Hypothesis Testing

Ishana Rana

2025-07-10

1) Loading the datasets

```
bertscore <- read.csv("C:\\Users\\babus\\OneDrive\\Documents\\uni
uzh\\FS25\\conversational speech processing\\mypaper\\Beyond-WER-in-
ASR\\data\\eval_results\\BERTScore_scores.csv", skip=1)
str(bertscore)
```

```
## 'data.frame':    6 obs. of  3 variables:
##  $ File         : chr  "EN2009c" "EN2009d" "ES2016a" "ES2016b" ...
##  $ with_punct   : num  84.7 88 84 91.1 94 ...
##  $ without_punct: num  86 89.3 86 93.7 95.8 ...
```

```
head(bertscore)
```

```
##      File with_punct without_punct
## 1 EN2009c      84.66         85.98
## 2 EN2009d      87.98         89.34
## 3 ES2016a      84.01         86.02
## 4 ES2016b      91.10         93.74
## 5 ES2016c      93.98         95.83
## 6 ES2016d      89.66         90.12
```

```
summary(bertscore)
```

```
##      File             with_punct    without_punct
##  Length:6           Min.   :84.01   Min.   :85.98
##  Class :character   1st Qu.:85.49   1st Qu.:86.85
##  Mode  :character   Median :88.82   Median :89.73
##                     Mean   :88.56   Mean   :90.17
##                     3rd Qu.:90.74   3rd Qu.:92.83
##                     Max.   :93.98   Max.   :95.83
```

```
bleu <- read.csv("C:\\Users\\babus\\OneDrive\\Documents\\uni
uzh\\FS25\\conversational speech processing\\mypaper\\Beyond-WER-in-
ASR\\data\\eval_results\\BLEU_scores.csv")
str(bleu)
```

```
## 'data.frame':    6 obs. of  3 variables:
##  $ File         : chr  "EN2009c" "EN2009d" "ES2016a" "ES2016b" ...
##  $ with_punct   : num  38.8 36.9 43.6 54.3 50 ...
##  $ without_punct: num  38.1 38.5 49.1 59.4 55.1 ...
```

```r
head(bleu)
```

```
##      File with_punct without_punct
## 1 EN2009c      38.77         38.14
## 2 EN2009d      36.93         38.54
## 3 ES2016a      43.61         49.07
## 4 ES2016b      54.26         59.40
## 5 ES2016c      49.96         55.11
## 6 ES2016d      30.65         31.81
```

```r
summary(bleu)
```

```
##      File             with_punct    without_punct
##  Length:6           Min.   :30.65   Min.   :31.81
##  Class :character   1st Qu.:37.39   1st Qu.:38.24
##  Mode  :character   Median :41.19   Median :43.80
##                     Mean   :42.36   Mean   :45.34
##                     3rd Qu.:48.37   3rd Qu.:53.60
##                     Max.   :54.26   Max.   :59.40
```

```r
# splitting dataset as it contains scores of ROUGE-1 and ROUGE-L
lines <- readLines("C:\\Users\\babus\\OneDrive\\Documents\\uni
uzh\\FS25\\conversational speech processing\\mypaper\\Beyond-WER-in-
ASR\\data\\eval_results\\ROUGE_scores.csv")
split_index <- grep("ROUGE-L", lines)

rouge1_lines <- lines[2:(split_index - 1)]
rougel_lines <- lines[(split_index + 1):length(lines)]

rouge1 <- read.csv(text = rouge1_lines)
rougel <- read.csv(text = rougel_lines)

str(rouge1)
```

```
## 'data.frame':    6 obs. of  3 variables:
##  $ File         : chr  "EN2009c" "EN2009d" "ES2016a" "ES2016b" ...
##  $ with_punct   : num  74.9 72.9 80.1 85 83 ...
##  $ without_punct: num  74.8 73 80 84.9 82.8 ...
```

```r
head(rouge1)
```

```
##      File with_punct without_punct
## 1 EN2009c      74.92         74.76
## 2 EN2009d      72.91         72.97
## 3 ES2016a      80.14         80.04
## 4 ES2016b      84.99         84.93
## 5 ES2016c      83.00         82.82
## 6 ES2016d      70.03         69.89
```

```r
summary(rouge1)
```

```
##      File             with_punct    without_punct
##  Length:6           Min.   :70.03   Min.   :69.89
##  Class :character   1st Qu.:73.41   1st Qu.:73.42
##  Mode  :character   Median :77.53   Median :77.40
##                     Mean   :77.67   Mean   :77.57
##                     3rd Qu.:82.28   3rd Qu.:82.12
##                     Max.   :84.99   Max.   :84.93
```

```
str(rougel)
```

```
## 'data.frame':    6 obs. of  3 variables:
##  $ File         : chr  "EN2009c" "EN2009d" "ES2016a" "ES2016b" ...
##  $ with_punct   : num  39.4 35.1 39.4 37.2 46.9 ...
##  $ without_punct: num  39.2 35.1 39.3 37.2 46.8 ...
```

```
head(rougel)
```

```
##      File with_punct without_punct
## 1 EN2009c      39.35         39.22
## 2 EN2009d      35.10         35.10
## 3 ES2016a      39.37         39.28
## 4 ES2016b      37.24         37.22
## 5 ES2016c      46.86         46.83
## 6 ES2016d      31.62         31.30
```

```
summary(rougel)
```

```
##      File             with_punct    without_punct
##  Length:6           Min.   :31.62   Min.   :31.30
##  Class :character   1st Qu.:35.63   1st Qu.:35.63
##  Mode  :character   Median :38.30   Median :38.22
##                     Mean   :38.26   Mean   :38.16
##                     3rd Qu.:39.37   3rd Qu.:39.27
##                     Max.   :46.86   Max.   :46.83
```

```
wer <- read.csv("C:\\Users\\babus\\OneDrive\\Documents\\uni
uzh\\FS25\\conversational speech processing\\mypaper\\Beyond-WER-in-
ASR\\data\\eval_results\\WER_scores.csv", skip=1)
str(wer)
```

```
## 'data.frame':    6 obs. of  3 variables:
##  $ File         : chr  "EN2009c" "EN2009d" "ES2016a" "ES2016b" ...
##  $ with_punct   : num  89.1 89.9 88.7 82.9 77 ...
##  $ without_punct: num  86.2 87.5 86.4 78.3 71.6 ...
```

```
head(wer)
```

```
##      File with_punct without_punct
## 1 EN2009c      89.12         86.16
## 2 EN2009d      89.94         87.55
## 3 ES2016a      88.73         86.38
## 4 ES2016b      82.87         78.33
```

```
## 5 ES2016c      77.02         71.61
## 6 ES2016d      86.95         83.21
```

```
summary(wer)
```

```
##      File             with_punct    without_punct
##  Length:6           Min.   :77.02   Min.   :71.61
##  Class :character   1st Qu.:83.89   1st Qu.:79.55
##  Mode  :character   Median :87.84   Median :84.69
##                     Mean   :85.77   Mean   :82.21
##                     3rd Qu.:89.02   3rd Qu.:86.33
##                     Max.   :89.94   Max.   :87.55
```

Statistical Testing for Difference

To assess whether punctuation and meeting type have a statistically significant effect on the metrics, we conduct two sets of tests: 1. Between conditions: with punctuation vs without punctuation 2. Between meeting types: scenario-based (scripted) vs natural speech (unscripted)

Given the small sample size (n = 6 meetings), the results must be interpreted with caution. Small samples increase the risk of both Type I and Type II errors and limit the generalizability of findings. To determine the appropriate statistical test, we first assess whether the normality assumption holds by applying the Shapiro–Wilk test and inspecting QQ plots. If the differences appear approximately normal, we proceed with parametric tests: the paired t-test for comparisons between conditions and the independent t-test for comparisons between meeting types (scenario vs. natural). If normality is violated, we instead use the corresponding non-parametric alternatives: the Wilcoxon signed-rank test for paired comparisons, and the Wilcoxon rank-sum test (Mann–Whitney U) for independent group comparisons.

These are the hypotheses for testing between conditions: - Null Hypothesis: There is no difference in the metric score between the 2 conditions - Alternative Hypothesis: There is a difference in the metric score between the 2 conditions These are the hypotheses for testing between meeting types: - Null Hypothesis: The mean of the metric score is the same for scenario and natural speech meetings - Alternative Hypothesis: The metric score has a significant difference between the two types All the tests are evaluated at an significance level of alpha = 5%.

WER

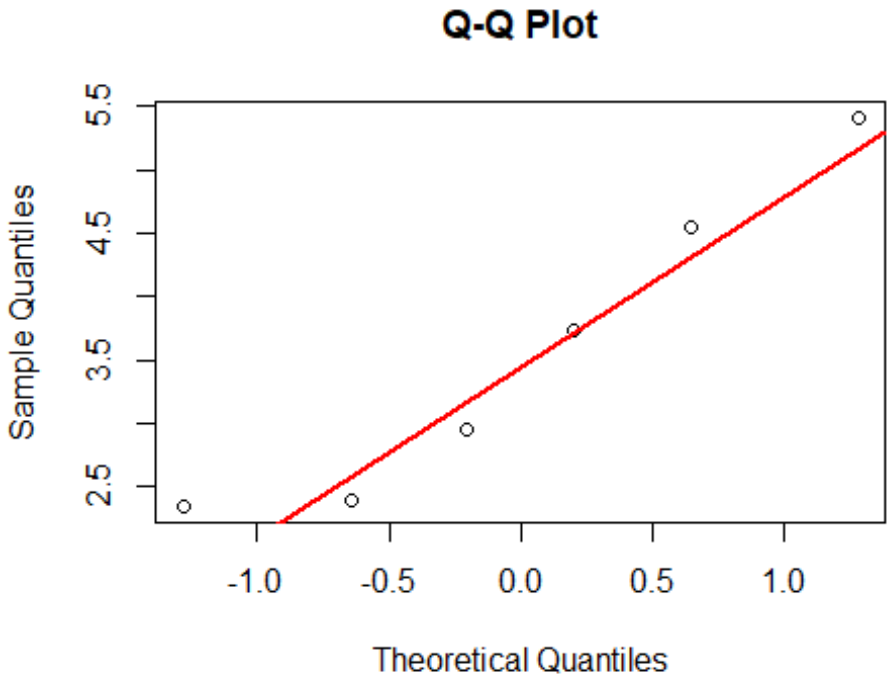
Testing between Conditions

Testing normality:

```
wer$diff <- wer$with_punct - wer$without_punct

shapiro_test <- shapiro.test(wer$diff)
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wer$diff
## W = 0.91597, p-value = 0.4768
```

```
qqnorm(wer$diff,
       main = "Q-Q Plot",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")

qqline(wer$diff, col = "red", lwd = 2)
```



As the p-value is greater than alpha, we fail to reject the Null Hypothesis. This means, normality can be assumed. This is supported by the QQ plot as there are no strong outliers or curvatures present. Therefore, we can proceed with the paired t-test:

Paired t-test:

```
t.test(wer$with_punct, wer$without_punct, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  wer$with_punct and wer$without_punct
## t = 7.0791, df = 5, p-value = 0.0008705
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
```

```
##  2.270475 4.859525
## sample estimates:
## mean difference
##           3.565
```

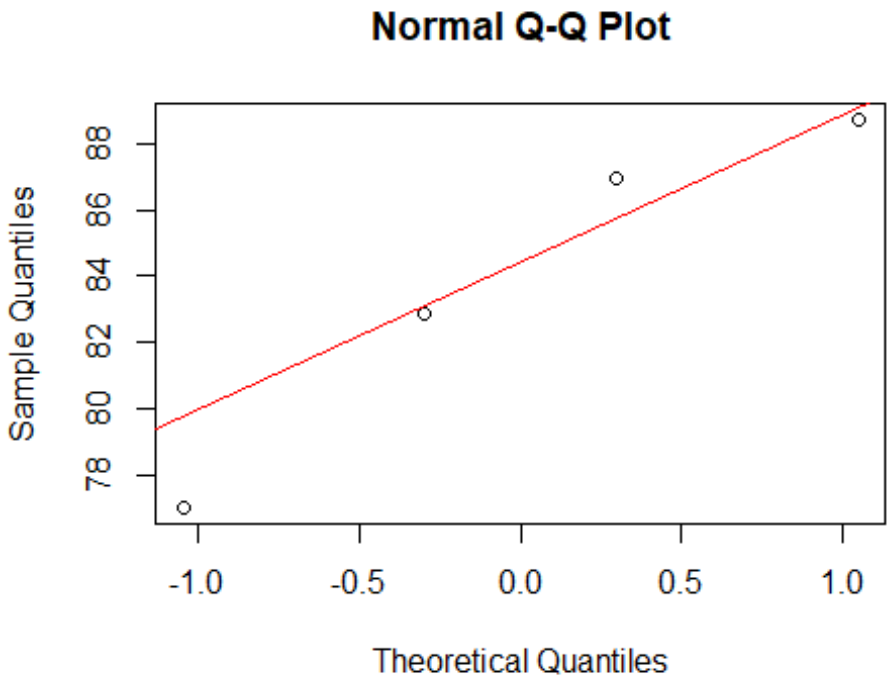
As the p-value is smaller than alpha, we can reject the Null Hypothesis. There IS a statistically difference between WER scores with and without punctuation.

Testing between Meeting types

Checking normality:

Since the sample size must be at least of size 3 for this test, we can only test it for the meeting type “scenario”.

```
wer$Type <- ifelse(grepl("^ES", wer$File), "scenario", "natural")

shapiro.test(wer$with_punct[wer$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  wer$with_punct[wer$Type == "scenario"]
## W = 0.93984, p-value = 0.6534

shapiro.test(wer$without_punct[wer$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  wer$without_punct[wer$Type == "scenario"]
## W = 0.96814, p-value = 0.8299

qqnorm(wer$with_punct[wer$Type == "scenario"])
qqline(wer$with_punct[wer$Type == "scenario"], col = "red")
```



For both cases, with and without punctuation, the normality is not violated as shown in the results. The p-value is larger than alpha, therefore, we fail to reject the Null Hypothesis. Therefore we will go ahead with the independent samples t-test:

```
# for with punctuation
t.test(with_punct ~ Type, data = wer)

##
##  Welch Two Sample t-test
##
## data:  with_punct by Type
## t = 2.143, df = 3.1454, p-value = 0.1173
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -2.519683 13.794683
## sample estimates:
##  mean in group natural mean in group scenario
##                89.5300                83.8925

# for without punctuation
t.test(without_punct ~ Type, data = wer)

##
##  Welch Two Sample t-test
##
## data:  without_punct by Type
## t = 2.119, df = 3.2653, p-value = 0.1169
```

```
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -3.03409 16.97909
## sample estimates:
##  mean in group natural mean in group scenario
##                86.8550                79.8825
```

As the p-values are greater than alpha, we fail to reject the null hypothesis in both cases. There is no statistically significant difference in WER scores between natural and scenario meetings with punctuation or without it. This result should be taken with caution, as the natural group only contained 2 samples. Therefore, the results do not hold much power.

BLEU

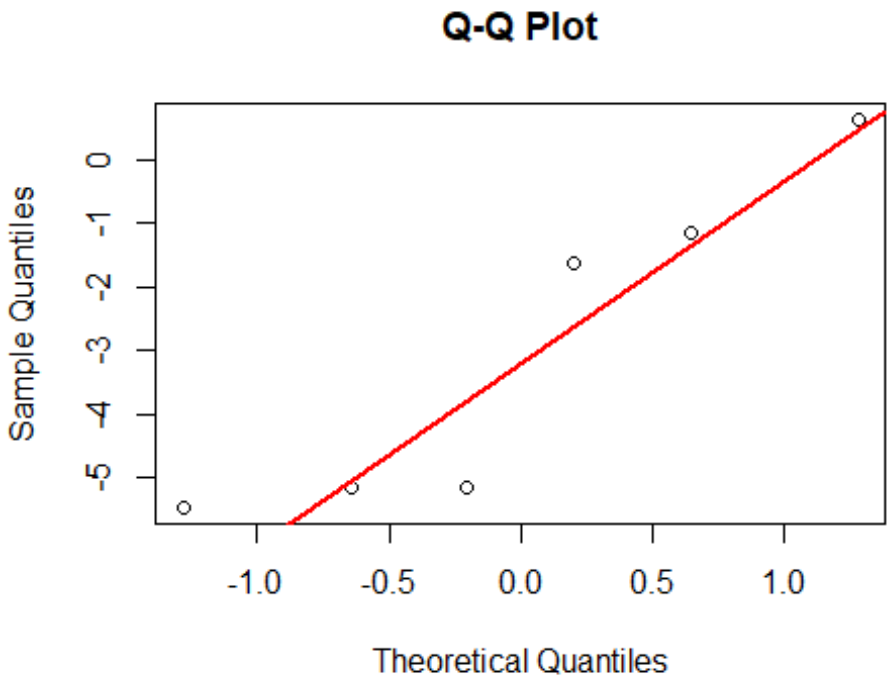
Testing between Conditions

Checking normality:

```r
bleu$diff <- bleu$with_punct - bleu$without_punct

shapiro_test <- shapiro.test(bleu$diff)
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bleu$diff
## W = 0.8473, p-value = 0.1497
```

```r
qqnorm(bleu$diff,
       main = "Q-Q Plot",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")

qqline(bleu$diff, col = "red", lwd = 2)
```



As the p-value is greater than alpha, we fail to reject the Null Hypothesis. This means, normality can be assumed. This is supported by the QQ plot as there are no strong outliers or curvatures present. Therefore, we can proceed with the paired t-test: ##### Paired t-test:

```
t.test(bleu$with_punct, bleu$without_punct, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  bleu$with_punct and bleu$without_punct
## t = -2.8113, df = 5, p-value = 0.03749
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -5.708030 -0.255303
## sample estimates:
## mean difference
##       -2.981667
```

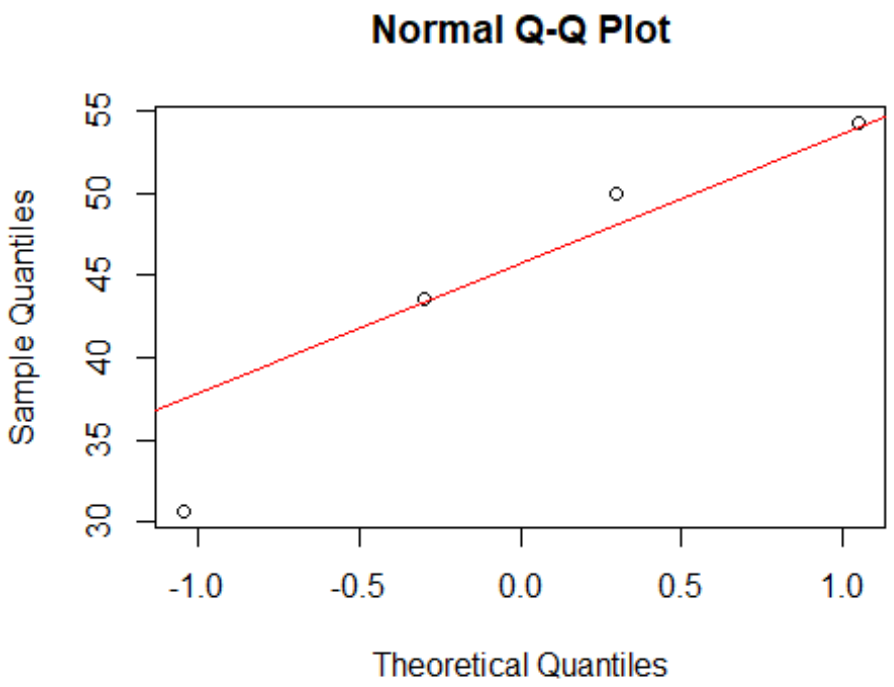
As the p-value is smaller than alpha, we can reject the Null Hypothesis. There IS a statistically difference between BLEU scores with and without punctuation.

Testing between Meeting types

Checking normality:

Since the sample size must be at least of size 3 for this test, we can only test it for the meeting type “scenario”.

```r
bleu$Type <- ifelse(grepl("^ES", bleu$File), "scenario", "natural")

shapiro.test(bleu$with_punct[bleu$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  bleu$with_punct[bleu$Type == "scenario"]
## W = 0.94037, p-value = 0.6566

shapiro.test(bleu$without_punct[bleu$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  bleu$without_punct[bleu$Type == "scenario"]
## W = 0.90418, p-value = 0.4521

qqnorm(bleu$with_punct[bleu$Type == "scenario"])
qqline(bleu$with_punct[bleu$Type == "scenario"], col = "red")
```



For both cases, with and without punctuation, the normality is not violated as shown in the results. The p-value is larger than alpha, therefore, we fail to reject the Null Hypothesis. Therefore we will go ahead with the independent samples t-test:

```r
# for with punctuation
t.test(with_punct ~ Type, data = wer)
```

```
##
##  Welch Two Sample t-test
##
## data:  with_punct by Type
## t = 2.143, df = 3.1454, p-value = 0.1173
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -2.519683 13.794683
## sample estimates:
##  mean in group natural mean in group scenario
##                89.5300                83.8925
```

```
# for without punctuation
t.test(without_punct ~ Type, data = wer)
```

```
##
##  Welch Two Sample t-test
##
## data:  without_punct by Type
## t = 2.119, df = 3.2653, p-value = 0.1169
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -3.03409 16.97909
## sample estimates:
##  mean in group natural mean in group scenario
##                86.8550                79.8825
```

As the p-values are greater than alpha, we fail to reject the null hypothesis in both cases. There is no statistically significant difference in BLEU scores between natural and scenario meetings with punctuation or without it. This result should be taken with caution, as the natural group only contained 2 samples. Therefore, the results do not hold much power.

ROUGE-1

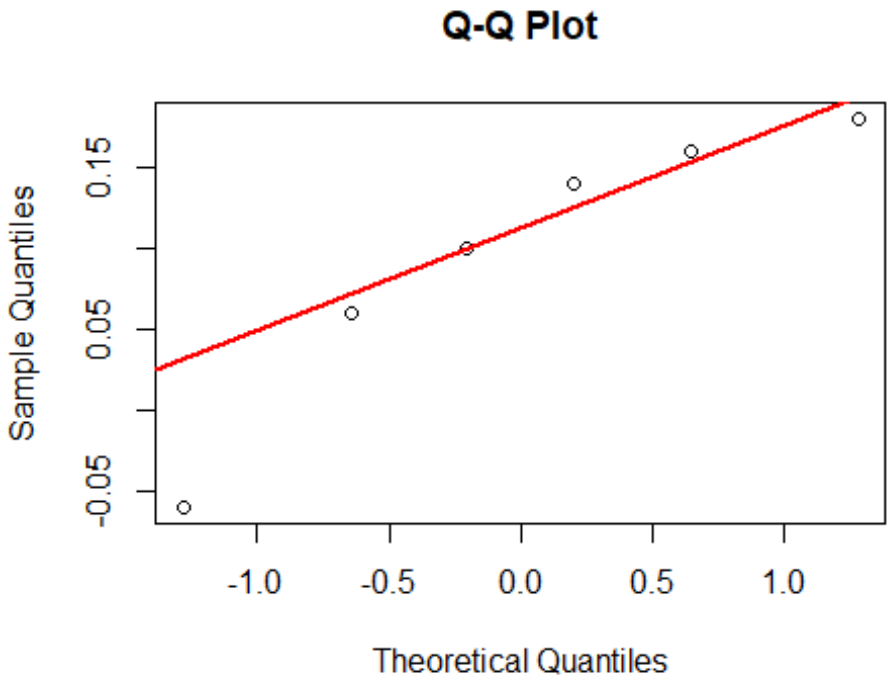
Testing between Conditions

Checking normality:

```
rouge1$diff <- rouge1$with_punct - rouge1$without_punct

shapiro_test <- shapiro.test(rouge1$diff)
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rouge1$diff
## W = 0.89236, p-value = 0.3308
```

```r
qqnorm(rouge1$diff,
       main = "Q-Q Plot",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")

qqline(rouge1$diff, col = "red", lwd = 2)
```



As the p-value is greater than alpha, we fail to reject the Null Hypothesis. This means, normality can be assumed. This is supported by the QQ plot as there are no strong outliers or curvatures present. Therefore, we can proceed with the paired t-test: ##### Paired t-test:

```r
t.test(rouge1$with_punct, rouge1$without_punct, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  rouge1$with_punct and rouge1$without_punct
## t = 2.6903, df = 5, p-value = 0.04328
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.004300431 0.189032902
## sample estimates:
## mean difference
##      0.09666667
```

As the p-value is smaller than alpha, we can reject the Null Hypothesis. There IS a statistically difference between ROUGE1 scores with and without punctuation. As the p-

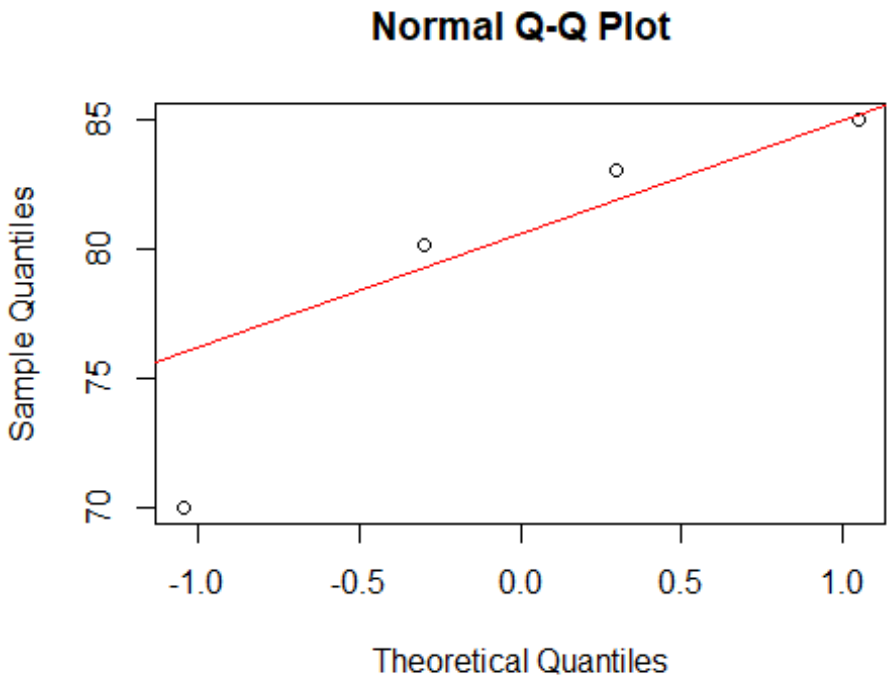
value is not that much smaller though, and the sample size is quite small, this result does not hold much power.

Testing between Meeting types

Checking normality:

Since the sample size must be at least of size 3 for this test, we can only test it for the meeting type “scenario”.

```
rouge1$Type <- ifelse(grepl("^ES", rouge1$File), "scenario", "natural")

shapiro.test(rouge1$with_punct[rouge1$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  rouge1$with_punct[rouge1$Type == "scenario"]
## W = 0.87351, p-value = 0.3116

shapiro.test(rouge1$without_punct[rouge1$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  rouge1$without_punct[rouge1$Type == "scenario"]
## W = 0.87598, p-value = 0.3218

qqnorm(rouge1$with_punct[rouge1$Type == "scenario"])
qqline(rouge1$with_punct[rouge1$Type == "scenario"], col = "red")
```



For both cases, with and without punctuation, the normality is not violated as shown in the results. The p-value is larger than alpha, therefore, we fail to reject the Null Hypothesis. Therefore we will go ahead with the independent samples t-test:

```
# for with punctuation
t.test(with_punct ~ Type, data = rouge1)

##
##  Welch Two Sample t-test
##
## data:  with_punct by Type
## t = -1.6205, df = 3.4865, p-value = 0.1908
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -15.849178   4.599178
## sample estimates:
##  mean in group natural mean in group scenario
##                 73.915                 79.540

# for without punctuation
t.test(without_punct ~ Type, data = rouge1)

##
##  Welch Two Sample t-test
##
## data:  without_punct by Type
## t = -1.6107, df = 3.3958, p-value = 0.195
```

```
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -15.841548   4.731548
## sample estimates:
##  mean in group natural mean in group scenario
##                  73.865                 79.420
```

As the p-values are greater than alpha, we fail to reject the null hypothesis in both cases. There is no statistically significant difference in ROUGE-1 scores between natural and scenario meetings with punctuation or without it. This result should be taken with caution, as the natural group only contained 2 samples. Therefore, the results do not hold much power.

ROUGE-L

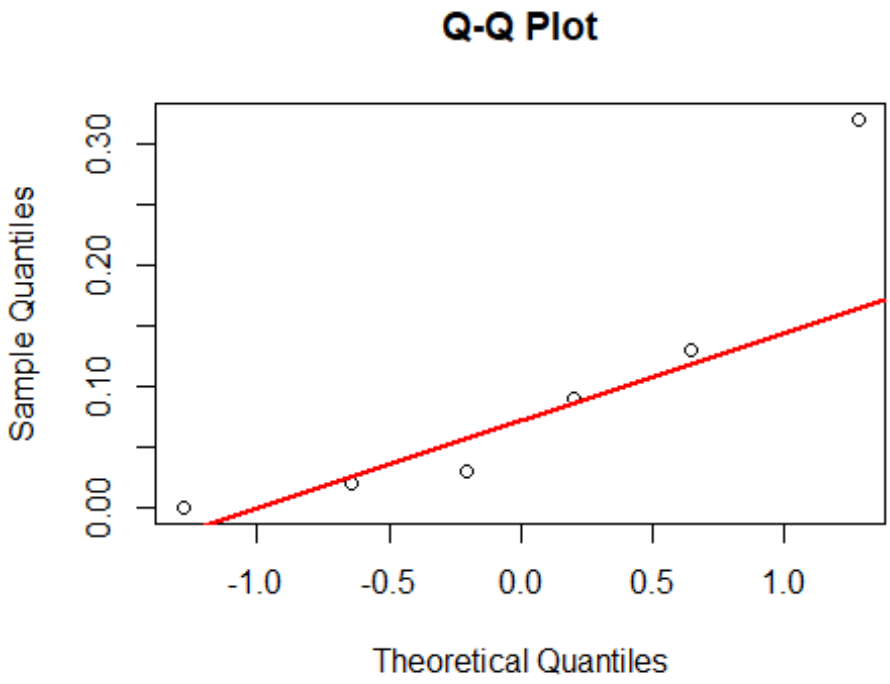
Testing between Conditions

Checking normality:

```r
rougel$diff <- rougel$with_punct - rougel$without_punct

shapiro_test <- shapiro.test(rougel$diff)
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rougel$diff
## W = 0.82802, p-value = 0.1034
```

```r
qqnorm(rougel$diff,
       main = "Q-Q Plot",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")

qqline(rougel$diff, col = "red", lwd = 2)
```



As the p-value is greater than alpha, we fail to reject the Null Hypothesis. This means, normality can be assumed. The QQ plot does contain one outlier compared to the other datasets. This should be kept in mind. We proceed with the paired t-test: ##### Paired t-test:

```
t.test(rougel$with_punct, rougel$without_punct, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  rougel$with_punct and rougel$without_punct
## t = 2.0258, df = 5, p-value = 0.09863
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.02644217  0.22310883
## sample estimates:
## mean difference
##      0.09833333
```

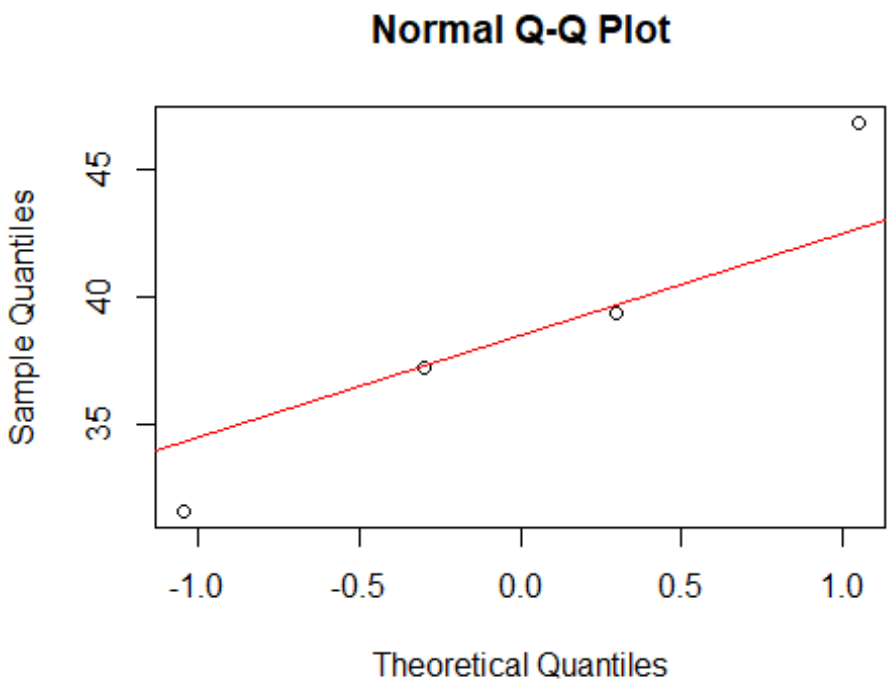
As the p-value is greater than alpha, we fail to reject the Null Hypothesis. There is NOT a statistically difference between ROUGE-L scores with and without punctuation.

Testing between Meeting types

Checking normality:

Since the sample size must be at least of size 3 for this test, we can only test it for the meeting type “scenario”.

```r
rougel$Type <- ifelse(grepl("^ES", rougel$File), "scenario", "natural")

shapiro.test(rougel$with_punct[rougel$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  rougel$with_punct[rougel$Type == "scenario"]
## W = 0.98305, p-value = 0.9197

shapiro.test(rougel$without_punct[rougel$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  rougel$without_punct[rougel$Type == "scenario"]
## W = 0.98357, p-value = 0.9226

qqnorm(rougel$with_punct[rougel$Type == "scenario"])
qqline(rougel$with_punct[rougel$Type == "scenario"], col = "red")
```



For both cases, with and without punctuation, the normality is not violated as shown in the results. In the QQ-plot we can notice one stronger outlier. The p-value is larger than alpha, therefore, we fail to reject the Null Hypothesis. Therefore we will go ahead with the independent samples t-test:

```r
# for with punctuation
t.test(with_punct ~ Type, data = rougel)
```

```
##
##  Welch Two Sample t-test
##
## data:  with_punct by Type
## t = -0.40703, df = 3.9186, p-value = 0.7052
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -12.190439   9.095439
## sample estimates:
##  mean in group natural mean in group scenario
##                37.2250                38.7725
```

```
# for without punctuation
t.test(without_punct ~ Type, data = rougel)
```

```
##
##  Welch Two Sample t-test
##
## data:  without_punct by Type
## t = -0.39292, df = 3.9624, p-value = 0.7146
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -12.118664   9.123664
## sample estimates:
##  mean in group natural mean in group scenario
##                37.1600                38.6575
```

As the p-values are greater than alpha, we fail to reject the null hypothesis in both cases. There is no statistically significant difference in ROUGE-L scores between natural and scenario meetings with punctuation or without it. This result should be taken with caution, as the natural group only contained 2 samples. Therefore, the results do not hold much power.

BERTScore

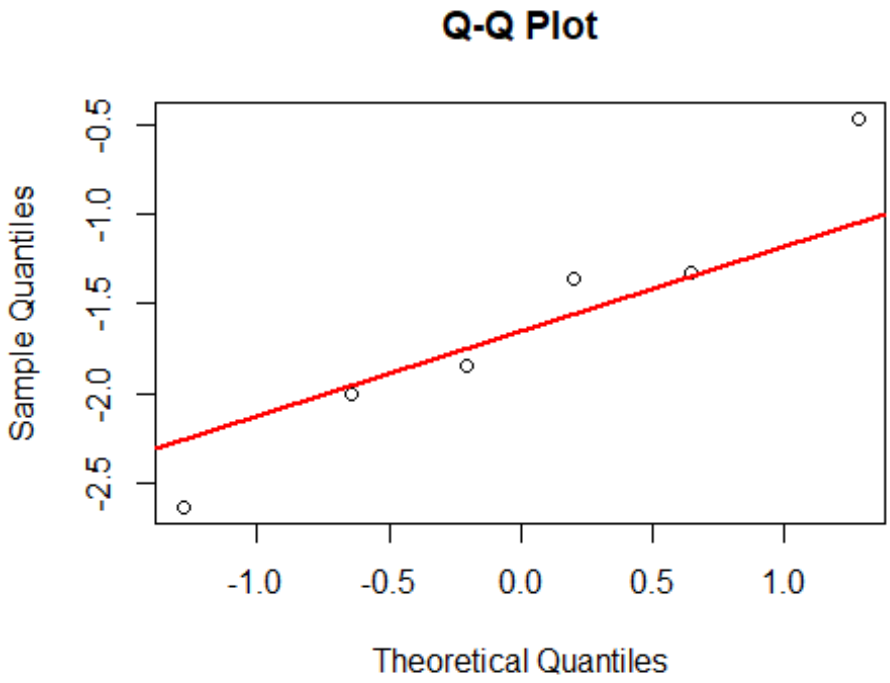
Testing between Conditions

Checking normality:

```
bertscore$diff <- bertscore$with_punct - bertscore$without_punct

shapiro_test <- shapiro.test(bertscore$diff)
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bertscore$diff
## W = 0.97726, p-value = 0.9371
```

```r
qqnorm(bertscore$diff,
       main = "Q-Q Plot",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")

qqline(bertscore$diff, col = "red", lwd = 2)
```



As the p-value is greater than alpha, we fail to reject the Null Hypothesis. This means, normality can be assumed. This is supported by the QQ plot as there are no curvatures present. There seems to be one stronger outlier, which is taken notice of.

Therefore, we can proceed with the paired t-test: ##### Paired t-test:

```r
t.test(bertscore$with_punct, bertscore$without_punct, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  bertscore$with_punct and bertscore$without_punct
## t = -5.309, df = 5, p-value = 0.003169
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -2.3845998 -0.8287335
## sample estimates:
## mean difference
##       -1.606667
```

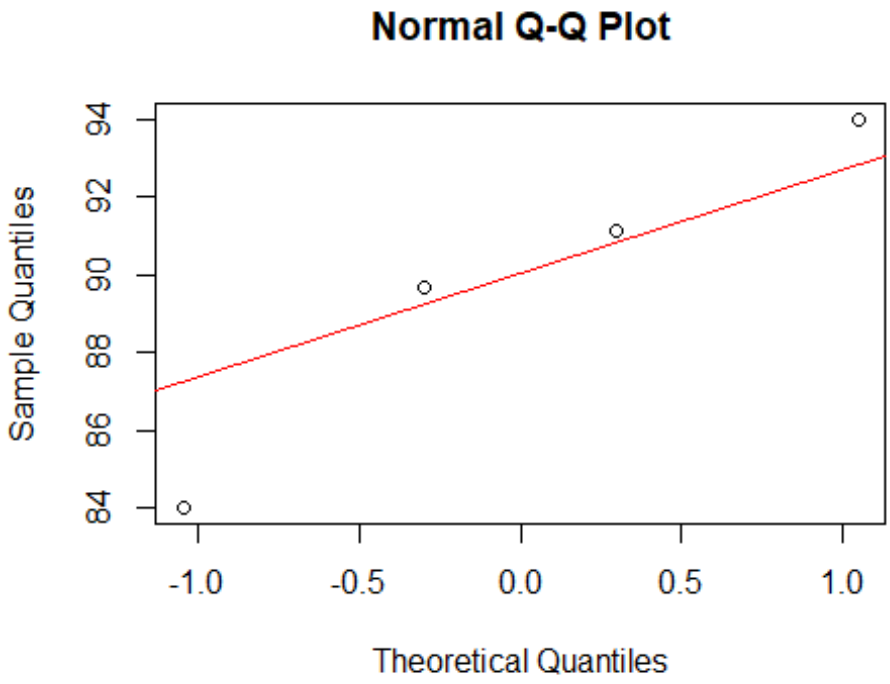
As the p-value is much smaller than alpha, we can reject the Null Hypothesis. There IS a statistically difference between BERTSCORE F1 scores with and without punctuation.

Testing between Meeting types

Checking normality:

Since the sample size must be at least of size 3 for this test, we can only test it for the meeting type “scenario”.

```
bertscore$Type <- ifelse(grepl("^ES", bertscore$File), "scenario", "natural")

shapiro.test(bertscore$with_punct[bertscore$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  bertscore$with_punct[bertscore$Type == "scenario"]
## W = 0.95508, p-value = 0.7479

shapiro.test(bertscore$without_punct[bertscore$Type == "scenario"])

##
##  Shapiro-Wilk normality test
##
## data:  bertscore$without_punct[bertscore$Type == "scenario"]
## W = 0.9687, p-value = 0.8334

qqnorm(bertscore$with_punct[bertscore$Type == "scenario"])
qqline(bertscore$with_punct[bertscore$Type == "scenario"], col = "red")
```



For both cases, with and without punctuation, the normality seems to be not violated as shown in the results. Once again, we take notice of one stronger outlier. The p-value is larger than alpha, therefore, we fail to reject the Null Hypothesis. Therefore we will go ahead with the independent samples t-test:

```
# for with punctuation
t.test(with_punct ~ Type, data = bertscore)

##
##  Welch Two Sample t-test
##
## data:  with_punct by Type
## t = -1.2599, df = 3.6421, p-value = 0.2824
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -11.084884   4.349884
## sample estimates:
##  mean in group natural mean in group scenario
##                86.3200                89.6875

# for without punctuation
t.test(without_punct ~ Type, data = bertscore)

##
##  Welch Two Sample t-test
##
```

```
## data:  without_punct by Type
## t = -1.3792, df = 3.6776, p-value = 0.2458
## alternative hypothesis: true difference in means between group natural and
group scenario is not equal to 0
## 95 percent confidence interval:
##  -11.621626   4.086626
## sample estimates:
##  mean in group natural mean in group scenario
##                87.6600                91.4275
```

As the p-values are greater than alpha, we fail to reject the null hypothesis in both cases. There is no statistically significant difference in BERTScore F1 scores between natural and scenario meetings with punctuation or without it. This result should be taken with caution, as the natural group only contained 2 samples. Therefore, the results do not hold much power.