Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

November 14, 2019

Topics

Objectives

Objectives:

- Carry out a chi-squared goodness of fit test.

Goodness of Fit Test

Introduction

- Recall that a ***binomial experiment*** involves n **trials**, each of which results in one of **two possible outcomes** (*success* or *failure*).

Goodness of Fit Test

Introduction

- Recall that a ***binomial experiment*** involves n **trials**, each of which results in one of **two possible outcomes** (*success* or *failure*).
- A ***multinomial experiment*** involves n **trials**, each of which results in one of $k > 2$ **possible outcomes** (or ***categories***).

Example

Suppose an unbalanced tetrahedral (4-sided) die lands on **1**, **2**, **3**, and **4** with probabilities $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$, and $p_4 = 0.4$.

Example

Suppose an unbalanced tetrahedral (4-sided) die lands on **1**, **2**, **3**, and **4** with probabilities $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$, and $p_4 = 0.4$.

If we roll the die $n = 10$ times and count how many times it lands on each side, that's a **multinomial experiment**.

- **Notation**:

$$\begin{aligned} \boldsymbol{n} &= \text{The number of trials (sample size).} \\ \boldsymbol{n_1, n_2, \ldots, n_k} &= \text{The } \underline{\textbf{\textit{observed counts}}} \text{ for the } k \\ &\quad \text{categories (or } \underline{\textbf{\textit{cells}}}\text{).} \\ \boldsymbol{p_1, p_2, \ldots, p_k} &= \text{The probabilities for the } k \text{ categories.} \end{aligned}$$

Example

If we roll the tetrahedral die described earlier $n = 10$ times, a one possible set of outcomes is:

	Outcome 1	2	3	4	Row Total
Observed Count	$n_1 = 2$	$n_2 = 2$	$n_3 = 3$	$n_4 = 3$	$n = 10$

(i.e. it landed on **1 twice**, on **2 twice**, on **3 three times**, and on **4 three times**).

- **Comment**: A **multinomial experiment** might consist of drawing a **sample** of size n from a **population** whose individuals each belong to one of k **categories.**

  In this case, $p_1, p_2, \ldots, p_k$ are the **population proportions** for the categories.

Example

A 1973 study at UCLA examined the ages of a sample of $n = 66$ individuals selected for grand juries in Alameda County, CA.

Example

A 1973 study at UCLA examined the ages of a sample of $n = 66$ individuals selected for grand juries in Alameda County, CA.

Each individual in the sample fell into one of $k = 4$ **age categories**:

21 to 40 years
41 to 50 years
51 to 60 years
61 years and older

Counting how many individuals fall into each category makes it a **multinomial experiment.**

The observed the **observed counts** n_1, n_2, n_3, and n_4 are:

	Age Category				Row
	21 to 40	41 to 50	51 to 60	61 and older	**Total**
Observed Count	$n_1 = 5$	$n_2 = 9$	$n_3 = 19$	$n_4 = 33$	$n = 66$

- Note:

$$\sum_i n_i = n \qquad \text{and} \qquad \sum_i p_i = 1.$$

- The **null hypothesis** is that the probabilities for the k categories are equal to some hypothesized values:

Null Hypothesis:

$$H_0: \quad p_1 = p_{10}, \quad p_2 = p_{20}, \quad \ldots, \quad p_k = p_{k0}$$

where $\boldsymbol{p_{10}, p_{20}, \ldots, p_{k0}}$ are hypothesized values for the true (unknown) probabilities $p_1, p_2, \ldots, p_k$.

- The **alternative hypothesis** is that *at least one probability* differs from its hypothesized value:

Alternative Hypothesis: The alternative hypothesis will be

$$H_a : \ p_i \neq p_{i0} \qquad \text{for at least one } i.$$

- To carry out the *goodness of fit test*, we'll need a new probability distribution.

Chi-Squared Distributions

- The ***chi-squared distribution*** has a single parameter ν, called the ***degrees of freedom***.

Chi-Squared Distributions

- The ***chi-squared distribution*** has a single parameter ν, called the ***degrees of freedom***.

  We denote the chi-squared distribution with ν **df** by $\chi^2(\nu)$.

- **Properties of Chi-Squared Distributions**:

- **Properties of Chi-Squared Distributions**:

  1. They're right skewed and lie entirely to the right of 0.

- **Properties of Chi-Squared Distributions**:

  1. They're right skewed and lie entirely to the right of 0.
  2. The one (and only) parameter of the distribution, its degrees of freedom, controls the distribution's shape, center, and spread.

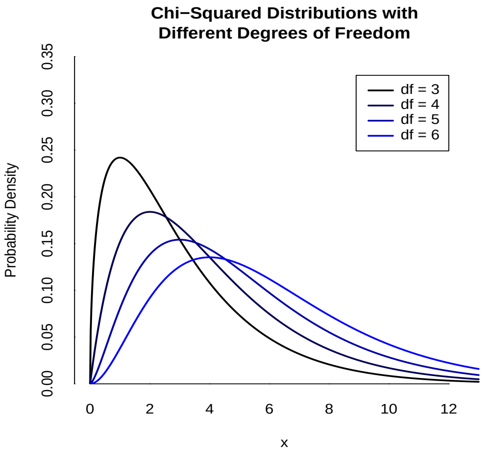

Figure: $\chi^2(\nu)$ distributions with different values of ν (df).

- If H_0 was true, we'd expect the sample proportions for the categories to be approximately equal to the null-hypothesized probabilities, i.e. we'd expect

$$\frac{n_i}{n} \approx p_{i0} ,$$

or equivalently, we'd expect the ***observed counts*** $\boldsymbol{n_i}$ to be

$$n_i \approx n \, p_{i0} .$$

- If H_0 was true, we'd expect the sample proportions for the categories to be approximately equal to the null-hypothesized probabilities, i.e. we'd expect

$$\frac{n_i}{n} \approx p_{i0},$$

or equivalently, we'd expect the ***observed counts*** $\boldsymbol{n_i}$ to be

$$n_i \approx n\, p_{i0}.$$

The right side above ($\boldsymbol{n\, p_{i0}}$) is called the ***expected count*** (under $\boldsymbol{H_0}$).

χ^2 Goodness of Fit Test Statistic:

$$\chi^2 = \sum_i \frac{(n_i - n\,p_{i0})^2}{n\,p_{i0}} = \sum_i \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

- The numerator of χ^2 will be large if the **observed** counts differ substantially from the counts **expected** under H_0, so ...

- The numerator of χ^2 will be large if the **observed** counts differ substantially from the counts **expected** under H_0, so ...

Large values of χ^2 provide **evidence against H_0 in favor of $H_a : p_i \neq p_{i0}$ for at least one i.**

Sampling Distribution of the Test Statistic Under $\boldsymbol{H_0}$:
If χ^2 is the goodness of fit test statistic, and the number of trials $\boldsymbol{n}$ in the *multinomial experiment* is ***large***, then when H_0 is true,

$$\chi^2 \sim \chi^2(k-1),$$

where $\boldsymbol{k}$ is the number possible outcomes (categories) for each trial.

- The **sample size** $\boldsymbol{n}$ is considered **large enough** as long as each of the **expected counts** is **five** (or higher).

- The $\chi^2(k-1)$ curve gives us:

- The $\chi^2(k-1)$ curve gives us:
  - The ***rejection region*** as the **extreme largest 100α% of χ^2 values.**

- The $\chi^2(k-1)$ curve gives us:
  - The ***rejection region*** as the **extreme largest 100α% of χ^2 values**.
  - The ***p-value*** as the **tail area to the right of the observed χ^2 value**.

- **Comment**: The **df** is $\boldsymbol{k-1}$ because only $k-1$ of the deviations $n_i - np_{i0}$ used to compute χ^2 are "free to vary" since the they sum to zero:

$$\sum_i n_i - np_{i0} \; = \; \sum_i n_i \; - \; n \sum_i p_{i0} \; = \; n \; - \; n \; = \; 0 \, .$$

Example

Refer to the study of ages of people selected for grand juries.

The Public Health Department listed the following age demographics for the general population in Alameda County, CA.

Age Category	Percentage of General Population
21 to 40	**42%**
41 to 50	**23%**
51 to 60	**16%**
61 and over	**19%**

If the $n = 66$ individuals selected for grand juries were a random sample from the population, they'd fall into the $k = 4$ **age categories** with **probabilities** given by the **percentages** above.

If the $n = 66$ individuals selected for grand juries were a random sample from the population, they'd fall into the $k = 4$ **age categories** with **probabilities** given by the **percentages** above.

We'll use the **observed counts** (from the earlier example) to test this claim with a **goodness of fit test.**

The **hypotheses** are

$$H_0: \quad p_1 = 0.42, \quad p_2 = 0.23, \quad p_3 = 0.16, \quad p_4 = 0.19$$

versus

$$H_a: \text{Not all } p_i\text{'s equal their null-hypothesized values.}$$

where $\boldsymbol{p_1, p_2, \ldots, p_k}$ are the true (unknown) probabilities of a **grand jury selectee** falling into the **four age categories.**

The **expected counts** (under H_0) are below next to the **observed counts**:

	Age Category				Row
	21 to 40	41 to 50	51 to 60	61 and older	**Total**
Observed Count	$n_1 = 5$	$n_2 = 9$	$n_3 = 19$	$n_4 = 33$	$n = 66$
Expected Count	$np_{10} = 27.7$	$np_{20} = 15.2$	$np_{30} = 10.6$	$np_{40} = 12.5$	$n = 66$

Is the **sample size** $n = 66$ is **large enough** to justify the use of the **goodness of fit test**?

The **expected counts** (under H_0) are below next to the **observed counts**:

	Age Category				Row
	21 to 40	41 to 50	51 to 60	61 and older	**Total**
Observed Count	$n_1 = 5$	$n_2 = 9$	$n_3 = 19$	$n_4 = 33$	$n = 66$
Expected Count	$np_{10} = 27.7$	$np_{20} = 15.2$	$np_{30} = 10.6$	$np_{40} = 12.5$	$n = 66$

Is the **sample size** $n = 66$ is **large enough** to justify the use of the **goodness of fit test**?

Answer: **Yes.** (Why?)

The **goodness of fit test statistic** is

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{k} \frac{(n_i - n \cdot p_{i0})^2}{n \cdot p_{i0}} \\
&= \frac{(5 - 27.72)^2}{27.72} + \frac{(9 - 15.18)^2}{15.18} \\
&\quad + \frac{(19 - 10.56)^2}{10.56} + \frac{(33 - 12.54)^2}{12.54} \\
&= \mathbf{61.27}.
\end{aligned}
$$

From a χ^2 distribution table using $k - 1 = 3$ **df**, the **p-value** is **less than 0.001**.

From a χ^2 distribution table using $\boldsymbol{k-1=3}$ **df**, the **p-value** is **less than 0.001**.

Using a level of significance $\alpha = 0.05$, we **reject** $\boldsymbol{H_0}$.

From a χ^2 distribution table using $\boldsymbol{k-1=3}$ **df**, the **p-value** is **less than 0.001**.

Using a level of significance $\alpha = 0.05$, we **reject** $\boldsymbol{H_0}$.

Grand jurors in Alameda County follow have a different **age distribution** than the general population.

Testing Whether a Sample Came From a Specific Distribution

- Consider an iid sample $X_1, X_2, \ldots, X_n$ from an **unknown** probability distribution whose **pdf** is $f(x)$.

Testing Whether a Sample Came From a Specific Distribution

- Consider an iid sample $X_1, X_2, \ldots, X_n$ from an **unknown** probability distribution whose **pdf** is $\boldsymbol{f(x)}$.

  We want to test

$$\begin{aligned} H_0 : f(x) &= f_0(x) \\ H_a : f(x) &\neq f_0(x) \end{aligned}$$

  where $\boldsymbol{f_0(x)}$ is the **pdf** of some **hypothesize** probability distribution.

- To carry out the test, we **subdivide** the **domain** (or ***support***) of $\boldsymbol{f_0(x)}$ into $\boldsymbol{k}$ intervals of the form

$$[a_0, a_1),\ \ [a_1, a_2),\ \ \ldots,\ \ [a_{k-1}, a_k)\,.$$

- To carry out the test, we **subdivide** the **domain** (or ***support***) of $\boldsymbol{f_0(x)}$ into $\boldsymbol{k}$ intervals of the form

$$[a_0, a_1),\ \ [a_1, a_2),\ \ \ldots,\ \ [a_{k-1}, a_k)\,.$$

Then **under H_0**, the **probability** of each X falling into the **ith interval** is

$$\boldsymbol{p_{i0}} \ = \ \int_{a_{i-1}}^{a_i} f_0(x)\,dx\,.$$

- We obtain the **observed counts** $\boldsymbol{n_1, n_2, \ldots, n_k}$ for each of the $\boldsymbol{k}$ **intervals**, and use them in **a goodness of fit test** of

$$H_0 : \quad p_1 = p_{10}, \quad p_2 = p_{20}, \quad \ldots, \quad p_k = p_{k0}$$

$$H_a : \; p_i \neq p_{i0} \qquad \text{for at least one } i$$

where each $\boldsymbol{p_i}$ is the true (unknown) **probability** of an X falling into the i**th interval**, i.e.

$$\boldsymbol{p_i} \;=\; \int_{a_{i-1}}^{a_i} f(x)\, dx\,.$$

Example

One way to test whether computer-generated ”pseudo random numbers” truly do follow the specified probability distribution is to carry out a **goodness of fit test.**

Example

One way to test whether computer-generated ”pseudo random numbers” truly do follow the specified probability distribution is to carry out a **goodness of fit test.**

A sample of $n = 1000$ ”pseudo random numbers” was generated from a standard normal distribution using R.

Example

One way to test whether computer-generated "pseudo random numbers" truly do follow the specified probability distribution is to carry out a **goodness of fit test.**

A sample of $n = 1000$ "pseudo random numbers" was generated from a standard normal distribution using R.

We'll use them to test:

$$\begin{aligned} H_0 : f(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ H_a : f(x) &\neq \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \end{aligned}$$

where $f(x)$ is the true (unknown) distribution of the "pseudo random numbers."

The observations were grouped into $\boldsymbol{k = 10}$ equi-probable **intervals** under the N$(0, 1)$ distribution.

The observations were grouped into $k = 10$ equi-probable **intervals** under the N$(0, 1)$ distribution.

The **observed counts**, theoretical probabilities, and **expected counts** are on the next slide.

Interval	N(0, 1) Probability p_{i0}	Expected Count $n\,p_{i0}$	Observed Count n_i
$-\infty$ to -1.28	0.1	$1000 \cdot 0.1 = 100$	104
-1.28 to -0.84	0.1	$1000 \cdot 0.1 = 100$	89
-0.84 to -0.52	0.1	$1000 \cdot 0.1 = 100$	105
-0.52 to -0.25	0.1	$1000 \cdot 0.1 = 100$	113
-0.25 to 0.00	0.1	$1000 \cdot 0.1 = 100$	90
0.00 to 0.25	0.1	$1000 \cdot 0.1 = 100$	84
0.25 to 0.52	0.1	$1000 \cdot 0.1 = 100$	96
0.52 to 0.84	0.1	$1000 \cdot 0.1 = 100$	110
0.84 to 1.28	0.1	$1000 \cdot 0.1 = 100$	122
1.28 to ∞	0.1	$1000 \cdot 0.1 = 100$	87
			$n = 1000$

We'll pose the problem as **goodness of fit test** of

$$H_0: \quad p_1 = 0.1, \quad p_2 = 0.1, \quad \ldots, \quad p_{10} = 0.1$$

versus

$$H_a: \ p_i \neq 0.1 \qquad \text{for at least one } i$$

where the $\boldsymbol{p_i}$'s are the true (unknown) probabilities of an R "pseudo random number" falling into the **ith interval.**

The **test statistic** is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\,p_{i0})^2}{n\,p_{i0}}$$
$$= \frac{(104-100)^2}{100} + \frac{(89-100)^2}{100} + \frac{(105-100)^2}{100} + \ldots + \frac{(87-100)^2}{100}$$
$$= \mathbf{14.56}.$$

The **p-value** is the area to the **right** of **14.56** under the χ^2 curve with $k - 1 = 9$ **df.**

The **p-value** is the area to the **right** of **14.56** under the $\boldsymbol{\chi^2}$ curve with $\boldsymbol{k - 1 = 9}$ **df.**

From a χ^2 distribution table, the **p-value** is **0.1037.**

The **p-value** is the area to the **right** of **14.56** under the $\boldsymbol{\chi^2}$ curve with $\boldsymbol{k-1=9}$ **df.**

From a χ^2 distribution table, the **p-value** is **0.1037.**

Using $\alpha = 0.05$, we **fail to reject** $\boldsymbol{H_0}$.

The **p-value** is the area to the **right** of **14.56** under the χ^2 curve with $k-1=9$ **df.**

From a χ^2 distribution table, the **p-value** is **0.1037**.

Using $\alpha = 0.05$, we **fail to reject** H_0.

We conclude that there is no evidence that the R "pseudo random numbers" behave any differently than they would if they were an iid sample from a standard normal distribution.