Homework 5

Jackson Dial

10/23/2021

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(naniar)

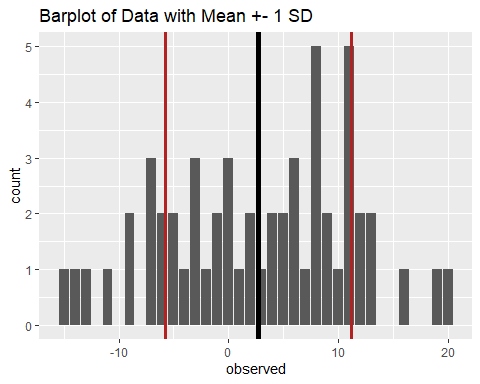
# Question 1

## Part A

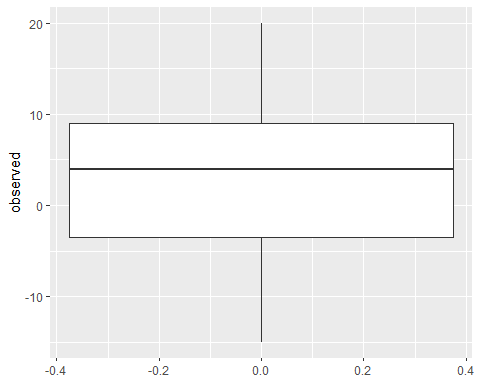
At the 0.05, is there enough evidence to suggest the population standard deviation change in heart rate is different from 7.5? Conduct a hypothesis test. Be sure to CHECK the assumptions.

#### Step 1: Know the Data

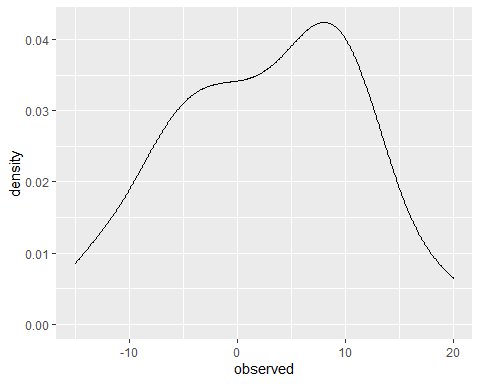
observed <- c(-4,11,0,0,8,5,-1,11,11,8,-9,1,10,2,-6,4,-11,13,8,-7,6,13,2,11,-14,-9,-5,12,  
 -6,11,-3,-1,7,8,-3,0,-5,8,9,19,-15,-7,20,9,-2,6,6,3,4,-3,-7,16,5,-13,12)  
m <- mean(observed)  
n <- length(observed)  
s <- sd(observed)  
  
df <- as.data.frame(observed)  
  
par(mfrow = c(1,3))  
ggplot(df, aes(x = observed))+  
 geom\_bar()+  
 geom\_vline(xintercept = m, size = 2)+  
 geom\_vline(xintercept = m + c(s,-s), size = 1.2, color = "firebrick")+  
 labs(title = "Barplot of Data with Mean +- 1 SD")



ggplot(df, aes(y = observed))+  
 geom\_boxplot()



ggplot(df, aes(x = observed))+  
 geom\_density()



We assume that our population is normally distributed with unknown mean and variance.

#### Step 2: Set up Statistical Hypotheses

We will use a significance level of for this hypothesis test where is the unknown population standard deviation.

#### Step 3: Choose a Test Statistic and Determine Sampling Distribution

This is a test for the true population standard deviation. We will use the following test statistic and distribution:

#### Step 4: Calculate the Test Statistic and P-value

sig <- 7.5  
  
ts <- (n-1)\*s^2 / sig^2  
pval <- pchisq(ts, n-1, lower.tail = FALSE)\*2

The test statistic is 68.97, and the two-sided p-value is 0.16

#### Step 5: Draw Inference Based on Significance Level

Based on our level of significance we fail to reject the null hypothesis that the true standard deviation is equal to 7.5.

## Part B

Construct a 95% CI for the population standard deviation. Interpret the CI in a way that a clinician can understand.

uts <- qchisq(.025, df = n-1)  
lts <- qchisq(.975, df = n-1)  
  
lower <- sqrt(((n-1)\*s^2)/lts)  
upper <- sqrt(((n-1)\*s^2)/uts)

The 95% confidence interval for the true standard deviation is (7.14, 10.44).

We are 95% confident that the standard deviation change in heart rate is between 7.14 and 10.44 for the true population.

# Question 2

Researchers are interested in incident rates of cataracts among individuals exposed to excess sunlight. In a prospective cohort study, they identified 200 people with excess sunlight exposure and followed them for 5 years and determined that 22 of them developed cataracts.

## Part A

Is the expected rate of cataracts over 5 years significantly lower than 30? Conduct a hypothesis test.

#### Step 1: Know the Data

Assume that the rate of cataracts among individuals exposed to excess sunlight follows a Poisson process with rate 𝜇 = 5𝜆, where “rate” is defined as count per something, or in this case, number of cataracts per 5 years. We adopt all of the assumptions of a Poisson process.

#### Step 2: Set up Statistical Hypotheses

We will use a level of significance of for this hypothesis test.

#### Step 3: Choose a Test Statistic and Determine Sampling Distribution

Our test statistic is:

Which can be transformed into a standard normal random variable, giving us a test statistic of:

#### Step 4: Calculate the Test Statistic and P-value

z <- (sqrt(22) - sqrt(30)) / (1/2)  
pval2 <- pnorm(z)

The test statistic for this hypothesis test is -1.57, and the one-sided p-value for this test is 0.06.

#### Step 5: Draw Inference Based on Significance Level

Based on our level of significance and p-value, we fail to reject the null hypothesis that the expected rate of cataracts over 5 years is equal to 30.

## Part B

Construct an exact and approximate one sided 90% CI for the population rate.

#### Approximate

z <- qnorm(.9)  
lower <- 0  
upper <- (sqrt(22) + z\*sqrt(1/4))^2

The approximate 90% confidence interval is (0, 28.42).

#### Exact

lower <- 0  
upper <- qchisq(.9, (22+1)\*2) /2

The exact 90% confidence interval is (0, 29.32).

# Final Project

# Question 1: Analysis Plan

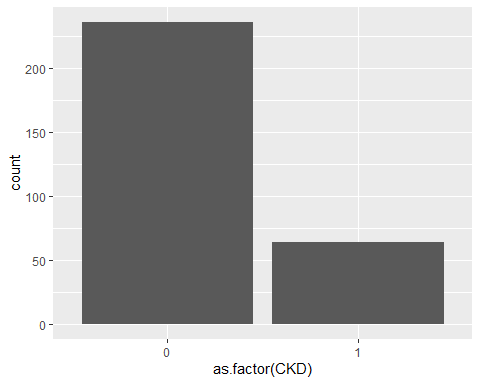
#### Read Data

setwd("C:/Users/jacks/OneDrive/Desktop/DUKE\_FALL2021/702/Final Project")  
dat <- read.csv(file="NHANES702.csv", header=TRUE, na.strings="",  
 stringsAsFactors=TRUE)  
  
dat["BMI"][which(dat$BMI > 70),] <- NA  
dat["SBP"][which(dat$SBP > 180),] <- NA  
dat["ACR"][which(dat$ACR >= 6000),] <- NA  
dat["A1C"][which(dat$A1C >= 11),] <- NA  
  
dat1 <- dat %>%  
 mutate(CKD = case\_when(  
 eGFR < 60 | ACR >= 30 ~ 1,  
 eGFR >= 60 & ACR < 30 ~ 0,  
 T ~ 2  
 ),  
 serumKCat = case\_when(  
 serumK >= 3 & serumK <= 4 ~ "low-normal",  
 serumK > 4 ~ "normal",  
 T ~ "NA"  
 ),  
 dietKCat = case\_when(  
 dietK1000 < 1534 ~ "inadequate intake",  
 dietK1000 >= 1534 & dietK1000 < 2238 ~ "borderline adequate intake",  
 dietK1000 >= 2238 ~ "adequate intake",  
 T ~ "NA"  
 )  
 ) %>%   
 replace\_with\_na(replace = list(CKD = 2,  
 serumKCat = "NA",  
 dietKCat = "NA"))  
  
ads <- dat1 %>%   
 filter(is.na(CKD) == FALSE, is.na(serumK) == FALSE)

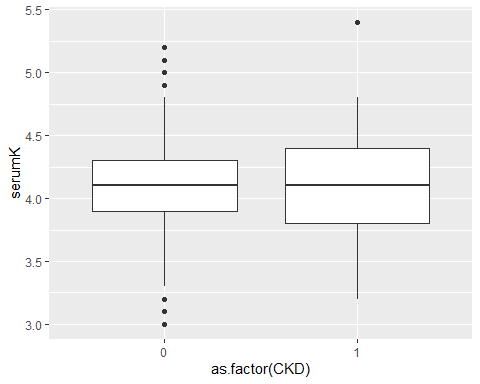
## Section 4.1

#ads is created from the "Final Project.Rmd" file   
ggplot(ads, aes(x = as.factor(CKD)))+  
 geom\_histogram(stat = "count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad



ggplot(ads, aes(x = as.factor(CKD), group = as.factor(CKD), y = serumK))+  
 geom\_boxplot()



The two populations we are examining, those with positive CKD status and those with negative CKD status, are both assumed to be normally distributed, and to be independent of each other.

We will assume that serum K among those with positive CKD status:

We will assume that serum K among those with negative CKD status:

ckdPos <- ads %>% filter(CKD == 1)  
ckdNeg <- ads %>% filter(CKD == 0)  
  
xbar1 <- mean(ckdPos$serumK)  
s1 <- sd(ckdPos$serumK)  
  
xbar2 <- mean(ckdNeg$serumK)  
s2 <- sd(ckdNeg$serumK)

A random sample of was taken and the sample mean +/- sample variance was found to be: 4.111 0.418.

A random sample of was taken and the sample mean +/- sample variance was found to be: 4.089 0.358.

The random variables of the mean serum K for respective populations of positive/negative CKD status are also assumed to be independent of each other.

We are performing the variance ratio test to determine if variance can be assumed to be equal for individuals with positive CKD status and those with negative CKD status.

A second way we could test for equal variances is to use the rule of thumb.

#### If variances are equal

Perform a two-sample t-test for the true difference in means, with the following parameters:

Note that the means are assumed to be different, but the variances are assumed to be equal, however both the means and the variance are assumed to be unknown.

Of interest is , but we use

We will assume:

Where is the unknown difference in mean serum K between those with positive CKD status and negative CKD status .

#### If variances not equal

Perform a two-sample t-test for the true difference in means, with the following test statistic:

This creates an issue, however, as does no longer follow a distribution. The exact distribution of this statistic is unknown, so we will use a Satterthwaite approximation instead, where:

and

This distribution is now used as an approximation of the distribution above and can be used to conduct hypothesis testing and construct confidence intervals.

We will make the following assumptions:

We will test the following hypotheses:

Where is the unknown difference in mean serum K between those with positive CKD status and negative CKD status .

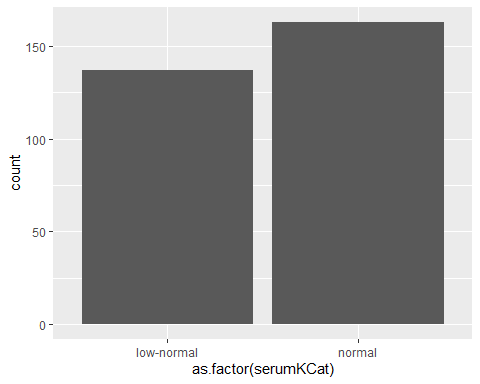
## Section 4.2

Similar analyses as described in Section 4.1 above will be performed, with serum K replaced with dietary K intake.

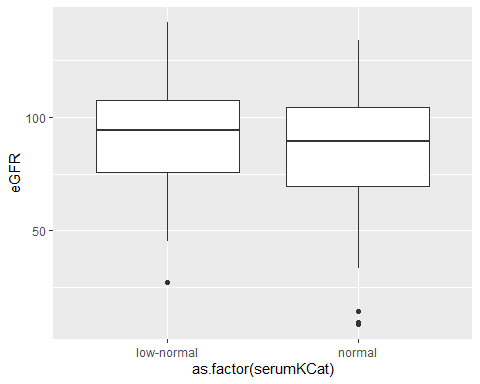
## Section 4.3

ggplot(ads, aes(x = as.factor(serumKCat)))+  
 geom\_histogram(stat = "count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad



ggplot(ads, aes(x = as.factor(serumKCat), group = as.factor(serumKCat), y = eGFR))+  
 geom\_boxplot()



The two populations we are examining, those with low-normal Serum K and those with normal Serum K, are both assumed to be normally distributed, and to be independent of each other.

We will assume that the population of those with low-normal Serum K:

We will assume that the population of those with normal Serum K:

lownor <- ads %>% filter(serumKCat == "low-normal")  
nor <- ads %>% filter(serumKCat == "normal")  
  
xbar1 <- mean(lownor$eGFR)  
s1 <- sd(lownor$eGFR)  
  
xbar2 <- mean(nor$eGFR)  
s2 <- sd(nor$eGFR)

A random sample of was taken and the sample mean +/- sample variance was found to be: 92.2997713 22.263878.

A random sample of was taken and the sample mean +/- sample variance was found to be: 86.8545451 24.8070673.

The random variables of the mean eGFR for respective populations of low-normal/normal serum K are also assumed to be independent of each other.

One way we can determine if variance is assumed to be normal is to perform the variance ratio test for individuals with low-normal serum K and normal serum K.

A second way we could test for equal variances is to compute a confidence interval for each of the population variances, and if they overlap at all then we have evidence to assume equal variance. If they do not overlap, we have evidence to say that they are unequal.

#### If variances are equal

Perform a two-sample t-test for the true difference in means, with the following parameters:

Note that the means are assumed to be different, but the variances are assumed to be equal, however both the means and the variance are assumed to be unknown.

Of interest is , but we use

We will assume:

Where is the unknown difference in mean eGFR between those with low-normal serum K and normal serum K .

#### If variances not equal

Perform a two-sample t-test for the true difference in means, with the following test statistic:

This creates an issue, however, as does no longer follow a distribution. The exact distribution of this statistic is unknown, so we will use a Satterthwaite approximation instead, where:

and

This distribution is now used as an approximation of the distribution above and can be used to conduct hypothesis testing and construct confidence intervals.

We will make the following assumptions:

We will test the following hypotheses:

Where is the unknown difference in mean eGFR between those with low-normal serum K and normal serum K .

## Section 4.4

Similar analyses as described in Section 4.3 above will be performed, with serum K replaced with dietary K intake.

# Question 2

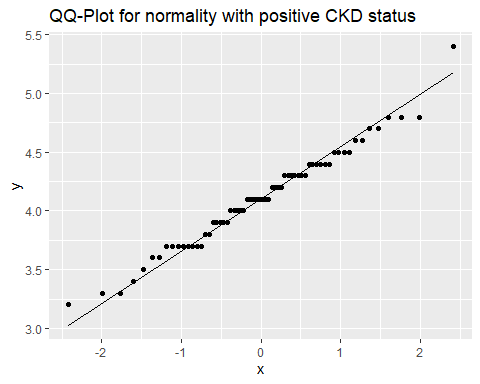
There are 64 individuals with positive CKD status, and 236 individuals with negative CKD status.

### Variance Ratio Test

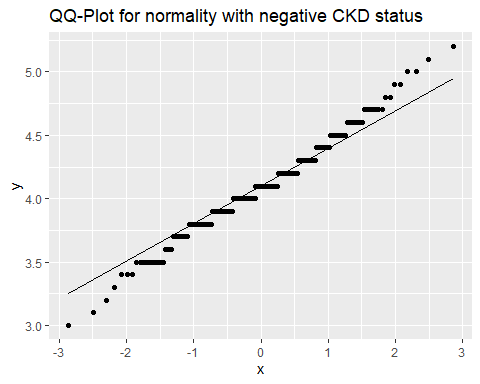
We will assume that our variance ratio test statistic follows this distribution:

Where and represents the distribution of individuals with negative CKD status, and represents the distribution of individuals with positive CKD status.

posvec <- ckdPos$serumK  
negvec <- ckdNeg$serumK  
ggplot(ckdPos) +  
 stat\_qq(aes(sample = posvec))+  
 stat\_qq\_line((aes(sample = posvec)))+  
 labs(title = "QQ-Plot for normality with positive CKD status")



ggplot(ckdNeg) +  
 stat\_qq(aes(sample = negvec))+  
 stat\_qq\_line((aes(sample = negvec)))+  
 labs(title = "QQ-Plot for normality with negative CKD status")



The hypotheses for this test are:

We will use a level of significance of

var.test(negvec, posvec, two.sided = TRUE)

##   
## F test to compare two variances  
##   
## data: negvec and posvec  
## F = 0.73237, num df = 235, denom df = 63, p-value = 0.1024  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.4817264 1.0636945  
## sample estimates:  
## ratio of variances   
## 0.7323675

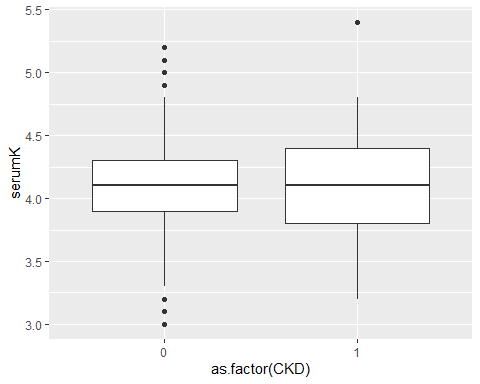
Since the variance test for equality produced a non-significant p-value, we fail to reject the null hypothesis and can conclude that the variances are equal.

### T-test

Now that we have proved that we can assume equal variance, we will run a two-sample t-test for the true difference of means.

#### Step 1: Know the data

ggplot(ads, aes(x = as.factor(CKD), y = serumK))+  
 geom\_boxplot()



We assume that our populations are both normally distributed with unknown mean and variances, and equal variance.

#### Step 2: Set up Statistical Hypotheses

Where and and This test will use a level of significance of

#### Step 3: Choose a Test Statistic and Determine Sampling Distribution

#### Step 4: Calculate the Test Statistic and P-value

mytest <- t.test(negvec, posvec, var.equal = TRUE)

The test statistic from our test is -0.4192926, and the p-value is 0.6753044.

#### Step 5: Draw Inference Based on Significance Level

Based on the results of our test, we fail to reject the null hypothesis that the difference in mean serum K between individuals with negative CKD status and positive CKD status is 0. This means that there is no significant difference in serum K levels when partitioned on positive/negative CKD status.

These results do not align with my original hypotheses, as I hypothesized that higher levels of serum K is associated with lower risk of CKD. Our results tell us there is no association.

#### Confidence Interval

The confidence interval for the true difference in mean serum K between individuals with positive and negative CKD status is [-0.124998, 0.0810891. As 0 is included in this interval, we can conclude there is no difference in mean serum K between individuals with positive and negative CKD status.

# Question 3

dietAdq <- ads %>%  
 mutate(dietK2Cat = case\_when(  
 dietKCat %in% c("borderline adequate intake", "adequate intake") ~ "adequate",  
 dietKCat == "inadequate intake" ~ "inadequate")) %>%   
 filter(dietK2Cat == "adequate")  
  
dietInadq <- ads %>%  
 mutate(dietK2Cat = case\_when(  
 dietKCat %in% c("borderline adequate intake", "adequate intake") ~ "adequate",  
 dietKCat == "inadequate intake" ~ "inadequate")) %>%   
 filter(dietK2Cat == "inadequate")

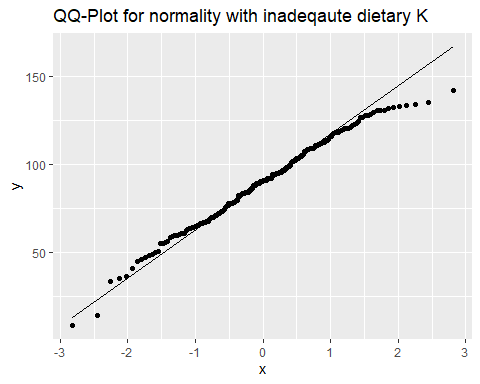
There are 207 individuals with inadequate dietary K, and 61 individuals with adequate dietary K.

### Variance Ratio Test

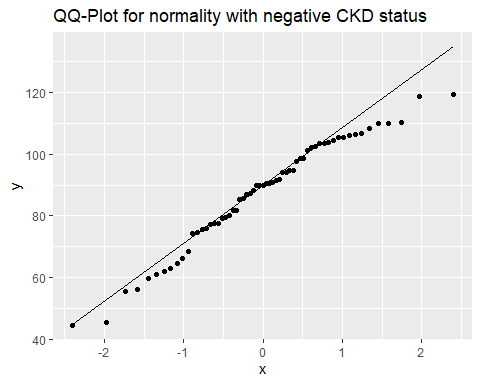
We will assume that our variance ratio test statistic follows this distribution:

Where and represents the distribution of individuals with inadequate dietary K, and represents the distribution of individuals with adequate serum K.

adqvec <- dietAdq$eGFR  
inadqvec <- dietInadq$eGFR  
ggplot(dietInadq) +  
 stat\_qq(aes(sample = inadqvec))+  
 stat\_qq\_line((aes(sample = inadqvec)))+  
 labs(title = "QQ-Plot for normality with inadeqaute dietary K")



ggplot(dietAdq) +  
 stat\_qq(aes(sample = adqvec))+  
 stat\_qq\_line((aes(sample = adqvec)))+  
 labs(title = "QQ-Plot for normality with negative CKD status")



The hypotheses for this test are:

We will use a level of significance of

var.test(inadqvec, adqvec, two.sided = TRUE)

##   
## F test to compare two variances  
##   
## data: inadqvec and adqvec  
## F = 1.9247, num df = 206, denom df = 60, p-value = 0.003524  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 1.248388 2.832656  
## sample estimates:  
## ratio of variances   
## 1.924709

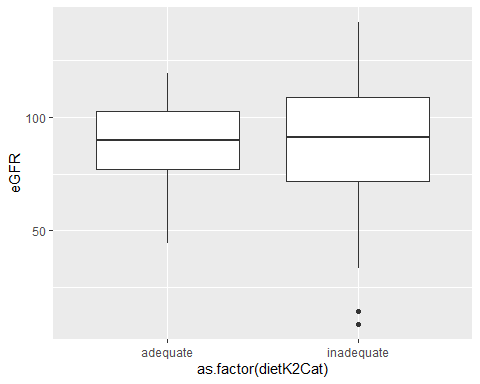
Since the variance test for equality produced a significant p-value, we reject the null hypothesis and can conclude that the variances are unequal.

### T-test

Now that we have assumed unequal variance, we must use a Satterthwaite approximation to approximate the degrees of freedom for our t distribution.

#### Step 1: Know the data

combined <- dietAdq %>% rbind(dietInadq)  
ggplot(combined, aes(x = as.factor(dietK2Cat), y = eGFR))+  
 geom\_boxplot()



We assume that our populations are both normally distributed with unknown mean and variances, and unequal variance.

#### Step 2: Set up Statistical Hypotheses

Where and and This test will use a level of significance of

#### Step 3: Choose a Test Statistic and Determine Sampling Distribution

#### Step 4: Calculate the Test Statistic and P-value

mytest2 <- t.test(inadqvec, adqvec, var.equal = FALSE)

The test statistic from our test is 0.7637319, and the p-value is 0.4463623.

#### Step 5: Draw Inference Based on Significance Level

Based on the results of our test, we fail to reject the null hypothesis that the difference in mean eGFR between individuals with inadequate dietary K and adequate dietary K is 0. This means that there is no significant difference in eGFR levels when partitioned on inadequate/adequate dietary K.

These results do not align with my original hypotheses, as I hypothesized that higher levels of dietary K is associated with lower risk of kidney function. Our results tell us there is no association.

#### Confidence Interval

The confidence interval for the true difference in mean eGFR between individuals with inadequate and adequate dietary K is [-3.4501637, 7.7912028. As 0 is included in this interval, we can conclude there is no difference in mean eGFR between individuals with inadequate and adequate dietary K.