Introduction to Probability and Statistics for Engineers and Scientists

Sheldon M Ross

Solutions

Anirudh Krishnan

September 20, 2021

Contents

Chapter 1

Goodness of Fit Tests and Categorical Data Analysis

"Ch11 Quote Here"

1.1 Introduction

The *a priori* assumption of a probability model governing an observed phenomenon is central to the analysis of samples from an underlying population. The measure appropriateness for this assumed probability model is done through *goodness-of-fit* tests.

The null hypothesis to be tested is that a sample has the specified probability distribution. The parameters of this probability distribution may not be fully specified, leading to a more complex problem.

1.2 Goodness of fit tests with all parameters specified

Consider a set of n independent random variables $\{Y_j\}$ each of which can take on discrete values in the integer set $\{1, \ldots, k\}$. The null hypothesis to test is that they all have the same underlying PMF, specified as

$$\begin{aligned} H_0 &: P\{Y = i\} = p_i \qquad \forall\ i \in \{1, \ldots, k\} \qquad (0.1) \\ H_1 &: P\{Y = i\} \neq p_i \qquad \text{for some } i \in \{1, \ldots, k\} \end{aligned}$$

Defining the set $\{X_i\}$ as the number of RVs $\{Y_j\}$ which have the value i, it follows that the set $\{X_i\}$ are independent binomial RVs with parameters $\{(n, p_i)\}$ under H_0

$$X_i \sim \texttt{Binom}(n,\ p_i) \tag{0.2}$$

$$\mathbb{E}[X_i] = np_i \tag{0.3}$$

A method of judging how close p_i is to the actual probability $P\{Y = i\}$, is to look at a standardized sum of squared errors, and use it to define a test statistic.

Using a significance threshold α, and the fact that T approaches a χ^2 RV with $(k-1)$ DOF as $n \to \infty$,

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{0.4}$$

$$\lim_{n\to\infty} T \to \chi^2_{k-1} \tag{0.5}$$

$$\begin{aligned} \text{reject } H_0 \text{ if} \qquad & T > \chi^2_{\alpha,k-1} \\ \text{accept } H_0 \qquad & \text{otherwise} \end{aligned} \tag{0.6}$$

A rule of thumb for sample sizes in the test above is to ensure that in the set $\{np_i\}$, all values exceed 1 and most exceed 5.

A simpler expression for T exploits the fact that $\sum X_i = n$ and $\sum p_i = 1$. This constraint on the X_i is also responsible for the χ^2 RV having $(k-1)$ DOF.

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{X_i^2}{np_i} \ - \ n \tag{0.7}$$

Simulation-based methods of determining critical region : Until the modern computer age led to computing power being cheap and widely available, the above χ^2 approximation was the only method of defining critical regions for the goodness-of-fit test.

Consider a set of randomly generated variables $\{Y_1^{(1)}, \ldots, Y_n^{(1)}\}$, each having the PMF $P\{Y_j^{(1)} = i\} = p_i$ for $i \in \{1, \ldots, k\}$.

Defining the set $\{X_i^{(1)}\}$ and test statistic $T^{(1)}$ as above,

$$X_i^{(1)} = \text{number of } j : Y_j^{(1)} = i \tag{0.8}$$

$$T^{(1)} = \sum_{i=1}^{k} \frac{(X_i^{(1)} - np_i)^2}{np_i} \tag{0.9}$$

Using the above procedure to generate a large number of test statistics $\{T^{(1)}, \ldots, T^{(r)}\}$ by repetition of the above procedure yields an approximation to the probability distribution of T.

$$P_{H_0}(T \geq t) \approxeq \frac{\text{number of } l : T^{(l)} \geq t}{r} \tag{0.10}$$

The above approximation becomes very accurate for large r and can also be used then to calculate a p-value for the test. The generation of a random set $\{Y^{(r)}\}$ exploits the Monte-Carlo system of using a standard uniform RV transformed using the set of probabilities $\{p_i\}$ to output the discrete value of $Y_j^{(r)}$.

1.3 Goodness of fit tests with some parameters unspecified

When the underlying probability distribution is not fully specified, a general strategy is to divide the possible continuous set of outcomes into a few discrete regions. Using the observed set of data points to calculate an estimate for the unspecified parameters, an estimated test statistic can then be calculated.

$$P\{Y_j = i\} = p_i \qquad \text{where} \qquad \widehat{p_i} \approx p_i \tag{0.11}$$

$$T = \sum_{i=1}^{k} \frac{(X_i - n\widehat{p_i})^2}{n\widehat{p_i}} \tag{0.12}$$

$$\lim_{n\to\infty} T \to \chi^2_{k-1-m} \tag{0.13}$$

$$\begin{aligned} \text{reject } H_0 \text{ if} \qquad & T > \chi^2_{\alpha,k-1-m} \\ \text{accept } H_0 \qquad & \text{otherwise} \end{aligned} \tag{0.14}$$

In the calculation of $\widehat{p_i}$ above, the CDF of the underlying probability distribution assumed by H_0, with estimated parameters $\{\widehat{\lambda}\}$ is used along with the user-defined discrete outcomes. The χ^2 RV has $(k-1-m)$ DOF if there are m unspecified parameters to be estimated.

For example, a set of observed data with a null hypothesis of the underlying distribution being Poisson, would involve estimating the unspecified mean of the Poisson distribution $\widehat{\lambda}$ using the observations and then calculating the test statistic and p-value.

1.4 Tests of independence in contingency tables

Consider a population whose members are governed by two characteristics (X, Y) each of which can take (r, s) possible values. The marginal probabilities can then be calculated as,

$$P_{ij} = P\{X = i, Y = j\} \qquad i \in \{1, \ldots, r\} \qquad j \in \{1, \ldots, s\} \tag{0.15}$$

$$p_i = P\{X = i\} = \sum_j P_{ij} \qquad i \in \{1, \ldots, r\} \tag{0.16}$$

$$q_j = P\{Y = j\} = \sum_i P_{ij} \qquad j \in \{1, \ldots, s\} \tag{0.17}$$

The null hypothesis of interest here is to test the independence of the X and Y characteristics.

$$
\begin{aligned}
H_0 : P_{ij} &= p_i q_j \qquad \forall \text{ possible pairs } (i, j) && (0.18)\\
H_1 : P_{ij} &\neq p_i q_j \qquad \text{for some pair } (i, j) && (0.19)
\end{aligned}
$$

Let the set of n observations be arranged into a *contingency table* where each element N_{ij} represents the number of observations with $X = i, Y = j$. The marginal probabilities can be estimated from the data set as

$$
\begin{aligned}
N_i &= \sum_j N_{ij} \qquad & \widehat{p}_i &= \frac{N_i}{n} && (0.20)\\
M_j &= \sum_i N_{ij} \qquad & \widehat{q}_j &= \frac{M_j}{n} && (0.21)
\end{aligned}
$$

When H_0 is true, a test statistic can be set up as,

$$
\begin{aligned}
\mathbb{E}[N_{ij}] = nP_{ij} &= p_i q_j \qquad \text{assuming } H_0 \text{ true} && (0.22)\\
T &= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n\widehat{p}_i\widehat{q}_j)^2}{n\widehat{p}_i\widehat{q}_j}\\
&= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{N_{ij}^2}{n\widehat{p}_i\widehat{q}_j} \; - \; n && (0.23)
\end{aligned}
$$

The reduction in DOF is $(1 + (r-1) + (s-1))$. This leads to $T \sim \chi^2_{(r-1)(s-1)}$ since there are a total of $r \times s$ possible categories into which each observation can belong.

$$
\lim_{n\to\infty} T \to \chi^2_{(r-1)(s-1)} \qquad (0.24)
$$

$$
\begin{aligned}
\text{reject } H_0 \text{ if} \qquad & T > \chi^2_{\alpha,(r-1)(s-1)}\\
\text{accept } H_0 \qquad & \text{otherwise} && (0.25)
\end{aligned}
$$

1.5 Tests of independence in contingency tables with fixed marginal totals

If the sample is chosen such that the row sum and/or column sum is fixed across all rows and/or columns, the procedure used in the above section is largely unchanged. Defining the sample incidence of each pair of characteristics $\widehat{e}_{ij}$, and then a test statistic,

$$\widehat{e}_{ij} = n\widehat{p}_i\widehat{q}_j = \frac{N_i M_j}{n} \tag{0.26}$$

$$T = \sum_i \sum_j \frac{(N_{ij} - \widehat{e}_{ij})^2}{\widehat{e}_{ij}} \tag{0.27}$$

Here, N_i and M_j are the row-sums and column-sums respectively. The rest of the test is also unchanged with the use of a χ^2 RV with $(r-1)(s-1)$ DOF used to calculate the critical regions.

An extension of the above procedure can be used to test the hypothesis that m populations with each member taking on one of n possible values, all have the same discrete population distribution.

Value	Population					Row Sum
	1	2	...	j	n	
1	N_{11}	N_{12}	...	N_{1j}	N_{1n}	M_1
2	N_{21}	N_{22}	...	N_{2j}	N_{2n}	M_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
i	N_{i1}	N_{i2}	...	N_{ij}	N_{in}	M_i
m	N_{m1}	N_{m2}	...	N_{mj}	N_{mn}	M_m
Column Sum	N_1	N_2	...	N_j	N_n	

The hypothesis above now reduces to the absence of a row effect in the table of observations. $H_0 : p_{1j} = p_{2j} = \cdots = p_{mj}$

1.6 Kolmogorov-Smirnov goodness of fit test for continuous data

Given a set of samples $\{Y_i\}$ from an underlying population distribution, the hypothesis testing whether this distribution is some continuous CDF given by F can be performed using the discretization procedure from the previous section.

Let the range $(-\infty, \infty)$ be broken up into k parts. Now, the observations can belong to one of these k categories as

$$Y_j^d = i \qquad \text{if } Y_j \in (y_{i-1}, y_i) \tag{0.28}$$

$$P\{Y_j^d = i\} = F(y_i) - F(y_{i-1}) \qquad \forall\ i \in \{1, \ldots, k\} \tag{0.29}$$

This creates a model that is amenable to the χ^2 goodness-of-fit test outlined in the above sections.

Consider the alternative approach which involves estimating the CDF as an empirical distribution function F_e,

$$F_e(x) = \frac{\text{number of } i : Y_i \leq x}{n} \tag{0.30}$$

Here, $F_e(x)$ is the proportion of observations that are less than or equal to x. Since $F_e(x)$ is an estimator of $F(x)$ when H_0 is true, the *Kolmogorov-Smirnov* test statistic is

$$D \equiv \max_x \left| F_e(x) - F(x) \right| \tag{0.31}$$

$F_e(x)$ is a step-like function with step size $1/n$ and jumps at each of the data points $\{Y_j\}$ after they have been rearranged into ascending order as $\{Y_{(j)}\}$.

$$F_e(x) = \begin{cases} 0 & \text{if } x \in (-\infty,\ Y_{(1)}) \\ 1/n & \text{if } x \in (Y_{(1)},\ Y_{(2)}) \\ \dots & \\ (n-1)/n & \text{if } x \in (Y_{(n-1)},\ Y_{(n)}) \\ 1 & \text{if } x \in (Y_{(n)},\ \infty) \end{cases} \tag{0.32}$$

Since $F(x)$ itself is a monotonically increasing function, the expression $|F_e(x) - F(x)|$ must have its maximum close to one of the points $x = \{Y_{(j)}\}$.

$$D = \max_j \left\{ \frac{j}{n} - F(y_{(j)}),\ F(y_{(j)}) - \frac{j-1}{n} \right\} \tag{0.33}$$

A p-value defined using this statistic does not depend on the choice of underlying distribution F,

$$p = P_F(D \geq d) = P_F \left\{ \max_x \left| \frac{\#i : Y_i \leq x}{n} - F(x) \right| \geq d \right\} \tag{0.34}$$

$$= P \left\{ \max_x \left| \frac{\#i : U_i \leq F(x)}{n} - F(x) \right| \geq d \right\} \tag{0.35}$$

The above uses the fact that if Y has a continuous CDF F, then $F(Y)$ is a standard uniform RV. This enables the use of independent standard uniform RVs $\{U_i\}$ to ease the Monte-Carlo simulation of the p-value.

The Monte-Carlo procedure involves defning $y = F(x)$ for the hypothesized CDF F and then performing many repeats of checking whether the following inequality holds,

$$\text{MC iteration is } \max_{0\le y\le 1}\left|\frac{\#i : U_i \le y}{n} - y\right| \ge d \tag{0.36}$$

$$\max_y \left|\frac{\#i : U_i \le y}{n} - y\right| = \max_j \left|\frac{j}{n} - U_{(j)},\ U_{(j)} - \frac{j-1}{n}\right| \tag{0.37}$$

Chapter 2

Goodness of Fit Tests and Categorical Data Analysis

1 Table of observations and results of goodness-of-fit tests are tabulated below.

	X_i	p_i
White	141	0.25
Pink	291	0.5
Red	132	0.25
Total	564	1

Goodness of Fit Test	
Test Statistic	8.62e-01
p value %	65.00
Significance (α) %	5.00
null hypothesis (H_0)	accepted
minimum np_i	141

2 Table of observations and results of goodness-of-fit tests are tabulated below.

	X_i	p_i
1	158	0.1667
2	172	0.1667
3	164	0.1667
4	181	0.1667
5	160	0.1667
6	165	0.1667
Total	1000	1

Goodness of Fit Test	
Test Statistic	1.98e+00
p value %	85.20
Significance (α) %	5.00
null hypothesis (H_0)	accepted
minimum np_i	167

3 Table of observations and results of goodness-of-fit tests for an underlying Poisson distribution are tabulated below.

Failures	X_i	p_i
0	0	0.015
1	5	0.063
2	22	0.1323
3	23	0.1852
4	32	0.1944
5	22	0.1633
6	19	0.1143
7	13	0.0686
8	6	0.036
9	4	0.0168
10	4	0.0071
11	0	0.0027
Total	150	1

Goodness of Fit Test	
Test Statistic	1.66e+01
p value %	12.16
Significance (α) %	5.00
null hypothesis (H_0)	accepted
minimum np_i	0

4 No data given.

5 Table of observations and results of goodness-of-fit tests for an underlying exponential distribution with mean 50 are tabulated below.

Lifetime	X_i	p_i
< 30	41	0.4512
30 - 60	31	0.2476
60 - 90	13	0.1359
> 90	15	0.1653
Total	100	1

Goodness of Fit Test	
Test Statistic	2.11e+00
p value %	54.89
Significance (α) %	5.00
null hypothesis (H_0)	accepted
minimum np_i	14

6 Table of observations and results of goodness-of-fit tests are tabulated below.

Grade	X_i	p_i
Top	234	0.4
High	117	0.3
Medium	81	0.2
Low	68	0.1
Total	500	1

Goodness of Fit Test	
Test Statistic	2.31e+01
p value %	0.00
Significance (α) %	5.00
null hypothesis (H_0)	rejected
minimum np_i	50