Quiz Week 6 - CLT and confidence intervals - Solutions

FIT5197 teaching team

Question 1

The relationship between the central limit theorem and confidence intervals for the true mean of the sample mean distribution. (These means of means are making me confused, I don't know what it means.

(a) Let $X_1, \ldots, X_n$ be i.i.d. Random Variables (RVs) with $\mathbb{E}[X_i] = \mu, \mathbb{V}[X_i] = \sigma^2$. From the sample mean version of the Central Limit Theorem (CLT) we know that as $n \to \infty$

$$\overline{X} \xrightarrow{d} \mathcal{N}\left(\mu, \sigma^2/n\right)$$

where $\bar{X}$ is the RV for the sample mean. Also note that when dealing with normal distributions (as above) we transform the normal distribution we are working with into the standard normal distribution in order to compute probabilities. I.e. in this case we define

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), z \in \mathbb{R}$$

Assuming the situation above with large n so the CLT can be applied, show why the two-sided confidence interval for the true population parameter for the sample mean, μ, provides bounds on this parameter that guarantees this parameter falls within these bounds with probability $1 - \alpha$. Moreover, specifically show that when we draw a specific sample $\mathbf{x} = \mathbf{x_1}, \ldots, \mathbf{x_n}$ from the RVs X_i and obtain an observed sample mean $\bar{x}$, that the two-sided confidence interval can be expressed as:

$$\mathrm{CI}_\mu(1-\alpha) = \left(\bar{x} - \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma, \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma\right).$$

Hint: recall that

$$\mathbb{P}\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) \approx 1 - \alpha.$$

where $z_{\alpha/2}$ is the value of z that satisfies this probability formula.

(b) Practical example of the above: Consider the following sample of weights of individuals,

$$\mathbf{w} = \{61, 55, 57, 70, 59, 65, 66, 58\}$$

and given that the variance of weight for an individual is known to be $\sigma^2 = 16$. Find the two-sided 95% confidence interval for the mean of the sample mean distribution for weight. Hint: You will probably need the z-table in the Formula Sheet (https://lms.monash.edu/mod/resource/view.php?id=7439150) for the unit.

(c) Why for the love of the universe do we have to calculate confidence intervals for the mean of a population?

Answer 1

(a) Confidence intervals are formed from an underlying probability interval for a random variable. In the present case, if σ is treated as known, and if n is large enough to justify the required distributional approximation, then you have the random variable:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

This result comes from application of the central limit theorem (CLT), assuming that the underlying distribution meets the requirements of the theorem (e.g., finite variance) and a sufficiently large value of n. Using this random variable and the hint above one can obtain the following probability interval:

$$\mathbb{P}\left(-z_{\alpha/2} \leqslant \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqslant z_{\alpha/2} \right) \approx 1 - \alpha$$

(Note that the value $z_{\alpha/2}$ is the critical value of the standard normal distribution having an uppertail probability of $\alpha/2.$) Re-arranging the inequalities inside the probability statement you obtain the equivalent probability statement:

$$\mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma \leqslant \mu \leqslant \bar{X} + \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma \right) \approx 1 - \alpha$$

This shows that there is a fixed probability that the unknown mean parameter μ will fall within the stated bounds. Note here that the sample mean $\bar{X}$ is the random quantity in the expression, so the statement expressed the probability that a fixed parameter value μ falls within the random bounds of the interval.

The confidence interval: From here, we form the confidence interval by substituting the observed sample mean, yielding the $1 - \alpha$ level confidence interval:

$$\text{CI}_\mu(1 - \alpha) = \left(\bar{x} - \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma, \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma \right).$$

We refer to this as a "confidence interval" (as opposed to a probability interval) since we have now substituted the random bounds with observed bounds. Note that the true mean parameter is treated as fixed, so the interval either does or does not contain the parameter when we consider the probability of the parameter being within the interval.

(b) First we know there are $n = 8$ samples and so we have to apply the CLT even though it might not be very accurate for this number of samples. Now we want the 95% confidence interval. This translates to an interval probability of $1 - \alpha = 0.95$, i.e. $\alpha = 0.05$. Noting the formula for confidence intervals in this case given above is:

$$\text{CI}_\mu(1 - \alpha) = \left(\bar{x} - \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma, \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \sigma \right).$$

we see we need to calculate the sample mean, $\bar{x}$, and $z_{\alpha/2} = z_{0.05/2} = z_{0.025}$. Given the weights above we find $\bar{x} = 61.375$. Looking up the z-table in the formula sheet for z-values greater than zero we need to find the z-value, $z_{\alpha/2}$, for which the probability $p = 1 - \alpha/2 = 1 - 0.05/2 = 1 - 0.025 = 0.975$. From the table this turns out to be $z_{0.025} = 1.96$. Now we can evaluate the confidence interval as:

$$\text{CI}_{\mu}(0.95) = \left(61.375 - \frac{1.96}{\sqrt{8}} \cdot 4, 61.375 + \frac{1.96}{\sqrt{8}} \cdot 4\right) = (58.60, 64.15).$$

(c) We want some confidence that we know the true value of the mean of a population. If we have enough samples we know by the law of large numbers that our sample mean will approximate the true mean well. If we have a low number of samples (common in practice) we need another way to characterise the true mean of the population. Under certain assumptions, confidence intervals help us do this by placing our true mean within a range of values.

Question 2

A small micro-loan bank has 500 loan customers. If the total annual loan repayments made by an individual is a random variable with mean \$750 and standard deviation \$900, approximate the probability that the average total annual repayments made across all customers is greater than \$755.

Answer 2

Apply the (sample mean version of the) CLT using $N(750, 900^2/500) = N(750, 1620)$.

The z-value for this is $\frac{755-750}{\sqrt{1620}} = 0.1242$.

Looking this up in the Z-table the closest value is 0.12 which corresponds to a probability of $P(Z < 0.12) = 0.5478$.

So the probability that the average total annual payments are greater than \$755 is approximately:

$P(payments > 755) = P(Z > 0.12) = 1 - P(Z < 0.12) = 1 - 0.5478 = 0.4522$.

Question 3

There are 10000 students participate in an exam and the exam score approximately follow a normal distribution. Given that 359 students get scores above 90 in the exam and 1151 students get lower than 60. If there are 2500 students pass the exam, find out the lowest score to pass.

Answer 3

Let X represents the exam score and $X \sim N(\mu, \sigma^2)$

`359` students get score above `90` $\Rightarrow P(X > 90) = P\left(\frac{X-\mu}{\sigma} > \frac{90-\mu}{\sigma}\right) = Pr\left(Z > \frac{90-\mu}{\sigma}\right) = 0.0359$

Similarly, $P(X < 60) = P\left(\frac{X-\mu}{\sigma} < \frac{60-\mu}{\sigma}\right) = Pr\left(Z < \frac{60-\mu}{\sigma}\right) = 0.1151$

with $Z \sim \mathcal{N}(0, 1)$

Looking up **Z table**, you can find that $\frac{90-\mu}{\sigma} \approx 1.8$ and, $\frac{60-\mu}{\sigma} \approx -1.2$

Soving for this, we can find that $\mu = 72, \sigma = 10$

Let k be the lowest score to pass and there are `2500` students pass the exam, then
$P(X \geq k) = P\left(\frac{X-\mu}{\sigma} \geq \frac{k-\mu}{\sigma}\right) = Pr\left(Z \geq \frac{k-\mu}{\sigma}\right) = 0.25$

Looking up **Z table**, you can find that $\frac{k-\mu}{\sigma} \geq 0.67$. Given that $\mu = 72, \sigma = 10$, we will be able to calculate that $k \geq 78.7$.

Thus, the lowest score to pass is 78.7

Note: the question looks a bit like it involves the CLT but you can see from the solution that we have not applied it. So it is kind of a trick question.

Question 4

The light bulb in Monash university has an average lifetime of 1000 hours with a standard deviation of 50 hours. How many of these light bulbs should Monash stock up so that it can guarrantee that the light will be on for at least 7200 hours with a probability of at least 98%?

Answer 4

$x \sim \mathcal{D}(\mu = 1000; \sigma = 50)$, **Note** we don't know what type of distribution the lightbulb lifetime, x, follows so we label it with the arbitrary label $\mathcal{D}$.

Thus, according to (the sample sum version of the) CLT, $\sum x \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$

$\Rightarrow \sum x \sim \mathcal{N}(1000n; 2500n)$

We need to find n that satisfies the following condition: $\Pr(\sum x \geq 7200) \geq 0.98$

We will need to convert to standardised distance:

$$\Pr\left(\textbf{Z-score} \geq \frac{7200 - 1000n}{\sqrt{2500n}}\right) \geq 0.98$$

Using the Z-table we have $\Rightarrow \frac{7200-1000n}{50\sqrt{n}} = 2.06$. Solving for this we have that $n \approx 6.93$ meaning that we need 7 light bulbs in total to guarantee that the light will be on for at least 7,200 hours at 98 percent confidence.

Question 5

During the European football championships in 2008, and the football World Cup in 2010, **an octopus called Paul** living at an aquarium in Oberhausen, Germany, was used to predict the outcome of football matches, mostly involving the German national football team. To obtain Pauls' predictions, his keepers at the aquarium would present him with two boxes of food before each match. Each box was covered in the flag of the two nations that were participating, and the box that Paul chose to feed from first determined which nation he predicted would win.

Paul was asked to predict the outcome of $\mathbf{14}$ matches, $\mathbf{12}$ of which involved Germany. He correctly predicted the outcomes of $\mathbf{12}$ matches, only incorrectly guessing that Germany would beat Croatia in the Euro 2008 group stage, and that Germany would beat Spain in the Euro 2008 final. Some people claimed he was an **"animal oracle"**:

Calculate an estimate of Paul's success rate at predicting football matches. Calculate a $95\ \%$ confidence interval for this estimate, and summarise/describe your results appropriately. Show all workings as required.

Answer 5

Paul estimated $\frac{12}{14}$ matches correctly, giving him an approximate success rate of 0.8571 , or 85.71%

So Paul will predict a given game outcome either correctly or incorrectly, so we can assume his predictions are a Bernoulli random variable.

Based on looking up the formula sheet, for the Bernoulli random variable with success probability parameter p (sometimes also called θ as in the lecture notes) unknown and applying the CLT, the two-sided confidence interval is $\hat{p} \pm Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ = $0.8571 \pm Z_{(1-0.95)/2}\sqrt{0.8571(1-0.8571)/14}$ = $0.8571 \pm 1.96\sqrt{0.8571(1-0.8571)/14}$ = $(0.674, 1)$. Note: in this answer we let our estimate of the success probability parameter $\hat{p}$ be 0.8571 as calculated based on the data and we also looked up $Z_{(1-0.95)/2} = Z_{0.025} = 1.96$ in the Z-table in the formula sheet.

R code hackers nail-biting challenge

Consider the standard normal distribution characterised by random variable:

$$Z \sim N(0,1), z \in \mathbb{R}$$

Using R, through simulation with $n = 200$ samples obtained with the R function rnorm and the calculation of normalised histograms show that

$$\mathbb{P}\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) \approx 1-\alpha.$$

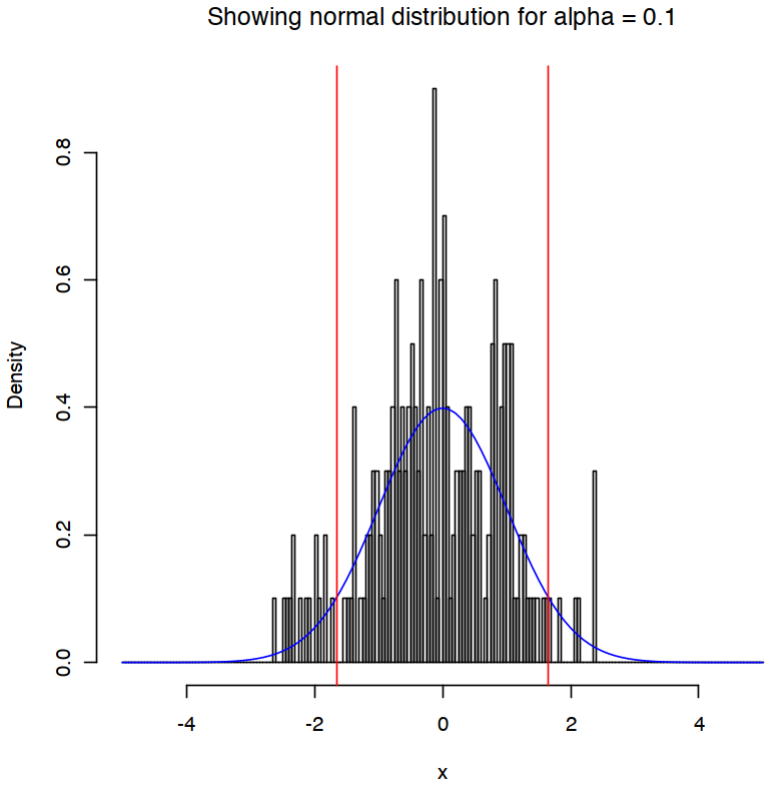
appears to be correct for $\alpha = 0.1$, 0.05 and 0.025. Note you will have to obtain the value of $z_{\alpha/2}$ using the R function qnorm. How do your simulated values for $\mathbb{P}\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right)$ compare to values for the same quantity computed using the R function pnorm. Which of these approaches gives a better estimate of $1-\alpha$? Why?

In [1]:
```
n <- 200
x <- rnorm(n)
```

In [2]:
```
#Following are the auxiliary functions for plotting and calculating probability.

plot_fn <- function(x, alpha, n, mean=0, sd=1, num_sds=5){
    #z-value is nothing but quantile for normal distribution.
    z_value <- qnorm(alpha/2)

    #Defining lower and upper limits for default 5 standard deviations from mean.
    lower_limit <- mean - num_sds*sd
    upper_limit <- mean + num_sds*sd
    #This would create a sequence of intervals to visualise proper plots for random variables for histogram
    breaks <- seq(lower_limit,upper_limit,by=alpha/2)

    h <- hist(x, breaks=breaks, plot=F) #putting hist object into variable to see density rather frequency
    h$density <- (h$counts/sum(h$counts))*2/alpha #dividing by alpha to normalise the density

    plot(h, freq=FALSE, main=paste("Showing normal distribution for alpha =", alpha))

    curve(dnorm(x,mean,sd), xlim=c(lower_limit,upper_limit), col="blue", add=T) #Normal curve

    #showing lower and upper limits depending upon alpha(or z-value)
    abline(v=(-z_value),col="red")
    abline(v=z_value,col="red")
}

prob_fn <- function(x, alpha){
    z_value <- qnorm(1-alpha/2)
    #Probability from simulation. Basically, calculating prob. getting all x that are between the alpha range.
    prob <- length(x[x>=-z_value & x<=z_value])/length(x)
    #Probability calculated using pnorm.
    pnorm_prob <- (pnorm(z_value) - pnorm(-z_value))

    print(paste("Showing values for alpha =", alpha))
    print(paste("Simulated value for probability, P(Z) =",prob))
    print(paste("Value of 1-alpha =", 1-alpha))
    print(paste("Probability calculated using pnorm =", pnorm_prob))
}
```

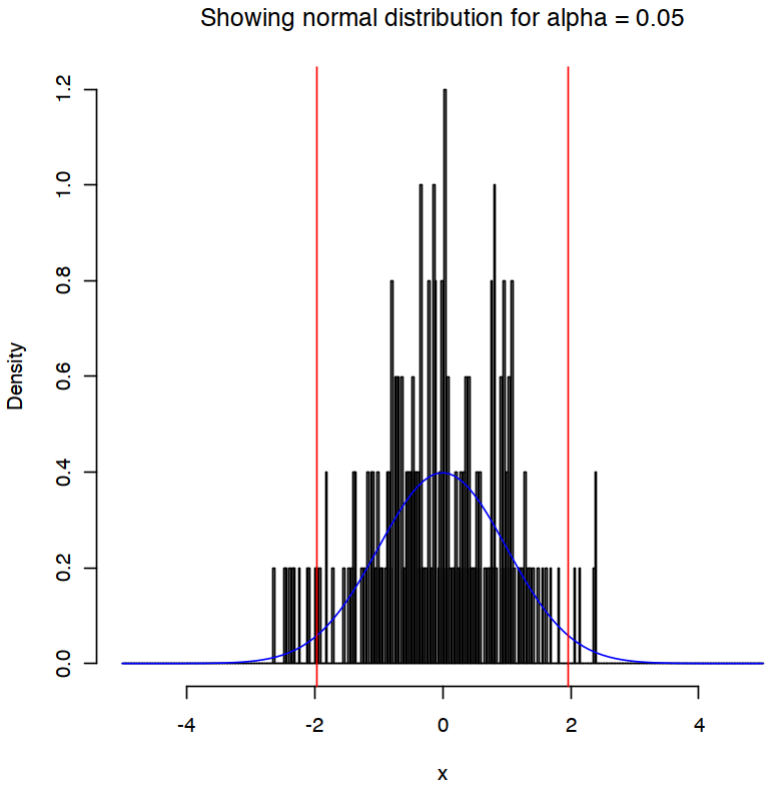
In [3]:
```
alpha <- 0.1
plot_fn(x, alpha, n)
prob_fn(x, alpha)
```

```
[1] "Showing values for alpha = 0.1"
[1] "Simulated value for probability, P(Z) = 0.89"
[1] "Value of 1-alpha = 0.9"
[1] "Probability calculated using pnorm = 0.9"
```



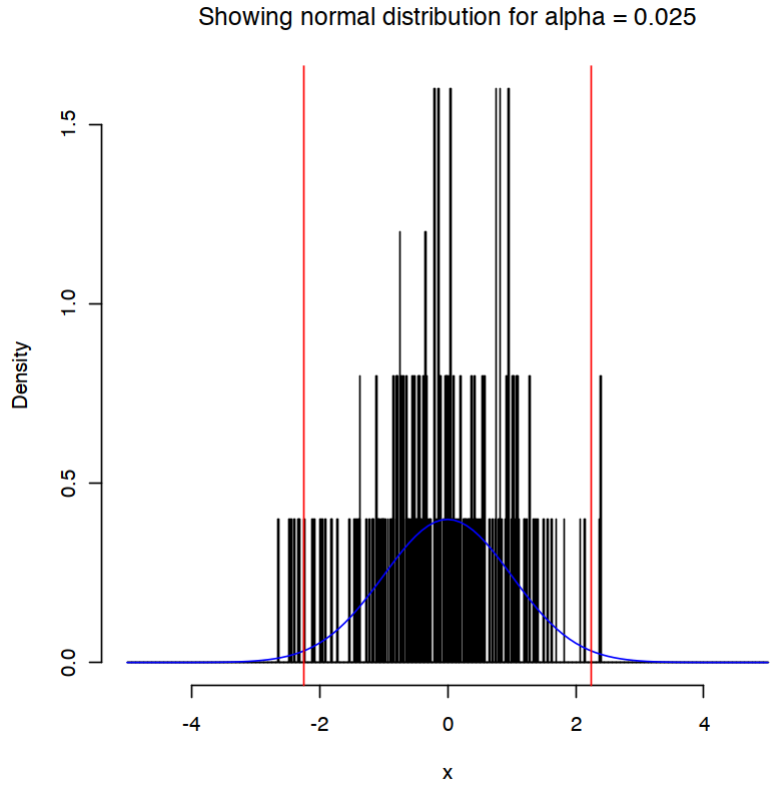
In [4]:
```
alpha <- 0.05
plot_fn(x, alpha, n)
prob_fn(x, alpha)
```

```
[1] "Showing values for alpha = 0.05"
[1] "Simulated value for probability, P(Z) = 0.925"
[1] "Value of 1-alpha = 0.95"
[1] "Probability calculated using pnorm = 0.95"
```



In [5]:
```
alpha <- 0.025
plot_fn(x, alpha, n)
prob_fn(x, alpha)
```

```
[1] "Showing values for alpha = 0.025"
[1] "Simulated value for probability, P(Z) = 0.955"
[1] "Value of 1-alpha = 0.975"
[1] "Probability calculated using pnorm = 0.975"
```



From these probability values, we can see that

$$\mathbb{P}\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) \approx 1-\alpha.$$

appears to be correct for $\alpha = 0.1$, 0.05 and 0.025.

Additionally, pnorm gives more accurate probabilities compared to simulation because pnorm implements the cumulative density function (cdf) directly. For the case of the normal distribution the error function is used based on numerical integration because an analytic expression for the cdf cannot be obtained. Nevertheless, numerical integration can perform accurate calculation of the cdf values. As we increase n, the probability calculated from simulation gets more closer to the pnorm value. The hist will also start looking more closer to normal curve. Try playing with the above code by increasing n.

Note

"Still don't understand sampling distributions and the CLT? Try this simulator (http://onlinestatbook.com/stat_sim/sampling_dist/) out which shows you how to get a simulated sampling distribution of several different statistics (e.g. mean, standard deviation) by sampling from random variables of different kinds (e.g. normal, uniform). The CLT corresponds to the case of the sample mean distribution when sampling from any kind of i.i.d. random variables."