Introduction to Probability and Statistics for Engineers and Scientists

Sheldon M Ross

Notes and Exercises

Anirudh Krishnan

September 19, 2021

Contents

Chapter 10

Analysis of Variance

"What do you mean 'Its a method used to compare mean values'?"

10.1 Introduction

Consider the problem of testing whether alternative approaches to solving a problem are equivalent. A simple null hypothesis is to ask if the average benefit of each problem-solving approach is the same.

A standard procedure to test this would be to randomly divide a population into many subgroups, each subjected to a different solution. The null hypothesis, under the assumption that the response of each individual tested has a variance independent of the solution itself, is simply the subgroup-means being equal.

10.2 Overview of the procedure

Consider a set of m populations from which samples of size n are drawn. Proving the equality of the means of these populations, which only depend on one parameter, namely the population the sample was from, is called a *one-way* analysis of variance.

This can also be extended to the more general case of the m different samples not being of equal size.

The more complex case of the variance of each variable depending on two factors, the population index as well as the sample index, can be visually represented as a two-dimensional array with row and column indices both determining the variance. This is called the *two-way* analysis of variance problem.

The above problem assumes no interaction between the two factors affecting the mean of the variable. Removing this assumption makes for a significantly more complicated problem, with possible non-linear interactions among the factors affecting the mean.

For the rest of this chapter, a general procedure can be outlined as follows,

- Assume samples are drawn from a normal population with unknown (but common) variance σ^2
- Find a valid estimator of σ^2, which is true regardless of the specific null hypothesis being true.
- Fina another estimator of σ^2 which is true only when H_0 holds.
- Design a test that rejects H_0 when the second estimator is sufficiently larger than the first (since it tends to always be larger).

Consider a set of N independent normal RVs $\{X_i\}$ with possibly different means but a common unknown variance.

$$\begin{aligned} \frac{X_i - \mu_i}{\sigma} &\sim Z_i \\ \sum_{i=1}^{N} \frac{(X_i - \mu_i)^2}{\sigma^2} &\sim \chi^2_N \end{aligned} \tag{10.1}$$

If each $\mu_i = \mathbb{E}[X_i]$ can be estimated using a linear combination of k parameters, then the resulting estimate of the means $\{\widehat{\mu_i}\}$, can be substituted into the above expression to make a χ^2 distribution with $(N - k)$ DOF.

10.3 One-way ANOVA

Consider a set of m samples, each of size n (*balanced samples*). The members of each sample are independent normal RVs and a hypothesis testing the equality of their means,

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \qquad i = 1, \ldots, m \qquad j = 1, \ldots, n \tag{10.2}$$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \qquad \text{vs} \qquad H_1 : \text{not all means are equal} \tag{10.3}$$

Using the set of sample means $\{X_{i\bullet}\}$ of the n populations as the estimators for $\{\mu_i\}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(X_{ij} - \mathbb{E}[X_{ij}])^2}{\sigma^2} \sim \chi^2_{nm} \tag{10.4}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(X_{ij} - X_{i\bullet})^2}{\sigma^2} \sim \chi^2_{nm-m} \tag{10.5}$$

Defining the *within-samples sum of squares* (SS_W) as

$$SS_W = \sum_i \sum_j (X_{ij} - X_{i\bullet})^2 \tag{10.6}$$

$$\sum_i \sum_j \frac{SS_W}{\sigma^2} \sim \chi^2_{nm-m} \tag{10.7}$$

$$\mathbb{E}\left[\frac{SS_W}{\sigma^2}\right] = nm - m \tag{10.8}$$

$$\sigma^2 = \frac{SS_W}{nm - m} \tag{10.9}$$

The above estimator of σ^2 is free of any assumptions about the H_0. Next, under the assumption that H_0 holds, and thus the set of population means $\{X_{i\bullet}\}$ are all normal RVs with common mean μ and common variance σ^2/n,

$$n \sum_{i=1}^{m} \frac{(X_{i\bullet} - \mu)^2}{\sigma^2} \sim \chi^2_m \tag{10.10}$$

$$X_{\bullet\bullet} = \frac{\sum_i \sum_j X_{ij}}{nm} = \frac{\sum_i X_{i\bullet}}{m} \tag{10.11}$$

$$\tag{10.12}$$

Since the above mean of all variables $X_{\bullet\bullet}$ is an estimator of the common mean μ which requires only 1 parameter to calculate, the *between-samples sum of squares* (SS_b) can be used to simplify the above expression,

$$SS_b = n \sum_{i=1}^{m} (X_{i\bullet} - X_{\bullet\bullet})^2 \tag{10.13}$$

$$\frac{SS_b}{\sigma^2} \sim \chi^2_{m-1}$$

$$\frac{SS_b}{(m-1)} = \sigma^2 \tag{10.14}$$

Since the conditional estimator of σ^2 tends to be larger and these two estimators are independent when H_0 holds, the ratio of these two χ^2 RVs to define an f-RV, the hypothesis test can be formulated as,

$$\frac{SS_b}{SS_W} \frac{(nm-m)}{(m-1)} \sim F_{m-1,nm-m}$$

$$H_0 : \text{all means are equal} \qquad \text{vs.} \qquad H_1 : \text{not all means are equal}$$

$$\text{reject } H_0 \text{ if} \qquad \frac{SS_b}{SS_W} \frac{(nm-m)}{(m-1)} > F_{\alpha,m-1,nm-m}$$

$$\text{accept } H_0 \qquad \text{otherwise} \tag{10.15}$$

A computational identity relating the two terms SS_b and SS_W is as follows,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2 = nm \; X_{\bullet\bullet}^2 + SS_b + SS_W \tag{10.16}$$

A proof of the relation between the two estimates of σ^2, uses the following variable transformation,

$$\mathbb{E}\left[\frac{SS_b}{(m-1)}\right] \geq \sigma^2 \qquad \text{equality only if } H_0 \text{ holds} \tag{10.17}$$

$$Y_i = X_{i\bullet} - \mu_i + \mu_\bullet$$

$$\mathbb{E}\left[\frac{\sum_i (X_{i\bullet} - X_{\bullet\bullet})^2}{(m-1)}\right] = \frac{\sigma^2}{n} + \sum_i \frac{(\mu_i - \mu_\bullet)^2}{(m-1)} \tag{10.18}$$

Multiple comparisons of sample means : When the null hypothesis above is rejected, a measure of how the different sample means are related is the T-method. This gives a confidence interval for the difference between all possible pairs $\mu_i - \mu_j$.

$$\mu_i - \mu_j \in X_{i\bullet} - X_{j\bullet} \pm W \tag{10.19}$$

$$\forall\ i \neq j \qquad \text{with probability } 1-\alpha$$

$$W = \frac{1}{\sqrt{n}}\ q(m, nm-m, \alpha)\ \sqrt{\frac{SS_W}{(nm-m)}} \tag{10.20}$$

Here, the studentized range distribution q has pre-calculated CDF tables. Further information in article - *Tukey's range test.*

Unequal sample sizes (unbalanced) : Since the assumption of sample sizes all being equal to n is no longer valid, it is replaced with sample sizes $\{n_1, n_2, \ldots, n_m\}$ for each of the m samples.

For convenience, define $N = \sum_i n_i$.

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(X_{ij} - \mu_i)^2}{\sigma^2} \sim \chi^2_N \tag{10.21}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(X_{ij} - X_{i\bullet})^2}{\sigma^2} \sim \chi^2_{N-m} \tag{10.22}$$

$$\tag{10.23}$$

Once again, defining SS_W and simplifying yields an unbiased estimator of σ^2 not dependent on H_0.

$$SS_W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - X_{i\bullet})^2 \tag{10.24}$$

$$\frac{SS_W}{N-m} = \sigma^2 \tag{10.25}$$

Next, under the assumption that H_0 holds, and the means are equal to a common value μ.

$$\begin{aligned} \mathbb{E}[X_{i\bullet}] &= \mu & \qquad \mathrm{Var}(X_{i\bullet}) &= \frac{\sigma^2}{n_i} \\ \sum_{i=1}^{m} \frac{(X_{i\bullet} - \mu)^2}{\sigma^2/n_i} &\sim \chi^2_m & \qquad \sum_{i=1}^{m} \frac{(X_{i\bullet} - X_{\bullet\bullet})^2}{\sigma^2/n_i} &\sim \chi^2_{m-1} \end{aligned} \tag{10.26}$$

$$SS_b = \sum_{i=1}^{m} n_i \, (X_{i\bullet} - X_{\bullet\bullet})^2 \tag{10.27}$$

When H_0 is true, both the estimators of σ^2 are unbiased and independent. This enables the construction of a level α hypothesis test

$$\frac{SS_b}{SS_W} \, \frac{(N-m)}{(m-1)} \sim F_{m-1,N-m}$$

$$\begin{aligned} H_0 : \text{all means are equal} \qquad & \text{vs.} \qquad H_1 : \text{not all means are equal} \\ \text{reject } H_0 \text{ if} \qquad & \frac{SS_b}{SS_W} \, \frac{(N-m)}{(m-1)} > F_{\alpha,m-1,N-m} \\ \text{accept } H_0 \qquad & \text{otherwise} \end{aligned} \tag{10.28}$$

Balanced samples are more robust to violations of the assumption of equal variances of the different populations and thus preferred over unbalanced samples.

10.4 Two-factor ANOVA

Consider the more complex where each data value in the data-set is affected by two factors, with (m, n) *levels* respectively. The dataset can be arranged into a two-dimensional array along these two indices.

$$\begin{bmatrix} X_{11} & X_{12} & X_{13} & \dots & X_{1n} \\ X_{21} & X_{22} & X_{23} & \dots & X_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & X_{i3} & \dots & X_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & X_{m3} & \dots & X_{mn} \end{bmatrix} \tag{10.29}$$

By convention the factors are called the *row-factor* and *column-factor* respectively. The mean value of a data point depends additively on both its row and column index.

Let $\mu_{ij} = \mathbb{E}[X_{ij}]$, where $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$.

$$\mu_{ij} = a_i + b_j \qquad \mu = \mu_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\mu_{ij}}{nm} \tag{10.30}$$

$$\mu_{\bullet j} = a_\bullet + b_j \qquad \mu_{i\bullet} = a_i + b_\bullet \tag{10.31}$$

$$\alpha_i = a_i - a_\bullet \qquad \beta_j = b_j - b_\bullet \tag{10.32}$$

The *grand mean* is defined as the mean of all terms $\mu \coloneqq \mu_{\bullet\bullet}$, and the terms α_i, β_j are the row and column *deviations* from the grand mean.

The model can now be rewritten as

$$\mu_{ij} = \mathbb{E}[X_{ij}] = \mu + \alpha_i + \beta_j \tag{10.33}$$

$$\sum_i \alpha_i = \sum_j \beta_j = 0 \tag{10.34}$$

Using the row-average, column-average and overall-average of the data points, $X_{i\bullet},\ X_{\bullet j},\ X_{\bullet\bullet}$ to calculate the estimators of α_i, β_j, μ respectively,

$$\mathbb{E}[X_{i\bullet}] = \mu + \alpha_i \qquad \mathbb{E}[X_{\bullet j}] = \mu + \beta_j \tag{10.35}$$

$$\mathbb{E}[X_{i\bullet} - X_{\bullet\bullet}] = \alpha_i \qquad \mathbb{E}[X_{\bullet j} - X_{\bullet\bullet}] = \beta_j \tag{10.36}$$

This leads to the following unbiased estimators of the model parameters.

$$\begin{aligned} \widehat{\mu} &= X_{\bullet\bullet} \\ \widehat{\alpha_i} &= X_{i\bullet} - X_{\bullet\bullet} \\ \widehat{\beta_j} &= X_{\bullet j} - X_{\bullet\bullet} \end{aligned} \tag{10.37}$$

10.5 Two-factor ANOVA : Hypothesis testing

The most common hypothesis to be tested is the absence of any effect due to the row-factors or the column-factors.

$$\begin{aligned} H_0 &: \text{all } \alpha_i \text{ are zero} \quad \text{vs} \quad H_1 : \text{not all } \alpha_i \text{ are zero} \\ H_0 &: \text{all } \beta_j \text{ are zero} \quad \text{vs} \quad H_1 : \text{not all } \beta_j \text{ are zero} \end{aligned}$$

Using the usual ANOVA procedure of comparing two different estimators of the variance,

$$\begin{aligned} \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(X_{ij} - \mathbb{E}[X_{ij}])^2}{\sigma^2} &\sim \chi^2_{mn} \\ \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(X_{ij} - \widehat{\mu} - \widehat{\alpha_i} - \widehat{\beta_j})^2}{\sigma^2} &= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(X_{ij} + X_{\bullet\bullet} - X_{i\bullet} - X_{\bullet j})^2}{\sigma^2} \end{aligned} \tag{10.38}$$

The number of parameters required to estimate the above 3 parameters is $1+(m-1)+(n-1)=(m+n-1)$. This leaves a χ^2 RV with $mn-(m+n-1)=(m-1)(n-1)$ DOF. Using the definiton of the *error sum of squares*,

$$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{\bullet\bullet} - X_{i\bullet} - X_{\bullet j})^2 \tag{10.39}$$

$$\frac{SS_e}{(m-1)(n-1)} = \sigma^2 \tag{10.40}$$

To find the H_0 conditioned estimator of σ^2, first consider the case where all of the row-factors are zero.

$$\mathbb{E}[X_{i\bullet}] = \mu + \alpha_i = \mu \qquad\qquad \mathrm{Var}(X_{i\bullet}) = \frac{\sigma^2}{n}$$

$$\text{if } H_0 \text{ is true,} \qquad\qquad n\sum_{i=1}^{m}\frac{(X_{i\bullet}-\mu)^2}{\sigma^2} \sim \chi^2_m \tag{10.41}$$

$$n\sum_{i=1}^{m}\frac{(X_{i\bullet}-X_{\bullet\bullet})^2}{\sigma^2} \sim \chi^2_{m-1} \tag{10.42}$$

Defining the *row sum of squares* and analogously the *column sum of squares*,

$$SS_r = n\sum_{i=1}^{m}(X_{i\bullet} - X_{\bullet\bullet})^2 \qquad\qquad SS_c = m\sum_{j=1}^{n}(X_{\bullet i} - X_{\bullet\bullet})^2 \tag{10.43}$$

Using the above definitions, a second estimator for σ^2 conditioned on H_0 being true is,

$$\frac{SS_r}{(m-1)} = \sigma^2 \qquad\qquad \frac{SS_r}{SS_e}\,\frac{(m-1)(n-1)}{(m-1)} \sim F_{m-1,(m-1)(n-1)} \tag{10.44}$$

The hypothesis test can now be constructed using this F-RV as follows,

$$\begin{aligned} H_0 : \text{all row-factors are zero} \qquad & \text{vs.} \qquad H_1 : \text{not all row-factors are zero} \\ \text{reject } H_0 \text{ if} \qquad & \frac{SS_r}{SS_e}\,\frac{(m-1)(n-1)}{(m-1)} > F_{\alpha,m-1,(m-1)(n-1)} \\ \text{accept } H_0 \qquad & \text{otherwise} \end{aligned} \tag{10.45}$$

10.6 Two-way ANOVA with interaction

Extending the model in the previous section to allow the possibility of some interaction between the row and column factors governing a particular data-point,

$$\mu_{ij} = \mathbb{E}[X_{ij}] \qquad \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{10.46}$$

$$\mu = \mu_{\bullet\bullet} \qquad \gamma_{ij} = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} \tag{10.47}$$

$$\alpha_i = \mu_{i\bullet} - \mu_{\bullet\bullet} \qquad \beta_j = \mu_{\bullet j} - \mu_{\bullet\bullet} \tag{10.48}$$

$$\sum_{i=1}^{m} \gamma_{ij} = 0 \qquad \sum_{j=1}^{n} \gamma_{ij} = 0 \tag{10.49}$$

In addition, the usual constraints on the set of $\{\alpha_i\}$ and $\{\beta_j\}$ summing to zero still apply. In this extended mode, the term γ_{ij} is called the *interaction of row i and column j*. It measures the difference between the mean of X_{ij} and the three terms indicating the grand mean, row effect and column effect.

For mathematical reasons, it is no longer sufficient to have just one observation for a given row and column index. Assume there are a set of l such observations corresponding to every possible pair of $\{i, j\}$. Since each data point X_{ijk} is still a normal RV with common unknown variance σ^2,

$$X_{ijk} \sim \mathcal{N}(\mu + \alpha_i + \beta_j + \gamma_{ij},\ \sigma^2) \qquad k \in \{1, \ldots, l\} \qquad (10.50)$$
$$i \in \{1, \ldots, m\} \qquad j \in \{1, \ldots n\}$$

To find the estimators for the three sets of parameters defined above, use,

$$\mathbb{E}[X_{ij\bullet}] = \mu_{ij} \qquad \mathbb{E}[X_{\bullet\bullet\bullet}] = \mu \qquad (10.51)$$
$$\mathbb{E}[X_{i\bullet\bullet}] = \mu + \alpha_i \qquad \mathbb{E}[X_{\bullet j\bullet}] = \mu + \beta_j \qquad (10.52)$$

Using the fact that substituting the parameters above with their unbiased estimators will lead to a reduction in DOF of the χ^2 RV,

$$\begin{aligned}
\widehat{\mu} &= X_{\bullet\bullet\bullet} \\
\widehat{\alpha_i} &= X_{i\bullet\bullet} - X_{\bullet\bullet\bullet} \\
\widehat{\beta_j} &= X_{\bullet j\bullet} - X_{\bullet\bullet\bullet} \\
\widehat{\gamma_{ij}} &= X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet}
\end{aligned} \qquad (10.53)$$

For hypothesis tests asking whether one of the three sets of parameters are all zeros, arrange χ^2RVs as follows,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l} \frac{(X_{ijk} - \mathbb{E}[X_{ijk}])^2}{\sigma^2} \sim \chi^2_{mnl}$$
$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l} \frac{(X_{ij} - \widehat{\mu} - \widehat{\alpha_i} - \widehat{\beta_j} - \widehat{\gamma_{ij}})^2}{\sigma^2} \sim \chi^2_{mnl-mn} \qquad (10.54)$$

The number of parameters needed to estimate $\widehat{\mu},\ \widehat{\alpha_i},\ \widehat{\beta_j},\ \widehat{\gamma_{ij}}$ are $1,\ (m-1),\ (n-1),\ (m-1)(n-1)$ respectively. This means that the loss in DOF is mn.

Now, defining the *error sum of squares* SS_e as

$$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(X_{ijk} - X_{ij\bullet})^2 \tag{10.55}$$

$$\frac{SS_e}{(mnl - mn)} = \sigma^2 \tag{10.56}$$

Having found the true estimator of the variance, now consider the estimator conditioned on the null hypothesis that there are no interaction terms. Now, each of the variables $X_{ij\bullet}$ is averaged over l data points, and thus

$$H_0^{int} : \gamma_{ij} = 0 \qquad \forall\ i, j$$

$$\mathbb{E}[X_{ij\bullet}] = \mu + \alpha_i + \beta_j \qquad \mathrm{Var}(X_{ij\bullet}) = \frac{\sigma^2}{l} \tag{10.57}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij\bullet} - \mu - \alpha_i - \beta_j)^2}{\sigma^2/l} \sim \chi^2_{mn} \tag{10.58}$$

From an earlier section, replacing the parameters above with their estimators involes an $m + n - 1$ loss in DOF. Defining the *non-interaction sum of squares* (SS_{int})

$$SS_{int} = \sum_{i=1}^{m}\sum_{j=1}^{n}\ l\ (X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2 \tag{10.59}$$

$$\frac{SS_{int}}{(n-1)(m-1)} = \sigma^2 \qquad \text{if } H_0^{int} \text{ holds} \tag{10.60}$$

Using the two estimators of σ^2 above, an F-RV can be created to test the null hypothesis at significance level a as follows,

$$H_0^{int} : \text{all } \gamma ij \text{ are zero} \qquad \text{vs.} \qquad H_1^{int}$$

$$\text{reject } H_0 \text{ if} \qquad \frac{SS_{int}}{SS_e} \frac{(mnl - mn)}{(n-1)(m-1)} > F_{a,(m-1)(n-1),mnl-mn}$$

$$\text{accept } H_0 \qquad \text{otherwise} \tag{10.61}$$

Similar hypothesis tests for row-factors or column-factors being all zero, involve the *row sum of squares* and *column sum of squares* respectively. For example the null hypothesis H_0^r : all α_i are zero can be tested using

$$SS_r = \sum_{i=1}^{m} nl \ (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet})^2 \tag{10.62}$$

$$\frac{SS_r}{SS_e} \frac{(mnl - mn)}{(m-1)} \sim F_{m-1,mnl-mn} \tag{10.63}$$

An analogous expression can be written for the null hypothesis H_0^c : all β_j are zero.

Note that the procedure in the later sections is a straightforward increase in complexity from the same procedure in earlier sections which was outlined in the introduction as the general ANOVA approach.

Chapter 11

Goodness of Fit Tests and Categorical Data Analysis

"Ch11 Quote Here"

11.1 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.

11.2 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

$$\begin{aligned} H_0 &: P\{Y = i\} = p_i \qquad \forall\ i \in \{1, \ldots, k\} \\ H_1 &: P\{Y = i\} \neq p_i \qquad \text{for some } i \in \{1, \ldots, k\} \end{aligned} \tag{11.1}$$

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

$$X_i \sim \texttt{Binom}(n,\ p_i) \tag{11.2}$$

$$\mathbb{E}[X_i] = np_i \tag{11.3}$$

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 $n \to \infty$,

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{11.4}$$

$$\lim_{n\to\infty} T \to \chi^2_{k-1} \tag{11.5}$$

$$\begin{aligned} \text{reject } H_0 \text{ if} \qquad & T > \chi^2_{\alpha,k-1} \\ \text{accept } H_0 \qquad & \text{otherwise} \end{aligned} \tag{11.6}$$

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.

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{X_i^2}{np_i} \ - \ n \tag{11.7}$$

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,

$$X_i^{(1)} = \text{number of } j : Y_j^{(1)} = i \tag{11.8}$$

$$T^{(1)} = \sum_{i=1}^{k} \frac{(X_i^{(1)} - np_i)^2}{np_i} \tag{11.9}$$

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.

$$P_{H_0}(T \geq t) \approxeq \frac{\text{number of } l : T^{(l)} \geq t}{r} \tag{11.10}$$

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 $Y_j^{(r)}$.

11.3 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.

$$P\{Y_j = i\} = p_i \qquad \text{where} \qquad \widehat{p_i} \approx p_i \tag{11.11}$$

$$T = \sum_{i=1}^{k} \frac{(X_i - n\widehat{p_i})^2}{n\widehat{p_i}} \tag{11.12}$$

$$\lim_{n\to\infty} T \to \chi^2_{k-1-m} \tag{11.13}$$

$$\begin{aligned} \text{reject } H_0 \text{ if} \qquad & T > \chi^2_{\alpha,k-1-m} \\ \text{accept } H_0 \qquad & \text{otherwise} \end{aligned} \tag{11.14}$$

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.

11.4 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,

$$P_{ij} = P\{X = i, Y = j\} \qquad i \in \{1, \ldots, r\} \qquad j \in \{1, \ldots, s\} \tag{11.15}$$

$$p_i = P\{X = i\} = \sum_j P_{ij} \qquad i \in \{1, \ldots, r\} \tag{11.16}$$

$$q_j = P\{Y = j\} = \sum_i P_{ij} \qquad j \in \{1, \ldots, s\} \tag{11.17}$$

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

$$H_0 : P_{ij} = p_i q_j \qquad \forall \text{ possible pairs } (i,j) \tag{11.18}$$

$$H_1 : P_{ij} \neq p_i q_j \qquad \text{for some pair } (i,j) \tag{11.19}$$

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

$$N_i = \sum_j N_{ij} \qquad\qquad \widehat{p_i} = \frac{N_i}{n} \tag{11.20}$$

$$M_j = \sum_i N_{ij} \qquad\qquad \widehat{q_j} = \frac{M_j}{n} \tag{11.21}$$

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

$$\mathbb{E}[N_{ij}] = nP_{ij} = p_i q_j \qquad \text{assuming } H_0 \text{ true} \tag{11.22}$$

$$\begin{aligned} T &= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n\widehat{p_i}\widehat{q_j})^2}{n\widehat{p_i}\widehat{q_j}} \\ &= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{N_{ij}^2}{n\widehat{p_i}\widehat{q_j}} \; - \; n \end{aligned} \tag{11.23}$$

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.

$$\lim_{n\to\infty} T \to \chi^2_{(r-1)(s-1)} \tag{11.24}$$

$$\begin{aligned} \text{reject } H_0 \text{ if} \qquad & T > \chi^2_{\alpha,(r-1)(s-1)} \\ \text{accept } H_0 \qquad & \text{otherwise} \end{aligned} \tag{11.25}$$

11.5 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,

$$\widehat{e}_{ij} = n\widehat{p}_i\widehat{q}_j = \frac{N_i M_j}{n} \tag{11.26}$$

$$T = \sum_i \sum_j \frac{(N_{ij} - \widehat{e}_{ij})^2}{\widehat{e}_{ij}} \tag{11.27}$$

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. $H_0 : p_{1j} = p_{2j} = \cdots = p_{mj}$

11.6 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

$$Y_j^d = i \qquad \text{if } Y_j \in (y_{i-1}, y_i) \tag{11.28}$$

$$P\{Y_j^d = i\} = F(y_i) - F(y_{i-1}) \qquad \forall\ i \in \{1, \ldots, k\} \tag{11.29}$$

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,

$$F_e(x) = \frac{\text{number of } i : Y_i \leq x}{n} \tag{11.30}$$

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

$$D \equiv \max_x \left| F_e(x) - F(x) \right| \tag{11.31}$$

$F_e(x)$ 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 $\{Y_{(j)}\}$.

$$F_e(x) = \begin{cases} 0 & \text{if } x \in (-\infty,\ Y_{(1)}) \\ 1/n & \text{if } x \in (Y_{(1)},\ Y_{(2)}) \\ \dots & \\ (n-1)/n & \text{if } x \in (Y_{(n-1)},\ Y_{(n)}) \\ 1 & \text{if } x \in (Y_{(n)},\ \infty) \end{cases} \tag{11.32}$$

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 $x = \{Y_{(j)}\}$.

$$D = \max_j \left\{ \frac{j}{n} - F(y_{(j)}),\ F(y_{(j)}) - \frac{j-1}{n} \right\} \tag{11.33}$$

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

$$p = P_F(D \geq d) = P_F \left\{ \max_x \left| \frac{\#i : Y_i \leq x}{n} - F(x) \right| \geq d \right\} \tag{11.34}$$

$$= P \left\{ \max_x \left| \frac{\#i : U_i \leq F(x)}{n} - F(x) \right| \geq d \right\} \tag{11.35}$$

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,

$$\text{MC iteration is } \max_{0\le y\le 1}\left|\frac{\#i : U_i \le y}{n} - y\right| \ge d \tag{11.36}$$

$$\max_{y}\left|\frac{\#i : U_i \le y}{n} - y\right| = \max_{j}\left|\frac{j}{n} - U_{(j)},\ U_{(j)} - \frac{j-1}{n}\right| \tag{11.37}$$