Math 40 Probability and Statistical Inference
Winter 2021
Lecture 16 Statistics as Inverse Probability (6.1)
Methods of Moments (6.2)
Method of Quantiles (6.3)

Yoonsang Lee (yoonsang.lee@dartmouth.edu)

6.1 What is statistics?

In Math 40, we focus on the following task:

- There are variables of interest, say θ, but unknown. Statistics aim for the **quantification of the unknown variables** using **Data**.
- Example: By measuring the heights of all students, calculate the average height of the students. In this case, the variable of interest is the average height.
- Example: By measuring the time series of a stock price, estimate the volatility of the stock. The variable of interest is the volatility, or the variance of the time series.
- That is, we are interested in **parametric** inference problems.

Example 6.1 A penny of 0.75 inches in diameter is randomly dropped 100 times on a lined paper. It intersected a line in 80 throws. Estimate the distance, say θ, between the lines.

Solution

- (Specify the variable of interest) The parameter of unknown is θ, the distance between the lines.
- (What do we know? What is the data?) The probability 0.80 that the coin intersects the lines.
- (What is the model to explain the probability?) Let X be the center of the coin, which we assume to be a uniform random variable in $(0, \theta)$.
- The coin does not intersect the lines if $X - 0.75/2 > 0$ or $X + 0.75/2 < \theta$.
- The probability of the previous event is given by $\frac{\theta - 0.75}{\theta}$
- The data says the probability, $\frac{\theta-0.75}{\theta}$, is equal to 0.80.
- We estimate that $\theta = \frac{0.75 \times 100}{80} \approx 0.94in.$

Example In a jar, there are balls numbered from 1 to θ, but θ is unknown. You are allowed to draw three balls out of the jar, and they turn out to be 4, 11,12. Estimate θ using the data.

Solution of Student A

As the maximum of the sample is 12, I estimate that $\theta = 12$.

Solution of Student B

Let X be a random variable uniform in $(0, \theta)$. We know that $E(X) = \frac{\theta}{2}$ and the sample mean is $\frac{4+11+12}{3} = 9$. Thus I estimate that $\theta = 9 \times 2 = 18$.

Solution of Student C

Well, I think I have a better idea. The median of X is $\frac{\theta}{2}$, and the sample median is 11. Thus, I estimate that $\theta = 22$.

Example In a jar, there are balls numbered from 1 to θ, but θ is unknown. You are allowed to draw three balls out of the jar, and they turn out to be 4, 11,12. Estimate θ using the data.

Solution of Student A
This approach is called the Maximum Likelihood Estimation (MLE; section 6.10)

Solution of Student B
This approach is called the Method of Moments (MM; section 6.2)

Solution of Student C
As you can guess, this approach is called the Method of Quantiles (MQ; section 6.3)

Note By coincidence, the Student A's approach is also equal to the method of quantiles using the 100% quantile.

There is an important question to ask. Which method is better (or more accurate)? (section 6.4)

This is a glimpse of our journey for Chapter 6, parameter estimation.

6.2 Method of moments (MM)

Review of probability (review of section 2.2) For a random variable X, we know how to calculate the central and non-central moments, μ_k and ν_k

$$\mu_k = E((X - \mu)^k, \quad \mu = E(X),$$

and

$$\nu_k = E(X^k).$$

- The probability density of X is parameterized by the unknown variable θ.
- There are two ways to calculate the moments of X, i) using the parameterized density or ii) sample moments.
- The method of moments compare the two types of moments (from the parameterized density and from the sample) to find θ.
- As you may guess, it is preferred to use the non-central moments (why?)

6.2 Method of moments (MM)

Example Let $\{X_i\}_{i=1}^n$ be IID from $\mathcal{N}(\mu, \sigma^2)$. Estimate μ and σ^2.

Solution

- $\hat{\mu} = E(X) = \frac{1}{n}\sum_i^n X_i = \overline{X}$
- $\hat{\mu}^2 + \hat{\sigma}^2 = E(X^2) = \frac{1}{n}\sum_i^n X_i^2$
- Solve for $\hat{\mu}$ and $\hat{\sigma}$, which yields

$$\hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_i^n (X_i - \overline{X})^2$$

Q What is the difference between

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i^n (X_i - \overline{X})^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X})^2?$$

Check R code `6_2_samp_var_test.R` on Canvas.

6.2 Method of moments (MM)

Example 6.7 Let $\{X_i\}_i^n$ be IID from $\mathcal{E}(\lambda)$. Use $M(x) = e^{-x}$ to estimate λ for an exponential distribution, where k is a given positive number.

Solution For an exponential distribution with λ,

$$E(M(X)) = E(e^{-X}) = \lambda \int_0^\infty e^{-x} e^{-\lambda x} dx = \frac{\lambda}{\lambda + 1}$$

(check the textbook for details). On the other hand, the sample approximation of $E(M(X))$ is

$$\frac{1}{n} \sum_i^n e^{-X_i}.$$

Therefore,

$$\lambda = \frac{-a}{a-1}$$

where

$$a = \frac{1}{n} \sum_i^n e^{-X_i}.$$

Check R code `6_2_expdist_estimate.R` on Canvas.

6.3 Method of quantiles (MQ)

- Key idea: compare the p-th quantile of the parameterized model with the sample p-th quantile.

- If `X` is a R array containing the data,

  `median(X)` is the median.

- For a general p-th quantile, use $quantile(X, p)$.

Example 6.9 Find the MQ estimator of the rate parameter λ in the exponential distribution using the median.

Solution

- Find the median in terms of λ. That is, we look for q such that

$$1 - e^{-\lambda q} = 0.5.$$

- As $\lambda = \frac{\ln 2}{q}$, we use the sample median $\hat{q}$ to estimate λ using

$$\hat{\lambda} = \frac{\ln 2}{\hat{q}}$$

Check R code `6_2_expdist_estimate.R` on Canvas.

6.3 Method of quantiles (MQ)

Example 6.10 Find the MQ estimator of μ and σ^2 of $\mathcal{N}(\mu, \sigma^2)$.

Solution

- The p-th quantile of $\mathcal{N}(\mu, \sigma^2)$, say q_p, is

$$q_p = \mu + \sigma q_p^{st}$$

  where q_p^{st} is the p-th quantile of the standard normal.

- As there are two parameters to estimate, we need to use two quantiles, say p_1- and p_2-th quantiles.

- Using two quantiles, p_1 and p_2, we need to solve for μ and σ using

$$q_{p_1} = \mu + \sigma q_{p_1}^{st}$$

$$q_{p_2} = \mu + \sigma q_{p_2}^{st}$$

  where q_{p_1} and q_{p_2} are approximated by the sample qunatiles.

Time to think in the context of data science
Q: What are we missing?

- How do you know the distribution type of your data?
- If you have different estimators, which one is your choice?