Applied Statistics (ECS764P) - Lab 3

Fredrik Dahlqvist

17 Nov 2022

1 Theory

1. It can be shown that the sum of two χ^2-distributions is a χ^2-distribution. Specifically:

$$\chi^2(k_1) + \chi^2(k_2) = \chi^2(k_1 + \chi_2)$$

Show that the sample mean of N independent and identically distributed χ^2 distributions $\chi^2(k)$ is exactly given by a Gamma distribution with shape parameter $\frac{Nk}{2}$ and scale parameter $\frac{2}{N}$, i.e.

$$\frac{1}{N}\sum_{i=1}^{N}\chi^2(k) = \text{Gamma}\left(\frac{Nk}{2}, \frac{2}{N}\right)$$

Hint: recall from the lecture notes that we've computed the density of αd (where α is some positive number and d is some distribution) in terms of the density of d. Look up the densities of the χ^2 and Gamma distribution online. The rest is just simple algebra.

2. In this exercise you will learn how to normalise/standardize a normal distribution, i.e. show that you can always reduce the computation of a probability mass under $\text{Norm}(\mu, \sigma)$ to the computation of a probability mass under $\text{Norm}(0, 1)$. You will then use this to prove the weak Law of Large Numbers for normal distributions.

   (a) Show from first principles (i.e. from the definition of the pushforward probability measure, see slides) that

$$\text{Norm}(\mu, \sigma)([a, b]) = (\text{Norm}(\mu, \sigma) - \mu)([a - \mu, b - \mu]) = \text{Norm}(0, \sigma)([a - \mu, b - \mu])$$

   (b) Show from first principles that

$$\text{Norm}(0, \sigma)([a, b]) = \frac{1}{\sigma}\text{Norm}(0, \sigma)\left(\left[\frac{a}{\sigma}, \frac{b}{\sigma}\right]\right) = \text{Norm}(0, 1)\left(\left[\frac{a}{\sigma}, \frac{b}{\sigma}\right]\right)$$

   (c) Conclude that

$$\text{Norm}(\mu, \sigma)([a, b]) = \text{Norm}(0, 1)\left(\left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]\right)$$

   and express this quantity in terms of the standard normal cdf Φ.

   (d) In the previous Lab you were asked to show that the distribution of sample means of n independent and identically distributed normal distributions is given by

$$\frac{1}{n}\sum_{i=1}^{n}\text{Norm}(\mu, \sigma) = \text{Norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

   Using this fact you can give a simple and direct proof of the weak Law of Large Number for the case of normal distributions. For this, fix $\varepsilon, \delta > 0$ and find an n such that

$$\frac{1}{n}\sum_{i=1}^{n}\text{Norm}(\mu, \sigma)([\mu - \varepsilon, \mu + \varepsilon]) > 1 - \delta$$

   In other words, find n such that the probably of a sample mean landing within ε of the sample mean is at least $1 - \delta$. (*Hint: use the inverse cdf/quantile function of the standard normal distribution, usually denoted Φ^{-1} .*)

2 Practice

You can assume that `numpy`, `matplotlib` *and* `scipy` *are installed on the machine of the person who will run and mark your notebook. There is no need to force an install with the* `!` *command. For textual answers please use a markdown cell.*

1. It can be shown (see theory exercises) that the sample mean of N independent and identically distributed χ^2-distributions $\chi^2(k)$ is *exactly* given by a Gamma distribution with shape parameter $\frac{Nk}{2}$ and scale parameter $\frac{2}{N}$ (see `https://en.wikipedia.org/wiki/Gamma_distribution`). Now follow these steps:

   (a) Create a 2-by-3 array of subplots. Fix $k = 3$ and instantiate an array $\texttt{N} = [5, 10, 50]$ and a variable $\texttt{size} = 100,000$.

   (b) Using a `for` loop, for each value `n` in `N` sample a `size` $\times$ `n` array of samples from the distribution $\chi^2(k)$.

   (c) Compute the sample average along each row (i.e. you should get `size` sample averages), and plot their histogram in a subplot.

   (d) Over the histogram (i.e. in the same subplot), plot the exact density of the distribution of sample averages which is given by the gamma distribution described above (*Hint: make sure you input the values of N and k correctly, and check how shape and scale parameters are input in the* `scipy.stats.gamma` *documentation*). You should get an extremely good agreement between the pdf and the histogram.

   (e) In a separate subplot, display the QQ plot of the sample means versus their exact (Gamma) distribution.

2. The Central Limit Theorem (CLT) shows that for sufficiently large values of N, the sample mean of N independent and identically distributed χ^2-distributions $\chi^2(k)$ is *approximately* given by a Normal Distribution with mean $E[\chi^2(k)] = k$ and variance $\frac{Var[\chi^2(k)]}{N} = \frac{2k}{N}$. Follow these steps (the first three are exactly like in Q1):

   (a) Create a 2-by-3 array of subplots. Fix $k = 3$ and instantiate an array $\texttt{N} = [5, 10, 50]$ and a variable $\texttt{size} = 100,000$.

   (b) Using a `for` loop, for each value `n` in `N` sample a `size` $\times$ `n` array of samples from the distribution $\chi^2(k)$

   (c) Compute the sample average along each row (i.e. you should get `size` sample averages), and plot their histogram in a subplot.

   (d) Over the histogram (i.e. in the same subplot), plot the *approximate* density of the distribution of sample averages which is given by the CLT as described above.

   (e) In a separate subplot, display the QQ plot of the sample means versus their *approximate* distribution

   For which value N is the *approximate* density of sample means given by the CLT indistinguishable from the *actual* density you've plotted in Q1?

3. Suppose that you're working for the UK Department of Health and Social Care, and that you have been tasked with establishing the prevalence of COVID-19 among primary school children. For the sake of simplicity we will assume that every class has 30 pupils. The number of COVID-positive children in a class is modelled by a binomial distribution $\text{Binom}(30, p)$, and the last time this study was done it was established that $p = 0.1$. You want to test if this has changed significantly, where significantly means with $\alpha = 95\%$. State your H_0.

   You start your study with 4 classes, test every children and report a total of 20 cases. Since $\text{Binom}(N, p)$ is a sum of N Bernoulli distributions with parameter p, it is clear that the sum of four $\text{Binom}(30, p)$ is $\text{Binom}(120, p)$. In your notebook compute:

   - The probability under H_0 of having observed 20 cases.
   - The probability under H_0 of having observed an outcome at least as extreme as 20 cases.

   Discuss whether to reject H_0 or not. What does it say about the epidemic?

   Your research gets extended, and you now test 300 classes. With so many classes, it is no longer practical to use an exact test with a Binomial distribution ($N = 9000$). Explain why in at most two sentences. (*Hint: think about what you have just computed.*)

You therefore use the CLT to test for the value of p. The variance of Binom $(30, p)$ can be easily computed exactly, so it makes sense to use the Z-test. Your confidence level remains at $\alpha = 95\%$ and you now observe a total of 945 cases.

Compute the Z-statistic. For this, make sure you can answer the following:

- What is the theoretical mean under H_0?
- What is the observed sample mean?
- What is the sample size?
- What is the variance of Binom $(30, p)$?

Now compute the p-value in your notebook and discuss whether you would you reject your H_0 or not and what is says about the epidemic and your previous, small-scale trial.