Assessing Goodness Of Fit

MIT 18.443

Dr. Kempthorne

Spring 2015

Outline

1. Assessing Goodness of Fit
   - Poisson Dispersion Test
   - Hanging Histograms/Chigrams/Rootograms
   - Probability Plots

Poisson Dispersion Test

Poisson Distribution

- Counts of *events* that occur at constant rate
- Counts in disjoint intervals/regions are independent
- If intervals/regions are constant in size, then identical distributions of counts
- Consider data:

$$X_1, X_2, \ldots, X_n \text{ } \textit{independent Poisson}(\lambda_i),$$

  and testing:

  - **Null Hypothesis**

    H_0: X_i are i.i.d. $\textit{Poisson}(\lambda)$.

  - **Alternate Hypothesis**

    H_1: X_i are independent $\textit{Poisson}(\lambda_i)$ (rates vary over i)

Apply Generalized Likelihood Ratio Test

Poisson Dispersion Test

Generalized Likelihood Ratio Test:

- MLE of λ under H_0: $X_1, \ldots, X_n$ *i.i.d. Poisson*(λ)
  $$Lik(\lambda) = \prod_{i=1}^n \left[\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right]$$
  $$\implies \hat{\lambda} = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}$$
- MLEs of λ_i under H_1: $X_i \sim Poisson(\lambda_i), i = 1, \ldots, n$
  $$Lik(\lambda_1, \ldots, \lambda_n) = \prod_{i=1}^n \left[\frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}\right]$$
  $$\implies \tilde{\lambda}_j = x_j, j = 1, \ldots, n$$
- Likelihood Ratio
  $$\begin{aligned} \Lambda &= Lik(\hat{\lambda})/Lik(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n) \\ &= \frac{\prod_{i=1}^n \left[\frac{\hat{\lambda}^{x_i}}{x_i!} e^{-\hat{\lambda}}\right]}{\prod_{i=1}^n \left[\frac{\tilde{\lambda}_i^{x_i}}{x_i!} e^{-\tilde{\lambda}_i}\right]} = \prod_{i=1}^n \left[\left(\frac{\hat{\lambda}}{\tilde{\lambda}_i}\right)^{x_i} e^{-\hat{\lambda}+\tilde{\lambda}_i}\right] \\ &= \prod_{i=1}^n \left(\frac{\overline{x}}{\tilde{x}_i}\right)^{x_i} \cdot e^{-n\overline{x}+\sum_1^n x_i} = \prod_{i=1}^n \left(\frac{\overline{x}}{x_i}\right)^{x_i} \end{aligned}$$

Poisson Dispersion Test

Generalized Likelihood Ratio Test (continued)

- GLR Test Statistic

$$\begin{aligned} LRStat &= -2 \times \log(\Lambda) \\ &= -2 \times \log(\textstyle\prod_{i=1}^{n} \left(\frac{\overline{x}}{x_i}\right)^{x_i}) \\ &= 2 \textstyle\sum_{i=1}^{n} x_j \ln(\frac{x_j}{\overline{x}}) \\ &\approx \frac{1}{\overline{x}} \textstyle\sum_{i=1}^{n} (x_i - \overline{x})^2 = n \times \left(\frac{\hat{\sigma}_x^2}{\overline{x}}\right) \end{aligned}$$

Note: Last line applies Taylor Series Approximation

$$f(x) = x \ln(\tfrac{x}{x_0}) \approx (x - x_0) + \tfrac{1}{2}(x - x_0)^2.$$

- Approximate Distribution under H_0:

$$LRStat \sim \chi^2_q, \text{ where } q = dim(\Theta) - dim(\Theta_0) = n - 1.$$

- H_0 is rejected when $LRStat$ is high $\iff \dfrac{\hat{\sigma}_x^2}{\overline{x}} >> 1$

(For a Poisson Distribution $Var(X) = E(X) = \lambda$.)

Poisson Dispersion Test

Example 9.6.A. Asbestos Fibers

Steel et al. 1980: Counts of asbestos fibers on filters
(from Example 8.4.A)

- Data: $x = c(31, 29, 19, 18, 31, 28, ..., 24)$ (23 values)
- Test Statistic:

$$\begin{aligned} LRStat &= 2\sum_1^n x_j \ln(x_j/\overline{x}) = 27.11 \\ &\approx \frac{1}{\overline{x}}\sum_1^n (x_i - \overline{x})^2 = n\left(\frac{\hat{\sigma}_x^2}{\overline{x}}\right) = 26.56 \end{aligned}$$

- Approximate P-Value:

Asymptotic Distribution: $LRStat \sim \chi^2_q$, with $q = n - 1 = 22$.
$P - Value = 0.2072$.

Outline

1. Assessing Goodness of Fit
   - Poisson Dispersion Test
   - Hanging Histograms/Chigrams/Rootograms
   - Probability Plots

Hanging Histograms

Histograms

- Random sample from distribution with cdf $F(x \mid \theta)$.

  Sample data: $x_1, x_2, \ldots, x_n$
- m interval bins in histogram:

  $bin_j = (b_j, b_{j+1}]$, for $j = 1, 2, \ldots, m$.
- m bin counts in histogram

  $n_j = \#(x_i \in bin_j) = \#(\{x_i : b_j < x_i \leq b_{j+1}\})$,
- Evaluate Goodness-of-Fit of $F(x \mid \theta)$
  - Expected Counts

    $$\begin{aligned} \hat{n}_j &= np_j \\ \text{where} \quad p_j &= F(b_{j+1} \mid \theta) - F(b_j \mid \theta) \end{aligned}$$
  - Observed Counts

    $n_j \sim Binomial(n, p_j)$
  - Hanging Histogram: Instead of plotting n_j,

    use $(n_j - \hat{n}_j)$ in Histogram.
  - Correct for non-constant $Var(n_j - \hat{n}_j) = np_j(1 - p_j)$

Hanging Histogram

Hanging Chigram

- Hanging Histogram: Instead of plotting n_j,
  use $(n_j - \hat{n}_j)$ in Histogram.
- Correct for non-constant $Var(n_j - \hat{n}_j) = np_j(1 - p_j)$
  use $\frac{(n_j - \hat{n}_j)}{\sqrt{\hat{n}_j}} \approx \frac{(n_j - \hat{n}_j)}{\sqrt{np_j(1-p_j)}}$

  Note: $\left[\frac{(n_j - \hat{n}_j)}{\sqrt{\hat{n}_j}}\right]^2 = \frac{(O_j - E_j)^2}{E_j}$

Hanging Histogram (continued)

Hanging Rootogram

- Hanging Rootogram: Instead of plotting n_j,
  use $\sqrt{n_j} - \sqrt{\hat{n}_j}$ in Histogram
- $g(x) = \sqrt{x}$ is a **Variance Stabilizing Transformation**
  For a r.v. X: $E[X] = \mu$ and $Var[X] \approx \sigma^2(\mu)$
  Then $Y = g(X)$ is a random variable with
  $$Var[Y] \approx [g'(\mu)]^2 \cdot Var[X] \approx const$$
  (if $g'(\mu) = 1/\sigma(\mu)$)

Outline

1. Assessing Goodness of Fit
   - Poisson Dispersion Test
   - Hanging Histograms/Chigrams/Rootograms
   - Probability Plots

Probability Plots

Sample from a Uniform$(0,1)$ Distribution

- $X_1, X_2, \ldots, X_n$ i.i.d. $Uniform(0,1)$.
- **Def:** Order Statistics ordered sample values
$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$
- CDF and PDF of $X_{(n)}$
$$\begin{aligned} F_{(n)}(x) &= P(\max X_i \leq x) = P(\text{ all } X_i \leq x) \\ &= [P(X_i \leq x)]^n = x^n \\ \Longrightarrow \quad f_{(n)}(x) &= \tfrac{d}{dx} F_{(n)}(x) \\ &= nx^{n-1}. \end{aligned}$$

Note: $E[X_{(n)}] = \int_0^1 x f_{(n)}(x) dx = \frac{n}{n+1}$

Probability Plots

Sample from a Uniform(0,1) Distribution

- CDF and PDF of $X_{(1)}$

$$\begin{aligned} 1 - F_{(1)}(x) &= P(\min X_i > x) = P(\text{ all } X_i > x) \\ &= [P(X_i > x)]^n = (1-x)^n \\ \implies F_{(1)}(x) &= 1 - (1-x)^n \\ \implies f_{(1)}(x) &= \frac{d}{dx}[1 - (1-x)^n] = n(1-x)^{n-1} \end{aligned}$$

Probability Plots

Order Statistics from a Uniform(0,1) Distribution

- PDF of $X_{(j)}$, the jth order statistic $(j = 1, 2, \dots, n)$
  Use the cdf of the original distribution: $F(x) = P(X \leq x)$
  $$f_{(j)}(x) = \left(\frac{n!}{(j-1)!1!(n-j)!} \right) [F(x)]^{(j-1)} f(x) [1 - F(x)]^{(n-j)}$$
- For $F(x) = x$, the cdf of the $Uniform(0,1)$ distribution
  $$\begin{aligned} f_{(j)}(x) &= \left(\frac{n!}{(j-1)!1!(n-j)!} \right) x^{(j-1)} \cdot 1 \cdot [1-x]^{(n-j)} \\ &= \frac{x^{j-1}(1-x)^{n-j+1-1}}{Beta(j, n-j+1)} \end{aligned}$$
  I.e., $X_{(j)} \sim Beta(j, (n-j)+1)$
- By properties of Beta integrals:
  Note: $$\begin{aligned} E[X_{(j)}] &= \frac{Beta(j+1,(n-j)+1)}{Beta(j,(n-j)+1)} = \frac{j}{n+1} \\ Var[X_{(j)}] &= \left(\frac{j}{n+1}\right) \cdot \left(1 - \frac{j}{n+1}\right) \cdot \left(\frac{1}{n+2}\right) \end{aligned}$$

Order Statistics from Sampling a Continuous Distribution

- $X_1, \ldots, X_n$ i.i.d. with cdf $F_X(x)$
  (assumption: $F_X(\cdot)$ is strictly increasing over its range)
- $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$, the order statistics

Definition: Probability Integral Transform

$$Y = F_X(X)$$

- $Y \sim Uniform(0,1)$ (See Rice Proposition C of Section 2.3)
- $Y_i = F_X(X_i)$ are i.i.d. $Uniform(0,1)$
- The jth order statistic $Y_{(j)}$ is
  $Beta(j, n-j+1)$ random variable and
  $$E[Y_{(j)}] = \frac{j}{n+1}$$

Probability Plots

Definition: Probability Plot

- Given a sample $X_1, \ldots, X_n$
- $H_0 : F(\cdot)$ is the cdf of each X_i
- Plot $y = F(X_{(k)})$ vs $x = \frac{k}{n+1}$
  The points should fall close to the line $y = x$ if H_0 is true.

QQ (Quantile-Quantile) Plots

- Plot $y = X_{(k)}$ vs $x = F^{-1}(\frac{k}{n+1})$
- The vertical axis is the observed **quantile** and the horizontal axis is the theoretical quantile of the distribution.

Probability Plots

Normal QQ Plots

- Plot $y = X_{(k)}$ vs $x = F^{-1}(\frac{k}{n+1})$ using $F(\cdot)$ for a $Normal(\mu, \sigma^2)$ distribution
- In terms of the $N(0,1)$ distribution cdf $\Phi(\cdot)$,
$$F(x \mid \mu, \sigma) = \Phi(\tfrac{x-\mu}{\sigma})$$
- Denote $z_k = \Phi^{-1}(\frac{k}{n+1}),\ k = 1, \ldots, n$
the theoretical quantiles of a $N(0,1)$ distribution, then
$$F^{-1}(\tfrac{k}{n+1}) = \mu + \sigma z_k$$
- If the $\{X_i\}$ are a $Normal(\mu, \sigma^2)$ the **Normal QQ Plot** can be graphed as
$$X_{(k)} \text{ versus } z_k \text{ (without using } \mu \text{ and } \sigma)$$
The plot will be close to linear with
$$\text{Intercept} = \mu \text{ and Slope} = \sigma$$
- Note: $F^{-1}(\frac{k}{n+1}) \approx E[X_{(k)}]$ (exact if F is uniform).

Testing for Normality

Goodness of Fit Tests for Normal Distributions

- Normal QQ Plots

  $$\begin{array}{l} \mathit{Filliben}(1975): \text{Accept Normal if} \\ \quad R{-}\text{Squared of } QQ - \mathit{Plot} \text{ is close to } 1. \end{array}$$

- **Skewness Coefficient**

  $$SKEW = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$$

  where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
  Reject Normality if $|SKEW|$ large.

- **Kurtosis Coefficient**

  $$KURT = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{s^4}$$

  where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
  Reject Normality if $|KURT|$ large.