CH5: Statistics and their distributions

Statistics

- A statistic is any quantity whose value can be calculated from sample data.
- A statistic can be thought of as a random variable.

Sampling Distribution

- Any statistic, being a random variable, has a probability distribution.
- The probability distribution of a statistic is sometimes referred to as its sampling distribution.

Note

- In this group of notes we will look at examples where we know the population and it’s parameters.
- This is to give us insight into how to proceed when we have large populations with unknown parameters (which is the more typical scenario).
The “Meta-Experiment”

- The “Meta-Experiment” consists of indefinitely many repetitions of the same experiment.
- If the experiment is taking a sample of 100 items from a population, the meta-experiment is to repeatedly take samples of 100 items from the population.
- This is a theoretical construct to help us understand the probabilities involved in our experiment.

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then

1. $E(\overline{X}) = \mu_{\overline{X}} = \mu$
2. $V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2 / n$

$SD(\overline{X}) = \sigma_{\overline{X}} = \sigma / \sqrt{n}$

Example: Random Rectangles

100 Rectangles with $\mu=7.42$ and $\sigma=5.26$. 

Histogram of Areas
So, the distribution of the sample mean based on samples of size 5, should have

1. \( E(X) = 7.42 \)
2. \( SD(X) = 5.26 / \sqrt{5} = 2.35 \)

Based on 68 random samples of size 5:
Mean of the sample means=7.33
SD of the sample means=1.88.

Normal Distributions

- Let \( X_1, X_2, \ldots, X_n \) be a random sample from a normal distribution with mean value \( \mu \) and standard deviation \( \sigma \).
- Then for any \( n \), \( \overline{X} \) is normally distributed (with mean \( \mu \) and standard deviation \( \sigma / \sqrt{n} \)).

Example: Women’s Heights

- It is known that women’s heights are normally distributed with population mean 64.5 inches and population standard deviation 2.5 inches.
- We will look at the distribution of sample means for various sample sizes.
- Since the population follows a normal distribution, the sampling distribution of \( \overline{X} \) is also normal regardless of sample size.
For \( n=9 \), the sample means will be normally distributed with mean=64.5 and standard deviation= \( 2.5 / \sqrt{9} = 0.83 \).

Distribution of Sample Means (n=9)

For \( n=25 \), the sample means will be normally distributed with mean=64.5 and standard deviation= \( 2.5 / \sqrt{25} = 0.5 \).

Distribution of Sample Means (n=25)

For \( n=100 \), the sample means will be normally distributed with mean=64.5 and standard deviation= \( 2.5 / \sqrt{100} = 0.25 \).

Distribution of Sample Means (n=100)

The Central Limit Theorem

- Let \( X_1, X_2, \ldots, X_n \) be a random sample from a distribution with mean value \( \mu \) and standard deviation \( \sigma \).
- Then if \( n \) is sufficiently large, \( \bar{X} \) has approximately a normal distribution with 
  \( E(\bar{X}) = \mu \) and \( V(\bar{X}) = \sigma^2 / n \).
Rule of Thumb

- If $n > 30$, the Central Limit Theorem can be used.

- For highly skewed data or data with extreme outliers, it may take sample of 40+ before the CLT starts “working”.

Sampling Distribution Simulation


Dependence on Sample Size

The sampling distribution of $\bar{X}$ depends on the sample size ($n$) in two ways:
1. The standard deviation is $\sigma_{\bar{X}} = \sigma / \sqrt{n}$, which is inversely proportional to $\sqrt{n}$.
2. If the population distribution is not normal, then the shape of the sampling distribution of $\bar{X}$ depends on $n$, being more nearly normal for larger $n$.

Why is the Central Limit Theorem Important?

- Every different type of population needs a different set of procedures (i.e. probability tables) to answer questions about probability.
- The CLT shows us that we can use the same procedure (normal probability procedures) for questions about probability and the sample mean, regardless of the shape of the original population.
- The only requirement is a “large” sample.
Proportions as Means

- Recall that a binomial RV $X$ is the number of successes in an experiment consisting of $n$ independent success/failure trials.
- Let $X_i = \begin{cases} 1 & \text{if the } i\text{th trial results in a success} \\ 0 & \text{if the } i\text{th trial results in a failure} \end{cases}$
- Then the sample proportion can be expressed as
  $$\hat{p} = \frac{\text{# successes}}{n} = \frac{\sum X_i}{n} = \bar{X}$$

Distribution of Sample Proportion

- The CLT implies that if $n$ is sufficiently large, then $\hat{p}$ has approximately a normal distribution with
  $$E(\hat{p}) = p \quad \text{and} \quad V(\hat{p}) = \frac{p(1-p)}{n}$$
  where $p$ is the true population proportion.
- In order for the approximation to hold we need $np \geq 10$ and $n(1-p) \geq 10$.

Example: Suppose we took samples of size 20 from a population where $p = 0.5$.

The sampling distribution of $\hat{p}$ is approximately Normal with mean $= 0.5$ and variance $= \frac{p(1-p)}{n} = \frac{0.5 \cdot 0.5}{20} = 0.0125$