The University of Sydney
School of Mathematics and Statistics

STUVAC Statistical Snacks 3

MATH1062/MATH1005: Mathematics 1B/Statistical Thinking With Data Semester 1, 2024

Lecturers: J. Baine, T. Cui, J. Spreer, M. Stewart

Two-sample T-tests The Marrickville Golf Club has been collecting weather data (on and off) since 1904. Daily rainfall data may be downloaded by selecting the link "`All years of data`" from **this page** within the Bureau of Meteorology website (**note:** the actual URL of the "`All years of data`" link seems to change from day to day).

```
mgc = read.csv("IDCJAC0009_066036_1800_Data.csv")
str(mgc)
```

```
## 'data.frame': 43980 obs. of  8 variables:
##  $ Product.code                                    : chr  "IDCJAC0009" "IDCJAC0009" "IDCJAC
##  $ Bureau.of.Meteorology.station.number            : int  66036 66036 66036 66036 66036 660
##  $ Year                                            : int  1904 1904 1904 1904 1904 1904 190
##  $ Month                                           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Day                                             : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Rainfall.amount..millimetres.                   : num  NA NA NA NA NA NA NA NA NA NA ..
##  $ Period.over.which.rainfall.was.measured..days.: int  NA NA NA NA NA NA NA NA NA NA ..
##  $ Quality                                         : chr  "" "" "" "" ...
```

The code below determines monthly totals of rainfall (note that `na.rm=T` ignores `NA`s, treating them as zero); rows are years, columns are months:

```
rf = mgc$Rainfall.amount..millimetres.
monthly.rf = tapply(rf, list(mgc$Year, mgc$Month), sum, na.rm=T)
monthly.rf
```

```
##          1     2     3     4     5     6     7     8     9    10    11    12
## 1904   0.0   0.0   0.0   0.0  87.3   0.5 303.1  37.3  26.7  38.6   0.0  25.4
## 1905  42.0  68.1 227.2 151.1 162.3  52.6   9.1   7.6  35.5  46.3   6.4  77.9
## 1906  46.1   7.0 134.5   8.9 115.3  28.0   3.8 121.7  35.2  41.2  93.4  59.7
## 1907  67.9  84.5 192.6  29.7  31.8 195.5   4.5   6.3   5.1   7.2  26.6  50.7
## 1908  25.8 191.5  31.0  55.0  46.4  10.6 212.7 219.3  53.0  17.0   2.0  15.2
## 1909  17.8 148.0  19.1  30.5  20.6 110.0  15.1  29.6 104.0  35.8  72.8  91.5
## 1910 135.2  16.6 156.4  73.2 107.5  61.9 224.8   5.6  48.2  53.3  19.1 149.9
## 1911 348.0 121.3  55.9  61.0  38.1   2.5 163.8 164.4  46.2  19.5  39.3  53.4
## 1912  27.9 158.6 112.1 139.7  80.1  43.2 218.3  54.6  14.0  21.1  74.6  44.1
## 1913  13.8  34.3 257.1 173.1 411.2   0.0 199.7   0.0  42.6  34.3  10.9   8.9
## 1914  14.4  25.0 203.7  40.1  96.8 131.1 220.8  54.4  97.1 148.9  76.1 157.6
## 1915  23.4  26.8  92.2 222.6  96.8  26.2 122.0  26.7  28.6  16.7   0.0  63.8
## 1916  21.8  54.3  52.6 120.1  38.4  42.4  70.6  97.0 123.5 312.4  69.7  78.9
## 1917  56.5 169.9   8.7 319.0  82.8 131.8   9.0  39.6 101.4  97.7 186.4  40.6
## 1918 229.5  90.8  29.0 161.9   7.4   3.9 211.0  46.9  69.9  19.2  20.7  11.4
## 1919  28.9  96.1  84.8  57.4 416.7  32.4  31.7   4.6  79.8  46.6  80.1  63.6
## 1920 137.5  32.3  27.7  59.3   5.1  57.4 125.9  24.7  29.4  24.3  42.2 341.0
## 1921  67.4  23.9  75.6 154.3 145.6  17.1 152.0  24.4  84.1  55.4  73.3 145.0
```



```
## 1922 133.5  75.1  41.4  29.2  89.0  25.0 243.1  41.0  99.5  49.3  14.6  40.7
## 1923  51.6  13.8  20.0 150.7  26.7 101.2 174.2 139.4  42.5  33.2  26.7  41.4
## 1924 111.9  58.5  95.8 134.3  49.3  48.2  38.4  57.1  74.9  25.9  79.6  67.4
## 1925  70.7  41.3  44.1  29.4 430.8 154.0   4.1  86.4  18.0  16.3  99.7  16.0
## 1926   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1927   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1928   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1929   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1930   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1931   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1932   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1933   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1934   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1935   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1936   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1937   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1938   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1939   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1940   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1941   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1942   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1943   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1944   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1945   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1946   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1947   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1948   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0  58.8
## 1949   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1950   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1951   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1952   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1953   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1954   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1955   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1956   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1957   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1958   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1959   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1960   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1961   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1962   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1963   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1964   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1965   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1966  18.8 113.9 203.3 183.7  50.1 113.3  11.4  67.6  50.6  38.0 140.2  68.2
## 1967 153.1 164.4 101.8  46.8  31.0 214.0  20.8 211.9  68.3  55.4  78.6  18.8
## 1968 109.5  15.0  86.1  11.2  85.8  20.3  47.8  22.2   3.1   4.2  17.5  76.1
## 1969  51.7 195.6 104.0 169.7  44.2 149.3  29.8 155.4  45.0  49.0 255.5  37.4
## 1970  94.1  60.4 140.1  58.7  15.2  27.2   0.0  31.9 132.9  18.1   0.0   0.0
## 1971   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1972   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
```

```
## 1973   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1974   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1975   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1976   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1977   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1978   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1979   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1980   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1981   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1982   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1983   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1984   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1985   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1986   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1987   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1988   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1989   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1990   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1991   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1992   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1993   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1994   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1995   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1996   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1997   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1998   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 1999   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 2000   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
## 2001   0.0   0.0   0.0   0.0   0.0   0.0   0.0  39.0  57.0  29.0  74.0  18.0
## 2002  67.0 313.0  33.0  26.0  75.0  18.0   9.0  13.0  20.0  19.0  19.0  88.0
## 2003   5.0  53.0  70.0 242.0 335.0  57.0  42.0  34.0   8.0  61.0  81.0  52.0
## 2004  51.0  43.0   4.0   2.0   8.0  27.0  28.0  86.0  54.0 200.0  42.0  71.0
## 2005  60.0  95.0  45.0  17.0  29.0  64.0  61.0   0.0  40.0  56.0  92.0  37.0
## 2006  61.0  33.0  29.0   2.0  20.0 157.0  98.0  68.0 147.0   8.0  22.0  13.0
## 2007  10.0 105.0  49.0 113.0  12.0 333.0  31.0  99.0  48.0  33.0 122.0  77.0
## 2008  53.0 325.0  64.0 155.0   4.0 109.0  55.0  34.0  73.0  44.0  44.0  73.0
## 2009  18.0 125.0  17.0  50.0 115.0  75.0  48.0   4.0  17.0 158.0  24.0  54.0
## 2010  24.0 164.0  50.0  32.0 149.0  93.0  30.0  18.0  46.0  79.0 165.0  75.0
## 2011  29.0  16.0 153.0 216.0 112.0  63.0 264.0  38.0  71.0  32.0 169.0 154.0
## 2012 112.0 140.0 221.0 163.0  23.0 211.0  52.0   8.0  20.0  24.0  48.0  38.0
## 2013 133.0 169.0  71.0 109.0   2.0 308.0  30.0  14.0  47.0  18.0 183.0  31.0
## 2014   7.0  41.0 117.0  61.0   9.0  84.0  12.0 243.0  48.0 131.0  18.0 161.0
## 2015 159.0 101.0  49.0 413.0 102.0  90.0  54.0  60.0  56.0  31.0  81.0  69.0
## 2016 251.0  29.0 117.0  86.0  13.0 300.0 108.0  82.0   0.0  27.0  34.0  66.0
## 2017  48.0 177.0 267.0  71.0  19.0 116.0  11.0  21.0   0.0  59.0  40.0  52.0
## 2018  25.0 116.0  77.0  15.0  14.0 146.0   8.0   6.0  80.0 181.0 110.0  93.0
## 2019  69.0  88.0 170.0  16.0   9.0 146.0  43.0  53.0 103.0  27.0  25.0   1.0
## 2020  66.0 435.0 138.0  25.0  99.0  82.0 140.0  60.0   6.0   4.0  45.0  89.0
## 2021  71.0 108.0 407.0  27.0  99.0  70.0  32.0  80.0  52.0  61.0 160.0  91.0
## 2022  91.0 394.0 626.0 221.0 178.0   4.0 348.0  42.0  79.0 173.0  43.0  42.0
## 2023 157.0 191.0  61.0 131.0  36.0  15.0  38.0  49.0   0.0  23.0 143.0  81.0
```

```
## 2024  76.0 180.0  51.0 249.0 163.0    NA    NA    NA    NA    NA    NA    NA
```

As can be seen, the rainfall was only recorded May 1904-Dec 1925, then Jan 1966-Oct 1970, then Aug 2001 to the present. Let us extract the rainfall for the month of May in the two periods 1904-1925 and 2002-2024:

```
yrs = as.numeric(rownames(monthly.rf))
yrs

##   [1] 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918
##  [16] 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933
##  [31] 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948
##  [46] 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963
##  [61] 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978
##  [76] 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993
##  [91] 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
## [106] 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023
## [121] 2024

yr.19 = (yrs < 1926)
yr.20 = (yrs > 2001)
may.19 = monthly.rf[yr.19,5]
may.19

##  1904  1905  1906  1907  1908  1909  1910  1911  1912  1913  1914  1915  1916
##  87.3 162.3 115.3  31.8  46.4  20.6 107.5  38.1  80.1 411.2  96.8  96.8  38.4
##  1917  1918  1919  1920  1921  1922  1923  1924  1925
##  82.8   7.4 416.7   5.1 145.6  89.0  26.7  49.3 430.8

may.20 =  monthly.rf[yr.20,5]
may.20

## 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
##   75  335    8   29   20   12    4  115  149  112   23    2    9  102   13   19
## 2018 2019 2020 2021 2022 2023 2024
##   14    9   99   99  178   36  163
```

```
m.19 = mean(may.19)
m.20 = mean(may.20)
s.19 = sd(may.19)
s.20 = sd(may.20)
n.19 = length(may.19)
n.20 = length(may.20)
cbind(m.19, s.19, n.19)

##          m.19     s.19 n.19
## [1,] 117.5455 129.8396   22

cbind(m.20, s.20, n.20)

##          m.20     s.20 n.20
## [1,] 70.65217 80.74967   23
```

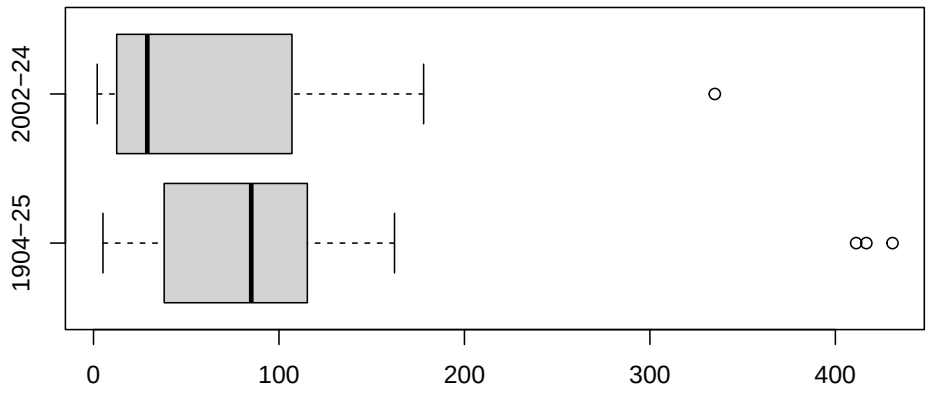
Are the differences in mean rainfall significant? To address this question we assume that

- the 1904-1925 May rainfall totals are like a random sample $X_1, \ldots, X_{22}$ taken with replacement from a box with mean μ_X;
- the 2002-2024 May rainfall totals are like a random sample $Y_1, \ldots, Y_{23}$ taken with replacement from a box with mean μ_X;
- both samples are taken independently of each other.

1. Classical two-sample T-test

(a) State the *additional* assumptions underlying the use of the Classical two-sample T-test. Based on the boxplots below, do these seem reasonable?

```
boxplot(may.19, may.20, names=c("1904-25", "2002-24"), horizontal=T)
```



(b) The test statistic used in the Classical two-sample T-test is the mean difference divided by a particular estimate of the standard error of the mean difference. Determine this estimated standard error. The R code below may be useful:

```
sp = sqrt(( (n.19-1)*(s.19^2) + (n.20-1)*(s.20^2))/(n.19+n.20-2))
sp
```

```
## [1] 107.5603
```

(c) Based on the R output below, is the apparent mean difference significantly different from zero at the 5% level? What about at the 10% level? Be sure to formally state the hypotheses being considered here. (**Hint:** is this a one-sided or two-sided test?)

```
qt(.9, df=43)
```

```
## [1] 1.301552
```

```
qt(.95, df=43)
```

```
## [1] 1.681071
```

```
qt(.975, df=43)
```

```
## [1] 2.016692
```

2. **Welch test**
   (a) This test relaxes one of the assumptions of the Classical two-sample T-test. Which one exactly?
   (b) This test uses a different estimate of the standard error of the mean difference, compared to the Classical two-sample T-test. Determine the value of this estimated standard error.

3. **Simulation-based two-sample T-test** Consider the R output below:

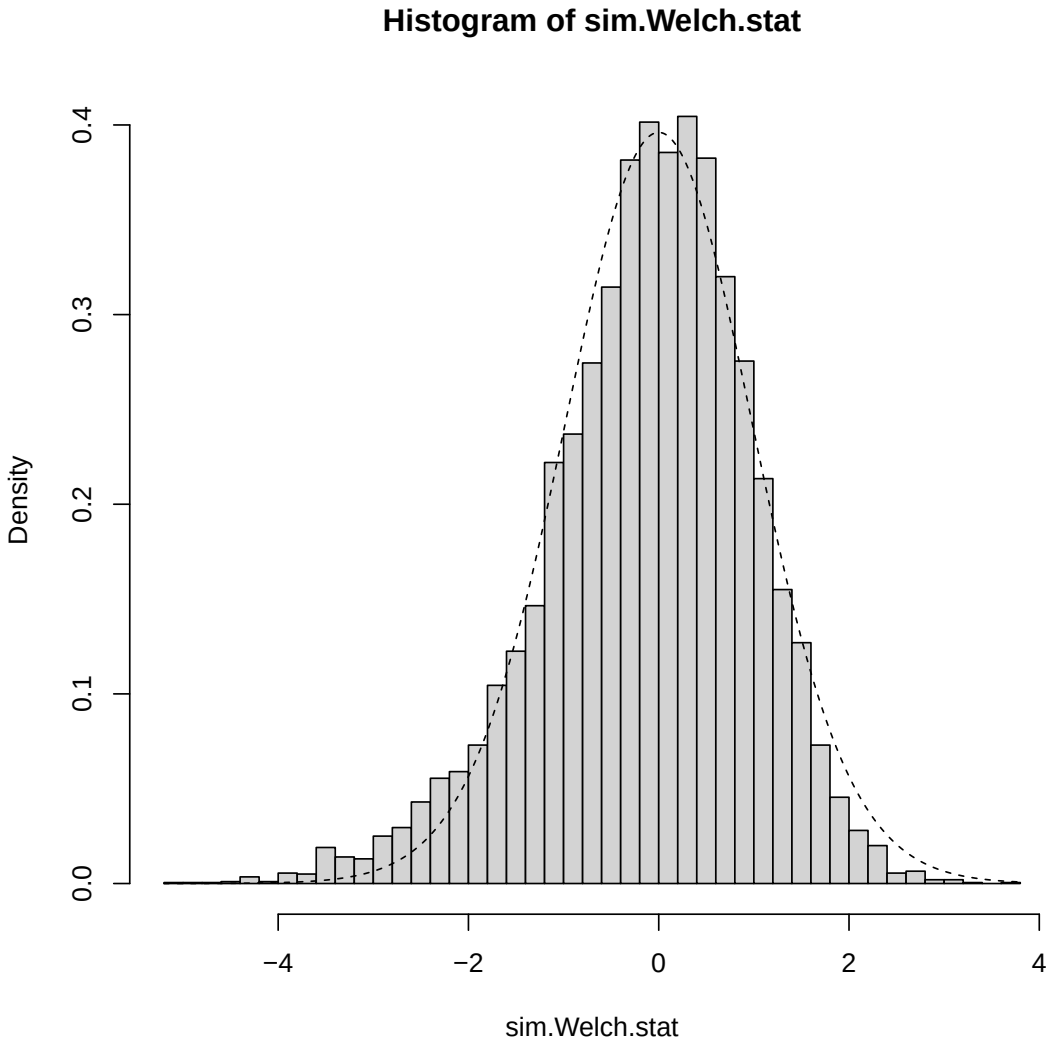
```
t.test(may.19, may.20)


##
##  Welch Two Sample t-test
##
## data:  may.19 and may.20
## t = 1.4473, df = 34.859, p-value = 0.1567
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.89264 112.67920
## sample estimates:
## mean of x mean of y
## 117.54545  70.65217
```

Note the "data-dependent" degrees of freedom here, 34.859, quite a bit smaller than 43, the degrees of freedom for the Classical two-sample T-test. This is perhaps a reflection of the fact that the sample SDs are quite different.

(a) Are the remaining assumptions underlying the Welch test reasonable in this case?

Consider the simulation performed using the code below:

```
n.19 = length(may.19)
n.20 = length(may.20)
box.19 = may.19-mean(may.19)
box.20 = may.20-mean(may.20)
sim.Welch.stat=0
for(i in 1:10000){
  samp.19 = sample(box.19, size=n.19, replace=T)
  samp.20 = sample(box.20, size=n.20, replace=T)
  sim.Welch.stat[i] = t.test(samp.19, samp.20)$stat
}
hist(sim.Welch.stat,pr=T, n=50)
curve(dt(x, df=34.436), lty=2, add=T, n=1001)
```




```
sum(sim.Welch.stat>=1.4473)
```

```
## [1] 558
```

```
sum(sim.Welch.stat<=-1.4473)
```

```
## [1] 1093
```

(b) Note that we define `box.19 = may.19-mean(may.19)` and `box.20 = may.20-mean(may.20)`. Why are the means subtracted off?

(c) We use the simulated values in `sim.Welch.stat` to approximate the distribution of the test statistic when the null hypothesis is true. What is a simulation-based P-value?