Math 3070, Applied Statistics

Section 1

October 21, 2019

Lecture Outline, 10/21

Section 5.3

- Central Limit Theorem
- Exercises

Central Limit Theorem

A result of probability theory guarantees that in general, if n is large ($n > 30$) , then the sample mean $\overline{X}$ is approximately normal:

Let $X_1, X_2, \dots$ be independent, identically distributed (iid) random variables with mean μ and standard deviation σ, then

$$\lim_{n\to\infty} P\left(a \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) = P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$$

where Z is a standard normal random variable.
Can also be used to approximate probability of events involving sums of many i.i.d. random variables.

A simple proof is beyond the scope of this course. Need to discuss moment generating functions.

$n > 40$ and $n > 50$ are sometimes used instead. This is an empirical rule.

Central Limit Theorem, Suggested Algorithm

1. Check sample size.
2. Find μ and σ.
3. Write down the event of interest.
4. Construct $(\overline{X} - \mu)/(\sigma/\sqrt{n})$.
5. Approximate using CLT.

Example, Normal Approximation to Binomial

Bob is a candidate for political office in a large city and must take more than half the votes in order to win the election. Suppose a poll finds that 170 of 300 randomly sampled voters favor Bob. Approximate the probability that the poll would find so many voters favoring Bob if the true proportion were only .5.

The number X of sampled voters favoring Bob is a hypergeometric random variable; but since the population is large, we may approximate X as binomial, $Bin(300, .5)$.

The Central Limit Theorem implies that X, a sum of i.i.d. Bernoulli random variables, is approximately normal, with mean $\mu = np = 150$ and standard deviation $\sigma = \sqrt{np(1-p)} = \sqrt{75} \approx 8.66$. $Z \sim N(0,1)$

$$\begin{aligned} P(X \geq 170) &\approx P(8.66Z + 150 \geq 170) \\ &\approx P(Z \geq 2.31) \\ &= P(Z \leq -2.31) = \Phi(-2.31) \approx .0104 \end{aligned}$$

Example, Normal Approximation to Poisson

A Geiger counter placed next to a certain radioactive specimen clicks an average of 90 times per minute. Over a 10 minute period, approximate the probability that it would click 800 times or less.

The number of clicks X in a 10 minute period has a Poisson distribution with mean $\mu = 900$. The variance of X is $V(X) = \mu = 900$, so the standard deviation is $\sigma = \sqrt{900} = 30$.

Since μ is large, the distribution of X is approximately normal. So we can approximate X as $30Z + 900$ where Z is a standard normal random variable.

$$\begin{aligned} P(X \leq 800) &\approx P(30Z + 900 \leq 800) \\ &\approx P(Z \leq -3.33) = \Phi(-3.33) \approx .0004 \end{aligned}$$

We can also think of X as a sum of 600 poisson random variables, sum of clicks in second. These have a mean of 9/6 ticks per minute. CLT gives the same result.

CLT Example 1

On any workday, the number of minutes T_i that a diligent employee arrives early to work is distributed as an exponential random variable with an expected value of 6 minutes. Assuming there are 250 work days in a year, approximate the probability that employee spends an extra full day at work.

$$T_i \sim exp(1/6) \qquad \mu = E[T_i] = 6, \;\; \sigma^2 = Var(T_i) = 36$$

Since $n = 250$ is large, the Central Limit Theorem applies,

$$\overline{T} \sim N(6, 36/250)$$

Note: this notation uses variance as the second parameter.

$$\text{extra time} = \sum_{i=1}^{250} T_i = 250\overline{T}$$

CLT Example 1

$$
\begin{aligned}
P(250\overline{T} > 24 * 60) = P(250\overline{T} > 1440) &= P\left(\overline{T} > \frac{1440}{250}\right) \\
&= P\left(\frac{\overline{T} - 6}{6/\sqrt{250}} > \frac{\frac{1440}{250} - 6}{6/\sqrt{250}}\right) \\
&\approx P(Z > -0.632455532) \\
&= P(Z \leq 0.632455532) \approx 0.7365
\end{aligned}
$$

$Z \sim N(0,1)$. Recall that the first $\approx$ is needed since the Central Limit Theorem is true when $n \to \infty$, but $n = 250$.

Tip: Only used the exponential distribution to extract a mean and variance. Same idea might apply to other distributions in CLT problems. A large sample size ($n > 30$) is the indication that CLT may be used.

CLT example 2

Suppose that one in a thousand have a rare illness. Estimate the probability that proportion of a simple random sample of 1000 is at least twice the population proprtion.

Consider bernoulli random variables X_i that are one when a person has the illness. In the simple random sample, $X_i \sim Bern(0.001)$.

$$E[X_i] = 0.001 \quad Var(X_i) = (0.001)(1 - 0.001) = 0.000999$$

$$\begin{aligned} P(\overline{X} > 0.002) &= P\left(\frac{\overline{X} - 0.001}{\sqrt{0.000999}\sqrt{1000}} > \frac{0.002 - 0.001}{\sqrt{0.000999}\sqrt{1000}} \right) \\ &\approx P(Z > 1.00050038) \\ &\approx 0.1587 \end{aligned}$$

CLT example 3

From past experience, it is known that the number of tickets purchased by a student standing in line at the ticket window for the football match of UU against BYU is at least one plus a number which follows a Poisson distribution with a mean equal to 0.2. Suppose that few hours before the start of one of these matches there are 100 eager students standing in line to purchase tickets. If only 121 tickets remain, what is the probability that all 100 students will be able to purchase the tickets they desire?

$$\text{tickets purchased by student } i = X_i + 1$$

where $X_i \sim Poi(1)$.

$$E[X_i + 1] = E[X_i] + 1 = 0.2 + 1 = 1.2$$

$$Var(X_i + 1) = Var(X_i) = 0.2$$

CLT Example 3

$$
\begin{aligned}
P\left(\sum_{i=1}^{100} X_i \leq 150\right) &= P\left(\frac{\sum_{i=1}^{100} X_i}{100} < 1.21\right) \\
&= P(\overline{X} < 1.21) \\
&= P\left(\frac{\overline{X} - 1.2}{0.2/\sqrt{100}} < \frac{1.21 - 1.2}{0.2/\sqrt{100}}\right) \\
&\approx P(Z < 0.223606798) \approx 0.5885
\end{aligned}
$$

$Z \sim N(0, 1)$

CLT Example 4

The duration of your daily commute to work W_i averages 25 minutes and the duration of the commute home H_i averages 35 minutes. You want to find a reasonable commute time to report on expense reports. Assuming that the commute times are exponential distributed and a 250 work days per year, estimate the value which at which your yearly commute times has a 75% of being below.

$$\text{Total commute time in a year } = T = \sum_{i=1}^{250} [H_i + W_i]$$

$$E[H_i + W_i] = E[H_i] + E[W_i] = 25 + 35 = 60$$

$$Var(H_i + W_i) = 25^5 + 35^2 = 1850$$

Want to find c so that

$$P(T < c) = 0.75$$

CLT Example 4

$$\begin{aligned} P(T < c) &= P\left(\overline{T} < \frac{c}{250}\right) \\ &= P\left(\frac{\overline{T} - 60}{\sqrt{1850}/\sqrt{250}} < \frac{\frac{c}{250} - 60}{\sqrt{1850}/\sqrt{250}}\right) \\ &\approx P\left(Z < \frac{\frac{c}{250} - 60}{\sqrt{1850}/\sqrt{250}}\right) \end{aligned}$$

$$\overline{T} = T/250, Z \sim N(0,1)$$

We see that $\Phi^{-1}(0.75) \approx 0.6745$ or $P(Z < 0.6745) \approx 0.75$ and use as to approximate c.

$$\begin{aligned} \frac{\frac{c}{250} - 60}{\sqrt{1850}/\sqrt{250}} &\approx 0.6745 \\ \Rightarrow c &\approx 250(0.6745\sqrt{1850}/\sqrt{250} + 60) \\ &\approx 15480 \text{ minutes or } 258 \text{ hours} \end{aligned}$$