Goodness of Fit

Bernd Schröder
Introduction
Introduction

1. Goodness of fit generalizes our tests of population proportions.
Introduction

1. Goodness of fit generalizes our tests of population proportions.
2. It can also be used to check normality assumptions etc.
Review and Task
Review and Task

1. In a multinomial experiment, each outcome has probability $p_i$. 
Review and Task

1. In a multinomial experiment, each outcome has probability $p_i$.
2. The sum of the probabilities $p_i$ is 1.
Review and Task

1. In a multinomial experiment, each outcome has probability $p_i$.
2. The sum of the probabilities $p_i$ is 1.
3. Alternatively, we have several sub-populations in a larger population and we want to find out the proportions of these populations in the overall population.
Review and Task

1. In a multinomial experiment, each outcome has probability $p_i$.
2. The sum of the probabilities $p_i$ is 1.
3. Alternatively, we have several sub-populations in a larger population and we want to find out the proportions of these populations in the overall population.
4. Will test a null hypothesis that specifies all probabilities against the alternative hypothesis that (at least) one given probability is wrong.
Review and Task

1. In a multinomial experiment, each outcome has probability $p_i$.
2. The sum of the probabilities $p_i$ is 1.
3. Alternatively, we have several sub-populations in a larger population and we want to find out the proportions of these populations in the overall population.
4. Will test a null hypothesis that specifies all probabilities against the alternative hypothesis that (at least) one given probability is wrong.
5. Example: Four different blood types occur with different frequencies.
Devising a Test
Devising a Test

1. We have $k$ categories/outcomes.
Devising a Test

1. We have $k$ categories/outcomes.
2. Denote the probabilities from the null hypothesis by $p_{i0}$. 
Devising a Test

1. We have \( k \) categories/outcomes.
2. Denote the probabilities from the null hypothesis by \( p_{i0} \).
3. \( N_i \) will be the number of observations of the \( i^{\text{th}} \) outcome and \( n \) is the sum of the \( N_i \).
Devising a Test

1. We have $k$ categories/outcomes.
2. Denote the probabilities from the null hypothesis by $p_{i0}$.
3. $N_i$ will be the number of observations of the $i^{th}$ outcome and $n$ is the sum of the $N_i$.
4. $E(N_i) = np_i$. 
Devising a Test

1. We have $k$ categories/outcomes.
2. Denote the probabilities from the null hypothesis by $p_{i0}$.
3. $N_i$ will be the number of observations of the $i^{th}$ outcome and $n$ is the sum of the $N_i$.
4. $E(N_i) = np_i$.
5. If $np_i \geq 5$ for all $i$, the variable $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ has approximately a $\chi^2$ distribution with $k - 1$ degrees of freedom.
Devising a Test

1. We have $k$ categories/outcomes.
2. Denote the probabilities from the null hypothesis by $p_{i0}$.
3. $N_i$ will be the number of observations of the $i^{th}$ outcome and $n$ is the sum of the $N_i$.
4. $E(N_i) = np_i$.
5. If $np_i \geq 5$ for all $i$, the variable $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ has approximately a $\chi^2$ distribution with $k - 1$ degrees of freedom.

This variable takes the sum of the relative squared deviations from the value we expect.
Devising a Test

1. We have $k$ categories/outcomes.
2. Denote the probabilities from the null hypothesis by $p_{i0}$.
3. $N_i$ will be the number of observations of the $i^{th}$ outcome and $n$ is the sum of the $N_i$.
4. $E(N_i) = np_i$.
5. If $np_i \geq 5$ for all $i$, the variable $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ has approximately a $\chi^2$ distribution with $k - 1$ degrees of freedom.

This variable takes the sum of the relative squared deviations from the value we expect. We have $k - 1$ degrees of freedom, because the last cell count is determined by the total.
Devising a Test

1. We have $k$ categories/outcomes.
2. Denote the probabilities from the null hypothesis by $p_{i0}$.
3. $N_i$ will be the number of observations of the $i^{th}$ outcome and $n$ is the sum of the $N_i$.
4. $E(N_i) = np_i$.
5. If $np_i \geq 5$ for all $i$, the variable $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ has approximately a $\chi^2$ distribution with $k - 1$ degrees of freedom.

This variable takes the sum of the relative squared deviations from the value we expect. We have $k - 1$ degrees of freedom, because the last cell count is determined by the total. (Another hard proof omitted.)
Devising a Test

To test $H_0: p_i = p_{i0}$ for all $i$ against $H_a: p_i \neq p_{i0}$ for some $i$, use the test statistic $\chi^2 = \sum_{i=1}^{k} \left( \frac{n_i - np_{i0}}{np_{i0}} \right)^2$ with a rejection region $\chi^2 \geq \chi^2_{\alpha}$.
Devising a Test

To test \( H_0 : p_i = p_{i0} \) for all \( i \) against \( H_a : p_i \neq p_{i0} \) for some \( i \)
Devising a Test

To test $H_0 : p_i = p_{i0}$ for all $i$ against $H_a : p_i \neq p_{i0}$ for some $i$, use the test statistic $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$.
Devising a Test

To test $H_0 : p_i = p_{i0}$ for all $i$ against $H_a : p_i \neq p_{i0}$ for some $i$, use

the test statistic $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$ with a rejection region

$\chi^2 \geq \chi^2_{\alpha,k-1}$
Example.

\[
\chi^2 = \frac{(101 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(91 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(8 - 200 \cdot \frac{2}{38})^2}{200 \cdot \frac{2}{38}} \approx 1.168
\]

\[
\chi^2 \approx 232.9
\]

Also computed the \( p \)-value, which is within 10\(^{-20} \) of 1.

So we definitely do not reject the null hypothesis.
**Example.** 200 spins of a roulette wheel resulted in 101 occurrences of red, 91 occurrences of black, 8 occurrences of 0 or 00. Test the hypothesis that the wheel is fair at the 5% level.

\[ \chi^2 = \left( \frac{101 - 200 \cdot \frac{18}{38}}{200 \cdot \frac{18}{38}} \right)^2 + \left( \frac{91 - 200 \cdot \frac{18}{38}}{200 \cdot \frac{18}{38}} \right)^2 + \left( \frac{8 - 200 \cdot \frac{2}{38}}{200 \cdot \frac{2}{38}} \right)^2 \approx 1.168 \]

\[ \chi^2_{0.05, 199} = 232.9 \]

Also computed the \( p \)-value, which is within 10\(^{-20} \) of 1. So we definitely do not reject the null hypothesis.
Example. 200 spins of a roulette wheel resulted in 101 occurrences of red, 91 occurrences of black, 8 occurrences of 0 or 00. Test the hypothesis that the wheel is fair at the 5% level.

\[ \chi^2 = \left( \frac{101 - 200 \cdot \frac{18}{38}}{200 \cdot \frac{18}{38}} \right)^2 + \left( \frac{91 - 200 \cdot \frac{18}{38}}{200 \cdot \frac{18}{38}} \right)^2 + \left( \frac{8 - 200 \cdot \frac{2}{38}}{200 \cdot \frac{2}{38}} \right)^2 \approx 1.168 \]
Example. 200 spins of a roulette wheel resulted in 101 occurrences of red, 91 occurrences of black, 8 occurrences of 0 or 00. Test the hypothesis that the wheel is fair at the 5% level.

\[ \chi^2 = \frac{(101 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(91 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(8 - 200 \cdot \frac{2}{38})^2}{200 \cdot \frac{2}{38}} \approx 1.168 \]

\[ \chi_{0.05,199}^2 = 232.9 \]
Example. 200 spins of a roulette wheel resulted in 101 occurrences of red, 91 occurrences of black, 8 occurrences of 0 or 00. Test the hypothesis that the wheel is fair at the 5% level.

\[
\chi^2 = \frac{(101 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(91 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(8 - 200 \cdot \frac{2}{38})^2}{200 \cdot \frac{2}{38}} \approx 1.168
\]

\[
\chi^2_{0.05,199} = 232.9
\]

Also computed the \( p \)-value, which is within \( 10^{-20} \) of 1.
**Example.** 200 spins of a roulette wheel resulted in 101 occurrences of red, 91 occurrences of black, 8 occurrences of 0 or 00. Test the hypothesis that the wheel is fair at the 5% level.

\[
\chi^2 = \frac{(101 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(91 - 200 \cdot \frac{18}{38})^2}{200 \cdot \frac{18}{38}} + \frac{(8 - 200 \cdot \frac{2}{38})^2}{200 \cdot \frac{2}{38}} \approx 1.168
\]

\[
\chi^2_{0.05,199} = 232.9
\]

Also computed the \( p \)-value, which is within \( 10^{-20} \) of 1. So we definitely do not reject the null hypothesis.
Chi-squared Test For Values From a Continuous Distribution

1. Subdivide the scale into subintervals $[a_{i-1}, a_i)$ and use $p_i^0 = \int_{a_{i-1}}^{a_i} f_0(x) \, dx$.
2. Choose cells so that $np_i^0 \geq 5$.
3. Often cells are chosen so that all $np_i^0$ are equal.
Chi-squared Test For Values From a Continuous Distribution

1. Subdivide the scale into subintervals \([a_{i-1}, a_i)\) and use

\[ p_{i0} = \int_{a_{i-1}}^{a_i} f_0(x) \, dx. \]
Chi-squared Test For Values From a Continuous Distribution

1. Subdivide the scale into subintervals \([a_{i-1}, a_i)\) and use
   \[ p_{i0} = \int_{a_{i-1}}^{a_i} f_0(x) \, dx. \]

2. Choose cells so that \(np_{i0} \geq 5\).
Chi-squared Test For Values From a Continuous Distribution

1. Subdivide the scale into subintervals \([a_{i-1}, a_i)\) and use
   
   \[ p_{i0} = \int_{a_{i-1}}^{a_i} f_0(x) \, dx. \]

2. Choose cells so that \(np_{i0} \geq 5\).

3. Often cells are chosen so that all \(np_{i0}\) are equal.
Example.
**Example.** Use a goodness of fit test to test (at the 5% level) the hypothesis that the 200 given data points are from an exponential distribution with parameter $\lambda = 1$. 
**Example.** Use a goodness of fit test to test (at the 5% level) the hypothesis that the 200 given data points are from an exponential distribution with parameter $\lambda = 1$.

$$ p_{i0} := \frac{5}{200} $$
Example. Use a goodness of fit test to test (at the 5% level) the hypothesis that the 200 given data points are from an exponential distribution with parameter $\lambda = 1$.

\[
p_{i0} := \frac{5}{200} = \frac{1}{40}
\]
**Example.** Use a goodness of fit test to test (at the 5% level) the hypothesis that the 200 given data points are from an exponential distribution with parameter $\lambda = 1$.

\[ p_{i0} := \frac{5}{200} = \frac{1}{40} \]

$k = 40$
Example. Use a goodness of fit test to test (at the 5% level) the hypothesis that the 200 given data points are from an exponential distribution with parameter $\lambda = 1$.

\[
p_{i0} := \frac{5}{200} = \frac{1}{40}
\]

\[
k = 40
\]

\[
\chi^2_{0.05,39} = 54.572
\]
Bin cutoffs.
Bin cutoffs.

\[ \int_{0}^{x} \lambda e^{-\lambda t} \, dt \]
Bin cutoffs.

\[ \int_{0}^{x} \lambda e^{-\lambda t} \, dt = -e^{-\lambda t} \bigg|_{0}^{x} \]
Bin cutoffs.

\[
\int_0^x \lambda e^{-\lambda t} \, dt = -e^{-\lambda t} \bigg|_0^x = 1 - e^{-\lambda x}
\]
Bin cutoffs.

\[ \int_{0}^{x} \lambda e^{-\lambda t} \, dt = -e^{-\lambda t} \bigg|_{0}^{x} = 1 - e^{-\lambda x} \]

\[ 1 - e^{-\lambda x} \doteq kp_{i0} \]
Bin cutoffs.

\[
\int_0^x \lambda e^{-\lambda t} \, dt = -e^{-\lambda t}\bigg|_0^x = 1 - e^{-\lambda x}
\]

\[
1 - e^{-\lambda x} \equiv kp_{i0}
\]

\[
e^{-\lambda x} = 1 - kp_{i0}
\]
Bin cutoffs.

\[ \int_0^x \lambda e^{-\lambda t} \, dt = -e^{-\lambda t} \bigg|_0^x = 1 - e^{-\lambda x} \]

\[ 1 - e^{-\lambda x} = kp_{i0} \]

\[ e^{-\lambda x} = 1 - kp_{i0} \]

\[ -\lambda x = \ln(1 - kp_{i0}) \]
Bin cutoffs.

\[
\int_{0}^{x} \lambda e^{-\lambda t} \, dt = -e^{-\lambda t} \bigg|_{0}^{x} = 1 - e^{-\lambda x}
\]

\[
1 - e^{-\lambda x} = kp_{i0}
\]

\[
e^{-\lambda x} = 1 - kp_{i0}
\]

\[
-\lambda x = \ln(1 - kp_{i0})
\]

\[
x = -\frac{1}{\lambda} \ln(1 - kp_{i0})
\]