Math 11: Discussion 7

Recap of discussion 7:

1. The power of the Law of Large Number, and the Central Limit Theorem.
2. Differences between having population statistics and inferring sample statistics and vice versa.
3. Building and (equally important) interpreting confidence interval.

Contents

7.1 Upcoming assignments . . . 1
7.2 Statistical Inference . . . 2
7.2.1 Law of Large Number . . . 2
7.2.2 Central Limit Theorem . . . 3
7.3 Confidence Interval . . . 6
7.3.1 Qualitative Data . . . 6
7.3.2 Quantitative Data . . . 6

7.1 Upcoming assignments

Assignments	Chapters	Deadlines
Quiz	Ch. 10	Tue. 05/12
Homework	Ch. 15	Wed. 05/13
Homework	Ch. 16	Wed. 05/13
Quiz	Ch. 15	Thu. 05/14
Quiz	Ch. 16	Thu. 05/14
Homework	Ch. 17	Mon. 05/18
Quiz	Ch. 17	Tue. 05/19
Lab 6		Fri. 05/15

Note: Beware of that there is another set of homework and quiz for Chapter 15.

Key concepts (not exhaustive):

1. *Statistical inference:* Law of Large Number, Central Limit Theorem
2. *Confidence interval:* building and interpreting

7.2 Statistical Inference

Motivation: Suppose we want to describe statistics about a population. However, as usually observed in real life, there is a multitude of reasons that obtaining the population statistics is practically impossible, instead, what we have are samples of the population. Our objective is to infer the population statistics from the sample statistics.

Now, suppose that we have a sample of $\{x_1, x_2, \ldots, x_n\}$, where each of x_i is *iid* (independent and identically distributed) sampled from the population. There are 2 types of statistics:

proportion (qualitative data)	mean (quantitative data)
$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=1}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

where $\mathbb{1}_{x_i=1}$ is an indicator r.v. of if $x_i = 1$ (a success). In other words, we are taking the naive estimates that p is the proportion of successes in our sample, and $\bar{x}$ the average of our sample data.

Importantly, we have the *expected value* and the *variance* of those estimates:

Estimates	*Expectation*	*Variance*	
$\hat{p}$	$\mathbb{E}[\hat{p}] = p$	$Var[\hat{p}] = \frac{p(1-p)}{n}$	$\hat{p} \sim \left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
$\bar{x}$	$\mathbb{E}[\bar{x}] = \mu$	$Var[\bar{x}] = \frac{\sigma^2}{n}$	$\bar{x} \sim \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Table 1: Expectation and Variance of the estimates $\hat{p}$ and $\bar{x}$.

where the last column is a concise expression of the *expectation* and *standard deviation* (note that though they may look similar to the notation for the *Normal* distribution, they are not and do not infer that the estimates are normally distributed).

7.2.1 Law of Large Number

As shown in table 1, the expected values of the estimates are the true (population) statistics. Indeed, as the sample size gets larger, the sample estimates will converge to the true statistics:

$$\lim_{n\to\infty} \hat{p} = p \qquad \text{and} \qquad \lim_{n\to\infty} \bar{x} = \mu$$

7.2.2 Central Limit Theorem

Idea: the sampling distribution of *any mean* becomes more *Normal* as the sample size increases.

The precise statement is:

- Suppose we have a sample of $\{x_1, x_2, \ldots, x_n\}$, where each x_i is *iid*. Regardless of the true distribution of the r.v. X, the sample mean is approximately *Normal* distributed with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \frac{Var[X]}{n}$ as n goes to infinity.

In other words, if we let $S_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, then:

$$S_n \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}\Big(\ \mu, \frac{\sigma}{\sqrt{n}}\ \Big)$$

where $\xrightarrow[n\to\infty]{\mathbb{P}}$ means as n goes to infinity, the distribution of S_n converges to.

Remarks:

1. In determining how *large* n needs to be before CLT applies, in general:
   (a) for *qualitative data*, we want both $np \geq 10$ and $n(1-p) \geq 10$
   (b) for *quantitative data*, we want $n \geq 30$; and larger the more severe the skewness of the data
2. CLT applies to the *mean*, and not any individual observations.

Put very loosely:

1. *Law of Large Number* implies that when the sample size gets large, our naive estimates get closer and closer to the true parameters.
2. *Central Limit Theoem* implies that when the sample size gets large enough, our naive estimates approximately have a normal distribution.

Math 11: Discussion 7

Example 1. *San Diego and California in general are known to have some of the best weather in the country (though this is also up to individual preferences). In particular, some like rain, while others not so much. Let's look at how often it rains in San Diego (just for fun: what you can rain in San Diego is nothing compared what rain is like in tropical countries, think South East Asia).*

Suppose the proportion of rainy days is .12*. Let's look at a sample of* 100 *days and calculate the probability of:*

(a) the proportion of rainy days is at least .15

(b) the total number of rainy days is at least 15

Solution:

(a) We first recognize that this is about proportion. Let $\hat{p}$ be the proportion of rainy days in those 100 days. From the Central Limit Theorem:

$$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{(1-p)p}{n}}\right)$$

Given the true proportion of .12, we have

$$\hat{p} \sim \mathcal{N}(.12, .032)$$

We are now ready to compute the probability of observing the proportion of at least .15:

$$\mathbb{P}(\hat{p} \geq .15) = \mathbb{P}\left(Z \geq \frac{.15 - .12}{.032}\right) = \mathbb{P}(Z \geq .9375) = 1 - \Phi(.9375) = .17$$

(b) We recognize that this is about mean. Let $\bar{x}$ be the total number of rainy days in those 100 days.

Here, we could either use the same argument as in part (a), or we could observe that given the proportion in part (a) and the sample size of 100, then

$$\mathbb{P}(\bar{x} \geq 15) = \mathbb{P}(\hat{p} \geq .15) = .17$$

Example 2. *Do you miss going to downtown San Diego? Have you ever wondered how long a green light at a crossing on* 1^{st} *street is?*

Suppose the length of a green light has a mean of $40s$ *and a standard deviation of* $15s$*. Suppose we take a sample of* 100 *green lights and look at the average of those green lights. Calculate the probability of these events:*

(a) the average is over $42s$

(b) the average is between $39s$ *and* $42s$

Solution:

We observe that this is about mean. Let $\bar{x}$ be the average of those 100 green lights. From the Central Limit Theorem, we know that

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Given our data in the question, we have

$$\bar{x} \sim \mathcal{N}(40, 1.5)$$

(a) the average is over $42s$:

$$\mathbb{P}(\bar{x} \geq 42) = \mathbb{P}\left(Z \geq \frac{42-40}{1.5}\right) = \mathbb{P}(Z \geq 4/3) = 1 - \Phi(4/3) = .09$$

(b) the average is between $39s$ and $42s$

$$\mathbb{P}(39 \leq \bar{x} \leq 42) = \mathbb{P}(\bar{x} \leq 42) - \mathbb{P}(\bar{x} \leq 39) = (1 - .09) - \Phi(-2/3) = .16$$

7.3 Confidence Interval

7.3.1 Qualitative Data

Recall that we estimate p via $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=1}$, it follows that the *standard error* of $\hat{p}$ is $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

In building a confidence interval, we start with some tolerance α (for example $\alpha = .05$), which in turn gives us the critical value z^* (based on the standard normal distribution); then the $100 \cdot (1-\alpha)\%$ CI is:

$$\Big(\hat{p} - z^* \cdot SE(\hat{p}),\ \hat{p} + z^* \cdot SE(\hat{p}) \Big) = \left(\hat{p} - z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Additionally, we define the *margin of error* as $ME = z^* \cdot SE(\hat{p})$ (note that ME is a function of α, or our CI).

Remarks:

1. **Interpretation of CI:**
   (a) Suppose we are working with a sample of size n, and built a 95% CI, this means that if we were to repeat this experiment many times, each of sample size n, then we would expect that 95% of the CI built would contain the true population statistics p.
   (b) In particular, it will be wrong to say that p has a 95% chance of being in our CI since p, the true statistics, is not a r.v. and hence it does not make sense to give a p a distribution.
2. The large the sample size n, the better the CI.

7.3.2 Quantitative Data

Recall that regardless of the distribution of x_i, we always have that

$$\bar{x} \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}\Big(\mu, \frac{\sigma}{\sqrt{n}} \Big) \tag{1}$$

However, (1) is only an approximation. Now, if we impose a condition that $x_i \sim^{iid} \mathcal{N}(\mu, \sigma)$, then

$$\frac{\bar{x} - \mu}{s/\sigma} \ \sim\ T_{n-1}, \qquad \text{where } \ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$$

which has the exact distribution for all n: t-distribution with $(n-1)$ degrees of freedom.

Now, when building a CI, similar to the procedure described in the case of *qualitative data*, we start with some tolerance α, which in turn gives us the critical value t^*_{n-1} (based on the t-distribution with $(n-1)$ d.f.); then the $100 \cdot (1-\alpha)\%$ CI is:

$$100 \cdot (1-\alpha)\%\ \text{CI} = \left(\bar{x} - t^*_{n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t^*_{n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

Similarly, we define the *margin of error* as $ME = t^*_{n-1} \cdot \frac{s}{\sqrt{n}}$ (note that ME is a function of α, or our CI).

Example 3. *We continue from Example* 1. *Suppose that after sampling* 100 *days randomly, we found that the proportion of rainy days was* .15.

(a) Build a 95% *confidence interval of the proportion of rainy days.*

(b) Interpret that interval. Be precise in your wording, ambiguity is considered wrong by default.

(c) We now want to cut the margin of error, ME, in half while keeping the same level of confidence, how many days do we need to sample?

(d) Similarly, how many days do we need to sample if we want ME of at most 1%*?*

Solution: Let $\hat{p}$ be the proportion of rainy days observed in the 100-day sample. We know that $\hat{p} = .15$.

(a) We recall from the Central Limit Theorem that $\hat{p}$ approximately has a normal distribution. As such, the critical value z^* is such that

$$\mathbb{P}(Z \le z^*) = .975 \quad \Longrightarrow \quad z^* = 1.96$$

The 95% confidence interval is thus:

$$\begin{aligned} 95\%\ CI &= \left(\hat{p} \ \pm \ z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \\ &= \left(.15 \ \pm \ 1.96 \cdot \frac{1}{10}\sqrt{(.15)(.85)}\right) = \big(.15 \ \pm \ .07\big) = \big(.08,\ .22\big) \end{aligned}$$

(b) *Interpretation:*

- If we were repeat the experiment many times (taking 100 observations, calculating the proportions, and building the confidence intervals), then we would expect that 95% of those confidence intervals would contain the true proportion p.

(c) We recall from the definition that $ME = z^* SE(\hat{p})$, where $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$. Let ME_1 be the ME associated with 100 samples while ME_2 is the desired new ME. We have

$$\begin{aligned} ME_2 &= \frac{1}{2} ME_1 \\ z^*\sqrt{\frac{\hat{p}(1-\hat{p}}{n}} &= \frac{1}{2} \cdot z^*\sqrt{\frac{\hat{p}(1-\hat{p}}{100}} \\ \frac{.15(.85)}{n} &= \frac{1}{4} \cdot \frac{.15(.85)}{100} \qquad (z^* \text{ is the same given the same confidence level}) \\ n &= 400 \end{aligned}$$

(d) Similar to part (c), assuming the same confidence level of 95% (and hence $z^* = 1.96$), we have

$$ME \le .01 \quad \Longrightarrow \quad 1.96\sqrt{\frac{.15(.85)}{n}} \le .01 \quad \Longrightarrow \quad .01^2 n \ge 1.96^2(.15)(.85) \quad \Longrightarrow \quad n \ge 4898.04$$

Hence, we need to sample at least 4899 days to ensure that ME is at most 1%.

Example 4. *We continue from Example* 2. *Suppose that after sampling* 100 *green lights randomly, we found that the green lights lasted on average* $42s$ *with a standard deviation of* $12s$.

Build a 95% *confidence interval for the length of green light.*

Solution: Let $\bar{x}$ be the average length of green lights observed in the 100-day sample. We know that $\bar{x} = 42$. Furthermore, we know that $s = 12$.

Since we are working with quantitative data, we can use the exact distribution via the t-distribution. Since $n = 100$, $df = 99$. As such the critical value t^* is such that

$$\mathbb{P}(t_{99} \leq t^*) = .975 \quad \implies \quad t^* = 1.98$$

The 95% confidence interval is thus:

$$\begin{aligned} 95\%\ CI &= \left(\bar{x} \ \pm \ t^* \frac{\sigma}{\sqrt{n}}\right) \\ &= \left(42 \ \pm \ 1.98 \cdot \frac{12}{10}\right) = \big(42 \ \pm \ 2.38\big) = \big(39.62, \ 44.38\big) \end{aligned}$$