Introductory Statistics

Tatiana Tyukina
tt51@leicester.ac.uk

Topic 4 - Goodness-of-Fit

- **Topic 0: Introduction**
- **Topic 1: Descriptive Statistics**
- **Topic 2: Estimators: Point estimation, Interval estimation**
  2.1 **Point estimation,**
  2.2 **Properties of estimators,**
  2.3 **Interval estimation,**
  2.4 **Inference based on Normal Distribution**
- **Topic 3: Hypothesis Testing**
  3.1 **Inference for two populations**
  3.2 **Hypothesis Testing Procedure**
- Topic 4: Goodness-of-Fit: The χ^2 test
- Topic 5: Bayesian Estimation

Hypothesis Testing: idea

The idea behind the chi-square goodness-of-fit test is

- to state a hypothesis about the distribution of data

HYPOTHESIS TESTING: IDEA

The idea behind the chi-square goodness-of-fit test is

- to state a hypothesis about the distribution of data
- to compare the observed frequencies of data to the frequencies that would be expected if the population has the hypothesised distribution

Hypothesis Testing: idea

The idea behind the chi-square goodness-of-fit test is

- to state a hypothesis about the distribution of data
- to compare the observed frequencies of data to the frequencies that would be expected if the population has the hypothesised distribution
- If the observed and expected frequencies match fairly well, we do not reject the null hypothesis; otherwise, we reject the null hypothesis.

GOODNESS-OF-FIT

DEFINITION

Suppose that we have outcomes of a multinomial experiment that consists of k mutually exclusive and exhaustive events $A_1, ..., A_k$. Let $p_i = P(A_i)$, $i = 1, 2, ..., k$ $(\sum_{i=1}^{k} p_i = 1)$.
Let the experiment be repeated n times, and let X_i ($i = 1, 2, ..., k$) represent the number of times the event A_i occurs $(\sum_{i=1}^{k} Xi = n)$. Then $(X_1, ..., X_k)$ have a ***multinomial distribution*** *with parameters n, $p_1, ..., p_k$.*

GOODNESS-OF-FIT

DEFINITION

Suppose that we have outcomes of a multinomial experiment that consists of k mutually exclusive and exhaustive events $A_1, ..., A_k$. Let $p_i = P(A_i)$, $i = 1, 2, ..., k$ $(\sum_{i=1}^{k} p_i = 1)$.
*Let the experiment be repeated n times, and let X_i ($i = 1, 2, ..., k$) represent the number of times the event A_i occurs $(\sum_{i=1}^{k} Xi = n)$. Then $(X_1, ..., X_k)$ have a **multinomial distribution** with parameters n, $p_1, ..., p_k$.*

EXAMPLE

GOODNESS-OF-FIT

DEFINITION

Suppose that we have outcomes of a multinomial experiment that consists of k mutually exclusive and exhaustive events $A_1, ..., A_k$. Let $p_i = P(A_i)$, $i = 1, 2, ..., k$ $(\sum_{i=1}^{k} p_i = 1)$.
Let the experiment be repeated n times, and let X_i ($i = 1, 2, ..., k$) represent the number of times the event A_i occurs $(\sum_{i=1}^{k} Xi = n)$. Then $(X_1, ..., X_k)$ have a ***multinomial distribution*** *with parameters n, $p_1, ..., p_k$.*

EXAMPLE

- Rolling of a 6-sided die: $k = 6, A_1 = 1, A_2 = 2, \ldots A_6 = 6,$
  $p_i = P(A_i = i) = 1/6$ for each $i = \overline{1, 6}$.
  The random sample of $n = 20$ throws can be
  $X_1 = 2, X_2 = 3, X_3 = 3, X_4 = 7, X_5 = 2, X_6 = 3.$

GOODNESS-OF-FIT

DEFINITION

Suppose that we have outcomes of a multinomial experiment that consists of k mutually exclusive and exhaustive events $A_1, ..., A_k$. Let $p_i = P(A_i)$, $i = 1, 2, ..., k$ $(\sum_{i=1}^{k} p_i = 1)$.
Let the experiment be repeated n times, and let X_i ($i = 1, 2, ..., k$) represent the number of times the event A_i occurs $(\sum_{i=1}^{k} Xi = n)$. Then $(X_1, ..., X_k)$ have a ***multinomial distribution*** *with parameters n, $p_1, ..., p_k$.*

EXAMPLE

- Rolling of a 6-sided die: $k = 6, A_1 = 1, A_2 = 2, ... A_6 = 6,$
  $p_i = P(A_i = i) = 1/6$ for each $i = \overline{1, 6}$.
  The random sample of $n = 20$ throws can be
  $X_1 = 2, X_2 = 3, X_3 = 3, X_4 = 7, X_5 = 2, X_6 = 3.$
- Suppose that in a three-way election, candidate A received 20% of the votes, candidate B – 30% of the votes, and candidate C – 50% of the votes.
  The events: $k = 3, A_1 = A, A_2 = B, A_3 = C, p_1 = P(A_1) = 0.2,$
  $p_2 = P(A_2) = 0.3$ and $p_3 = P(A_3) = 0.5$.

Goodness-of-Fit

Theorem

Suppose that a random sample of n observations is taken from $f_X(x)$ [or $p_X(x)$], a pdf having t unknown parameters.
Let $r_1, r_2, ..., r_k$ be a set of mutually exclusive ranges (or qualitative outcomes) associated with each of the n observations.
*Let $\hat{p}_i$ = **estimated probability** of r_i , $i = 1, 2, ..., k$ (as calculated from $f_X(x)$ [or $p_X(x)$] after the pdf's t unknown parameters have been replaced by their maximum likelihood estimates).*
Let X_i denote the number of times that r_i occurs, $i = 1, 2, ..., k$.
Then the random variable

$$G = \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

has approximately a χ^2 distribution with $k - 1 - t$ degrees of freedom.

GOODNESS-OF-FIT

PROOF.

Consider a special case $k = 2$: assume, $A_1 = 1, A_2 = 0$, and $p_1 = P(A_1)$ is known. In this case, for the sample of size n, $X_1 = x_1$ and $X_2 = x_2$ being observed frequencies, we can write that $x_2 = n - x_1$ and $p_2 = 1 - p_1$, so the statistic is

$$G = \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2} = \frac{(x_1 - np_1)^2}{np_1} + \frac{(n - x_1 - n(1-p_1))^2}{n(1-p_1)}$$

$$G = \frac{(x_1 - np_1)^2}{np_1(1-p_1)} = \left(\frac{x_1/n - p_1}{\sqrt{p_1(1-p_1)/n}}\right)^2$$

The sampling distribution of $X_1 \sim Bin(n, p_1)$, so as $n \to \infty$ the sampling distribution of $\frac{x_1}{n} \to N(p_1, \frac{p_1(1-p_1)}{n})$.
This means the sampling distribution of

$$Y = \frac{x_1/n - p_1}{\sqrt{p_1(1-p_1)/n}} \sim N(0,1)$$

The distribution of our statistic $G = Y^2 \sim \chi^2_{2-1-0}$ for $n \to \infty$. $\square$

Goodness-of-Fit

Definition

The statistic G defined as

$$G = \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

is called the ***chi-squared statistic****.*
The statistic G is a measure of how close our observed frequencies come to the expected frequencies and is referred to as a measure of ***goodness-of-fit****.*
Smaller values of G indicate better fit.

$$\boxed{G = \sum_{i=1}^{k} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}}$$

Goodness-of-Fit: Hypothesis Testing

For hypothesis test at the α level of significance:

- the null hypothesis is usually stated as "cell probabilities" :

$$H0 : p_1 = p_{1_0}, p_2 = p_{2_0}, ..., p_k = p_{k_0},$$

where p_{i_0} denotes a specified value for p_i.

GOODNESS-OF-FIT: HYPOTHESIS TESTING

For hypothesis test at the α level of significance:

- the null hypothesis is usually stated as "cell probabilities" :

$$H0 : p_1 = p_{1_0}, p_2 = p_{2_0}, ..., p_k = p_{k_0},$$

  where p_{i_0} denotes a specified value for p_i.

- the alternative hypothesis is the general one that states that at least one of the equalities does not hold:

$$HA : p_j \neq p_{j_0}$$

Goodness-of-Fit: Hypothesis Testing

For hypothesis test at the α level of significance:

- the null hypothesis is usually stated as "cell probabilities" :

$$H0 : p_1 = p_{1_0}, p_2 = p_{2_0}, ..., p_k = p_{k_0},$$

  where p_{i_0} denotes a specified value for p_i.

- the alternative hypothesis is the general one that states that at least one of the equalities does not hold:

$$HA : p_j \neq p_{j_0}$$

- calculate the statistics

$$g = \sum_{i=1}^{k} \frac{(x_i - n\hat{p}_{i_0})^2}{n\hat{p}_{i_0}}$$

  where $x_1, x_2, ..., x_k$ are the observed frequencies of $r_1, r_2, ..., r_k$, respectively, and $n\hat{p}_{1_0}, n\hat{p}_{2_0}, ..., n\hat{p}_{k_0}$ are the corresponding estimated expected frequencies based on the null hypothesis. (The r_i's should be defined so that $n\hat{p}_{i_0} \geq 5$ for all i.)

Goodness-of-Fit: Hypothesis Testing

For hypothesis test at the α level of significance:

- the null hypothesis is usually stated as "cell probabilities" :

$$H0 : p_1 = p_{1_0}, p_2 = p_{2_0}, ..., p_k = p_{k_0},$$

  where p_{i_0} denotes a specified value for p_i.

- the alternative hypothesis is the general one that states that at least one of the equalities does not hold:

$$HA : p_j \neq p_{j_0}$$

- calculate the statistics

$$g = \sum_{i=1}^{k} \frac{(x_i - n\hat{p}_{i_0})^2}{n\hat{p}_{i_0}}$$

  where $x_1, x_2, ..., x_k$ are the observed frequencies of $r_1, r_2, ..., r_k$, respectively, and $n\hat{p}_{1_0}, n\hat{p}_{2_0}, ..., n\hat{p}_{k_0}$ are the corresponding estimated expected frequencies based on the null hypothesis. (The r_i's should be defined so that $n\hat{p}_{i_0} \geq 5$ for all i.)

- If $g \geq \chi^2_{1-\alpha,k-1-t}$ $H0$ should be rejected.

GOODNESS-OF-FIT: EXAMPLE

EXAMPLE

Do you hate Mondays?
Researchers in Germany concluded that the risk of heart attack on a Monday for a working person may be as much as 50% greater than on any other day.
The researchers kept track of heart attacks and coronary arrests over a period of 5 years among $330,000$ people who lived near Augsberg, Germany.
In an attempt to verify the researcher's claim, 200 working people who had recently had heart attacks were surveyed.

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
24	36	27	26	32	26	29

Do these data present sufficient evidence to indicate that there is a difference in the percentages of heart attacks that occur on different days of the week? Test using $\alpha = 0.05$.

Goodness-of-Fit: Example

Example

Solution:
The events $A_1 = 1$ (Sunday), $A_2 = 2$ (Monday),..., $A_7 = 7$ (Saturday), $k = 7$.
The sample size $n = 200$, observed X_i are listed in the table.
The hypotheses:

$$H0 : p_1 = p_2 = ... = p_7 = 1/7$$

$$HA : p_2 > 1/7$$

Assumption $np_i \geq 5$ is satisfied.
The goodness-of-fit statistic is:

$$g = \frac{(24 - 200/7)^2}{200/7} + \frac{(36 - 200/7)^2}{200/7} + ... + \frac{(29 - 200/7)^2}{200/7} = 3.63$$

The critical value for $\chi^2_{0.95, 7-1-0} = 12.59$.
Hence, $H0$ cannot be rejected.

Goodness-of-Fit: for Quantitative RV

To test the hypothesis:

$H0$: The given data follow a specific probability distribution F_0

versus

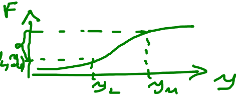
HA : The data do not follow the specified probability distribution

Specify ranges as $r_i = (Y_{i_L}, Y_{i_U}]$,where $i = 1, 2, ..., k$ is the number of classes where Y_{i_L} and Y_{i_U} the lower limit upper limits of class i, respectively;
Then the goodness-of-fit statistics is defined as

$$G = \sum_{i=1}^{k} \frac{(X_i - nE_i)^2}{nE_i},$$

where X_i is the ith observed outcome frequency (in class i), E_i is the ith expected (theoretical) *relative* frequency calculated by,

$$E_i = F_0(Y_{i_U}) - F_0(Y_{i_L}),$$ $E_i = P(y_L, y_u)$



where F_0 is the cumulative probability distribution that is being tested (assumed) to determine if the given data follows (fits) this probability distribution, and n is the sample size.

Goodness-of-Fit: Example

Example

Question 1 from MockTest 1

Data was collected for the mileage ratings of 40 cars of a new car model determined for an environmental survey. The frequency distribution is presented in the table:

$(0, 32]$	$(32, 34]$	$(34, 36]$	$(36, 38]$	$(38, 40]$	$(40, 42]$	$(42, \infty)$
3	6	8	9	8	4	2

Test the hypothesis that the distribution is normal $N(\mu, 10)$, for $\alpha = 0.05$

GOODNESS-OF-FIT: EXAMPLE

EXAMPLE

Question 1 from MockTest 1

Data was collected for the mileage ratings of 40 cars of a new car model determined for an environmental survey. The frequency distribution is presented in the table:

$(0, 32]$	$(32, 34]$	$(34, 36]$	$(36, 38]$	$(38, 40]$	$(40, 42]$	$(42, \infty)$
3	6	8	9	8	4	2

Test the hypothesis that the distribution is normal $N(\mu, 10)$, for $\alpha = 0.05$

Solution: As we need to test for $N(\mu, 10)$, where μ is unknown, then according to the theorem the value of μ for the null hypothesis can be find using the MLE for μ.

We can show that MLE for μ for population with normal distribution is $\bar{X}$, hence, as $\bar{x} = 36.765$:

$$H0: \; F_0 \text{ distribution is } N(36.765, 10)$$

$$HA: \; F_0 \text{ distribution is not } \; N(36.765, 10)$$

continued ...

Goodness-of-Fit: Example

Example

Using the formula for cumulative normal distribution:

$$F_0(X < x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt$$

or using software, for example, in MATLAB, *normcdf(x, mu,sigma)* we can find expected frequencies for each interval and estimate G statistics.
The information is collected in the table:

continued ...

GOODNESS-OF-FIT: EXAMPLE

EXAMPLE

Using the formula for cumulative normal distribution: $N(36.8, 10)$

$$F_0(X < x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

or using software, for example, in MATLAB, *normcdf(x, mu,sigma)* we can find expected frequencies for each interval and estimate G statistics.
The information is collected in the table:

$(Y_l, Y_U]$	$(0, 34]$	$(34, 36]$	$(36, 38]$	$(38, 40]$	$(40, \infty)$
X_i	9	8	9	8	6
$F_0(Y_L)$	0.0000	0.1910	0.4044	0.6519	0.8468
$F_0(Y_U)$	0.1910	0.4044	0.6519	0.8468	1.0000
E_i	0.1910	0.2135	0.2475	0.1949	0.1532
nE_i	7.64	8.5386	9.9003	7.7965	6.128

$n = 40$

continued ...

GOODNESS-OF-FIT: EXAMPLE

EXAMPLE

$(Y_l, Y_U]$	$(0, 34]$	$(34, 36]$	$(36, 38]$	$(38, 40]$	$(40, \infty)$
X_i	9	8	9	8	6
$F_0(Y_L)$	0.0000	0.1910	0.4044	0.6519	0.8468
$F_0(Y_U)$	0.1910	0.4044	0.6519	0.8468	1.0000
E_i	0.1910	0.2135	0.2475	0.1949	0.1533
nE_i	7.64	8.5386	9.9003	7.7965	6.128

For the goodness-of-fit statistics:

$$G = \sum_{i=1}^{k} \frac{(X_i - nE_i)^2}{nE_i},$$

Goodness-of-Fit: Example

Example

$(Y_l, Y_U]$	$(0, 34]$	$(34, 36]$	$(36, 38]$	$(38, 40]$	$(40, \infty)$
X_i	9	8	9	8	6
$F_0(Y_L)$	0.0000	0.1910	0.4044	0.6519	0.8468
$F_0(Y_U)$	0.1910	0.4044	0.6519	0.8468	1.0000
E_i	0.1910	0.2135	0.2475	0.1949	0.1533
nE_i	7.64	8.5386	9.9003	7.7965	6.128

For the goodness-of-fit statistics:

$$G = \sum_{i=1}^{k} \frac{(X_i - nE_i)^2}{nE_i},$$

the observed $g = 0.374$, the corresponding $\chi^2_{crit} = \chi^2_{1-0.05,5-1-1} = 7.81$, therefore, we cannot reject the null hypothesis, and so it is likely that the sample was obtained from population with normal distribution. N(36.8, 10)

GOODNESS-OF-FIT: EXERCISE

Exercises for practice:

1. The speeds of vehicles (in mph) passing through a section of *Highway 75* are recorded for a random sample of 150 vehicles and are given below. Test the hypothesis that the speeds are normally distributed with a mean of 70 and a standard deviation of 4. Use $\alpha = 0.01$.

Range	$40-55$	$56-65$	$66-75$	$76-85$	> 85
Number	12	14	78	40	6

2. Based on the sample data of 50 days contained in the following table, test the hypothesis that the daily mean temperatures in the City of Tampa are normally distributed with mean 77 and variance 6. Use $\alpha = 5\%$.

Temperature	$46-55$	$56-65$	$66-75$	$76-85$	$86-95$
Number of days	4	6	13	23	4

QUESTIONS TO TAKE HOME

1. What is meant by saying that a variable has a chi-square distribution?

Questions to take home

1. What is meant by saying that a variable has a chi-square distribution?
2. How do you identify different chi-square distributions?

QUESTIONS TO TAKE HOME

1. What is meant by saying that a variable has a chi-square distribution?
2. How do you identify different chi-square distributions?
3. What is the difference between the observed frequencies and the expected frequencies?

Questions to take home

1. What is meant by saying that a variable has a chi-square distribution?
2. How do you identify different chi-square distributions?
3. What is the difference between the observed frequencies and the expected frequencies?
4. How do you define the observed/expected frequencies for nominal data?

Questions to take home

1. What is meant by saying that a variable has a chi-square distribution?
2. How do you identify different chi-square distributions?
3. What is the difference between the observed frequencies and the expected frequencies?
4. How do you define the observed/expected frequencies for nominal data?
5. How do you define the observed/expected frequencies for quantitative data?