STAT 391
Homework 5
Out Tuesday May 1, 2018
Due Tuesday May 8, 2018

©Marina Meilă
mmp@stat.washington.edu

Problem 1 – Bias and variance for the Poisson distribution
This is a thought experiment of a kind statisticians often do. Imagine that... n data points $x_{1:n}$ are sampled i.i.d. from a Poisson distribution with parameter λ *(and because it's a thought experiment, we assume we know* λ*.)* Recall also that $\lambda^{ML} = \sum_{i=1}^{n} x_i/n$.

a. Calculate $E[\lambda^{ML}]$ as a function of λ. *All expectations are under the true distribution of the data.*

b. Calculate $Var(\lambda^{ML})$ as a function of λ.

c. Assume that n is large enough for the Central Limit Theorem to apply. Express the probability that $\lambda^{ML} \geq \lambda + 1$ as a function of n, λ and Φ the CDF of the standard normal.

d. Numerical answer for $\lambda = 10$, $n = 100$. *Use a table/calculator/computer, don't submit code.*

Problem 2 – Confidence Intervals and Boostrap
Use the data set $\mathcal{D}$ from `hw5-data1.dat`, generated by an unknown density f, to answer the following questions. *Formula and numerical result OK, no proofs needed.*

a. Estimate μ^{ML} the mean of f.

b. Estimate $\sigma^{2,ML}$ the Maximum Likelihood variance of f. Then calculate $\sigma^{2,C}$ the unbiased estimator of $Var\, f$.

c. Estimate the variance and standard deviation of μ^{ML}, pretending that $\sigma^{2,C}$ is the true variance $Var\, f$.

d. Use the CLT approximation to obtain the Confidence Interval (CI) for confidence level $1 - \delta$, for $\delta = 0.01$.

e. Now estimate the variance of μ^{ML} by Bootstrap; denote this by $\sigma^{2,B}$. Take $B = 1000$ bootstrap samples, and calculate from them the numerical value of $\sigma^{2,B}$.

f. Use the CLT approximation again to obtain the Confidence Interval (CI) for confidence level $1 - \delta$, for $\delta = 0.01$, from the new variance estimator $\sigma^{2,B}$. *You will obtain slightly different values in* **c** + **d**, *vs* **e** + **f**. *How different are they? The natural unit of measure in probability is the standard deviation of the quantity measured.* Take $SD_C(\mu^{ML})$, the standard deviation of μ^{ML} (NOT SQUARED) computed in (c) as the unit. Then, calculate

$$e_r = \frac{|SD_C(\mu^{ML}) - \sigma^B|}{SD_C(\mu^{ML})}.$$

If this is "small" (that is, much smaller than 1), then the two CI's are "close". *You can also measure the overlap of the intervals, relative to the length of one of them. Note that this will be exactly equal to* $1 - e_r$.

Problem 3 – Median of Means (MOM)
A recent method for estimating the mean of a distribution is MOM[1].

The MOM estimator of the mean is computed as follows: 1. Divide the data set into K equal subsets of equal size m (assume that $n = mK$ exactly). 2. Compute μ_k the mean of subset k, $k = 1 : K$. 3. Compute the median of the μ_k's.

$$\mu^{MOM} \;=\; \text{median}(\mu_1, \ldots \mu_K), \quad \text{where } \mu_k = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} x_i, \quad \text{and } m = \frac{n}{K}. \tag{1}$$

[1]See e.g. `http://www.ub.edu/focm2017/slides/Lugosi.pdf`.

Why the trouble? It can be proved that μ^{MOM} is **robust**, that is, it is less influenced by outliers than μ^{ML} the mean of the data. Verify this on the data set `hw5-data2.dat`.

a. Compute μ^{ML} the mean of the data, and μ^{MOM} for $K = 56$ ($n = 2800$ for this data set).

b. Extract $B = 1000$ bootstrap samples, and compute $\mu^{MOM,b}$ and $\mu^{ML,b}$ for $b = 1 : B$. Then estimate the variance of μ^{MOM}, μ^{ML} by bootstrap. Does the experiment agree witht the theory?

[**c. Extra credit**] Repeat **a, b** for the data from the previous problem. Do you observe any difference?

Some theory The choice of K depends on the desired *confidence level* δ. If $K = 8 \log_2 \frac{1}{\delta}$ then

$$|\mu^{MOM} - \mu| \;\leq\; 2\frac{\sigma}{\sqrt{m}}, \quad \text{with probability } \geq 1 - \delta, \tag{2}$$

where $m = n/K$, and μ, σ are the unknown mean and standard deviation.

So, $K = 56$ corresponds to $\delta = \frac{1}{128}$.