Introduction to Medical Statistics 2026
Exercises Class 5
Statistics for the comparison of groups (continuous data)

Nguyen Lam Vuong and the Biostatistics team

2026-03-23

Exercise i)

We use the dataset cmTbmData.csv. It contains information on 201 patients with meningitis from 4 different patient groups. For this session, we will restrict attention to HIV-positive patients.

1. Import the dataset and create a new dataset cmTbm.hiv which contains HIV-positive patients only. We also load the **ggplot2** package to plot beautiful graphs, and the **patchwork** package which allows us to easily plot graphs next to each other.

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(patchwork)
```

```
Warning: package 'patchwork' was built under R version 4.4.3
```

```r
# Import the dataset
cmTbm <- read.csv("https://raw.githubusercontent.com/oucru-biostats/IntroductionToBios

# Create a new dataset cmTbm.hiv that contains HIV-positive patients only
cmTbm.hiv <- subset(cmTbm,hiv==1)
```

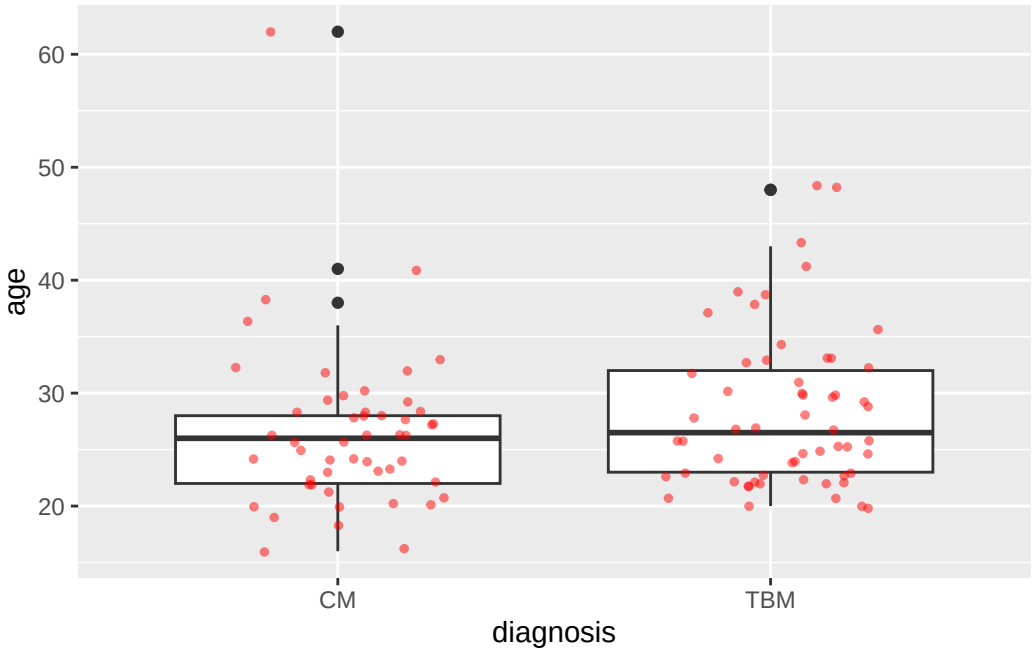
2. We test whether there a difference in age between HIV-positive patients with cryptococcal meningitis (CM) or tuberculous meningitis (TBM). What should we do?

a. First, visualise and compare the age distribution in the two groups using a box-plot and a density plot. Do you think it is reasonable to assume that the age distribution is normal in both groups?

```r
ggplot(cmTbm.hiv, aes(diagnosis,age)) + geom_boxplot() +
  geom_jitter(size = 1, alpha = 0.5, width = 0.25, colour = 'red')
```

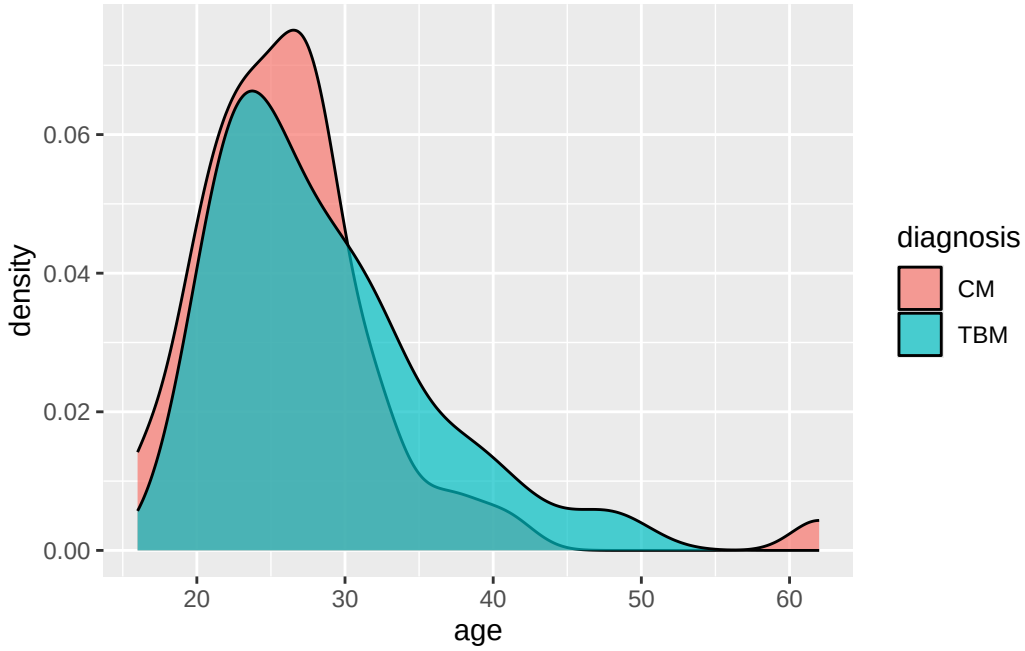
```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



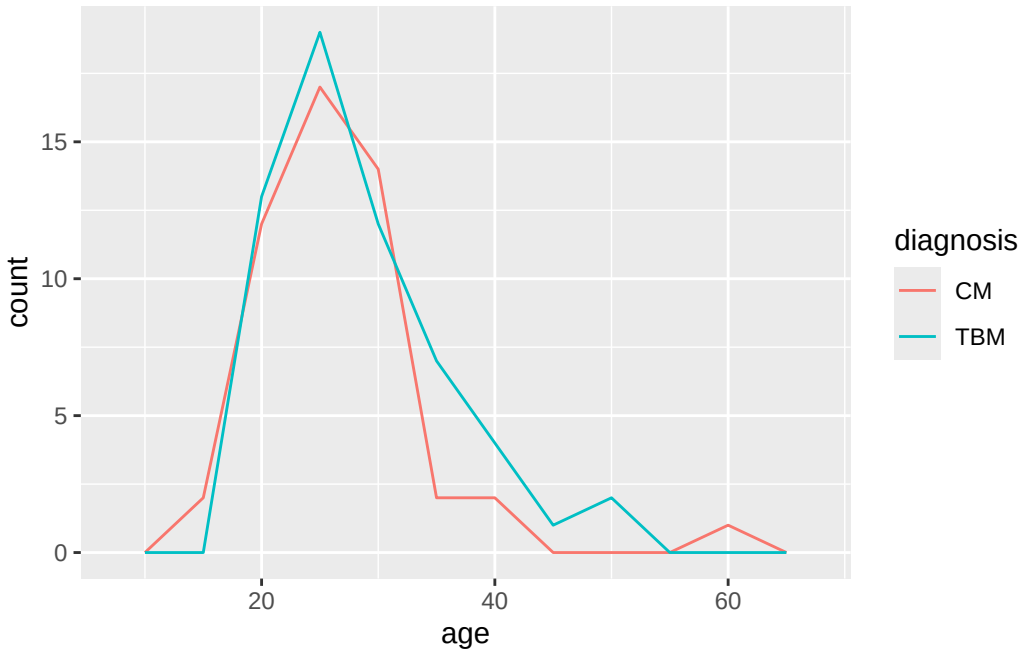
```
ggplot(cmTbm.hiv, aes(x = age, fill = diagnosis)) + geom_density(alpha = 0.7)
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_density()`).
```




```
## alternative: frequency polygon (not available via ggplotgui)
ggplot(cmTbm.hiv, aes(x = age)) + geom_freqpoly(aes(colour = diagnosis),binwidth = 5)
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_bin()`).
```



Answer: *The distribution of age in each diagnosis group is slightly skewed to the right. there is one outlier in the CM group, a person aged 62.*

b. Perform a t-test to check the hypothesis that both age distributions are equal. Use the *t.test* function. You can use the formula specification *age~diagnosis* as first argument, with the outcome (dependent) variable on the left hand side, and the grouping (independent) variable on the right hand side. Try to understand the output of the *t.test* function. What do you conclude?

Next, remove the person aged 62 and check whether anything changes.

```
t.test(age ~ diagnosis, data = cmTbm.hiv)
```

```
    Welch Two Sample t-test

data:  age by diagnosis
t = -1.363, df = 101.48, p-value = 0.1759
alternative hypothesis: true difference in means between group CM and group TBM is not
95 percent confidence interval:
 -4.5856262  0.8504538
sample estimates:
 mean in group CM mean in group TBM
         26.46000          28.32759
```

```r
# remove the person aged 62
t.test(age ~ diagnosis, data = cmTbm.hiv, subset = age<60)
```

```
    Welch Two Sample t-test

data:  age by diagnosis
t = -2.212, df = 104.03, p-value = 0.02915
alternative hypothesis: true difference in means between group CM and group TBM is not
95 percent confidence interval:
 -4.9173369 -0.2684478
sample estimates:
 mean in group CM mean in group TBM
         25.73469          28.32759
```

Answer: *The p-value of 0.18 gives no strong indication of a difference in age. If we remove the person aged 62, the p-value changes quite a lot and based on the p-value there is a suggestion of a difference.*

c. Test the hypothesis again, now with the Wilcoxon rank-sum test, using the *wilcox.test* function (here we can use the formula structure as well). What happens if we remove the person aged 62 in the test?

```r
# Wilcoxon rank-sum test
wilcox.test(age~diagnosis, data=cmTbm.hiv)
```

```
    Wilcoxon rank sum test with continuity correction

data:  age by diagnosis
W = 1210, p-value = 0.1392
alternative hypothesis: true location shift is not equal to 0
```

```r
# remove the person aged 62
wilcox.test(age~diagnosis, data=cmTbm.hiv, subset=age<60)
```

```
    Wilcoxon rank sum test with continuity correction

data:  age by diagnosis
W = 1152, p-value = 0.09244
alternative hypothesis: true location shift is not equal to 0
```

Answer: *If we do the Wilcoxon rank-sum test, p-value is more in correspondence with the t.test on the full data set. Removing the one person lowers the p-value, but not as much as with the t-test.*

We could conclude that there's no strong indication that the age distribution is different. There are two further things to consider:

Is the difference in mean age clinically relevant?

What is the interpretation of the test? Do we test for a difference in age between hiv positive individuals with CM or TBM in some larger population? If so, which population?

If the age difference is specific for this sample, performing the test and computing confidence intervals and p-values is irrelevant. We only compare the age distribution in this specific sample and it cannot be generalized to another setting and a larger population.

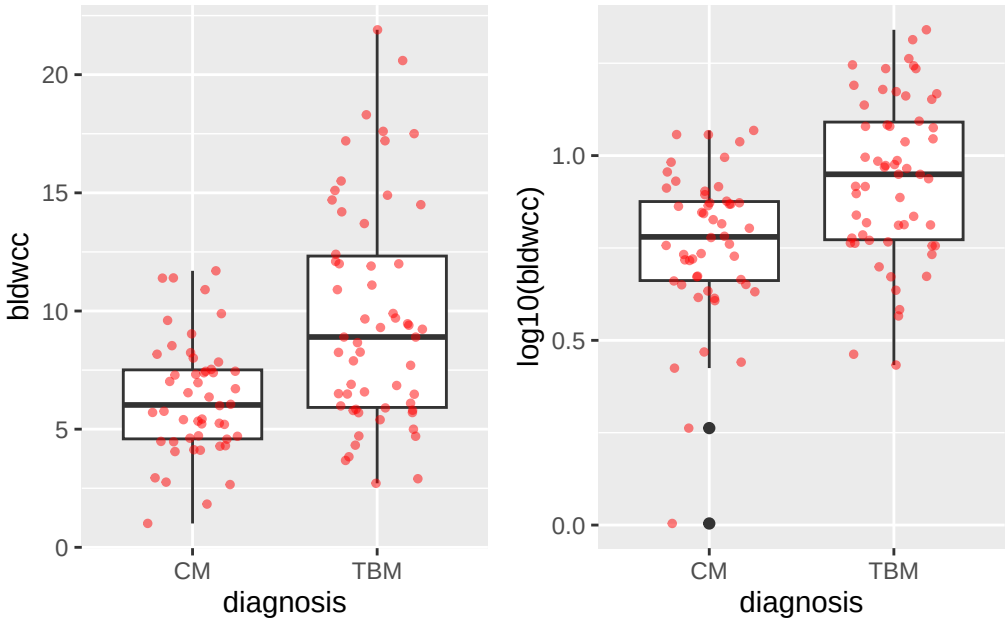
3. We compare white cell count both in blood (bldwcc) and in CSF (csfwcc) between the two groups, using both the t-test and the Wilcoxon rank-sum test. Are there significant differences between HIV-positive patients with CM or TBM? Do the variables need to be transformed before performing a t-test? If yes, please do so.

```
# First, draw boxplots and calculate several summary statistics
summary(cmTbm.hiv$bldwcc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.010   5.317   7.155   8.143   9.748  21.900       1
```

```
p1 <- ggplot(cmTbm.hiv, aes(diagnosis, bldwcc)) + geom_boxplot() +
  geom_jitter(size = 1, alpha = 0.5, width = 0.25, colour = 'red')
p2 <- ggplot(cmTbm.hiv, aes(diagnosis, log10(bldwcc))) + geom_boxplot() +
  geom_jitter(size = 1, alpha = 0.5, width = 0.25, colour = 'red')
p1 + p2
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```

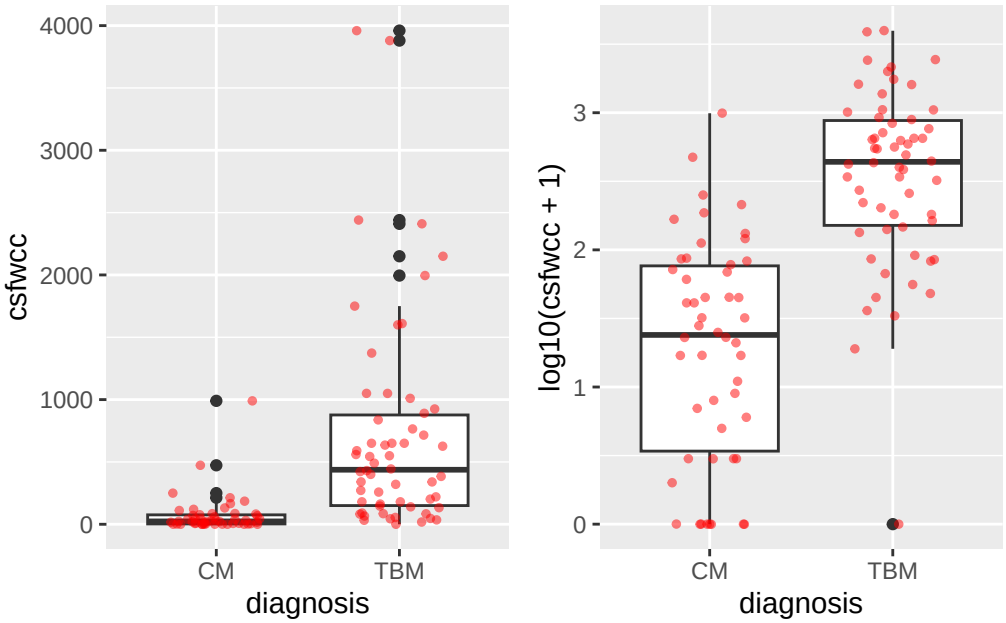



```
summary(cmTbm.hiv$csfwcc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0    23.5   115.5   422.4   545.5  3960.0       1
```

```
p1 <- ggplot(cmTbm.hiv, aes(diagnosis, csfwcc)) + geom_boxplot() +
  geom_jitter(size = 1, alpha = 0.5, width = 0.25, colour = 'red')
p2 <- ggplot(cmTbm.hiv, aes(diagnosis, log10(csfwcc + 1))) + geom_boxplot() +
  geom_jitter(size = 1, alpha = 0.5, width = 0.25, colour = 'red')
p1 + p2
```

```
Warning: Removed 1 row containing non-finite outside the scale range (`stat_boxplot()`
Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



Answer: *Log-transformation is a good idea for csfwcc, because it has a very skewed distribution. Note that there's still an outlier in the TBM group. For bldwcc a transformation is not needed. Notice that for csfwcc there are 0 values –> if you decide to log transform, please add a small value into the original data.*

```r
# t test (csfwcc we only use log transformed data)
t.test(bldwcc ~ diagnosis, data = cmTbm.hiv)
```

```
    Welch Two Sample t-test

data:  bldwcc by diagnosis
```

```
t = -4.8479, df = 88.283, p-value = 5.309e-06
alternative hypothesis: true difference in means between group CM and group TBM is not
95 percent confidence interval:
 -4.833328 -2.022907
sample estimates:
 mean in group CM mean in group TBM
         6.302400          9.730517
```

```
t.test(log10(bldwcc) ~ diagnosis, data = cmTbm.hiv)
```

```
    Welch Two Sample t-test

data:  log10(bldwcc) by diagnosis
t = -4.3806, df = 105.34, p-value = 2.803e-05
alternative hypothesis: true difference in means between group CM and group TBM is not
95 percent confidence interval:
 -0.25643133 -0.09663046
sample estimates:
 mean in group CM mean in group TBM
        0.7606388         0.9371697
```

```
t.test(log10(csfwcc + 1) ~ diagnosis, data = cmTbm.hiv)
```

```
    Welch Two Sample t-test

data:  log10(csfwcc + 1) by diagnosis
t = -8.8917, df = 93.14, p-value = 4.47e-14
alternative hypothesis: true difference in means between group CM and group TBM is not
95 percent confidence interval:
 -1.5481555 -0.9829012
sample estimates:
 mean in group CM mean in group TBM
         1.270431          2.535959
```

```
# Wilcoxon test
wilcox.test(log10(bldwcc) ~ diagnosis, data = cmTbm.hiv)
```

```
    Wilcoxon rank sum test with continuity correction

data:  log10(bldwcc) by diagnosis
W = 804.5, p-value = 7.062e-05
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(bldwcc ~ diagnosis, data = cmTbm.hiv)
```

```
    Wilcoxon rank sum test with continuity correction

data:  bldwcc by diagnosis
W = 804.5, p-value = 7.062e-05
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(log10(csfwcc + 1) ~ diagnosis, data = cmTbm.hiv)
```

```
    Wilcoxon rank sum test with continuity correction

data:  log10(csfwcc + 1) by diagnosis
W = 298.5, p-value = 1.299e-12
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(csfwcc ~ diagnosis, data = cmTbm.hiv)
```

```
    Wilcoxon rank sum test with continuity correction

data:  csfwcc by diagnosis
W = 298.5, p-value = 1.299e-12
alternative hypothesis: true location shift is not equal to 0
```

Both markers seem to differ by infection type. All tests convincingly show that both bldwcc and csfwcc are higher in TBM patients (all p<0.0001).

t.test gives you the CI (but note that this is a CI for the difference of the log-transformed data if we use the log-transformed values). For the Wilcoxon test it doesn't matter whether data is transformed or not.The wilcox.test function does not provide CI's.

Exercise ii)

We use the dataset bmData.csv, containing selected variables from 300 patients with confirmed bacterial meningitis who were randomized to either adjunctive dexamethasone therapy or placebo.

1. Import the dataset with both treatment groups, check the distribution of CSF total white cell count at baseline and follow-up. Compute the transformed variables if needed.

```
# Import dataset
bmData <- read.csv("https://raw.githubusercontent.com/oucru-biostats/IntroductionToBio

summary(bmData$wc.csf)
```

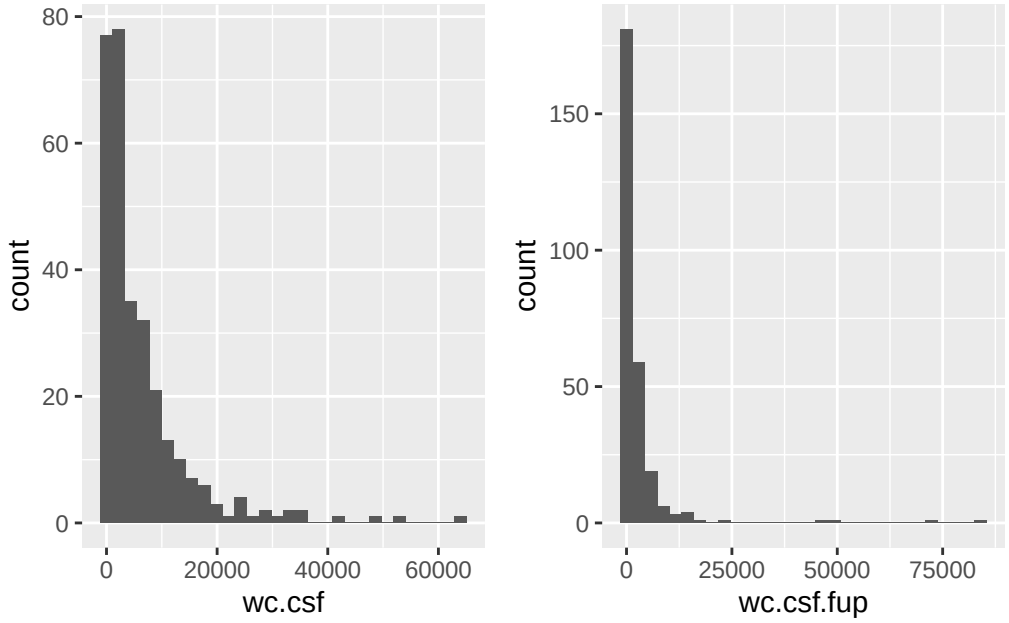
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      1    1085    3160    6252    7830   64000       1
```

```
summary(bmData$wc.csf.fup)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   16.0   352.5   902.5  2883.4  2375.0 84000.0      22
```

```
# draw histogram
p1 <- ggplot(bmData, aes(wc.csf)) + geom_histogram()
p2 <- ggplot(bmData, aes(wc.csf.fup)) + geom_histogram()
p1 + p2
```

```
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_bin()`).

`stat_bin()` using `bins = 30`. Pick better value `binwidth`.

Warning: Removed 22 rows containing non-finite outside the scale range
(`stat_bin()`).
```



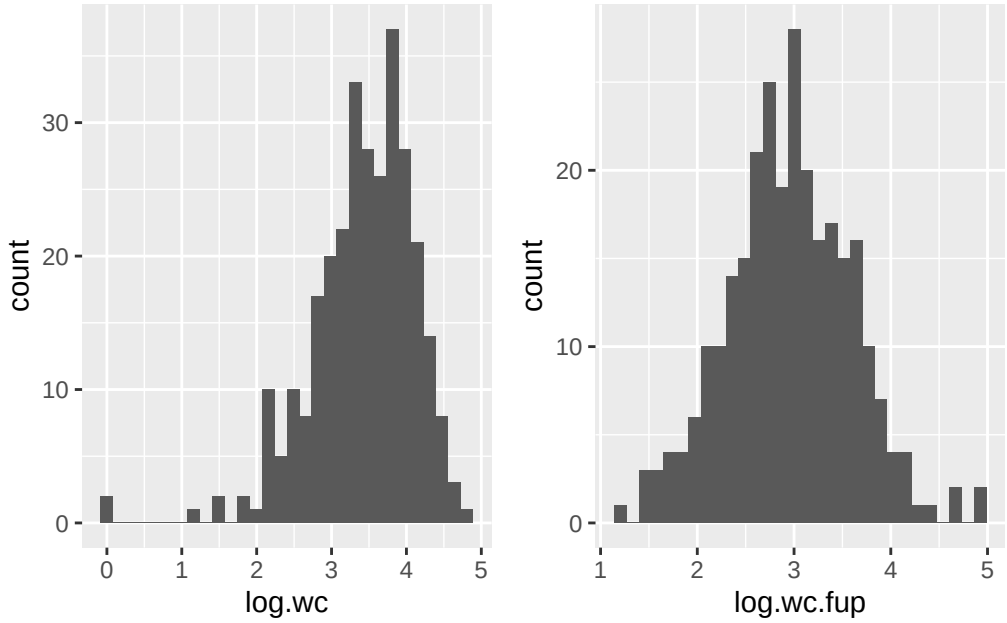

```
# perform a log-transformation
bmData$log.wc <- log10(bmData$wc.csf)
bmData$log.wc.fup <- log10(bmData$wc.csf.fup)

# draw histogram of the log-transformed variables
p1 <- ggplot(bmData, aes(log.wc)) + geom_histogram()
p2 <- ggplot(bmData, aes(log.wc.fup)) + geom_histogram()
p1 + p2
```

```
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_bin()`).
```

```
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```

```
Warning: Removed 22 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Answer: *We saw that CSF total white cell count at baseline and follow-up are highly skewed in the dexamethasone group. Therefore, we perform a log-transformation first (note that all values were >0). Be careful that these are paired variables (WCC in CSF measured at baseline and at follow-up are from the same patients).*

2. Test whether the change in value differs from zero in the dexamethasone group, using the paired t-test. Compare the result with the one based on the one-sample t-test for the difference.

```
# compare
t.test(subset(bmData, group=="dexamethasone")$log.wc, subset(bmData, group=="dexametha
```

```
    Paired t-test

data:  subset(bmData, group == "dexamethasone")$log.wc and subset(bmData, group == "de
t = 7.1351, df = 133, p-value = 5.623e-11
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.374681 0.662089
sample estimates:
mean difference
       0.518385
```

```r
# alternative formulation
with( subset(bmData, group=="dexamethasone"), t.test(log.wc, log.wc.fup, paired = TRUE
```

```
    Paired t-test

data:  log.wc and log.wc.fup
t = 7.1351, df = 133, p-value = 5.623e-11
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.374681 0.662089
sample estimates:
mean difference
       0.518385
```

```r
# Equivalent alternative: One sample t.test to test whether difference is 0
t.test(subset(bmData, group=="dexamethasone")$log.wc - subset(bmData, group=="dexameth
```

```
    One Sample t-test

data:  subset(bmData, group == "dexamethasone")$log.wc - subset(bmData, group == "dexa
t = 7.1351, df = 133, p-value = 5.623e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.374681 0.662089
sample estimates:
mean of x
 0.518385
```

```r
# this can also be formulated as
t.test(log.wc.fup - log.wc ~ 1, data = bmData, subset = group=="dexamethasone")
```

```
    One Sample t-test

data:  log.wc.fup - log.wc
t = -7.1351, df = 133, p-value = 5.623e-11
```

```
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.662089 -0.374681
sample estimates:
mean of x
-0.518385
```

Answer: *To perform the test for the dexamethasone group, we could create a separate data set first using subset(). That would save some writing, hence make the code more readable. But it's not needed. We can also apply the subset() function when performing the test. We could use a paired t-test or a one sample t.test to test whether difference is 0. The latter has the advantage that we can use the formula structure of the function and select the dexamethasone subgroup via the subset argument.*

The data strongly suggest that the value is higher in the follow-up measurement.

3. Does the change in CSF total white cell counts differ between the two randomized groups?

```r
# t-test using the formula structure, with the difference as outcome.
t.test(log.wc.fup - log.wc ~ group, data = bmData)
```

```
    Welch Two Sample t-test

data:  log.wc.fup - log.wc by group
t = -0.88503, df = 265.48, p-value = 0.3769
alternative hypothesis: true difference in means between group dexamethasone and group
95 percent confidence interval:
 -0.2726660  0.1035552
sample estimates:
mean in group dexamethasone       mean in group placebo
                 -0.5183850                  -0.4338296
```

```r
# Or Wilcoxon test
wilcox.test(log.wc.fup-log.wc ~ group, data = bmData)
```

```
    Wilcoxon rank sum test with continuity correction
```

```
data:  log.wc.fup - log.wc by group
W = 8980, p-value = 0.3674
alternative hypothesis: true location shift is not equal to 0
```

Answer: *The change seems to be comparable in both groups.*