Detecting Causal Structure in Time Series Data

Lawrence Tray
Ioannis Kontoyiannis

October 22, 2020

Abstract

1 Introduction

2 Hypothesis Testing

2.1 Single Sample against known mean

We start with the simple case of determining whether or not a coin is fair. Each coin flip can be represented by a random variable with a Bernoulli distribution $X_i \sim Bern(p)$. Each coin can result in either a Tails or a Heads denoted by $X_i \in \{0, 1\}$. We toss the coin n times leading us to a set $\{X_i\}_{i=1}^n$. We wish to test the null hypothesis $p = p_0$ against the alternative. We shall keep the p_0 notation for generality, though for a fair coin we require $p_0 = 1/2$.

$$\begin{aligned} H_0 : \quad & p = p_0 \\ H_1 : \quad & p \neq p_0 \end{aligned}$$

We denote the number of heads with the random variable $K \coloneqq \sum_{i=1}^n X_i \sim Bern(p, n)$. For a particular experiment we observe $K = k$. We employ a standard likelihood ratio test. The test statistic t_n is calculated as the log-likelihood ratio of observing $K = k$ under H_1 and H_0.

$$t_n \coloneqq \log \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \log \frac{\max_p P(K = k|p)}{P(K = k|p = p_0)} \tag{1}$$

We note that since K is distributed as a binomial, $P(K = k|p) = \binom{n}{k} p^k (1-p)^{n-k}$. If we can vary p, this probability is maximised for $p = \hat{p} \coloneqq k/n$. Therefore, the test statistic is given by.

$$t_n = \log \frac{\binom{n}{k} \hat{p}^k (1-\hat{p})^{n-k}}{\binom{n}{k} p_0^k (1-p_0)^{n-k}} = \log \frac{\hat{p}^k (1-\hat{p})^{n-k}}{p_0^k (1-p_0)^{n-k}} \tag{2}$$

The combinatoric term $\binom{n}{k}$ cancels out. This implies that the order in which the heads land does not matter when determining the fairness of the coin. We can work the above expression into a more usable form:

$$\begin{aligned} t_n =& k \log \frac{\hat{p}}{p_0} + (n-k) \log \frac{1-\hat{p}}{1-p_0} \\ =& n \left(\hat{p} \log \frac{\hat{p}}{p_0} + (1-\hat{p}) \log \frac{1-\hat{p}}{1-p_0} \right) \\ =& n \mathcal{D} \left(Bern(\hat{p}) || Bern(p_0) \right) \end{aligned} \tag{3}$$

Where $\mathcal{D}(f|g) \coloneqq \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)}$ is the Kullback-Leibler divergence between two arbitrary probability mass functions f and g. This is also called the relative entropy. The KL divergence has the property that:

$$\mathcal{D}(f||g) \geq 0 \quad \text{with equality iff} \quad f(x) = g(x) \quad \forall x \in \mathcal{X} \tag{4}$$

We can exploit this result for the case that $f \approx g$ to obtain a simplified expression for the KL divergence. We begin by defining $\delta(x) \coloneqq f(x) - g(x)$. We are interested in the region where δ is small. We start by substituting for $f = \delta + g$ and then taking the Taylor expansion of $\log 1 + x$.

$$\begin{aligned}
\mathcal{D}(f||g) &= \sum_{x\in\mathcal{X}} (\delta + g)\log\left(1 + \frac{\delta}{g}\right) \\
&= \sum_{x\in\mathcal{X}} (\delta + g)\left(\frac{\delta}{g} - \frac{\delta^2}{2g^2} + O(\delta^3)\right) \\
&= \sum_{x\in\mathcal{X}} \delta + \frac{1}{2}\sum_{x\in\mathcal{X}} \frac{\delta^2}{g} + O(\delta^3) \\
&= \frac{1}{2}\sum_{x\in\mathcal{X}} \frac{\delta^2}{g} + O(\delta^3) \\
&= \frac{1}{2}\chi^2(f||g) + O(\delta^3)
\end{aligned}$$

Where the summation over δ evaluates to 0 because δ is the difference of two valid p.m.f's which each sum to 1 over $x \in \mathcal{X}$. We are able to neglect the $O(\delta^3)$ terms for f very close to g we shall see what this means later. $\chi^2(f||g)$ is known as the chi-squared distance between two distributions and is defined simply as $\chi^2(f||g) \coloneqq \sum_{x\in\mathcal{X}}(f-g)^2/g$. We now investigate the chi-squared divergence for $f = Bern(p)$ and $g = Bern(q)$.

$$\begin{aligned}
\chi^2(Bern(p)||Bern(q)) &= \frac{(p-q)^2}{q} + \frac{((1-p)-(1-q))^2}{1-q} \\
&= \frac{(p-q)^2}{q(1-q)} \\
&= \left(\frac{p-q}{\sqrt{q(1-q)}}\right)^2
\end{aligned}$$

Now we can exploit these results to get a workable expression for the test statistic t_n.

$$\begin{aligned}
t_n &= n\mathcal{D}\left(Bern(\hat{p})||Bern(p_0)\right) \\
&= \frac{n}{2}\chi^2\left(Bern(\hat{p})||Bern(p_0)\right) + nO(\delta^3) \\
&= \frac{1}{2}\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)^2 + \epsilon
\end{aligned}$$

So far we have been treating t_n as deterministic but it is merely an observation a random variable. To make this distinction clear we shall use upper-case to refer to random variables. Therefore, we have that the random variable T_n is a function of $\hat{P} \coloneqq K/n$. For large n, we can find the distribution of $\hat{P}$ by the Central Limit Theorem.

$$\begin{gathered}
\hat{P} = \frac{K}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \\
H_0 : \mathbb{E}[X_i] = p_0, Var(X_i) = p_0(1-p_0) \\
\text{CLT, as } n \to \infty : \hat{P} \sim \mathcal{N}\left(\mu = p_0, \sigma^2 = p_0(1-p_0)/n\right) \\
\therefore \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} = Z \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)
\end{gathered}$$

Therefore neglecting the error term ϵ[1], we have that under H_0, for sufficiently large n.

$$T_n = \frac{1}{2}Z^2 \sim \frac{1}{2}\chi_1^2 \tag{5}$$

By the definition of the chi-squared distribution with one degree of freedom. To reject the null hypothesis H_0 at the $100(1-\alpha)\%$ confidence level, we require that $P(T_n = t_n|H_0) < \alpha$. In other words, a low probability of observing this result under the null hypothesis.

[1] See the Appendix for proof of negligibility

2.2 Two samples equality of means

We now complicate things by introducing a second coin, with throws denoted by $\{Y_i\}_{i=1}^m$ where we assume each throw is i.i.d Bernoulli with parameter q ($Y_i \sim Bern(q)$). Note that the population sizes n and m may be different. We define $L \coloneqq \sum_{i=1}^m Y_i$ which is the analogue of K. We now set up our hypotheses to be:

$$\begin{aligned} H_0 :& \quad p = q \\ H_1 :& \quad p \neq q \end{aligned}$$

Proceeding as before we can derive a formula for the test statistic. This time we denote the test statistic by t_N where $N \coloneqq n + m$.

$$\begin{aligned} t_N \coloneqq \log \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} &= \log \frac{\max_{p,q} P(K = k|p)P(L = l|q)}{\max_p P(K = k|p)P(L = l|q = p)} \\ &= \log \frac{\binom{n}{k}\hat{p}^k(1-\hat{p})^{n-k}\binom{m}{l}\hat{q}^k(1-\hat{q})^{m-l}}{\binom{n}{k}\hat{r}^k(1-\hat{r})^{n-k}\binom{m}{l}\hat{r}^l(1-\hat{r})^{m-l}} \\ &= \log \frac{\hat{p}^k(1-\hat{p})^{n-k}}{\hat{r}^k(1-\hat{r})^{n-k}} + \log \frac{\hat{q}^k(1-\hat{q})^{m-l}}{\hat{r}^l(1-\hat{r})^{m-l}} \end{aligned}$$

Where, $\hat{p} \coloneqq k/n, \hat{q} \coloneqq l/m, \hat{r} \coloneqq (k+l)/(n+m) = (n\hat{p} + m\hat{q})/(n+m)$. Using the same tricks as before, we can express this in terms of the chi-squared distance between the various parameters:

$$\begin{aligned} t_N &= \frac{n}{2}\chi^2\left(Bern(\hat{p}||\hat{r})\right) + \frac{m}{2}\chi^2\left(Bern(\hat{q}||\hat{r})\right) + \epsilon \\ &\approx \frac{1}{2\hat{r}(1-\hat{r})}\left(n(\hat{p}-\hat{r})^2 + m(\hat{q}-\hat{r})^2\right) \\ &= \frac{1}{2\hat{r}(1-\hat{r})}\left(n\left(\frac{m(\hat{p}-\hat{q})}{n+m}\right)^2 + m\left(\frac{n(\hat{q}-\hat{p})}{n+m}\right)^2\right) \\ &= \frac{nm(\hat{p}-\hat{q})^2}{2\hat{r}(1-\hat{r})(n+m)} \\ &= \frac{1}{2}\left(\sqrt{\frac{nm}{\hat{r}(1-\hat{r})(n+m)}}(\hat{p}-\hat{q})\right)^2 \end{aligned}$$

Under the null hypothesis H_0, we require $p = q (= \mu)$; we introduce this third variable μ to refer to the true mean to avoid ambiguity. Applying the central limit theorem (for sufficiently large n and m) and combining Gaussians in the standard way, we have that:

$$\begin{aligned} \hat{P} &\sim \mathcal{N}\left(\mu, \frac{\mu(1-\mu)}{n}\right) \\ \hat{Q} &\sim \mathcal{N}\left(\mu, \frac{\mu(1-\mu)}{m}\right) \\ \hat{R} &\sim \mathcal{N}\left(\mu, \frac{\mu(1-\mu)}{n+m}\right) \\ \therefore \sqrt{\frac{nm}{\mu(1-\mu)(n+m)}}(\hat{P}-\hat{Q}) &= Z \sim \mathcal{N}(0,1) \\ \delta\hat{R} \coloneqq \hat{R} - \mu &\sim \mathcal{N}\left(0, \frac{\mu(1-\mu)}{n+m}\right) \end{aligned}$$

We almost have T_n we just need to demonstrate that $\hat{R}(1-\hat{R})$ is sufficiently close to $\mu(1-\mu)$ for our purposes.

$$\begin{aligned} \hat{R}(1-\hat{R}) &= (\mu + \delta\hat{R})(1 - (\mu + \delta\hat{R})) \\ &= \mu(1-\mu) + O(\delta\hat{R}) \\ \therefore \frac{1}{\hat{R}(1-\hat{R})} &= \frac{1}{\mu(1-\mu)}\left(\frac{1}{1 + O(\delta\hat{R})}\right) \\ &= \frac{1}{\mu(1-\mu)}(1 + O(\delta\hat{R})) \\ &\approx \frac{1}{\mu(1-\mu)} \end{aligned}$$

We can neglect the terms of order $\delta \hat{R}$ and higher powers, as it is zero mean and for sufficiently large $n+m$ the variance approaches 0. Therefore, we have the desired expression for T_N.

$$T_N \approx \frac{1}{2}\left(\sqrt{\frac{nm}{\mu(1-\mu)(n+m)}}(\hat{P}-\hat{Q})\right)^2 = \frac{1}{2}Z^2 \sim \frac{1}{2}\chi_1^2 \tag{6}$$

For this test we can also use the z-statistic instead of the t-statistic. Since the former is distributed like a Gaussian it may be easier to deal with.

$$z_N = \frac{\hat{p}-\hat{q}}{\sqrt{\hat{r}(1-\hat{r})\,(1/n+1/m)}} \sim \mathcal{N}(0,1) \tag{7}$$

3 Appendix

3.1 Proving H.O.T can be neglected

We have shown that under the null hypothesis, that for sufficiently large n, $(\hat{P}-p_0) \sim \mathcal{N}(0, \beta/n)$ for some positive finite constant β (in this case $\beta = p_0(1-p_0)$ but we just require finiteness for this proof). Therefore, $Q := \sqrt{n}(\hat{P}-p_0) \sim \mathcal{N}(0,\beta)$. Manipulating our expression for the error term ϵ we can show that it can be expressed as a sum of

$$\begin{gathered}
\epsilon = nO(\delta^3) \\
\therefore |\epsilon| \leq n\sum_{i=0}^{\infty}\alpha_i|\hat{P}-p_0|^{3+i} \quad \text{for some finite constants } \alpha_i \geq 0 \\
|\epsilon| \leq \sum_{i=0}^{\infty}\alpha_i n^{-\frac{1+i}{2}}(\sqrt{n}|\hat{P}-p_0|)^{3+i} \\
|\epsilon| \leq \sum_{i=0}^{\infty}\alpha_i n^{-\frac{1+i}{2}}|Q|^{3+i}
\end{gathered} \tag{8}$$

We know that Q is a Gaussian of zero mean and finite variance, therefore $|Q|$ will be a finite value. However, we see that it is scaled by a negative power of n, therefore for sufficiently large n we have that the error term asymptotes to 0. To be precise:

$$\lim_{n\to\infty} P(|\epsilon| < \eta) = 1 \text{ for arbitrarily small } \eta \geq 0 \tag{9}$$