

Math 362: Mathematical Statistics II

Le Chen

le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on April 13, 2021

2021 Spring

Chapter 10. Goodness-of-fit Tests

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.4 Goodness-of-Fit Tests: Parameters Unknown

§ 10.5 Contingency Tables

Plan

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.4 Goodness-of-Fit Tests: Parameters Unknown

§ 10.5 Contingency Tables

Chapter 10. Goodness-of-fit Tests

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.4 Goodness-of-Fit Tests: Parameters Unknown

§ 10.5 Contingency Tables

Rationale

! We want to test if the c.d.f. $F_Y(\cdot)$ is given by the true c.d.f. $F_0(\cdot)$, i.e.,

$$H_0 : F_Y(y) = F_0(y) \quad v.s. \quad H_1 : F_Y(y) \neq F_0(y)$$

~ By properly partitioning the domain, the random sample should follow
an induced multinomial distribution.

⇒ Then testing $F_Y(\cdot) = F_0(\cdot)$ reduces to testing the induced multinomial distribution of the following form:

$$H_0 : p_1 = p'_1, \dots, p_n = p'_n$$

v.s.

$$H_1 : p_i \neq p'_i \quad \text{for at least one } i$$

Rationale

! We want to test if the c.d.f. $F_Y(\cdot)$ is given by the true c.d.f. $F_0(\cdot)$, i.e.,

$$H_0 : F_Y(y) = F_0(y) \quad \text{v.s.} \quad H_1 : F_Y(y) \neq F_0(y)$$

~ By properly partitioning the domain, the random sample should follow
an induced multinomial distribution.

⇒ Then testing $F_Y(\cdot) = F_0(\cdot)$ reduces to testing the induced multinomial distribution of the following form:

$$\begin{aligned} H_0 : p_1 = p'_1, \dots, p_n = p'_n \\ \text{v.s.} \\ H_1 : p_i \neq p'_i \quad \text{for at least one } i \end{aligned}$$

Rationale

! We want to test if the c.d.f. $F_Y(\cdot)$ is given by the true c.d.f. $F_0(\cdot)$, i.e.,

$$H_0 : F_Y(y) = F_0(y) \quad \text{v.s.} \quad H_1 : F_Y(y) \neq F_0(y)$$

~ By properly partitioning the domain, the random sample should follow
an induced multinomial distribution.

⇒ Then testing $F_Y(\cdot) = F_0(\cdot)$ reduces to testing the induced multinomial distribution of the following form:

$$H_0 : p_1 = p'_1, \dots, p_n = p'_n$$

v.s.

$$H_1 : p_i \neq p'_i \quad \text{for at least one } i$$

Rationale

! We want to test if the c.d.f. $F_Y(\cdot)$ is given by the true c.d.f. $F_0(\cdot)$, i.e.,

$$H_0 : F_Y(y) = F_0(y) \quad \text{v.s.} \quad H_1 : F_Y(y) \neq F_0(y)$$

~ By properly partitioning the domain, the random sample should follow
an induced multinomial distribution.

⇒ Then testing $F_Y(\cdot) = F_0(\cdot)$ reduces to testing the induced multinomial distribution of the following form:

$$H_0 : p_1 = p'_1, \dots, p_n = p'_n$$

v.s.

$$H_1 : p_i \neq p'_i \quad \text{for at least one } i$$

How

1. Suppose we are sampling from the c.d.f. $F(y)$
2. Divide the range of the distribution into k mutually exclusive and exhaustive intervals, say I_1, \dots, I_k .
3. Let $\pi_i = \mathbb{P}(X \in I_i)$, $i = 1, \dots, k$.
4. Let O_1, \dots, O_k be the respective observed numbers of the observations X_1, \dots, X_n in the intervals I_1, \dots, I_k .
5. Then $O = (O_1, \dots, O_k) \sim$ multinomial distribution with (π_1, \dots, π_k) , i.e.,

$$\mathbb{P}(O_1 = o_1, \dots, O_k = o_k) = \frac{n!}{\prod_{i=1}^k o_i!} \prod_{i=1}^k \pi_i^{o_i}$$

with $\sum_{i=1}^k \pi_i = 1$, $\sum_{i=1}^k o_i = n$, and

$$\mathbb{E}[O_i] = n\pi_i =: e_i, \quad \text{Var}(O_i) = n\pi_i(1 - \pi_i)$$

How

1. Suppose we are sampling from the c.d.f. $F(y)$
2. Divide the range of the distribution into k mutually exclusive and exhaustive intervals, say I_1, \dots, I_k .
3. Let $\pi_i = \mathbb{P}(X \in I_i)$, $i = 1, \dots, k$.
4. Let O_1, \dots, O_k be the respective observed numbers of the observations X_1, \dots, X_n in the intervals I_1, \dots, I_k .
5. Then $O = (O_1, \dots, O_k) \sim$ multinomial distribution with (π_1, \dots, π_k) , i.e.,

$$\mathbb{P}(O_1 = o_1, \dots, O_k = o_k) = \frac{n!}{\prod_{i=1}^k o_i!} \prod_{i=1}^k \pi_i^{o_i}$$

with $\sum_{i=1}^k \pi_i = 1$, $\sum_{i=1}^k o_i = n$, and

$$\mathbb{E}[O_i] = n\pi_i =: e_i, \quad \text{Var}(O_i) = n\pi_i(1 - \pi_i)$$

How

1. Suppose we are sampling from the c.d.f. $F(y)$
2. Divide the range of the distribution into k mutually exclusive and exhaustive intervals, say I_1, \dots, I_k .
3. Let $\pi_i = \mathbb{P}(X \in I_i)$, $i = 1, \dots, k$.
4. Let O_1, \dots, O_k be the respective observed numbers of the observations X_1, \dots, X_n in the intervals I_1, \dots, I_k .
5. Then $O = (O_1, \dots, O_k) \sim$ multinomial distribution with (π_1, \dots, π_k) , i.e.,

$$\mathbb{P}(O_1 = o_1, \dots, O_k = o_k) = \frac{n!}{\prod_{i=1}^k o_i!} \prod_{i=1}^k \pi_i^{o_i}$$

with $\sum_{i=1}^k \pi_i = 1$, $\sum_{i=1}^k o_i = n$, and

$$\mathbb{E}[O_i] = n\pi_i =: e_i, \quad \text{Var}(O_i) = n\pi_i(1 - \pi_i)$$

How

1. Suppose we are sampling from the c.d.f. $F(y)$
2. Divide the range of the distribution into k mutually exclusive and exhaustive intervals, say I_1, \dots, I_k .
3. Let $\pi_i = \mathbb{P}(X \in I_i)$, $i = 1, \dots, k$.
4. Let O_1, \dots, O_k be the respective observed numbers of the observations X_1, \dots, X_n in the intervals I_1, \dots, I_k .
5. Then $O = (O_1, \dots, O_k) \sim$ multinomial distribution with (π_1, \dots, π_k) , i.e.,

$$\mathbb{P}(O_1 = o_1, \dots, O_k = o_k) = \frac{n!}{\prod_{i=1}^k o_i!} \prod_{j=1}^k \pi_j^{o_j}$$

with $\sum_{i=1}^k \pi_i = 1$, $\sum_{i=1}^k o_i = n$, and

$$\mathbb{E}[O_i] = n\pi_i =: e_i, \quad \text{Var}(O_i) = n\pi_i(1 - \pi_i)$$

How

1. Suppose we are sampling from the c.d.f. $F(y)$
2. Divide the range of the distribution into k mutually exclusive and exhaustive intervals, say I_1, \dots, I_k .
3. Let $\pi_i = \mathbb{P}(X \in I_i)$, $i = 1, \dots, k$.
4. Let O_1, \dots, O_k be the respective observed numbers of the observations X_1, \dots, X_n in the intervals I_1, \dots, I_k .
5. Then $O = (O_1, \dots, O_k) \sim$ multinomial distribution with (π_1, \dots, π_k) , i.e.,

$$\mathbb{P}(O_1 = o_1, \dots, O_k = o_k) = \frac{n!}{\prod_{i=1}^k o_i!} \prod_{j=1}^k \pi_j^{o_j}$$

with $\sum_{i=1}^k \pi_i = 1$, $\sum_{i=1}^k o_i = n$, and

$$\mathbb{E}[O_i] = n\pi_i =: e_i, \quad \text{Var}(O_i) = n\pi_i(1 - \pi_i)$$

How

1. Suppose we are sampling from the c.d.f. $F(y)$
2. Divide the range of the distribution into k mutually exclusive and exhaustive intervals, say I_1, \dots, I_k .
3. Let $\pi_i = \mathbb{P}(X \in I_i)$, $i = 1, \dots, k$.
4. Let O_1, \dots, O_k be the respective observed numbers of the observations X_1, \dots, X_n in the intervals I_1, \dots, I_k .
5. Then $O = (O_1, \dots, O_k) \sim$ multinomial distribution with (π_1, \dots, π_k) , i.e.,

$$\mathbb{P}(O_1 = o_1, \dots, O_k = o_k) = \frac{n!}{\prod_{i=1}^k o_i!} \prod_{i=1}^k \pi_i^{o_i}$$

with $\sum_{i=1}^k \pi_i = 1$, $\sum_{i=1}^k o_i = n$, and

$$\mathbb{E}[O_i] = n\pi_i =: e_i, \quad \text{Var}(O_i) = n\pi_i(1 - \pi_i)$$

6. When $k = 2$, by CLT, as $n \rightarrow \infty$,

$$\begin{aligned} \frac{O_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} &\xrightarrow{d} N(0, 1) \quad \implies \quad \frac{(O_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \xrightarrow{d} \chi_1^2 \\ & \parallel \\ & \frac{(O_1 - n\pi_1)^2}{n\pi_1} + \frac{(O_2 - n\pi_2)^2}{n\pi_2} \\ & \parallel \\ & \frac{(O_1 - \theta_1)^2}{\theta_1} + \frac{(O_2 - \theta_2)^2}{\theta_2} \end{aligned}$$

Hence, as $n \rightarrow \infty$,

$$\sum_{i=1}^k \frac{(O_i - \theta_i)^2}{\theta_i} \xrightarrow{d} \chi_{k-1}^2$$

6. When $k = 2$, by CLT, as $n \rightarrow \infty$,

$$\begin{aligned} \frac{O_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} &\xrightarrow{d} N(0, 1) \quad \Longrightarrow \quad \frac{(O_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \xrightarrow{d} \chi_1^2 \\ & \parallel \\ & \frac{(O_1 - n\pi_1)^2}{n\pi_1} + \frac{(O_2 - n\pi_2)^2}{n\pi_2} \\ & \parallel \\ & \frac{(O_1 - \theta_1)^2}{\theta_1} + \frac{(O_2 - \theta_2)^2}{\theta_2} \end{aligned}$$

Hence, as $n \rightarrow \infty$,

$$\sum_{i=1}^k \frac{(O_i - \theta_i)^2}{\theta_i} \xrightarrow{d} \chi_{k-1}^2$$

6. When $k = 2$, by CLT, as $n \rightarrow \infty$,

$$\begin{aligned} \frac{O_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} &\xrightarrow{d} N(0, 1) \quad \implies \quad \frac{(O_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \xrightarrow{d} \chi_1^2 \\ &\quad \parallel \\ &\frac{(O_1 - n\pi_1)^2}{n\pi_1} + \frac{(O_2 - n\pi_2)^2}{n\pi_2} \\ &\quad \parallel \\ &\frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2} \end{aligned}$$

Hence, as $n \rightarrow \infty$,

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi_{k-1}^2$$

6. When $k = 2$, by CLT, as $n \rightarrow \infty$,

$$\begin{aligned} \frac{O_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} &\xrightarrow{d} N(0, 1) \implies \frac{(O_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \xrightarrow{d} \chi_1^2 \\ &\parallel \\ &\frac{(O_1 - n\pi_1)^2}{n\pi_1} + \frac{(O_2 - n\pi_2)^2}{n\pi_2} \\ &\parallel \\ &\frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2} \end{aligned}$$

Hence, as $n \rightarrow \infty$,

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi_{k-1}^2$$

7. For general k ,

$$\sum_{i=1}^k \frac{(O_i - n\pi_i)^2}{n\pi_i} = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

follows a complicated, but exact, distribution, from which, one can show

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi_{k-1}^2$$

↓

Thm. When n is large enough, namely, when $n\pi_i \geq 5$ for all i ,

$$D = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \underset{\text{appr.}}{\sim} \chi_{k-1}^2.$$

Rmk: The above is called Pearson's chi-square test. It is asymptotically equivalent to the generalized likelihood ratio test.

7. For general k ,

$$\sum_{i=1}^k \frac{(O_i - n\pi_i)^2}{n\pi_i} = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

follows a complicated, but exact, distribution, from which, one can show

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi_{k-1}^2$$

↓

Thm. When n is large enough, namely, when $n\pi_i \geq 5$ for all i ,

$$D = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \underset{\text{appr.}}{\sim} \chi_{k-1}^2.$$

Rmk: The above is called Pearson's chi-square test. It is asymptotically equivalent to the generalized likelihood ratio test.

7. For general k ,

$$\sum_{i=1}^k \frac{(O_i - n\pi_i)^2}{n\pi_i} = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

follows a complicated, but exact, distribution, from which, one can show

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi_{k-1}^2$$

↓

Thm. When n is large enough, namely, when $n\pi_i \geq 5$ for all i ,

$$D = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \overset{\text{appr.}}{\sim} \chi_{k-1}^2.$$

Rmk: The above is called Pearson's chi-square test. It is asymptotically equivalent to the generalized likelihood ratio test.

7. For general k ,

$$\sum_{i=1}^k \frac{(O_i - n\pi_i)^2}{n\pi_i} = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

follows a complicated, but exact, distribution, from which, one can show

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi_{k-1}^2$$

↓

Thm. When n is large enough, namely, when $n\pi_i \geq 5$ for all i ,

$$D = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i} \underset{\text{appr.}}{\sim} \chi_{k-1}^2.$$

Rmk: The above is called Pearson's chi-square test. It is asymptotically equivalent to the generalized likelihood ratio test.

Alternative: G-test

– the likelihood ratio test for multinomial model

1. Under $H_0 : \pi_i = p_i, i = 1, \dots, k$, the MLE of π_i are

$$\tilde{\pi}_i = p_i = \frac{np_i}{n} = \frac{e_i}{n}, \quad \forall i.$$

2. When there are no constraints, for $i = 1, \dots, k-1$,

$$\frac{\partial}{\partial \pi_i} \ln L(\pi_1, \dots, \pi_{k-1} | o_1, \dots, o_k) = 0, \quad 1 \leq i \leq k-1$$

\Downarrow

$$\frac{o_i}{\hat{\pi}_i} = \frac{o_k}{1 - \hat{\pi}_1 - \dots - \hat{\pi}_{k-1}}, \quad 1 \leq i \leq k-1$$

\Downarrow

$$\hat{\pi}_i = \frac{o_i}{n}, \quad 1 \leq i \leq k.$$

Alternative: G-test

– the likelihood ratio test for multinomial model

1. Under $H_0 : \pi_i = p_i, i = 1, \dots, k$, the MLE of π_i are

$$\tilde{\pi}_i = p_i = \frac{np_i}{n} = \frac{e_i}{n}, \quad \forall i.$$

2. When there are no constraints, for $i = 1, \dots, k - 1$,

$$\frac{\partial}{\partial \pi_i} \ln L(\pi_1, \dots, \pi_{k-1} | o_1, \dots, o_k) = 0, \quad 1 \leq i \leq k - 1$$

\Leftrightarrow

$$\frac{o_i}{\hat{\pi}_i} = \frac{o_k}{1 - \hat{\pi}_1 - \dots - \hat{\pi}_{k-1}}, \quad 1 \leq i \leq k - 1$$

\Leftrightarrow

$$\hat{\pi}_i = \frac{o_i}{n}, \quad 1 \leq i \leq k.$$

⇒

$$\begin{aligned}\lambda &:= \ln \left(\frac{L(\tilde{\pi}_1, \dots, \tilde{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)}{L(\hat{\pi}_1, \dots, \hat{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)} \right) = \log \left(\frac{\prod_{i=1}^k \tilde{\pi}_i^{o_i}}{\prod_{i=1}^k \hat{\pi}_i^{o_i}} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{\theta_i}{\phi_i} \right)\end{aligned}$$

Critical region: $\lambda < \lambda_* < 0$.

Def.

$$G := -2\lambda = -2 \sum_{i=1}^k o_i \ln \left(\frac{\phi_i}{\theta_i} \right) = 2 \sum_{i=1}^k o_i \ln \left(\frac{\theta_i}{\phi_i} \right)$$

$G \stackrel{\text{approx.}}{\sim} \chi_{k-1}^2$ for large n .

Critical region: $G \geq G_* = \chi_{1-\alpha, k-1}^2$.

⇒

$$\begin{aligned}\lambda &:= \ln \left(\frac{L(\tilde{\pi}_1, \dots, \tilde{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)}{L(\hat{\pi}_1, \dots, \hat{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)} \right) = \log \left(\frac{\prod_{i=1}^k \tilde{\pi}_i^{o_i}}{\prod_{i=1}^k \hat{\pi}_i^{o_i}} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right)\end{aligned}$$

Critical region: $\lambda < \lambda_* < 0$.

Def.

$$G := -2\lambda = -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) = 2 \sum_{i=1}^k o_i \ln \left(\frac{o_i}{e_i} \right)$$

$G \stackrel{\text{approx.}}{\sim} \chi_{k-1}^2$ for large n .

Critical region: $G \geq G_* = \chi_{1-\alpha, k-1}^2$.

⇒

$$\begin{aligned}\lambda &:= \ln \left(\frac{L(\tilde{\pi}_1, \dots, \tilde{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)}{L(\hat{\pi}_1, \dots, \hat{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)} \right) = \log \left(\frac{\prod_{i=1}^k \tilde{\pi}_i^{o_i}}{\prod_{i=1}^k \hat{\pi}_i^{o_i}} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right)\end{aligned}$$

Critical region: $\lambda < \lambda_* < 0$.

Def.

$$G := -2\lambda = -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) = 2 \sum_{i=1}^k o_i \ln \left(\frac{o_i}{e_i} \right)$$

$G \stackrel{\text{approx.}}{\sim} \chi_{k-1}^2$ for large n .

Critical region: $G \geq G_* = \chi_{1-\alpha, k-1}^2$.

⇒

$$\begin{aligned}\lambda &:= \ln \left(\frac{L(\tilde{\pi}_1, \dots, \tilde{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)}{L(\hat{\pi}_1, \dots, \hat{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)} \right) = \log \left(\frac{\prod_{i=1}^k \tilde{\pi}_i^{o_i}}{\prod_{i=1}^k \hat{\pi}_i^{o_i}} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right)\end{aligned}$$

Critical region: $\lambda < \lambda_* < 0$.

Def.

$$G := -2\lambda = -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) = 2 \sum_{i=1}^k o_i \ln \left(\frac{o_i}{e_i} \right)$$

$G \stackrel{\text{approx.}}{\sim} \chi_{k-1}^2$ for large n .

Critical region: $G \geq G_* = \chi_{1-\alpha, k-1}^2$.

⇒

$$\begin{aligned}\lambda &:= \ln \left(\frac{L(\tilde{\pi}_1, \dots, \tilde{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)}{L(\hat{\pi}_1, \dots, \hat{\pi}_{k-1} | \mathbf{o}_1, \dots, \mathbf{o}_k)} \right) = \log \left(\frac{\prod_{i=1}^k \tilde{\pi}_i^{o_i}}{\prod_{i=1}^k \hat{\pi}_i^{o_i}} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right) \\ &= \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right)\end{aligned}$$

Critical region: $\lambda < \lambda_* < 0$.

Def.

$$\mathbf{G} := -2\lambda = -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) = 2 \sum_{i=1}^k o_i \ln \left(\frac{o_i}{e_i} \right)$$

$\mathbf{G} \overset{\text{approx.}}{\sim} \chi_{k-1}^2$ for large n .

Critical region: $\mathbf{G} \geq \mathbf{G}_* = \chi_{1-\alpha, k-1}^2$.

Relation G-test and Pearson's Chi square test

By second order Taylor expansion around 1,

$$\begin{aligned} G &= -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) \\ &\approx -2 \sum_{i=1}^k o_i \left[\left(\frac{e_i}{o_i} - 1 \right) - \frac{1}{2} \left(\frac{e_i}{o_i} - 1 \right)^2 \right] \\ &= -2 \sum_{i=1}^k (e_i - o_i) + \sum_{i=1}^k o_i \left(\left(1 - \frac{o_i}{e_i} \right) + \frac{o_i}{e_i} \right) \left(\frac{e_i}{o_i} - 1 \right)^2 \\ &= 0 + \sum_{i=1}^n \frac{o_i^2}{e_i} \left(1 - \frac{o_i}{e_i} \right)^3 + \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\ &\approx \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\ &\quad \parallel \\ &\quad D \end{aligned}$$

∴ Pearson's Chi-square test is an approximation of G-test.

Relation G-test and Pearson's Chi square test

By second order Taylor expansion around 1,

$$\begin{aligned} G &= -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) \\ &\approx -2 \sum_{i=1}^k o_i \left[\left(\frac{e_i}{o_i} - 1 \right) - \frac{1}{2} \left(\frac{e_i}{o_i} - 1 \right)^2 \right] \\ &= -2 \sum_{i=1}^k (e_i - o_i) + \sum_{i=1}^k o_i \left(\left(1 - \frac{o_i}{e_i} \right) + \frac{o_i}{e_i} \right) \left(\frac{e_i}{o_i} - 1 \right)^2 \\ &= 0 + \sum_{i=1}^k \frac{o_i^2}{e_i} \left(1 - \frac{o_i}{e_i} \right)^3 + \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\ &\approx \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\ &\quad \parallel \\ &\quad D \end{aligned}$$

∴ Pearson's Chi-square test is an approximation of G-test.

Relation G-test and Pearson's Chi square test

By second order Taylor expansion around 1,

$$\begin{aligned}G &= -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) \\&\approx -2 \sum_{i=1}^k o_i \left[\left(\frac{e_i}{o_i} - 1 \right) - \frac{1}{2} \left(\frac{e_i}{o_i} - 1 \right)^2 \right] \\&= -2 \sum_{i=1}^k (e_i - o_i) + \sum_{i=1}^k o_i \left(\left(1 - \frac{o_i}{e_i} \right) + \frac{o_i}{e_i} \right) \left(\frac{e_i}{o_i} - 1 \right)^2 \\&= 0 + \sum_{i=1}^k \frac{o_i^2}{e_i} \left(1 - \frac{o_i}{e_i} \right)^3 + \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\&\approx \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\&\quad \parallel \\&\quad D\end{aligned}$$

∴ Pearson's Chi-square test is an approximation of G-test.

Relation G-test and Pearson's Chi square test

By second order Taylor expansion around 1,

$$\begin{aligned} G &= -2 \sum_{i=1}^k o_i \ln \left(\frac{e_i}{o_i} \right) \\ &\approx -2 \sum_{i=1}^k o_i \left[\left(\frac{e_i}{o_i} - 1 \right) - \frac{1}{2} \left(\frac{e_i}{o_i} - 1 \right)^2 \right] \\ &= -2 \sum_{i=1}^k (e_i - o_i) + \sum_{i=1}^k o_i \left(\left(1 - \frac{o_i}{e_i} \right) + \frac{o_i}{e_i} \right) \left(\frac{e_i}{o_i} - 1 \right)^2 \\ &= 0 + \sum_{i=1}^k \frac{o_i^2}{e_i} \left(1 - \frac{o_i}{e_i} \right)^3 + \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\ &\approx \sum_{i=1}^k \frac{(e_i - o_i)^2}{e_i} \\ &\quad \parallel \\ &\quad D \end{aligned}$$

\therefore Pearson's Chi-square test is an approximation of G-test.

E.g. 1 *Benford's law*:

Initial digits

Use this law to check whether the bookkeepers have made up entries.

Assume that bookkeepers are not aware of Benford's law.

E.g. 1 *Benford's law*:

| Table 10.3.1 | |
|---------------------|-----------------------------------|
| Digit, i | $\log_{10}(i + 1) - \log_{10}(i)$ |
| 1 | 0.301 |
| 2 | 0.176 |
| 3 | 0.125 |
| 4 | 0.097 |
| 5 | 0.079 |
| 6 | 0.067 |
| 7 | 0.058 |
| 8 | 0.051 |
| 9 | 0.046 |

Initial digits

| Digit | Observed, k_i |
|-------|-----------------|
| 1 | 111 |
| 2 | 60 |
| 3 | 46 |
| 4 | 29 |
| 5 | 26 |
| 6 | 22 |
| 7 | 21 |
| 8 | 20 |
| 9 | 20 |
| | <hr/> |
| | 355 |

Use this law to check whether the bookkeepers have made up entries.

Assume that bookkeepers are not aware of Benford's law.

Sol. The test should be

$$H_0 : p_1 = p_{10}, \dots, p_9 = p_{90}$$

v.s.

$$H_1 : p_i \neq p_{i0} \quad \text{for at least one } i = 1, \dots, 9.$$

Critical region: $(\chi^2_{95,8}, \infty) = (15.507, \infty)$.

Sol. The test should be

$$H_0 : p_1 = p_{10}, \dots, p_9 = p_{90}$$

v.s.

$$H_1 : p_i \neq p_{i0} \quad \text{for at least one } i = 1, \dots, 9.$$

Critical region: $(\chi^2_{.95,8}, \infty) = (15.507, \infty)$.

Compute the D and G scores:

| Digit | o_i | p_i | e_i | $(o_i - e_i)^2 / e_i$ | $2o_i \ln(e_i / o_i)$ |
|-------|-------|-------|-------|--------------------------------|--------------------------------|
| 1 | 111 | 0.301 | | | |
| 2 | 60 | 0.176 | | | |
| 3 | 46 | 0.125 | | | |
| 4 | 29 | 0.097 | | | |
| 5 | 26 | 0.079 | | | |
| 6 | 22 | 0.067 | | | |
| 7 | 21 | 0.058 | | | |
| 8 | 20 | 0.051 | | | |
| 9 | 20 | 0.046 | | | |
| sum | 355 | 1 | 355 | $d = \underline{\hspace{2cm}}$ | $g = \underline{\hspace{2cm}}$ |

| Digit | o_i | p_i | e_i | $(o_i - e_i)^2 / e_i$ | $2o_i \ln(e_i / o_i)$ |
|-------|-------|-------|-------|------------------------|------------------------|
| 1 | 111 | 0.301 | 106.9 | 0.16 | 8.449 |
| 2 | 60 | 0.176 | 62.5 | 0.10 | -4.860 |
| 3 | 46 | 0.125 | 44.4 | 0.06 | 3.309 |
| 4 | 29 | 0.097 | 34.4 | 0.86 | -9.963 |
| 5 | 26 | 0.079 | 28.0 | 0.15 | -3.937 |
| 6 | 22 | 0.067 | 23.8 | 0.13 | -3.433 |
| 7 | 21 | 0.058 | 20.6 | 0.01 | 0.828 |
| 8 | 20 | 0.051 | 18.1 | 0.20 | 3.982 |
| 9 | 20 | 0.046 | 16.3 | 0.82 | 8.109 |
| sum | 355 | 1 | 355 | $d = \underline{2.49}$ | $g = \underline{2.48}$ |

Conclusion: Fail to reject.

```

1 > # EX 10.3.2
2 > library(data.table)
3 > mydat <- fread('http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_
  10-3-2.data')
4 trying URL 'http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_10-3-2.
  data'
5 Content type 'unknown' length 153 bytes
6 =====
7 downloaded 153 bytes
8
9 > head(mydat)
10   Digit Oi   Pi
11 1:     1 111 0.301
12 2:     2  60 0.176
13 3:     3  46 0.125
14 4:     4  29 0.097
15 > pi = mydat[,3]
16 > oi = mydat[,2]
17 > n = sum(oi)
18 > ei = n*pi
19 > di = (ei-oi)^2/ei
20 > gi = 2*oi*log(oi/ei)
21 > print(paste("Using Pearson's test, D value is equal to ", round(sum(di),3)))
22 [1] "Using Pearson's test, D value is equal to 2.491"
23 > print(paste("Using the G-test, G value is equal to ", round(sum(gi),3)))
24 [1] "Using the G-test, G value is equal to 2.484"

```

Codes available

http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_10-3-2.R

E.g. 2 Test for randomness

Is the following sample of size 40 from $f_Y(y) = 6y(1 - y)$, $y \in [0, 1]$?

E.g. 2 Test for randomness

Is the following sample of size 40 from $f_Y(y) = 6y(1 - y)$, $y \in [0, 1]$?

| | | | | |
|------|------|------|------|------|
| 0.18 | 0.06 | 0.27 | 0.58 | 0.98 |
| 0.55 | 0.24 | 0.58 | 0.97 | 0.36 |
| 0.48 | 0.11 | 0.59 | 0.15 | 0.53 |
| 0.29 | 0.46 | 0.21 | 0.39 | 0.89 |
| 0.34 | 0.09 | 0.64 | 0.52 | 0.64 |
| 0.71 | 0.56 | 0.48 | 0.44 | 0.40 |
| 0.80 | 0.83 | 0.02 | 0.10 | 0.51 |
| 0.43 | 0.14 | 0.74 | 0.75 | 0.22 |

Sol. Test continuous pdf \rightarrow reduce to a set of classes:

$$d = \dots = 1.84.$$

Sol. Test continuous pdf \rightarrow reduce to a set of classes:

| Class | Observed Frequency, k_i | P_{i_o} | $40 p_{i_o}$ |
|----------------------|---------------------------|-----------|--------------|
| $0 \leq y < 0.20$ | 8 | 0.104 | 4.16 |
| $0.20 \leq y < 0.40$ | 8 | 0.248 | 9.92 |
| $0.40 \leq y < 0.60$ | 14 | 0.296 | 11.84 |
| $0.60 \leq y < 0.80$ | 5 | 0.248 | 9.92 |
| $0.80 \leq y < 1.00$ | 5 | 0.104 | 4.16 |

$$d = \dots = 1.84.$$

Sol. Test continuous pdf \rightarrow reduce to a set of classes:

| Table 10.3.5 | | | |
|----------------------|---------------------------|-----------|--------------|
| Class | Observed Frequency, k_i | P_{i_o} | $40 p_{i_o}$ |
| $0 \leq y < 0.20$ | 8 | 0.104 | 4.16 |
| $0.20 \leq y < 0.40$ | 8 | 0.248 | 9.92 |
| $0.40 \leq y < 0.60$ | 14 | 0.296 | 11.84 |
| $0.60 \leq y < 0.80$ | 5 | 0.248 | 9.92 |
| $0.80 \leq y < 1.00$ | 5 | 0.104 | 4.16 |

| Table 10.3.6 | | | |
|-------------------------|---------------------------|-----------|--------------|
| Class | Observed Frequency, k_i | P_{i_o} | $40 p_{i_o}$ |
| $0 \leq y < 0.40$ | 16 | 0.352 | 14.08 |
| $0.40 \leq y < 0.60$ | 14 | 0.296 | 11.84 |
| $0.60 \leq y \leq 1.00$ | 10 | 0.352 | 14.08 |

$$d = \dots = 1.84.$$

Sol. Test continuous pdf \rightarrow reduce to a set of classes:

| Table 10.3.5 | | | |
|----------------------|---------------------------|-----------|--------------|
| Class | Observed Frequency, k_i | P_{i_o} | $40 p_{i_o}$ |
| $0 \leq y < 0.20$ | 8 | 0.104 | 4.16 |
| $0.20 \leq y < 0.40$ | 8 | 0.248 | 9.92 |
| $0.40 \leq y < 0.60$ | 14 | 0.296 | 11.84 |
| $0.60 \leq y < 0.80$ | 5 | 0.248 | 9.92 |
| $0.80 \leq y < 1.00$ | 5 | 0.104 | 4.16 |

| Table 10.3.6 | | | |
|-------------------------|---------------------------|-----------|--------------|
| Class | Observed Frequency, k_i | P_{i_o} | $40 p_{i_o}$ |
| $0 \leq y < 0.40$ | 16 | 0.352 | 14.08 |
| $0.40 \leq y < 0.60$ | 14 | 0.296 | 11.84 |
| $0.60 \leq y \leq 1.00$ | 10 | 0.352 | 14.08 |

$$d = \dots = 1.84.$$

Critical region: $(\chi_{.95,2}^2, \infty) = (5.992, \infty)$.

Conclusion: Fail to reject.

Sol. Test continuous pdf \rightarrow reduce to a set of classes:

| Table 10.3.5 | | | |
|----------------------|---------------------------|-----------|--------------|
| Class | Observed Frequency, k_i | P_{i_o} | $40 p_{i_o}$ |
| $0 \leq y < 0.20$ | 8 | 0.104 | 4.16 |
| $0.20 \leq y < 0.40$ | 8 | 0.248 | 9.92 |
| $0.40 \leq y < 0.60$ | 14 | 0.296 | 11.84 |
| $0.60 \leq y < 0.80$ | 5 | 0.248 | 9.92 |
| $0.80 \leq y < 1.00$ | 5 | 0.104 | 4.16 |

| Table 10.3.6 | | | |
|-------------------------|---------------------------|-----------|--------------|
| Class | Observed Frequency, k_i | P_{i_o} | $40 p_{i_o}$ |
| $0 \leq y < 0.40$ | 16 | 0.352 | 14.08 |
| $0.40 \leq y < 0.60$ | 14 | 0.296 | 11.84 |
| $0.60 \leq y \leq 1.00$ | 10 | 0.352 | 14.08 |

$$d = \dots = 1.84.$$

Critical region: $(\chi_{.95,2}^2, \infty) = (5.992, \infty)$.

Conclusion: Fail to reject.

```

1 > # Case Study 10.3.2
2 > # Read data from the URL link
3 > library(data.table)
4 > mydat <- fread('http://math.emory.edu/~lchen41/teaching/2020_Spring/EX_
      10-3-1.data')
5 trying URL 'http://math.emory.edu/~lchen41/teaching/2020_Spring/EX_10-3-1.
      data'
6 Content type 'unknown' length 234 bytes
7 =====
8 downloaded 234 bytes
9
10 >d(mydat)
11   Col1 Col2 Col3 Col4 Col5
12 1: 0.18 0.06 0.27 0.58 0.98
13 2: 0.55 0.24 0.58 0.97 0.36
14 3: 0.48 0.11 0.59 0.15 0.53
15 4: 0.29 0.46 0.21 0.39 0.89
16 5: 0.34 0.09 0.64 0.52 0.64
17 6: 0.71 0.56 0.48 0.44 0.40
18 # Conditions for lower bounds
19 > lb=c(0,0.40,0.60)
20 > # Conditions for upper bounds
21 > up=c(0.40,0.60,1.00)
22 > # Store the results in d
23 > oi <- seq(1:length(lb))
24 > pi <- seq(1:length(lb))
25 > integrand <- function(y) {6*y*(1-y)}
26 > for (i in c(1:length(lb))) {
27 +   oi[i] <- table(mydat>=lb[i] & mydat<up[i])[2]
28 +   pi[i] <- integrate(integrand, lb[i], up[i])$value[1]
29 +   print(paste("the", i,"th bin has", oi[i],
30 +     "entries and pi is equal to", pi[i]))
31 + }

```



```

1 [1] "the 1 th bin has 16 entries and pi is equal to 0.352"
2 [1] "the 2 th bin has 14 entries and pi is equal to 0.296"
3 [1] "the 3 th bin has 10 entries and pi is equal to 0.352"
4 > pi <- unlist(pi)
5 > n <- sum(oi)
6 > ei <- n*pi
7 > di <- (ei-oi)^2/ei
8 > gi <- 2*oi*log(oi/ei)
9 > rbind(oi,pi,ei,di,gi)
10      [,1]      [,2]      [,3]
11 oi 16.0000000 14.0000000 10.0000000
12 pi  0.3520000 0.2960000 0.3520000
13 ei 14.0800000 11.8400000 14.0800000
14 di  0.2618182 0.3940541 1.182273
15 gi  4.0906679 4.6920636 -6.843405
16 > print(paste("Using Pearson's test, D value is equal to ",round(sum(
    di),3)))
17 [1] "Using Pearson's test, D value is equal to 1.838"
18 > print(paste("Using the G-test, G value is equal to ", round(sum(gi
    ),3)))
19 [1] "Using the G-test, G value is equal to 1.939"<Paste>

```

http://math.emory.edu/~lchen41/teaching/2020_Spring/EX_10-3-1.R

E.g. 3 Fisher's suspicion on Mendel's experiments on 1866:

$$d = \dots = 0.47$$

$$P\text{-value} = \mathbb{P}(\chi_3^2 \leq 0.47) = 0.0746.$$

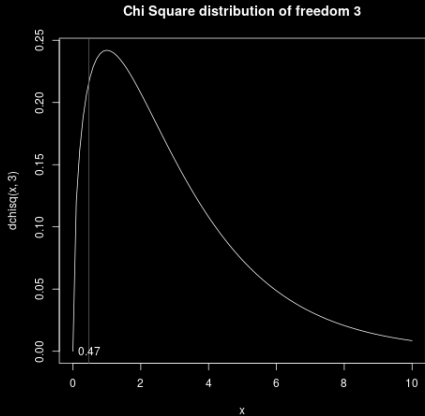
E.g. 3 Fisher's suspicion on Mendel's experiments on 1866:

| Phenotype | Obs. Freq. | Mendel's Model | Exp. Freq. |
|-------------------|------------|----------------|------------|
| (round, yellow) | 315 | 9/16 | 312.75 |
| (round, green) | 108 | 3/16 | 104.25 |
| (angular, yellow) | 101 | 3/16 | 104.25 |
| (angular, green) | 32 | 1/16 | 34.75 |

$$d = \dots = 0.47$$

$$P\text{-value} = \mathbb{P}(\chi_3^2 \leq 0.47) = 0.0746.$$

```
1 > # Case Study 10.3.3
2 > x=seq(0,10,0.1)
3 > plot(x,dchisq(x,3),type = "l")
4 > abline(v=0.47,col = "gray60")
5 > text(0.47,0,"0.47")
6 > title("Chi Square distribution
7 +   of freedom 3")
8 > pchisq(0.47,3)
9 [1] 0.07456892
```



E.g. 2' A second look at the random generator in E.g. 2.

Does it fit the model too well? Find the P -value.

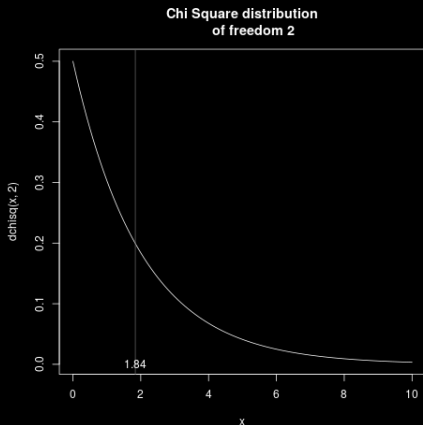
```
1 > # Example 10.3.1
2 > x=seq(0,10,0.1)
3 > plot(x,dchisq(x,2),type = "l")
4 > abline(v=1.84,col = "gray60")
5 > text(1.84,0,"1.84")
6 > title("Chi Square distribution
7 +     of freedom 2")
8 > pchisq(1.84,2)
9 [1] 0.601481
```

P -value = 0.601 \implies No.

E.g. 2' A second look at the random generator in E.g. 2.

Does it fit the model too well? Find the P -value.

```
1 > # Example 10.3.1
2 > x=seq(0,10,0.1)
3 > plot(x,dchisq(x,2),type = "l")
4 > abline(v=1.84,col = "gray60")
5 > text(1.84,0,"1.84")
6 > title("Chi Square distribution
7 +   of freedom 2")
8 > pchisq(1.84,2)
9 [1] 0.601481
```



P -value = 0.601 \implies No.