower.tail = FALSE` returns the **upper-tail area** and `x` = chi-square test statistic.

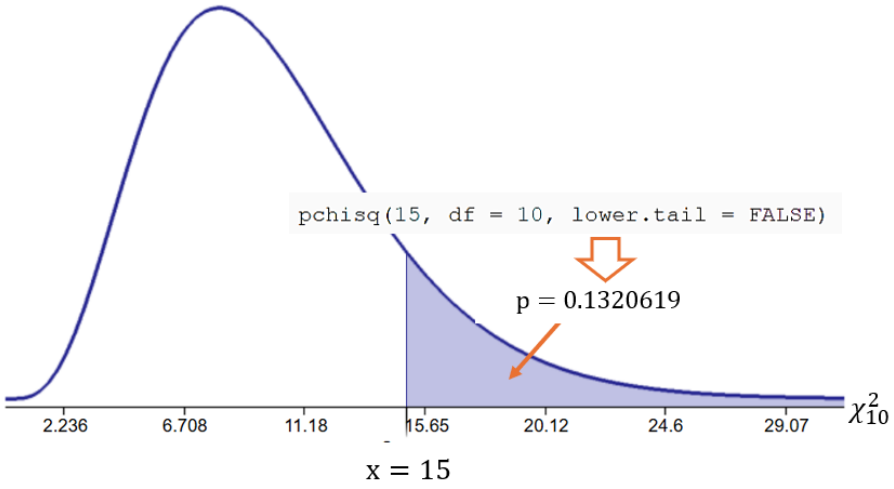
- **Example 1**: Find $P(X \leq 5)$ for $df = 10$.




```
pchisq(5, df = 10)
```

```
## [1] 0.108822
```

- **Example 2**: Find $P(X \geq 15)$ for $df = 10$.




```
pchisq(15, df = 10, lower.tail = FALSE)
```

```
## [1] 0.1320619
```

Finding Percentile (given the area under the curve)

- **Goal**: Calculate the cut χ^2 value (percentile) that defines the tail region with given area (i.e., the tail probability).
- **Typical Question**: What is the value x_0 that satisfies $P(X \leq x_0) = p_0$ or $P(X \geq x_0) = p_0$ for a χ^2-tail area (probablity) p_0?
- **R Command**:

```
qchisq(p, df, lower.tail = TRUE)
 # p: the probability (area under the curve)
```

```
# df: degrees of freedom
# lower.tail:
#   TRUE (default) for left-tail area
#   FALSE for right-tail area
```

Example 3: find the 95-th percentile of the χ^2 distribution with 12 degrees of freedom.

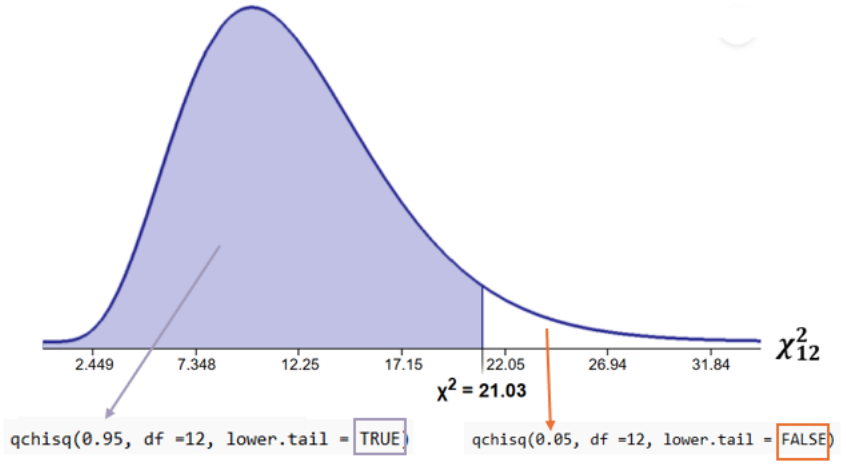
Answer: 95% percentile is the cut-off value that splits the region under the χ^2_{12} density curve into two regions. The left region has an area of 0.95, and the right region has an area of 0.05. This means we use either `qchi(0.95, df =12, lower.tail = TRUE)` or `qchi(0.05, df = 12, lower.tail = FALSE)` to find the same percentile in R. The code is in the following.

```
qchisq(0.95, df =12, lower.tail = TRUE)
```

```
## [1] 21.02607
```

```
qchisq(0.05, df =12, lower.tail = FALSE)
```

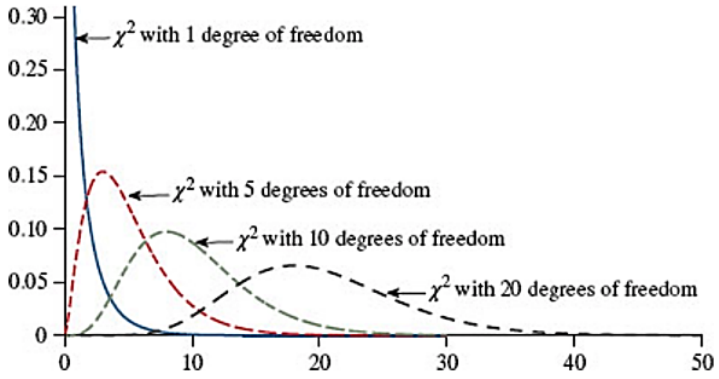
```
## [1] 21.02607
```



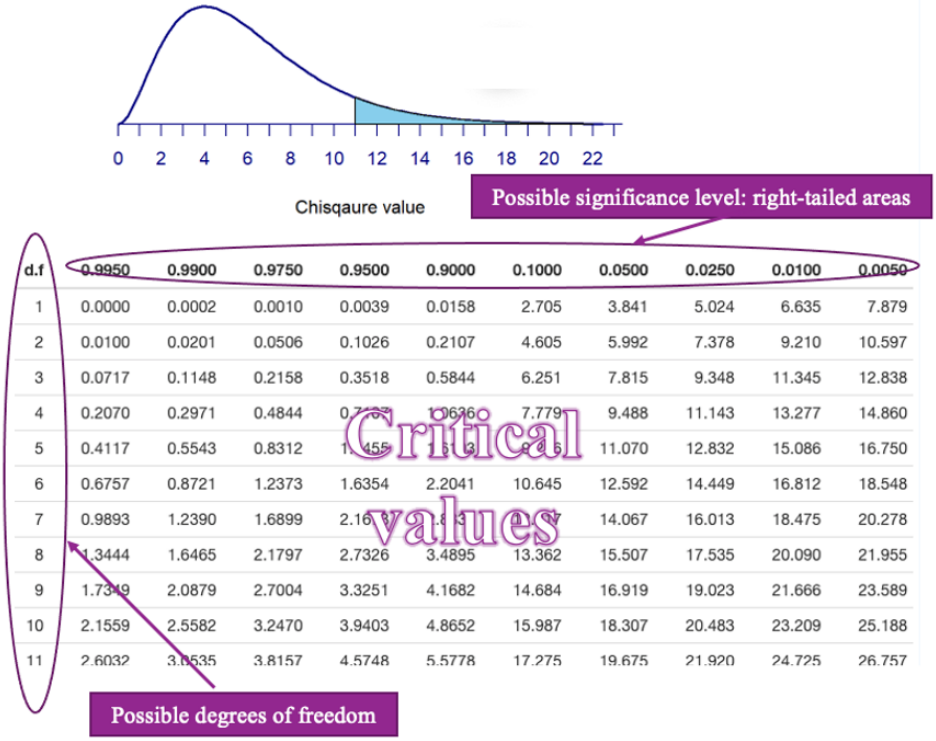
2.3 Chisquare Table

We can also use a chi-square table to find the critical value based on a given significance level. When using a table, we always use the critical value method. The p-value method is used when using a software program.

The chi-square distribution is used to characterize a positive random variable. Unlike normal and t distributions that have symmetric density curves, the chi-square distributions (dependent on the degrees of freedom) have skewed density curves.



We can find the critical value of the chi-square distribution from the chi-square table that is available on the course web page. The structure of the chi-square table is similar to the t-table.

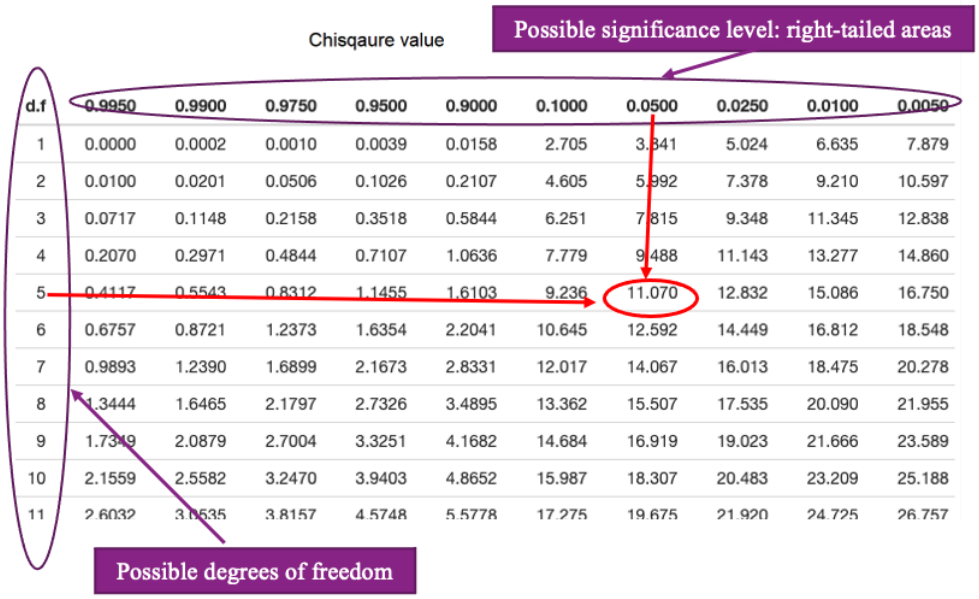


d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.210	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.2070	0.2971	0.4844	0.7	1.0636	7.779	9.488	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.455	1.6	9.2	11.070	12.832	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.16	2.8	12.0	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.090	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.920	24.725	26.757

The possible degrees of freedom are listed in the first column, the possible right-tail areas are listed in the top row, and the critical values are listed in the main body of the table.

The steps for finding the critical values are the same as those we followed for finding the t-critical values.

Example 2. Find the critical value of the chi-square distribution with 5 degrees of freedom with a significance level of 0.05.



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.210	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.017	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.090	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.920	24.725	26.757

The above figure shows how to find the critical value, denoted by $CV = \chi^2_{5,0.05} = 11.071$. The first subscript denotes 5 degrees of freedom, and the second subscript is the significance level of 0.05.

3 Chi-square Test of Goodness-of-fit

A goodness-of-fit test of a distribution is a testing procedure that justifies whether the null hypothesis that the specified distribution is correct, based on sample information.

For a single categorical variable, the null hypothesis should specify the cell probabilities. In other words, if the category has k categories, then $(p_1, p_2 \cdots, p_k)$ must be specified in the null hypothesis.

3.1 A Motivational Example

As a special case, we look at the following example of testing the proportion problem.

Example 1. We want to justify a claim that about 30% of WCU students are STEM majors. That is, we test the following hypotheses.

$$H_0: \quad p = 0.3 \quad v.s. \quad H_a: \quad p \neq 0.3.$$

We took a random sample of 100 students and recorded their majors, and found that 33 of them claimed a major in STEM. This means 67 of them are non-STEM majors. We have introduced a procedure to test the above hypotheses with the test statistic.

$$TS = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

That compares the claimed proportion with the sample proportion.

Note that the proportion of STEM majors contains the number of majors (frequencies) in both STEM and non-STEM disciplines. We can think about using (observed)sample frequencies and null (expected) frequencies to define the test statistic.

- **Under** H_0, we would **expect** to have 30 STEM majors and 70 non-STEM majors.

- We **observed** 33 STEM majors and 67 STEM majors in the random sample.

The above observed and expected numbers of STEM and non-STEM majors are summarized in the following table.

	Observed Values (from the sample)	Expected Value (under Ho)
STEM	$O_1 = 33$	$E_1 = 30$
non-STEM	$O_2 = 67$	$E_2 = 70$

In fact, a test statistic that measures the "distance" between the observed and expected frequency tables and has a χ^2 (chi-square) distribution is defined below

$$TS = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \to \chi^2_1.$$

The value of the test statistic in this example

$$TS = \frac{(33 - 30)^2}{30} + \frac{(67 - 70)^2}{70} = 9/30 + 49/70 = 3/10 + 7/10 = 1$$

With the above value of the test statistic, we can make a statistical decision on H_0 based on a given significance level.

3.2 Formulation of Chi-square Test of Goodness-of-fit

Let k be the number of categories of categorical variable Y. The category labels are $C_1, C_2, \cdots, C_k$. Let $P_1 = Pr(C_1), P_2 = Pr(C_2), \cdots, P_k = Pr(C_k)$. The null hypothesis claims that the categorical variable follows a specific distribution, and the alternative hypothesis claims that the categorical variable does NOT follow the distribution specified in the null hypothesis. That is,

$$H_0: \quad P_1 = p_1, P_2 = p_2, \cdots, P_k = p_k \quad v.s. \quad H_a: \quad \textit{the distribution in } H_0 \textit{ is not correct.}$$

The N is the sample size. We can then calculate the **expected cell frequency** of each category using formulas: $E_1 = N \times p_1, E_2 = N \times p_2, \cdots, E_k = N \times p_k$. The **observed cell frequency**, denoted by O_i (for $i = 1, 2, \cdots, k$), of each category is obtained from the data set. The **expected** and **observed** frequencies are summarized in the following table.

	Category 1	Category 1	⋯	Category (k-1)	Category k
Observed	O_1	O_2	...	O_{k-1}	O_k
Expected	E_1	E_2	...	E_{k-1}	E_k

The chi-square statistic is

$$G^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} \to \chi^2_{k-1}$$

A small G^2 indicates a lack of evidence for rejecting the null hypothesis. This implies that **the Pearson chi-square test of goodness is always right-tailed**. The degrees of freedom are always $(k-1)$ if the categorical factor variable has k levels.

Example 3. A gambler wants to test a die to determine whether it is fair. The gambler rolls a die that has six possible outcomes: 1, 2, 3, 4, 5, and 6; and the die is fair if each of these outcomes is equally likely. The gambler rolls the die 60 times and counts the number of times each number comes up. These counts, which are called the observed frequencies, are

Outcome	1	2	3	4	5	6
Observed	12	7	14	15	4	8

Solution: The null hypothesis is that the six-sided die is fair. That is equivalent to

$$H_0: \quad p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6 \quad v.s. \quad H_a: \quad The\ die\ is\ NOT\ fair.$$

Based on the observed frequency table, the size of the sample is 60. Using the *cell probabilities* in H_0, we have **expected frequencies** of the 6 categories to be equal to 10. We summarize the **expected** and **observed** frequencies in the following table.

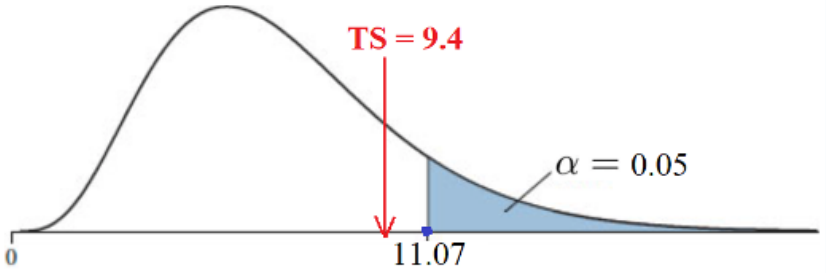
Outcome	1	2	3	4	5	6	Total
Observed	12	7	14	15	4	8	N=60
Expected	10	10	10	10	10	10	

$Np_i = E_i$ ⇨ $60 \times \frac{1}{6}$ | $60 \times \frac{1}{6}$ | $60 \times \frac{1}{6}$ | $60 \times \frac{1}{6}$ | $60 \times \frac{1}{6}$ | $60 \times \frac{1}{6}$

The test statistic for testing H_0 is defined by

$$G^2 = \frac{(12-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(14-10)^2}{10} + \frac{(15-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(8-10)^2}{10} = (4+9+16+25+36+4)/10 = 9.4$$

Since the categorical variable (number of dots on the faces of the 6-sided die) has 6 categories. The test statistic has 5 degrees of freedom. The critical value based on the significance level of 0.05 is given in the following figure.



Degrees of Freedom	Area in Right Tail									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278

The test statistic is NOT in the rejection region. We fail to reject the null hypothesis. The die is fair.

Example 4. Grade distribution: A statistics teacher claims that, on average, 20% of her students get a grade of A, 35% get a B, 25% get a C, 10% get a D, and 10% get an F. The grades of a random sample of 100 students were recorded. The following table presents the sample frequencies of each grade.

Grade	A	B	C	D	F
Observed	29	42	20	5	4

Solution: Based on the given information. $N = 100$. The null hypothesis and the alternative hypothesis are

$$H_0: \quad p_A = 0.2, p_B = 0.35, p_C = 0.25, p_D = 0.1, p_F = 0.1 \quad v.s. H_a: \quad \textit{The claimed distribution is wrong.}$$

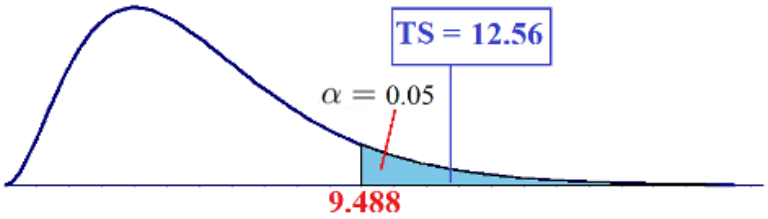
Under the null hypothesis, the expected table is given by

Grade	A	B	C	D	F
Expected	20	35	25	10	10

The test statistic has 4 degrees of freedom, and the value is calculated as follows.

$$G^2 = \frac{(29-20)^2}{20} + \frac{(42-35)^2}{35} + \frac{(20-25)^2}{25} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} = 12.56$$

The critical value with a significance level of 0.05 is given by



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832

Since the test statistic is inside the rejection region, we reject the null hypothesis. That is, the sample does not support the claimed grade distribution in H_0.

Remark: In the chi-square goodness-of-fit test, the crucial step is to find the *expected** frequency table under the null hypothesis.

The next YouTube video gives another example with detailed manual calculation.

4 Performing Goodness-of-fit Test in R

The two pieces of information required in the chi-square goodness-of-fit test are: observed frequency and the Hypothetical distribution under H_0. Next, we will use the formulas introduced earlier and built-in R

functions, respectively. For convenience, we use the following example to implement the χ^2 goodness-of-fit test.

Testing M&M Color Distribution: A company claims their M&Ms are distributed with these proportions:

Brown - 30%; Yellow - 20%; Red - 20%; Blue - 10%; Orange - 10%; Green - 10%

A random sample of 200 M&Ms was taken and observed:

Brown - 52; Yellow - 48; Red - 38; Blue - 22; Orange - 20; Green - 20

The null and alternative hypotheses are:

$$H_0 : p_1 = 0.3, \quad p_2 = 0.2, \quad p_3 = 0.2, \quad p_4 = 0.1, \quad p_5 = 0.1, \quad p_6 = 0.1$$

v.s.

at least one of the hypothetical probabilities is not true.

Remark: In general, χ^2 tests are all right-tailed. This means that the critical value is on the right tail of the chi-square density curve, and the p-value is defined as the tail area to the right of the test statistic.

Define R Data:

```
observed <- c(52, 48, 38, 22, 20, 20)
expected.props <- c(0.30, 0.20, 0.20, 0.10, 0.10, 0.10)
```

4.1 Formula Approoach

Recall that the χ^2 test statistic is given by

$$G^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} \to \chi^2_{k-1}.$$

To evaluate the test statistic G^2, we need to find the expected number of colors based on the data defined earlier.

```
# observed and expected.props were defined earlier
# total count
tot <- sum(observed)
expected.counts <- tot * expected.props   # this gives (52x0.30, 48x0.2, 38x0.2, ....)
                                          # and returns a vector of observed frequencies
## test statistics
G.sq <- sum ((observed - expected.counts)^2/expected.counts)  # sum() sum of resulting components
## degrees of freedom
df <- length(observed) - 1    # length() counts the number of cells in the vector
## all chi.sq test are right-tailed, using `lower.tail = FALSE`  to return the upper tail area.
p.value <- pchisq(G.sq, df =df, lower.tail = FALSE)   #
## combine G.sq and p-value using cbind()
cbind(G.sq = G.sq, p.value = p.value)
```

```
##          G.sq  p.value
## [1,] 2.966667 0.705125
```

Since the p-value is greater than 0.05, the null hypothesis is **not** rejected. That is, the observed frequency distribution is **not** significantly different from the hypothetical probability distribution.

4.2 Built-in Function Approach

The package {stats} coming with base R has a function chisq.test() to perform the chi-square test. The following single line of code returns the results of the χ^2 test.

```
## CAUTION: the second argument must be in the form of: p = hypothetical probabilities!
##          when performing goodness-of-fit tests. OTHERWISE, R will perform independence tests.
chisq.test(observed, p = expected.props)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 2.9667, df = 5, p-value = 0.7051
```

5 Practice Exercises

College Sports: A University conducted a survey of its recent graduates to collect demographic and health information for future planning purposes as well as to assess students' satisfaction with their undergraduate experiences. The survey revealed that a substantial proportion of students were not engaging in regular exercise, many felt their nutrition was poor, and a substantial number were smoking. In response to a question on regular exercise, 60% of all graduates reported getting no regular exercise, 25% reported exercising sporadically, and 15% reported exercising regularly as undergraduates. The next year, the University launched a health promotion campaign on campus in an attempt to increase health behaviors among undergraduates. The program included modules on exercise, nutrition, and smoking cessation. To evaluate the impact of the program, the University again surveyed graduates and asked the same questions. The survey was completed by 470 graduates, and the following data were collected on the exercise question:

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Number of Students	255	125	90	470

We specifically want to compare the distribution of responses in the sample to the distribution reported the previous year (i.e., 60%, 25%, 15% reporting no, sporadic, and regular exercise, respectively). Whether the data supports the above distribution at a significance level of 0.05.