Permutation-Based and Rank-Based Methods

Yibi Huang

Data	Methods		
	Normal-based	Permutation-based	Rank-based
Two-sample	two-sample t-test Welch t-test	permutation test	rank-sum test
Multi-sample	ANOVA F-test	permutation test	Kruskal-Wallis test
Matched-pair	paired t-test	permutation test	signed-rank test

Two-Sample t-Test when $\sigma_1^2 = \sigma_2^2$ (Review)

Two-Sample Data:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij}\text{'s are i.i.d.} \sim N(0, \sigma_i^2)$$
$$\text{for } i = 1, 2,\ j = 1, \ldots, n_i$$

To test H$_0$: $\mu_1 = \mu_2$ v.s. H$_a$: $\mu_1 \neq \mu_2$, assuming $\sigma_1^2 = \sigma_2^2$, the t-statistic is

$$t = \frac{\overline{y}_{1\bullet} - \overline{y}_{2\bullet}}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \text{ under } \mathrm{H}_0,$$

where

$$s_p^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_{1\bullet})^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y}_{2\bullet})^2}{n_1 + n_2 - 2} = MSE,$$

called the "pooled sample variance", is an estimate of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$.

Welch t-Test when $\sigma_1^2 \neq \sigma_2^2$ (Review)

When $\sigma_1^2 \neq \sigma_2^2$, we use the Welch t-statistic

$$t = \frac{\overline{y}_{1\bullet} - \overline{y}_{2\bullet}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where σ_1^2 and σ_2^2 are estimated separately using the sample variances, s_1^2 and s_2^2, of the 2 groups,

$$s_1^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_{1\bullet})^2}{n_1 - 1} \quad \text{and} \quad s_2^2 = \frac{\sum_{j=1}^{n_2}(y_{2j} - \overline{y}_{2\bullet})^2}{n_2 - 1}$$

The Welch t-statistic has an approximate (not exact) t-distribution with df $= \nu$ where

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2}.$$

Limitations of Two-Sample t-test and Welch t-test

When can the two-sample t-test or the Welch t-test be used?

- First of all, the observations must be **independent**

Limitations of Two-Sample *t*-test and Welch *t*-test

When can the two-sample *t*-test or the Welch *t*-test be used?

- First of all, the observations must be **independent**
- Conditions on normality and sample size:

	Normal Data?	
	Yes	*No*
Large sample	✓	✓
Small sample	✓	No!

Limitations of Two-Sample *t*-test and Welch *t*-test

When can the two-sample *t*-test or the Welch *t*-test be used?

- First of all, the observations must be **independent**
- Conditions on normality and sample size:

	Normal Data?	
	Yes	*No*
Large sample	✓	✓
Small sample	✓	No!

- Large experiments are often time-consuming and expensive and hence may not be affordable.
  Need tools for small-sample data

Limitations of Two-Sample t-test and Welch t-test

When can the two-sample t-test or the Welch t-test be used?

- First of all, the observations must be **independent**
- Conditions on normality and sample size:

	Normal Data?	
	Yes	*No*
Large sample	✓	✓
Small sample	✓	No!

- Large experiments are often time-consuming and expensive and hence may not be affordable.
  Need tools for small-sample data
- Practically, we are almost never certain about the normality of data.
  Need at least a moderate sample size to check normality
  Hard to check normality when the sample size is small

Permutation Test

Example: Rat's Diet Experiment

- Objective: to investigate the effect of high protein diet on weight gain.

- 8 rats available, randomly choose 4 to be fed with beef, the remaining 4 fed with cereal.

- Response: weight gain (in grams) over a period of time.

- Results:

Protein source	*Weight gain*				*Mean*	*SD*
Cereal	111	56	86	92	86.25	22.81
Beef	104	118	117	111	112.50	6.54

- Questions: Does beef diet yield higher weight gain than cereal diet?

- *t*-tests is not reliable as the sample size is very small

Permutation Test

Under the H_0 that beef or cereal diet makes no difference, the weight gain of 8 rats would remain to be

$$\{111, 56, 86, 92, 104, 118, 117, 111\}$$

no matter they were given beef or cereal. The variation in weight gain is simply the natural rat-to-rat variation. Some rats grow faster, some slower.

[1] To distinguish the two rats of the same weight gain 111 g, we write their weight gains as 111a and 111b.

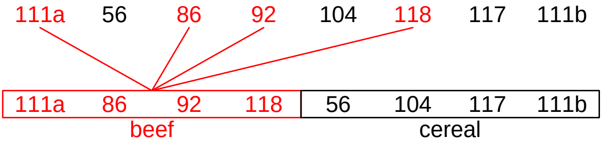
Permutation Test

Under the H_0 that beef or cereal diet makes no difference, the weight gain of 8 rats would remain to be

$$\{111, 56, 86, 92, 104, 118, 117, 111\}$$

no matter they were given beef or cereal. The variation in weight gain is simply the natural rat-to-rat variation. Some rats grow faster, some slower.

As the 8 rats were **randomly** allocated to the beef group or the cereal group (4 rats each), one possible allocation and outcome could be[1]



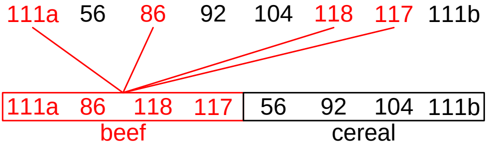
111a 56 86 92 104 118 117 111b

111a	86	92	118	56	104	117	111b

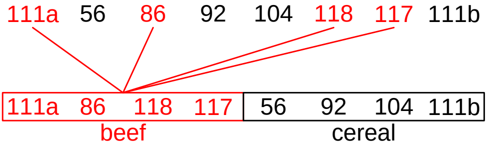
beef cereal

[1] To distinguish the two rats of the same weight gain 111 g, we write their weight gains as 111a and 111b.

Another possible allocation/outcome:



Another possible allocation/outcome:

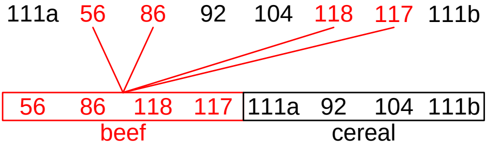


111a 56 86 92 104 118 117 111b

111a 86 118 117 | 56 92 104 111b

beef | cereal

One more possible allocation/outcome:

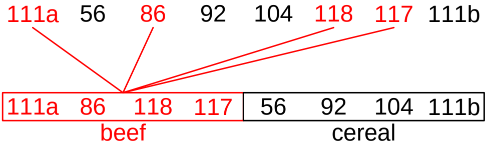


111a 56 86 92 104 118 117 111b

56 86 118 117 | 111a 92 104 111b

beef | cereal

Another possible allocation/outcome:

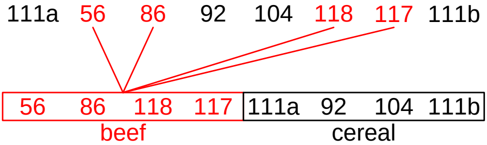


111a 56 86 92 104 118 117 111b

111a 86 118 117 | 56 92 104 111b

beef cereal

One more possible allocation/outcome:



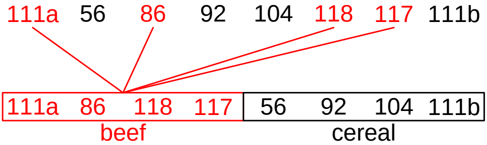
111a 56 86 92 104 118 117 111b

56 86 118 117 | 111a 92 104 111b

beef cereal

- In total, there are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ possible allocations, shown in the next page.

Another possible allocation/outcome:

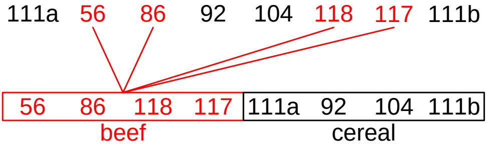

111a 56 86 92 104 118 117 111b

111a 86 118 117 | 56 92 104 111b

beef | cereal

$$D = \overline{y}_{beef} - \overline{y}_{cereal}$$
$$= \frac{111+86+118+117}{4} - \frac{56+92+104+111}{4}$$
$$= -17.25$$

One more possible allocation/outcome:


111a 56 86 92 104 118 117 111b

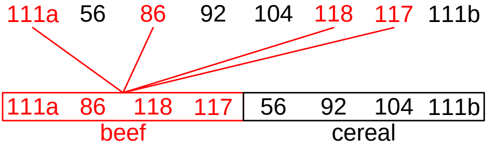
56 86 118 117 | 111a 92 104 111b

beef | cereal

- In total, there are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ possible allocations, shown in the next page.
- For each allocation, we calculate the test statistic

$$D = \overline{y}_{beef} - \overline{y}_{cereal}$$

Another possible allocation/outcome:

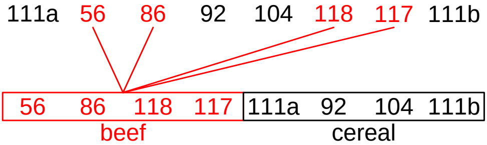

111a 56 86 92 104 118 117 111b

111a 86 118 117 | 56 92 104 111b

beef | cereal

$$\begin{aligned} D &= \overline{y}_{beef} - \overline{y}_{cereal} \\ &= \frac{111+86+118+117}{4} - \frac{56+92+104+111}{4} \\ &= -17.25 \end{aligned}$$

One more possible allocation/outcome:


111a 56 86 92 104 118 117 111b

56 86 118 117 | 111a 92 104 111b

beef | cereal

$$\begin{aligned} D &= \overline{y}_{beef} - \overline{y}_{cereal} \\ &= \frac{56+86+118+117}{4} - \frac{111+92+104+111}{4} \\ &= -10.25 \end{aligned}$$

- In total, there are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ possible allocations, shown in the next page.
- For each allocation, we calculate the test statistic

$$D = \overline{y}_{beef} - \overline{y}_{cereal}$$

									if 2 groups swapped
beef diet				cereal diet				$\overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$	$\overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$
118	117	111a	111b	104	92	86	56	29.75	−29.75
118	117	111a	104	111b	92	86	56	26.25	−26.25
118	117	111a	92	111b	104	86	56	20.25	−20.25
118	117	111a	86	111b	104	92	56	17.25	−17.25
118	117	111a	56	111b	104	92	86	2.25	−2.25
118	117	111b	104	111a	92	86	56	26.25	−26.25
118	117	111b	92	111a	104	86	56	20.25	−20.25
118	117	111b	86	111a	104	92	56	17.25	−17.25
118	117	111b	56	111a	104	92	86	2.25	− 2.25
118	117	104	92	111a	111b	86	56	16.75	−16.75
118	117	104	86	111a	111b	92	56	13.75	−13.75
118	117	104	56	111a	111b	92	86	−1.25	1.25
118	117	92	86	111a	111b	104	56	7.75	−7.75
118	117	92	56	111a	111b	104	86	−7.25	7.25
118	117	86	56	111a	111b	104	92	−10.25	10.25
118	111a	111b	104	117	92	86	56	23.25	−23.25
118	111a	111b	92	117	104	86	56	17.25	−17.25
118	111a	111b	86	117	104	92	56	14.25	−14.25
118	111a	111b	56	117	104	92	86	−0.75	0.75
118	111a	104	92	117	111b	86	56	13.75	−13.75
118	111a	104	86	117	111b	92	56	10.75	−10.75
118	111a	104	56	117	111b	92	86	−4.25	4.25
118	111a	92	86	117	111b	104	56	4.75	−4.75
118	111a	92	56	117	111b	104	86	−10.25	10.25
118	111a	86	56	117	111b	104	92	−13.25	13.75
118	111b	104	92	117	111a	86	56	13.75	−13.75
118	111b	104	86	117	111a	92	56	10.75	−10.75
118	111b	104	56	117	111a	92	86	−4.25	4.25
118	111b	92	86	117	111a	104	56	4.75	−4.75
118	111b	92	56	117	111a	104	86	−10.25	10.25
118	111b	86	56	117	111a	104	92	−13.25	13.25
118	104	92	86	117	111a	111b	56	1.25	−1.25
118	104	92	56	117	111a	111b	86	−13.75	13.75
118	104	86	56	117	111a	111b	92	−16.75	16.75
118	92	86	56	117	111a	111b	104	−22.75	22.75

These are 35 of the 70 possible allocations. The remaining 35 can be obtained by swapping the beef and cereal group of these 35 allocations, and the corresponding test-statistic $D = \overline{y}_{beef} - \overline{y}_{cereal}$ are of the opposite sign.

beef diet				cereal diet				$\overline{y}_{\text{beef}}-\overline{y}_{\text{cereal}}$	if 2 groups swapped $\overline{y}_{\text{beef}}-\overline{y}_{\text{cereal}}$
118	117	111a	111b	104	92	86	56	29.75	−29.75
118	117	111a	104	111b	92	86	56	26.25	−26.25
118	117	111a	92	111b	104	86	56	20.25	−20.25
118	117	111a	86	111b	104	92	56	17.25	−17.25
118	117	111a	56	111b	104	92	86	2.25	−2.25
118	117	111b	104	111a	92	86	56	26.25	−26.25
118	117	111b	92	111a	104	86	56	20.25	−20.25
118	117	111b	86	111a	104	92	56	17.25	−17.25
118	117	111b	56	111a	104	92	86	2.25	− 2.25
118	117	104	92	111a	111b	86	56	16.75	−16.75
118	117	104	86	111a	111b	92	56	13.75	−13.75
118	117	104	56	111a	111b	92	86	−1.25	1.25
118	117	92	86	111a	111b	104	56	7.75	−7.75
118	117	92	56	111a	111b	104	86	−7.25	7.25
118	117	86	56	111a	111b	104	92	−10.25	10.25
118	111a	111b	104	117	92	86	56	23.25	−23.25
118	111a	111b	92	117	104	86	56	17.25	−17.25
118	111a	111b	86	117	104	92	56	14.25	−14.25
118	111a	111b	56	117	104	92	86	−0.75	0.75
118	111a	104	92	117	111b	86	56	13.75	−13.75
118	111a	104	86	117	111b	92	56	10.75	−10.75
118	111a	104	56	117	111b	92	86	−4.25	4.25
118	111a	92	86	117	111b	104	56	4.75	−4.75
118	111a	92	56	117	111b	104	86	−10.25	10.25
118	111a	86	56	117	111b	104	92	−13.25	13.75
118	111b	104	92	117	111a	86	56	13.75	−13.75
118	111b	104	86	117	111a	92	56	10.75	−10.75
118	111b	104	56	117	111a	92	86	−4.25	4.25
118	111b	92	86	117	111a	104	56	4.75	−4.75
118	111b	92	56	117	111a	104	86	−10.25	10.25
118	111b	86	56	117	111a	104	92	−13.25	13.25
118	104	92	86	117	111a	111b	56	1.25	−1.25
118	104	92	56	117	111a	111b	86	−13.75	13.75
118	104	86	56	117	111a	111b	92	−16.75	16.75
118	92	86	56	117	111a	111b	104	−22.75	22.75

As the 70 possible allocations were equally likely to occur, we can obtain the sampling distribution of

$$D = \overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$$

as follows.

Value of D	Prob.	Value of D	Prob.
−29.75	1/70	29.75	1/70
−26.25	2/70	26.25	2/70
−23.25	1/70	23.25	1/70
−22.75	1/70	22.75	1/70
−20.25	2/70	20.25	2/70
−17.25	3/70	17.25	3/70
−16.75	2/70	16.75	2/70
−14.25	1/70	14.25	1/70
−13.75	4/70	13.75	4/70
−13.25	2/70	13.25	2/70
−10.75	2/70	10.75	2/70
−10.25	3/70	10.25	3/70
−7.75	1/70	7.75	1/70
−7.25	1/70	7.25	1/70
−4.75	2/70	4.75	2/70
−4.25	2/70	4.25	2/70
−2.25	2/70	2.25	2/70
−1.25	2/70	1.25	2/70
−0.75	1/70	0.75	1/70

beef diet				cereal diet				if 2 groups swapped	
								$\overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$	$\overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$
118	117	111a	111b	104	92	86	56	29.75	−29.75
118	117	111a	104	111b	92	86	56	26.25	−26.25
118	117	111a	92	111b	104	86	56	20.25	−20.25
118	117	111a	86	111b	104	92	56	17.25	−17.25
118	117	111a	56	111b	104	92	86	2.25	−2.25
118	117	111b	104	111a	92	86	56	26.25	−26.25
118	117	111b	92	111a	104	86	56	20.25	−20.25
118	117	111b	86	111a	104	92	56	17.25	−17.25
118	117	111b	56	111a	104	92	86	2.25	− 2.25
118	117	104	92	111a	111b	86	56	16.75	−16.75
118	117	104	86	111a	111b	92	56	13.75	−13.75
118	117	104	56	111a	111b	92	86	−1.25	1.25
118	117	92	86	111a	111b	104	56	7.75	−7.75
118	117	92	56	111a	111b	104	86	−7.25	7.25
118	117	86	56	111a	111b	104	92	−10.25	10.25
118	111a	111b	104	117	92	86	56	23.25	−23.25
118	111a	111b	92	117	104	86	56	17.25	−17.25
118	111a	111b	86	117	104	92	56	14.25	−14.25
118	111a	111b	56	117	104	92	86	−0.75	0.75
118	111a	104	92	117	111b	86	56	13.75	−13.75
118	111a	104	86	117	111b	92	56	10.75	−10.75
118	111a	104	56	117	111b	92	86	−4.25	4.25
118	111a	92	86	117	111b	104	56	4.75	−4.75
118	111a	92	56	117	111b	104	86	−10.25	10.25
118	111a	86	56	117	111b	104	92	−13.25	13.75
118	111b	104	92	117	111a	86	56	13.75	−13.75
118	111b	104	86	117	111a	92	56	10.75	−10.75
118	111b	104	56	117	111a	92	86	−4.25	4.25
118	111b	92	86	117	111a	104	56	4.75	−4.75
118	111b	92	56	117	111a	104	86	−10.25	10.25
118	111b	86	56	117	111a	104	92	−13.25	13.25
118	104	92	86	117	111a	111b	56	1.25	−1.25
118	104	92	56	117	111a	111b	86	−13.75	13.75
118	104	86	56	117	111a	111b	92	−16.75	16.75
118	92	86	56	117	111a	111b	104	−22.75	22.75

As the 70 possible allocations were equally likely to occur, we can obtain the sampling distribution of

$$D = \overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$$

as follows.

Value of D	Prob.	Value of D	Prob.
−29.75	1/70	29.75	1/70
−26.25	2/70	26.25	2/70
−23.25	1/70	23.25	1/70
−22.75	1/70	22.75	1/70
−20.25	2/70	20.25	2/70
−17.25	3/70	17.25	3/70
−16.75	2/70	16.75	2/70
−14.25	1/70	14.25	1/70
−13.75	4/70	13.75	4/70
−13.25	2/70	13.25	2/70
−10.75	2/70	10.75	2/70
−10.25	3/70	10.25	3/70
−7.75	1/70	7.75	1/70
−7.25	1/70	7.25	1/70
−4.75	2/70	4.75	2/70
−4.25	2/70	4.25	2/70
−2.25	2/70	2.25	2/70
−1.25	2/70	1.25	2/70
−0.75	1/70	0.75	1/70

beef diet				cereal diet					if 2 groups swapped
								$\bar{y}_{\text{beef}} - \bar{y}_{\text{cereal}}$	$\bar{y}_{\text{beef}} - \bar{y}_{\text{cereal}}$
118	117	111a	111b	104	92	86	56	29.75	−29.75
118	117	111a	104	111b	92	86	56	26.25	−26.25
118	117	111a	92	111b	104	86	56	20.25	−20.25
118	117	111a	86	111b	104	92	56	17.25	−17.25
118	117	111a	56	111b	104	92	86	2.25	−2.25
118	117	111b	104	111a	92	86	56	26.25	−26.25
118	117	111b	92	111a	104	86	56	20.25	−20.25
118	117	111b	86	111a	104	92	56	17.25	−17.25
118	117	111b	56	111a	104	92	86	2.25	− 2.25
118	117	104	92	111a	111b	86	56	16.75	−16.75
118	117	104	86	111a	111b	92	56	13.75	−13.75
118	117	104	56	111a	111b	92	86	−1.25	1.25
118	117	92	86	111a	111b	104	56	7.75	−7.75
118	117	92	56	111a	111b	104	86	−7.25	7.25
118	117	86	56	111a	111b	104	92	−10.25	10.25
118	111a	111b	104	117	92	86	56	23.25	−23.25
118	111a	111b	92	117	104	86	56	17.25	−17.25
118	111a	111b	86	117	104	92	56	14.25	−14.25
118	111a	111b	56	117	104	92	86	−0.75	0.75
118	111a	104	92	117	111b	86	56	13.75	−13.75
118	111a	104	86	117	111b	92	56	10.75	−10.75
118	111a	104	56	117	111b	92	86	−4.25	4.25
118	111a	92	86	117	111b	104	56	4.75	−4.75
118	111a	92	56	117	111b	104	86	−10.25	10.25
118	111a	86	56	117	111b	104	92	−13.25	13.75
118	111b	104	92	117	111a	86	56	13.75	−13.75
118	111b	104	86	117	111a	92	56	10.75	−10.75
118	111b	104	56	117	111a	92	86	−4.25	4.25
118	111b	92	86	117	111a	104	56	4.75	−4.75
118	111b	92	56	117	111a	104	86	−10.25	10.25
118	111b	86	56	117	111a	104	92	−13.25	13.25
118	104	92	86	117	111a	111b	56	1.25	−1.25
118	104	92	56	117	111a	111b	86	−13.75	13.75
118	104	86	56	117	111a	111b	92	−16.75	16.75
118	92	86	56	117	111a	111b	104	−22.75	22.75

As the 70 possible allocations were equally likely to occur, we can obtain the sampling distribution of

$$D = \bar{y}_{\text{beef}} - \bar{y}_{\text{cereal}}$$

as follows.

Value of D	Prob.	Value of D	Prob.
−29.75	1/70	29.75	1/70
−26.25	2/70	26.25	2/70
−23.25	1/70	23.25	1/70
−22.75	1/70	22.75	1/70
−20.25	2/70	20.25	2/70
−17.25	3/70	17.25	3/70
−16.75	2/70	16.75	2/70
−14.25	1/70	14.25	1/70
−13.75	4/70	13.75	4/70
−13.25	2/70	13.25	2/70
−10.75	2/70	10.75	2/70
−10.25	3/70	10.25	3/70
−7.75	1/70	7.75	1/70
−7.25	1/70	7.25	1/70
−4.75	2/70	4.75	2/70
−4.25	2/70	4.25	2/70
−2.25	2/70	2.25	2/70
−1.25	2/70	1.25	2/70
−0.75	1/70	0.75	1/70

beef diet				cereal diet					if 2 groups swapped
								$\overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$	$\overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$
118	117	111a	111b	104	92	86	56	29.75	−29.75
118	117	111a	104	111b	92	86	56	26.25	−26.25
118	117	111a	92	111b	104	86	56	20.25	−20.25
118	117	111a	86	111b	104	92	56	17.25	−17.25
118	117	111a	56	111b	104	92	86	2.25	−2.25
118	117	111b	104	111a	92	86	56	26.25	−26.25
118	117	111b	92	111a	104	86	56	20.25	−20.25
118	117	111b	86	111a	104	92	56	17.25	−17.25
118	117	111b	56	111a	104	92	86	2.25	− 2.25
118	117	104	92	111a	111b	86	56	16.75	−16.75
118	117	104	86	111a	111b	92	56	13.75	−13.75
118	117	104	56	111a	111b	92	86	−1.25	1.25
118	117	92	86	111a	111b	104	56	7.75	−7.75
118	117	92	56	111a	111b	104	86	−7.25	7.25
118	117	86	56	111a	111b	104	92	−10.25	10.25
118	111a	111b	104	117	92	86	56	23.25	−23.25
118	111a	111b	92	117	104	86	56	17.25	−17.25
118	111a	111b	86	117	104	92	56	14.25	−14.25
118	111a	111b	56	117	104	92	86	−0.75	0.75
118	111a	104	92	117	111b	86	56	13.75	−13.75
118	111a	104	86	117	111b	92	56	10.75	−10.75
118	111a	104	56	117	111b	92	86	−4.25	4.25
118	111a	92	86	117	111b	104	56	4.75	−4.75
118	111a	92	56	117	111b	104	86	−10.25	10.25
118	111a	86	56	117	111b	104	92	−13.25	13.75
118	111b	104	92	117	111a	86	56	13.75	−13.75
118	111b	104	86	117	111a	92	56	10.75	−10.75
118	111b	104	56	117	111a	92	86	−4.25	4.25
118	111b	92	86	117	111a	104	56	4.75	−4.75
118	111b	92	56	117	111a	104	86	−10.25	10.25
118	111b	86	56	117	111a	104	92	−13.25	13.25
118	104	92	86	117	111a	111b	56	1.25	−1.25
118	104	92	56	117	111a	111b	86	−13.75	13.75
118	104	86	56	117	111a	111b	92	−16.75	16.75
118	92	86	56	117	111a	111b	104	−22.75	22.75

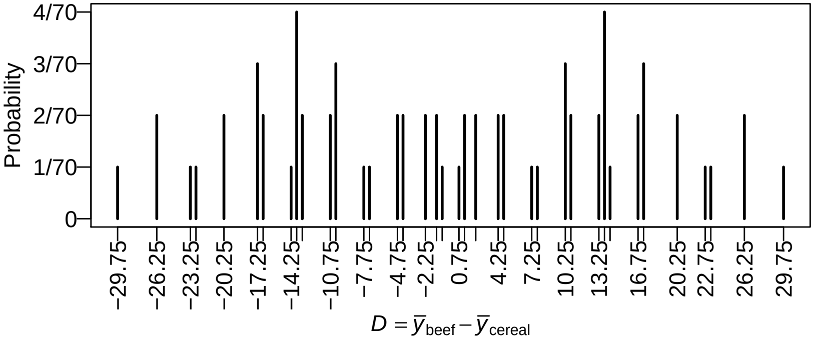
As the 70 possible allocations were equally likely to occur, we can obtain the sampling distribution of

$$D = \overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$$

as follows.

Value of D	Prob.	Value of D	Prob.
−29.75	1/70	29.75	1/70
−26.25	2/70	26.25	2/70
−23.25	1/70	23.25	1/70
−22.75	1/70	22.75	1/70
−20.25	2/70	20.25	2/70
−17.25	3/70	17.25	3/70
−16.75	2/70	16.75	2/70
−14.25	1/70	14.25	1/70
−13.75	4/70	13.75	4/70
−13.25	2/70	13.25	2/70
−10.75	2/70	10.75	2/70
−10.25	3/70	10.25	3/70
−7.75	1/70	7.75	1/70
−7.25	1/70	7.25	1/70
−4.75	2/70	4.75	2/70
−4.25	2/70	4.25	2/70
−2.25	2/70	2.25	2/70
−1.25	2/70	1.25	2/70
−0.75	1/70	0.75	1/70

Plot of the Sampling Distribution of $D = \bar{y}_{\text{beef}} - \bar{y}_{\text{cereal}}$



Observe the sampling distribution of D doesn't look normal.

To test against a one-sided alternative

H_a: Rats given beef diet had a higher mean weight gain

Larger values of $D = \overline{y}_{beef} - \overline{y}_{cereal}$ are evidence for H_a.

To test against a one-sided alternative

> H_a: Rats given beef diet had a higher mean weight gain

Larger values of $D = \overline{y}_{beef} - \overline{y}_{cereal}$ are evidence for H_a.

For the observed allocation, the D value is

$$\begin{aligned} D &= \frac{104+118+117+111}{4} - \frac{111+56+86+92}{4} \\ &= 112.5 - 86.25 = 26.25 \end{aligned}$$

To test against a one-sided alternative

H$_a$: Rats given beef diet had a higher mean weight gain

Larger values of $D = \overline{y}_{beef} - \overline{y}_{cereal}$ are evidence for H$_a$.

For the observed allocation, the D value is

$$D = \frac{104+118+117+111}{4} - \frac{111+56+86+92}{4}$$
$$= 112.5 - 86.25 = 26.25$$

Only 3 of the 70 possible allocations give a D as high or higher the observed $D = 26.25$. Hence the one-sided P-value is $3/70$.

beef diet				cereal diet				$D = \overline{y}_{beef} - \overline{y}_{cereal}$	
118	117	111a	111b	104	92	86	56	29.75	
118	117	111a	104	111b	92	86	56	26.25	
118	117	104	111b	111a	92	86	56	26.25	← observed

To test against a one-sided alternative

H_a: Rats given beef diet had a higher mean weight gain

Larger values of $D = \overline{y}_{beef} - \overline{y}_{cereal}$ are evidence for H_a.

For the observed allocation, the D value is

$$\begin{aligned} D &= \frac{104+118+117+111}{4} - \frac{111+56+86+92}{4} \\ &= 112.5 - 86.25 = 26.25 \end{aligned}$$

Only 3 of the 70 possible allocations give a D as high or higher the observed $D = 26.25$. Hence the one-sided P-value is $3/70$.

beef diet				cereal diet				$D = \overline{y}_{beef} - \overline{y}_{cereal}$	
118	117	111a	111b	104	92	86	56	29.75	
118	117	111a	104	111b	92	86	56	26.25	
118	117	104	111b	111a	92	86	56	26.25	← observed

Practically, there is **no need to check all 70 allocations**. One just needs to count the number of allocations that gives a D-value $\geq$ the observed D.

Two-Sided Permutation Test

For a two-sided test

H_0 : beef or cereal diet makes no difference on rats' weight gain

H_a : the two diets make some difference on rats' weight gain

a reasonable test statistic is $|D| = |\overline{y}_{beef} - \overline{y}_{cereal}|$. Larger values of $|D|$ are evidence for H$_a$.

Two-Sided Permutation Test

For a two-sided test

H_0 : beef or cereal diet makes no difference on rats' weight gain

H_a : the two diets make some difference on rats' weight gain

a reasonable test statistic is $|D| = |\overline{y}_{beef} - \overline{y}_{cereal}|$. Larger values of $|D|$ are evidence for H$_a$.

By swapping rats in the two groups, we get allocations in the other extreme.

beef diet				cereal diet				$D = \overline{y}_{\text{beef}} - \overline{y}_{\text{cereal}}$	
118	117	111a	111b	104	92	86	56	29.75	
118	117	111a	104	111b	92	86	56	26.25	
118	117	104	111b	111a	92	86	56	26.25	← observed
111a	92	86	56	118	117	104	111b	−26.25	
111b	92	86	56	118	117	111a	104	−26.25	
104	92	86	56	118	117	111a	111b	−29.75	

The **two-sided** P-value is thus $6/70 = 6/70 \approx 8.6\%$.

Test Procedures of a Permutation Test (1)

Data:	Sample 1 (Treatment 1)	$y_{11}, y_{12}, \ldots, y_{1n_1}$
	Sample 2 (Treatment 2)	$y_{21}, y_{22}, \ldots\ldots\ldots, y_{2n_2}$

For randomized experiments

H$_0$: the treatments make no difference

For observational studies

H$_0$: The two populations have an identical distributions

Under H$_0$, any n_1 of the total of $n_1 + n_2$ observations is as likely to be our observations in Sample 1/ Treatment group 1.

Test Procedures of a Permutation Test (2)

1. Find the observed difference in means: $D_{observed} = \bar{y}_1 - \bar{y}_2$.
2. For **one-sided** tests, list all the possible allocations of units to a group of size n_1 and another group of size n_2 that the difference in means

   $$D_{new} = \bar{y}_{1,new} - \bar{y}_{2,new}$$

   is $\geq$ the observed difference in means $D_{observed}$. The one-sided P-value is the count of such allocations over $\binom{n_1+n_2}{n_1}$
3. If **two-sided**, list all the possible allocations of units to a group of size n_1 and another group of size n_2 that the absolute difference in means

   $$|D_{new}| = |\bar{y}_{1,new} - \bar{y}_{2,new}|$$

   is $\geq$ the observed absolute difference in means $|D_{observed}|$. The two-sided P-value is the number of such allocations over $\binom{n_1+n_2}{n_1}$.

Two-Sided P-value is Not Always 2x One-Sided P-value

- When the size of the two groups are **equal** $n_1 = n_2$, the two-sided P-value is twice the one-sided P-value, since one can obtain allocations in the other extreme by swapping the cases in the two groups

Two-Sided *P*-value is Not Always 2x One-Sided *P*-value

- When the size of the two groups are **equal** $n_1 = n_2$, the two-sided *P*-value is twice the one-sided *P*-value, since one can obtain allocations in the other extreme by swapping the cases in the two groups
- When two groups are of **different sizes** $n_1 \neq n_2$, the two-sided *P*-value may NOT be 2x the one-sided *P*-value, e.g., when $n_1 = 2$, $n_2 = 3$,

Group 1		Group 2			$D = \overline{y}_1 - \overline{y}_2$	
1	2	3	4	6	$(1+2)/2-(3+4+6)/3 \approx -2.83$	
1	3	2	4	6	$(1+3)/2-(2+4+6)/3 \approx -2$	
1	4	2	3	6	$(1+4)/2-(2+3+6)/3 \approx -1.17$	
2	3	1	4	6	$(2+3)/2-(1+4+6)/3 \approx -0.33$	
2	4	1	3	6	$(2+4)/2-(1+3+6)/3 \approx -0.33$	
1	6	2	3	4	$(1+6)/2-(2+3+4)/3 \approx 0.5$	
3	4	1	2	6	$(3+4)/2-(1+2+6)/3 \approx 0.5$	
2	6	1	3	4	$(2+6)/2-(1+3+4)/3 \approx 1.33$	$\rightarrow$ observed
3	6	1	2	4	$(3+6)/2-(1+2+4)/3 \approx 2.17$	
4	6	1	2	3	$(4+6)/2-(1+2+3)/3 \approx 3$	

Two-Sided *P*-value is Not Always 2x One-Sided *P*-value

- When the size of the two groups are **equal** $n_1 = n_2$, the two-sided P-value is twice the one-sided P-value, since one can obtain allocations in the other extreme by swapping the cases in the two groups
- When two groups are of **different sizes** $n_1 \neq n_2$, the two-sided P-value may NOT be 2x the one-sided P-value, e.g., when $n_1 = 2$, $n_2 = 3$,

Group 1		Group 2			$D = \overline{y}_1 - \overline{y}_2$	
1	2	3	4	6	$(1+2)/2-(3+4+6)/3 \approx -2.83$	
1	3	2	4	6	$(1+3)/2-(2+4+6)/3 \approx -2$	
1	4	2	3	6	$(1+4)/2-(2+3+6)/3 \approx -1.17$	
2	3	1	4	6	$(2+3)/2-(1+4+6)/3 \approx -0.33$	
2	4	1	3	6	$(2+4)/2-(1+3+6)/3 \approx -0.33$	
1	6	2	3	4	$(1+6)/2-(2+3+4)/3 \approx 0.5$	
3	4	1	2	6	$(3+4)/2-(1+2+6)/3 \approx 0.5$	
2	6	1	3	4	$(2+6)/2-(1+3+4)/3 \approx 1.33$	→ observed
3	6	1	2	4	$(3+6)/2-(1+2+4)/3 \approx 2.17$	
4	6	1	2	3	$(4+6)/2-(1+2+3)/3 \approx 3$	

One-sided P-value $= 3/10$;

Two-Sided P-value is Not Always 2x One-Sided P-value

- When the size of the two groups are **equal** $n_1 = n_2$, the two-sided P-value is twice the one-sided P-value, since one can obtain allocations in the other extreme by swapping the cases in the two groups
- When two groups are of **different sizes** $n_1 \neq n_2$, the two-sided P-value may NOT be 2x the one-sided P-value, e.g., when $n_1 = 2$, $n_2 = 3$,

Group 1		Group 2			$D = \overline{y}_1 - \overline{y}_2$	
1	2	3	4	6	$(1+2)/2 - (3+4+6)/3 \approx -2.83$	
1	3	2	4	6	$(1+3)/2 - (2+4+6)/3 \approx -2$	
1	4	2	3	6	$(1+4)/2 - (2+3+6)/3 \approx -1.17$	
2	3	1	4	6	$(2+3)/2 - (1+4+6)/3 \approx -0.33$	
2	4	1	3	6	$(2+4)/2 - (1+3+6)/3 \approx -0.33$	
1	6	2	3	4	$(1+6)/2 - (2+3+4)/3 \approx 0.5$	
3	4	1	2	6	$(3+4)/2 - (1+2+6)/3 \approx 0.5$	
2	6	1	3	4	$(2+6)/2 - (1+3+4)/3 \approx 1.33$	→ observed
3	6	1	2	4	$(3+6)/2 - (1+2+4)/3 \approx 2.17$	
4	6	1	2	3	$(4+6)/2 - (1+2+3)/3 \approx 3$	

One-sided P-value $= 3/10$; Two-sided P-value $= 5/10$

Approximate P-value for Permutation Test

When the sample sizes n_1 and n_2 are large, it is labor-intensive to to find the exact P-value by counting the extreme cases. Nonetheless, one can estimate the exact P-value by sampling from the possible permutations. We will demonstrate using the Rats' diet experiment.

1. Sample $n_1 = 4$ observations without replacement from the set of all observations $\{111, 56, 86, 92, 104, 118, 117, 111\}$ as the beef group, and the rest as the cereal group. Find the mean differences of the two groups $D_{new} = \bar{y}_{1,new} - \bar{y}_{2,new}$.
2. Repeat the first step a huge number M of times and get a mean difference D_{new} for every repetition.
3. Count the number k of repetitions that produce mean difference $D_{new} \geq$ the mean difference of the original grouping.
4. When M is large enough, k/M is an approximate P-value.

Permutation Test in R

There is no build-in R function to do the permutation test (at least to my knowledge), but it's not hard to write our own code to find an approximate *P*-value.

The `sample()` function in R can randomly permute the observations.

```
> wtgain = c(104, 118, 117, 111, 111, 56, 86, 92)
> newwtgain = sample(wtgain); newwtgain
[1] 111 118  86  56 104 117 111  92
```

After permutation, regard the first $n_1 = 4$ observations, `newwtgain[1:4]`, as the beef group, and the rest `newwtgain[5:8]` as the cereal group, and then compute the mean difference of the two group.

```
> D = mean(newwtgain[1:4])-mean(newwtgain[5:8]); D
[1] -13.25
```

Permutation Test in R

Let's repeat the previous step $M = 10000$ times.

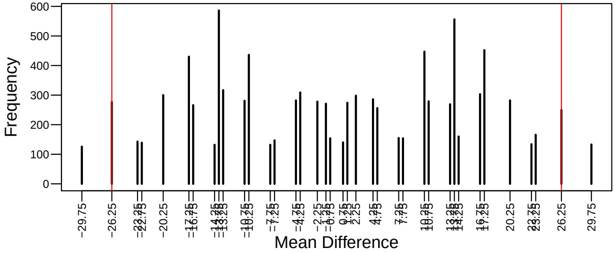
```
M = 10000
D = vector("numeric",length=M)
for(i in 1:M){
  newwtgain = sample(wtgain)
  D[i] = mean(newwtgain[1:4])-mean(newwtgain[5:8])
}
```

Let's take a look at the frequencies of the values of the mean difference D we obtained. (The result may vary from simulation to simulation).

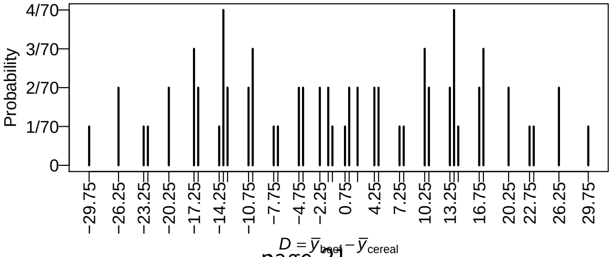
```
> table(D)
D
-29.75 -26.25 -23.25 -22.75 -20.25 -17.25 -16.75 -14.25 -13.75 -13.25
   147    275    125    143    277    399    296    138    540    297
-10.75 -10.25  -7.75  -7.25  -4.75  -4.25  -2.25  -1.25  -0.75   0.75
   265    455    144    131    302    309    302    273    138    134
  1.25   2.25   4.25   4.75   7.25   7.75  10.25  10.75  13.25  13.75
   299    264    267    298    133    144    413    308    317    631
 14.25  16.75  17.25  20.25  22.75  23.25  26.25  29.75
   144    282    451    257    120    150    283    149
> D.obs = mean(wtgain[1:4])-mean(wtgain[5:8]); D.obs
[1] 26.25             # observed mean difference
> sum(abs(D) >= D.obs)
[1] 854
```

Among the 10000 mean differences, we see $147 + 275 + 283 + 149 = 854$ of them have absolute values $\geq$ the observed mean difference 26.25. So the 2-sided P-value is estimated to be $854/10000 = 0.0854$, not far from the exact P-value, $6/70 \approx 0.0857$.

```
> plot(table(D), ylab="Frequency", xlab="Mean Difference")
```



Observe the (simulated) sampling distribution for D look pretty close to the exact sampling distribution of D below. This is why we can approx. the exact P-value by simulation.



Remarks on Permutation Tests

- Permutation tests applied on randomized experiments are called the randomization tests. This is the name used Chapter 2 in Oehlert's textbook.
- The test statistic of permutation tests can also be difference in medians, 25th percentiles, etc, between the 2 groups, not necessarily the means.
- The sampling distribution of the test statistic of a permutation test is obtained by considering all possible random allocations of units to groups, making no assumption on the form of the population distribution.

When to Use Permutation Test?

- When sample sizes are very small, and hence it's hard to check the normality assumption, permutation test is a nice alternative to the conventional t-test or Welch t-test.

When to Use Permutation Test?

- When sample sizes are very small, and hence it's hard to check the normality assumption, permutation test is a nice alternative to the conventional t-test or Welch t-test.
- When sample sizes are large, P-values of permutation tests are usually close to P-values of t-tests

When to Use Permutation Test?

- When sample sizes are very small, and hence it's hard to check the normality assumption, permutation test is a nice alternative to the conventional t-test or Welch t-test.
- When sample sizes are large, P-values of permutation tests are usually close to P-values of t-tests
- Permutation test is still subject to the effect of outliers.

When to Use Permutation Test?

- When sample sizes are very small, and hence it's hard to check the normality assumption, permutation test is a nice alternative to the conventional *t*-test or Welch *t*-test.
- When sample sizes are large, *P*-values of permutation tests are usually close to *P*-values of *t*-tests
- Permutation test is still subject to the effect of outliers.
- When applied on observational studies, the H_0 of the permutation test assumes the two population have identical distributions, which implies they have identical SDs, even though it makes no assumption on the form of the population distribution

When to Use Permutation Test?

- When sample sizes are very small, and hence it's hard to check the normality assumption, permutation test is a nice alternative to the conventional t-test or Welch t-test.
- When sample sizes are large, P-values of permutation tests are usually close to P-values of t-tests
- Permutation test is still subject to the effect of outliers.
- When applied on observational studies, the H_0 of the permutation test assumes the two population have identical distributions, which implies they have identical SDs, even though it makes no assumption on the form of the population distribution
- When one group appear to have greater variabilities than the other group, comparison of two groups is not simply the comparison of the two means. One may transform the data to mitigate the unequal variance problem before applying the permutation test

Rank-Sum Test

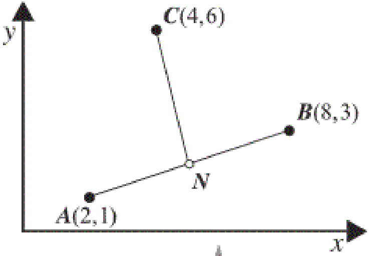
Example: Cognitive Load Theory in Teaching

— A Randomized Experiment

Consider the following problem in coordinate geometry.

> *Point A has coordinates* $(2,1)$*, point B has* $(8,3)$*, and point C has* $(4,6)$*. What is the slope of the line that connects C to the midpoint between A and B?*

Presenting the solution as a worked problem, a conventional textbook shows a picture of the layout, gives a discussion in the text, and then provides the lines of algebraic manipulation leading to the right answer. (See the next slide). Recent theoretical developments in cognitive science suggest that splitting the presentation into the 3 distinct units of diagram, text, and algebra imposes a heavy, extraneous cognitive load on the student. The requirement that the student organize and process the separate elements constitutes a cognitive load.



Solution: *The coordinates of* N *are:*

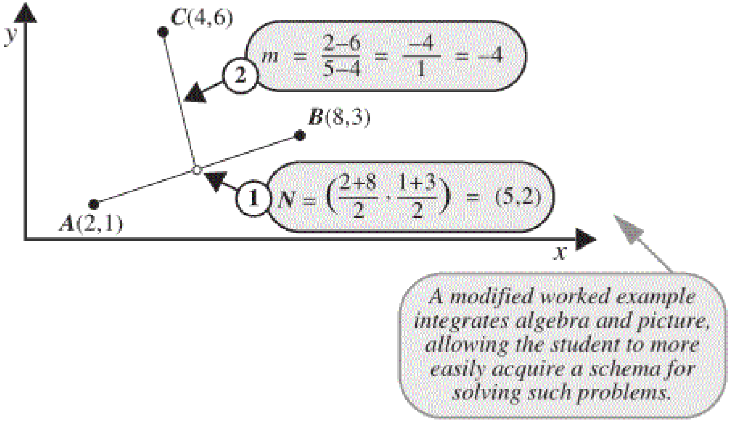
$$N = \left(\frac{2+8}{2}, \frac{1+3}{2}\right)$$
$$= (5,2)$$

The slope of CN *is:*

$$m = \frac{2-6}{5-4}$$
$$= \frac{-4}{1} = -4$$

In a conventional worked example, algebra and diagram are separated, giving students an extraneous cognitive load of having to assimilate the two.

In a test of this theory, researchers compared the effectiveness of conventional textbook worked examples to modified worked examples, which present the algebraic manipulations and explanation as part of the graphical display.


C(4,6)
y
2
$m = \frac{2-6}{5-4} = \frac{-4}{1} = -4$
B(8,3)
1
$N = \left(\frac{2+8}{2}, \frac{1+3}{2}\right) = (5,2)$
A(2,1)
x

A modified worked example integrates algebra and picture, allowing the student to more easily acquire a schema for solving such problems.

Example: Cognitive Load Theory in Teaching — Study Design

- Subjects: 28 ninth-year students in Sydney, Australia, with no previous exposure to coordinate geometry but have adequate math to deal with the problems given
- The 28 subjects were randomized to self-study one of two instructional materials. The two materials covered exactly the same problems, presented differently. Students were given as much time as they wished to study the material, but not allowed to ask questions.
- Following the instructional phase, all students were tested with a common examination over 3 problems of different difficulty.
- Response: the time (in seconds) required to arrive at a solution to the moderately difficult problem.

Example: Cognitive Load Theory in Teaching — Data

Modified Group:

68, 70, 73, 75, 77, 80, 80, 132, 148, 155, 183, 197, 206, 210

Conventional Group:

130, 139, 146, 150, 161, 177, 228, 242, 265, 300^*, 300^*,300^*,300^*,300^*

Note the response is censored at 300 seconds because the time allotment for the problem is 5 minutes Five students did not complete the problem in the 5-minute (300 seconds) time allotment.

The Rank Transformation

- Whenever the data contains outliers, it's a headache considering whether to remove the outlier(s).

The Rank Transformation

- Whenever the data contains outliers, it's a headache considering whether to remove the outlier(s).
- If transformations cannot make outliers less extreme, an effective and widely used approach is to work with the **ranks** of the data rather than the original values.

The Rank Transformation

- Whenever the data contains outliers, it's a headache considering whether to remove the outlier(s).
- If transformations cannot make outliers less extreme, an effective and widely used approach is to work with the **ranks** of the data rather than the original values.
- By ranking the data, **the impact of outliers is mitigated**: regardless of how extreme an outlier is, it receives the same rank as if it were just slightly larger than the second-largest observation, e.g.,

value	0.04	0.86	1.3	2.2	3.8	8.0	10.7	11.6	61.8
	↓	↓	↓	↓	↓	↓	↓	↓	↓
rank	1	2	3	4	5	6	7	8	9

The Rank Transformation

- Whenever the data contains outliers, it's a headache considering whether to remove the outlier(s).
- If transformations cannot make outliers less extreme, an effective and widely used approach is to work with the **ranks** of the data rather than the original values.
- By ranking the data, **the impact of outliers is mitigated**: regardless of how extreme an outlier is, it receives the same rank as if it were just slightly larger than the second-largest observation, e.g.,

value	0.04	0.86	1.3	2.2	3.8	8.0	10.7	11.6	61.8
	↓	↓	↓	↓	↓	↓	↓	↓	↓
rank	1	2	3	4	5	6	7	8	9

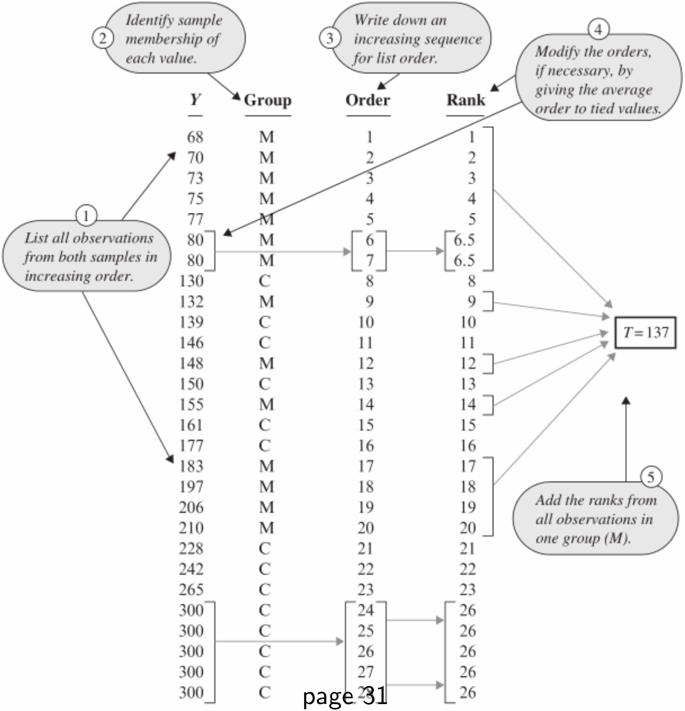
- **Skewness also is mitigated**, sine all ranks are equally far apart from each other.

The Rank Transformation

- Whenever the data contains outliers, it's a headache considering whether to remove the outlier(s).
- If transformations cannot make outliers less extreme, an effective and widely used approach is to work with the **ranks** of the data rather than the original values.
- By ranking the data, **the impact of outliers is mitigated**: regardless of how extreme an outlier is, it receives the same rank as if it were just slightly larger than the second-largest observation, e.g.,

value	0.04	0.86	1.3	2.2	3.8	8.0	10.7	11.6	61.8
	↓	↓	↓	↓	↓	↓	↓	↓	↓
rank	1	2	3	4	5	6	7	8	9

- **Skewness also is mitigated**, sine all ranks are equally far apart from each other.
- Another attractive feature of the rank transformation is its ability to deal with censored observations as in the cognitive load experiment.



1 *List all observations from both samples in increasing order.*

2 *Identify sample membership of each value.*

3 *Write down an increasing sequence for list order.*

4 *Modify the orders, if necessary, by giving the average order to tied values.*

5 *Add the ranks from all observations in one group (M).*

Y	Group	Order	Rank
68	M	1	1
70	M	2	2
73	M	3	3
75	M	4	4
77	M	5	5
80	M	6	6.5
80	M	7	6.5
130	C	8	8
132	M	9	9
139	C	10	10
146	C	11	11
148	M	12	12
150	C	13	13
155	M	14	14
161	C	15	15
177	C	16	16
183	M	17	17
197	M	18	18
206	M	19	19
210	M	20	20
228	C	21	21
242	C	22	22
265	C	23	23
300	C	24	26
300	C	25	26
300	C	26	26
300	C	27	26
300	C	28	26

$T = 137$

The Rank Sum Statistic

First transform the data to their ranks.

1. List all observations from both samples in increasing order.
2. Identify which sample each observation came from.
3. Create a new column labeled "order", as a straight sequence of numbers from 1 to $n_1 + n_2$.
4. Search for ties — that is, duplicated values — in the combined data set. The ranks for tied observations are taken to be the average of the orders for those cases.

The rank-sum statistic, T, is the sum of all the ranks in one group, called "group 1." Group 1 is conventionally the group with the smaller sample size (because that requires less computation). The choice, however, is arbitrary.

Rank Sum Test = Permutation Test Performed on Ranks

Recall the permutation test uses the $\boxed{\text{diff. in means}}$ as the test statistic. Performed on ranks, the test statistic would be

$$\begin{pmatrix}\text{Mean ranks}\\ \text{for group 1}\end{pmatrix} - \begin{pmatrix}\text{Mean ranks}\\ \text{for group 2}\end{pmatrix} = \frac{T_1}{n_1} - \frac{T_2}{n_2}$$

where n_i = sizes of Group i, and T_i = rank-sum for Group i, $i = 1, 2$

Rank Sum Test = Permutation Test Performed on Ranks

Recall the permutation test uses the $\boxed{\text{diff. in means}}$ as the test statistic. Performed on ranks, the test statistic would be

$$\begin{pmatrix}\text{Mean ranks}\\ \text{for group 1}\end{pmatrix} - \begin{pmatrix}\text{Mean ranks}\\ \text{for group 2}\end{pmatrix} = \frac{T_1}{n_1} - \frac{T_2}{n_2}$$

where n_i = sizes of Group i, and T_i = rank-sum for Group i, $i = 1, 2$

- Observe that $T_1 + T_2$ equals the rank-sum of all observations

  $$1+2+3+\cdots+(n_1+n_2) = \frac{(n_1+n_2)(n_1+n_2+1)}{2} = \text{rank sum of all}$$

  which doesn't change from permutation to permutation

Rank Sum Test = Permutation Test Performed on Ranks

Recall the permutation test uses the $\boxed{\text{diff. in means}}$ as the test statistic. Performed on ranks, the test statistic would be

$$\begin{pmatrix}\text{Mean ranks}\\ \text{for group 1}\end{pmatrix} - \begin{pmatrix}\text{Mean ranks}\\ \text{for group 2}\end{pmatrix} = \frac{T_1}{n_1} - \frac{T_2}{n_2}$$

where n_i = sizes of Group i, and T_i = rank-sum for Group i, $i = 1, 2$

- Observe that $T_1 + T_2$ equals the rank-sum of all observations

  $$1+2+3+\cdots+(n_1+n_2) = \frac{(n_1+n_2)(n_1+n_2+1)}{2} = \text{rank sum of all}$$

  which doesn't change from permutation to permutation

So diff. in mean ranks equals

$$\frac{T_1}{n_1} - \frac{T_2}{n_2} = \frac{T_1}{n_1} - \frac{(\text{rank sum of all}) - T_1}{n_2} = \frac{T_1}{n_1} + \frac{T_1}{n_2} - \underbrace{\frac{\text{rank sum of all}}{n_2}}_{\text{not change w/ permutation}}$$

Thus

$$\begin{aligned}\text{large value of } T_1 &\iff \text{large diff. in mean ranks}\\ &\iff \text{stronger evidence against } \mathrm{H}_0\end{aligned}$$

Using T_1 as the test-statistic is equivalent to using the diff in mean ranks as the test-statistic

Example: Rat's Diet Revisit

The rank sum test is equivalent to a permutation test performed on the ranks of the data.

	beef diet				cereal diet				Rank Sum Statistic
Data	118	117	104	111	111	92	86	56	
	↓	↓	↓	↓	↓	↓	↓	↓	
Ranks	8	7	4	5.5	5.5	3	2	1	$T = 8 + 7 + 4 + 5.5 = 24.5$

One can find the rank-sum P-value by counting the permutations of ranks in the two groups that give a rank-sum $\geq$ the observed rank-sum statistic, and divide it by $\binom{n_1+n_2}{n_1}$

beef diet				cereal diet				Rank Sum Statistic
8	7	5.5a	5.5b	4	3	2	1	$T = 8 + 7 + 5.5 + 5.5 = 26$
8	7	5.5a	4	5.5b	3	2	1	$T = 8 + 7 + 5.5 + 4 = 24.5$
8	7	5.5b	4	5.5a	3	2	1	$T = 8 + 7 + 5.5 + 4 = 24.5$

The one-sided P-value for the Rank-Sum test is $3/70$ as there are only 3 permutations with a rank-sum $T \geq$ the observed $T = 24.5$

Normal Approximation to the Rank-Sum Statistic

- When n_1 and n_2 are small, one can compute the exact P-value for a rank-sum test by counting the number of permutations of ranks in the two groups that give a rank-sum $\geq$ the observed rank-sum statistic, and dividing it by $\binom{n_1+n_2}{n_1}$.
- However, when n_1 and n_2 get moderately large, computation of the exact P-value of permutation tests is labor-intensive.
- Fortunately, because conversion to ranks avoids absurd distributional anomalies, the **sampling distribution of rank sum statistic can be approximated accurately by a normal distribution**, unless
  - when at least one sample is small (say, under 5),
  - or when large numbers of ties occur.

  See next page.

Normal Approximation to the Rank-Sum Statistic

The rank sum statistic T is approximately Normal

$$T \text{ is approx.} \sim N\left(n_1\overline{R},\ s_R\sqrt{\frac{n_1 n_2}{n_1+n_2}}\right)$$

where $\overline{R}$ and s_R, are the average and the sample SD, respectively, for the combined set of n_1+n_2 ranks.

- In fact, $\overline{R} = (1+n_1+n_2)/2$
- If no ties, $s_R = \sqrt{\dfrac{(n_1+n_2)(n_1+n_2+1)}{12}}$.

Normal Approximation to the Rank-Sum Statistic

The rank sum statistic T is approximately Normal

$$T \text{ is approx.} \sim N\left(n_1\overline{R},\ s_R\sqrt{\frac{n_1 n_2}{n_1+n_2}}\right)$$

where $\overline{R}$ and s_R, are the average and the sample SD, respectively, for the combined set of n_1+n_2 ranks.

- In fact, $\overline{R}=(1+n_1+n_2)/2$
- If no ties, $s_R=\sqrt{\dfrac{(n_1+n_2)(n_1+n_2+1)}{12}}$.

Do NOT use this approximation when

- at least one sample is small (say, under 5), or
- large numbers of ties occur.

In those cases, find the P-value by listing the extreme permutations or by simulation.

Example: Cognitive Load Theory in Teaching

First we find the ranks of the data

```
> Time = c(68,70,73,75,77,80,80,132,148,155,183,197,206,210,
          130,139,146,150,161,177,228,242,265,300,300,300,300,300)
> Treatment = c(rep("Modified", 14), rep("Conventional",14))
> obsrank = rank(Time, ties.method = "average")
> obsrank
 [1]  1.0  2.0  3.0  4.0  5.0  6.5  6.5  9.0 12.0 14.0 17.0 18.0
[13] 19.0 20.0  8.0 10.0 11.0 13.0 15.0 16.0 21.0 22.0 23.0 26.0
[25] 26.0 26.0 26.0 26.0
```

The rank sum statistic T is

```
> T = sum(obsrank[1:14]); T
[1] 137
```

The average $\overline{R}$ and the sample standard deviation s_R, respectively, for the combined set of $n_1 + n_2$ ranks are

```
> meanR = mean(obsrank); meanR
[1] 14.5
> SR = sd(obsrank); SR
[1] 8.202303
```

Example: Cognitive Load Theory in Teaching

As both groups have 14 observations, $n_1 = n_2 = 14$. The mean and SD of the rank sum statistic is

```
> n1=14
> n2=14
> n1*meanR
[1] 203
> sqrt(n1*n2/(n1+n2))*SR
[1] 21.70125
```

The *P*-value is hence

```
> 2*pnorm(137, mean=203, sd=21.70125)
[1] 0.002355599
```

Rank-Sum Test in R — Cognitive Load Experiment

```
> wilcox.test(Time ~ Treatment)

        Wilcoxon rank sum test with continuity correction

data:  Time by Treatment
W = 164, p-value = 0.002542
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(130, 139, 146, 150, 161, 177, 228,  :
  cannot compute exact p-value with ties
> wilcox.test(Time ~ Treatment, correct=F)

        Wilcoxon rank sum test

data:  Time by Treatment
W = 164, p-value = 0.002356
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(130, 139, 146, 150, 161, 177, 228,  :
  cannot compute exact p-value with ties
```

Rank-Sum Test in R — Rat's Diet Experiment

```
> beef = c(104, 118, 117, 111)
> cereal = c(111, 56, 86, 92)
> wilcox.test(beef, cereal, alternative = "greater")

Wilcoxon rank sum test with continuity correction

data:  beef and cereal
W = 14.5, p-value = 0.04071
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(beef, cereal, alternative = "greater") :
  cannot compute exact p-value with ties
```

Note that R calculate the P-value using normal approximation when there are ties. Hence the P-value 0.04071 given is not equal to the exact P-value $3/70 \approx 0.04286$.

Comparisons of the 3 Tests

The rank-sum test, just like the permutation test, is a nonparametric or distribution-free statistical tool, meaning there are no specific distributional assumptions required.

	Two-Sample t-test Welch t-test	Permutation Test	Rank-Sum Test
Requires Normality?	Yes for small samples No for large samples	No	No
Robust to Outliers?	No	No	Yes

When the sample sizes are large, the 3 tests usually give similar p-values and hence will give the same conclusion.

Other names for the rank-sum test are the Wilcoxon test and the Mann-Whitney test. The different names refer to originators of different forms of the test statistic.

Permutation Test for One-Way ANOVA Data

Permutation Test for One-Way ANOVA

Permutation tests for two-sample data can be extended to multi-sample data.

The test-statistic can be the F-statistic or SS_{trt}, which are actually equivalent since

$$F = \frac{\text{SS}_{trt}/(g-1)}{\text{SSE}/(N-g)} = \frac{\text{SS}_{trt}/(g-1)}{(\text{SST} - \text{SS}_{trt})/(N-g)}.$$

as SST doesn't change with permutation. Thus

$$\begin{aligned} \text{large value of } \text{SS}_{trt} &\iff \text{large value of } F\text{-statistic} \\ &\iff \text{stronger evidence against } \text{H}_0\text{: all } \mu_i\text{'s are equa} \end{aligned}$$

So we just use SS_{trt} as the test-statistic.

Permutation Test for One-Way ANOVA

1. Compute the observed SS_{trt}.
2. Permutate the observations among groups, while keeping the size of each group as in the original data. For each permutation, compute the SS_{trt} for that permutation.
3. If the size of the groups are: $n_1, n_2, \ldots, n_g$, the total number of possible permutations are $M = \dfrac{(n_1 + n_2 + \cdots + n_g)!}{n_1!n_2! \cdots n_g!}$. If k out of the M permutations have SS_{trt} greater or equal to the SS_{trt} of the original data set, then the exact P-value is k/M.

Permutation Test in R for One-Way ANOVA Data

Just like two-sample data, it is labor-intensive to to find the exact P-value of permutation test by counting of more extreme cases. Usually we can only estimate the exact P-value by sampling from the possible permutations. We will demonstrate using the Hodgkin's disease data.

First we compute the observed SS_{trt}.

```
> hodgkins = read.table("Hodgkins.txt", header=T)
> anova(lm(BradyLevel ~ Hodgkins, data=hodgkins))
Analysis of Variance Table

Response: BradyLevel
          Df  Sum Sq Mean Sq F value    Pr(>F)
Hodgkins   2  65.893  32.946   10.67 0.0001042 ***
Residuals 62 191.449   3.088
```

Note the (1,2) entry of the ANOVA output is the value of SS_{trt}.

```
> obsSStrt = anova(lm(BradyLevel ~ Hodgkins, data=hodgkins))[1,2]
> obsSStrt
[1] 65.8928
```

Permutation Test in R for One-Way ANOVA Data

Next we permute the response `BradyLevel` using the `sample()` function, and then compute the SS_{trt} for the permuted data.

```
> anova(lm(sample(BradyLevel) ~ Hodgkins, data=hodgkins))[1,2]
[1] 4.580726
```

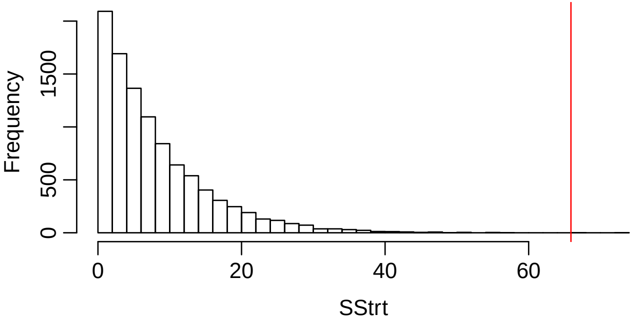
Next, we repeat the previous step a huge number of times. For each repetition, we permute the response and then obtain the SS_{trt} for the permuted data.

```
M = 10000
SStrt = vector("numeric",length=M)
for(i in 1:M){
  SStrt[i]=anova(lm(sample(BradyLevel)~Hodgkins, data=hodgkins))[1,2]
}
```

Permutation Test in R for One-Way ANOVA Data

Now let's take a look at the frequencies of the values of the SS_{trt} we obtained. (The result may vary from simulation to simulation).

```
> hist(SStrt, nclass=50, xlab="SStrt",main="")
> abline(v=obsSStrt, col=2)
> sum(SStrt >= obsSStrt)
[1] 2
```



Among the 10000 SS_{trt}'s, we see only 2 of them are $\geq$ the observed SS_{trt}. So the P-value is estimated to be $2/10000 = 0.0002$.

Kruskal-Wallis test for One-Way ANOVA

Kruskal-Wallis test is simply the rank-sum test extended to one-way ANOVA data.

- First, convert the original data values with their ranks in the entire data set. The smallest value gets a rank of 1, the second-smallest gets a rank of 2, etc. Tied observations get average ranks.
- Though omputation of the exact sampling distribution and the exact P-value of permutation tests are labor-intensive, since ranks are more well-behaved than the original data (unless there are a large number of ties), an accurate approximation of the permutation distribution is

$$\frac{\mathrm{SS}_{trt}}{\sigma_R^2} \sim \chi^2_{g-1}$$

where σ_R^2 is the sample variance of all N ranks and where N is the total number of observations in all groups.

```
> hodgkins = read.table("Hodgkins.txt", header=T)
> obsrank = rank(hodgkins$BradyLevel, ties.method = "average")
> anova(lm(obsrank ~ Hodgkins, data=hodgkins))
Analysis of Variance Table

Response: obsrank
          Df  Sum Sq Mean Sq F value    Pr(>F)
Hodgkins   2  6840.2  3420.1  13.222 1.649e-05 ***
Residuals 62 16036.8   258.7
```

We obtained $\text{SS}_{trt} = 6840.2$.

```
> hodgkins = read.table("Hodgkins.txt", header=T)
> obsrank = rank(hodgkins$BradyLevel, ties.method = "average")
> anova(lm(obsrank ~ Hodgkins, data=hodgkins))
Analysis of Variance Table

Response: obsrank
          Df  Sum Sq Mean Sq F value    Pr(>F)
Hodgkins   2  6840.2  3420.1  13.222 1.649e-05 ***
Residuals 62 16036.8   258.7
```

We obtained $\text{SS}_{trt} = 6840.2$.
The sample variance of the ranks $\sigma_R^2 = 357.4531$ can be obtained

```
> var(obsrank)
[1] 357.4531
```

```
> hodgkins = read.table("Hodgkins.txt", header=T)
> obsrank = rank(hodgkins$BradyLevel, ties.method = "average")
> anova(lm(obsrank ~ Hodgkins, data=hodgkins))
Analysis of Variance Table

Response: obsrank
          Df  Sum Sq Mean Sq F value    Pr(>F)
Hodgkins   2  6840.2  3420.1  13.222 1.649e-05 ***
Residuals 62 16036.8   258.7
```

We obtained $\text{SS}_{trt} = 6840.2$.
The sample variance of the ranks $\sigma_R^2 = 357.4531$ can be obtained

```
> var(obsrank)
[1] 357.4531
```

The Kruskal-Wallis test statistic is

$$\frac{\text{SS}_{trt}}{\sigma_R^2} = \frac{6840.2}{357.4531} = 19.1359 \sim \chi^2_{3-1}$$

The approximate P-value is 6.99×10^{-5}.

```
> pchisq(6840.2/357.4531, df=2, lower.tail=F)
[1] 6.993331e-05
```

Kruskal-Wallis test in R

```
> kruskal.test(BradyLevel~Hodgkins, data=hodgkins)

Kruskal-Wallis rank sum test

data:  BradyLevel by Hodgkins
Kruskal-Wallis chi-squared = 19.136, df = 2, p-value =
6.994e-05
```

Matched-Pair Designs

- Matched-pair designs
- t-test for matched-pair designs
- Randomization test for matched-pair designs
- Wilcoxon signed-rank test

Example: Coffee & Blood Flow During Exercise

Doctors studying healthy men measured myocardial blood flow (MBF)[2] during bicycle exercise after giving the subjects a placebo or a dose of 200 mg of caffeine that was equivalent to drinking two cups of coffee.

Two possible designs:

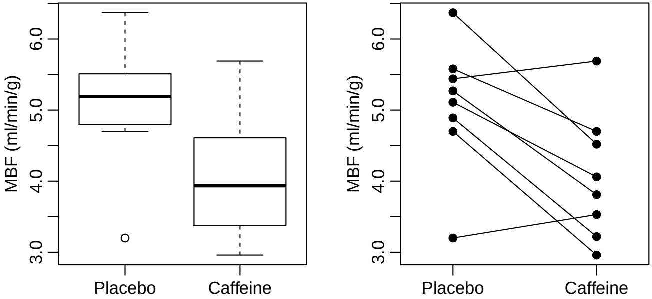
- **Completely Randomized Design**: 16 subjects. Randomly choose 8 subjects to be given caffeine, the other 8 placebo
- **Matched Pairs Design**: 8 subjects, each is tested twice. Randomly choose 4 subjects to receive caffeine in the first test and placebo in the second test; the other 4 receive placebo first and caffeine second. There is a 24-hour gap between the two tests.

Both designs will result in 16 measurements, 8 for caffeine and 8 for placebo. Which design would be more efficient?

[2] MBF was measured by taking positron emission tomography (PET) images after oxygen-15 labeled water was infused in the patients.

Data for Matched-Pair Design

	MBF (ml/min/g)	
Subject	Placebo	Caffeine
1	6.37	4.52
2	5.44	5.69
3	5.58	4.70
4	5.27	3.81
5	5.11	4.06
6	4.89	3.22
7	4.70	2.96
8	3.20	3.53



Matched-pair data cannot be analyzed like 2 independent samples since the 2 measurements on the same subject are *dependent*.

Method: take <u>differences</u> and analyze like **one-sample data**.

	MBF (ml/min/g)		
Subject	Placebo	Caffeine	Difference
j	y_{2j}	y_{1j}	$d_j = y_{2j} - y_{1j}$
1	6.37	4.52	1.85
2	5.44	5.69	−0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	4.06	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.20	3.53	−0.33
Mean	5.07	4.06	1.01
SD	0.91	0.89	0.87

To test H_0: $\mu_1 = \mu_2$, the test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

where

$$s_d = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(d_j - \bar{d})^2}$$

In this example, $\bar{d} = 1.01$, $s_d = 0.87$, $t = \frac{1.01-0}{0.87/\sqrt{8}} \approx 3.28$. The 2-sided P-value is

```
> 2*pt(-3.28,df=7)
[1] 0.01348706
```

Permutation Test for Matched-Pair Data (1)

Under the H_0 that treatments make no difference, if we had reversed the order that Subject 1 received caffeine and placebo, the measurement would be 4.52 for placebo and 6.37 for caffeine, and the difference would become Placebo − Caffeine = −1.85 rather than 1.85.

	MBF (ml/min/g)		
Subject	Placebo	Caffeine	Difference
j	y_{2j}	y_{1j}	$d_j = y_{2j} - y_{1j}$
1	6.37	4.52	1.85
2	5.44	5.69	−0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	5.11	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.20	3.53	−0.33

Permutation Test for Matched-Pair Data (1)

Under the H_0 that treatments make no difference, if we had reversed the order that Subject 1 received caffeine and placebo, the measurement would be 4.52 for placebo and 6.37 for caffeine, and the difference would become Placebo − Caffeine = −1.85 rather than 1.85.

	MBF (ml/min/g)			
Subject	Placebo		Caffeine	Difference
j	y_{2j}		y_{1j}	$d_j = y_{2j} - y_{1j}$
1	4.52	↔	6.37	−1.85
2	5.44		5.69	−0.25
3	5.58		4.70	0.88
4	5.27		3.81	1.46
5	5.11		5.11	1.05
6	4.89		3.22	1.67
7	4.70		2.96	1.74
8	3.20		3.53	−0.33

Permutation Test for Matched-Pair Data (1)

Under the H_0 that treatments make no difference, if we had reversed the order that Subject 1 received caffeine and placebo, the measurement would be 4.52 for placebo and 6.37 for caffeine, and the difference would become Placebo $-$ Caffeine $= -1.85$ rather than 1.85.

	MBF (ml/min/g)			
Subject	Placebo		Caffeine	Difference
j	y_{2j}		y_{1j}	$d_j = y_{2j} - y_{1j}$
1	6.37		4.52	1.85
2	5.44		5.69	-0.25
3	4.70	↔	5.58	-0.88
4	3.81	↔	5.27	-1.46
5	5.11		5.11	1.05
6	4.89		3.22	1.67
7	4.70		2.96	1.74
8	3.53	↔	3.20	$+0.33$

Under H_0, the two measurements for each pair would remain the same no matter whether caffeine or placebo was applied first. Only the order could be swapped. So the difference between the pair would be of the same magnitude but could change sign.

Permutation Test for Matched-Pair Data (1)

Under the H_0 that treatments make no difference, if we had reversed the order that Subject 1 received caffeine and placebo, the measurement would be 4.52 for placebo and 6.37 for caffeine, and the difference would become Placebo − Caffeine = −1.85 rather than 1.85.

	MBF (ml/min/g)			
Subject	Placebo		Caffeine	Difference
j	y_{2j}		y_{1j}	$d_j = y_{2j} - y_{1j}$
1	4.52	↔	6.37	−1.85
2	5.69	↔	5.44	+0.25
3	5.58		4.70	0.88
4	5.27		3.81	1.46
5	4.06	↔	5.69	−1.05
6	4.89		3.22	1.67
7	2.96	↔	4.70	−1.74
8	3.53	↔	3.20	+0.33

Under H_0, the two measurements for each pair would remain the same no matter whether caffeine or placebo was applied first. Only the order could be swapped. So the difference between the pair would be of the same magnitude but could change sign.

Permutation Test for Matched-Pair Data (2)

So for a permutation test, permutation of observations can only be done **within each pair**. No cross-pair permutations are allowed.

Permutation Test for Matched-Pair Data (2)

So for a permutation test, permutation of observations can only be done **within each pair**. No cross-pair permutations are allowed.For all permutations, the magnitude of the differences would remain the same, but the signs could change.

$$\pm 1.85,\ \pm 0.25,\ \pm 0.88,\ \pm 1.46,\ \pm 1.05,\ \pm 1.67,\ \pm 1.74,\ \pm 0.33$$

Permutation Test for Matched-Pair Data (2)

So for a permutation test, permutation of observations can only be done **within each pair**. No cross-pair permutations are allowed.For all permutations, the magnitude of the differences would remain the same, but the signs could change.

$$\pm 1.85,\ \pm 0.25,\ \pm 0.88,\ \pm 1.46,\ \pm 1.05,\ \pm 1.67,\ \pm 1.74,\ \pm 0.33$$

There are $2^8 = 256$ ways of changing the signs of $d_1, d_2, \ldots, d_8$.

Permutation								
1.85	0.25	0.88	1.46	1.05	1.67	1.74	0.33	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	0.33	
1.85	0.25	0.88	1.46	1.05	1.67	1.74	-0.33	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	-0.33	← observed
1.85	-0.25	-0.88	1.46	1.05	1.67	1.74	0.33	
		⋮			⋮			⋮
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	

Test Statistic for the Permutation Test for Paired Data

We use the sum of the differences as the test statistic:

$$\sum_{j=1}^{8} d_j = d_1 + d_2 + \ldots + d_8$$

since the greater the sum of the differences, the stronger the evidences against the H_0.

For the observed randomization, the value of the test-statistic is

$$\begin{aligned} T &= \sum_{j=1}^{8} d_j \\ &= 1.85 - 0.25 + 0.88 + 1.46 + 1.05 + 1.67 + 1.74 - 0.33 \\ &= 8.07 \end{aligned}$$

One-Sided Permutation Tests for Matched-Pair Data

The one-sided P-value is the count of permutations that result in a test statistics $\sum_{j=1}^{8} d_j$ that is at least as great as the test statistic for the observed data, divided by 2^n, where $n =$ the number of pairs.

Permutation								test-statistic $\sum_j d_j$	
1.85	0.25	0.88	1.46	1.05	1.67	1.74	0.33	9.23	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	0.33	8.73	
1.85	0.25	0.88	1.46	1.05	1.67	1.74	-0.33	8.57	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	-0.33	8.07	← observed
1.85	-0.25	-0.88	1.46	1.05	1.67	1.74	0.33	7.47	
		⋮			⋮			⋮	
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	-8.07	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	-8.57	
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	-8.73	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	-9.23	

As there are 4 permutations with $\sum_j d_j \geq$ the observed $\sum_j d_j = 8.07$, the one-sided P-value is $4/2^8 = 0.015625$.

Two-Sided Permutation Tests for Matched-Pair Data

For a two-sided test, the test statistic would be $|\sum_j d_j|$.

The two-sided P-value is the count of permutations that result in a test statistics $|\sum_j d_j|$ that is at least as great as the observed $|\sum_j d_j|$, and then divide the count by 2^n, where $n = \#$ of pairs.

Permutation								test-statistic $\sum_j d_j$	
1.85	0.25	0.88	1.46	1.05	1.67	1.74	0.33	9.23	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	0.33	8.73	
1.85	0.25	0.88	1.46	1.05	1.67	1.74	-0.33	8.57	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	-0.33	8.07	← observed
1.85	-0.25	-0.88	1.46	1.05	1.67	1.74	0.33	7.47	
		⋮			⋮			⋮	
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	−8.07	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	−8.57	
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	−8.73	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	−9.23	

There are 8 permutations with $|\sum_j d_j| \geq$ the observed $|\sum_j d_j| = 8.07$, so the two-sided P-value is $8/2^8 = 0.03125$.

Only Count the Extreme Cases

Like permutation test for two-sample data, practically there is no need to list all 2^n possible permutations.
One just need to list all possible permutations (sign changes) that resulted in a sum of differences $\sum_j d_j$ as large or larger than the one for the observed data.

Permutation Test for Matched-Pair Data in R

As it is labor-intensive to count the more extreme cases, we can estimate the exact *P*-value by sampling from the possible permutations.

First we compute the difference for each pair.

```
> placebo = c(6.37,5.44,5.58,5.27,5.11,4.89,4.70,3.20)
> caffeine = c(4.52,5.69,4.70,3.81,4.06,3.22,2.96,3.53)
> diff = placebo - caffeine
```

We then select the set of pairs to swap (so the difference changes sign).

```
> swap = rbinom(8, size=1, p=0.5); swap
[1] 1 0 0 1 0 0 1 0
```

The sum of differences after permutation is

```
> diffsum = sum(diff[swap==1]) - sum(diff[swap==0]); diffsum
[1] 2.03
```

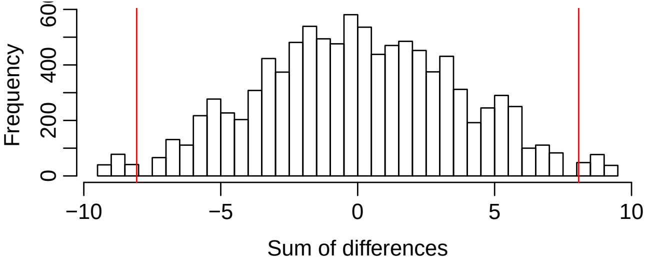
Permutation Test for Matched-Pair Data in R

We repeat the previous step a huge number of times. For each repetition, we swap some of the pairs and then obtain the sum of differences for the permuted data.

```
M = 10000
diffsum = vector("numeric",length=M)
for(i in 1:M){
  swap = rbinom(8, size=1, p=0.5)
  diffsum[i] = sum(diff[swap==1]) - sum(diff[swap==0])
}
```

Let's take a look at the frequencies of the sums of differences we obtained, (which may vary from simulation to simulation).

```
> hist(diffsum, nclass=50,xlab="Sum of difference",main="")
> abline(v=sum(diff), col=2)
> abline(v=-sum(diff), col=2)
> sum(abs(diffsum)>= abs(sum(diff)))
[1] 320
```



Among the 10000 sums of differences, we see 320 of them are $\geq$ the observed $\sum_i d_i = 8.07$. So the two-sided P-value is estimated to be $320/10000 = 0.032$, close to the exact two-sided P-value $8/2^8 = 0.03125$.

Wilcoxon Signed-Rank Test

1. Compute the difference in each of the n pairs.
2. Drop zeros from the list (i.e., drop pairs with no difference).
3. Order the absolute differences from smallest to largest and assign them their ranks $1, \ldots, n$ (or average rank for ties).
4. The signed-rank statistic, S, is the sum of the ranks from the pairs for which the difference is positive.

See the next slide for an example.

Example: Wilcoxon Signed-Rank Test

Subject	MBF (ml/min/g)				
	Placebo	Caffeine	Difference	Rank	Signed Rank
j	y_{2j}	y_{1j}	$d_j = y_{2j} - y_{1j}$		
2	5.44	5.69	-0.25	1	-1
8	3.20	3.53	-0.33	2	-2
3	5.58	4.70	0.88	3	3
5	5.11	4.06	1.05	4	4
4	5.27	3.81	1.46	5	5
6	4.89	3.22	1.67	6	6
7	4.70	2.96	1.74	7	7
1	6.37	4.52	1.85	8	8

$$\begin{aligned} \text{Signed-rank statistic } S &= \text{Sum of ranks for positive differences} \\ &= 3 + 4 + 5 + 6 + 7 + 8 = 33 \end{aligned}$$

Exact *p*-Value of the Wilcoxon Signed-Rank Test

An exact P-value for the signed-rank test is the proportion of all permutations of outcomes within each pair that lead to a test statistic as extreme as or more extreme than the one observed.

- Permutations refers to switching the group status of the two observations within each pair. Within a single pair there are two possible permutations, so with n pairs there are a total of 2^n possible permutations
- The P-value is therefore the number of possible permutations that provide a sum of positive ranks as extreme as or more extreme than the observed one, divided by 2^n

Exact P-Value of the Wilcoxon Signed-Rank Test

Permutation								Signed-Rank Statistic	
1	2	3	4	5	6	7	8	36	
-1	2	3	4	5	6	7	8	35	
1	-2	3	4	5	6	7	8	34	
1	2	-3	4	5	6	7	8	33	
-1	-2	3	4	5	6	7	8	33	← observed
		⋮			⋮			⋮	
-1	-2	3	-4	-5	-6	-7	-8	3	
1	2	-3	-4	-5	-6	-7	-8	3	
-1	2	-3	-4	-5	-6	-7	-8	2	
1	-2	-3	-4	-5	-6	-7	-8	1	
-1	-2	-3	-4	-5	-6	-7	-8	0	

There are 5 permutations that result in a signed-rank statistic $\geq$ the one observed 33. So the one sided P-value is $5/2^8 \approx 0.01953$. For a two-sided test, there are 10 permutations, P-value $= 10/2^8 \approx 0.039$.

Wilcoxon Signed-Rank Test in R

```
> placebo = c(6.37,5.44,5.58,5.27,5.11,4.89,4.70,3.20)
> caffeine = c(4.52,5.69,4.70,3.81,4.06,3.22,2.96,3.53)
> wilcox.test(placebo, caffeine, paired=TRUE, alternative="greater")

        Wilcoxon signed rank test

data:  placebo and caffeine
V = 33, p-value = 0.01953
alternative hypothesis: true location shift is greater than 0

> wilcox.test(placebo, caffeine, paired=TRUE, alternative="two.sided",

        Wilcoxon signed rank test

data:  placebo and caffeine
V = 33, p-value = 0.03906
alternative hypothesis: true location shift is not equal to 0
```

By default, if there are no ties and the number of pairs $n < 50$, R can produce the exact P-value. When there are ties or $n \geq 50$, R will use a normal approximation to calculate an approx. P-value.

Wilcoxon Signed-Rank Test in R

The R command `wilcox.test()` can perform both the signed-rank test for paired data and the rank-sum test for two-sample data.

Without specifying `paired=TRUE`, `wilcox.test()` will perform the rank-sum test for two-sample data.

```
> wilcox.test(placebo, caffeine, alternative="greater", exact=TRUE)

        Wilcoxon rank sum test with continuity correction

data:  placebo and caffeine
W = 50.5, p-value = 0.02926
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(placebo, caffeine, alternative = "greater",  :
  cannot compute exact p-value with ties
```

Normal Approximated P-value

Finding the exact P-value by counting of more extreme cases is lots of work. A normal approximation for convenient computation of an approximate P-value is available. Signed-rank statistic S is approximately

$$N\left(\mu = \frac{n(n+1)}{4},\ \sigma = \sqrt{\frac{\sum_i R_i^2}{4}}\right)$$

where n is the number of pairs (excluding pairs with no difference).

- R_i's are the (unsigned) ranks of the absolute differences of the pairs
- When there are no tie, $\sum_i R_i^2 = n(n+1)(2n+1)/6$.
- This normal approximation works well for $n \geq 20$.

Example

Suppose for some matched-pair data, the signed-ranks are

$$1, -2, -3, -4.5, 4.5, -6, -7, -8, -9, 10, 11, 12, 13, 14, 15.$$

There are $n = 15$ pairs and the signed-rank statistic is

$$S = 1 + 4.5 + 7 + 10 + 11 + 12 + 13 + 14 + 15 = 87.5$$

The mean and SD of the normal approximation are

$$\mu = \frac{n(n+1)}{4} = \frac{15 \times 16}{4} = 60$$

$$\sigma = \sqrt{\frac{\sum_i R_i^2}{4}} = \sqrt{\frac{1^2+2^2+3^2+4.5^2+4.5^2+6^2+\cdots+15^2}{4}} = \sqrt{\frac{1239.5}{4}} \approx 17.60$$

So S is approx. $N(\mu = 60, \sigma \approx 17.60)$. The one-sided P-value is about

$$P(S \geq 87.5) = P\left(Z \geq \frac{87.5 - 60}{17.60}\right) = P(Z \geq 1.56) \approx 0.059.$$

Wilcoxon Signed-Rank Test in R

When there are ties, Wilcoxon signed-rank test in R always uses a normal approximation to calculate an approximate *P*-value.

```
> d = c(1,-2,-3,-4.5,4.5,-6,7,-8,-9,10,11,12,13,14,15)
> wilcox.test(d, alternative="greater", correct=F)

        Wilcoxon signed rank test

data:  d
V = 87.5, p-value = 0.05912
alternative hypothesis: true location is greater than 0

Warning message:
In wilcox.test.default(d, alternative = "greater", correct = F) :
  cannot compute exact p-value with ties
```

Parametric v.s. Nonparametric

- Permutation tests and rank-based tests require less assumptions about the population distribution than t- or F-tests, and hence are more reliable
- But there is no free lunch, permutation tests and rank-based tests have less power than t- or F-tests, in particular when the sample sizes are very small
- For example, consider the following made-up data: the response is 1,2,3 in one group and 101,102,103 in the other group.
  - The t-test gives the two-sided P-value of 3×10^{-8}
  - However, permutation test and rank-sum test only comes up with a two-sided p-value of $2/\binom{6}{3} = 0.1$.

Parametric v.s. Nonparametric

- Don't read too much into this, however
- The difference in power is far less dramatic when the sample size is larger (for large sample sizes, the rank-sum test is about 95% as powerful as the t-test, even when the outcome is normally distributed)
- Furthermore, when outliers/skewness are present, nonparametric methods can be much more powerful than t-tests or F-tests
- Parametric vs. nonparametric:
  - Parametric advantages: More powerful when parametric assumptions hold, straightforward confidence intervals
  - Nonparametric advantages: Minimal assumptions, more powerful when parametric assumptions are wrong