Example 13.7.3:

We are given the following table of data that gives the number of astronomical observations whiich fall into a given velocity interval.

Observed and expected frequencies for chi-square goodness-of-fit test with estimates $\hat{\mu} = -20.3$ and $\hat{\sigma} = 12.7$

Intervals of Velocities	o_j		$\hat{p}_{jo}$		$\hat{e}_j$
(−80, −70)	1	15	.000	.224	17.92
(−70, −60)	2		.001		
(−60, −50)	2		.009		
(−50, −40)	2		.051		
(−40, −30)	8		.163		
(−30, −20)	24		.284		22.72
(−20, −10)	26		.283		22.64
(−10, 0)	11	15	.154	.209	16.72
(0, 10)	2		.046		
(10, 20)	1		.008		
(20, 30)	1		.001		
	80		1.000		80.00

Where o_j represents the number of observed occurrences in the interval $\{a_j, a_{j+1}\}^1$. A test for normality proceeds as follows; assume $X_i \sim N[\mu, \sigma]$. Let x_{ij} denote the i^{th} observation in cell (velocity interval) j. We have $\sum_i x_{i,j} = o_j$. Being as we are testing for a normal distribution the central limit theorem is not strictly necessary, but for reasons which will soon become apparent it is required for the Chi-squared test. So we set $z = \frac{x-\mu}{\sigma}$ which gives the probability density function (pdf) of the j^{th} cell,

$$p_j = \left[\frac{1}{\sqrt{2\pi\sigma}} \int_{z_j}^{z_{j+1}} e^{-\frac{1}{2}z^2} dz \right]^{o_j} = [\Phi(z_{j+1}) - \Phi(z_j)]^{o_j}$$

Where Φ denotes the integration of the standard normal. We've integrated over the cell $\{a_j, a_{j+1}\}$ because we are only concerned with whether a given observation occurred *anywhere* within this interval. By taking this integral to the power o_j we've assumed observations within each cell are independent of one another, which is to say the fact one observation was made in this cell had no relation to the fact that some other observation was also made in it.

In order to perorm a Chi-squared goodness of fit test it is of use to relate this pdf to the multinomial distribution. In fact we've already done the bulk of the work required to do this by integrating over the cells in which observations are contained. This has the effect of reducing the expected values of the cells to np_j where $n = \sum_j o_j$.

We now make use of the multinational expansion on the sum of probabilities:

$$(p_0 + p_1 + ... + p_c) = \frac{n!}{o_1! o_2! .... o_c!} pq * p2 * ... * pc \tag{1}$$

[1] The table suggests that $\{a_j, a_{j+1}\}$ is a velocity interval, but considering the math to follow, unit analysis requires that a_j must be taken to represent a boundary of o – the continuous equivalent of o_j, i.e. a_j is a unitless number. Considering the variable of interest is actually $\frac{o_1 - np}{\sqrt{npq}}$ (c.f. eq 2) there is not *neccessarily* a contradiction. Dividing numerator and denominator by t we infer the central limit theorem is being applied to the number of observations per unit time (hence the title of the graph – freqeuncies for...).

Where c represents the number of cells ($c = 10$ in our case). The multinomial distribution has expected values $E[o_j] = \mu_j = np_j$ and variance values $\sigma_j = np_j(1-p_j) = np_jq_j$.

Chi-squared test

Now consider for just a moment the case of $c = 1$ which is to say there is only one cell or velocity interval in which observations are made. We perform n trials in which we attempt to observe an asteroid moving with the range of velocities that define this cell. There are only two outcomes; an asteroid was observed to moving at some velocity within the given velocity range of our cell on trial i (successful trial), or an asteroid was not observed to move within the given velocity range on trial i (failed trial), i.e. these are Bernoulli trials. If $x_i \in 0, 1$ is our Bernoulli variable, we have $\sum_i x_i = o_1$ as the total number of successful trials, and $o_2 = n - 0_1$ unsuccessful trials. The probability of success is $p_1 = p$.

Making the central limit approximation on a sum of n Bernoulli trials gives for the integration variable,

$$z = \frac{\sum_i x_i - n\mu}{\sqrt{n}\sigma} = \frac{o_1 - np}{\sqrt{npq}} \tag{2}$$

Now it is well known that A) for sufficiently large n a variable of the form of equation 2 is distributed as $N[0,1]$, and B) for $Z \sim N[0,1]$ we have $z^2 \sim \mathcal{X}^2(1)$ where $\mathcal{X}^2(v)$ denotes the Chi-squared distribution with v degrees of freedom. We can decompose z^2 as follows,

$$\begin{aligned} z^2 = \frac{(o_1 - np)^2}{np(1-p)} &= \frac{(o_1 - np)^2}{np(1-p)} + \frac{(o_1 - np)^2}{n(1-p)} \\ &= \frac{(o_1 - np)^2}{np} + \frac{((n - o_1) - n(1-p))^2}{n(1-p)} \\ &= \sum_{i=1}^{2} \frac{(o_i - np)^2}{np} \end{aligned} \tag{3}$$

where $p_1 = p$ is again the probability of a successful trial and $p_2 = 1 - p$ is the probability of a failed trial. In terms of expected values e_1, e_2 for o_1, o_2,

$$z^2 = \sum_{j=1}^{2} \frac{(o_i - e_j)^2}{e_j} \sim \mathcal{X}^2(1) \tag{4}$$

So, for two distinct outcomes we got only *one* degree of freedom. In general, for k different outcomes we have that,

$$\sum_{j=1}^{k} \frac{(o_i - np_j)^2}{np_j} \sim \mathcal{X}^2(k-1) \tag{5}$$

This is known as the *Pearson Chi-squared statistic*

Question; what would the corresponding normal variable(s) look like for the case of $k > 2$? Attempting the above decomposition for even the relatively simple case of $k = 3$ is difficult. There are however, a number of advanced theorems which show that equation 2 holds. Reference [1] provides no less than seven different proofs.

Likewise, there are subtleties in calculating the degrees of freedom. Noting that a sum of Chi-squared variables is distributed as a Chi-squared statistic with degrees of freedom equal to the sum of the degrees of freedom of the summed variables ($\sum_i^c \mathcal{X}^2(v_i) \sim \mathcal{X}^2(\sum_i^c v_i)$), from the above analysis it is tempting to say that c cells with k possible outcomes would produce $c(k-1)$ degrees of freedom. In fact we'll have for our total degrees of freedom $c-1$. Why? A detailed proof can be found in [1][2] We will settle for noting that if an asteroid was not observed in the velocity interval of interest, then implicitly we've assumed there was an asteroid, it just was moving at some other velocity, ergo, it must exist in another cell! So the number of different outcomes k is not in all cases unrelated to the number of cells c.

If there are any parameters which need to be estimated then we'll have to subtract a degree of freedom for each variable estimated. Being as we've not been given values for the parameters μ, σ we'll need to estimate these parameters and so we will subtract 2 giving $c - 1 - 2 = c - 3$ degrees of freedom.

Lastly, note that it sometimes occurs that cells need to be pooled together because a criteria of the Pearson Chi-squared test is that the expected values are $e_j > 5$. In such cases we of course count the number of *grouped* cells when calculating the degrees of freedom. Being as $e_j = np_j$ the criteria that $e_j > 5$ amounts to $p_j > 5/n$.

1 Estimating μ, σ

We need to estimate the parameters μ, σ by finding the maximum likelihood estimates (MLE's). Taking the natural log of equation 1 and differentiating with respect to the variables μ and σ then setting the two resulting equations to zero gives,

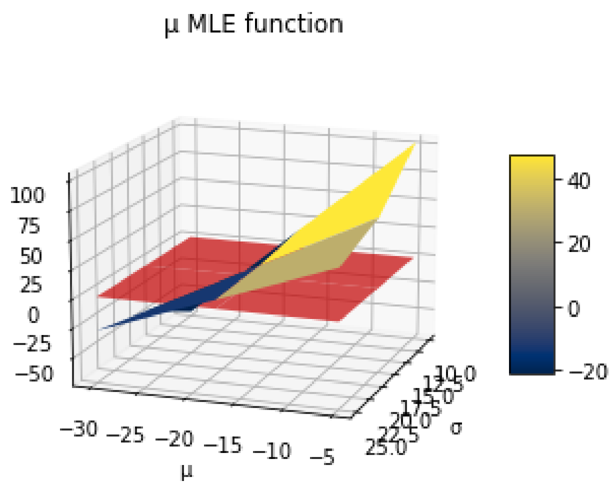
$$\sum_j o_j \frac{\Phi'(z_{j+1}) - \Phi'(z_j)}{\Phi(z_{j+1}) - \Phi(z_j)} = 0 \tag{6}$$

$$\sum_j o_j \frac{z_{j+1}\Phi'(z_{j+1}) - z_j\Phi'(z_j)}{\Phi(z_{j+1}) - \Phi(z_j)} = 0 \tag{7}$$

where Φ' denotes the derivative of an integral, so it is simply the standard normal distribution evaluated at its given argument. Equations 6 and 7 are solved via the two dimensional Newton-Raphson algorithm contained in *raphson.py*. The Newton-Raphson method is among the more primitive root-finding algorithms, yet with a little trial and error so as to [graphically or otherwise] guess the vicinity of a root, it serves its purpose in most cases.

First we take a guess. As described in the code contained in $Ch\_test\_for\_normality\_interval\_data.py$ we first run the program with $PLOT = True$ and $RAPHSON = False$. The graph of equation 6 gives a rough idea of what the values of μ, σ are. We ascertain this estimate by looking for where the three dimensional plane is zero. After some trial and error with what range of values over which we ought to plot μ, σ (and also different viewing angles) we get the following plot,

[2] see *Sixth Proof: Generic induction with De Moivre-Laplace theorem.*



So we estimate $\mu, \sigma = -20, 5$. Now we turn $PLOT = False, RAPHSON = True$ and run the code again with this estimate to see if we can get an even better estimate from the Newton-raphson algorithm contained in $raphson.py$.

After some trial and error we find that [integral] equations 6 and 7 are very touchy, so we reluctantly increase the integration steps to $1/10,000$ and we set the step size for calculating derivatives in the Newton-Raphson algorithm to $1/10,000$ as well. Noting that equations 6 and 7 are not just integrated equations, but are a *sum* of integrated equations, and furthermore that we ought to expect the Newton-Raphson algorithm to run $\tilde{5}$-20 times before finding a root, we let the program run while we take a coffee break. We come back the program has estimated the values of μ, σ as,

$$\mu, \sigma \approx -19.867115036443728, 11.2626087946973$$

Which are not *too* far off from the reported values in [2][3] which gives $\mu, \sigma \approx -21.3, 12.7$.

These estimates enter into the Chi test via the calculation of the probabilities. We get for our Chi-squared statistic,

$$\sum_{j=1}^{k} \frac{(o_i - np_j)^2}{np_j} = 1.302 \tag{8}$$

The calculated value in [2] is 1.22.

Finally, we need a critical value to which we can compare our result to. It turns out we need to pool a number of cells in this example and we are left with $c = 4$ cells. With degrees of freedom $v = c - 1 - 2 = 1$ we have for a critical value $\mathcal{X}^2(1) \approx 3.927$. As noted in the code, the Chi-squared distribution explodes at the origin when $v = 1$ so this is a very touchy integration in which we again need to make very small and time consuming integration steps. The result given in [2] are $\mathcal{X}^2(1) = 3.84$.

Recalling that we originally set out to test $Ho : X_i \sim N[\mu, \sigma^2]$. We reject Ho If our test statistic is greater than the critical value. In our case $1.302 < 3.927$ so we do not reject this hypothesis.

[3]See top of page 457

2 Alternative approach to estimating μ, σ: Regression analysis

As a much more expedient alternative, we can estimate the parameters μ, σ with a least squared regression. To this end we further manipulate our statistics into the form of a uniform distribution by considering the variables,

$$
\begin{aligned}
F_0 &= \frac{0}{n} \\
F_1 &= \frac{o_1}{n} \\
F_2 &= \frac{o_1}{n} + \frac{o_2}{n} \\
F_3 &= \frac{o_1}{n} + \frac{o_2}{n} + \frac{o_3}{n} \\
&\vdots \\
F_n &= 1
\end{aligned}
$$

Making the crude approximation that $\Phi(z_j) \approx F_j$ we have $\Phi^{-1}(F_j) \approx z_j = -\frac{\mu}{\sigma} + \frac{a_j}{\sigma} = \beta_0 + \beta_1 a_j$. Now let $x_j = a_j, y_j = \Phi^{-1}(F_j)$. This is now in the form to perform a linear regression.

But first, how exactly do we go about calculating $\Phi^{-1}(F_j)$? Mathematically, this is the inverse error function. Computationally it is nothing fancy; we do what we often do whenever we want to find a critical values $z_{1-\alpha}$ given a percentile α; we integrate the standard normal until the cumulative total reaches α and return whatever value z was at when this sum first reached α, ergo the function which performs this trick is called $Percentile$.

Omitting the last variable $F_n = 1$, a linear least squares regression yields the coefficients $B0 = -1.87000, B1 = -0.08162$ which in terms of μ, σ give $\mu = -22.9106, \sigma = -12.2516$.

References

[1] Eric Benhamou, Valentin Melot *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation.* AMS 1991 subject classification: 62E10, 62E15

[2] Bain, Engelhardt *Introduction to Probability and Mathematical Statistics [second ed.]*. Brooks/ Cole, Belmont, CA