Write-up

Yui Hang, Wong (Harris) K21115599

Library

```
suppressPackageStartupMessages(library(here))
suppressPackageStartupMessages(library(tidyverse))
library(haven)
library(patchwork)
```

1 Manipulating dataframes

1.1

```
file <- here("data", "ESS.sav")
ess_raw <- haven::read_sav(file)
```

1.2

```
ess <- ess_raw %>% select(
        idno, cntry, stfeco, gincdif, advbach, gndr, yrbrn,
        wkhtot, jbscr, mainact, hltherb
)
```

1.3

```
ess <- ess %>% rename(
        idno = idno,
        country = cntry,
        satisfied_econ = stfeco,
        reduce_diff = gincdif,
        first_if_backache = advbach,
        gender = gndr,
        yob = yrbrn,
        working_hours = wkhtot,
        job_secure = jbscr,
        main_activity = mainact,
        herbal_remedies = hltherb
)
```

1.4

```
year_of_collection <- 2004
ess <- ess %>% mutate(age = year_of_collection - yob)

ess$age[1:10]
```

```
##  [1] 33 79 27 15 16 55 33 38 37 46
```

1.5

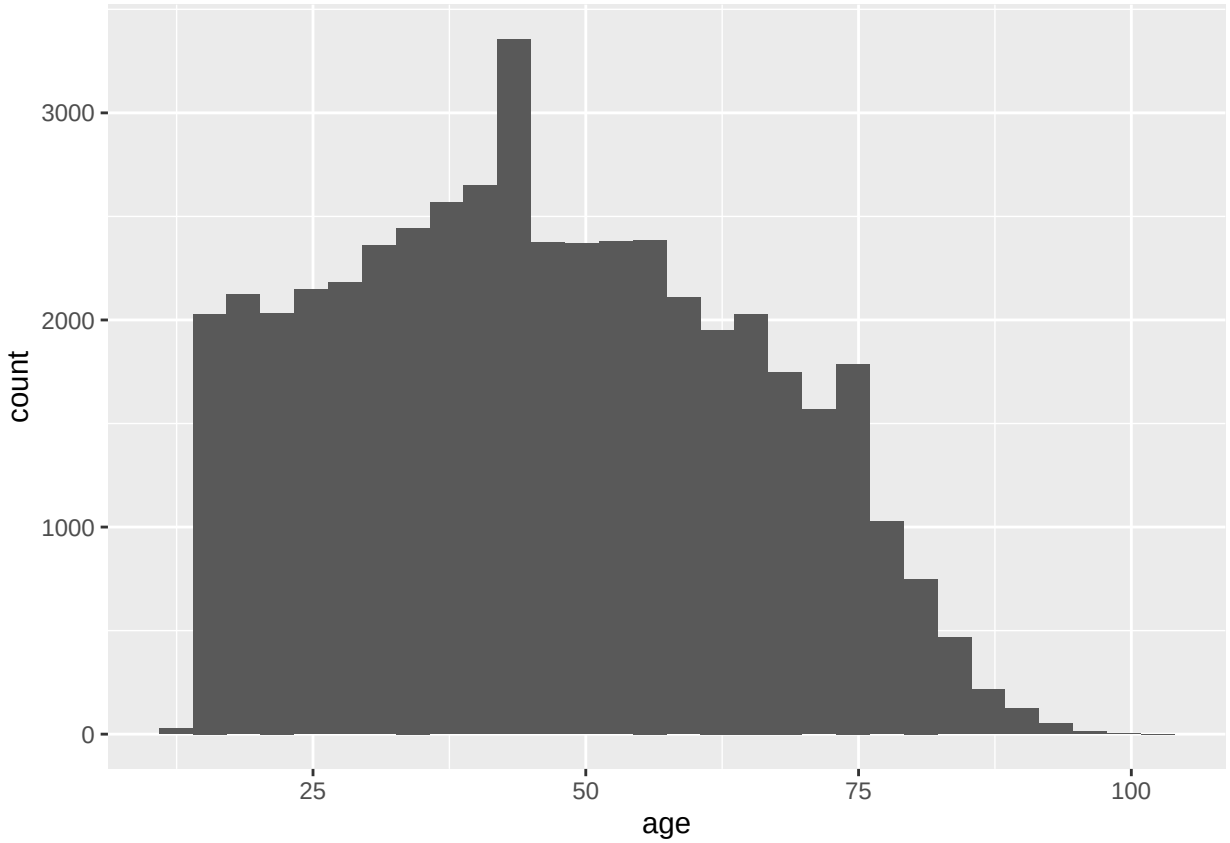
```
ess %>%
        summarise(
                sd = sd(age, na.rm = T),
                mean = mean(age, na.rm = T),
                oldest = max(age, na.rm = T),
                youngest = min(age, na.rm = T)
        )
```

```
## # A tibble: 1 x 4
##      sd  mean oldest youngest
##   <dbl> <dbl>  <dbl>    <dbl>
## 1  18.5  46.2    102       12
```

1.6

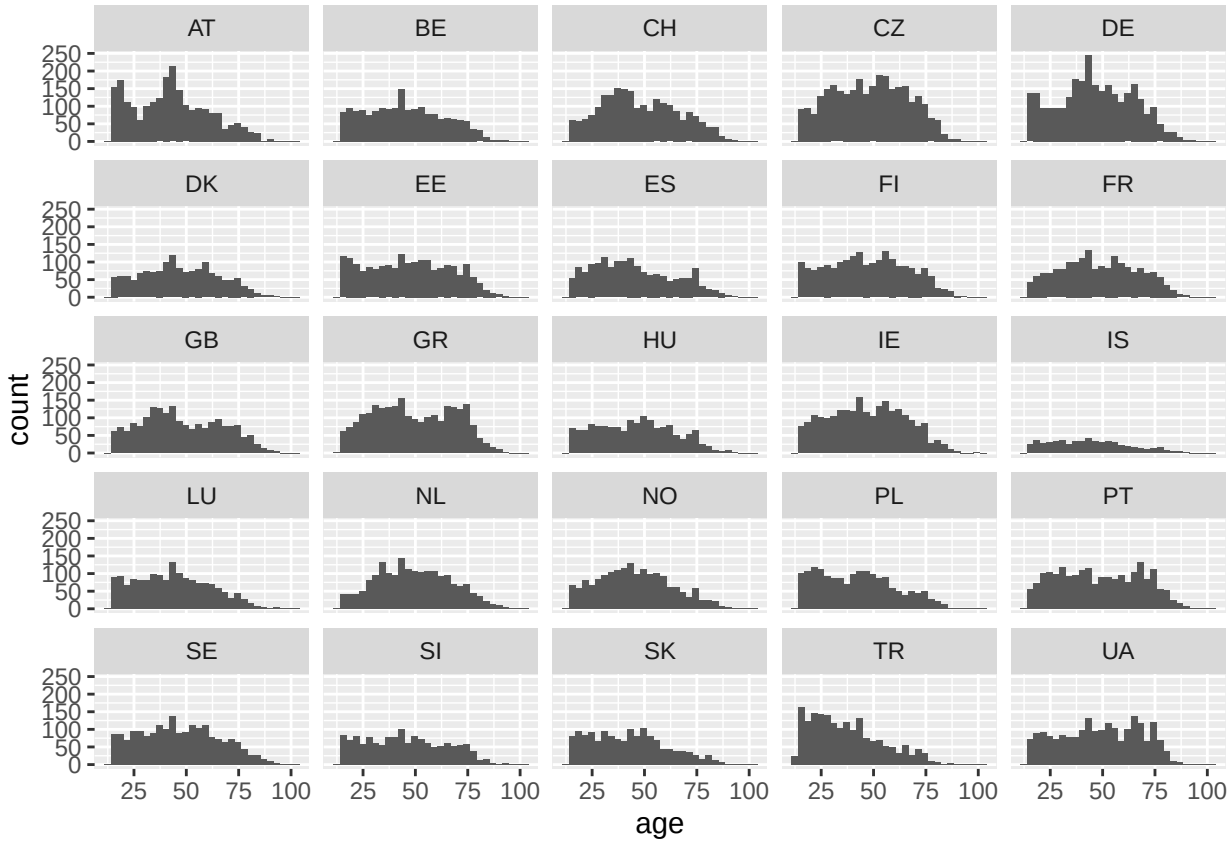
```
p1.6.1 <- ess %>%
        ggplot(aes(x = age)) +
        geom_histogram(bins = 30)

p1.6.2 <- p1.6.1 + facet_wrap(~country)

p1.6.1
```

```
## Warning: Removed 273 rows containing non-finite values (stat_bin).
```




```
p1.6.2
```

```
## Warning: Removed 273 rows containing non-finite values (stat_bin).
```



1.7

```
#! how to treat NA?
ess <- ess %>%
        mutate(female = as.numeric(gender == 2))
```

1.8

```
ess <- ess %>%
        mutate(work_over_50h = working_hours  > 50)
```

1.9

```
#! there must be a simpler way!
ess %>%
        count(work_over_50h) %>%
        mutate(prop = n/sum(n)) %>%
        filter(work_over_50h)
```

```
## # A tibble: 1 x 3
##   work_over_50h     n  prop
##   <lgl>         <int> <dbl>
## 1 TRUE           4938 0.104
```

```
ess %>%
        group_by(country) %>%
        count(work_over_50h) %>%
        mutate(prop = n / sum(n)) %>%
        filter(work_over_50h)
```

```
## # A tibble: 25 x 4
## # Groups:   country [25]
##    country          work_over_50h     n   prop
##    <chr+lbl>        <lgl>         <int>  <dbl>
##  1 AT [Austria]     TRUE            174 0.0771
##  2 BE [Belgium]     TRUE            170 0.0956
##  3 CH [Switzerland] TRUE            262 0.122
##  4 CZ [Czechia]     TRUE            324 0.107
##  5 DE [Germany]     TRUE            287 0.1
##  6 DK [Denmark]     TRUE             96 0.0646
##  7 EE [Estonia]     TRUE            134 0.0674
##  8 ES [Spain]       TRUE            168 0.101
##  9 FI [Finland]     TRUE            181 0.0895
## 10 FR [France]      TRUE            161 0.0891
## # ... with 15 more rows
```

1.10

```
ess <- ess %>%
        mutate(first_if_backache = as.numeric(first_if_backache %in% 4:5))
```

1.11

```
bycountry <- ess %>%
        group_by(country) %>%
        count(first_if_backache) %>%
        mutate(prop = n / sum(n)) %>%
        filter(first_if_backache == 1) %>%
        select(country, prop) %>%
        rename(seekdoctor = prop) %>%
        ungroup()
```

1.12

```
load(here("data", "health.Rdata"))

health_long <- health %>%
        pivot_longer(
                cols = -measure,
                names_to = "cntry"
        ) %>%
        tidyr::separate(measure, into = c(NA, "feature", "year"), sep = "_") %>%
        pivot_wider(names_from = "feature",
                    values_from = "value") %>%
        rename(healthspend = pcexhe, doctors = phyrape)
```

1.13

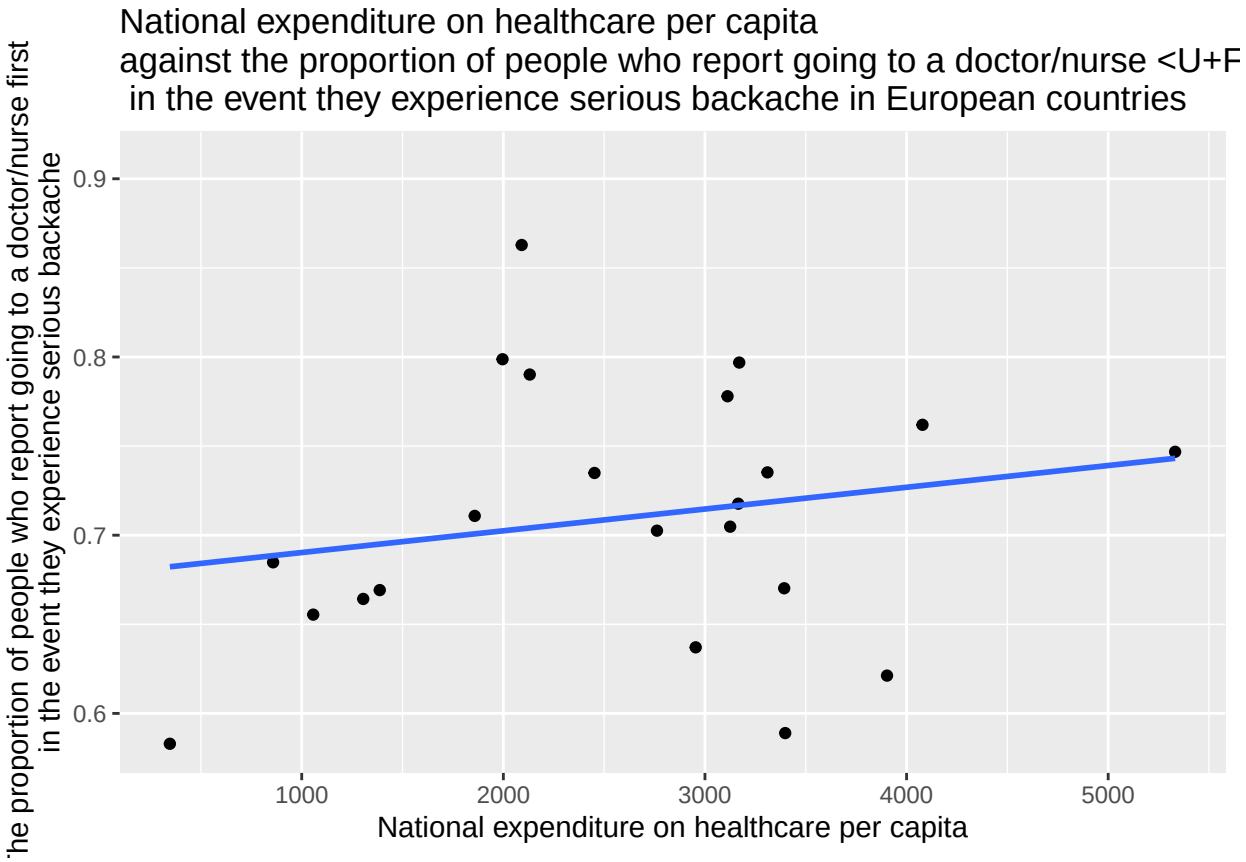
```
year_of_collection <- 2004

merged <- health_long %>%
        filter(year == year_of_collection) %>%
        select(-year) %>%
        full_join(bycountry, by = c("cntry" = "country"))

merged %>%
        summarise(
                both = sum(!is.na(healthspend) & !is.na(doctors) & !is.na(seekdoctor)),
                only_left = sum(!is.na(healthspend) & !is.na(doctors)),
                only_right = sum(!is.na(seekdoctor))
        )
```

```
## # A tibble: 1 x 3
##    both only_left only_right
##   <int>     <int>      <int>
## 1    20        27         25
```

1.14

```
p1.14 <- merged %>%
        ggplot(aes(x = healthspend, y = seekdoctor)) +
        geom_point() +
        ggplot2::stat_smooth(method = "lm", se = F) +
        labs(title = "National expenditure on healthcare per capita \nagainst the proportion of people w
             x = "National expenditure on healthcare per capita",
             y = "The proportion of people who report going to a doctor/nurse first \n in the event they
p1.14
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 11 rows containing non-finite values (stat_smooth).

## Warning: Removed 11 rows containing missing values (geom_point).
```



2 Random Numbers, Skewness and Kurtosis

2.1

```
n.seq <- c(100, 200, 10^3, 10^4, 10^5)
```

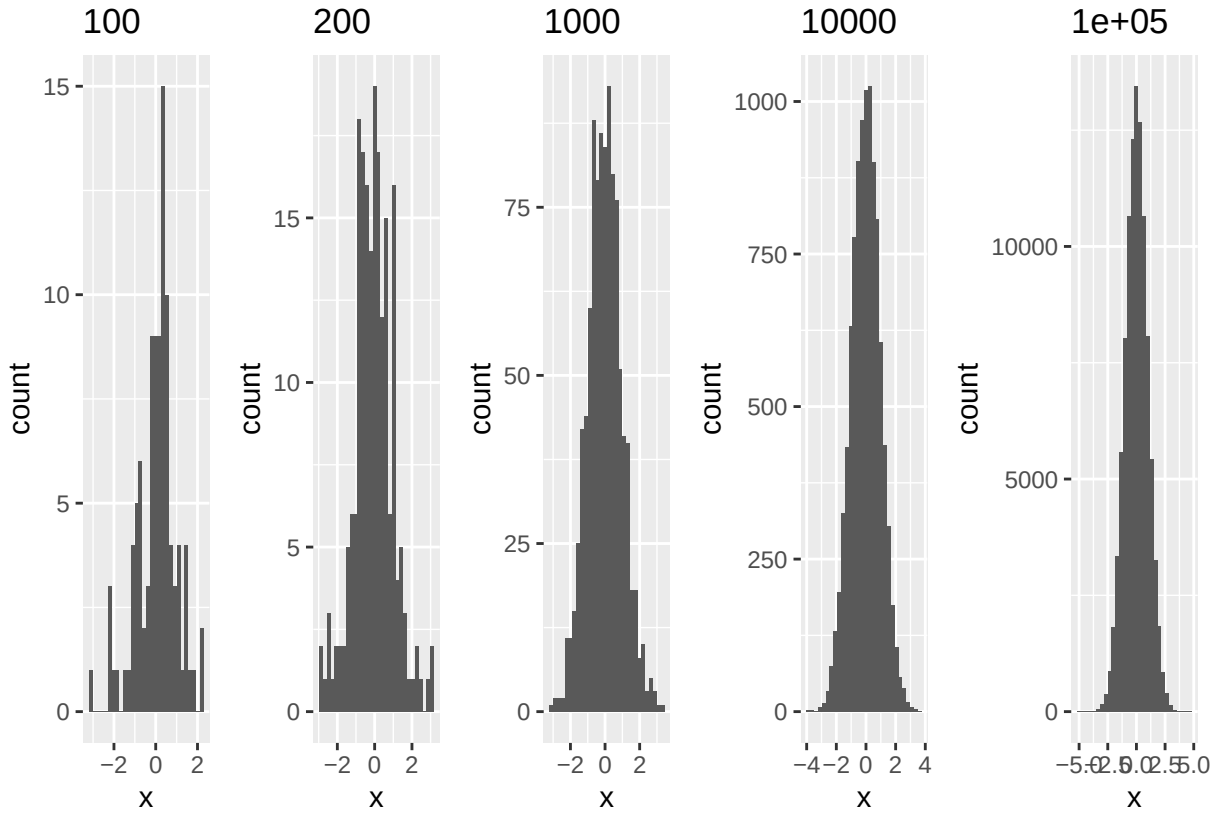
2.2

```
samples <- n.seq %>%
        map(rnorm, mean = 0, sd = 1)

# A list can be used to store vectors of different lengths.
```

2.3

```
plot_histo <- function(x, title = NULL) {
        p <- x %>%
                as.data.frame() %>%
```

```
                ggplot(aes(x = x)) +
                geom_histogram(bins = 30) +
                labs(title = title)
        return(p)
}

plot_histo_panel <- function(distributions, titles) {
        plot_list <- map2(
                .x = distributions,
                .y = titles,
                .f = plot_histo
        )
        pw <- plot_list %>% patchwork::wrap_plots(nrow = 1)
        return(pw)
}

titles <- as.character(n.seq)
p2.3 <- plot_histo_panel(samples, titles)
p2.3
```

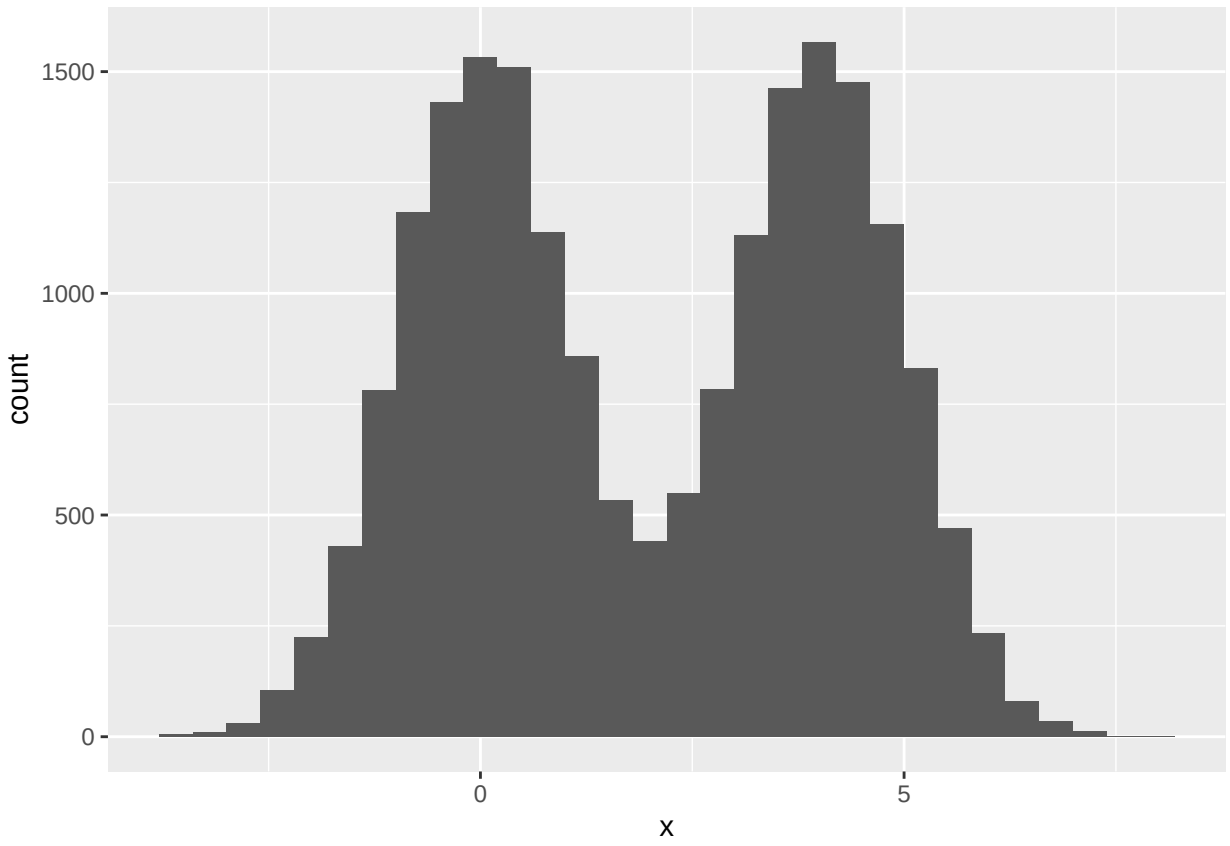


2.4

```
#! central limit theorem? How to describe?
```

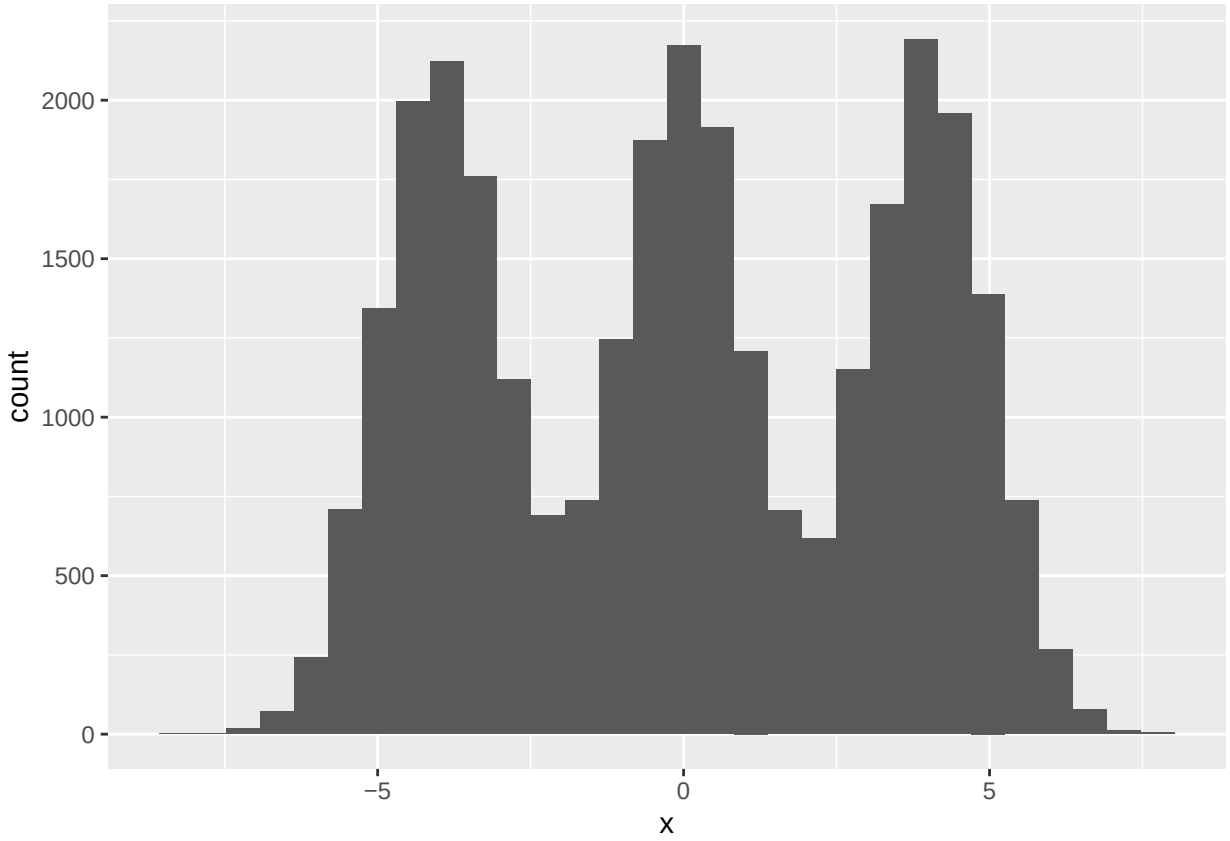
2.5

```
bimodal <- c(
        rnorm(mean = 0, sd = 1, n = 10000),
        rnorm(mean = 4, sd = 1, n = 10000)
        )
p2.5 <- plot_histo(bimodal)
p2.5
```



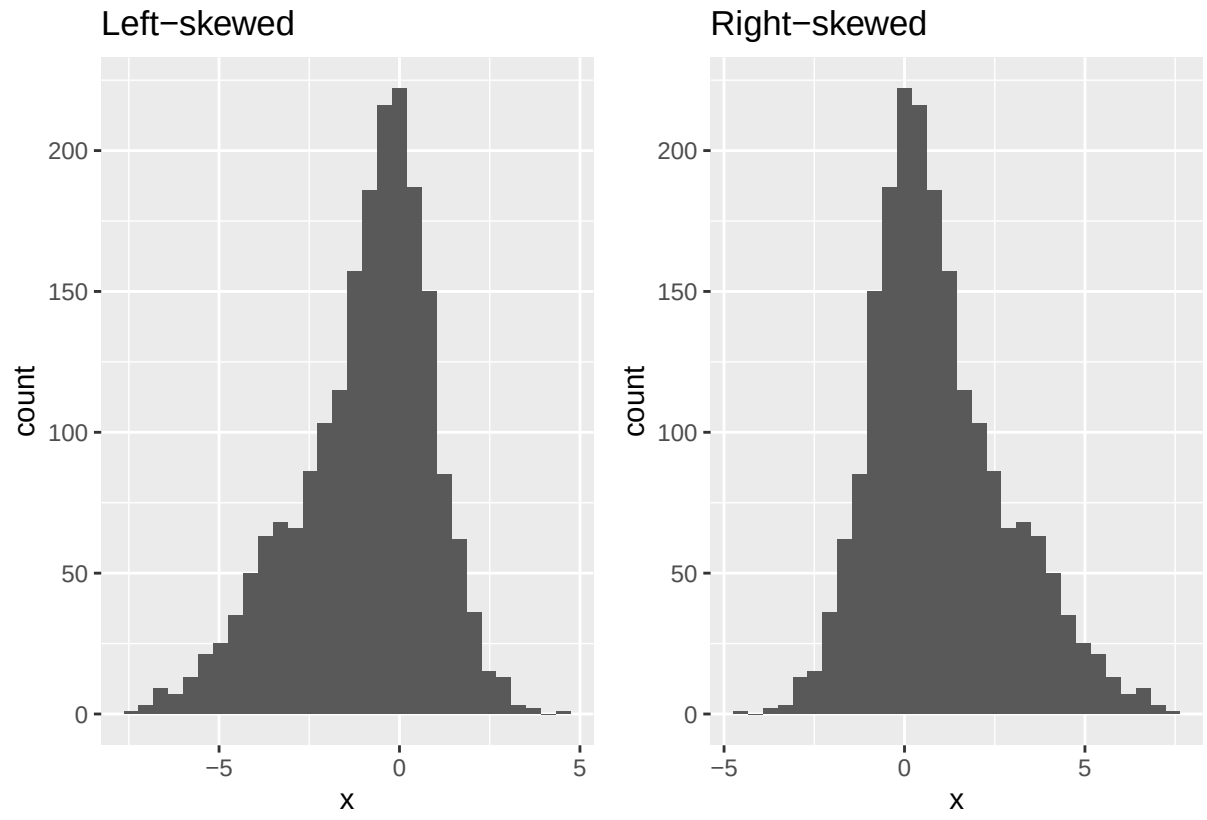
2.6

```
trimodal <- c(
        rnorm(mean = -4, sd = 1, n = 10000),
        rnorm(mean = 0,  sd = 1, n = 10000),
        rnorm(mean = 4,  sd = 1, n = 10000)
        )
p2.6 <- plot_histo(trimodal)
p2.6
```



2.7

```
right_skewed <- c(
        rnorm(mean = 0, sd = 1, n = 1000),
        rnorm(mean = 2, sd = 2, n = 1000)
)
left_skewed <- -right_skewed

skewed_distributions <- list(left_skewed, right_skewed)

titles <- c("Left-skewed", "Right-skewed")
plot_histo_panel(skewed_distributions, titles)
```

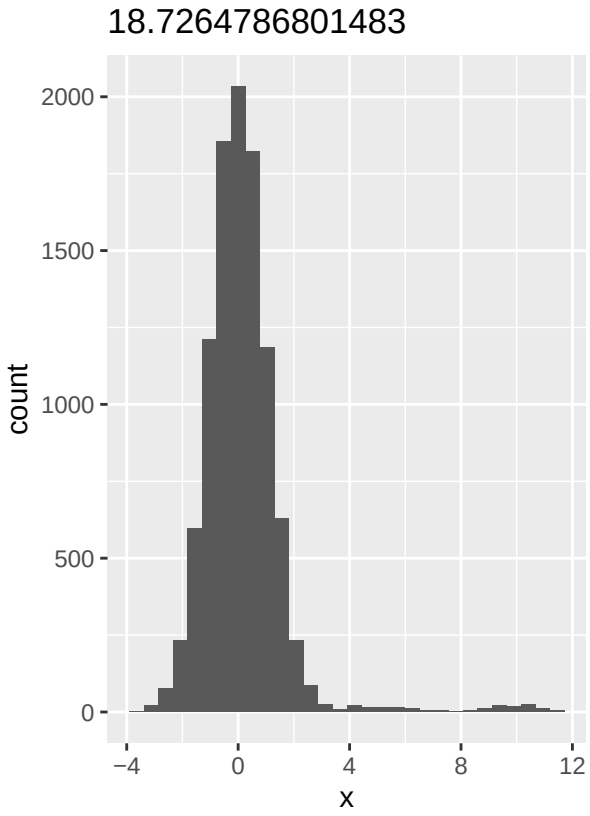


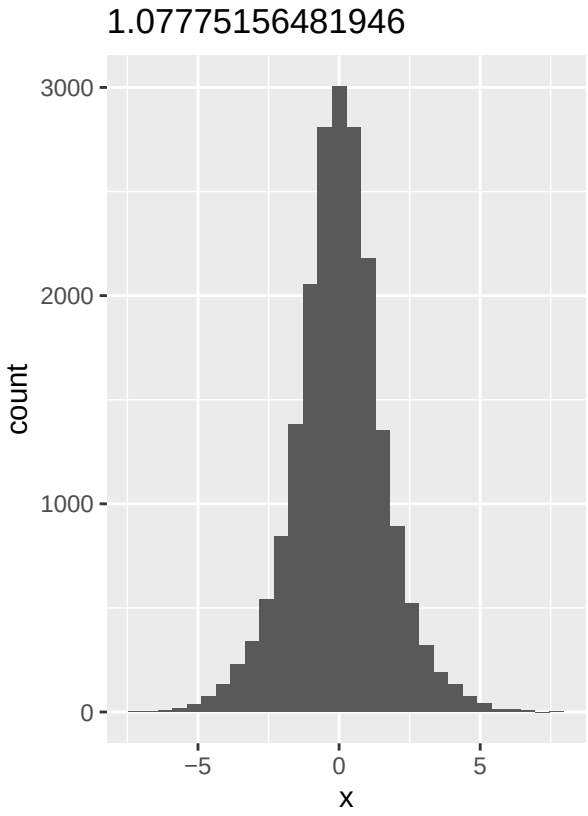

```
skewed_distributions %>%
        map_dbl(e1071::skewness)
```

```
## [1] -0.6350482  0.6350482
```

2.8

```
leptokurtic <- c(
        rnorm(10000, mean = 0, sd = 1),
        rnorm(100, mean = 5,  sd = 1),
        rnorm(100, mean = 10, sd = 1)
)
platykurtic <- c(
        rnorm(10000, mean = 0, sd = 1),
        rnorm(10000, mean = 0, sd = 2)
        )

kurtosis_distributions <- list(leptokurtic, platykurtic)

titles <- kurtosis_distributions %>% map(e1071::kurtosis)
p2.8 <- plot_histo_panel(kurtosis_distributions, titles)
p2.8
```


18.7264786801483
2000
1500
1000
500
0
count
−4
0
4
8
12
x


1.07775156481946
3000
2000
1000
0
count
−5
0
5
x