MONASH University

ETC5521: Exploratory Data Analysis

Using computational tools to determine whether what is seen in the data can be assumed to apply more broadly

Lecturer: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 11 - Session 1

Revisiting hypothesis testing

Testing coin bias Part 1/2

- Suppose I have a coin that I'm going to flip 
- If the coin is unbiased, what is the probability it will show heads?
- Yup, the probability should be 0.5.
- So how would I test if a coin is biased or unbiased?
- We'll collect some data.
- **Experiment 1**: I flipped the coin 10 times and this is the result:

- The result is 7 head and 3 tails. So 70% are heads.
- Do you believe the coin is biased based on this data?

Testing coin bias Part 2/2

- **Experiment 2**: Suppose now I flip the coin 100 times and this is the outcome:

- We observe 70 heads and 30 tails. So again 70% are heads.
- Based on this data, do you think the coin is biased?

(Frequentist) hypotheses testing framework

- Suppose X is the number of heads out of n independent tosses.
- Let p be the probability of getting a head for this coin.

Hypotheses	$H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$
Assumptions	Each toss is independent with equal chance of getting a head.
Test statistic	$X \sim B(n, p)$. Recall $E(X) = np$. The observed test statistic is denoted x.
P-value (or critical value or confidence interval)	$P(\mid X - np \mid \geq \mid x - np \mid)$
Conclusion	Reject null hypothesis when the p-value is less than some significance level α. Usually $\alpha = 0.05$.

- The p-value for experiment 1 is $P(\|X - 5\| \geq 2) \approx 0.34$.
- The p-value for experiment 2 is $P(\|X - 50\| \geq 20) \approx 0.00008$.

Judicial system

		Jury's verdict	
		Not guilty	Guilty
Defendant's true status	Innocent	Correct decision	Convicted an innocent person
	Guilty	Freed a criminal	Correct decision

	Fail to reject Ho	Reject Ho
Ho is true	Correct decision	Type I error
Ho is false	Type II error	Correct decision

- Evidence by test statistic
- Judgement by p-value, critical value or confidence interval

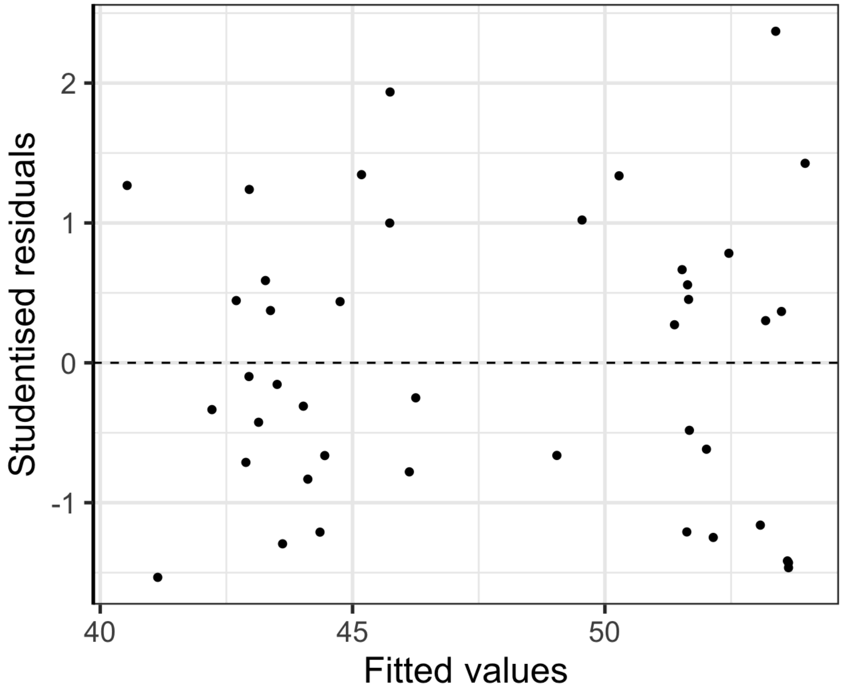
Does the test statistic have to be a *numerical summary statistics*?

Visual inference

Visual inference

- Hypothesis testing in visual inference framework is where:
  - the *test statistic is a plot* and
  - judgement is by human perceptions.
- You (and many other people) actually do visual inference many times but generally in an informal fashion.
- Here, we are making an inference on whether the residual plot has any patterns based on a single data plot.

From Exercise 4 in week 9 tutorial: a residual plot after modelling high-density lipoprotein in human blood.



- Data plots tend to be over-interpreted

- Reading data plots require calibration

Visual inference more formally

1. State your null and alternate hypotheses.
2. Define a **visual test statistic**, $V(.)$, i.e. a function of a sample to a plot.
3. Define a method to generate **null data**, $\boldsymbol{y}_0$.
4. $V(\boldsymbol{y})$ maps the actual data, $\boldsymbol{y}$, to the plot. We call this the **data plot**.
5. $V(\boldsymbol{y}_0)$ maps a null data to a plot of the same form. We call this the **null plot**. We repeat this $m-1$ times to generate $m-1$ null plots.
6. A **lineup** displays these m plots in a random order.
7. Ask n human viewers to select a plot in the lineup that looks different to others without any context given.

Suppose x out of n people detected the data plot from a lineup, then

- the **visual inference p-value** is given as

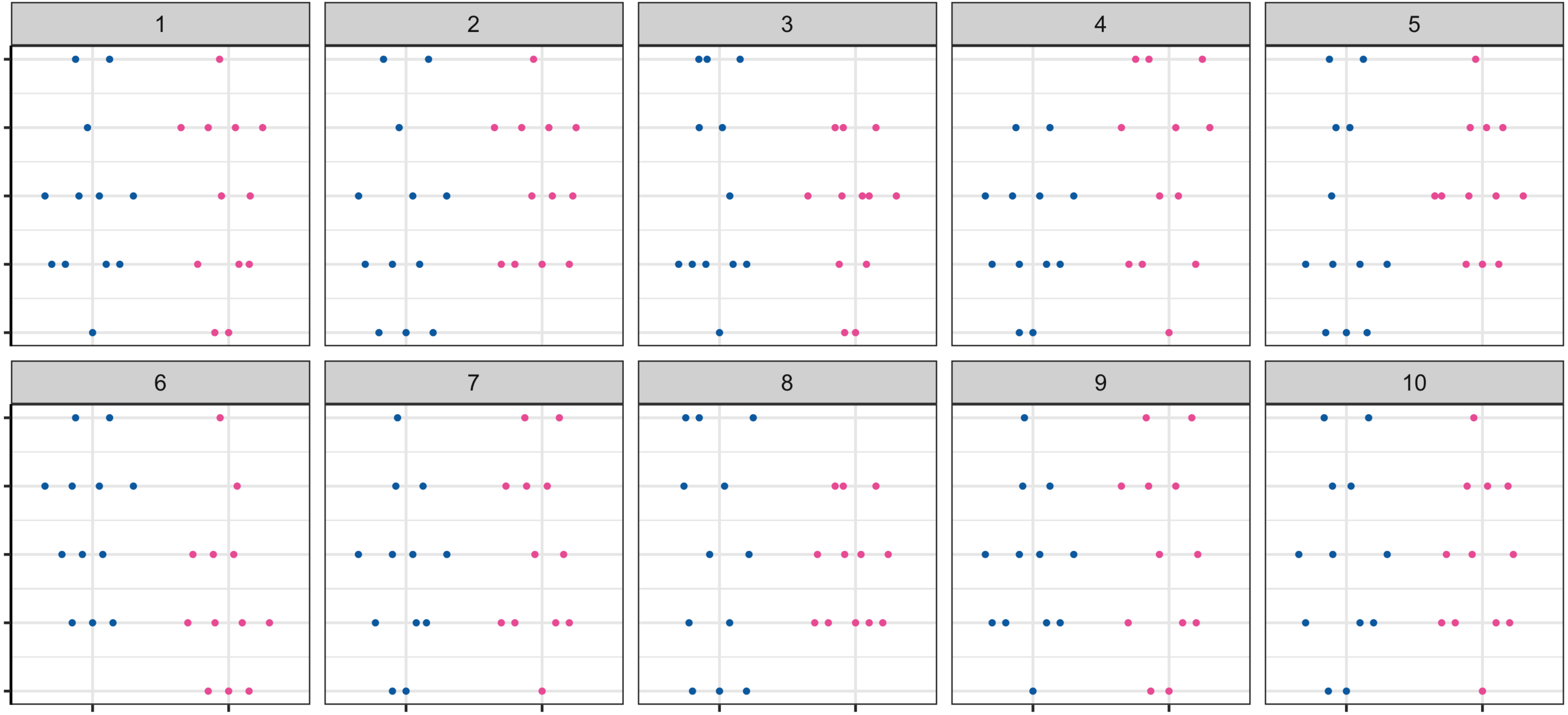
$$P(X \geq x)$$

where $X \sim B(n, 1/m)$, and

- the **power of a lineup** is estimated as x/n.

Lineup 1 In which plot is the pink group higher than the blue group?

- Note: there is no correct answer here.



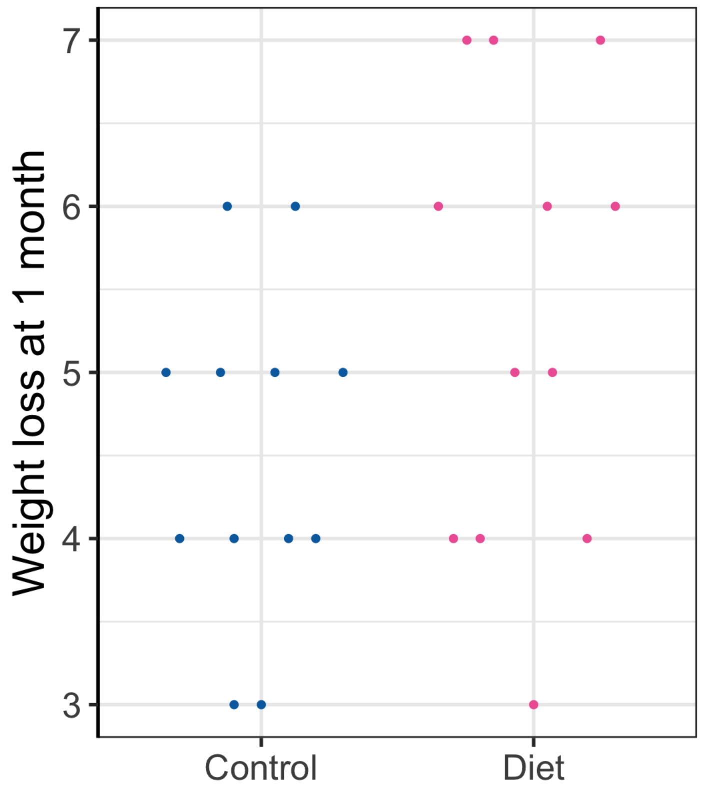
Visual inference p-value (or "see"-value)

- So x out of n of you chose the data plot.
- So the visual inference p-value is $P(X \geq x)$ where $X \sim B(n, 1/10)$.
- In R, this is

```r
1 - pbinom(x - 1, n, 1/10)
# OR
nullabor::pvisual(x, n, 10)
```

Case study 1 Weight loss by diet

📊 data R



This is actually Plot 4 in the previous lineup.

- Is weight loss greater with diet after 1 month?

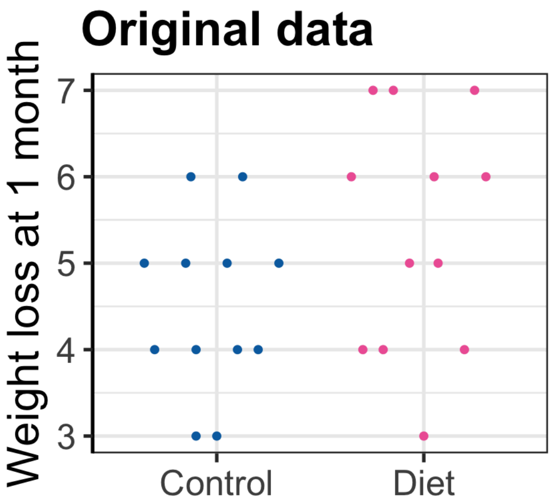
Group	N	Mean	Std. Dev
Control	12	4.50	1.00
Diet	12	5.33	1.37

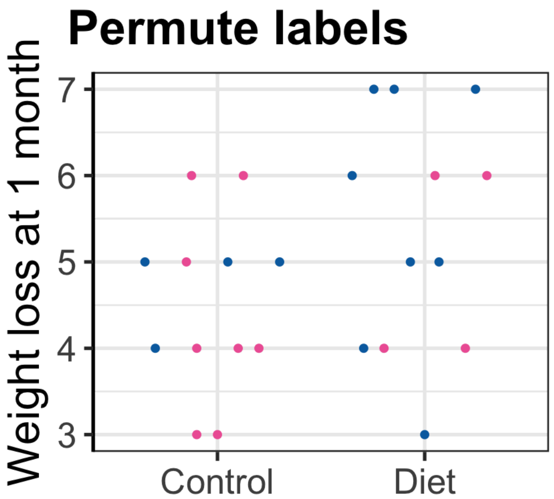
```
with(df,
     t.test(wl1[group=="Diet"], wl1[group=="Control"],
            alternative = "greater"))
```

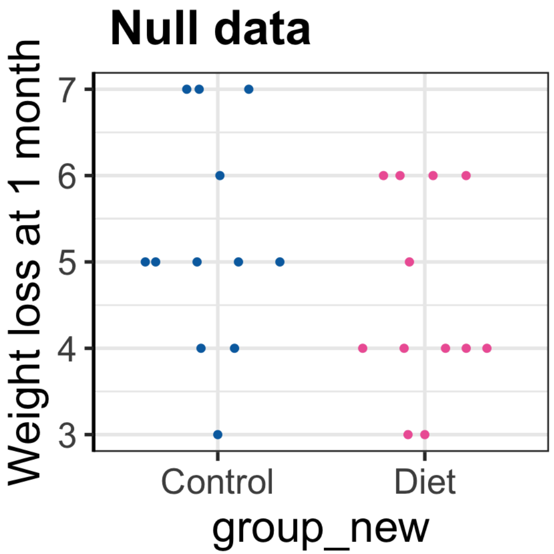
```
##
##     Welch Two Sample t-test
##
## data:  wl1[group == "Diet"] and wl1[group == "Control"]
## t = 1.7014, df = 20.125, p-value = 0.05213
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.01117097         Inf
## sample estimates:
## mean of x mean of y
##  5.333333  4.500000
```

Null data generation method

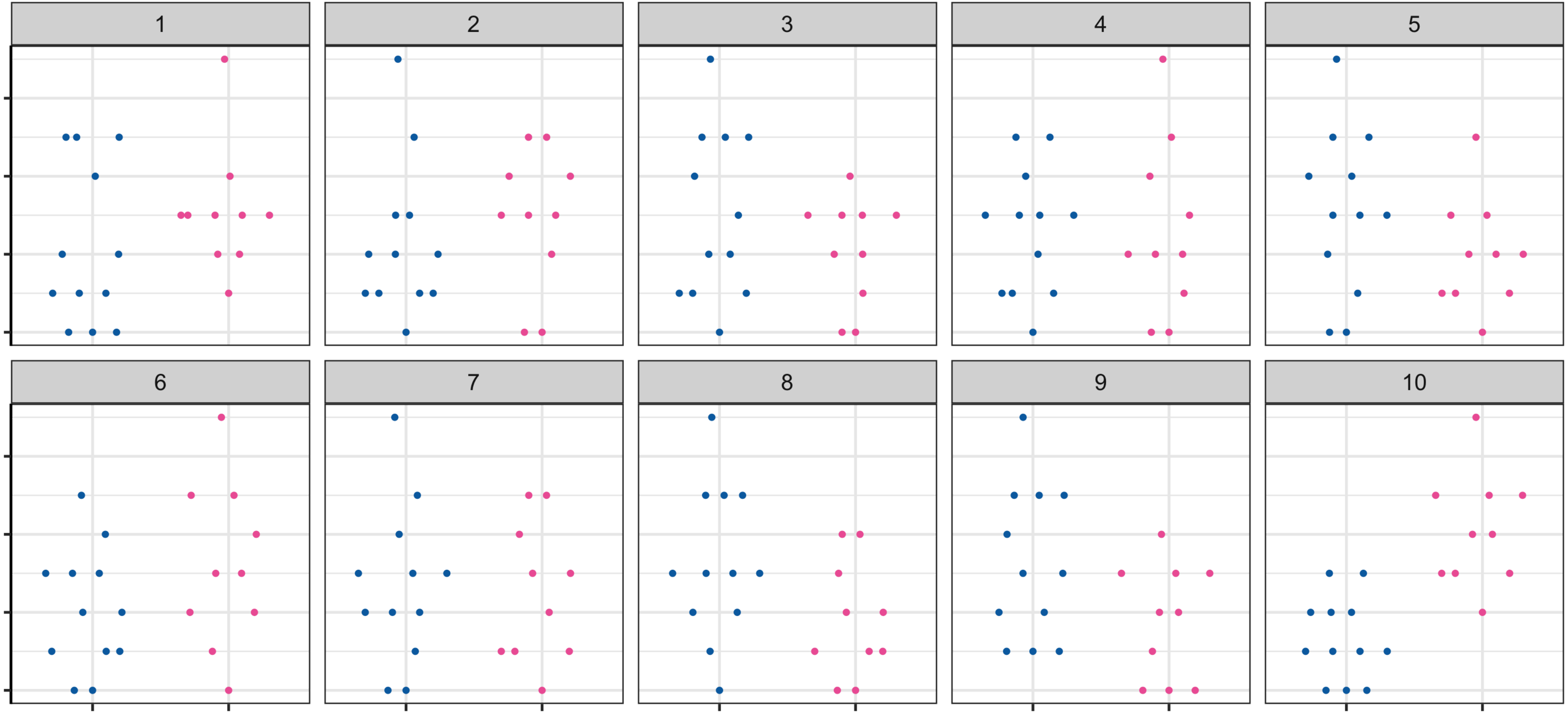
- We are testing $H_0 : \mu_{diet} = \mu_{control}$ vs. $H_1 : \mu_{diet} > \mu_{control}$ where μ_{diet} and $\mu_{control}$ are the average weight loss for population on diet and no diet, respectively.
- There are a number of ways to generate null data under H_0, e.g.
  - we could assume a parametric distribution of the data and estimate the parameters from the data, or
  - we could permute the labels for the diet and control group.





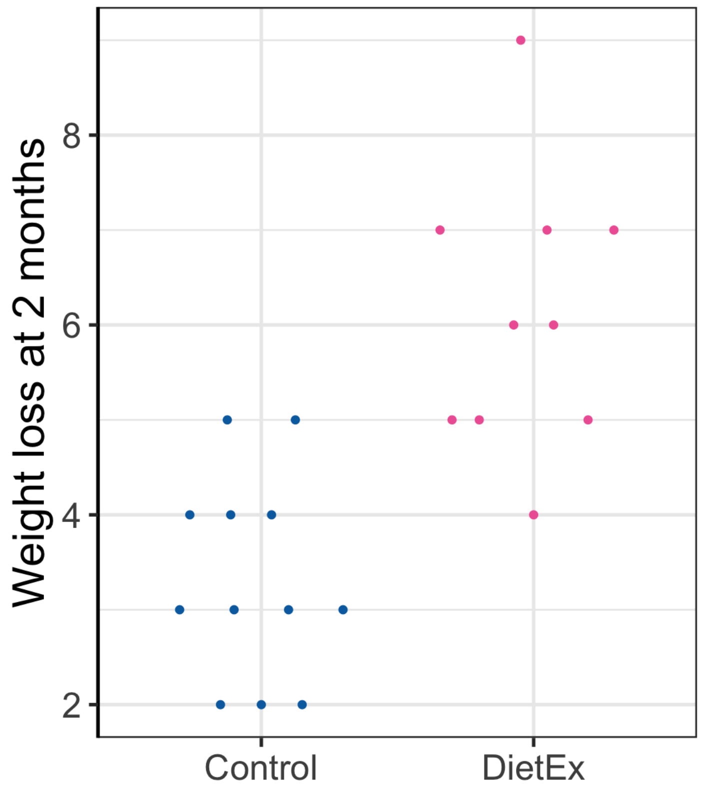


Lineup 2 In which plot is the pink group higher than the blue group?



Case study 1 Weight loss by diet and exercise

📊 data R



- Is weight loss greater with diet *and exercise after 2 months*?

Group	N	Mean	Std. Dev
Control	12	3.33	1.07
DietEx	10	6.10	1.45

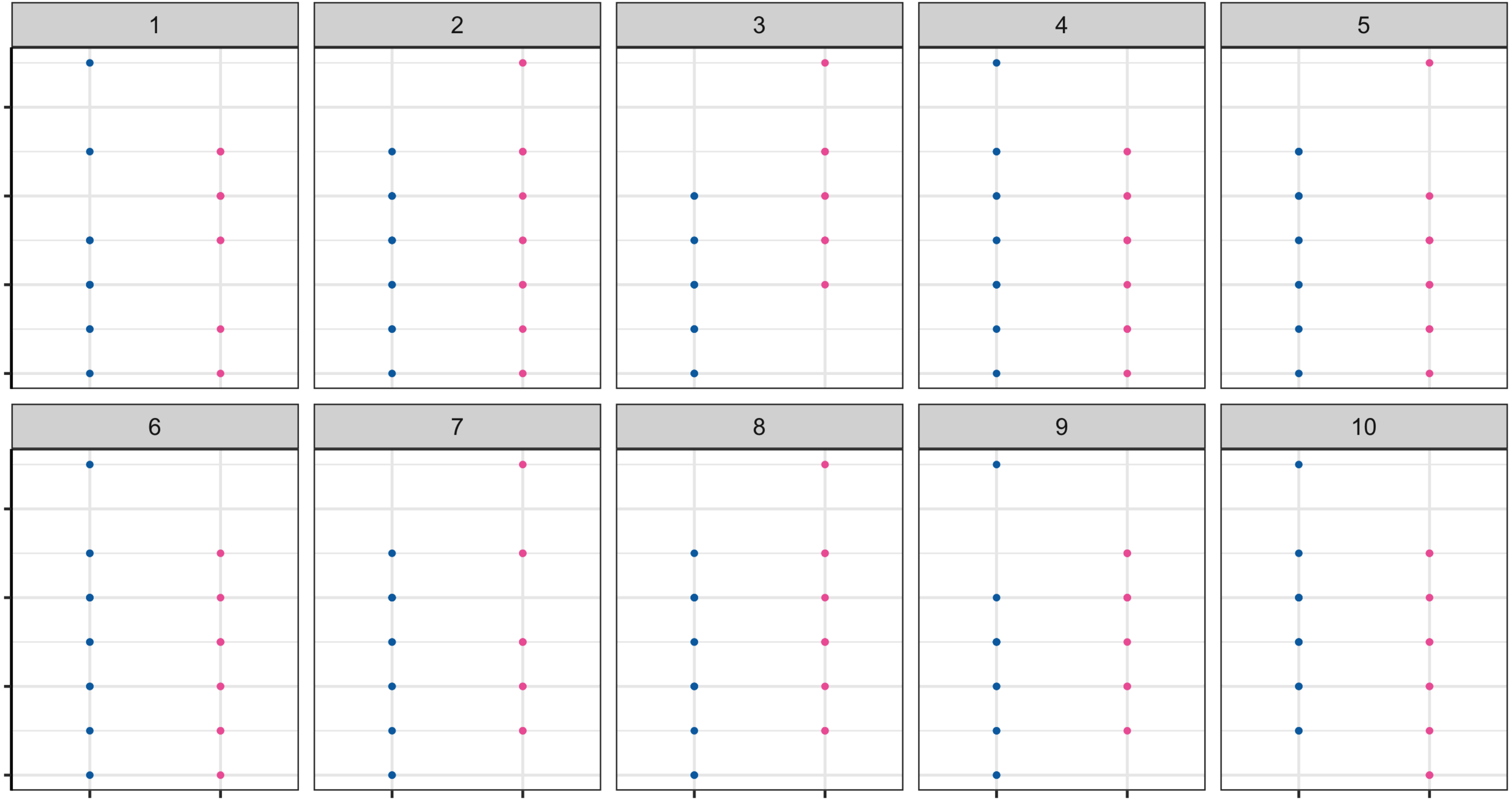
```
with(df2,
     t.test(wl2[group=="DietEx"], wl2[group=="Control"],
            alternative = "greater"))

##
##     Welch Two Sample t-test
##
## data:  wl2[group == "DietEx"] and wl2[group == "Control"]
## t = 5.0018, df = 16.317, p-value = 6.155e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.802104      Inf
## sample estimates:
## mean of x mean of y
##  6.100000  3.333333
```

What about if we change the visual test statistic?

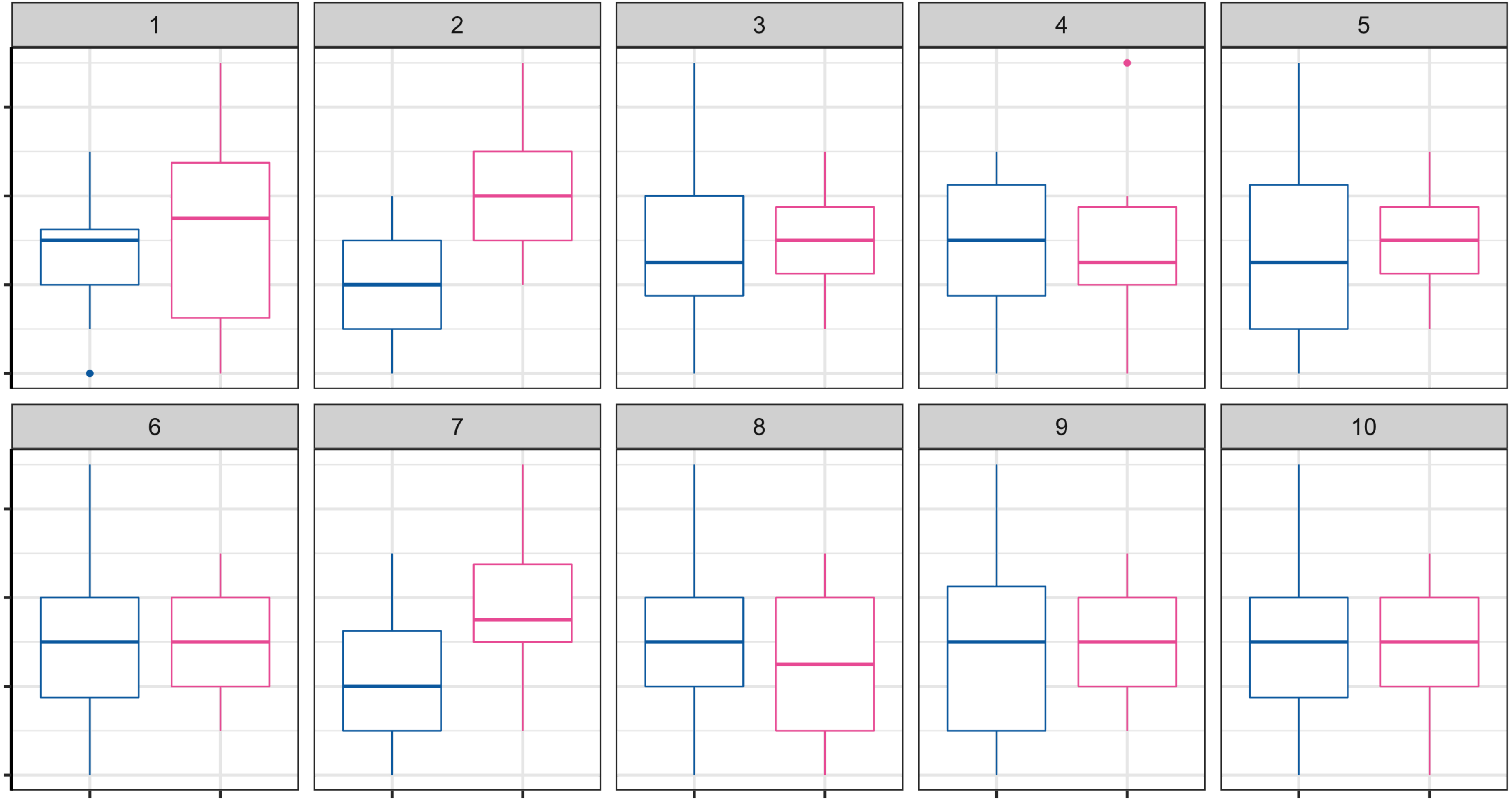
Lineup 3 In which plot is the pink group higher than the blue group?

`geom_point()`



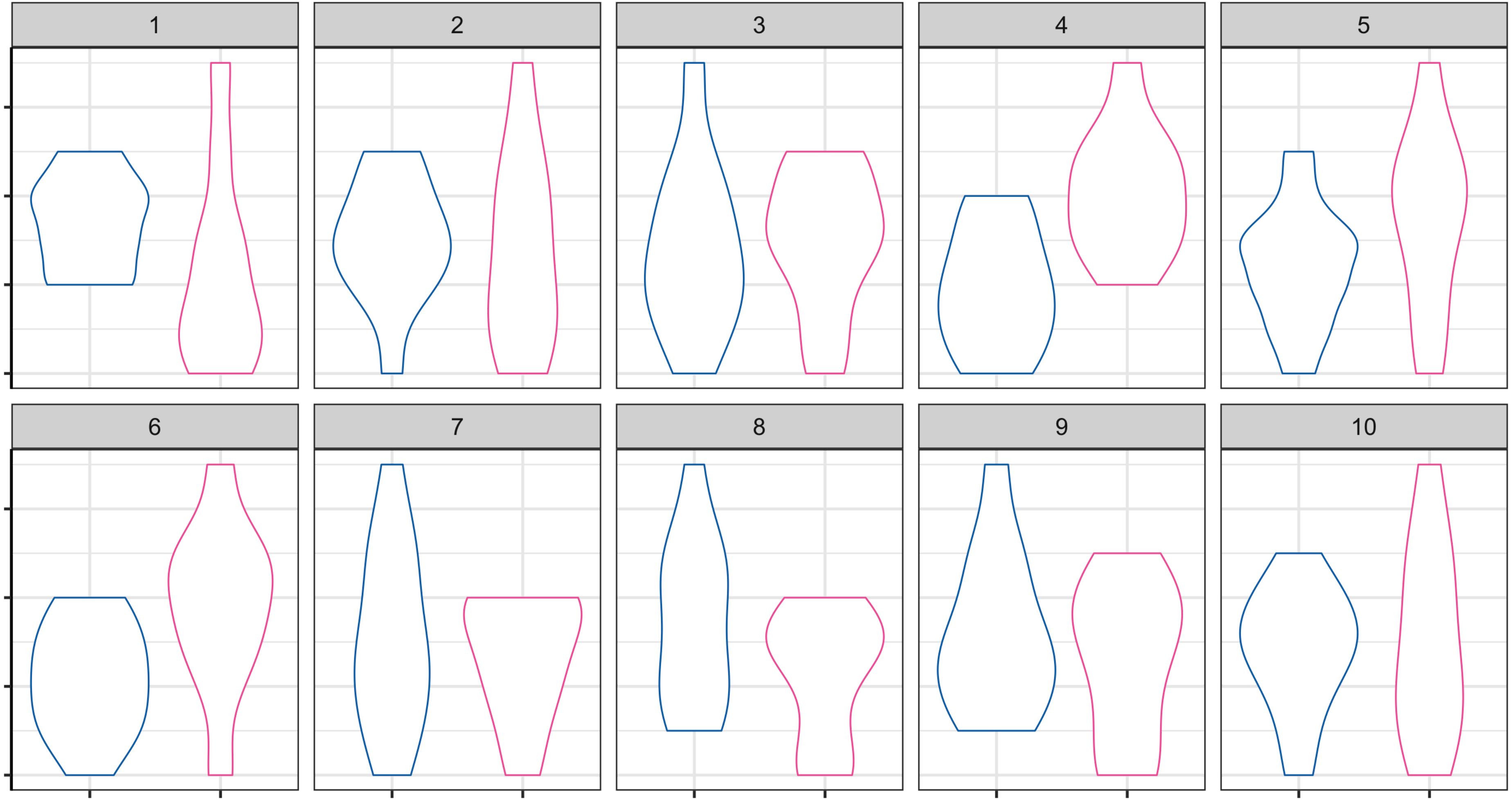
Lineup 4 In which plot is the pink group higher than the blue group?

`geom_boxplot()`



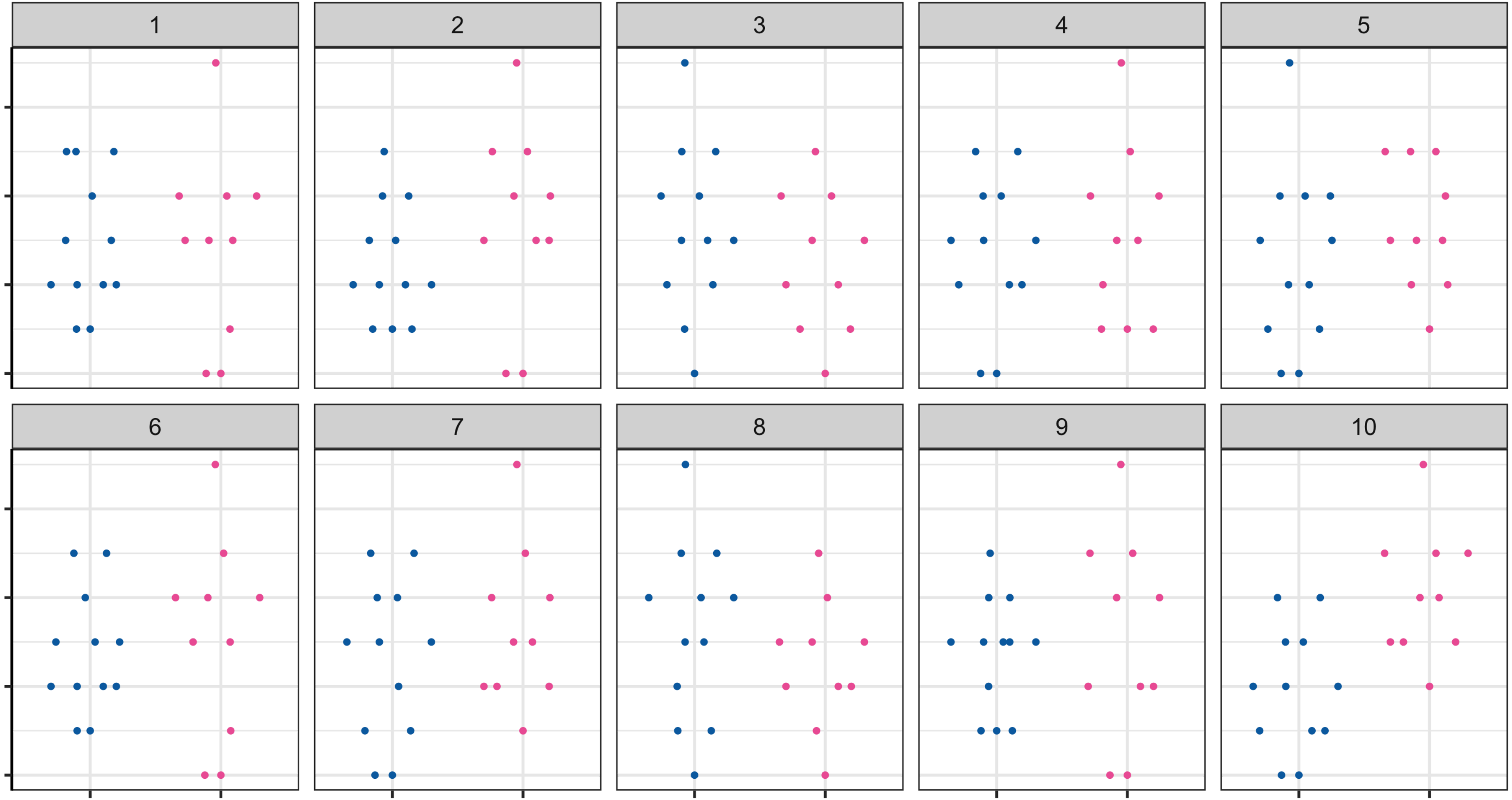
Lineup 5 In which plot is the pink group higher than the blue group?

`geom_violin()`



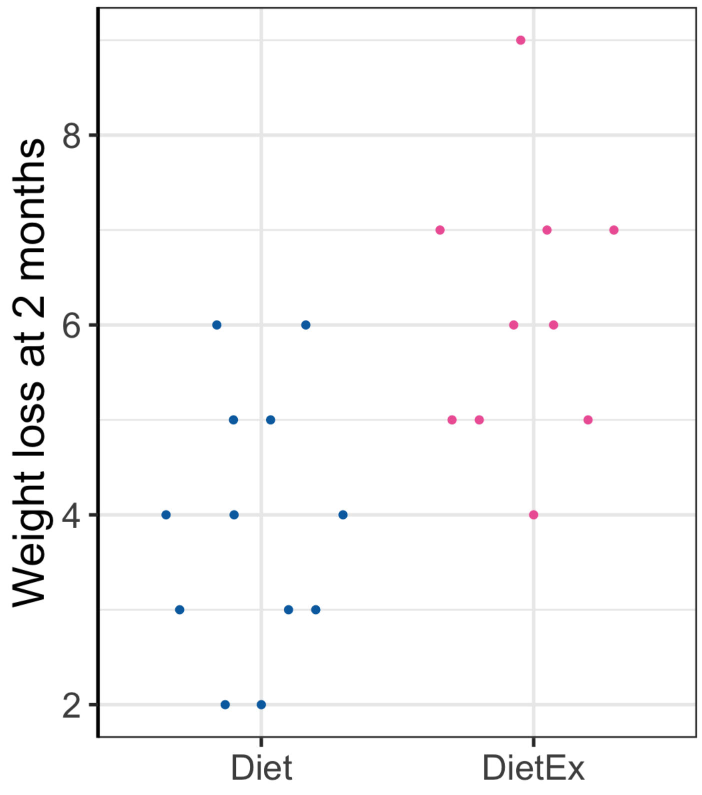
Lineup 6 In which plot is the pink group higher than the blue group?

`ggbeeswarm::geom_quasirandom()`



Case study 1 Weight loss by exercise

📊 data R



- Is weight loss greater with *exercise* after 2 months?

Group	N	Mean	Std. Dev
Diet	12	3.92	1.38
DietEx	10	6.10	1.45

```
with(df3,
     t.test(wl2[group=="DietEx"], wl2[group=="Diet"],
            alternative = "greater"))

##
##     Welch Two Sample t-test
##
## data:  wl2[group == "DietEx"] and wl2[group == "Diet"]
## t = 3.5969, df = 18.901, p-value = 0.0009675
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.133454      Inf
## sample estimates:
## mean of x mean of y
##  6.100000  3.916667
```

Power of a lineup

- The power of a lineup is calculated as x/n where x is the number of people who detected the data plot out of n people

Plot type	x	n	Power
geom_point	x_1	n_1	x_1/n_1
geom_boxplot	x_2	n_2	x_2/n_2
geom_violin	x_3	n_3	x_3/n_3
ggbeeswarm::geom_quasirandom	x_4	n_4	x_4/n_4

- The plot type with a higher power is preferable
- You can use this framework to find the optimal plot design

Hofmann, H., L. Follett, M. Majumder, and D. Cook. 2012. "Graphical Tests for Power Comparison of Competing Designs." IEEE Transactions on Visualization and Computer Graphics 18 (12): 2441–48.

Some considerations in visual inference

- In practice you don't want to bias the judgement of the human viewers so for a proper visual inference:
  - you should *not* show the data plot before the lineup
  - you should *not* give the context of the data
  - you should remove labels in plots
- You can crowd source these by paying for services like:
  - Amazon Mechanical Turk,
  - Appen (formerly Figure Eight) and
  - LABVANCED.
- If the data is for research purposes, then you may need ethics approval for publication.

Resources and Acknowledgement

- Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906): 4361–83.
- Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. "Graphical Inference for Infovis." IEEE Transactions on Visualization and Computer Graphics 16 (6): 973–79.
- Hofmann, H., L. Follett, M. Majumder, and D. Cook. 2012. "Graphical Tests for Power Comparison of Competing Designs." IEEE Transactions on Visualization and Computer Graphics 18 (12): 2441–48.
- Majumder, M., Heiki Hofmann, and Dianne Cook. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." Journal of the American Statistical Association 108 (503): 942–56.

- Data coding using `tidyverse` suite of R packages
- Slides constructed with `xaringan`, remark.js, `knitr`, and R Markdown.

Lecturer: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 11 - Session 1