Lecture 11

Part 2 Probability and Distributions

Sampling Distribution Part 2

Example 2

- Given a normal distribution with unknown mean μ and variance equal to 100, find the probability that the sample mean of a sample of 25 observations will differ from the population mean by less than 4 units.
- Let $\bar{X}$ denote the sample mean of the 25 observations.
- From the CLT we know that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} = \frac{100}{25}\right)$$

Solution

- We only care that the distance between the sample mean and population mean is less than 4 units.
- So we want to find $P(|\bar{X} - \mu| < 4)$:

$$
\begin{aligned}
P(|\bar{X} - \mu| < 4) &= P(-4 < \bar{X} - \mu < 4) \\
&= P\left(\frac{-4}{\frac{10}{\sqrt{25}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{4}{\frac{10}{\sqrt{25}}}\right) \\
&= P(-2 < Z < 2) \\
&= 0.9544
\end{aligned}
$$

Example 3

- The number of accidents per week at a hazardous intersection varies randomly with mean 2.2 and standard deviation 1.4.
- This distribution is discrete and certainly not normal.
- What is the approximate probability that there are fewer than 100 accidents at the intersection in a year?

Solution

- Let X_i be the number of accidents that occur in the ith week of the year.
- We know $E(X_i) = 2.2$ and $V(X_i) = 1.4^2$ for all i.
- We want to find the probability of fewer than 100 accidents in the year, i.e.:

$$P\left(\sum_{i=1}^{52} X_i < 100\right) = P\left(\frac{\sum_{i=1}^{52} X_i}{52} < \frac{100}{52}\right)$$

$$= P(\bar{X} < 1.923)$$

Solution

- But we know from the CLT that

$$\bar{X} \sim N\left(\mu = 2.2, \frac{\sigma^2}{n} = \frac{1.4^2}{52}\right)$$

- Therefore,

$$P(\bar{X} < 1.923) = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{1.923 - 2.2}{\frac{1.4}{\sqrt{52}}}\right)$$
$$= P(Z < -1.43)$$
$$= 0.0764$$

Binomial Distribution

- Recall the binomial distribution that was introduced earlier.
- If $X \sim Bin(n, p)$, then X was counting the number of successes in n independent Bernoulli trials.
- We assumed that p was known.

Sample Proportion

- But in reality, p could be an unknown population parameter.
- Therefore, just like we did with the population mean, we need to use a sample to estimate the population proportion p.
- For example, suppose we are interested in whether Coke or Pepsi is the more popular soft drink.
- Let p denote the population proportion of people who prefer Coke over Pepsi.

Sample Proportion

- If X denotes the number of people who prefer Coke over Pepsi in a randomly selected sample, what is a reasonable estimate of p?
- If n is the size of the sample, then a reasonable estimate of p is the **sample proportion** of people who prefer Coke over Pepsi:

$$\hat{p} = \frac{X}{n}$$

- Let's investigate $\hat{p}$ in a little more detail...

Sample Proportion

- Recall that we can also write $X = \sum_{i=1}^{n} X_i$, where

$$X_i = \begin{cases} 1 & \text{person } i \text{ prefers Coke over Pepsi} \\ 0 & \text{otherwise} \end{cases}$$

- So $\hat{p}$ is just the sample mean of a sample of independent Bernoulli random variables:

$$\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sample Proportion

- Which means we can apply the CLT to find the sampling distribution of $\hat{p}$.
- And from last lecture we know that the mean and variance of $\hat{p}$ are equal to:

$$\mu_{\hat{p}} = E(\hat{p}) = \mu$$

$$\sigma^2_{\hat{p}} = V(\hat{p}) = \frac{\sigma^2}{n}$$

where μ and σ^2 are the mean and variance, respectively, of the Bernoulli population (X_i).

Probability Distribution of X_i

x_i	0	1
$p(x_i)$	$1-p$	p

$$\begin{aligned}\mu &= E(X_i)\\ &= 0\times(1-p)+1\times p\\ &= p\end{aligned}$$

$$\begin{aligned}\sigma^2 &= V(X_i)\\ &= E(X_i^2)-(E(X_i))^2\\ &= (0^2\times(1-p)+1^2\times p)-p^2\\ &= p(1-p)\end{aligned}$$

Central Limit Theorem

- The CLT tells us that for sufficiently large n (i.e., when both np and $n(1-p)$ are ≥ 5):

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}\right)$$

- If we standardise $\hat{p}$, we then get:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = Z \sim N(0, 1)$$

Example 1

A psychologist believes that 80% of male drivers when lost continue to drive hoping to find the location they seek rather than ask directions. To examine this belief, he took a random sample of 350 male drivers and asked each what they did when lost.
If the belief is true, determine the probability that less than 75% said they continue driving.

Solution

- Let $\hat{p}$ be the sample proportion of drivers that keep driving.
- We know that $n = 350$ and the population proportion is $p = 0.8$.
- We want to find $P(\hat{p} < 0.75)$.
- Since $n = 350$ is very large, we can apply the CLT and conclude that:

$$\hat{p} \sim N\left(p = 0.8, \frac{p(1-p)}{n} = \frac{0.8(1-0.8)}{350}\right)$$

Solution

- Therefore:

$$P(\hat{p} < 0.75) = P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0.75 - 0.8}{\sqrt{\frac{0.8(1-0.8)}{350}}}\right)$$

$$= P(Z < -2.34)$$

$$= 0.0096$$

Example 2

An accounting professor claims that no more than one-quarter of undergraduate business students will major in accounting. What is the probability that in a random sample of 1200 undergraduate business students, 336 or more will major in accounting?

Solution

- Let X be the number of students that major in accounting out of the random sample of 1200.
- Then $X \sim Bin(n = 1200, p = 0.25)$.
- We want:

$$P(X \geq 336) = P(X = 336) + P(X = 337) + \dots$$
$$\dots + P(X = 1200)$$

Solution

- We can use the CLT to approximate this probability:

$$
\begin{aligned}
P(X \geq 336) &= P\left(\frac{X}{1200} \geq \frac{336}{1200}\right) \\
&\approx P(\hat{p} > 0.28) \\
&= P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0.28 - 0.25}{\sqrt{\frac{0.25(1-0.25)}{1200}}}\right) \\
&= P(Z > 2.40) \\
&= 1 - 0.9918 \\
&= 0.0082
\end{aligned}
$$