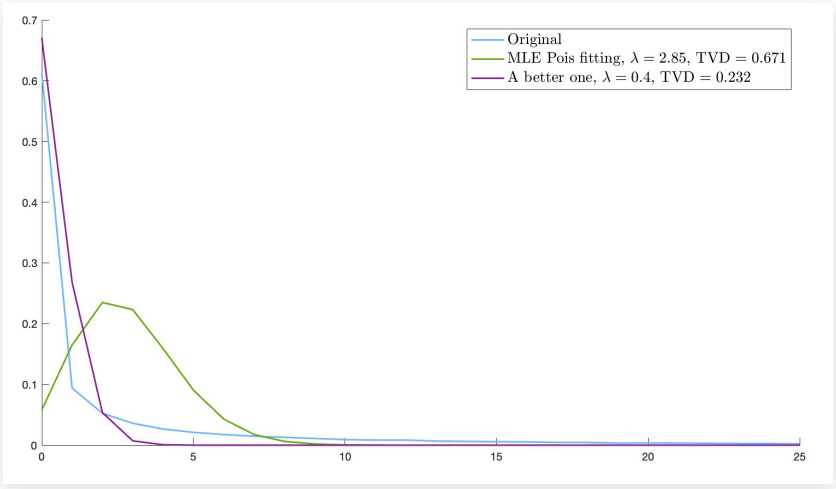
# MATH 408 Homework 5

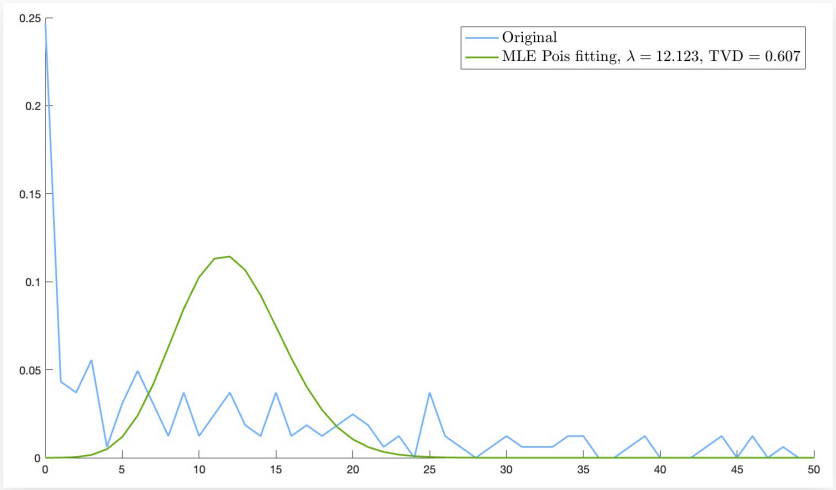
Qilin Ye

October 28, 2021

**Problem: Graphs for Problem 1**




Poisson distribution fitting for the baseball example




Poisson distribution fitting for the WNBA example

**Problem 3**

Suppose you flip a coin $1000$ times, resulting in $560$ heads and $440$ heads. Is it reasonable to conclude that the coin is fair?

*Solution.* Flipping a coin can be thought of as a Bernoulli random variable with parameter $\lambda \in [0,1]$. Let $X_i \sim \text{Bern}(\lambda)$ and define $S := \sum_{i=1}^{1000} X_i = \#\text{heads}$. Define the rejection region to be $\{(x_1,...,x_{1000}) \in \mathbb{R}^{1000} : |S-500| > c\}$ so that $t(X) := |\#\text{heads} - 500|$. $\Theta_0$ is simply $\{1/2\}$ as we assume that $H_0$ states the coin is fair. Then,

$$p(\text{560 heads}) = \mathbb{P}_{1/2}(S < 440 \text{ or } S > 560).$$

By CLT, $S$ is approximately $\mathcal{N}(500, \sqrt{1000 \cdot (1/2) \cdot (1-1/2)}) = \mathcal{N}(500, 5\sqrt{10})$. Hence

$$p(\text{560 heads}) = \mathbb{P}(\mathcal{N}(500,5\sqrt{10}) < 440 \text{ or } > 560) \approx \mathbb{P}\left(Z < \frac{60}{5\sqrt{10}} \text{ or } > \frac{60}{5\sqrt{10}}\right) = 2\Phi(12/\sqrt{10})$$

which is extremely small. (Here $Z$ denotes the standard normal.) Therefore we reject the null hypothesis, i.e., we claim the coin is biased.

**Problem 4**

Suppose the number of typos in my notes in a given year follows a Poisson distribution. In the last few years, the average number of typos was $15$, and this year, I had $10$ typos in my notes. Is it reasonable to conclude that the rate of typos has dropped this year?

*Solution.* Let $X \sim \text{Pr}(\lambda)$ be the random variable describing Heilman's typos. Since we are only interested in whether the rate has dropped, $\Theta = (0,15]$, $\Theta_0 = \{15\}$, and $\Theta_1 = (0,15)$. Hence $H_0$ is $\{\lambda = 15\}$ and $H_1$ is $\{0 < \lambda < 15\}$. Let the rejection region be defined by $\{x : t(x) \geqslant c\}$ where $t(X) := 15 - X$. Then if $10$ typos are observed, the $p$-value is

$$p(10) = \mathbb{P}_{15}(15 - X \geqslant 15 - 10) = \mathbb{P}_{15}(X \leqslant 10) = \sum_{k=0}^{10} \frac{15^k e^{-15}}{k!} \approx 0.118,$$

so it is likely that Heilman indeed made some improvements in preventing typos.

**Problem 5**

Suppose $X$ is a Gaussian distributed random variable with known variance $\sigma^2 > 0$ but unknown mean $\mu$. Fix $\mu, \mu \in \mathbb{R}$ with $\mu_0 > \mu_1$. We want to test the hypothesis $H_0$ that $\mu = m_0$ versus the hypothesis $H_1$ that $\mu = m_1$. Fix $\alpha \in (0,1)$. Explicitly describe the UMP test for the class of tests whose significance level is at most $\alpha$.

*Solution.* In this case $\Theta = \{\mu_0, \mu_1\}$, with $H_0 : \{\mu = \mu_0\}$ and $H_1 : \{\mu = \mu_1\}$. Neyman-Pearson tells us that the UMP test for all tests with significance level $\leqslant \alpha$ is the likelihood ratio test with level $= \alpha$. Recall that the rejection region is defined by

$$C := \{x \in \mathbb{R}^n : f_{\mu_1}(x) > k f_{\mu_0}(x)\}.$$

Since $f_{\mu_0}(x) > 0$, we may write it in quotient form $f_{\mu_1}(x)/f_{\mu_0}(x) > k$, that is,

$$\text{LR} := \frac{\exp(-(x-\mu_1)^2/2\sigma^2)}{\exp(-(x-\mu_0)^2/2\sigma^2)} = \exp\left(-\frac{(x-\mu_1)^2-(x-\mu_0)^2}{2\sigma^2}\right) > k.$$

Since

$$(x-\mu_1)^2-(x-\mu_0)^2 = \mu_1^2-\mu_0^2-2x(\mu_1-\mu_0) = (\mu_1-\mu_0)(\mu_1+\mu_0-2x),$$

we have

$$\text{LR} = \exp\left(\frac{\mu_1-\mu_0}{\sigma^2}\left(x-\frac{\mu_1+\mu_0}{2}\right)\right).$$

Hence LR $> k$ if and only if $\log(\text{LR}) > \log k$, i.e.,

$$(\mu_1-\mu_0)\left(x-\frac{\mu_1+\mu_0}{2}\right) > \sigma^2\log k,$$

or equivalently

$$x < \frac{\sigma^2\log k}{\mu_1-\mu_0}+\frac{\mu_1+\mu_0}{2} =: \tilde{k}.$$

It remains to find the $\tilde{k}$ such that $\alpha = \mathbb{P}_{\mu_0}(X \in C) = \mathbb{P}_{\mu_0}(X < \tilde{k}) = \alpha$. Since $X \sim \mathcal{N}(\mu_0, \sigma^2)$, this becomes

$$\mathbb{P}_{\mu_0}\left(\frac{X-\mu_0}{\sigma} \leqslant \frac{\tilde{k}-\mu_0}{\sigma}\right) = \Phi\left(\frac{\tilde{k}-\mu_0}{\sigma}\right) = \alpha.$$

Hence

$$\tilde{k} = \sigma\Phi^{-1}(\alpha)+\mu_0.$$

Therefore, the UMP test rejects $H_0$ if and only if

$$x > \tilde{k} = \sigma\Phi^{-1}(\alpha)+\mu_0.$$

**Problem 6**

Let $X_1, \ldots, X_n$ be i.i.d. random variables and denote $X = (X_1, \ldots, X_n)$. Assume $X$ has distribution $f_\theta \in \{f_\theta : \theta \in \Theta\}$. Suppose $\Theta$ consists of two points, i.e., $\Theta = \{\theta_0, \theta_1\}$. Let $Z$ be sufficient for $\theta$. Consider the likelihood ratio test of the null hypothesis $H_0$ that $\{\theta = \theta_0\}$ versus the alternative $H_1$ that $\theta = \theta_1$. Show that the likelihood ratio is a function of $Z$.

*Proof.* Since $Z$ is sufficient for $\theta$, by factorization theorem,

$$f_\theta(x) = g_\theta(z)h(x)$$

for some functions $g_\theta(z)$ and $h(x)$. The likelihood ratio is given by

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = \frac{g_{\theta_1}(z)h(x)}{g_{\theta_0}(z)h(x)} := J(z).$$

Since $f_\theta$ is a PDF, whenever it is nonzero, $h(x)$ is also nonzero, so cancellation holds. This shows that the likelihood ratio is a function of $Z$. □

**Problem 7**

Suppose $X$ is a binomial random variable with parameters $n = 100$ and $\theta \in [0,1]$ unknown. suppose we want to test the hypothesis $H_0$ that $\theta = 1/2$ versus the alternative $H_1$ that $\theta \neq 1/2$. Consider the hypothesis test that rejects $H_0$ if and only if $|X - 50| > 10$.
Using e.g. the CLT, do the following:

(1) Give an approximation to the significance level $\alpha$ of this test.

(2) Plot an approximation of the power function $\beta(\theta)$ as a function of $\theta$.

(3) Estimate $p$-values for this test when $X = 50, 70$, and $90$.

*Solution.* (1) By definition,

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \notin [40,60]) = \mathbb{P}_{1/2}(X \notin [40,60]) = 2\mathbb{P}_{1/2}(X > 60).$$

By CLT, $X$ is approximately $\mathcal{N}(100 \cdot 0.5, \sqrt{100 \cdot 0.5 \cdot (1-0.5)}) = \mathcal{N}(50,5)$, so

$$\alpha = 2\mathbb{P}(Z > 2) \approx 1 - 0.9545 = 0.0455.$$

(2) As $\theta$ varies, $X$ is now approximated by $\mathcal{N}(100\theta, 10\sqrt{\theta(1-\theta)})$, so

$$\begin{aligned}\beta(\theta) &= \mathbb{P}_\theta(X < 40) + \mathbb{P}_\theta(X > 60) \\ &= \mathbb{P}\left(Z < \frac{40 - 100\theta}{10\sqrt{\theta(1-\theta)}}\right) + \mathbb{P}\left(Z > \frac{60 - 100\theta}{10\sqrt{\theta(1-\theta)}}\right)\end{aligned}$$

A plot using the following code:

```
x = (0:0.01:1);
y = normcdf((40-100*x)./(10*sqrt(x.*(1-x)))) + normcdf((-60+100*x)./(10*sqrt(x.*(1-x))));
plot(x,y);
```

(3) The $p$-values are

$$p(50) = \mathbb{P}_{1/2}(|X - 50| > |50 - 50|) = 1,$$

$$p(70) = \mathbb{P}_{1/2}(|X - 50| > |70 - 50|) \approx \mathbb{P}(|Z| > 4) = 2\Phi(-4) \approx 6.334 \cdot 10^{-5},$$

and

$$p(90) = \mathbb{P}_{1/2}(|X - 50| > |90 - 50|) \approx \mathbb{P}(|Z| > 8) = 2\Phi(-8) \approx 1.244 \cdot 10^{-15}.$$