STA 114: Statistics

Notes 20. Chi-square tests for Independence, Goodness-of-fit

Two-way tables: category counts based on two attributes

A two-way contingency table is a special kind of categorical data where n units are split into categories determined by two attributes. In the following table, 231 humans are categorized according to their hair color and eye color.

		Eye color				
		Blue	Green	Brown	Black	Total
Hair color	Blonde	20	15	18	14	67
	Red	11	4	24	2	41
	Brown	9	11	36	18	74
	Black	8	17	20	4	49
	Total	48	47	98	38	231

The table contains counts data X on 16 categories and so we can model $X \sim \mathsf{Multinomial}(n, p)$ with p in the 16-dimensional simplex.

However, as shown in the table, it is more convenient and informative to present the counts as a two-dimensional array, with the labels of one attribute along each dimension. Therefore, we would interpret, with R = number of rows and C = number of columns of the table,

$$X = (X_{11}, \cdots, X_{R1}, X_{12}, \cdots, X_{R2}, \cdots, X_{1C}, \cdots, X_{RC})$$
$$p = (p_{11}, \cdots, p_{R1}, p_{12}, \cdots, p_{R2}, \cdots, p_{1C}, \cdots, p_{RC}),$$

i.e., x_{ij} is the count in the (i, j)-th cell of the table and p_{ij} is the probability that any unit would land up in that cell. Observed data would be denoted in the same manner: $x = (x_{11}, \cdots, x_{R1}, x_{12}, \cdots, x_{R2}, \cdots, x_{1C}, \cdots, x_{RC})$.

To facilitate discussion, introduce the following margin counts

$$\begin{aligned} X_{i\cdot} &= X_{i1} + X_{i2} + \cdots + X_{iC}, \quad i = 1, 2, \cdots, R \\ X_{\cdot j} &= X_{1j} + X_{2j} + \cdots + X_{Rj}, \quad j = 1, 2, \cdots, C \end{aligned}$$

and use similar definitions on p to get $p_{i\cdot}$'s, $p_{\cdot j}$'s and on x to get $x_{i\cdot\cdot}$'s and $x_{\cdot j}$'s. For the hair-eye color table, $x_{1\cdot} = 67$, $x_{\cdot 3} = 98$ and so on.

The row counts $X^R = (X_{1\cdot}, \cdots, X_{R\cdot})$ are simply the count data when the units are only split according to the row attribute (hair color). Similarly, the columns counts $X^C = (X_{\cdot 1}, \cdots, X_{\cdot C})$ give the one-way categorization according to the column-attribute (eye color). Under the model $X \sim \mathsf{Multinomial}(n, p)$, we must have $X^R \sim \mathsf{Multinomial}(n, p^R)$ and $X^C \sim \mathsf{Multinomial}(n, p^C)$. And if we observe data $X = x$ then the observation for X^R is simply x^R and that for X^C is x^C.

Test of attribute independence

Two way tables are widely used to test for independence between the two attributes. For our hair-eye color example, we could test whether what eye color we have is independent of our hair color. Under independence of the two attributes we can write:

$$\begin{aligned} p_{ij} &= P(\text{row attribute label} = i, \text{column attribute label} = j) \\ &= P(\text{row attribute label} = i) \times P(\text{column attribute label} = j) \\ &= p_{i\cdot} p_{\cdot j}. \end{aligned}$$

Therefore, if we know $p^R = (p_{1\cdot}, \cdots, p_{R\cdot})$ and $p^C = (p_{\cdot 1}, \cdots, p_{\cdot C})$, then under the assumption of independence we can construct all RC elements of$p$ by taking products $p_{ij} = p_{i\cdot} p_{\cdot j}$. We'll use $p = \text{out}(p^R, p^C)$ to denote such an "outer product" construction of $p \in \Delta_{RC}$ from a $p^R \in \Delta_R$ and a $p^C \in \Delta_C$. So, we can formalize the hypothesis of independence as

$$H_0 : p \in \Delta_{RC}^{\text{out}}, \quad H_1 : p \notin \Delta_{RC}^{\text{out}}$$

where $\Delta_{RC}^{\text{out}} = \{p = \text{out}(p^R, p^C) : p^R \in \Delta_R,\ p^C \in \Delta_C\}$.

Pearson's test of independence

From our discussion in the last handout, we can test the above H_0 as follows. We first construct

$$Q(x) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(x_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

where $\hat{e}_{ij} = n\hat{p}_{ij}^{\text{out}}$ is the expected cell counts under H_0, with $\hat{p}^{\text{out}}$ denoting the restricted MLE of p under H_0. To find $\hat{p}^{\text{out}}$ notice that for a $p = \text{out}(p^R, p^C)$, the likelihood function simplifies to

$$L_x(p) = \text{const} \times \left[\prod_{i=1}^{R} p_{i\cdot}^{x_{i\cdot}} \right] \times \left[\prod_{j=1}^{C} p_{\cdot j}^{x_{\cdot j}} \right]$$

and hence it is maximized at $\hat{p}_{i\cdot} = x_{i\cdot}/n$, $i = 1, 2, \cdots, R$ and $\hat{p}_{\cdot j} = x_{\cdot j}/n$, $j = 1, 2, \cdots, C$. That is, $\hat{p}_{ij}^{\text{out}} = (x_{i\cdot} x_{\cdot j})/n^2$ and consequently, $\hat{e}_{ij} = (x_{i\cdot} x_{\cdot j})/n$.

With $Q(x)$ constructed, Pearson's chi-square size-α test would reject H_0 if $Q(x) > F_{k-1-r}^{-1}(1-\alpha)$ where $k = RC$ is the total number of categories, and r is the number of free parameters in Δ_{RC}^{out}. Note that elements of this set are made of a $p^R \in \Delta_R$ with $R-1$ free elements, plus with a $p^C \in \Delta_C$ with another $C-1$ free elements. So the number of free elements in Δ_{RC}^{out} is $r = R-1+C-1$. Combining, we get $k-1-r = RC-R-C+1 = (R-1)(C-1)$.

So in summary, we'd perform a size-α Pearson's chi-square test of independence by rejecting H_0 if $Q(x) > F_{(R-1)(C-1)}^{-1}(1-\alpha)$. The p-value based on these tests is given by $1 - F_{(R-1)(C-1)}(Q(x))$.

Example (Hair-eye color)**.** To perform the test on the hair/eye color data, we'd insert in each cell on the table the expected count $x_{i\cdot} x_{\cdot j}/n$ (shown in parentheses below).

		Eye color				
		Blue	Green	Brown	Black	Total
Hair color	Blonde	20 (13.9)	15 (13.6)	18 (28.4)	14 (11.0)	67
	Red	11(8.5)	4 (8.3)	24 (17.4)	2 (6.7)	41
	Brown	9 (15.4)	11 (15.1)	36 (31.4)	18 (12.2)	74
	Black	8 (10.2)	17 (10.0)	20 (20.8)	4 (8.1)	49
	Total	48	47	98	38	231

and find $Q(x) = 30.9$. So the p-value is $1 - F_{(4-1)(4-1)}(30.9) = 1 - F_9(30.9) \approx 0$.

Testing goodness-of-fit

So far in our course we have always accepted a given statistical model to describe our data. Can we ascertain whether a model is adequate for the data we observe? Surprisingly, Pearson's chi-square test provides quite a compelling way of doing it.

Let's start with a simpler problem. Suppose we model $Z = (Z_1, \cdots, Z_n)$ as $Z_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0, 1)$. Once we observe data $z = (z_1, \cdots, z_n)$ can we ascertain if they really came from a $\mathsf{Normal}(0, 1)$ distribution. We have done this in our labs many times. We simply make a histogram plot of our data and then overlay with the pdf of the $\mathsf{Normal}(0, 1)$. If the histogram and the pdf look similar, we're reasonably happy that the $\mathsf{Normal}(0, 1)$ model fits the observations.

In matching the histogram with the theoretical pdf, we essentially did the following. We split the range of our observations into bins, then counted number of observations in each bin and checked whether these counts are in agreement with the expected counts in the bins if they really came from the $\mathsf{Normal}(0, 1)$ distribution. That is, we first turn our original data into a count data (with multinomial pmf) and then checked whether they agree with the bin probabilities under the $\mathsf{Normal}(0, 1)$ pdf.

Formalization through bin count and Pearson's test

We can do this testing formally as follows. Fix k bins $(a_0 = -\infty, a_1]$, $(a_1, a_2]$, $\cdots$, $(a_{k-2}, a_{k-1}]$, $(a_{k-1}, a_k = \infty)$ covering the whole of $(-\infty, \infty)$, and define X_1 as the number of Z_i's in the first bin, X_2 as the number of Z_i's in the second bin and so on. Together we get k bin counts, and $X = (X_1, \cdots, X_n)$ must have a $\mathsf{Multinomial}(n, p)$ where p_l denotes the probability that Z_1 (or any other Z_i) would be in the l-th bin $(a_{l-1}, a_l]$.

If indeed $Z_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0, 1)$ then we should have $p = p_0$ where $p_{0l} = \Phi(a_l) - \Phi(a_{l-1})$ for $l = 1, \cdots, k$. Here $\Phi(z)$ denotes the $\mathsf{Normal}(0, 1)$ cdf. So we can rephrase our question as testing $H_0 : p = p_0$ against $H_1 : p \neq p_0$. If we reject H_0, then that would mean that we have evidence against Z_i's being IID $\mathsf{Normal}(0, 1)$.

Once the problem is set up this way, testing H_0 is straightforward by using Pearson's chi-square test for a point null. That is, we calculate $Q(x) = \sum_{l=1}^{k} (x_l - e_l)^2/e_l$ where $e_l = np_{0l} = n(\Phi(a_l) - \Phi(a_{l-1})$ and reject H_0 at level α if $Q(x) > F_{k-1}^{-1}(1 - \alpha)$, or report the p-value $1 - F_{k-1}(Q(x))$.

The only non-trivial thing here is how we choose the bins (how many and where). More bins would mean a finer comparison between the histogram shape and the $\mathsf{Normal}(0, 1)$ pdf, and hence better power. But by choosing too many bins, we may bring down the count in

each to very small numbers, making the test unreliable. It is usually recommended that each bin should have expected count (under the null) of at least 5. So we can take k to be the integer just larger than or equal to $n/5$.

We saw in the lab that the power of a Pearson's chi-square test against a point null is larger when the point null is $p_0 = (1/k, \cdots, 1/k)$. So we can choose a_l's to make the bins receive equal probabilities under the $\mathsf{Normal}(0,1)$ pdf. This is achieved by setting $a_1 = \Phi^{-1}(1/k)$, $a_2 = \Phi^{-1}(2/k)$, $\cdots$, $a_{k-1} = \Phi^{-1}((k-1)/k)$. With this choice each $e_l = n/k$ so $Q(x) = \sum_{l=1}^{k}(x_l - n/k)^2/(n/k) = \frac{k}{n}\sum_{l=1}^{k}(x_l - n/k)^2$.

Similarly, if we wanted to test if $Z_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ for some fixed μ and σ^2 then we would do exactly as above, but choose the bins according to the inverse of the $\mathsf{Normal}(\mu, \sigma^2)$ cdf, i.e., $a_1 = \mu + \sigma\Phi^{-1}(1/k)$, $a_2 = \mu + \sigma\Phi^{-1}(2/k)$, $\cdots$, $a_{k-1} = \mu + \sigma\Phi^{-1}((k-1)/k)$. Once we get the bin counts $x_1, \cdots, x_k$ we have a similar construction of the test statistic $Q(x) = \frac{k}{n}\sum_{l=1}^{k}(x_l - n/k)^2$. As before, $Q(X)$ is approximately $\chi^2(k-1)$ under the null. Therefore a size α test rejects H_0 if $Q(x) > F_{k-1}^{-1}(1-\alpha)$ and the p-value based on such tests is $1 - F_{k-1}(Q(x))$.

Testing goodness-of-fit of the normal model

However, our real interest is in checking if $Z_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ for some μ and σ^2 that are not necessarily known to us. To tackle this we can follow the same recipe above but with μ and σ estimated by $\hat{\mu} = \bar{z}$ and $\hat{\sigma} = s_z$. That is, we use bins $a_1 = \bar{z} + s_z\Phi^{-1}(1/k)$, $a_2 = \bar{z} + s_z\Phi^{-1}(2/k)$, etc to get our bin counts $x_1, \cdots, x_k$. We again construct $Q(x) = \frac{k}{n}\sum_{l=1}^{k}(x_l - n/k)^2$.

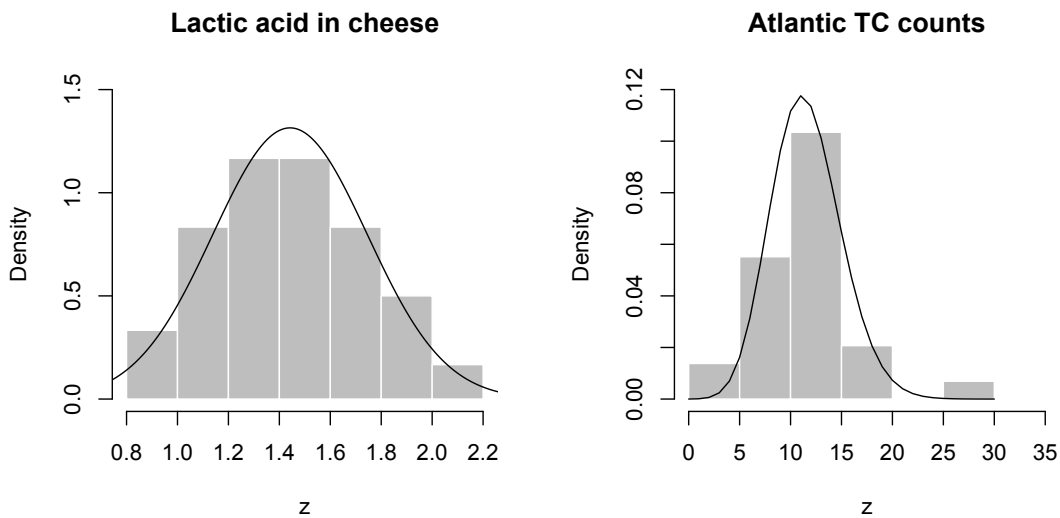
What should be the distribution of this $Q(X)$ under the null? The number of free parameters under the null is $r = 2$ because we had both μ and σ^2 unknown. Should we have $Q(X) \sim \chi^2(k-1-2) = \chi^2(k-3)$? This approximation, however, does not work so well. But it turns out that under the null $P(Q(X) > c$ lies between $1 - F_{k-3}(c)$ and $1 - F_{k-1}(c)$. Consequently the p-value based on "reject H_0 if $Q(x) > c$" type of tests lies between $1 - F_{k-3}(Q(x))$ and $1 - F_{k-1}(Q(x))$. So we have a range of p-values.

Example (Lactic acid concentration)**.** 30 samples from a cheese slab are measured to have the following lactic acid concentrations:

```
0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58,
1.68, 1.90, 1.06, 1.30, 1.52, 1.74, 1.16, 1.49, 1.63, 1.99,
1.15, 1.33, 1.44, 2.01, 1.31, 1.46, 1.72, 1.25, 1.08, 1.25
```

We want to test whether these observations are IID draws from a $\mathsf{Normal}(\mu, \sigma^2)$ distibution, with no assumption made on μ and σ^2. The figure below (left) shows the data histogram overlaid with the $\mathsf{Normal}(\hat{\mu}, \hat{\sigma}^2)$ pdf where $\hat{\mu} = \bar{z} = 1.44$ and $\hat{\sigma} = s_z = 0.30$. There seems a fairly good agreement between the two.

Because we have $n = 30$, we can use $k = n/5 = 6$ bins of the form $(a_{l-1}, a_l]$ with $a_l = \bar{z} + s_z\Phi^{-1}(l/k)$. The bins are found to be $(-\infty, 1.15]$, (1.15, 1.31], (1.31, 1.44], (1.44, 1.57], (1.57, 1.74], $(1.74, \infty)$ with bin counts $5, 8, 2, 5, 4, 6$. We calculate $Q(x) = 4$. Therefore the p-value ranges between $1 - F_3(4) = 0.26$ and $1 - F_5(4) = 0.55$. So we would fail to accept the null hypotheses at nominal levels of 5%, 10% or 1%.



Testing goodness-of-fit of a general parametric model

The same concepts apply for a general model $Z_i \overset{\text{IID}}{\sim} g(z_i||\theta)$, $\theta \in \Theta$ for data $Z = (Z_1, \cdots, Z_n)$. We first find an estimate $\hat{\theta}(z)$ of θ, usually the MLE, and then form $k \geq n/5$ bins $(a_{l-1}, a_l]$, $l = 1, \cdots, k$ with $a_l = G^{-1}(l/k|\theta = \hat{\theta}(x))$, where $G(z_i|\theta)$ denotes the CDF of $g(z_i|\theta)$. We get bin counts $x_1, \cdots, x_k$ and constrict $Q(x) = \frac{k}{n}\sum_{l=1}^{k}(x_l - n/k)^2$ and find a p-value range between $1 - F_{k-1-r}(Q(x))$ and $1 - F_{k-1}(Q(x))$ where r is the dimension of θ (i.e., number of free parameters we needed to estimate).

Example (Annual TC counts in the north Atlantic)**.** The annual tropical cyclone counts in the north Atlantic between 1980 and 2009 are as below (30 years).

```
11 11  5  4 12 11  6  7 12 11
14  8  6  8  7 19 13  7 14 12
14 15 12 16 14 27 10 14 16 11
```

Are these data from a $\mathsf{Poisson}(\mu)$ distribution for some $\mu > 0$? To test this, we first estimate μ by $\hat{\mu}_{\text{MLE}}(z) = 11.57$. The figure above (right) shows the data histogram overlaid with the $\mathsf{Poisson}(11.57)$ pmf. A total of $n/5 = 6$ bins are found by using the inverse CDF of this Poisson pmf. These equal [0, 8], (8, 10], (10. 11], (11, 13], (13, 15], $(15, \infty)$ with bin counts 9 1 5 5 6 4. For these counts, we get $Q(x) = 6.8$. Consequently the p-value ranges between $1 - F_4(6.8) = 0.147$ to $1 - F_5(6.8) = 0.236$. So again there is no strong evidence against the null hypothesis.