A4: Central Limit Theorem, Confidence Intervals and Bootstrap

EPIB 607 - FALL 2021

your name and McGill ID

compiled on October 18, 2021

1 (25 points) Concordance between PCR-based extraction-free saliva and nasopharyngeal swabs for SARS-CoV-2 testing - PART I

(a)

Table (2 points), Coding the numbers and not just copying them (3 points) (-0.5 for not adding % agreement)

```
## ---- Question-1 -------------------------------------------------------------
dt_sum <- dt %>%
  transmute(naso = ifelse(is.na(Nasopharyngeal), "negative", "positive"),
            sal = ifelse(is.na(Saliva), "negative", "positive")) %>%
  group_by(naso, sal) %>%
  summarize(n = n())
total_row <- dt_sum %>%
  group_by(naso) %>%
  summarize(n = sum(n))

col1 <- c( "Salivadirect - Saliva", "Salivadirect - Saliva", "Total")
col2 <- c("Positive", "Negative", "")
col3 <- c(dt_sum$n[4], dt_sum$n[3], total_row$n[2])
col4 <- c(dt_sum$n[2], dt_sum$n[1], total_row$n[1])

tab <- cbind.data.frame(col1, col2, col3, col4)
names(tab) <- c("", "", "Positive", "Negative")
tab <- tab %>%
  rbind.data.frame(c("Positive agreement",
                     "94.3 % (95 % CI 87.2-97.5 %)", "", "")) %>%
  rbind.data.frame(c("Negative agreement",
                     "95.9 % (95 % CI 92.4-97.8 %)", "", ""))

kable(tab, booktabs = T, align = "c") %>%
  collapse_rows(columns = 1, latex_hline = "major", valign = "middle") %>%
  add_header_above(c(" " = 2, "Nasopharyngeal swab" = 2))
```

		Nasopharyngeal swab	
		Positive	Negative
Salivadirect – Saliva	Positive	82	9
	Negative	5	212
Total		87	221
Positive agreement	94.3 % (95 % CI 87.2-97.5 %)		
Negative agreement	95.9 % (95 % CI 92.4-97.8 %)		

(b)

Percent agreement (2.5 points each)

```
# Positive agreement = 82 / 87
pa <- round((dt_sum$n[4] / total_row$n[2]) * 100, 1); pa
```

```
## [1] 94.3
```

```
# Negative agreement = 212 / 221
na <- round((dt_sum$n[1] / total_row$n[1]) * 100, 1); na
```

```
## [1] 95.9
```

(c)

No they didn't mention what method they used in the paper. (0.5 point)

Using an appropriate method (bootstrap, binomial) (2.5 points)

Comparison with the ones in the paper (0.5 point)

Stating Assumptions (1.5 points)

```
set.seed(295)
pa_boot <- na_boot <- rep(NA, 1000)

for (i in 1:1000){

  dt_boot <- sample_n(dt, size = nrow(dt),  replace = TRUE) %>%
    transmute(naso = ifelse(is.na(Nasopharyngeal), "negative","positive"),
              sal = ifelse(is.na(Saliva), "negative", "positive")) %>%
    group_by(naso, sal) %>%
    summarize(n = n())

  total_boot <- dt_boot %>%
  group_by(naso) %>%
  summarize(n = sum(n))

  pa_boot[i] <- round((dt_boot$n[4] / total_boot$n[2]) * 100, 1)
  na_boot[i] <- round((dt_boot$n[1] / total_boot$n[1]) * 100, 1)

}

pa_ci_95 <- quantile(pa_boot, probs = c(0.025, 0.975), na.rm = T)
na_ci_95 <- quantile(na_boot, probs = c(0.025, 0.975))
```

95% bootstrap CI for Positive agreement (PA) is [88.6%, 98.6125%]
95% bootstrap CI for Negative agreement (NA) is [93.2%, 98.2%]

Both 95% bootstrap CIs for PA and NA are similar to those calculated in the paper, they have the same width and are asymmetric.
Bootstrap requires a SRS, however it doesn't require any assumptions about sampling distribution of PA or NA unlike other traditional methods that require CLT or knowing the exact distribution of the parameter.

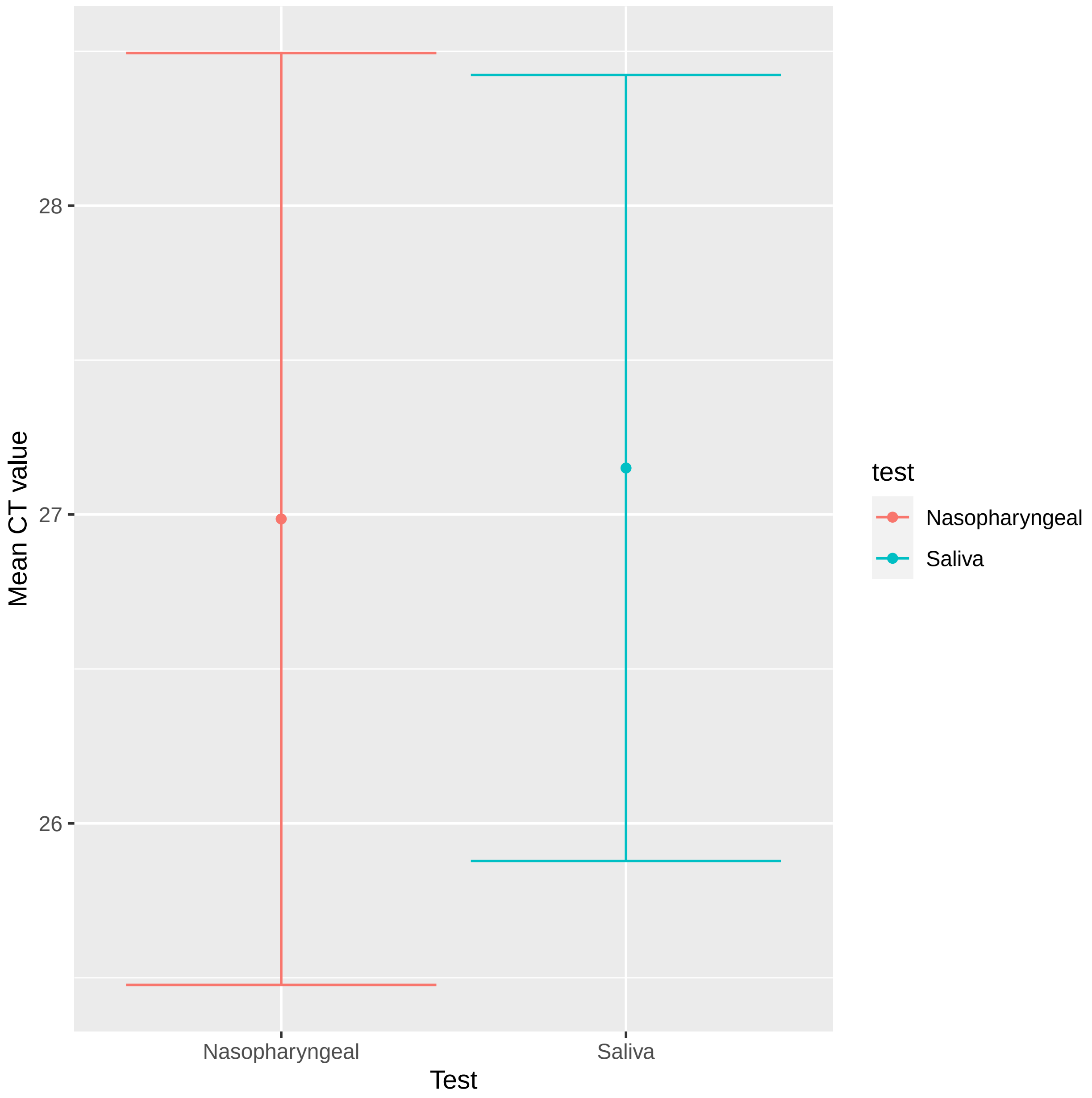
(d)

For the first 2 figures (fig A): tidy format
Saliva and NPS are categories of the same variable (type of test probably), they are mapped on the same aesthetic (position: x-axis). The data used to plot this graph consisted of two variables: **test** and its values were "Saliva" and "NPS"; and **result** and its values were CT values. (3 points)

For figure B: wide format
Saliva and NPS are different variables, mapped on two separate aesthetics (one on each axis). The data used to plot this graph consisted of two variables: **Saliva** and **NPS** and their values were CT values. (2 points)

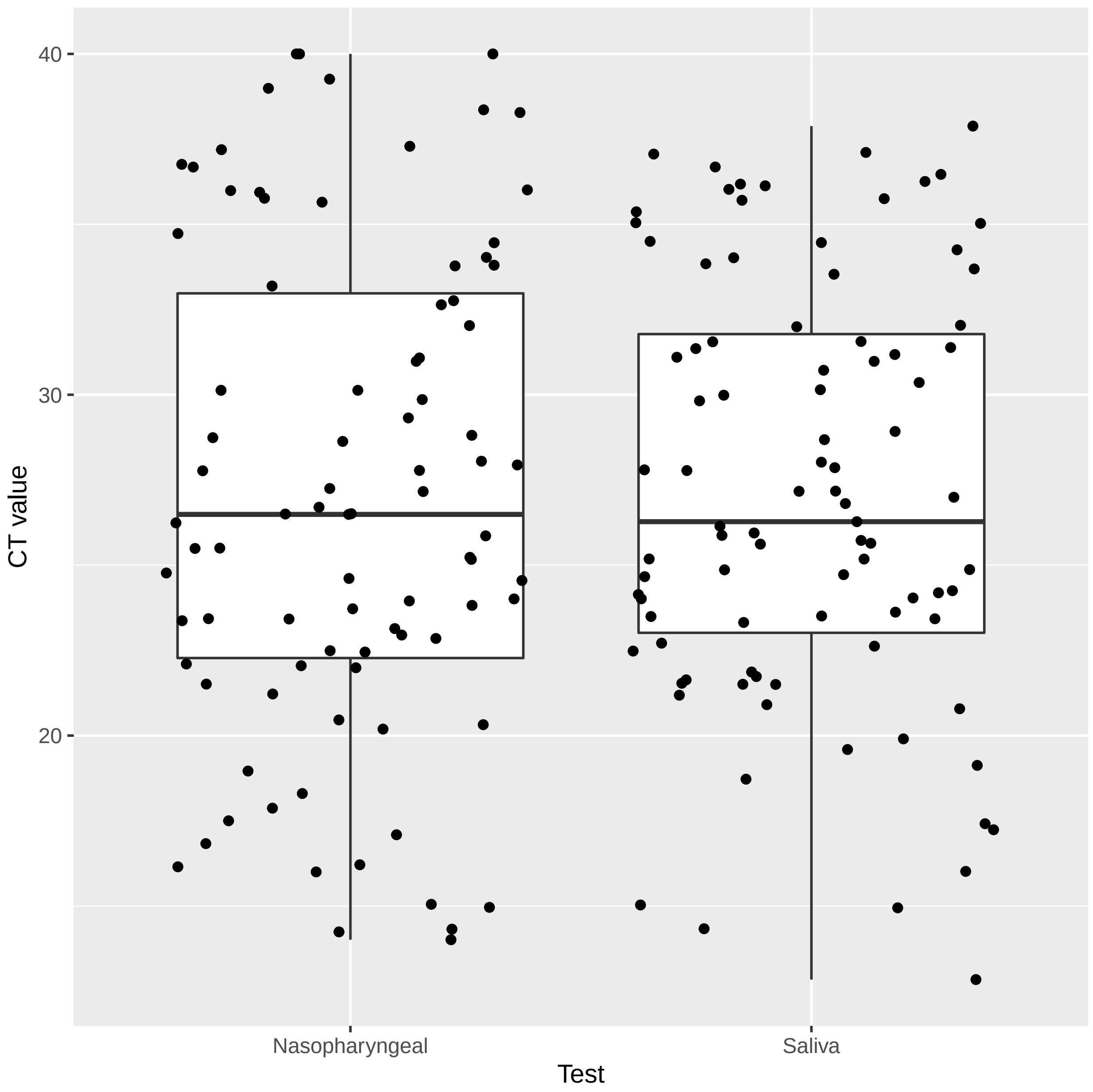
(e)

An appropriate graph (2.5 points)
Justification (2.5 points)



The 95% CI for both means overlap which supports the claim that the difference between the means isn't statistically significant.

OR



There isn't much difference between the distribution of the ct values for both tests since median is almost similar, and even though the IRQs are slightly different they don't seem to be notably different. Thus, due to similar overlapping distributions of the two tests, the mean Ct values is expected to be small and not be statistically significant.

2 (25 points) Concordance between PCR-based extraction-free saliva and nasopha- ryngeal swabs for SARS-CoV-2 testing - PART II

Notice: Since the questions are not very specific, you can calculate the mean and 95%CI based on the complete data in each group, not the concordant subjects as well.

(a)

Assumptions: (2.5 points) (1) The samples are simple random samples. (Identical and independent.) (2) We can apply the CLT to the sample mean because we assume the distribution of a sample mean approximates a normal distribution. (Or t-distribution is you want to be more conservative)

```r
mean_q2a1 <- mean(dt_symp$Nasopharyngeal)
se_q2a1 <- sd(dt_symp$Nasopharyngeal)/sqrt(length(dt_symp$Nasopharyngeal))

upper_lim_q2a1 = mean_q2a1 + 1.96*se_q2a1  # 1.96 is the 0.975 quantile for a standard
lower_lim_q2a1 = mean_q2a1 - 1.96*se_q2a1

cat("the 95% CI for Nasopharyngeal group is: (", lower_lim_q2a1, ",",upper_lim_q2a1,").
```

```
## the 95% CI for Nasopharyngeal group is: ( 24.82852 , 27.89024 ).
```

```r
mean_q2a2 <- mean(dt_symp$Saliva)
se_q2a2 <- sd(dt_symp$Saliva)/sqrt(length(dt_symp$Saliva))

upper_lim_q2a2 = mean_q2a2 + 1.96*se_q2a2
lower_lim_q2a2 = mean_q2a2 - 1.96*se_q2a2

cat(" the 95% CI for Saliva group is: (", lower_lim_q2a2, ",",upper_lim_q2a2,"). ")
```

```
##  the 95% CI for Saliva group is: ( 25.17912 , 27.80547 ).
```

Give the correct standard error and 95% CI (2.5 points)
Important: Present your result in words! Marks will be deducted if you don't specify which variable the confidence interval belongs to.

(b)

By looking at the 2 CIs, I'm convinced that the two means have no statistical difference. The 95 CIs seem to largely overlap with each other.

(As long as your explanation is reasonable and no theoretical mistake. 5 points)

(c)

We can use the built-in (paired-up) t-test function to calculate the 95% CI for the mean difference.
The assumptions for running a t-test is: (2 points)
1. the data is collected from a representative, randomly selected portion of the total population. (It's a simple random sample)
2. when sample size large, the data (difference) will follow a normal distribution, bell-shaped distribution curve
3. Equal variance exists when the standard deviations of both samples are approximately equal.

```
t_result <- t.test(dt_symp$Nasopharyngeal,dt_symp$Saliva,paired = TRUE)
t_result
```

```
##
##  Paired t-test
##
## data:  dt_symp$Nasopharyngeal and dt_symp$Saliva
## t = -0.17725, df = 80, p-value = 0.8598
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.625169  1.359352
## sample estimates:
## mean of the differences
##              -0.1329084
```

```
cat("The 95% CI for the mean difference is: (",t_result$conf.int[1], ",",t_result$conf.i
```

```
## The 95% CI for the mean difference is: ( -1.625169 , 1.359352 )
```

Build-in function: (1 points, you can used other functions as long as it's correct)
Correct 95% CI (2 points)

(d)

```
set.seed(1129)

sample_difference <- dt_symp$Nasopharyngeal - dt_symp$Saliva

bootstrap_difference <- replicate(10000,sample(sample_difference,replace=TRUE))
```

```
mean_difference <- apply(bootstrap_difference,2,mean)

bootstrap_quantile <- quantile(mean_difference,c(0.025,0.975))
cat("The 95% bootstrap CI for the mean difference is: (",bootstrap_quantile[1], ",",boot
```

```
## The 95% bootstrap CI for the mean difference is: ( -1.587693 , 1.303271 )
```

Correct bootstrap code (various options) and reasonable 95% bootstrap CI. (5 points)

(e)

The standard error calculated by hand is:

```
se_manual <- sd(sample_difference)/sqrt(length(sample_difference))
se_manual
```

```
## [1] 0.7498556
```

```
## The se used in (c) is: 0.7498556 . The se in bootstrap sample is: 0.7404514
```

Thus, our calculated standard error is the one used in part (c), and it's close (but not equal) to the one calculated from part (d). Part (a) investigates the standard errors for the 2 groups.

Correct equation and standard error. (3 points)
Comparison to the se in (a), (c) and (d). (2 points)

3 (25 points) Concordance between PCR-based extraction-free saliva and nasopha- ryngeal swabs for SARS-CoV-2 testing - PART III

(a)

The sensitivity (probability that someone who is positive for Covid on the NPS test is positive on the saliva test) and specificity (probability that someone who is negative for Covid on the NPS test is negative on the saliva test) in the asymptomatic cohort are both 100%.

```
## [1] 1
```

```
## [1] 1
```

(b)

The confidence interval given in the article is (5%, 100%). I disagree with the reviewer's statement that the confidence interval is very wide. There was only one person who had Covid in the asymptomatic cohort, so we have a very small sample size of n=1 to calculate sensitivity with (since sensitivity is only calculated using individuals who are truly positive/positive on the NPS test). Therefore, it is to be expected that we would have a very wide confidence interval.

(c)

No, we are unable to calculate a 95% bootstrap CI for sensitivity in the asymptomatic cohort. Only one individual was actually positive for Covid using the NPS test. We would only be able to sample that one individual, and therefore each bootstrap replication would result in the same bootstrap sample with no variability and an estimated sensitivity of 100%. Since bootstrapping relies on variability between samples, we would not be able to estimate a confidence interval.

(d)

We can think of specificity as a binomial proportion (proportion of true negatives among true negatives and false positives). Since we have less than 5 false positives, we need to use an exact or assymmetric method to calculate the confidence interval. Using an exact binomial method (Clopper-Pearson), we get a 95% CI of (98.0%, 100%). Using the Wilson score method, we get a 95% CI of (97.9%, 100%), which is the same as the interval provided in the article. Therefore, we can assume the authors provided the Wilson score interval.

Note that we would not be able to bootstrap a CI for specificity because there are no false positives to be sampled in the dataset, nor would we be able to use a normal approximation (Wald interval) due to the low number of non-events (the upper bound of the CI would exceed 100%). Also of note is that the CI for specificity is much narrower than that for sensitivity because we have a much larger sample size of 180 individuals who tested negative on the NPS test.

4 (25 points) How deep is the ocean?

(a)

The student_id is fake, so the ocean depth is manually randomly selected.

```
source("https://github.com/sahirbhatnagar/epib607/raw/master/inst/labs/003-ocean-depths/
index_n5 <- 2008:2012
sample_5 <- automate_water_task(index = index_n5, student_id = 260610974)
```

```
mean_5 <- mean(sample_5)
se_5 <- sd(sample_5)/sqrt(5)

#By formula
lower_formula_5 <- mean_5 - qnorm(0.975)*se_5
upper_formula_5 <- mean_5 + qnorm(0.975)*se_5

#By qnorm()
CI_qnorm_5 <- qnorm(c(0.025,0.975),mean = mean_5,sd = se_5)

#By bootstrap
bootstrap_5 <- replicate(10000,mean(sample(sample_5,replace = TRUE)))
CI_bootstrap_5 <- quantile(bootstrap_5,c(0.025,0.975))
```

```
cat("For n = 5, the caculated 95% Confidence interval using the ± formula is: (", lower_
```

```
## For n = 5, the caculated 95% Confidence interval using the ± formula is: ( 1291.786 ,
```

```
cat("using qnorm() is: (",CI_qnorm_5[1],",",CI_qnorm_5[2],");")
```

```
## using qnorm() is: ( 1291.786 , 4104.214 );
```

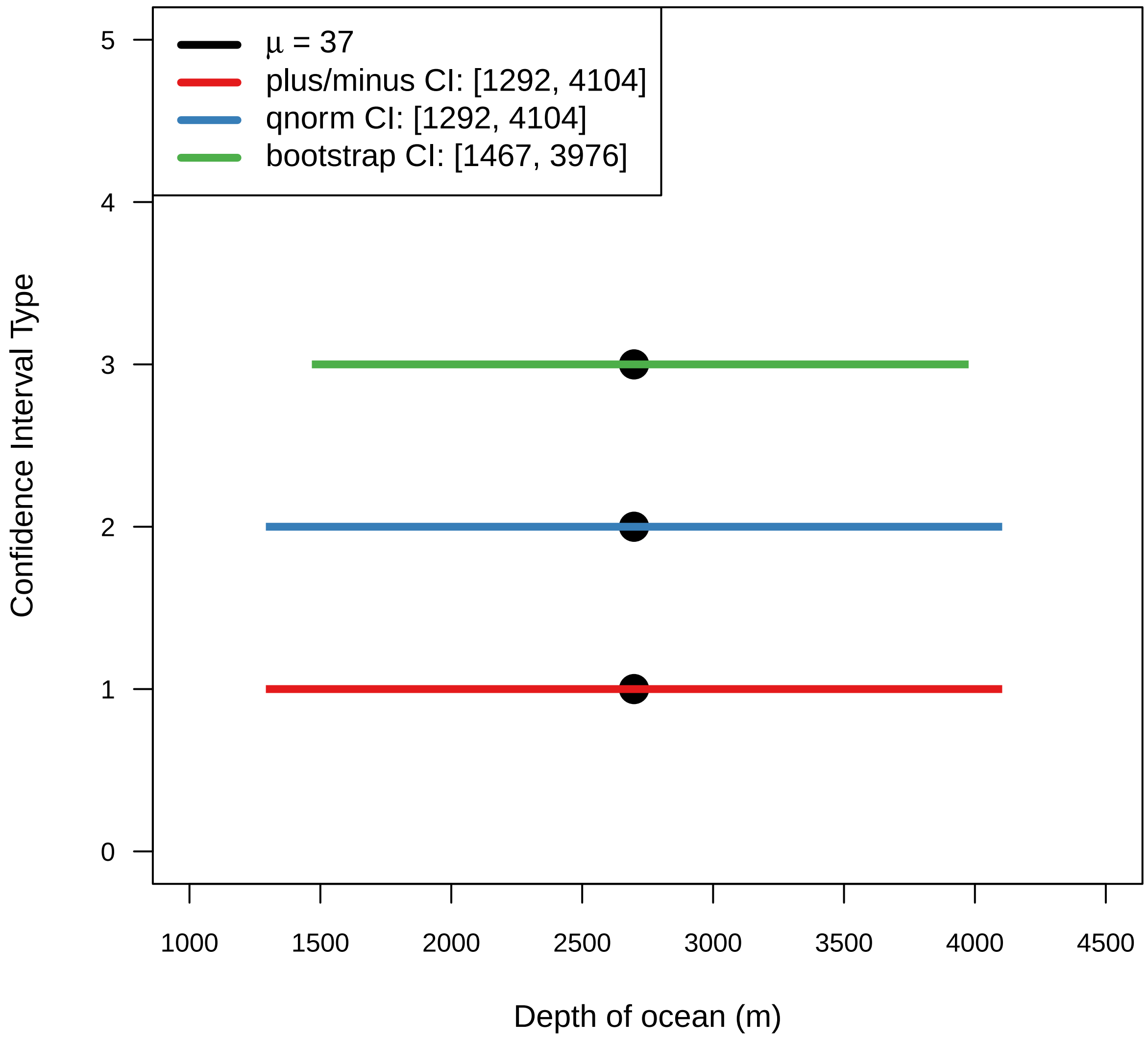
```
cat("using bootstrap is: (",CI_bootstrap_5[1],",",CI_bootstrap_5[2],")")
```

```
## using bootstrap is: ( 1467.4 , 3976 )
```

Correct code for "formula calculation" (1 pt); "qnorm() calculation" (2 pts); "Bootstrap sampling" (2 pts)

(b)

```
compare_CI(ybar = mean_5, PM = c(lower_formula_5, upper_formula_5), QNORM = CI_qnorm_5,
```



The 95% CI calculated using the formula is equal to that using the qnorm() function. (1 pt, if you use a t-distribution for the formula, then the 95% CI will be wider than that using the qnorm())
The 95% CI calculated with bootstrap samples is not symmetric, while the other two are. (1 pt) The bootstrap 95% CI is narrower than the other two. (1 pt)

Correct confidence interval plot. (2 pts)

(c)

Repeat with n = 20.

```
mean_20 <- mean(sample_20)
se_20 <- sd(sample_20)/sqrt(20)

#By formula
lower_formula_20 <- mean_20 - qnorm(0.975)*se_20
upper_formula_20 <- mean_20 + qnorm(0.975)*se_20

#By qnorm()
CI_qnorm_20 <- qnorm(c(0.025,0.975),mean = mean_20,sd = se_20)

#By bootstrap
bootstrap_20 <- replicate(10000,mean(sample(sample_20,replace = TRUE)))
CI_bootstrap_20 <- quantile(bootstrap_20,c(0.025,0.975))
```

```
cat("For n = 20, the caculated 95% Confidence interval using the ± formula is: (", lower
```

```
## For n = 20, the caculated 95% Confidence interval using the ± formula is: ( 2265.677
```

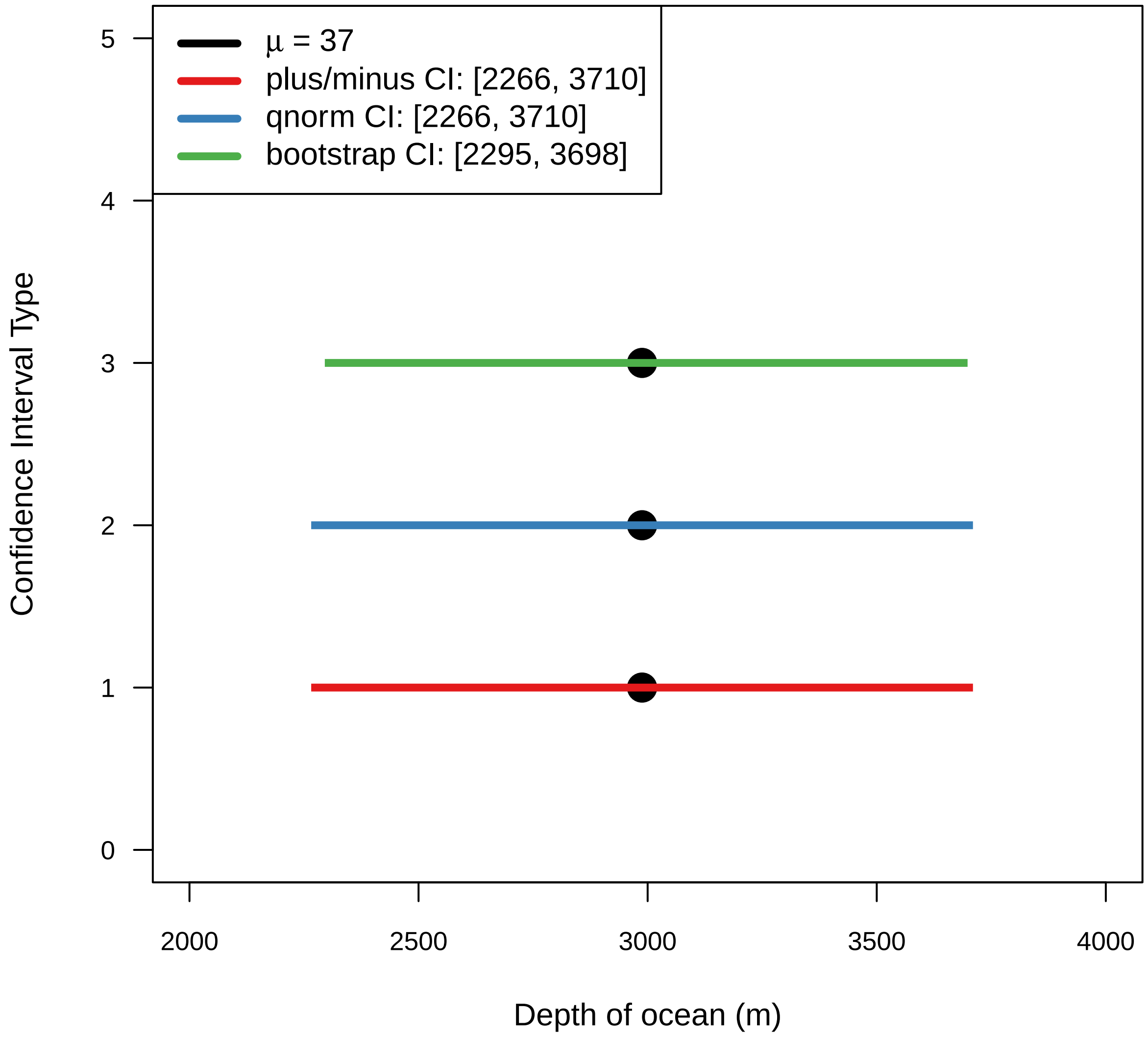
```
cat("using qnorm() is: (",CI_qnorm_20[1],",",CI_qnorm_20[2],");")
```

```
## using qnorm() is: ( 2265.677 , 3710.023 );
```

```
cat("using bootstrap is: (",CI_bootstrap_20[1],",",CI_bootstrap_20[2],")")
```

```
## using bootstrap is: ( 2295.398 , 3698.262 )
```

```
compare_CI(ybar = mean_20, PM = c(lower_formula_20, upper_formula_20), QNORM = CI_qnorm_
```



Correct calculation (3 pts) and correct plot (2 pts).

In general, with n = 20, the standard error of the sample is smaller than that with n = 5. Thus, all 95% CI are narrower than those calculated with n = 5. (2 pts)
Similar to (b), the 95% confidence interval calculated using the first two method is the same, while that by bootstrap is narrower. (2 pts)
However, when we have n = 20, the difference 95% CI given by bootstrap and qnorm() is smaller compared to what we have when n = 5. (CLT probably has kicked in when n = 20) (1 pt)

(d)

N = 60 represents the sampling procedure is repeated by 60 students, which represents the number of sample groups. n = 5 or n = 20 indicates how many data values are drawn by each student, which is the sample size. (3 pts)
When n gets larger, the sample mean is more likely to follow a normal distribution such that the Central-Limit-Theorem will kick in (2 pts), and the estimates made with bootstrap samples will converge to the true values.