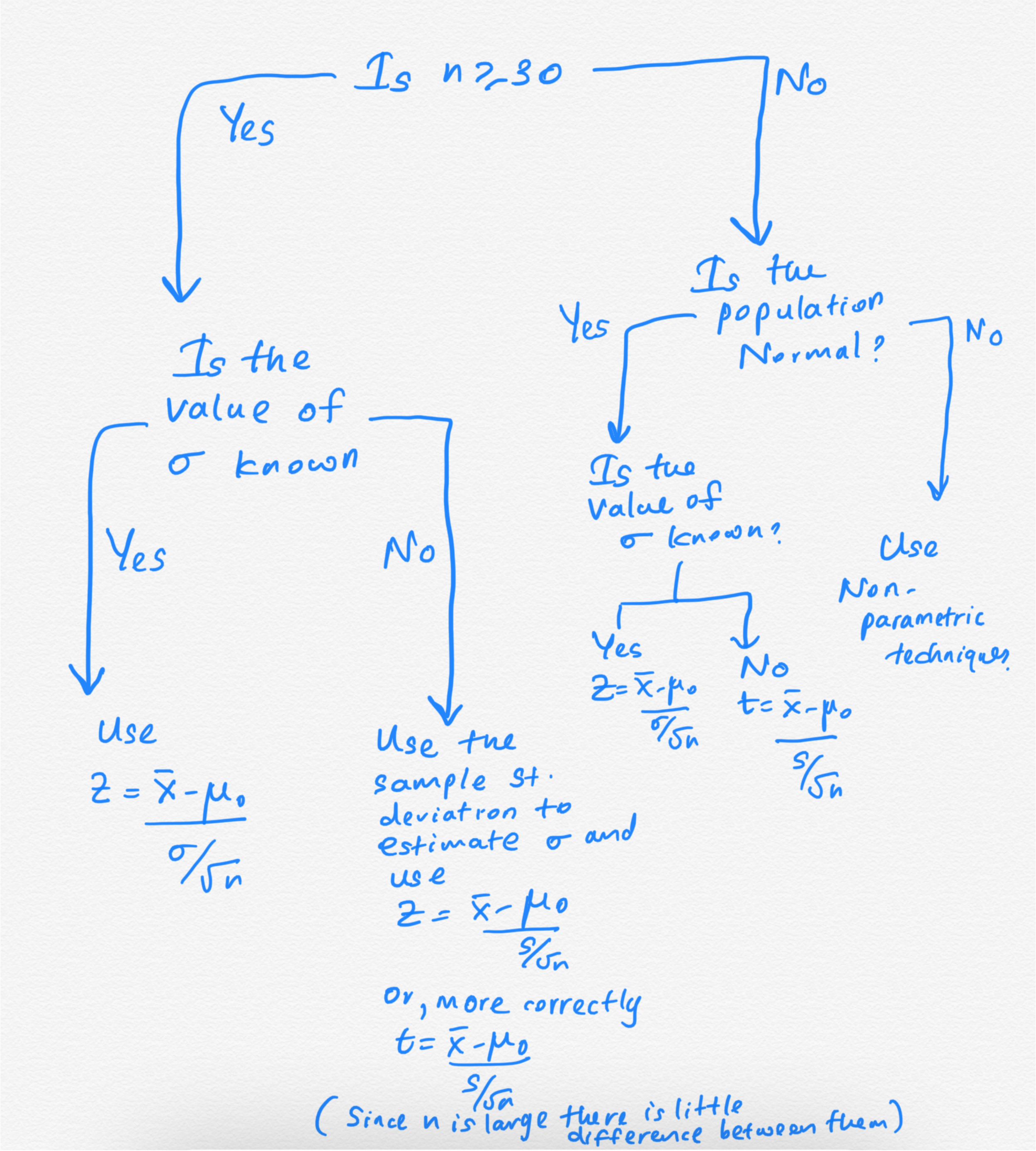
Hypothesis Testing: Comparing Means

Dr Thiyanga Talagala

Contents

1. One Sample - mean


Is n ≥ 30
Yes
No
Is the value of σ known
Yes
No
Use
$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
Use the sample st. deviatron to estimate σ and use
$z = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
or, more correctly
$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
(Since n is large there is little difference between them)
Is the population Normal?
Yes
No
Is the value of σ known?
Yes
$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
No
$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
Use Non-parametric techniques

1.2 Parametric

1.2.1 Z-test (σ known)

As reported by the US National Centre for Health Statistics, the mean serum high density (HDL) cholesterol of female 20 - 29 years old is 53. Dr Jack Hall claims that the HDL Cholesterol level of female 20 - 29 years old is greater than 53. He uses the following data, randomly gathered from 22 individuals.

```
HDL <- c(65, 47, 51, 54, 70, 55, 44, 48, 36, 53, 45, 34, 59, 45, 54, 50, 40, 60, 53, 53, 54, 55)
```

It is known from past research that the distribution of the HDL cholesterol is normally distributed and the corresponding population variance is 81. Test the claim that the HDL level is greater than 53 at $\alpha = 0.01$ level of significance.

```
HDL.df <- data.frame(HDL=HDL)
ggplot(HDL.df, aes(y=HDL, x="")) +
  geom_boxplot(outlier.shape = NA, fill="forestgreen", alpha=0.5) +
  geom_jitter(alpha=0.5) + labs(x = "")
```

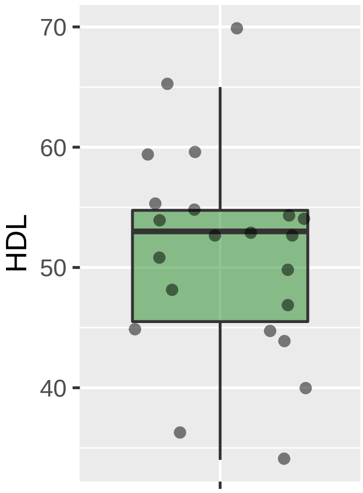


Figure 1: Distribution of HDL level

Hypothesis

H0:

H1:

μ -

```
z.test <- function(data, mu, var, alternative){
   z = (mean(data) - mu) / (sqrt(var / length(data)))
   if(alternative =="greater"){
     1-pnorm(z)

   } else if (alternative =="less"){

     pnorm(z)

   } else {

     pnorm(-1*abs(z)) * 2

   }

}
z.test(HDL.df$HDL, 53, 81,"greater")
```

```
[1] 0.8342875
```

Decision:

Conclusion:

1.2.2 t-test (σ unknown)

A chemist wants to measure the bias in a pH meter. She uses the meter to measure the pH in 14 neutral substances (pH=7) and obtains the data below.

```
ph <- c( 7.01, 7.04, 6.97, 7.00, 6.99, 6.97, 7.04, 7.04, 7.01, 7.00, 6.99, 7.04, 7.07, 6.97)
```

Is there sufficient evidence to support the claim that the pH meter is not correctly calibrated at the $\alpha = 0.05$ level of significance?

Answer:

```
ph.df <- data.frame(pH=ph)
ggplot(ph.df, aes(y=pH, x="")) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha=0.5) +
  labs(x = "")
```

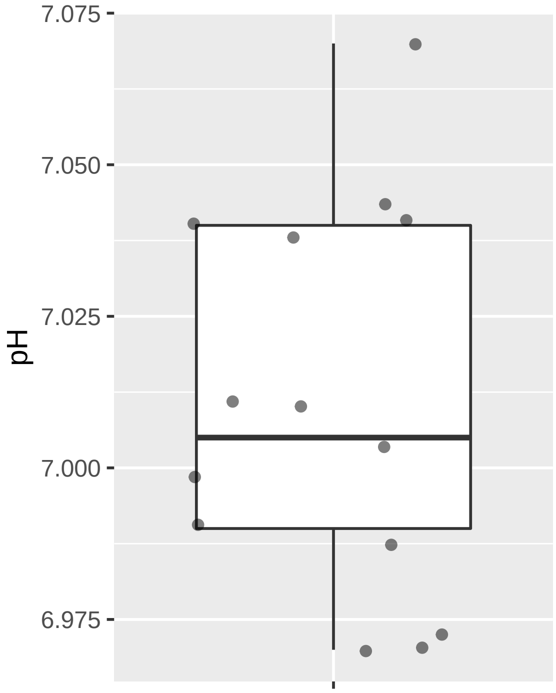


Figure 2: Distribution of pH values

In this case, we have only sixteen observations, meaning that the Central Limit Theorem does not apply. With a small sample, we should only use the t-test if we can reasonably assume that the population is normally distributed. Hence, we must first verify that pH is normally distributed.

```
ggplot(ph.df,
       aes(sample=pH))+
  stat_qq() + stat_qq_line()+labs(x="Theoretical Quantiles", y="Sample Quantiles")
```

```
shapiro.test(ph.df$pH)
```

```
    Shapiro-Wilk normality test

data:  ph.df$pH
W = 0.91603, p-value = 0.1927
```

Hypothesis to be tested:

H0: Data are normally distributed.

H1: Data are not normally distributed.

According to the Shapiro-Wilk normality test p-value, 0.19 > 0.05. Hence, we do not reject H0 at the 0.05 level of significance. We can conclude data are normally distributed.

Now we can proceed with the t.test.

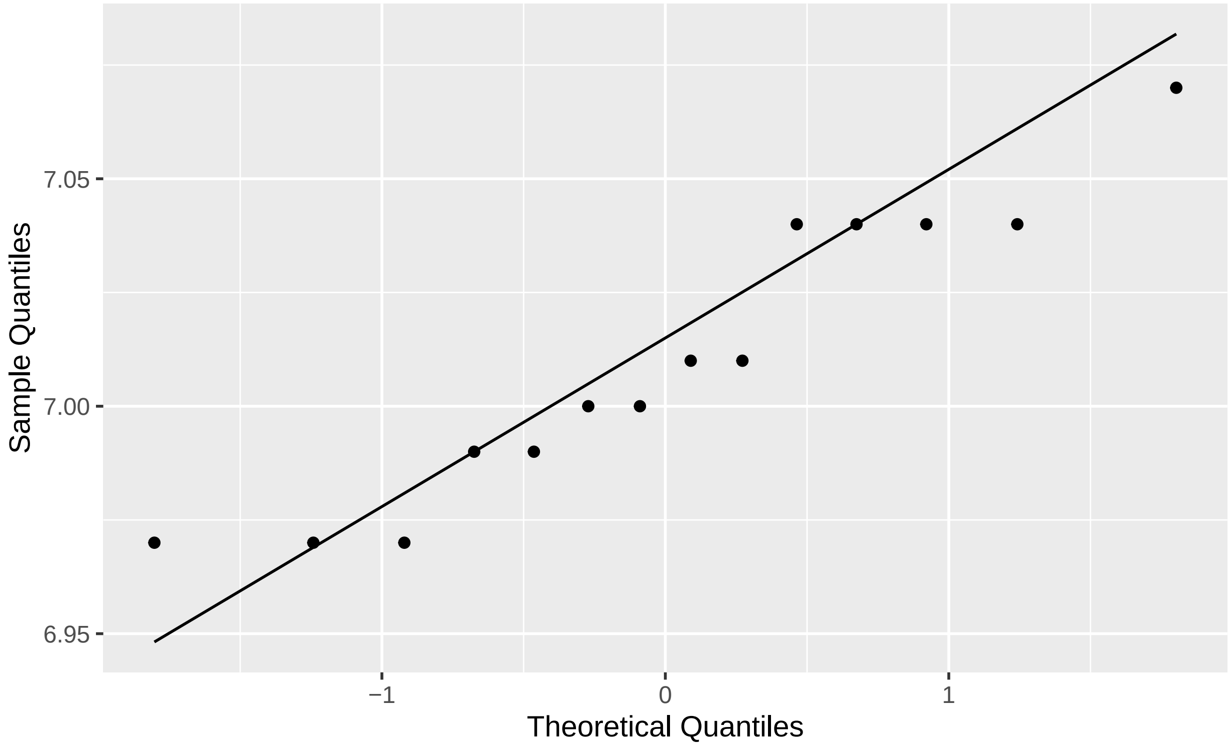


Figure 3: Normal probability plot of pH values

Hypothesis to be tested.

H0: $\mu = 7$

H1: $\mu \neq 7$

μ - Population mean pH value (in neutral substances).

`t.test` syntax

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

```
t.test(ph.df$pH, alternative = "two.sided", mu=7)
```

```

    One Sample t-test

data:  ph.df$pH
t = 1.1832, df = 13, p-value = 0.2579
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.991742 7.028258
sample estimates:
mean of x
     7.01
```

Decision: p-value (0.258) > $\alpha = 0.05$. Hence, we do not reject Ho.

Conclusion: We do not have enough evidence to conclude that the population mean pH level is different from 7 at the 0.05 level of significance.

2. Two sample - mean

2.1 Dependent (paired)

Approach 1

A dietician hopes to reduce a person's cholesterol level by using a special diet supplemented with a combination of vitamin pills. Twenty (20) subjects were pre-tested and then placed on diet for two weeks. Their cholesterol levels were checked after the two week period. The results are shown below. Cholesterol levels are measured in milligrams per decilitre.

i) Test the claim that the Cholesterol level before the special diet is greater than the Cholesterol level after the special diet at $\alpha = 0.01$ level of significance.

ii) Construct 99% confidence interval for the difference in mean cholesterol levels. Assume that the cholesterol levels are normally distributed both before and after.

```
id <- 1:20
before <- c(210, 235, 208, 190, 172, 244, 211, 235, 210,
            190, 175, 250, 200, 270, 222, 203, 209, 220, 250, 280)
after <- c(190, 170, 210, 188, 173, 195, 228, 200, 210, 184,
           196, 208, 211, 212, 205, 221, 240, 250, 230, 220)
cholesterol_1 <- data.frame(id=id, before=before, after=after)
head(cholesterol_1)
```

```
  id before after
1  1    210   190
2  2    235   170
3  3    208   210
4  4    190   188
5  5    172   173
6  6    244   195
```

```
cholesterol_2 <- pivot_longer(cholesterol_1, before:after, "type", "value")
head(cholesterol_2)
```

```
# A tibble: 6 x 3
     id type   value
  <int> <chr>  <dbl>
1     1 before   210
2     1 after    190
3     2 before   235
4     2 after    170
5     3 before   208
6     3 after    210
```

```
ggplot(data= cholesterol_2, aes(x=type, y=value)) +
  geom_boxplot(outlier.shape = NA, aes(fill=type), alpha=0.5) +
  geom_jitter(aes(fill=type))
```

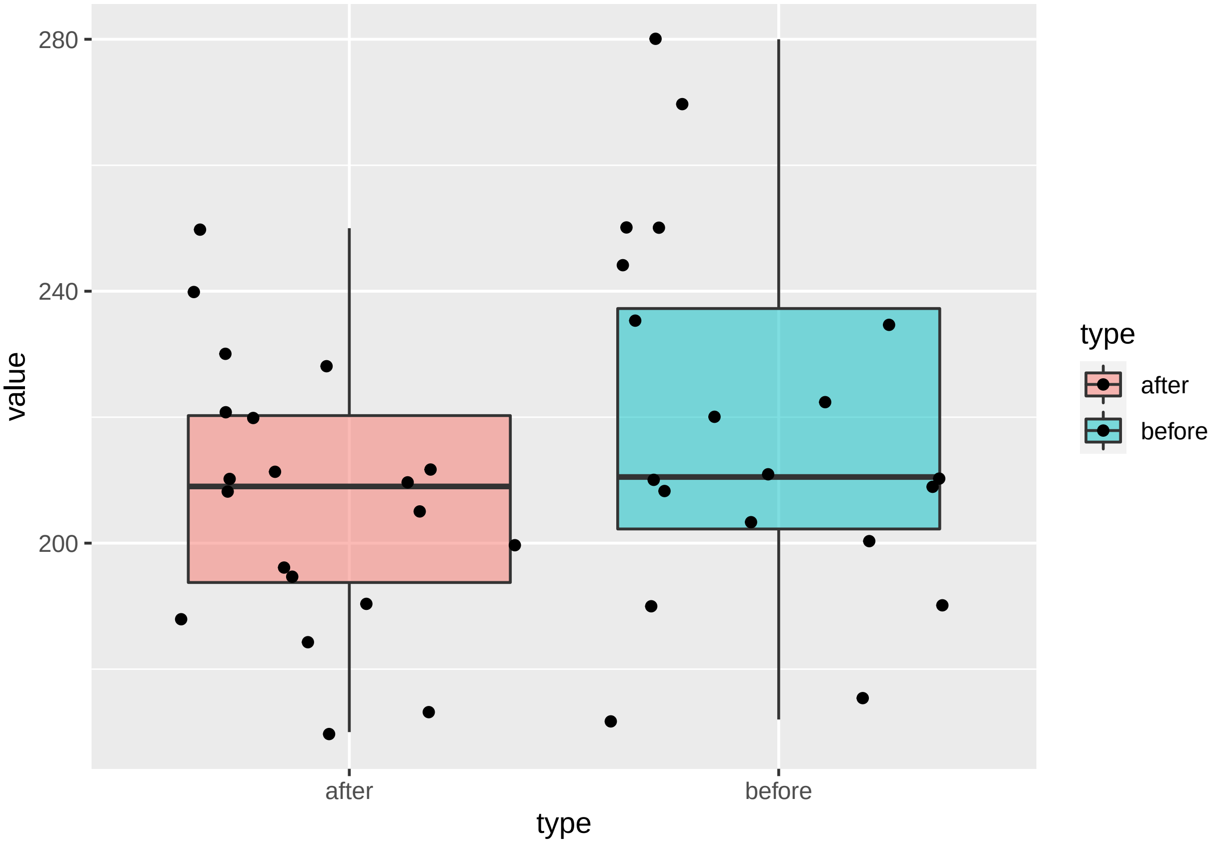


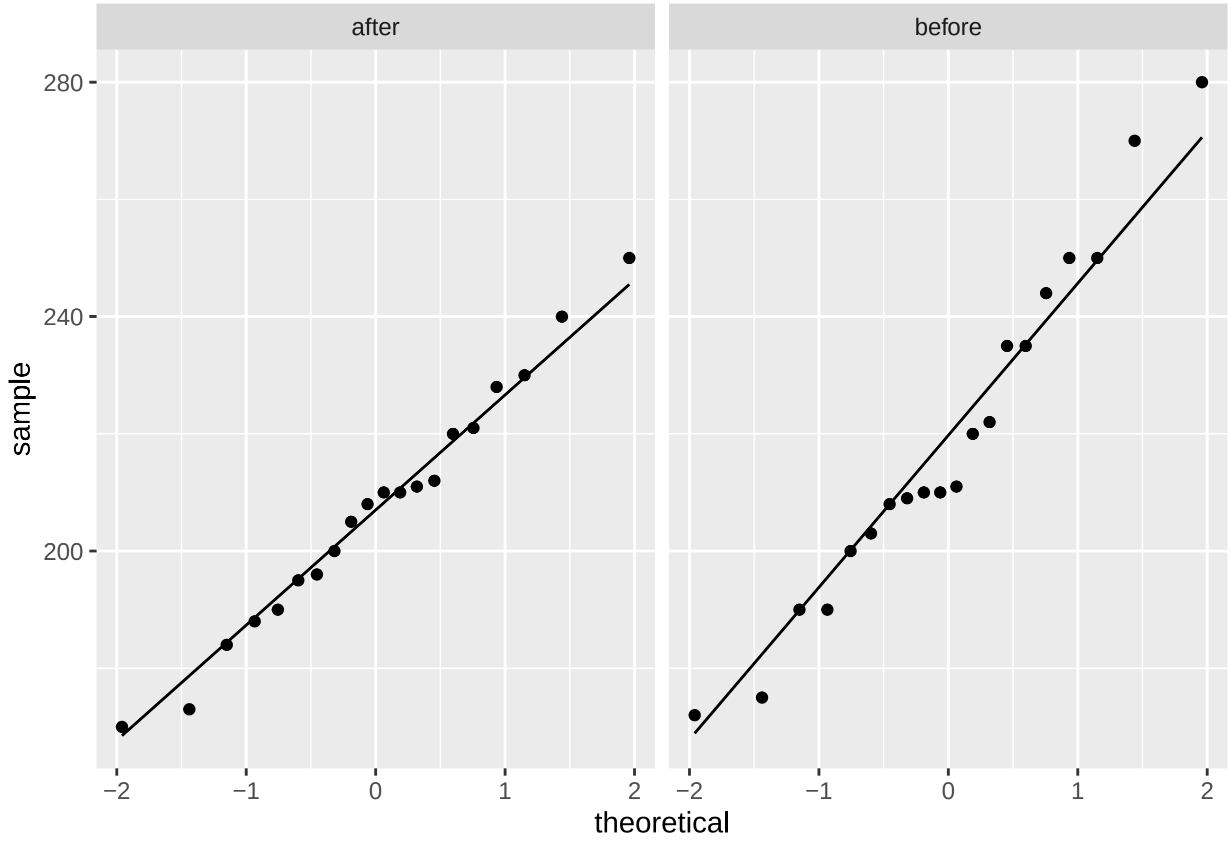
Figure 4: Distribution of cholesterol levels after and before the special diet

```
cholesterol_2 %>%
  group_by(type) %>%
  summarize(mean = round(mean(value), 2),
            sd = round(sd(value), 2))
```

```
# A tibble: 2 x 3
  type    mean    sd
  <chr>  <dbl> <dbl>
1 after   207.  21.0
2 before  219.  29.3
```

2.1.1 Testing for Normality

```
ggplot(data = cholesterol_2, aes(sample = value)) +
  stat_qq() +
  stat_qq_line() +
  facet_grid(. ~ type)
```



2.1.2 Paired t-test

Hypothesis:

H0: $\mu_{before} \leq \mu_{after}$

H1: $\mu_{before} > \mu_{after}$

μ_{before} - population mean cholesterol level before the special diet

μ_{after} - population mean cholesterol level after the special diet

```
t.test(before, after, data=cholesterol_1, "greater", paired=TRUE)
```

```

	Paired t-test

data:  before and after
t = 1.7754, df = 19, p-value = 0.04593
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3167385       Inf
sample estimates:
mean of the differences 
                  12.15 
```

Decision: ______________________________

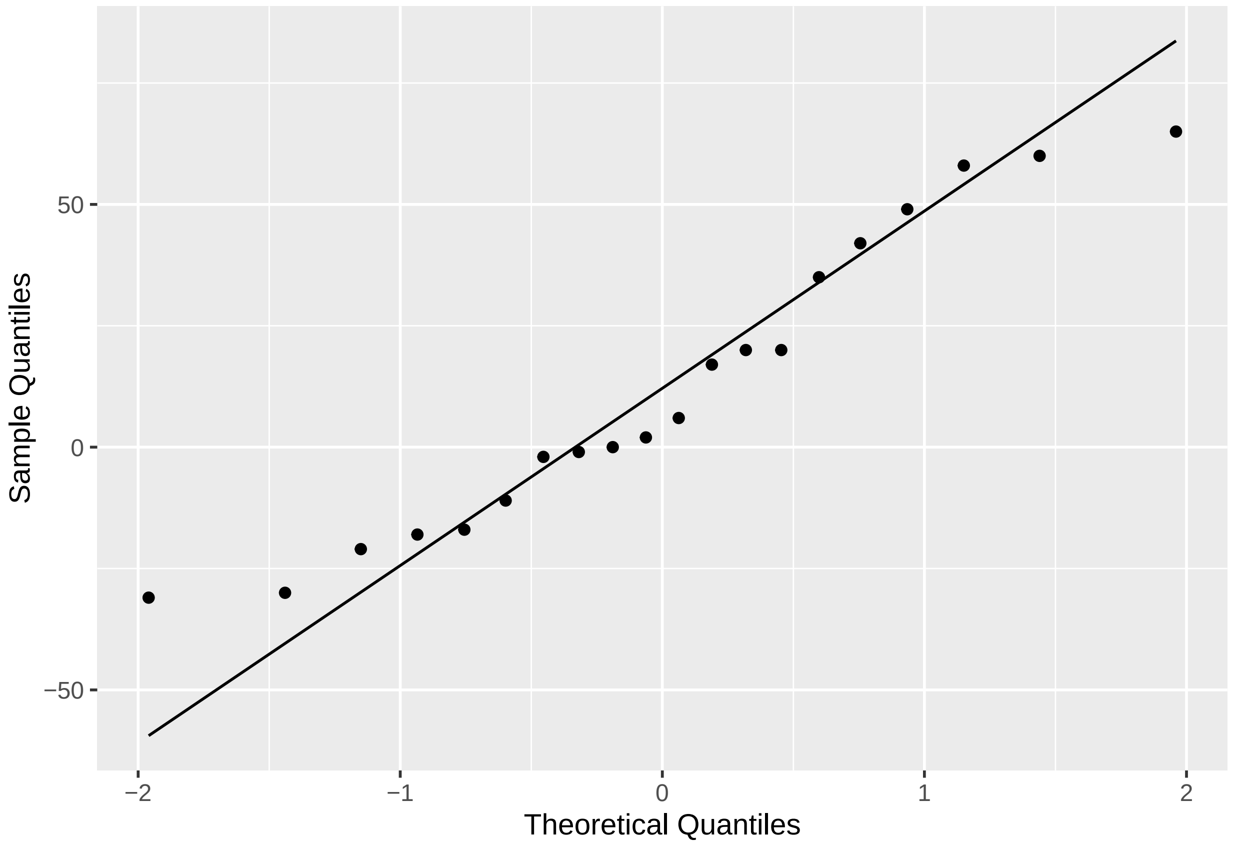
Conclusion: ____________________________

Approach 2

```
approach2_tbl <- tibble(diff = cholesterol_1$before - cholesterol_1$after)
```

2.1.3 Testing for Normality

```
ggplot(approach2_tbl,
       aes(sample=diff))+
  stat_qq() + stat_qq_line()+
  labs(x="Theoretical Quantiles", y="Sample Quantiles")
```



```
shapiro.test(approach2_tbl$diff)
```

```

    Shapiro-Wilk normality test

data:  approach2_tbl$diff
W = 0.93729, p-value = 0.213
```

H0: $\mu_d \leq 0$

H0: $\mu_d > 0$,

where: $\mu_d = \mu_{before} - \mu_{after}$

```r
t.test(x = approach2_tbl$diff,  alternative = c("greater"), mu=0)
```

```

	One Sample t-test

data:  approach2_tbl$diff
t = 1.7754, df = 19, p-value = 0.04593
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.3167385       Inf
sample estimates:
mean of x 
    12.15 
```

Decision: ______________

Conclusion: ___________________

2.1.4 Confidence intervals

To obtain confidence intervals

```r
t.test(before, after, data=cholesterol_1, "two.sided", paired=TRUE)
```

```

	Paired t-test

data:  before and after
t = 1.7754, df = 19, p-value = 0.09185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.173539 26.473539
sample estimates:
mean of the differences 
                  12.15 
```

95% CI for $\mu_{before} - \mu_{after}$: ____

```r
t.test(before, after, data=cholesterol_1, "two.sided", paired=TRUE, conf.level = 0.99)
```

```

	Paired t-test

data:  before and after
t = 1.7754, df = 19, p-value = 0.09185
alternative hypothesis: true difference in means is not equal to 0
```

```
99 percent confidence interval:
 -7.428709 31.728709
sample estimates:
mean of the differences
                  12.15
```

99% CI for $\mu_{before} - \mu_{after}$: ______________________________________

2.2 Independent

```
birthwt <- as_tibble(MASS::birthwt)
head(birthwt)
```

```
# A tibble: 6 x 10
    low   age   lwt  race smoke   ptl    ht    ui   ftv   bwt
  <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
1     0    19   182     2     0     0     0     1     0  2523
2     0    33   155     3     0     0     0     0     3  2551
3     0    20   105     1     1     0     0     0     1  2557
4     0    21   108     1     1     0     0     1     2  2594
5     0    18   107     1     1     0     0     1     0  2600
6     0    21   124     3     0     0     0     0     0  2622
```

```
?birthwt
```

smoke: smoking status during pregnancy.

(0 - No, 1 - Yes)

Is there a significant difference in birth weights between mothers who smoked during pregnancy and those who did not?

Data Wrangling

```
birthwt <- as_tibble(MASS::birthwt)

# Rename variables
birthwt <- birthwt %>%
  rename(smoking.status = smoke,
         birthwt.grams = bwt)

# Change factor level names
birthwt <- birthwt %>%
  mutate_at(c("smoking.status"),
            ~ recode_factor(.x, `0` = "no", `1` = "yes"))
head(birthwt)
```

```
# A tibble: 6 x 10
    low   age   lwt  race smoking.status   ptl    ht    ui   ftv birthwt.grams
  <int> <int> <int> <int> <fct>          <int> <int> <int> <int>         <int>
1     0    19   182     2 no                 0     0     1     0          2523
2     0    33   155     3 no                 0     0     0     3          2551
```

```
3      0     20    105      1 yes                    0      0      0      1          2557
4      0     21    108      1 yes                    0      0      1      2          2594
5      0     18    107      1 yes                    0      0      1      0          2600
6      0     21    124      3 no                     0      0      0      0          2622
```

```
ggplot(birthwt, aes(x=smoking.status, y=birthwt.grams))+
  geom_boxplot(outlier.shape=NA, aes(fill=smoking.status), alpha=0.05) +
  geom_jitter(aes(colour=smoking.status)) +
  scale_colour_manual(values = c("#d95f02", "#7570b3"))
```

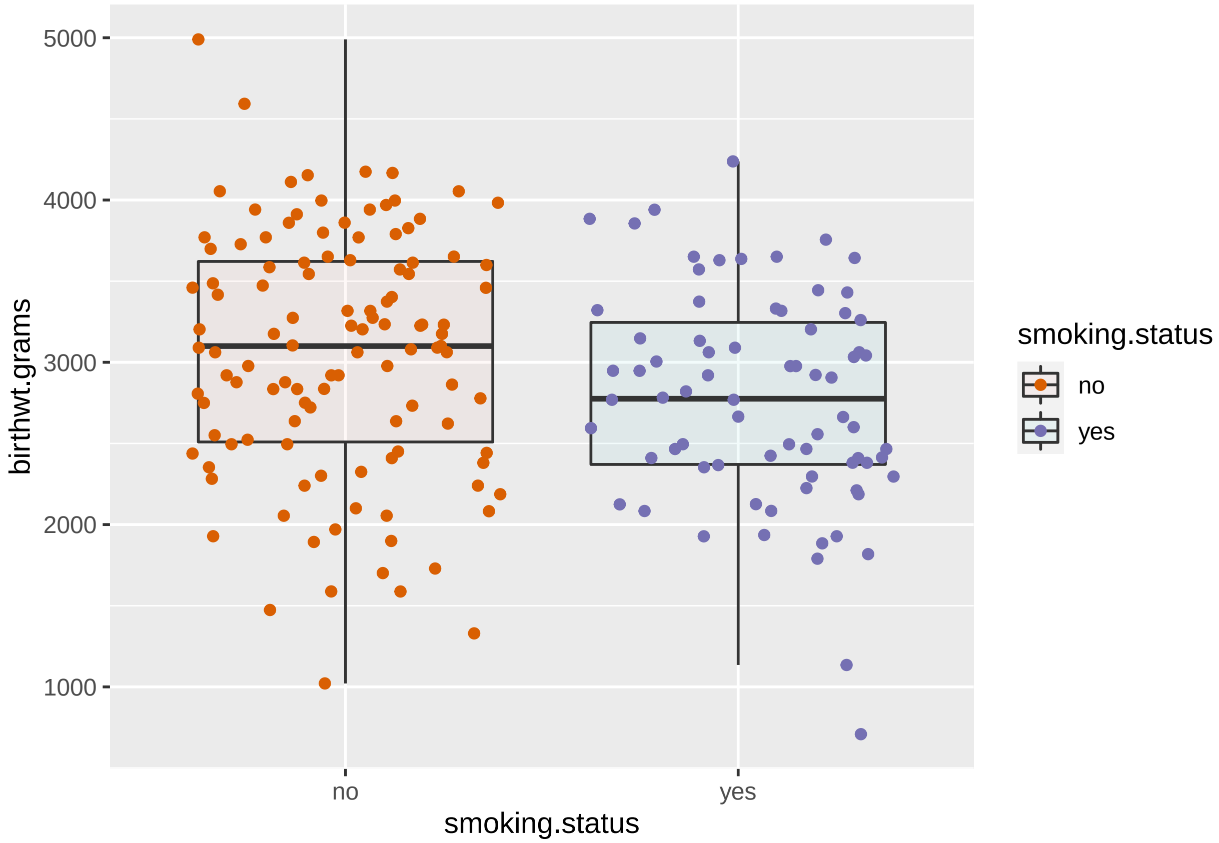


Figure 5: Distribution of infants birth weight by mothers' smoking status

```
birthwt %>%
  group_by(smoking.status) %>%
  summarize(mean = round(mean(birthwt.grams), 1),
            sd = round(sd(birthwt.grams), 1),
            max. = round(max(birthwt.grams), 1),
            min = round(min(birthwt.grams), 1),
            missing= sum(is.na(birthwt.grams)),
            count= sum(is.na(birthwt.grams)==FALSE))
```

```
# A tibble: 2 x 7
  smoking.status  mean    sd  max.   min missing count
  <fct>          <dbl> <dbl> <dbl> <dbl>   <int> <int>
```

```
1 no              3056.  753.  4990  1021       0   115
2 yes             2772.  660.  4238   709       0    74
```

```
se <- function(data){
  sd(data)/sqrt(length(data))
}
```

```
birthwt %>%
  group_by(smoking.status) %>%
  summarize(mean = round(mean(birthwt.grams), 1),
            sd = round(sd(birthwt.grams), 1),
            max. = round(max(birthwt.grams), 1),
            min = round(min(birthwt.grams), 1),
            missing= sum(is.na(birthwt.grams)),
            count= sum(is.na(birthwt.grams)==FALSE),
            se = se(birthwt.grams))
```

```
# A tibble: 2 x 8
  smoking.status  mean    sd  max.   min missing count    se
  <fct>          <dbl> <dbl> <dbl> <dbl>   <int> <int> <dbl>
1 no              3056.  753.  4990  1021       0   115  70.2
2 yes             2772.  660.  4238   709       0    74  76.7
```

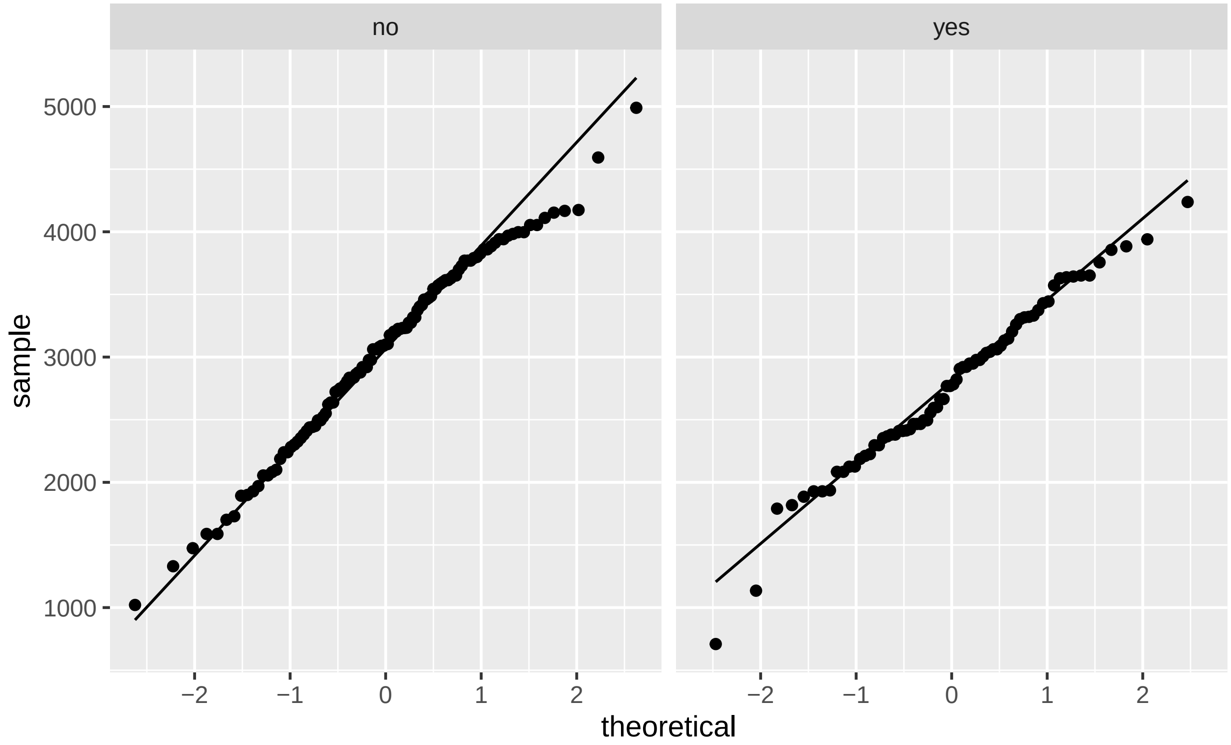
```
birthwt %>%
  group_by(smoking.status) %>%
  summarize(num.obs = n(),
            mean.birthwt = round(mean(birthwt.grams), 0),
            sd.birthwt = round(sd(birthwt.grams), 0),
            se.birthwt = round(sd(birthwt.grams) / sqrt(num.obs), 0))
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

```
# A tibble: 2 x 5
  smoking.status num.obs mean.birthwt sd.birthwt se.birthwt
  <fct>            <int>        <dbl>      <dbl>      <dbl>
1 no                 115         3056        753         70
2 yes                 74         2772        660         77
```

2.2.1 Testing for Normality

```
 ggplot(data = birthwt, aes(sample = birthwt.grams)) +
  stat_qq() +
  stat_qq_line() +
  facet_grid(. ~ smoking.status)
```



```
mother_yes_birthwt <- birthwt %>% filter(smoking.status=="yes")
dim(mother_yes_birthwt)
```

```
[1] 74 10
```

```
shapiro.test(mother_yes_birthwt$birthwt.grams)
```

```

    Shapiro-Wilk normality test

data:  mother_yes_birthwt$birthwt.grams
W = 0.98296, p-value = 0.4195
```

Hypothesis:

H0:

H1:

Decision: ____________________

Conclusion: _____________________

```
mother_no_birthwt <- birthwt %>% filter(smoking.status=="no")
dim(mother_no_birthwt)
```

```
[1] 115  10
```

```r
shapiro.test(mother_no_birthwt$birthwt.grams)
```

```
    Shapiro-Wilk normality test

data:  mother_no_birthwt$birthwt.grams
W = 0.98694, p-value = 0.3337
```

Hypothesis:

H0:

H1:

Decision: ___________________

Conclusion: ____________________

2.2.2 Equality of variance

The equality of variances between two samples can be tested using the F test.

Hypothesis:

H0: _________________

H1: _________________

σ_1^2 -

σ_2^2 -

```r
var.test(birthwt.grams ~ smoking.status, data = birthwt,
         alternative = "two.sided")
```

```
    F test to compare two variances

data:  birthwt.grams by smoking.status
F = 1.3019, num df = 114, denom df = 73, p-value = 0.2254
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8486407 1.9589574
sample estimates:
ratio of variances
          1.301927
```

2.2.3 How can we assess whether the mean difference is statistically significant?

Hypothesis

H0: ______________

H1: ______________

where,

μ_1 -

μ_2 -

```r
t.test(birthwt.grams ~ smoking.status, data = birthwt,
       alternative = c("two.sided"),
       var.equal = TRUE)
```

```

    Two Sample t-test

data:  birthwt.grams by smoking.status
t = 2.6529, df = 187, p-value = 0.008667
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  72.75612 494.79735
sample estimates:
 mean in group no mean in group yes 
         3055.696          2771.919 
```

3. Other test functions

- `fisher.test` Fisher's exact test for counts
- `t.test(data)` 1 sample t test
- `t.test(data1,data2)` 2 sample t test
- `t.test(pre,post,paired=TRU E)` paired sample t test
- `wilcox.test(data)` Wilcox test
- `cor.test(data1,data2)` correlation test
- `chisq.test(data)` Chi square test
- `shapiro.test(data)` Shapiro test
- `aov()` ANOVA
- etc.