Continuous Random Variables Lecture 4

Central limit theorem

Normal random variables are widely used in light of the following theorem.

Thm (The Central Limit theorem): Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then for n large, the distribution $X_1 + \ldots + X_n$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

Observe that $E[X_1 + \ldots + X_n] = E[X_1] + \ldots + E[X_n]$

$= \mu + \ldots + \mu = n\mu$.

Also by independence, $Var(X_1 + \ldots + X_n) = Var(X_1) + \ldots + Var(X_n) = \sigma^2 + \ldots + \sigma^2 = n\sigma^2$

Thus, calculating $P(X_1 + \ldots + X_n < b)$ amounts to

$$P(X_1 + \ldots + X_n < b) = P\left(\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\,\sigma} < \frac{b - n\mu}{\sqrt{n}\,\sigma}\right)$$

$$\approx P\left(Z < \frac{b - n\mu}{\sqrt{n}\,\sigma}\right) = \Phi\left(\frac{b - n\mu}{\sqrt{n}\,\sigma}\right)$$

Ex. Insurance company has 25,000 automobile policy holders. If yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

Solution: Total sum of claims $T = X_1 + \ldots + X_{25,000}$

$$P(T > 8.3 \times 10^6) = P\left(\frac{T - 25,000 \cdot 320}{540 \cdot \sqrt{25,000}} > \frac{8.3 \times 10^6 - 25,000 \cdot 320}{540 \cdot \sqrt{25,000}}\right)$$

$$\approx P\left(Z > \frac{3 \times 10^5}{85381.5}\right) \approx P(Z > 3.51) = 1 - \varphi(3.51)$$

$$\approx 1 - 0.9990 \approx 0.001.$$

Remark: It can be shown that the sum of two independent normally distributed random variables X and Y with means μ_x and μ_y and standard deviations σ_x and σ_y is a normal random variable with mean $\mu = \mu_x + \mu_y$ and standard deviation $\sqrt{\sigma_x^2 + \sigma_y^2}$.

Ex. The amount of weight in units of 1000 pounds that a particular bridge can withstand is a normal random variable W with mean 400 and standard deviation 40. If the weight of a car is a random variable with mean 3 and standard deviation 0.3, how many cars would have to be on the bridge so that the probability of structural damage would exceed 0.1?

Solution: Suppose there are n cars on the bridge. The total weight of these cars is $T = X_1 + \ldots + X_n$

Damage occurs if $T > W$.

$P(T > W) = P(T - W > 0)$ where $E[T-W] = 3n - 400$

and $Var(T-W) = Var(T) + Var(W) = n\,Var(X_1) + 40^2$

$= (0.3)^2 n + 40^2 = 0.09n + 1{,}600$

$$P(T - W > 0) = P\left(\frac{T - W - (3n-400)}{\sqrt{0.09n + 1600}} > \frac{-3n + 400}{\sqrt{0.09n + 1600}}\right)$$

$$\approx P\left(Z > \frac{400 - 3n}{\sqrt{0.09n + 1600}}\right) = 1 - \Phi\left(\frac{400-3n}{\sqrt{0.09n + 1600}}\right)$$

We want to find a Z score z such that

$1 - \Phi(z) \geq 0.1$ or $1 - 0.1 = 0.9 \geq \Phi(z)$

$\Longrightarrow$ the score we're looking for is $\Phi^{-1}(0.9)$

The score we're looking for is $z = 1.28$.
Hence the number of cars n must be large enough to insure
$$\frac{400 - 3n}{\sqrt{0.09n + 1600}} \leq 1.28$$
or $n \geq 117$.

Ex. A coin is tossed 1000 times.
(a) How likely are we to get 550 or more outcomes where the coin landed on heads, given that it is fair.
(b) If the coin lands 550 times on heads, is the evidence enough to conclude that the coin isn't fair?

Solution:
(a) Let $X = X_1 + \ldots + X_{1000}$ be the number of heads in 1000 tosses. $E[X] = 1000 \cdot E[X_1] = 1000 \cdot \frac{1}{2} = 500$

$Var(X) = Var(X_1) + \ldots + Var(X_{1000}) = 1000\, Var(X_1) = 1000 \cdot \frac{1}{4}$
$= 250$

$$P(X \geq 550) = P\left(\frac{X - 500}{\sqrt{250}} \geq \frac{550 - 500}{\sqrt{250}}\right) \approx P\left(Z \geq \frac{50}{5\sqrt{10}}\right)$$
$$\approx P(Z \geq \sqrt{10}) \approx P(Z \geq 3.16) = 1 - \varphi(3.16) \approx 1 - 0.9992$$
$$= 0.0008 = \frac{8}{10{,}000}$$

(b) From (a) we have $P(X \geq 550 \mid \text{Coin is fair}) \approx \frac{8}{10{,}000}$ which is rather unlikely. It is tempting to reject the hypothesis that the coin is fair. However, first consider the following example

<u>Ex.</u> A teacher working in "Children are My Sunshine" inner city school, but the 101st pupil is Dennis the Menace. After a long work day, the teacher sits to rest in her favorite chair, only to discover much too late that it is littered with board pins.

Furious, the teacher goes at once to see the principle and demans that Dennis be immediately expelled. She is armed with the following indisputable evidence:

(i) The prank was perpetrated by one and only one of her 101 students.

(ii) If Dennis were innocent, the probability of such prank would be 0.001.

If it is the school's policy to regard a pupil guilty of a misdemeanor if the probability of innocence is below 0.005, what should the principle do? Find Dennis guilty of the misdemeanor or dismiss the case against him for lack of evidence? Explain.

Solution: The principle looks into his directory on juvenile delinquents and finds that out of 9001 young sociopaths (on the average) only 1 will place board pins as a prank. It is also estimated that a good pupil is likely to commit this prank with probability 10^{-5}. Then letting I be the event that Dennis is innocent and E be the event this crime is commited.

$$P(I) = \frac{9000}{9001}, \quad P(E|I) = 10^{-3}, \quad P(E|I^c) = 1.$$

where $P(E|I)$ = probability that one of the 100 good pupils commited the prank.

Thus $$P(I|E) = \frac{P(E|I)\,P(I)}{P(E|I)\,P(I) + P(E|I^c)\,P(I^c)}$$

$$= \frac{10^{-3}\,\frac{9000}{9001}}{10^{-3}\,\frac{9000}{9001} + 1\cdot\frac{1}{9001}} = 0.9$$

It is therefore 90% likely that Dennis is innocent! This example illustrates the prosecutor paradox.

Here is an actual case:

Ex. In 1998, Sally Clark was tried for murder after two of her sons died shortly after birth. During the trial, an expert witness for the prosecution testified that the probability of a newborn dying of sudden infant death syndrome (SIDS) was $\frac{1}{8500}$, so the probability of two deaths due to SIDS in one family was $\left(\frac{1}{8500}\right)^2$, or about one in 73 million. Therefore, he continued, the probability of Clark's innocence was one in 73 million.

Based on this evidence, would you have found the defendant guilty? As it happens Clark was convicted of murder and sent to prison.

Solution:

(i) The death of the two children isn't necessarily independent if it is caused by SIDS. The death of one sibling might make the death of the other very likely.

(ii) The expert witness speaks about the likelyhood of evidence under the assumption of innocence $P(E|I)$. What is needed is $P(I|E) =$

$$= \frac{P(E|I)\,P(I)}{P(E|I)\,P(I) + P(E|I^c)\,P(I^c)}$$

As seen in the last example, the latter probability can be fairly heigh. That would happen if probability of innocence is very heigh.

Ex. Ideal size of a first-year class at a particular college is 150 students. On average 30% of those accepted for admission will actually attend. College approves the applications of 450 students. What is the probability that more than 150 first-year students attend this college?

Solution: Set $X = X_1 + \ldots + X_{450}$;

$$X_k = \begin{cases} 1 & \text{if } k^{th} \text{ student accepts} \\ 0 & \text{otherwise.} \end{cases}$$

$$P(X \geqslant 151) = P\left(\frac{X - 450 \cdot 0.3}{\sqrt{450 \cdot 0.3 \cdot 0.7}} \geqslant \frac{151 - 450 \cdot 0.3}{\sqrt{450 \cdot 0.3 \cdot 0.7}} \right)$$

$\approx P(Z \geqslant 1.65) \approx 0.05$. Hence about 5%

Ex. Due to the atmospheric disturbances, the exact distance to a star, d, cannot be determined exactly. As a result, the astronomer has decided to make a series of measurements and then use their average to estimate the actual distance. If the values of the successive measurements are independent random variables with mean

d of light years and standard deviation of 2 light years, how many measurements need to be made to be at least 95% certain that the estimate is accurate to within ± 0.5 light years?

Solution: Let $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$.

$$E[\bar{X}] = \frac{1}{n}\left(E[X_1] + \ldots + E[X_n]\right) = d.$$

$$Var(\bar{X}) = \frac{1}{n^2} Var(X_1 + \ldots + X_n) = \frac{1}{n^2}\, n\, Var(X_1) = \frac{2^2}{n}$$

Hence $P(d - 0.5 < \bar{X} < d + 0.5) =$

$$= P\left(\frac{-0.5}{\sqrt{\frac{2^2}{n}}} < \frac{\bar{X} - d}{\sqrt{\frac{2^2}{n}}} < \frac{0.5}{\sqrt{\frac{2^2}{n}}}\right)$$

$$\approx P\left(\frac{-0.5}{2}\sqrt{n} < Z < \frac{0.5}{2}\sqrt{n}\right) =$$

$$= \varphi\left(\frac{0.5}{2}\sqrt{n}\right) - \varphi\left(\frac{-0.5}{2}\sqrt{n}\right) = 2\varphi\left(\frac{\sqrt{n}}{4}\right) - 1$$

$$\geq 0.95. \implies \varphi\left(\frac{\sqrt{n}}{4}\right) \geq \frac{1.95}{2} = 0.975$$

Since $\varphi(1.96) \approx 0.975$ we have $\frac{\sqrt{n}}{4} \geq 1.96$

or $n \geq 62$ observations/measurements are necessary.