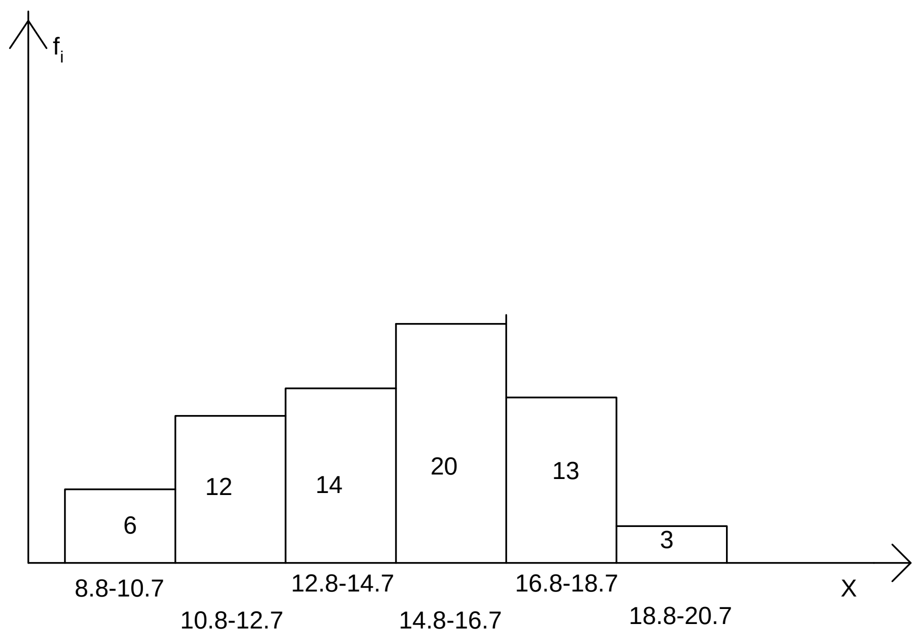
Testing for Goodness-of-Fit **Maghsoodloo**

Reference: Chapter 14 of Devore's 7th Edition

An Example of a Normal Distribution GOF to a Grouped Data

The following histogram describes the empirical distribution of the length of 68 fish caught from a nearby lake, in measured inches.



We wish to test the null hypothesis that the above data have originated from a normal population with unknown μ and σ^2. Since these two parameters are unknown, we must estimate them by the sample statistics $\hat{\mu} = \bar{x}$ and sample variance $\hat{\sigma}^2$, respectively, where $\bar{x} = \sum_{i=1}^{6} m_i f_i / 68$ =14.6618 inches, m_1 = 9.75, m_2 = 11.75, ..., m_6 =19.75 represent the subgroup midpoints, and

$$\hat{\sigma}^2 = \frac{1}{68}\left[\sum_{i=1}^{6} m_i^2 f_i - \left(\sum m_i f_i\right)^2 / n\right] = 7.110 .$$

Therefore, our null hypothesis becomes H_0: X ~ N(14.6618, 7.110 inches^2), or

$$H_0: F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{u-14.6618}{2.66643}\right)^2} du ,$$

where f(x) is the hypothesized underlying distribution of the rv X, and $\hat{\sigma}_x$ = $\sqrt{7.110}$ = 2.66643 is the moment (and also the maximum likelihood) estimator of σ. It is generally best to estimate the parameters from the frequency distribution rather than the raw data.

The expected frequencies must be computed under the above null hypothesis, i.e., assuming that H_0 is true. Thus, $E_1 = np_1$, where p_1 = Pr(Z ≤ (10.75 − 14.6618)/2.66643), Z ~ N(0, 1) and 10.75 equals the upper boundary (UB) of the 1st subgroup. Excel computations led to the summary in Table 30 below, where the last cell expectation, E_6, must be obtained from $E_6 = 68 - \sum_{i=1}^{5} E_i$, and UB = upper boundary.

Table 30

subgroup	f_i	UB	Z_i	$\Phi(Z_i)$	p_i	E_i
8.8– 10.7"	6	10.75	-1.4671	0.07118	0.07118	4.8403
10.8–12.7	12	12.75	-0.7170	0.23670	0.16551	11.2547
12.8–14.7	14	14.75	0.0331	0.51319	0.27650	18.8022
14.8–16.7	20	16.75	0.7831	0.78323	0.27004	18.3624
16.8–18.7	13	18.75	1.5332	0.93739	0.15416	10.4828
18.8–20.7	3	∞	N/A	1.00000	0.06261	4.2576
Sum	68				1.0000	68.0000

In Table 30, f_i represents the observed frequency of the i^{th} subgroup, while $E_i = 68 \times p_i$ is the corresponding expected frequency computed under H_0, where $p_i = \Phi(Z_i) - \Phi(Z_{i-1})$. For example, $p_2 = \Phi(Z_2) - \Phi(Z_{2-1})$ = 0.23670 − 0.07118 = 0.16551,

where $\Phi(Z_0) = \Phi(-\infty) = 0$ and $\Phi(Z)$ is the cdf of a unit normal density. The value of the chi-square Goodness-Of-Fit (GOF) statistic is given by

$$\chi_0^2 = \sum_{i=1}^{k=6}\left[\frac{(f_i - E_i)^2}{E_i}\right] = \sum_{i=1}^{k=6}\left[\frac{(n_i - E_i)^2}{E_i}\right] = \sum_{i=1}^{k=6}\left[\frac{(O_i - E_i)^2}{E_i}\right] = 2.6757. \qquad (84)$$

Note that some authors use n_i or O_i for observed frequencies, but the most prevalent notation for expected frequencies under H_0 is E_i. The use of e_i or $\hat{e}_i$ for the i^{th} expected frequency by some authors can be confusing because in statistical literature, e_i generally stands for the i^{th} residual.

Since the df of the above chi-square statistic is $\nu = k - 1 -$ number of parameters estimated $= 6 - 1 - 2 = 3$, and only the larger values of χ_0^2 lead to the rejection of H_0, we compare the above $\chi_0^2 = 2.6757$ against the 5 percentage point of the rv χ_3^2, which from Table A.7, page 673, is $\chi^2_{0.05,3} = 7.815$. Note that the exact pdf of the GOF statistic in Eq. (84) is not χ^2_ν, but chi-square provides a good approximation to the SMD of $\sum_{i=1}^{k=6}[(f_i - E_i)^2 / E_i]$ when each $E_i \geq 5$. This implies that the sample size n has to be sufficiently large so that each $E_i = n \times p_i \geq 5$, i = 1, 2, ..., k. Generally, grouping the data into different classes for conducting a χ^2 GOF test requires sample sizes $n \geq 30$. However, if E_i's are all equal (i.e., equally-probable classes), then one can group data with an n as small as 20.

Since $\chi_0^2 = 2.6757$ does not exceed the threshold value of 7.815, we cannot reject the assumption of normality. This does not at all imply that the length of fish from this lake is normally distributed, but that the 68 observations do not provide sufficient evidence to the contrary (i.e., the Goodness-Of-Fit of the normal distribution to the data cannot be rejected). The ***P-value*** of the test is

given by $\hat{\alpha}$ = Pr($\chi_3^2 \geq$ 2.6757) = 0.44437, which far exceeds α = 0.05. Note that the larger the *P-value* is, the better the fit! When the *P-value* = 1, the fit is perfect!

<u>Exercise 128.</u> (a) Verify the values of $\bar{x}$, $\hat{\sigma}$, and the values in the 14.8-16.7 subgroup of the above Table 30. Then, test the GOF of the above grouped data to a N(15.0", σ^2) at α = 0.05. (b) Use a spreadsheet to verify the values in the Table 30 above.

For moderate sample sizes, 20 < n $\leq$ 50, grouping the data for testing the GOF is recommended only with equiprobable intervals; see the Example 14.10 on pages 582-583 of Devore to test for normality with 7 equiprobable intervals. For small sample sizes n < 20, only the nonparametric Kolmogorov–Smirnov GOF test is appropriate.

<u>Exercise 129.</u> (a) Study pages 576-585 of Devore and rework the Example 14.10 on pp. 582-583 of Devore's 7th edition, assuming equal probability intervals. (b) Work Exercise 14.23 on page 587 of Devore's 7th edition by dividing the data into 5 equiprobable intervals.

GOF for Testing Discrete Distributions

The χ^2 GOF is applicable when cell Prs depend on unknown parameters, provided that one df is deducted for every parameter that is replaced by its Maximum Likelihood Estimate (MLE). For our purposes, all we need is that the MLE of μ is $\hat{\mu}$ = $\bar{x}$ and the MLE of σ is $\hat{\sigma} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 / n}$; please note that the divisor of the sample variance $\hat{\sigma}^2$ is indeed n [and not (n – 1)]. Therefore, the net df for the GOF statistic χ_0^2 in equation (84) is ν = k – 1 – m, where m is the number of unknown parameters that are estimated from the data by their MLEs. Further, it is necessary that all cell $E_i \geq 5$; it turns out that if all E_i's are nearly equal, then the constraint $E_i \geq 5$ can be relaxed to $E_i \geq 3$. Since the SMD of the

GOF is only approximately χ^2 and approximation improving with increasing n, a more accurate *P-value* can be estimated by computing the *P-value* for the χ^2_{k-1} and immediately rejecting H_0 if $\hat{\alpha}_{k-1} < 0.05$. Next, the *P-value* should be computed using the χ^2_{k-1-m} distribution from $\hat{\alpha}_{k-1-m} = \Pr(\chi^2_{k-1-m} \geq \chi^2_0)$. If $\hat{\alpha}_{k-1-m}$ > 0.05, the null hypothesis of a good fit should immediately be accepted. If χ^2_0 lies in the indecision interval ($\chi^2_{0.05,k-1-m}$, $\chi^2_{0.05,k-1}$), the test of GOF should be declared inconclusive. Thus, for the Fish Example on pp. 221-223 of these notes, the actual *P*-value, denoted $\hat{\alpha}_a$, lies in the interval $0.44437 \leq \hat{\alpha}_a \leq 0.74983$, where $0.74983 = \Pr(\chi^2_5 \geq 2.6757)$. Further, the indecision interval for the Fish Example is given by $7.8147 \leq \chi^2_0 \leq 11.0705$.

<u>Example 47.</u> Outgoing lots of size N = 500 are inspected for number of defectives before shipment to customers. The results for a random sample of size n = 150 lots (each of size N = 500) are tabulated below, where the random variable X represents the number of defectives observed per lot. We wish to test if a Poisson distribution is a plausible model for the pmf (pr mass function) of X.

X (rv)	0	1	2	3	4	5	6	7
f_i	23	39	43	23	10	7	4	1

The above table shows that 23 of the 150 outgoing lots, each of size 500, to customers had no defectives, 39 had exactly 1 defective, 43 had exactly 2 defectives, etc, and only one lot had 7 defectives. The Poisson pmm (Pr mass model) is given by

$$P(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, 3, ...,$$

where the unknown parameter $\mu = E(X) = V(X)$. Therefore, a MLE of μ is given by the sample average number of defectives per lot given below. Clearly, $\bar{x}$ =

$$\frac{1}{150}\sum_{i=1}^{k=8} x_i f_i = \frac{1}{150}(39 + 86 + 69 + ... + 7) = \frac{300}{150} = 2.0 \text{ defectives per lot.}$$

Therefore, our null hypothesis is constrained to H_0 : $p(x;\mu) = \frac{2^x}{x!}e^{-2}$, x = 0, 1, 2, 3, ... The expected frequencies E_i, i = 0, 1, 2, 3, 4, 5, 6, 7 must be computed under H_0, i.e., $E_i = np_i$, where $p_i = \Pr(X = i) = \frac{2^i}{i!}e^{-2}$, i = 0, 1, 2, ..., 7. Under H_0, $E_0 = 150\times e^{-2} = 20.3$; $E_1 = 150\times 2e^{-2} = 40.6$, $E_2 = 150\times[\frac{2^2}{(2)!}e^{-2}] = 40.6$, $E_3 = 27.1$, and $E_4 = 13.5$. Thus far, the $\sum_{i=0}^{4} E_i = 142.1020$, which leaves 7.90 number of defectives for cells 5, 6, and 7. The only way that we can have all $E_i \geq 5$ is to combine the last 3 adjacent cells, as shown in the following table.

X	0	1	2	3	4	$x \geq 5$	Sums
f_i	23	39	43	23	10	12	150
E_i	20.30	40.60	40.60	27.07	13.53	7.90	150
$f_i - E_i$	2.70	– 1.60	2.40	– 4.07	– 3.53	4.10	0

The above table clearly shows that the $\sum_{i=0}^{5} E_i$ has been constrained by necessity to equal to $\sum_{i=0}^{5} f_i = 150 = n$. Hence, there are 2 constraints on the GOF statistic χ_0^2: (1) $\sum_{i=0}^{5}(f_i - E_i) \equiv 0$, and (2) The value of process mean has been constrained to $\mu = \bar{x} = 2$. Hence, the df of the GOF statistic

$$\chi_0^2 = \sum_{i=0}^{5}\frac{(f_i - E_i)^2}{E_i} = \sum_{i=0}^{5}\frac{(n_i - E_i)^2}{E_i} = 4.228$$

is equal to $\nu = 6 - 1 - m = 4$, where the one parameter μ is being replaced by its MLE. This yields the *P-value* =

Pr($\chi^2_4 \geq$ 4.228) = 0.37601, which far exceeds α = 0.05. Therefore, we cannot reject the null hypothesis of a Poisson fit to the data at levels of significance even as high as 0.37. Put differently, the Poisson pmf with μ = 2 does provide an acceptable fit to the pmf of the number of defectives per lot. The actual P-*value* lies in the interval 0.37601 $\leq \hat{\alpha}_a \leq$ 0.5171, where 0.5171 = Pr($\chi^2_5 \geq$ 4.228).

<u>Exercise 130.</u> (a) Work Exercise 15 on page 586 of Devore. ANS: $0.44424 \leq \hat{\alpha}_a \leq 0.65424$. (b) A computer generates the base-10 numbers 0, 1, 2, 3, ..., 9 completely at random (i.e., the discrete uniform distribution). If 1000 trials are made, how large should the $\sum_{i=0}^{9} f_i^2 = \sum_{i=0}^{9} n_i^2$ be so that the null hypothesis of randomness (i.e., equal Prs for all 10 cells) can be rejected at the 5% LOS?

<u>Exercise 131.</u> Work Exercises 16 and 18 on page 586 of Devore.

<u>Exercise 132.</u> Mendelian theory claims that 4 types of plants α, β, γ, and δ should occur in the ratio of 9:3:3:1. Does the following data support his theory? Write the null hypothesis and use your *P-value* to make a judgment about the GOF of the data to the p_i's theorized under H_0.

Plant Type	α	β	γ	δ	Total
f_i	120	48	36	13	217

Answer: χ^2_0 = 1.913; $\hat{\alpha}$ = *P-value* = Pr($\chi^2_3 \geq$ 1.913) = 0.5907.

It is paramount to become realistic and be concerned about the fact that in almost all real-life situations, the experimenter has no clue as to what type of underlying distribution function (except perhaps for discreteness or continuity) the collected data have originated from! There are three steps that the

experimenter must go through to come to some sort of decision regarding the underlying distribution for the collected data.

<u>Step 1.</u> **Compute the 1st four moments of the collected data, i.e., compute the values of $\bar{x}$, S, $\hat{\alpha}_3$, and $\hat{\alpha}_4$, where the sample skewness**

$$\hat{\alpha}_3 \cong \left[\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(x_i-\bar{x})^3\right]/s^3,$$

and the sample standardized fourth moment

$$\hat{\alpha}_4 = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(x_i-\bar{x})^4\right]/s^4,$$

and the corresponding kurtosis = $\hat{\alpha}_4 - \dfrac{3(n-1)^2}{(n-2)(n-3)} \cong \hat{\alpha}_4 - 3$ **(for n > 30).**

<u>Step 2.</u> **If there are no information about the values of μ and σ, then assume that the underlying distribution, which is being fitted to the data, has the approximate mean $\bar{x}$ and the approximate standard deviation $\hat{\sigma}$. This implies that we can always obtain a perfect fit for the 1st two moments of the data with that of the theoretical distribution being fitted to the data! This was the reason why we lost 2 df in the χ_0^2 test statistic in equation (84) for using the point estimates of μ and σ, because the true values of μ and σ were unknown.**

<u>Step 3.</u> **Compare the values of $\hat{\alpha}_3$ and $\hat{\alpha}_4$ of the data with α_3 and α_4, respectively, of the known statistical distributions, which are summarized in Table 31 below. Then, apply the GOF procedure to the distribution that is listed in Table 31, whose α_3 and α_4 are closest to those of the data. If there are 2 candidate statistical distributions, whose α_3 and α_4 are close to those of $\hat{\alpha}_3$**

and $\hat{\alpha}_4$, then more emphasis must be placed on the skewness $\hat{\alpha}_3$ than the kurtosis $\hat{\alpha}_4 - 3$. In Table 31, q = 1 – p, and also we have listed some information

Table 31 (Skewness and Kurtosis of Selected Statistical Distributions)

Discrete pmf's	α_3 (Skewness)	α_4 = Kurtosis +3
Binomial	$(q - p)/(npq)^{1/2}$	$[3pq\,(n - 2) + 1]/(npq)$
Geometric	$(1+q)\,/(q^{1/2})$	$(p^2 - 9p + 9)/q$
Poisson	$1/\mu^{0.50}$	$3 + (1/\mu)$
Continuous pdf's	α_3	α_4
Uniform	0	1.80
Triangular	$-\sqrt{0.32} \le \alpha_3 \le \sqrt{0.32}$	2.40
Normal (Gaussian)	0	3.00
Exponential	2	9
Gamma	$2\,/(n^{1/2})$	$3 + (6/n)$
Beta	$2(b - a)(a+b+1)^{1/2}/[(a+b+2)\times (ab)^{1/2}]$	

about the standard Beta distribution because of its applications to the fields of engineering and QC are widespread. Almost invariably, the pdf of a sample proportion (or FNC, $\hat{p}$) can be represented by the standard Beta distribution given by

$$f(\hat{p}) = \begin{cases} \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\hat{p}^{\,a-1}(1-\hat{p})^{b-1} & ,0 \le \hat{p} \le 1, \\ 0, \text{elsewhere} \end{cases}$$

where the rv $\hat{p}$ = a sample proportion, the parameters a , b > 0, E ($\hat{p}$) = a/(a+b) , and V ($\hat{p}$) = $ab/[(a+b)^2(a+b+1)]$.

Further, the Beta distribution has also widespread applications in the field of Bayesian Statistics and project management. In Table 31, the standardized 4th moment for the standard Beta pdf is given by

$$\alpha_4\,(\text{Beta}) = \frac{3(a+b+1)(a^2b+ab^2+2a^2+2b^2-2ab)}{ab(a+b+2)(a+b+3)}.$$

Further, when a = b, the skewness of the Beta distribution $\alpha_3 = 0$ and its kurtosis reduces to $\alpha_4 - 3 = 3(2a+1)/(2a+3) - 3 = -\frac{6}{2a+3} < 0$.

<u>Exercise 133.</u> Show that the skewness and kurtosis of the Binomial, Poisson, and the Gamma frequency functions approach those of the normal as the sample size n $\to\infty$. The kurtosis of the binomial is $\alpha_4 - 3 = \frac{1-6pq}{npq}$; the kurtosis of the Poisson distribution is $1/\mu$; the kurtosis of Gamma density is 6/n.

Finally, the ranges of α_3 and α_4 are, respectively, $-\infty < \alpha_3 < +\infty$, $1 < \alpha_4 < +\infty$, and my conjecture is that $\alpha_4 \geq 1 + \alpha_3^2$ for all statistical distributions! Further, the value of kurtosis (β_4) for all Triangular distributions in the universe is exactly equal to -0.6000, but the skewness value for all Triangular distributions lies within the interval $-\sqrt{0.32} \leq \alpha_3 \leq \sqrt{0.32}$.

Contingency Tables

A two-way contingency table consists of r rows and c columns, in which case it is called an r×c contingency table. Each unit in the sample is classified according to 2 categories described by row and column headings. As such, contingency tables have two major applications: (1) There are r distinct populations from which samples of sizes $n_1, n_2, \ldots, n_r$ are drawn and each unit is classified according to category 1, category 2, ..., category c. In this case, the null hypothesis is that the proportion of population i belonging to category j is homogeneous for all r populations, i.e., H_0: $p_{1j} = p_{2j} = \ldots = p_{rj} = p_{.j}$ for all j = 1, 2,

..., c versus the alternative that at least 2 of the r populations have different proportions in the j^{th} category of classification. (2) There is a single population from which N members are selected at random and each unit in the sample is classified according to both characteristics X and Y. In this case, the null hypothesis is that the X and Y classifications are independent.

1. Testing for Homogeneity of Proportions

<u>Example 48.</u> Random samples of sizes $n_1 = 80$, $n_2 = 60$, $n_3 = 70$, and $n_4 = 40$ are selected at random from a university's Freshman, Sophomore, Junior, and Senior classes, respectively. Note that 4 different frames were used to select the 4 samples at random from the r = 4 populations. The objective was to determine if the proportion of students belonging to the c = 3 categories of CGPA $2.0 \le X < 2.5$, $2.5 \le X < 3.4$, and the 3^{rd} category $3.4 \le X \le 4.0$ are the same for the four populations. The data are displayed in the Table 32. Table 32 clearly indicates

Table 32 (A contingency Table with fixed rows but random columns)

X / Populations	$2.0 \le X < 2.5$	$2.5 \le X < 3.4$	$3.4 \le X \le 4.0$	n_i
Freshmen	$n_{11} = 50$ (38.4)	18 (24.96)	12 (16.64)	80
Sophomores	$n_{21} = 30$ (28.8)	20 (18.72)	$n_{23} = 10$ (12.48)	60
Juniors	28 (33.6)	$n_{32} = 22$ (21.84)	20 (14.56)	70
Seniors	12 (19.2)	18 (12.48)	10 (8.32)	40
C_j	120	78	52	N = 250

that the row totals n_i, i = 1, 2, 3, 4 are fixed a priori, i.e., the experimenter has to decide what specific sample sizes are needed from each of the 4 populations so that 4 separate frames were used to draw the 4 random samples. The null hypothesis is $H_0 : p_{1j} = p_{2j} = p_{3j} = p_{4j} = p_{.j}$ for j = 1, 2, 3 vs the alternative $H_1 : p_{ij} \neq$

p_{kj} for some j and some pair i and k. We next compute the expected frequencies, E_{ij}, under H_0 in order to compare them against the observed frequencies $n_{11} = 50$, $n_{12} = 18$, ..., $n_{43} = 10$. Clearly, $\sum_{j=1}^{3} p_{ij} = 1$ for all i = 1, 2, 3, 4, yielding 4 constraints. Table 32 shows that under H_0, $\hat{p}_{.1} = 120/250 = 0.48$, $\hat{p}_{.2} = 78/250 = 0.312$, and $\hat{p}_{.3} = 1 - \hat{p}_{.1} - \hat{p}_{.2} = 0.208$. Hence, $E_{11} = n_1 \times \hat{p}_{.1} = 80 \times 0.48 = n_1 \times C_1/N = 38.4$, $E_{12} = n_1 \times \hat{p}_{.2} = 80 \times 0.312 = 80 \times 78/250 = 24.96$, and $E_{13} = 80 - E_{11} - E_{12} = 80 - 38.4 - 24.96 = 16.64$. As you have observed, it turns out that $E_{ij} = n_i \times C_j/N$. Similar computations, as done for the Freshmen population, leads to the expected frequencies for the other 3 populations, which are listed in parentheses in Table 32. Further, note that the sum of expectations of each row is constrained to equal to the corresponding n_i; this is why the last expectation in each column must be obtained by subtraction. Further, we also estimated two parameters, namely $p_{.1}$ and $p_{.2}$, and hence there are 12cells – 6 constraints = 6 df. Clearly if the null hypothesis is true, then we expect E_{ij}'s to be close to the corresponding

n_{ij}'s so that the statistic $n_{ij} - E_{ij}$ is a measure of the validity of H_0. The closer $(n_{ij} - E_{ij})$'s for all i & j are to zero, the stronger our belief will be in the validity of H_0. Hence, we may again use the GOF statistic

$$\chi_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left[\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \right] = \sum_{i=1}^{r} \sum_{j=1}^{c} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] \qquad (85)$$

in order to test H_0 and will reject H_0 if χ_0^2 is too large. Note that some authors use O_{ij} to denote the observed frequency of the (i, j) cell. How large should χ_0^2 be depends on the LOS α, which is generally taken to be 0.05. It can easily be argued that the df for the equation (85) is always equal to (r –1)(c – 1). For the example above, two parameters have to be estimated (namely $p_{.1}$ and $p_{.2}$) and we must require that $\sum_{j=1}^{3} E_{1j} = 80$, $\sum_{j=1}^{3} E_{2j} = 60$, $\sum_{j=1}^{3} E_{3j} = 70$, and $\sum_{j=1}^{3} E_{4j} = 40$, which

yield a total of 6 constraints. Hence, the df = 12 cells − 6 constraints = 12 − 6 = 3×2 = 6. For the Example 48, the statistic χ_0^2 has an approximate chi-square distribution with 3×2 = 6 df. You may easily verify that

$$\chi_0^2 = \frac{(50-38.4)^2}{38.4} + \frac{(18-24.96)^2}{24.96} + \ldots + \frac{(10-8.32)^2}{8.32} = 15.817,$$

which easily exceeds the 5 percentage point of chi-square with 6 df, $\chi^2_{0.05,6}$ = 12.592. Hence, we may reject H_0 at the LOS as small as *P-value* = $\hat{\alpha}$ = Pr($\chi^2_6 \geq$ 15.817) = 0.014771, and conclude that the proportions of students belonging to the 3 categories of college performance are not homogeneous for the 4 college classes. Put differently, college classification significantly impacts grades.

<u>Exercise 134.</u> Show that the chi-square statistic in (85) reduces to χ_0^2 = $\sum_{i=1}^{r}\sum_{j=1}^{c}[\frac{n_{ij}^2}{E_{ij}}]$ − N. Then use this last computational form to recompute the value of the test statistic for the Example 48. Verify the *P-value* = 0.014771. (b) Work Exercises 27, 29, and 30 on pp. 593-594 of Devore's 7th edition.

2. Testing for Independence in a Two-Way Classification

<u>Example 49.</u> A psychologist wished to determine if there were any relationships between a person's educational level, X, and the same persons adjustment to marriage, Y, i.e., he wished to test the null hypothesis that X and Y are independent. Accordingly, in a survey he selected N = 400 individuals at random (from a single frame) and measured the values of both random variables X and Y from each individual. The data are displayed in Table 33. The null hypothesis that X and Y classifications are independent can be formally written as $H_0 : p_{ij} = p_{i.}\times p_{.j}$ versus the alternative $H_1 : p_{ij} \neq p_{i.}\times p_{.j}$ for at least one pair (i, j). Without the assumption of independence, it follows that $p_{ij} = p_{i.}\times p_{j|i}$ for all i & j, where $p_{j|i}$ denotes the conditional Pr of j given i. By independence in a contingency table, we mean that the proportion out of each row total that belongs to the j^{th} column, $n_{ij}/n_{i.}$, is the same for all rows I = 1, 2, ..., r, and vice a

versa, i.e., $p_{j|i} = p_{.j}$. Therefore, under H_0 each cell expectation can be estimated as follows:

$$E_{ij} = N\times p_{ij} = N\times(p_{i.}\times p_{.j}) \cong N\times(\hat{p}_{i.}\times\hat{p}_{.j}) = N\times(\frac{n_{i.}}{N})\times(\frac{n_{.j}}{N}) = \frac{n_{i.}\times n_{.j}}{N}. \qquad (86)$$

Table 33. A Contingency Table with both Random Rows and Columns

X \ Y	very low	Low	High	Very High	$n_{i.}$
College educated	18	29	70	115	232
HS graduate	17	28	30	41	116
Grades	11	10	11	20	52
$n_{.j}$	46	67	111	176	N = 400

Applying equation (86) to the data of Table 33 yields $E_{11} = \frac{232\times46}{400} = 26.68$,

$E_{12} = \frac{232\times67}{400} = 38.86$, $E_{13} = \frac{232\times111}{400} = 64.38$, and $E_{14} = 232 - \sum_{j=1}^{3} E_{1j} = 102.08$.

The remaining E_{ij}'s are computed similarly, and their values are $E_{21} = 13.34$, $E_{22} = 19.43$, $E_{23} = 32.19$, $E_{24} = 116 - 13.34 - 19.43 - 32.19 = 51.04$, $E_{31} = 46 - 26.68 - 13.34 = 5.98$, $E_{32} = 67 - 38.86 - 19.43 = 8.71$, $E_{33} = 111 - 64.38 - 32.19 = 14.43$, $E_{34} = 176 - 102.08 - 51.04 = 22.88$. Therefore, the chi-square statistic is

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}(\frac{n_{ij}^2}{E_{ij}}) - N = \sum_{i=1}^{r}\sum_{j=1}^{c}(\frac{o_{ij}^2}{E_{ij}}) - N = \sum_{i=1}^{r}\sum_{j=1}^{c}(\frac{f_{ij}^2}{E_{ij}}) - N$$

$$= \frac{18^2}{26.68} + \frac{29^2}{38.86} + \ldots + \frac{20^2}{22.88} - 400 = 19.94265$$

The *P-value* for the above statistic is computed from $\hat{\alpha} = \Pr(\chi_6^2 \geq 19.94265) = 0.00284$, which is much less than 0.05. Hence, we may reject H_0 at the LOS as small as $\hat{\alpha} = 0.00284$ and conclude that X and Y are not independent. This implies the data indicates that adjustment to marriage is somehow related to

educational level of individuals. The GOF statistic, χ_0^2, has 6 df in this example because 5 parameters $p_{1.}$, $p_{2.}$, $p_{.1}$, $p_{.2}$, $p_{.3}$ have to be estimated form the data and we must force $\sum_{i=1}^{3}\sum_{j=1}^{4}\hat{p}_{ij} = \sum_{i=1}^{3}\sum_{j=1}^{4}\hat{p}_{i.}\times\hat{p}_{.j} = 1$. Put differently, the 6 constraints are $\sum_{j=1}^{4}E_{1j} = 232$, $\sum_{j=1}^{4}E_{2j} = 116$, $\sum_{i=1}^{3}E_{i1} = 46$, $\sum_{i=1}^{3}E_{i2} = 67$, $\sum_{i=1}^{3}E_{i3} = 111$ and $\sum_{i=1}^{3}\sum_{j=1}^{4}E_{ij} = 400 \rightarrow$ df = 12Cells − 6Constraints = 6.

<u>Exercise 135.</u> A psychologist obtained the following data on human eye and hair color in order to ascertain if eye (X) and hair (Y) colors are independent?

X \ Y	Light hair	Dark hair	Red hair	$n_{i.}$
Blue eyes	35	15	10	60
Brown eyes	20	30	10	60
Green eyes	10	10	20	40
$n_{.j}$	65	55	40	N = 160

Test the null hypothesis that eye color (X) and hair color (Y) are independent at the LOS α = 0.01. Answer: $\chi_0^2 \cong 27.9720$, $\hat{\alpha}$ = *P-value* = 0.000012637.

<u>Exercise 136.</u> Work exercises 32, p. 594, and 42 on page 596 of Devore.

We have now come to the end of the course STAT 3610, but I need to emphasize what you should do in order to prepare well for the Final Exam in the STAT 3610. The STAT 3610 Final will have a small closed-book and closed-notes (True or False) section followed by only open-downloaded-notes section. You may not bring solutions to homework problems to the final exam with you, but you do need the book for tables. Therefore, you need to review the notes on SLREG&CORR, MLREG, Chi-square GOF test, and contingency tables carefully!