Math 6070-1, Spring 2014
Partial Solutions to Assignment #5

1. ...

2. *A random sample of 100 people from a certain population resulted in the following:*

Age Group	No. of samples
0–16	27
17-24	26
25–34	15
35–49	22
50–100	10

Perform a χ^2 test in order to see if the data is distributed uniformly across the mentioned age groups.

Solution. Add an extra column, using χ^2-testing notation, and the fact that the total sample size is 100:

Group	Observed	Expected
1	27	20
2	26	20
3	15	20
4	22	20
5	10	20

Then,

$$\chi^2 = \sum_i \frac{(\mathrm{Obs}_i - \mathrm{Exp}_i)^2}{\mathrm{Exp}_i} = \frac{49}{27} + \frac{36}{26} + \frac{25}{15} + \frac{4}{22} + \frac{100}{10} \approx 15.68.$$

With four degrees of freedom, this test statistic yields a p-value < 0.005. In other words, we declare the distribution as non uniform at say 95% confidence level.

3. *Consider two finite populations: One has respective proportions $\theta_1, \ldots, \theta_m$ for its individuals of types $1, \ldots, m$. The other has respective proportions $p_1, \ldots, p_m$ for its individuals of type $1, \ldots, m$. Let $\boldsymbol{\theta} := (\theta_1\,, \ldots, \theta_m)'$ and $\boldsymbol{p} := (p_1\,, \ldots, p_m)'$ be the respective probability vectors. We assume that $\boldsymbol{p}$ and $\boldsymbol{\theta}$ are unknown.*

   Independent samples are taken from each population [independently from one another]. Let the sample sizes be n_1 and n_2 respectively, and denote by $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{p}}$ the sample-proportion vectors of types.

   (a) *Prove that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ in probability in the following sense:*

   $$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\| \xrightarrow{\mathrm{P}} 0 \quad \textit{as } n_1 \to \infty.$$

   This, of course, would also prove that $\hat{\mathbf{p}}$ converges to $\boldsymbol{p}$ in probability as $n_2 \to \infty$, since the problem is symmetric in the two populations.

   Solution. $\hat{\theta_k} = n_1^{-2} \sum_{i=1}^{n_1} X_i$, where $X_i = 1$ if the kth coordinate of the ith sample from the $\boldsymbol{\theta}$-population is one, and $X_i = 0$ otherwise. Therefore, $\hat{\theta_k} \xrightarrow{\mathrm{P}} \mathrm{E}(X_1)$ as $n_1 \to \infty$ by the law of large numbers. This does the job, because $\mathrm{E}(X_1) = \theta_k$.

   (b) *Prove that the random vector $\sqrt{n_1}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\}$ has a limiting distribution, as $n_1 \to \infty$. Identify that limiting distribution. Perform the analogous analysis for $\sqrt{n_2}\{\hat{\mathbf{p}} - \boldsymbol{p}\}$ [you do not need to reproduce the work; just work out the statement].*

   Solution. $\hat{\boldsymbol{\theta}}$ has a multinormial distribution, and the material of the lecture notes shows that

   $$\sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}\,, \boldsymbol{Q}) \qquad \text{as } n_1 \to \infty,$$

   where

   $$\boldsymbol{Q} = \begin{pmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_{m-1} & -\theta_1\theta_m \\ -\theta_2\theta_1 & \theta_2(1-\theta_2) & \cdots & -\theta_2\theta_{m-1} & -\theta_2\theta_m \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\theta_m\theta_1 & -\theta_m\theta_2 & \cdots & -\theta_m\theta_{m-1} & \theta_m(1-\theta_m) \end{pmatrix}.$$

   Similarly,

   $$\sqrt{n_2}\,(\hat{\mathbf{p}} - \boldsymbol{p}) \xrightarrow{d} N_m(\mathbf{0}\,, \boldsymbol{W}) \qquad \text{as } n_2 \to \infty,$$

   where

   $$\boldsymbol{W} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{m-1} & -p_1p_m \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_{m-1} & -p_2p_m \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -p_mp_1 & -p_mp_2 & \cdots & -p_mp_{m-1} & p_m(1-p_m) \end{pmatrix}.$$

(c) *Consider the null hypothesis, $H_0 : \boldsymbol{\theta} = \boldsymbol{p}$ against its two-sided alternative. Describe a condition under which you can ensure that the distribution of $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{p}}$ has a large-sample asymptotically normal approximation. Carefully state your central limit theorem.*

Solution. We know that when n_1 and n_2 are large:

- $\hat{\boldsymbol{\theta}} \approx N_m(\boldsymbol{\theta}\,, n_1^{-1}\boldsymbol{Q})$ in distribution; and
- $\hat{\boldsymbol{p}} \approx N_m(\boldsymbol{p}\,, n_2^{-1}\boldsymbol{Q})$ in distribution.

Since the sum of two independent multivariate normals is a multivariate normal, it follows that

$$\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{p}} \approx N_m(\boldsymbol{\theta} - \boldsymbol{p}\,, n_1^{-1}\boldsymbol{Q} + n_2^{-1}\boldsymbol{W}) \quad \text{in distribution.}$$

Under $H_0 : \boldsymbol{\theta} = \boldsymbol{p}$, $\boldsymbol{Q} = \boldsymbol{W}$, and the preceding simplifies to

$$\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{p}} \approx N_m(\boldsymbol{0}\,, \left[n_1^{-1} + n_2^{-1}\right] \boldsymbol{W}) \quad \text{in distribution.}$$

Now consider the special condition that $n_1 = n_2$, and call their common value n. That is,

$$n := n_1 = n_2.$$

Under this condition, $n_1^{-1} + n_2^{-1} = 2n^{-1}$, and hence

$$\sqrt{\frac{n}{2}}[\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{p}}] \xrightarrow{d} N_m(\boldsymbol{0}\,, \boldsymbol{W}),$$

under H_0. We could also write $N_m(\boldsymbol{0}\,, \boldsymbol{Q})$ here.

It is more convenient to rewrite the preceding in the following, equivalent, form:

$$\sqrt{\frac{1}{2n}}[n\hat{\boldsymbol{\theta}} - n\hat{\boldsymbol{p}}] \xrightarrow{d} N_m(\boldsymbol{0}\,, \boldsymbol{W}),$$

under H_0. This is because $n\hat{\boldsymbol{\theta}}$ and $n\hat{\boldsymbol{p}}$ are vectors of observed counts [think of χ^2 tests!].

(d) *Use your central limit theorem to devise a χ^2 test for $H_0 : \boldsymbol{\theta} = \boldsymbol{p}$.*

Solution. The form of the solution depends on the relative behavor of n_1 and n_2. But first recall from the lecture notes that if $\boldsymbol{X} \sim N_m(\boldsymbol{0}\,, \boldsymbol{W})$, then the exact distribution of $\sum_{i=1}^m (X_i/p_i)^2$ is

$$\sum_{i=1}^m \frac{X_i^2}{p_i^2} \sim \chi^2_{m-1}. \tag{Chi}$$

[Under H_0, $p_i^2 = \theta_i^2$.] Therefore, the preceding part tells us that

$$\frac{1}{2n}\sum_{i=1}^m \frac{[n\hat{\theta}_i - n\hat{p}_i]^2}{p_i} \xrightarrow{d} \chi^2_{m-1} \qquad \text{as } n \to \infty,$$

under H_0. Under H_0, the pooled estimate $\hat{\theta}_i + \hat{p}_i$ converges to $(\theta_i + p_i) = 2p_i$ in probability as $n \to \infty$, thanks to the law of large numbers. Therefore, Slutzky's theorem tells us that

$$\sum_{i=1}^m \frac{[n\hat{\theta}_i - n\hat{p}_i]^2}{n\hat{\theta}_i + n\hat{p}_i} \xrightarrow{d} \chi^2_{m-1} \qquad \text{as } n \to \infty,$$

In other words,

$$\boldsymbol{\chi}^2 := \sum_{i=1}^m \frac{(\mathrm{Obs}_i^\theta - \mathrm{Obs}_i^p)^2}{\mathrm{Obs}_i^\theta + \mathrm{Obs}_i^p} \approx \chi^2_{m-1},$$

where:

i. $\mathrm{Obs}_i^\theta := n\hat{\theta}_i$ for all $i = 1, \ldots, m$; and
ii. $\mathrm{Obs}_i^p := n\hat{p}_i$ for all $i = 1, \ldots, m$.

The rest is easy: If $\boldsymbol{\chi}^2 > \chi^2_{m-1}(1-\alpha)$ then reject H_0. This produces an asymptotically-correct level-$(1-\alpha) \times 100\%$ test for $H_0 : \boldsymbol{\theta} = \boldsymbol{p}$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{p}$.