MTH 224, Spring 2024

Instructor: Bahman Angoshtari

Lecture 20

Sections 6.1 and 6.2: Sample mean and variance, sample median and quartiles.

20.1. The central limit theorem (continued)

Exercise 20.1. (Predicting the outcome of elections). In order to predict the outcome of a presidential election, a poll is taken. The predicted percentage of votes for candidate A is computed from the poll. How many people need to be included in the poll to validate the following claim:

"The predicted percentage is accurate within 1% with probability of at least ≥ 0.95" ?

Solution. Let $p =$ actual percentage, $n =$ poll size. Let $S_n =$ the number of people in favor of candidate A. Note that $S_n \sim \text{Bin}(n, p)$ and that the predicted percentage is $\frac{S_n}{n}$.

We need to find n large enough so that $P\left(\left|\frac{S_n}{n} - p\right| \leq 0.01\right) \geq 0.95$. We have

$$P\left(\left|\frac{S_n}{n} - p\right| \leq 0.01\right) = P\left(|S_n - np| \leq 0.01n\right) = P\left(\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| \leq 0.01 \cdot \sqrt{\frac{n}{p(1-p)}}\right)$$

$$\approx \phi\left(0.01 \cdot \sqrt{\frac{n}{p(1-p)}}\right) - \phi\left(-0.01 \cdot \sqrt{\frac{n}{p(1-p)}}\right) = 2\phi\left(0.01 \cdot \sqrt{\frac{n}{p(1-p)}}\right) - 1.$$

We need to find n such that

$$2\phi\left(0.01 \cdot \sqrt{\frac{n}{p(1-p)}}\right) - 1 \geq 0.95 \iff \phi\left(0.01 \cdot \sqrt{\frac{n}{p(1-p)}}\right) \geq \frac{1.95}{2} = 0.975.$$

Applying the inverse function ϕ^{-1} to both side yields,

$$0.01 \cdot \sqrt{\frac{n}{p(1-p)}} \geq \phi^{-1}(0.975) \approx 1.96 \implies n \geq 38,416 \cdot p(1-p).$$

Only one problem remains, we do not know the value of p! However, we can check that maximum value of $p(1-p)$ is $\frac{1}{4}$ (when $p = \frac{1}{2}$). Thus, regardless of what the actual value of p is, we have

$$38,416 \cdot p(1-p) \leq \frac{38,416}{4} = 9604.$$

Finally, if $\boxed{n \geq 9604}$, we would be certain that $n \geq 38,416 \cdot p(1-p)$ (regardless of the value of p) which, in turn, yields that $P\left(\left|\frac{S_n}{n} - p\right| \leq 0.01\right) \geq 0.95$.

20.2. simple random samples (SRS)

- A **population** is the entire collection of objects or outcomes about which information is sought.

- A **sample** is a subset of a population, containing the objects or outcomes that are actually observed.
- A **simple random sample (SRS)** of size n is a sample chosen by a method in which each collection of n population items is equally likely to comprise the sample, just as in a lottery.
- If measurements are based on an SRS, then we can assume that they are i.i.d. (independent and identically distributed).

20.3. Sample mean, sample variance, and sample standard deviation

- A sample is often a long list of numbers. To help make the important features of a sample stand out, we compute summary statistics. The **sample mean** gives an indication of the center of the data, and the **sample standard deviation** gives an indication of how spread out the data are.
- Let $X_1, X_2, \ldots, X_N$ be a sample. Then:

$$\text{Sample mean:} \qquad \overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n = \frac{X_1 + \cdots + X_N}{N}$$

$$\text{Sample variance:} \qquad S_X^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X})^2 = \frac{1}{N-1}\left[\left(\sum_{n=1}^{N} X_n^2\right) - N\overline{X}^2\right]$$

$$\text{Sample standard deviation:} \qquad S_X = \sqrt{S_X^2} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X})^2}$$

Note that for sample variance, we divide by $N-1$, while for sample mean, we divide by N. The reason will be explained shortly.

Example 20.2.

A sample of 100 cars driving on a freeway during a morning commute was drawn, and the number of occupants in each car was recorded. The results were as follows:

Occupants	1	2	3	4	5
Number of Cars	70	15	10	3	2

Find the sample mean and sample standard deviation of the number of occupants.

Solution.

$$\overline{X} = \frac{\sum_{i=1}^{100} X_i}{100} = \frac{70 \times 1 + 15 \times 2 + 10 \times 3 + 3 \times 4 + 2 \times 5}{100} = 1.52$$

$$S_X^2 = \frac{\sum_{i=1}^{1}(X_i - 1.52)^2}{99} = \frac{1}{99}\left[\left(\sum_{i=1}^{100} X_i^2\right) - 100 \times 1.52^2\right]$$

$$= \frac{1}{99}\left[70 \times 1^2 + 15 \times 2^2 + 10 \times 3^2 + 3 \times 4^2 + 2 \times 5^2 - 1.52^2\right] \approx 0.8783838$$

$$S_X = \sqrt{S_X^2} \approx 0.937$$

- Let $X_1, X_2, \ldots, X_N$ be a sample and define $Y_n = a + bX_n$ for $n = 1, \ldots, N$, in which a and b are constants. We then have that:

$$\overline{Y} = a + b\overline{X}$$
$$S_Y^2 = b^2 S_X^2$$

  Show these by using the definition of sample mean and variance!
  Note that a affects $\overline{Y}$ but does not affect S_Y^2. Can you explain why?
- Assume that $X_1, X_2, \ldots, X_N$ are i.i.d. with mean μ and variance σ^2.
- What is $\mathbb{E}[\overline{X}]$?

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[X_i\right] = \frac{N\mu}{N} = \mu$$

- What is $\mathbb{E}[S_X^2]$?

$$\begin{aligned}\mathbb{E}[S_X^2] = \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2\right] &= \frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}\left[X_i^2 + \overline{X}^2 - 2X_i\overline{X}\right]\\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[X_i^2\right] + \mathbb{E}\left[\overline{X}^2\right] - 2\mathbb{E}\left[X_i\overline{X}\right]\right)\end{aligned}$$

We have:

$$\mathbb{E}\left[X_i^2\right] = \text{Var}(X_i) + (\mathbb{E}[X_i])^2 = \sigma^2 + \mu^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\overline{X}^2\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}X_i\right)^2\right] &= \mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}X_iX_j\right]\\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[X_iX_j\right] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big(\text{Cov}(X_i, X_j) + \mathbb{E}[X_i]\mathbb{E}[X_j]\Big)\\ &= \frac{1}{N^2}\left(\sum_{i=1}^{N}\text{Var}(X_i) + \sum\sum_{i\neq j}\text{Cov}(X_i, X_j) + \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[X_i]\mathbb{E}[X_j]\right)\\ &= \frac{1}{N^2}\left(N\sigma^2 + 0 + N^2\mu^2\right) = \frac{\sigma^2}{N} + \mu^2\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[X_i\overline{X}\right] = \mathbb{E}\left[X_i\left(\frac{1}{N}\sum_{j=1}^{N}X_j\right)\right] &= \frac{1}{N}\mathbb{E}\left[X_i^2 + \sum_{j\neq i}X_iX_j\right]\\ &= \frac{1}{N}\left(\text{Var}(X_i) + (\mathbb{E}[X_i])^2 + \sum_{j\neq i}\left(\text{Cov}(X_i, X_j) + \mathbb{E}[X_i]\mathbb{E}[X_j]\right)\right)\\ &= \frac{1}{N}\left(\sigma^2 + \mu^2 + \sum_{j\neq i}\left(0 + \mu^2\right)\right) = \frac{\sigma^2}{N} + \mu^2.\end{aligned}$$

Finally, we obtain that

$$\begin{aligned}\mathbb{E}[S_X^2] &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[X_i^2\right]+\mathbb{E}\left[\overline{X}^2\right]-2\mathbb{E}\left[X_i\overline{X}\right]\right)\\ &= \frac{N}{N-1}\left(\sigma^2+\mu^2+\frac{\sigma^2}{N}+\mu^2-2\frac{\sigma^2}{N}-2\mu^2\right)\\ &= \frac{N}{N-1}\left(\sigma^2-\frac{\sigma^2}{N}\right)=\frac{N}{N-1}\frac{N-1}{N}\sigma^2=\sigma^2.\end{aligned}$$

- So, we have:

$$\mathbb{E}[\overline{X}]=\mu \qquad \text{and} \qquad \mathbb{E}[S_X^2]=\sigma^2.$$

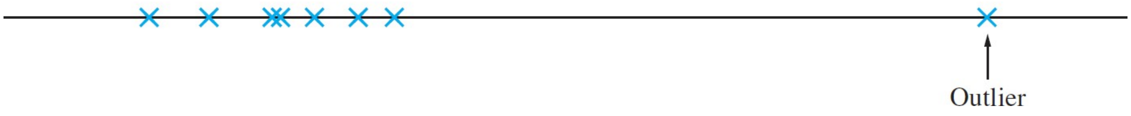
In particular, we have shown

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})^2\right]=\mathbb{E}\left[\frac{N-1}{N}S_X^2\right]=\frac{N-1}{N}\sigma^2$$

which explains why, in the definition of S_X^2, we divide by $N-1$ and not N.

20.4. Outliers

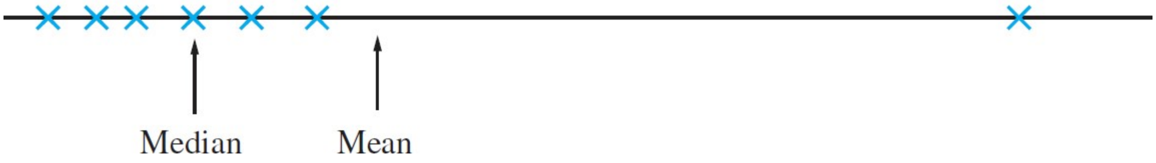
- **Outliers** are points that are much larger or smaller than the rest of the sample points. Outliers may be data entry errors or they may be points that really are different from the rest. Outliers should not be deleted without considerable thought. The best approach is to consider statistical analysis with and without them and then compare if they presence is significant. If so, further investigations is needed.



- Some statistical techniques are sensitive to outliers (that is, their results change significantly if an outlier is removed). Sample mean, sample variance, and sample standard deviations are particularly sensitive to outliers. Next, we learn about other summary statistics that are more resilient to outliers.

20.5. Sample ordered statistics

- The **median**, like the mean, is another measure of center. The sample median is the middle number of the ordered data values. Therefore, it is less sensitive to outliers. Can you explain why?



- To find the Sample Median:
  - Order the N data points from smallest to largest.
  - If N is odd, the sample median is the number in position $\frac{N+1}{2}$.
  - If N is even, the sample median is the average of the numbers in positions $\frac{N}{2}$ and $\frac{N}{2}+1$.
- **Sample quartiles** divide the data as nearly as possible into quarters.

  - The **first sample quartile** is the median of the lower half of the data. To find the first quartile of N observations, compute $\frac{N+1}{4}$. If this is an integer, then the sample value in that position is the first quartile. If not, take the average of the sample values on either side of this value.
  - The **second sample quartile** is simply the sample median.
  - The **third sample quartile** is the median of the upper half of the data. Can you explain how to find it?
- Quantiles (and percentiles) generalize the concept of quartiles. For a fraction $q \in [0, 1]$, the **q-quantile** (or **100q-th percentile**) of a sample divides the sample so that as nearly as possible q fraction of the sample values are less than the q-quantile, and $(1 - q)$ fraction of the values are greater.

Example 20.3.

Recall Example 1.4 from Lecture 1:

Occupants	1	2	3	4	5
Number of Cars	70	15	10	3	2

Find the sample median number of occupants. Compute the first and third quartiles of the number of occupants.

Solution. The sample median is the average of the 50th and 51st value, after arranging the data. Both of these values are equal to 1, so the median is 1.

The first quartile = the average of the 25th and 26th value when arranged in order = 1

The third quartile = the average of the 75th and 76th value when arranged in order = 2