Topic 7 – Non-parametric tests

ENVX1002 Introduction to Statistical Methods

Januar Harianto

The University of Sydney

Dec 2024

Evaluation task (Week 8)

- Testing materials in weeks 2, 5 and 6 (excludes this week).
- See this Ed announcement for all information.
- **No GenAI**: we will check.
- Practice before the test using the Practice Assignment on Canvas

Parametric vs non-parametric methods

Overview

Parametric methods

Depends on the assumption that the data is normally distributed with mean μ and standard deviation σ ,e.g. t -test, ANOVA, linear regression.

Non-parametric methods

Do **not** make *any* assumptions about the distribution of the data.

Uses other properties e.g. ranking of the data, e.g. Wilcoxon signed-rank test, Mann-Whitney U test, Kruskal-Wallis test.

Rank-based tests

General idea

- Rank the data e.g., from smallest to largest.
- Replace the data with their ranks.
- Perform the test on the ranks.

It’s *kind of* like a transformation…

For the **Wilcoxon signed-rank test** suppose we have the following data:

sample:	12	10	8	6	4	10	8	6	10

We arrange the data in ascending order (*similar values are given the same colour for illustration*):

ordered:	4	6	6	8	8	10	10	10	12

Then, we rank the data:

ordered ranks:	1	2	3	4	5	6	7	8	9

Finally, ranks that are *tied* are given the average rank:

final rank:	1	2.5	2.5	4.5	4.5	7	7	7	9

These ranks are then used to perform the test, instead of the original data.

Use case

Two-sample *t*-test

Consider two sets of **identical** data that compares between a group **A** and **B**, where one contains an outlier.

Data:

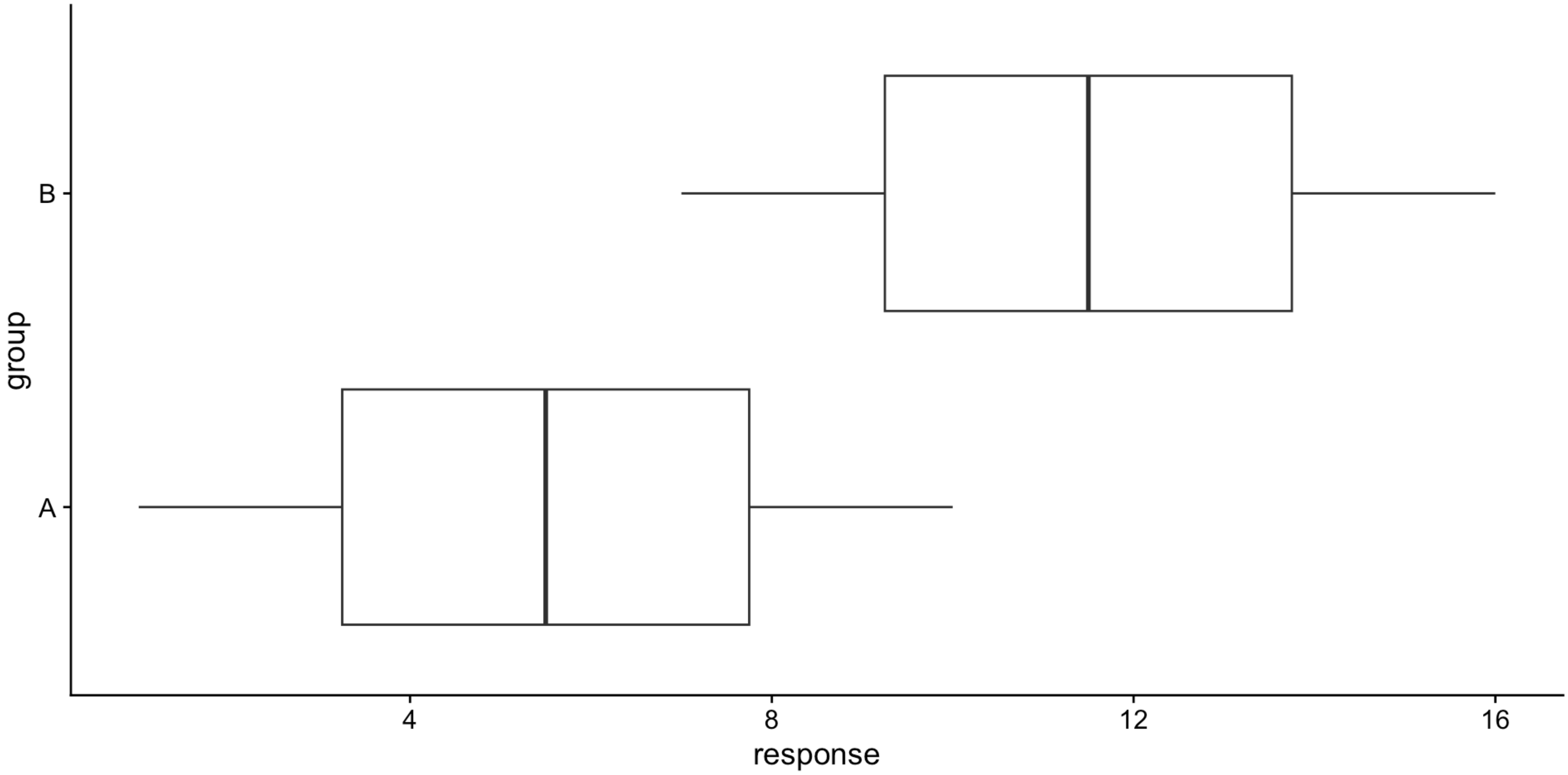
A	B
1	7
2	8
3	9
4	10
5	11
6	12
7	13
8	14
9	15
10	16

Data *with* outlier:

A	B
1	7
2	8
3	9
4	10
5	11
6	12
7	13
8	14
9	15
10	200

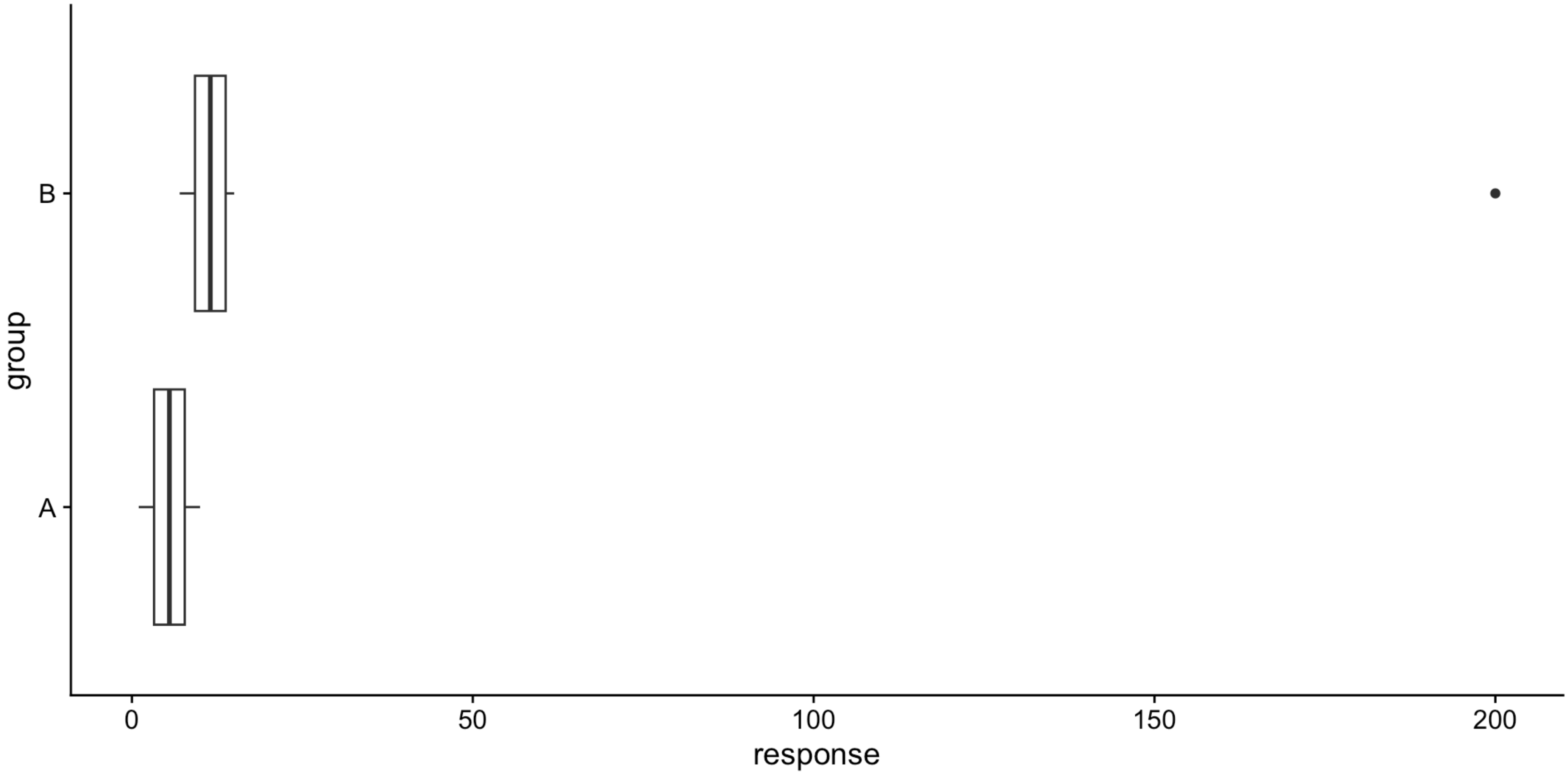
Should there be a difference?

Without the outlier, the data would have been normally distributed.



Outlier

The same data, but with a single outlier in group **B**:



Analysis

If we perform t-tests on both data sets, we get the following results:

```
	Welch Two Sample t-test

data:  response by group
t = -4.4313, df = 18, p-value = 0.0003224
alternative hypothesis: true difference in means
between group A and group B is not equal to 0
95 percent confidence interval:
 -8.844662 -3.155338
sample estimates:
mean in group A mean in group B 
            5.5            11.5 
```

Results indicate that there is a statistically significant difference between the two groups (t_{18} = -4.4, $p < 0.05$).

```
	Welch Two Sample t-test

data:  response by group
t = -1.2882, df = 9.0461, p-value = 0.2297
alternative hypothesis: true difference in means
between group A and group B is not equal to 0
95 percent confidence interval:
 -67.21615  18.41615
sample estimates:
mean in group A mean in group B 
            5.5            29.9 
```

Results indicate that the two groups are **not** significantly different (t_{18} = -2.1, p = 0.23).

The real difference between the two groups is *obscured* by the outlier. **Type II error** (false negative)?

Non-parametric alternatives

When to use

1. If assumptions are met (normality, homogeneity of variance), use parametric tests as they are more powerful and efficient than non-parametric tests.
2. If the normality assumption is violated, transform the data and check for normality again (*optional*).
3. **Non-parametric tests are a good way to deal with circumstances in which parametric tests perform "poorly".**

What to use

Parametric tests	Non-parametric counterpart
One-sample t-test	Wilcoxon signed-rank test
Two-sample t-test	Mann-Whitney U test
ANOVA	Kruskal-Wallis test
Pearson's correlation	Spearman's rank correlation

All of the non-parametric techniques above convert the data into **ranks** before performing the test.

ⓘ **Note**

We will focus on the **Wilcoxon signed-rank test** and the **Mann-Whitney U test**.

Wilcoxon signed-rank test

Altenrative to the one-sample t-test and the paired t-test.

Overview

The Wilcoxon signed-rank test is a non-parametric test used to compare two related samples, matched pairs, or repeated measures on a single sample.

Is an alternative to:

- One-sample t-test
- Paired t-test

Assumptions

- Data comes from the same population
- Data are randomly and independently sampled

Basically, used in same situations as the one-sample or paired t-test, but when the data is not normally distributed but **still symmetric**.

Calculating ranks

If comparing two groups, the ranks are calculated as follows:

1. Calculate the difference D between the two groups.
2. Rank the absolute values of the differences in ascending order.
3. Assign the sign of the difference to the rank.
4. Sum the ranks for each group – **zero differences are ignored**.

ⓘ **Note**

See Slide 5 to recall how ranks are calculated, but we will show another example in the next slide.

Example: Paired data

Weight gain

We measured weight gain in chickens before and after a diet.

chicken	weight	weight_after
1	2.5	4.0
2	3.5	5.0
3	3.5	5.0
4	3.4	4.6

Is there a significant increase in weight gain after the diet?

Rank values

chicken	weight	weight_after	D	Sign	rank	Signed rank
1	2.5	4.0	1.5	+	3	3
2	3.5	5.0	1.5	+	3	3
3	3.5	5.0	1.5	+	3	3
4	3.4	4.6	1.2	+	1	1

Note

The order of the ranks is based on the **absolute** values of the differences; the signs are assigned afterward.

Hypothesis

> Is there a significant increase in weight gain after the diet?

$$H_0 : \mu_{\text{before}} = \mu_{\text{after}}$$

$$H_1 : \mu_{\text{before}} < \mu_{\text{after}}$$

In words:

- H_0: There is no difference in weight gain before and after the diet.
- H_1: There is an increase in weight gain after the diet.

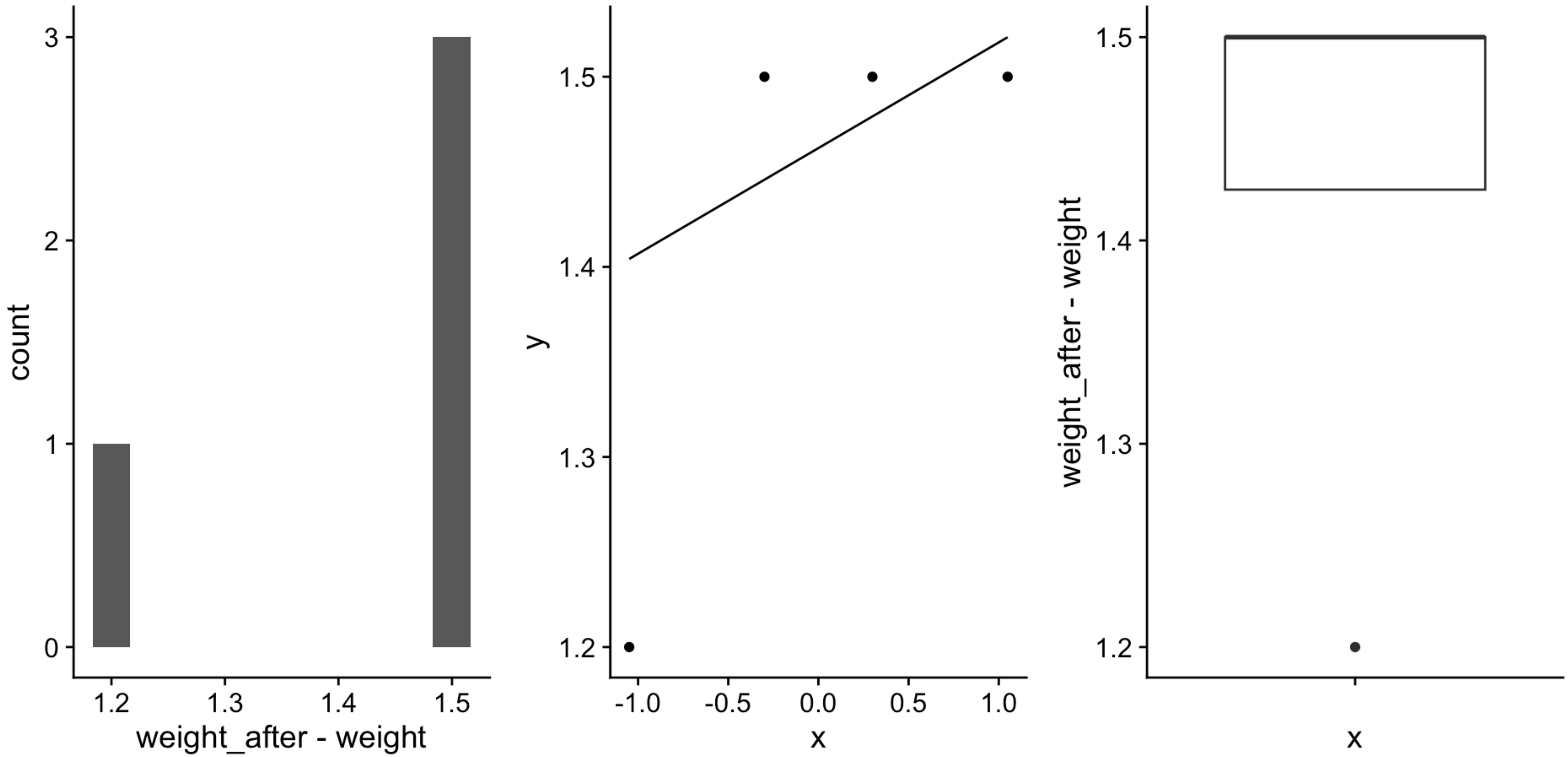
Alternatively, since the data is paired, we may also consider hypotheses based on the differences between the two groups:

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D > 0$$

where D is the difference between the two groups.

Assumptions



With so few data points, we may want to use a formal test to check for normality.

Assumptions

```

    Shapiro-Wilk normality test

data:  df$weight_after - df$weight
W = 0.62978, p-value = 0.001241
```

Results indicate that the data significantly deviates from normality (W = 0.63, p < 0.05). We will use the Wilcoxon signed-rank test.

Performing the test

In R

```
    Wilcoxon signed rank test with continuity correction

data:  df$weight_after and df$weight
V = 10, p-value = 0.04449
alternative hypothesis: true location shift is greater than 0
```

where `V` is the sum of the signed ranks.

The results indicate that there is a **significant increase in weight gain** after the diet ($V = 10$, $p < 0.05$).

Example: One-sample data

Beetle consumption in lizards

Researchers investigated differences in beetle consumption between two size classes of eastern horned lizard (*Phrynosoma douglassi brevirostre*)

- **Larger class**: adult females.
- **Smaller class**: adult males, yearling females.

Focusing on just the smaller size class (for now) – it was hypothesised that this size class would eat a minimum of 100 beetles per day.

Hypothesis

Does the average smaller size class lizard eat about 100 beetles per day?

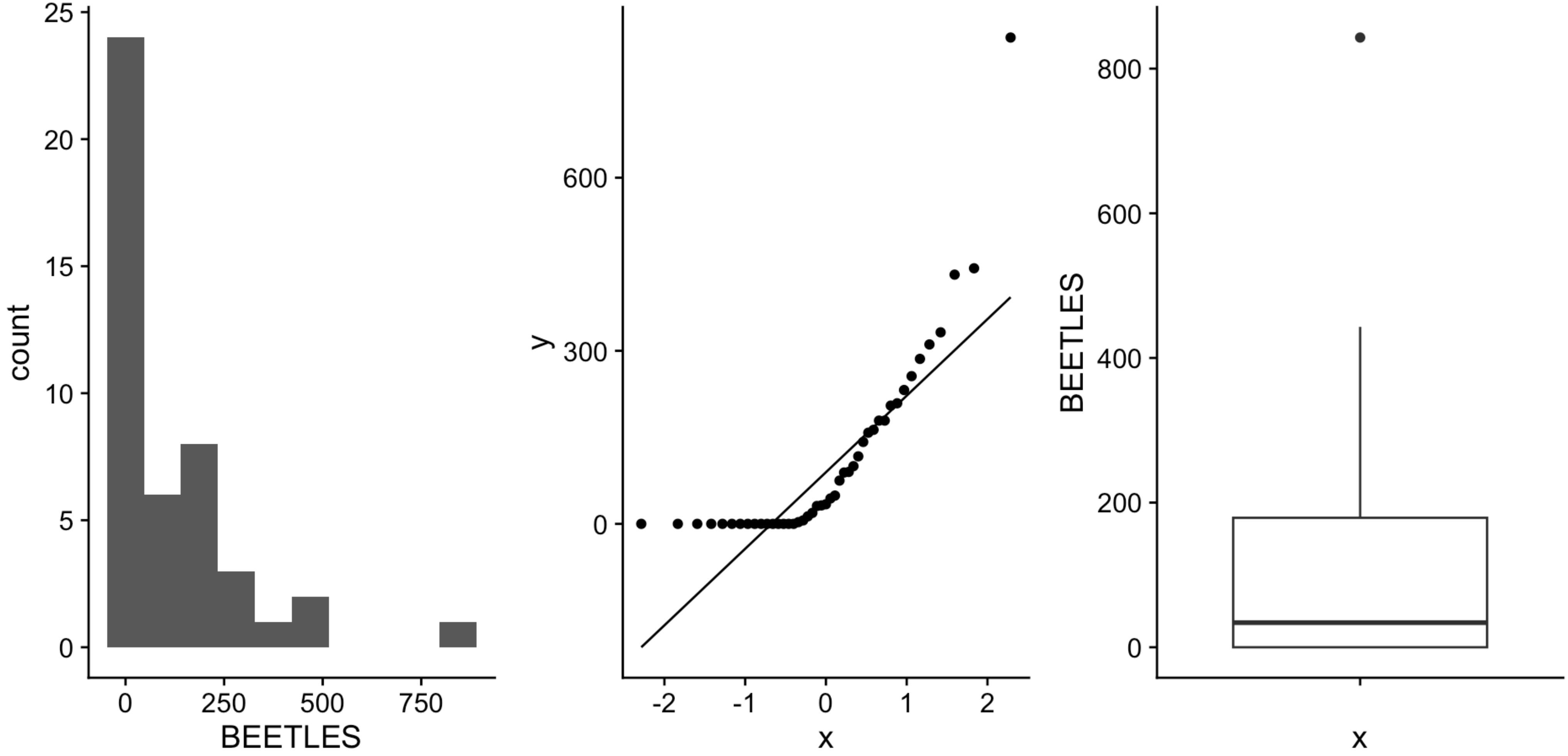
$$H_0 : \mu = 100$$

$$H_1 : \mu \neq 100$$

Dataset Download

```
Rows: 45
Columns: 2
$ SIZE    <chr> "small", "small", "small", "small", "small", "small", "small",…
$ BEETLES <dbl> 256, 209, 0, 0, 0, 44, 49, 117, 6, 0, 0, 75, 34, 13, 0, 90, 0,…
```

First, check assumptions



Is it normally distributed?

Run the test

The Wilcoxon signed-rank test for one sample can be performed as follows:

```
    Wilcoxon signed rank test with continuity correction

data:  .
V = 92, p-value = 0.09755
alternative hypothesis: true location is not equal to 100
```

Results indicate that the average number of beetles consumed by the smaller size class lizard is **not significantly different from 100** (V = 92, p = 0.1).

Important

We are unable to make a conclusion about effect size from non-parametric tests as the information is lost when the data is transformed into ranks.

Mann-Whitney U test

Alternative to the **two-sample t-test**.

Also called the Mann–Whitney–Wilcoxon (MWW/MWU), ***Wilcoxon rank-sum test (what R calls it)****, or Wilcoxon–Mann–Whitney test*.

About

- A **non-parametric test** used to compare two independent samples similar to the two-sample t-test.
- Like the Wilcoxon signed-rank test, it uses **ranks** to perform the test and does not assume normality.
- It is also more *relaxed* in that it does not assume symmetry in the distribution of the data – instead, it assumes that the two groups have the same shape/distribution.

Example: Back to the lizards

Beetle consumption in lizards

Researchers investigated differences in beetle consumption between two size classes of eastern horned lizard (*Phrynosoma douglassi brevirostre*)

- **Larger class**: adult females.
- **Smaller class**: adult males, yearling females.

We will now compare the number of beetles consumed by the **larger** and **smaller** size classes of lizards.

Hypotheses

> Are the number of beetles consumed by the larger and smaller size classes of lizards different?

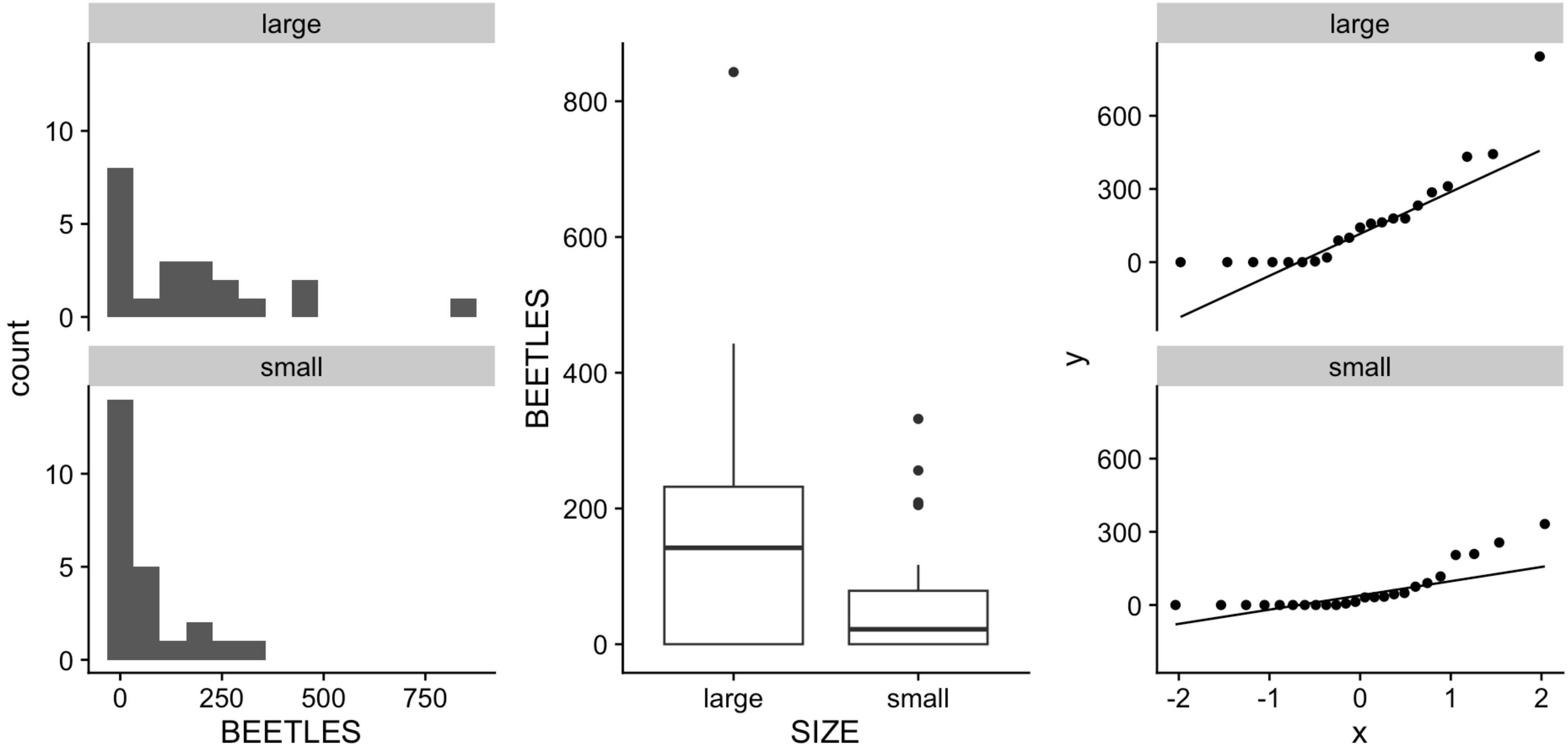
Loosely speaking, because we are not assuming symmetry, the most appropriate summary statistic to use when comparing the two groups is the **median**.

$$H_0 : median_{\text{larger}} = median_{\text{smaller}}$$

$$H_1 : median_{\text{larger}} \neq median_{\text{smaller}}$$

More accurately, we are testing for a difference in the *distribution* of the two groups.

Assumptions



Data does not meet the normality assumption.

Test statistic

The same function `wilcox.test()` can be used to perform the Mann-Whitney U test.

```

	Wilcoxon rank sum test with continuity correction

data:  BEETLES by SIZE
W = 329, p-value = 0.07494
alternative hypothesis: true location shift is not equal to 0
```

- `W` is the sum of the ranks of the *smaller* group.
- The “true location shift” is the median of the larger group minus the median of the smaller group.

The results indicate that the number of beetles consumed by the larger and smaller size classes of lizards is **not significantly different** (W = 329, p = 0.07).

What about transformations?

Transform, or non-parametric?

- As usual, there is ongoing debate on whether to transform the data or use non-parametric tests, but the general consensus is to always prefer parametric tests and transformations when assumptions are met using those techniques.
  - ➟ e.g. Parametric analysis of transformed data is more powerful than non-parametric analysis
- Some argue that non-parametric tests **must** be decided during experimental design and not when the data fails to meet the normality assumption: **as the decision to rank data has implications on the interpretation of the results.**
  - ➟ e.g. Graphpad Advice: Don't automate the decision to use a nonparametric test
- The conventional wisdom is to **transform the data** and check for normality if the assumption is not met. If the data is *still* not normal, then use non-parametric tests (after considering the implications on interpretation).
  - ➟ Or, use bootstrapping (next week!).

Summary

- **Wilcoxon signed-rank test**: alternative to the one-sample t-test and paired t-test.
- **Mann-Whitney U test**: alternative to the two-sample t-test.
- **Advantages**: **Robust** to outliers, skewness, and non-normality.
- **Drawbacks**: Less powerful than parametric tests when assumptions are met, provide no insight into the size of the effect.

Questions?