DATA2002

Wilcoxon rank-sum test

Garth Tarr

THE UNIVERSITY OF
SYDNEY

Wilcoxon rank-sum test

Wilcoxon rank-sum test

Also known as Mann-Whitney U test or
Mann-Whitney-Wilcoxon test or
Wilcoxon-Mann-Whitney test

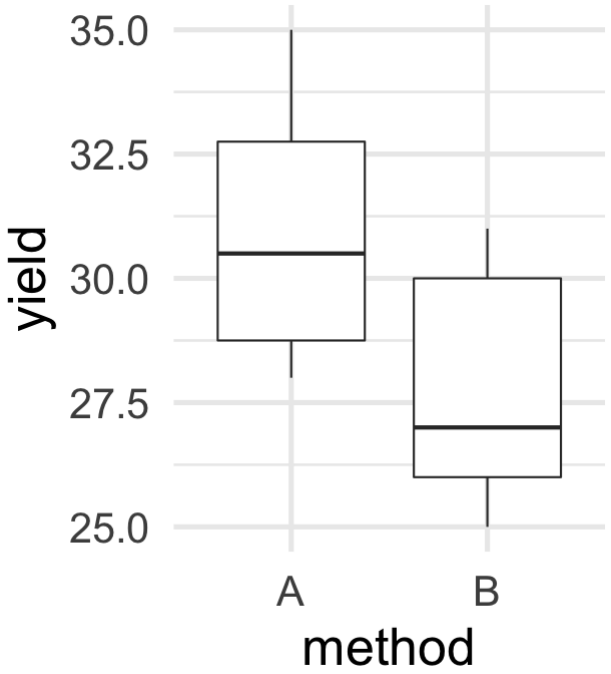
Yield

The following data yield measurements by two different methods.

```
A = c(32, 29, 35, 28)
B = c(27, 31, 26, 25, 30)
dat = data.frame(
  yield = c(A,B),
  method = c(rep("A", length(A)),
             rep("B", length(B)))
)
```

If the normality assumptions are in doubt, does the data present sufficient evidence to indicate a difference in the methods A and B?

```
# boxplot(dat$yield~dat$method, ylab = "Yield")
library(ggplot2)
ggplot(dat, aes(x = method, y = yield)) +
  geom_boxplot() +
  theme_minimal(base_size = 26)
```



Wilcoxon rank-sum test

- A non-parametric test to compare means of two independent samples
- Relaxes normality assumption (like sign and Wilcoxon sign-rank tests)
- Also relaxes the assumption of symmetry

Wilcoxon rank-sum test

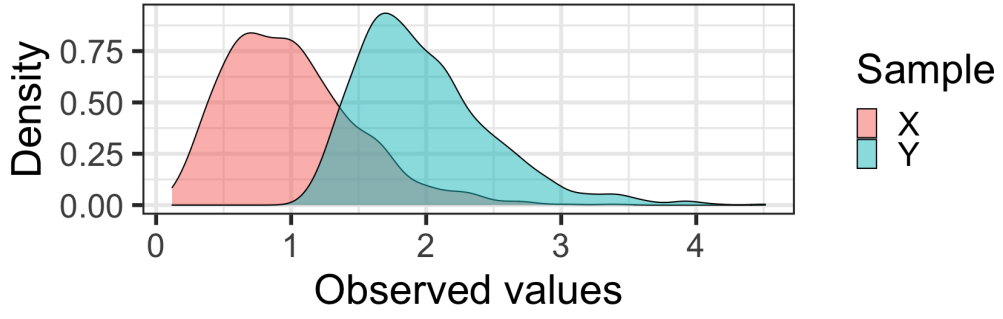
Suppose the samples $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are taken from two distinct populations that follow the same kind of distribution but differ in location.

That is, $\mu_x = \mu_y + \theta$, where μ_x is the population mean of X, μ_y is the population mean of Y and θ is a location shift parameter.

```
library(tidyverse)
N = 1000
theta = 1
loc = data.frame(
  obs = c(rgamma(N,4,4), rgamma(N,4,4)+theta),
  group = rep(c("X","Y"), each = N)
)
loc %>% group_by(group) %>%
  summarise(Mean = mean(obs), SD = sd(obs))
```

```
## # A tibble: 2 × 3
##   group  Mean    SD
##   <chr> <dbl> <dbl>
## 1 X      1.00 0.487
## 2 Y      1.98 0.501
```

```
loc %>% ggplot() +
  aes(x = obs, fill = group) +
  geom_density(alpha = 0.5) +
  labs(x = "Observed values",
       y = "Density",
       fill = "Sample")+
  theme_bw(base_size = 34)
```



Wilcoxon rank-sum test

Let $R_1, R_2, \ldots, R_N$ with $N = n_x + n_y$ be the ranks of combined sample: $X_1, X_2, \ldots, X_{n_x}, Y_1, Y_2, \ldots, Y_{n_y}$.

- For one sample **Wilcoxon signed-rank** test, the ranks are summed over positive side of the differences
- For two sample **Wilcoxon rank-sum** test, the ranks are summed over one of the samples.
- i.e. $W = R_1 + R_2 + \ldots + R_{n_x}$

If H_0 is true, then W should be close to its expected value

$$\mathrm{E}(W) = \text{Proportion} \times \text{Total rank sum} = \frac{n_x}{N} \times \frac{N(N+1)}{2} = \frac{n_x(N+1)}{2}.$$

If W is small (large), we expect $\mu_x < \mu_y$ ($\mu_x > \mu_y$).

Suppose $X_1, \ldots, X_n$ are drawn from some population. Given a significance level α, we want to test on the population mean.

- **Hypothesis:** H_0: $\mu_x = \mu_y$ vs H_1: $\mu_x > \mu_y$, $\mu_x < \mu_y$, $\mu_x \neq \mu_y$
- **Assumptions:** X_i and Y_i are independent and follow the same distribution but differ by a shift.
- **Test statistic:** $W = R_1 + R_2 + \ldots + R_{n_x}$. Under H_0, W follows the $WRS(n_X, n_Y)$ distribution.
- **Observed test statistic:** $w = r_1 + r_2 + \ldots + r_{n_x}$
- **p-value:**

$$P(W \geq w) \text{ for } H_1\text{: } \mu_x > \mu_y \quad \text{or} \quad P(W \leq w) \text{ for } H_1\text{: } \mu_x < \mu_y$$

$$2P(W \geq w) \text{ if } w > \frac{n_x(N+1)}{2} \text{ and } H_1\text{: } \mu_x \neq \mu_y$$

$$2P(W \leq w) \text{ if } w < \frac{n_x(N+1)}{2} \text{ and } H_1\text{: } \mu_x \neq \mu_y$$

- **Decision:** If p-value is less than α, there is evidence against H_0. If the p-value is greater than α, the data are consistent with H_0.

Calculate p-value: no ties on the data

The exact p-value $P(W \leq w)$ is given by in R by

```
pwilcox(w - minw, m = nx, n = ny)
```

where $\min(W) = \underbrace{1 + 2 + \ldots + n_x}_{n_x} = \frac{n_x(n_x + 1)}{2}$ (the smallest possible sum of ranks for the X sample).

The distribution that `pwilcox()` uses is for the distribution of $W - \min(W)$ (and so starts at 0).

Yield example

The following data yield measurements by two different methods.

```
A = c(32, 29, 35, 28)
B = c(27, 31, 26, 25, 30)
dat = data.frame(
  yield = c(A,B),
  method = c(rep("A", length(A)),
             rep("B", length(B)))
)
```

If the normality assumptions are in doubt, does the data present sufficient evidence to indicate a difference in the methods A and B?

```
dat = dat %>% mutate(rank = rank(yield))
dat
```

```
##   yield method rank
## 1    32      A    8
## 2    29      A    5
## 3    35      A    9
## 4    28      A    4
## 5    27      B    3
## 6    31      B    7
## 7    26      B    2
## 8    25      B    1
## 9    30      B    6
```

```
w_A = dat %>%
  filter(method == "A") %>%
  pull(rank) %>%
  sum()
w_A
```

```
## [1] 26
```

- **Hypothesis:** H_0: $\mu_A = \mu_B$ vs H_1: $\mu_A \neq \mu_B$
- **Assumptions:** A_i and B_i are independent and follow the same kind of distribution but differ by a shift.
- **Test statistic:** $W = R_1 + R_2 + \ldots + R_{n_A}$. Under H_0, W follows the $WRS(4, 5)$ distribution.
- **Observed test statistic:** $w = 26$ (sum of the ranks associated with method A).
- **P-value:** $2P(W \geq w) = 0.19$ because $w = 26 > \frac{n_A(N+1)}{2} = 20$ so we're looking in the upper tail.
- **Decision:** As the p-value is greater than 0.05, the data are consistent with H_0.

```r
sum_dat = dat %>%
  group_by(method) %>%
  summarise(n = n(),
            w = sum(rank))
sum_dat
```

```
## # A tibble: 2 × 3
##   method     n     w
##   <chr>  <int> <dbl>
## 1 A          4    26
## 2 B          5    19
```

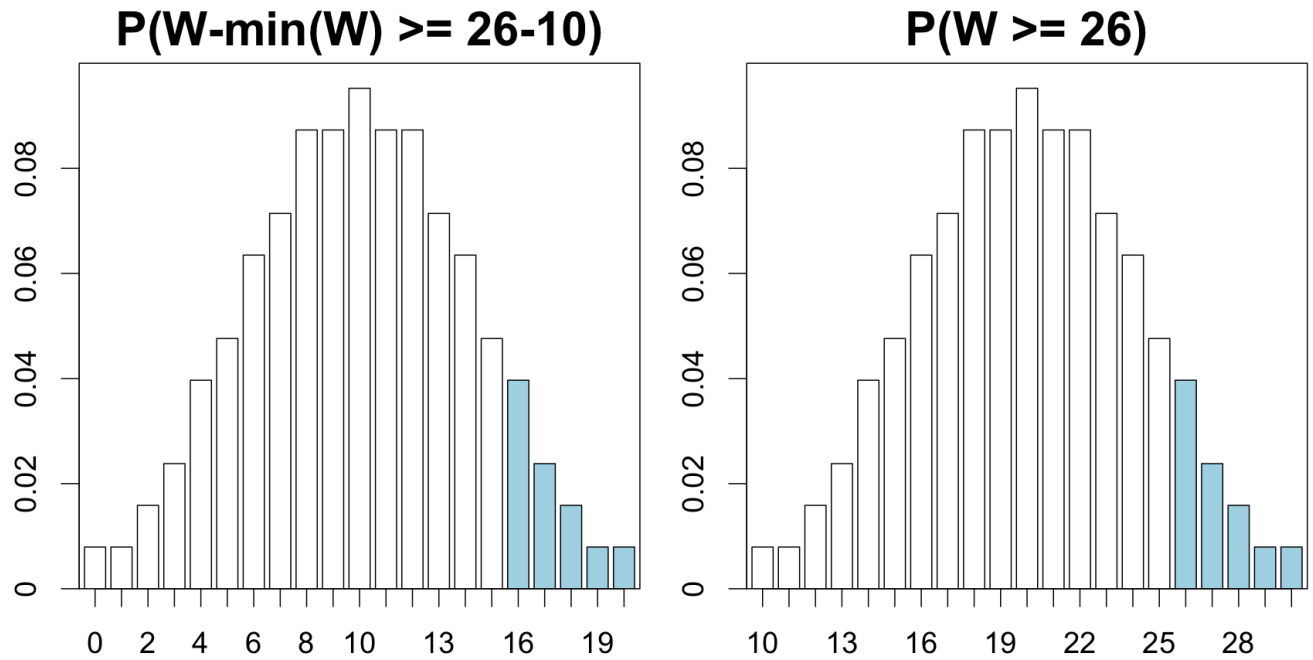
```r
n_A = sum_dat %>%
  filter(method == "A") %>%
  pull(n)
n_B = sum_dat %>%
  filter(method == "B") %>%
  pull(n)
# using the sums of the A sample
w_A = sum_dat %>%
  filter(method == "A") %>%
  pull(w)
ew_A = n_A * (n_A + n_B + 1)/2
minw_A = n_A * (n_A + 1)/2
```

```r
c(minw_A, w_A, ew_A) # w_A > ew_A
```

```
## [1] 10 26 20
```

```r
# looking in the upper tail, so use lower.tail = FALSE
2 * pwilcox(w_A - minw_A - 1, n_A, n_B, lower.tail = FALSE)
```

```
## [1] 0.1904762
```



What if we use the sums of the ranks of the B sample?

```
sum_dat
```

```
## # A tibble: 2 × 3
##   method     n     w
##   <chr>  <int> <dbl>
## 1 A          4    26
## 2 B          5    19
```

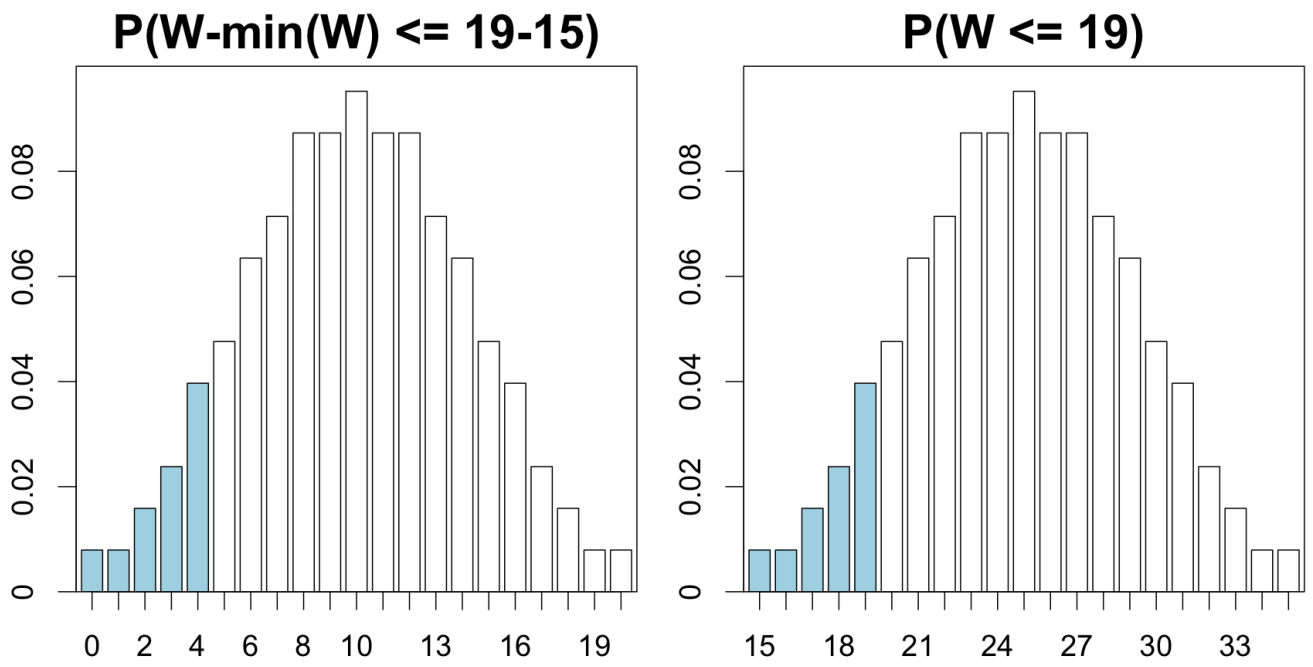
```
# using the sums of the B sample
w_B = sum_dat %>%
  filter(method == "B") %>%
  pull(w)
ew_B = n_B * (n_B + n_A + 1)/2
minw_B = n_B * (n_B + 1)/2
c(minw_B, w_B, ew_B)
```

```
## [1] 15 19 25
```

Note that $w_B < E(W_B)$ so we need to look in the lower tail of the distribution.

```
# now looking in the lower tail
2 * pwilcox(w_B - minw_B, n_B, n_A)
```

```
## [1] 0.1904762
```



Yield

```
wilcox.test(A, B) # wilcox.test(yield ~ method, data = dat)
```

```
##
##     Wilcoxon rank sum exact test
##
## data:  A and B
## W = 16, p-value = 0.1905
## alternative hypothesis: true location shift is not equal to 0
```

```
t.test(A, B) # t.test(yield ~ method, data = dat)
```

```
##
##     Welch Two Sample t-test
##
## data:  A and B
## t = 1.633, df = 5.8232, p-value = 0.1551
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.630468  8.030468
## sample estimates:
## mean of x mean of y
##      31.0      27.8
```

p-value calculation: when there are ties

- The p-value can be calculated using normal approximation to the distribution of test statistic:

$$T = \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \sim \mathcal{N}(0,1) \text{ approximately,}$$

where $\mathrm{E}(W) = \dfrac{n_x(N+1)}{2}$ and $\mathrm{Var}(W) = \dfrac{n_x n_y}{N(N-1)} \left(\sum_{i=1}^{N} r_i^2 - \dfrac{N(N+1)^2}{4} \right)$.

- Our p-value calculations are:
- $\text{p-value} \approx P\left(Z \geq \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \right)$ for $H_1\colon\ \mu_x > \mu_y$
- $\text{p-value} \approx P\left(Z \leq \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \right)$ for $H_1\colon\ \mu_x < \mu_y$
- $\text{p-value} \approx 2P\left(Z \geq \left| \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \right| \right)$ for $H_1\colon\ \mu_x \neq \mu_y$.

- We use normal approximation for the test statistic W NOT the data $X_i,\ Y_i$.
- As we do not consider sign, zero measurements should be ranked in the same way as other measurements (clear membership to either sample).

Latent heat of fusion

Natrella (1963; page 3-23) presents data from two methods that were used in a study of the latent heat latent heat of fusion of ice. Both method A (digital method) and Method B (method of mixtures) were conducted with the specimens cooled to -0.72°C. The data represent the change in total heat from -0.72°C to water at 0°C, in calories per gram of mass.

> Does the data support the hypothesis that the electrical method (method A) gives larger results?

```
A = c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
      79.97, 80.05, 80.03, 80.02, 80.00, 80.02)
B = c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95,
      79.97)
heat = data.frame(
  energy = c(A,B),
  method = rep(c("A","B"), c(length(A), length(B))))
```

```
heat = heat %>%
  dplyr::mutate(r = rank(energy))
# or equivalently
# heat$rank = rank(heat$energy)
heat %>% arrange(r) %>%
  rmarkdown::paged_table(
    options = list(rows.print = 8))
```

energy <dbl>	method <chr>	r <dbl>
79.94	B	1.0
79.95	B	2.0
79.97	A	4.5
79.97	B	4.5
79.97	B	4.5
79.97	B	4.5
79.98	A	7.5
79.98	B	7.5

1-8 of 21 r... Previous **1** 2 3 Next

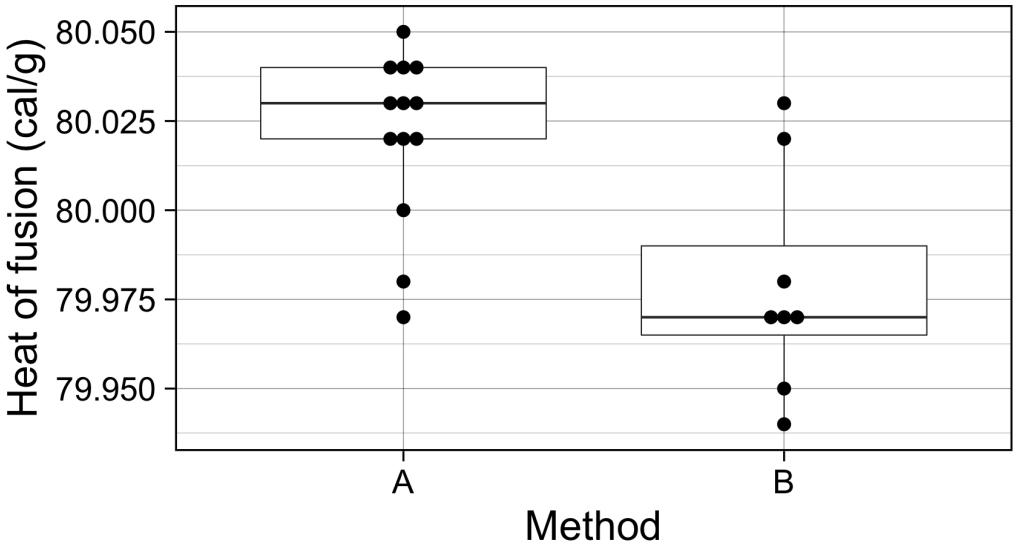
Source: Rice (2006; page 423)

Latent heat of fusion

```
heat %>%
  dplyr::group_by(method) %>%
  dplyr::summarise(
    w = sum(r),
    xbar = mean(energy),
    s = sd(energy),
    n = n()
  ) %>%
  knitr::kable(format = "markdown",
               digits = 3)
```

method	w	xbar	s	n
A	180	80.021	0.024	13
B	51	79.979	0.031	8

```
ggplot(heat, aes(x = method, y = energy)) +
  geom_boxplot() +
  geom_dotplot(stackdir = "center",
               binaxis = "y") +
  theme_linedraw(base_size = 34) +
  labs(y = "Heat of fusion (cal/g)",
       x = "Method")
```



We have $n_A = 13, n_B = 8$ and $N = 21$. The Wilcoxon rank-sum test for the difference between methods A and B is

- **Hypothesis:** H_0: $\mu_A = \mu_B$ vs H_1: $\mu_A > \mu_B$
- **Assumptions:** A_i and B_i are independent and follow the same kind of distribution but differ by a shift.
- **Test statistic:** $W = R_1 + R_2 + \ldots + R_{n_A}$ (the sum of the ranks of observations in method A). Under H_0, $W \sim \text{WRS}'(13, 8)$, the WRS distribution with sizes 13, 8 and with ties as shown.
- **Observed Test statistic:** $w = r_1 + r_2 + \ldots + r_{n_A} = 180$
- **p-value:** As the exact $\text{WRS}'(13, 8)$ distribution with ties is unknown, we use a normal approximation to this distribution with $\text{E}(W) = \frac{n_x(N+1)}{2} = \frac{13 \times (13+8+1)}{2} = 143$ and $\text{Var}(W) = \frac{n_x n_y}{N(N-1)} \left(\sum_{i=1}^{N} r_i^2 - \frac{N(N+1)^2}{4} \right) = \frac{13(8)(3293.5-2541)}{21(20)} = 186.33$
- $\text{p-value} = P(W \geq w) \simeq P\left(Z \geq \frac{w - \text{E}(W)}{\sqrt{\text{Var}(W)}} \right) = P\left(Z \geq \frac{180-143}{\sqrt{186.33}} \right) = P(Z > 2.7) = 0.003$
- **Decision:** As the p-value is less than 0.05, there is sufficient evidence to reject H_0.

```
heat_sum = heat %>%
  dplyr::group_by(method) %>%
  dplyr::summarise(
    w = sum(r),
    xbar = mean(energy),
    s = sd(energy),
    n = n()
  )
heat_sum
```

```
## # A tibble: 2 × 5
##   method     w  xbar      s     n
##   <chr>  <dbl> <dbl>  <dbl> <int>
## 1 A        180  80.0 0.0240    13
## 2 B         51  80.0 0.0314     8
```

```
na = heat_sum$n[heat_sum$method == "A"]
nb = heat_sum$n[heat_sum$method == "B"]
N = na + nb # total number of observations
c(na, nb, N)
```

```
## [1] 13  8 21
```

```
w = heat_sum$w[heat_sum$method == "A"]
EW = na * (N + 1)/2
c(w, EW)
```

```
## [1] 180 143
```

```
sumsqrank = sum(heat$r^2)
g = N * (N + 1)^2/4
varW = na * nb * (sumsqrank - g)/(N * (N - 1))
t0 = (w - EW)/sqrt(varW)
t0
```

```
## [1] 2.710544
```

```
1 - pnorm(t0)
```

```
## [1] 0.003358648
```

```r
wilcox.test(A, B, alternative = 'greater', correct = FALSE)
```

```
## Warning in wilcox.test.default(A, B, alternative =
## "greater", correct = FALSE): cannot compute exact p-value
## with ties
```

```
##
##     Wilcoxon rank sum test
##
## data:  A and B
## W = 89, p-value = 0.003359
## alternative hypothesis: true location shift is greater than 0
```

```r
t.test(A, B, alternative = 'greater')
```

```
##
##     Welch Two Sample t-test
##
## data:  A and B
## t = 3.2499, df = 12.027, p-value = 0.00347
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.01897943        Inf
## sample estimates:
## mean of x mean of y
```

Robustness properties

What happens if there is an outlier in the data?

```
# change the first value for the B method
heat1 = heat
heat1$energy[14] = 80.20 # instead of 80.02
# recalculate ranks
heat1 = heat1 %>% dplyr::mutate(
  r = rank(energy)
)
```

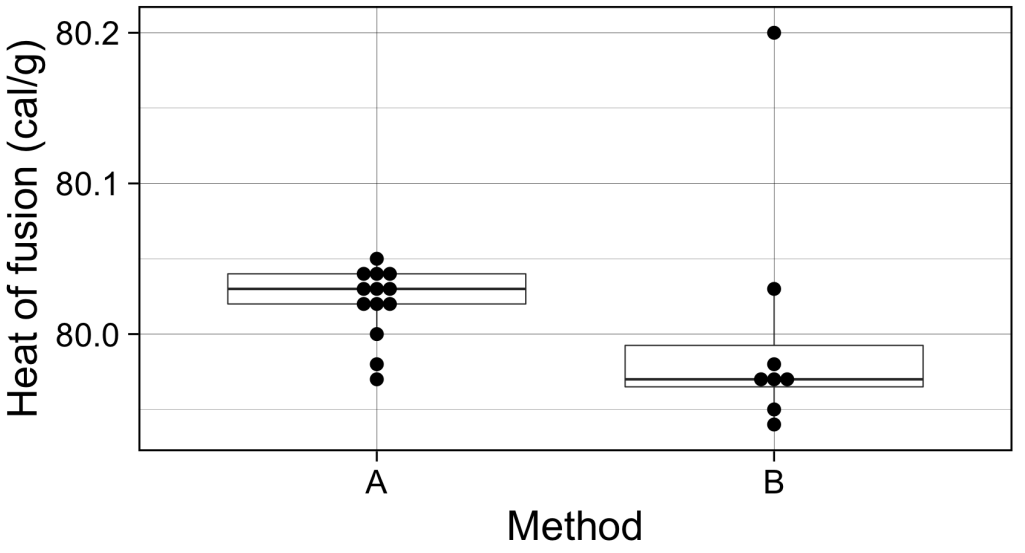
```
heat1
```

```
##    energy method    r
## 1   79.98      A  7.5
## 2   80.04      A 18.0
## 3   80.02      A 11.0
## 4   80.04      A 18.0
## 5   80.03      A 14.5
## 6   80.03      A 14.5
## 7   80.04      A 18.0
## 8   79.97      A  4.5
## 9   80.05      A 20.0
## 10  80.03      A 14.5
## 11  80.02      A 11.0
## 12  80.00      A  9.0
## 13  80.02      A 11.0
## 14  80.20      B 21.0
## 15  79.94      B  1.0
## 16  79.98      B  7.5
## 17  79.97      B  4.5
## 18  79.97      B  4.5
## 19  80.03      B 14.5
## 20  79.95      B  2.0
```

```
heat1 %>%
  dplyr::group_by(method) %>%
  dplyr::summarise(
    w = sum(r),
    xbar = mean(energy),
    s = sd(energy),
    n = n()
  ) %>%
  knitr::kable(format = "markdown",
               digits = 3)
```

method	w	xbar	s	n
A	171.5	80.021	0.024	13
B	59.5	80.001	0.085	8

```
ggplot(heat1, aes(x = method, y = energy)) +
  geom_boxplot() +
  geom_dotplot(stackdir = "center",
               binaxis = "y") +
  theme_linedraw(base_size = 34) +
  labs(y = "Heat of fusion (cal/g)",
       x = "Method")
```



```r
wilcox.test(energy ~ method, data = heat1, alternative = 'greater', correct = FALSE)
```

```
## Warning in wilcox.test.default(x = c(79.98, 80.04, 80.02,
## 80.04, 80.03, : cannot compute exact p-value with ties
```

```
##
##     Wilcoxon rank sum test
##
## data:  energy by method
## W = 80.5, p-value = 0.01859
## alternative hypothesis: true location shift is greater than 0
```

```r
t.test(energy ~ method, data = heat1, alternative = 'greater')
```

```
##
##     Welch Two Sample t-test
##
## data:  energy by method
## t = 0.63712, df = 7.6977, p-value = 0.2713
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.03774242         Inf
## sample estimates:
## mean in group A mean in group B
##        80.02077        80.00125
```

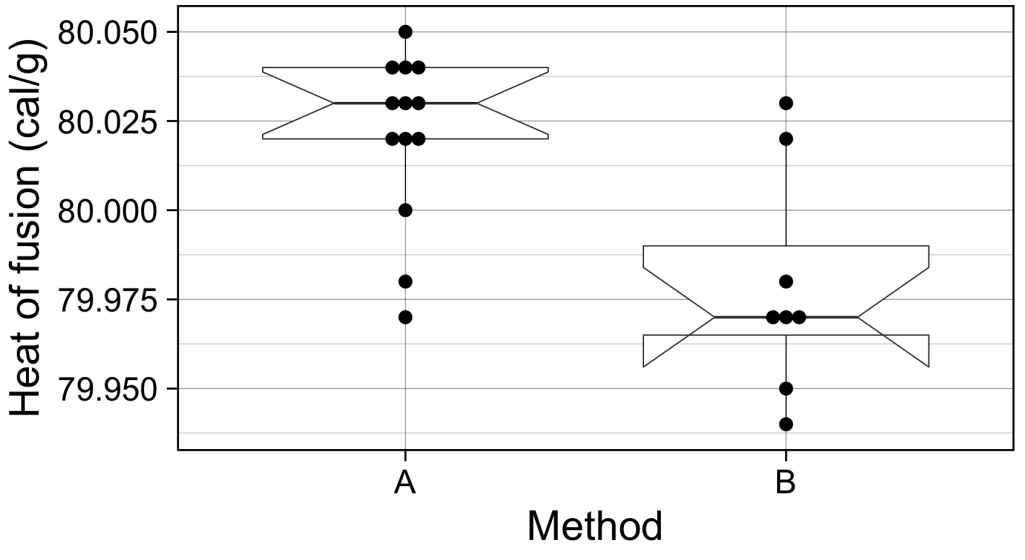
A heuristic for testing for differences: notched boxplot

The upper and lower edges of the notches are at

$$\text{median} \pm 1.58 \times \frac{\text{IQR}}{\sqrt{n}}$$

Rule of thumb: if the notches of two boxes do not overlap, this suggests that the medians are significantly different (Mcgill, Tukey, and Larsen, 1978).

```
ggplot(heat, aes(x = method, y = energy)) +
  geom_boxplot(notch = TRUE) +
  geom_dotplot(stackdir = "center",
               binaxis = "y") +
  theme_linedraw(base_size = 34) +
  labs(y = "Heat of fusion (cal/g)",
       x = "Method")
```



Final comments

- The Wilcoxon rank-sum test is valid for data from any distribution, whether normal or not, and is much less sensitive to outliers than the two-sample t-test.
- If one is primarily interested in differences in location between the two distributions, the Wilcoxon test has the disadvantage of also reacting to other differences between the distributions such as differences in shape.
- When the assumptions of the two-sample t-test hold, the Wilcoxon rank-sum test is somewhat less likely to detect a location shift than is the two-sample t-test (i.e. less powerful).
- In a practical situation in which we are uneasy about the applicability of two-sample t-test, we use both them and the Wilcoxon and feel happiest when both give similar conclusions. 😄

Further reading

Larsen and Marx (2012; section 14.3).

Larsen, R. J. and M. L. Marx (2012). *An Introduction to Mathematical Statistics and its Applications*. 5th ed. Boston, MA: Prentice Hall. ISBN: 978-0-321-69394-5.

Mcgill, R., J. W. Tukey, and W. A. Larsen (1978). "Variations of Box Plots". In: *The American Statistician* 32.1, pp. 12-16. DOI: 10.1080/00031305.1978.10479236.

Natrella, M. G. (1963). *Experimental Statistics*. Vol. 91. National Bureau of Standards Handbook. United States Department of Commerce.

Rice, J. (2006). *Mathematical Statistics and Data Analysis*. Advanced series. Cengage Learning. ISBN: 9780534399429.