Central Limit Theorem

- Before we can begin computing confidence intervals, we must introduce the ***Central Limit Theorem***.
- Suppose random sample of size n are drawn from any distribution, with the distribution having a mean of μ (equivalently $E(X)$) and variance of σ^2 (i.e. standard deviation of σ).
- Also suppose that the sample size is large (i.e. $n > 30$).
- The sample means tend to form a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- We call the standard deviation of the sample means the ***standard error***
- Standard error is commonly denoted as $S.E.$

Central Limit Theorem

- Recall from earlier lectures, an experiment was carried out where the sum of 100 throws of a die were recorded.
- The underlying distribution of the die values is not normally distributed. (Actually discrete uniform between 1 and 6.)
- Nonetheless the distribution of the sum of 100 throws was normally distributed. Necessarily the distribution of the average score for 100 throws is normally distributed.

Distribution of means

Exercise

From previous lecture, we know the following properties of the dice distribution.
(Remark: In this case we know the variance, but that is not always the case.)

- Mean (Expected Value) $E(X) = \mu = 3.5$
- Variance $V(X) = \sigma^2 = 2.9166$
- Standard deviation $= \sigma = 1.707$

Compute the standard error $S.E.(\bar{x})$ for the mean value $\bar{x}$ of die values:

- when the die is thrown 25 times
- when the die is thrown 225 times.

Exercise

- When the die is thrown 25 times n = 25
- Therefore the standard error is

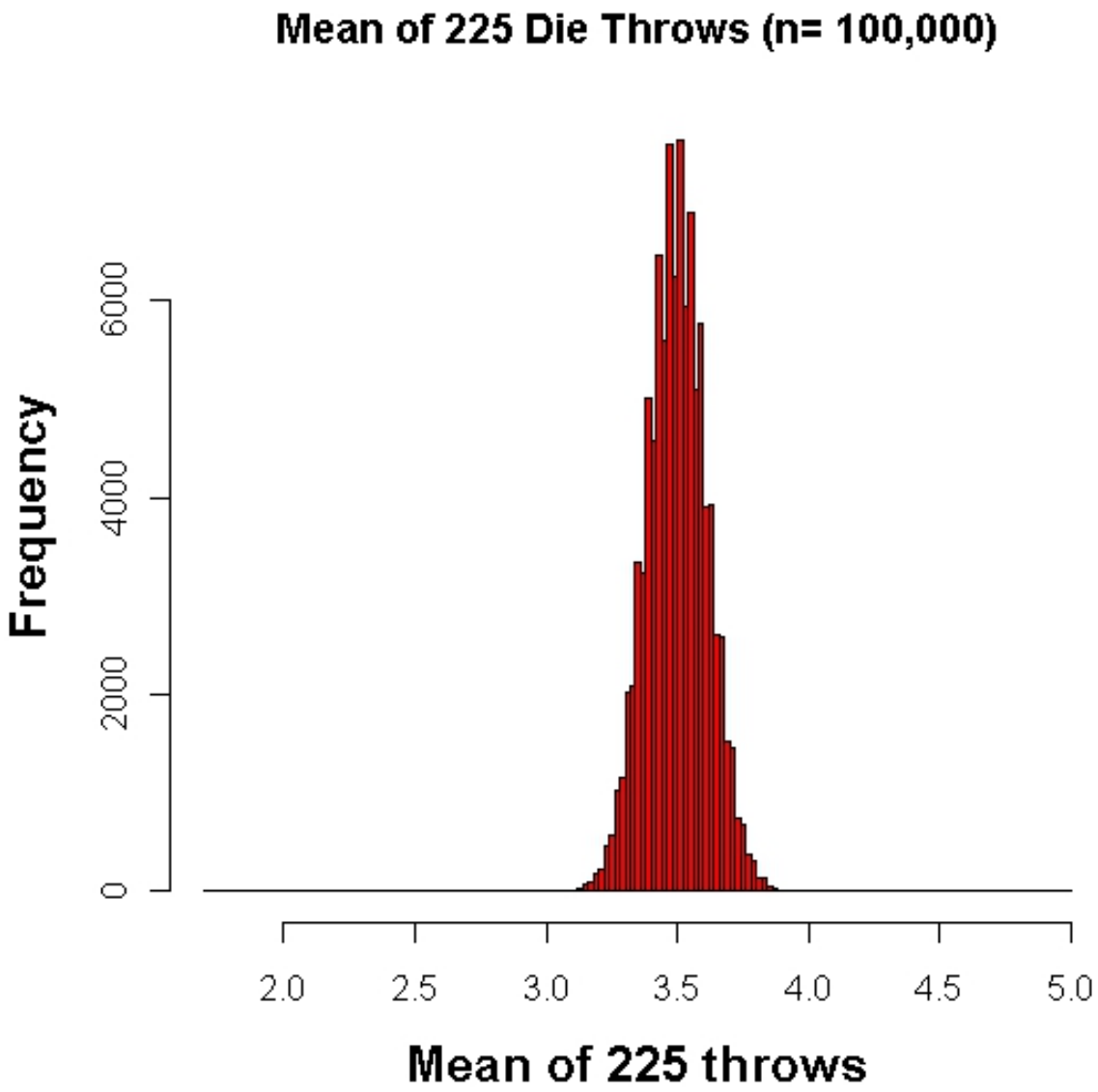
$$\frac{\sigma}{\sqrt{n}} = \frac{1.707}{\sqrt{25}} = \frac{1.707}{5} = 0.3415.$$

- When the die is thrown 225 times: n = 225
- Therefore the standard error is

$$\frac{\sigma}{\sqrt{n}} = \frac{1.707}{\sqrt{225}} = \frac{1.707}{15} = 0.1138.$$

Distribution of means

Distribution of means



Exercise

- Compare the two histograms on the previous slides. These horizontal range of value is the same for both histograms.
- We can see that with a larger sample size ($n = 225$), the distribution of sample means are clustered closely around the 3.5 mark, and have much less dispersion than distribution of sample means with a sample size $n = 25$.

Confidence Intervals (Revision)

- The 95% confidence interval is a range of values which contain the true population parameter (i.e. mean, proportion etc) with a probability of 95%.
- We can expect that a 95% confidence interval will not include the true parameter values 5% of the time.
- A confidence level of 95% is commonly used for computing confidence interval, but we could also have confidence levels of 90%,99% and 99.9%.

Confidence Level

- A confidence level for an interval is denoted to $1-\alpha$ (in percentages: $100(1-\alpha)\%$) for some value α.
- A confidence level of 95% corresponds to $\alpha = 0.05$.
- $100(1-\alpha)\% = 100(1-0.05)\% = 100(0.95)\% = 95\%$
- For a confidence level of 99%, $\alpha = 0.01$.
- Knowing the correct value for α is important when determining quantiles.

The Central Limit Theorem

- This theorem states that as sample size n is increased, the sampling distribution of the mean (and for other sample statistics as well) approaches the normal distribution in form, regardless of the form of the population distribution from which the sample was taken.
- For practical purposes, the sampling distribution of the mean can be assumed to be approximately normally distributed, even for the most non-normal populations or processes, whenever the sample size is $n > 30$.
- (For populations that are only somewhat non-normal, even a smaller sample size will suffice. A variation of the normal distribution can be used for such circumstances.)

Computing Confidence Intervals

Confidence limits are the lower and upper boundaries / values of a confidence interval, that is, the values which define the range of a confidence interval. The general structure of a confidence interval is as follows:

$$\text{Point Estimate} \pm [\text{Quantile} \times \text{Standard Error}]$$

- Point Estimate: estimate for population parameter of interest, i.e. sample mean, sample proportion.
- Quantile: a value from a probability distribution that scales the intervals according to the specified confidence level.
- Standard Error: measures the dispersion of the sampling distribution for a given sample size n.

Point Estimates (1)

- Point estimates are generally straightforward calculations.
- Sometimes they will even be stated directly in the questions.
- When considering the population mean μ, the appropriate point estimate is the sample mean $\bar{x}$.
- When considering the population proportion π, the appropriate point estimate is the sample proportion $\hat{p}$.

Point Estimates (2)

Sample percentage

$$\hat{p} = \frac{x}{n} \times 100\%$$

- $\hat{p}$ - sample proportion.
- x - number of "successes".
- n - the sample size.

Point Estimates (3)

Of a sample of 160 computer programmers, 56 reported than Python was their primary programming language.
Let π be the proportion of all programmers who regard Python as their programming language. What is the point estimate for π?

$$\hat{p} = \frac{x}{n} \times 100\% = \frac{56}{160} = 35\%$$