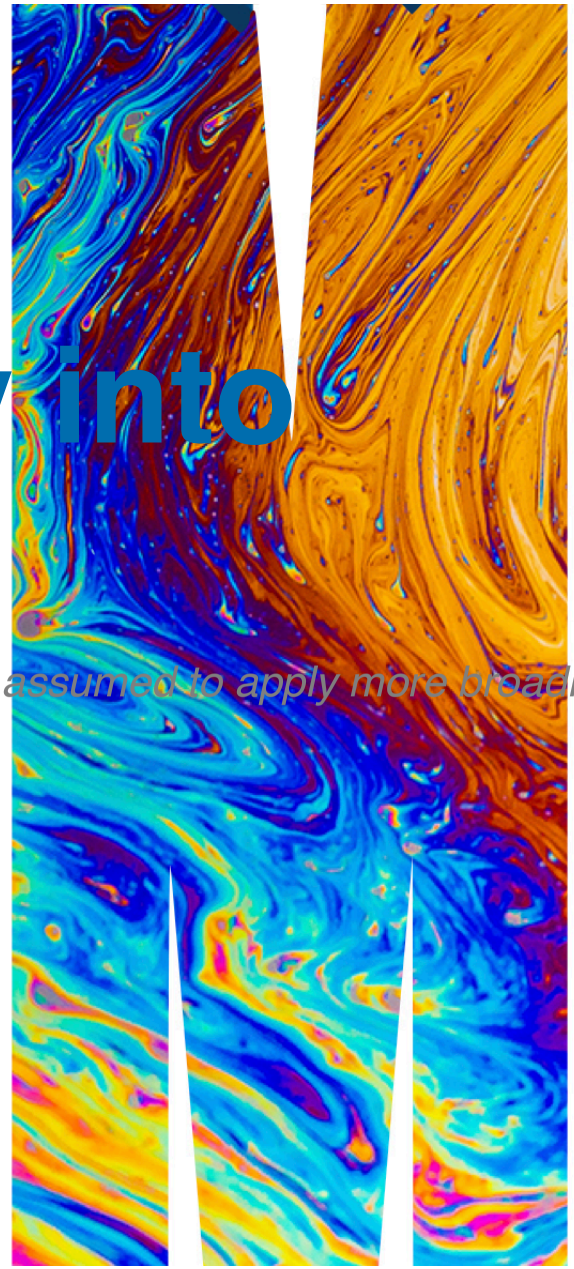
MONASH University

ETC5521: Diving Deeply into Data Exploration

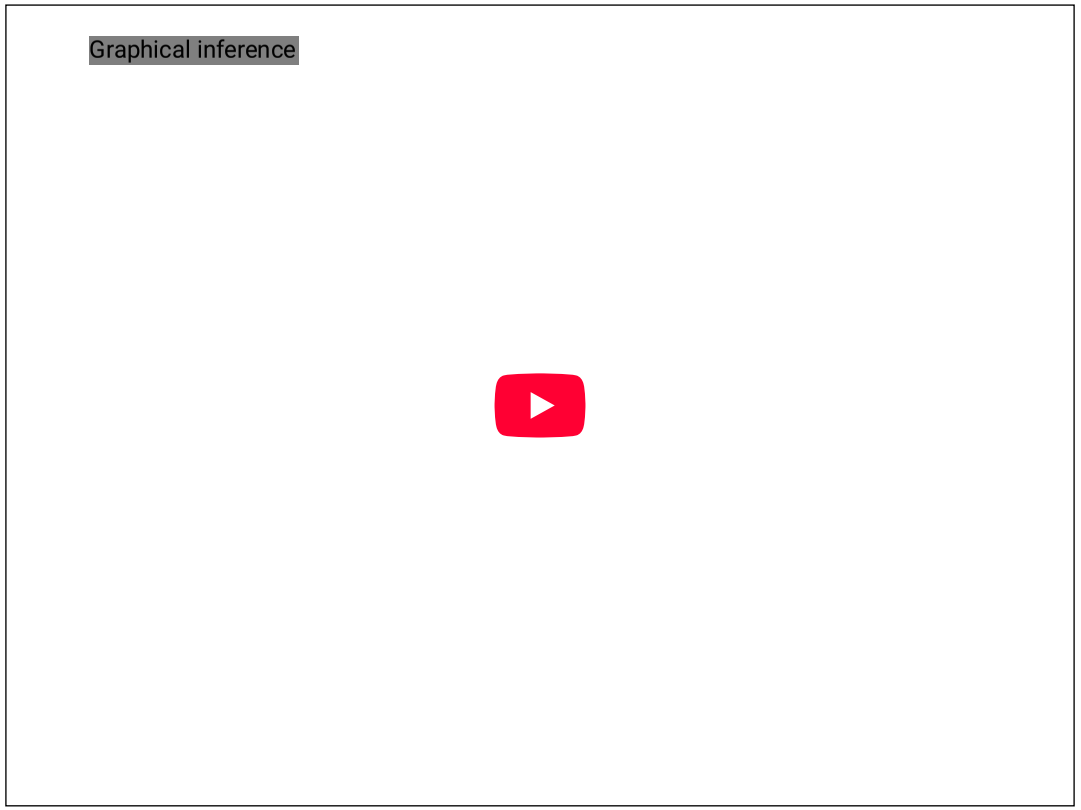
Using computational tools to determine whether what is seen in the data can be assumed to apply more broadly



Professor Di Cook

Department of Econometrics and Business Statistics

What this class is about



Revisiting hypothesis testing

(Frequentist) hypothesis testing framework

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

Hypotheses

$H_0 : p = 0.5$ vs. $H_a : p > 0.5$. Note $p_0 = 0.5$.
Alternative H_a is saying we believe that the coin is biased to heads.

NOTE: Alternative needs to be decided before seeing data.

Assumptions Each toss is independent with equal chance of getting a head.

Test statistic

$X \sim B(n, p)$. Recall $E(X \mid H_0) = np_0$.
We observe $n, x, \hat{p}$. Test statistic is $\hat{p} - p_0$.

P-value

(or critical value or confidence interval) $P(X \geq x \mid H_0)$

Conclusion Reject null hypothesis when the p-value is less than some significance level α. Usually $\alpha = 0.05$.

Testing coin bias (1/4)

- 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.

Testing coin bias (2/4)

- **Experiment 1**: I flipped the coin 10 times and this is the result:

```
set.seed(924)
samp10 <- sample(rep(c(head, tail), c(7, 3)))
cat(paste0(samp10, collapse = ""))
```

- 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 (3/4)

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

```
samp100 <- sample(rep(c(head, tail), c(70, 30)))
cat(paste0(samp100, collapse = ""))
```

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

Testing coin bias (4/4)

Calculate it

Experiment 1 (n=10)

- We observed $x = 7$, or $\hat{p} = 0.7$.
- Assuming H_0 is true, we expect $np = 10 \times 0.5 = 5$.
- Calculate the $P(X \geq 7)$

```
1 sum(dbinom(7:10, 10, 0.5))
```

```
[1] 0.17
```

Experiment 2 (n=100)

- We observed $x = 70$, or $\hat{p} = 0.7$.
- Assuming H_0 is true, we expect $np = 100 \times 0.5 = 50$.
- Calculate the $P(X \geq 70)$

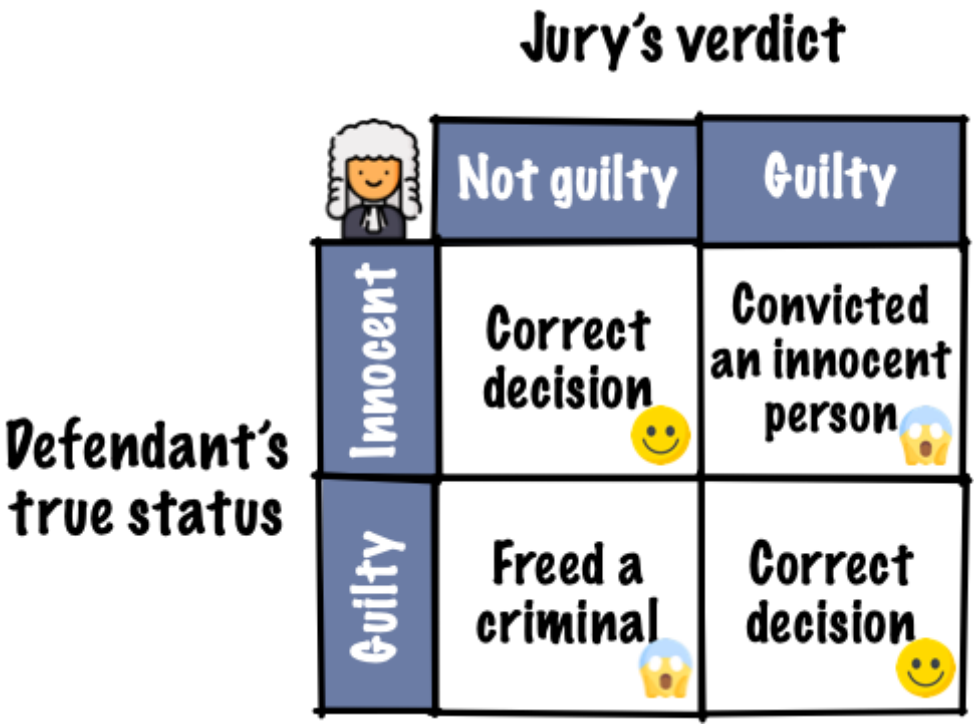
```
1 sum(dbinom(70:100, 100, 0.5))
```

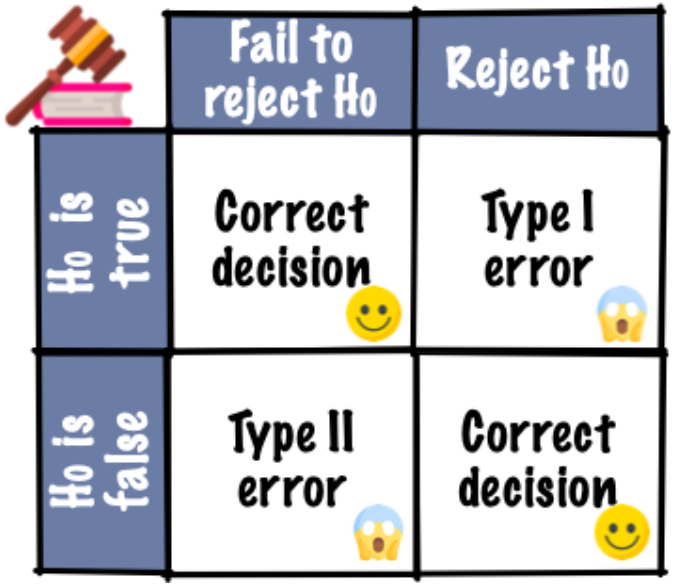
```
[1] 3.9e-05
```

Why is the null hypothesis always specific?

You need to be able to calculate the probability of something happening, if the null was true.

Judicial system





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

Does the test statistic have to be *numerical*?

Visual inference

Visual inference

- Hypothesis testing in a visual inference framework is where:
  - the *test statistic is a plot* and
  - judgement is by human visual perception.

Why is the plot a test statistic? We'll see why soon.

- *You, we, me* actually do visual inference many times but generally in an *informal* fashion.
- The problem with doing this is we are making an inference on whether the plot has any patterns based on *a single data plot*.
- The single data plot needs to be examined in the context of *what might this look like if different samples were shown.*

Reasons to use visual inference

- Data plots tend to be over-interpreted.
- Reading data plots requires 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.

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

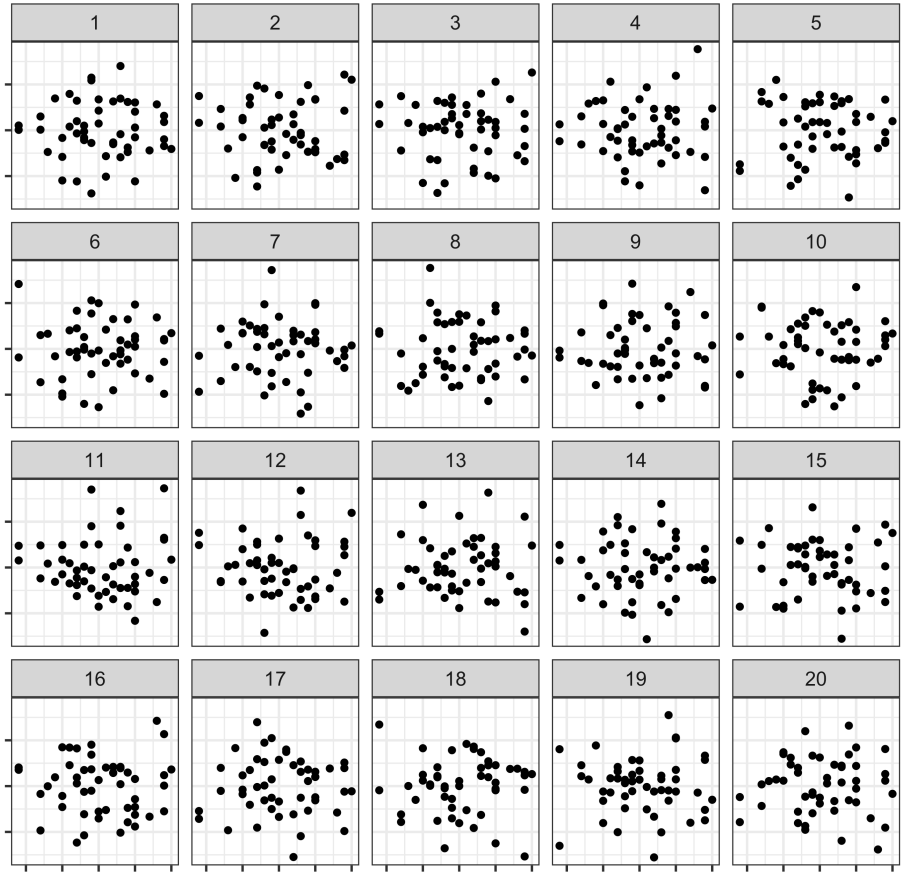
$$P(X \geq x)$$

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

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

Two residual plots examples seen last week

Lineup: Which plot has a pattern that is different from other plots?



Residuals from `dist~speed` using `datasets::cars` (week 3).

```
1  lm(dist ~ speed, data = cars)
```

- This is a lineup of the residual plot
- Which plot (if any) looks different from the others?
- Why do you think it looks different?

```
> decrypt("clZx bKhK oL 3OHohoOL 0B")
[1] "True data in position  11"
```

How do we calculate statistical significance from this?

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

- So x out of n people 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

```
1 1 - pbinom(x - 1, n, 1/20)
2 # OR
3 nullabor::pvisual(x, n, 20)
```

- The calculation is made with the assumption that the chance of a single observer randomly chooses the true plot is 1/20.

Suppose $x = 2$ out of $n = 16$ people chose plot 11 (previous slide).

The probability that this happens by random guessing (p-value) is

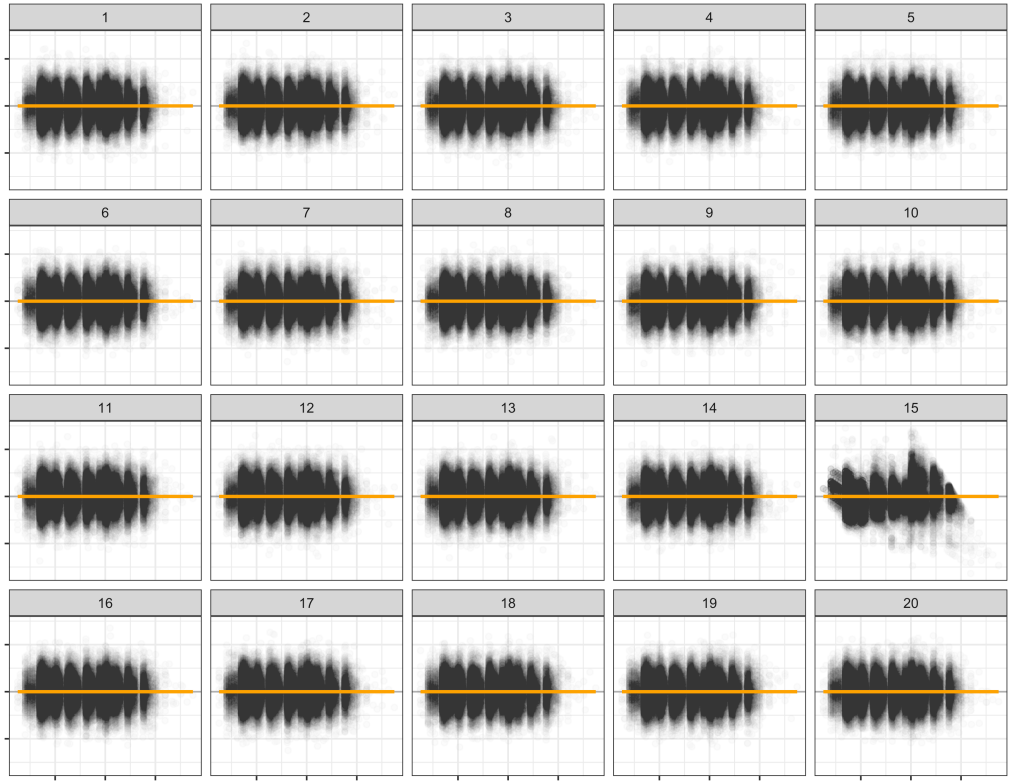
```
1 1 - pbinom(2 - 1, 16, 1/20)
```

```
[1] 0.19
```

```
1 nullabor::pvisual(2, 16, 20)
```

```
     x simulated binom
[1,] 2       0.2  0.19
```

Lineup: Which plot has a pattern that is different from other plots?



Residuals from log-transformed `price~carat ggplot2::diamonds` (week 3).

```
1 d_fit <- lm(lprice ~ lcarat, data=diamonds)
```

- This is a lineup of the residual plot for the model where both carat and price are log-transformed
- Which plot (if any) looks different from the others?
- Why do you think it looks different?

```
> decrypt("clZx bKhK oL 3OHohoOL 0Q")
[1] "True data in position  15"
```

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

Suppose $x = 8$ out of $n = 12$ people chose plot 15 (previous slide).

The probability that this happens by random guessing (p-value) is

```
1 1 - pbinom(8 - 1, 12, 1/20)
```

```
[1] 1.6e-08
```

```
1 nullabor::pvisual(8, 12, 20)
```

```
     x simulated   binom
[1,] 8         0 1.6e-08
```

This is basically impossible to happen by chance.

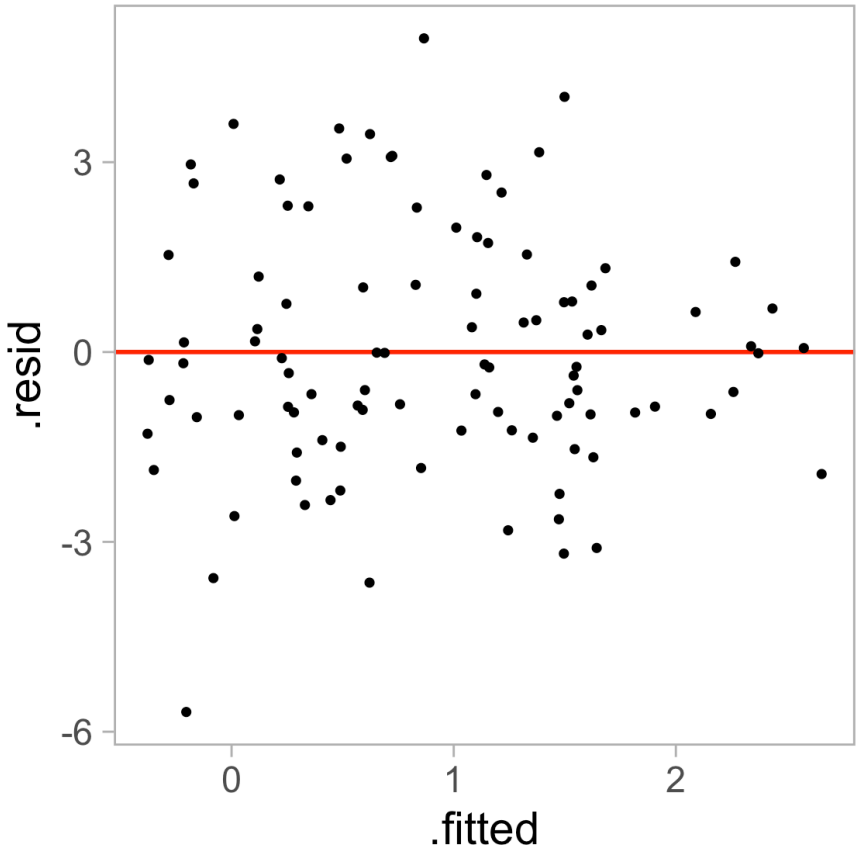
Next, how the residuals are different from "good" residuals has to be determined by the follow-up question: how did you decide your chosen plot was different?

Plot 15 has a different variance pattern, it's not the regular up-down pattern seen in the other plots. This suggests that there is some heteroskedasticity in the data that is not captured by the error distribution in the model.

New residual plot examples

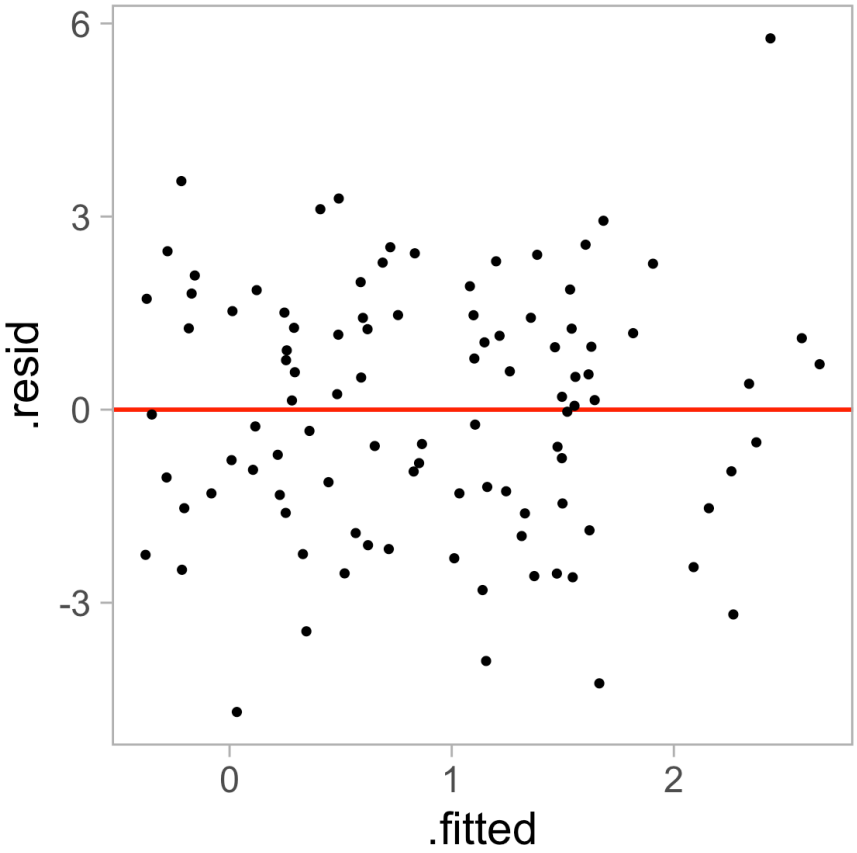
Residual plot (1/3)

Is there a problem with the model?



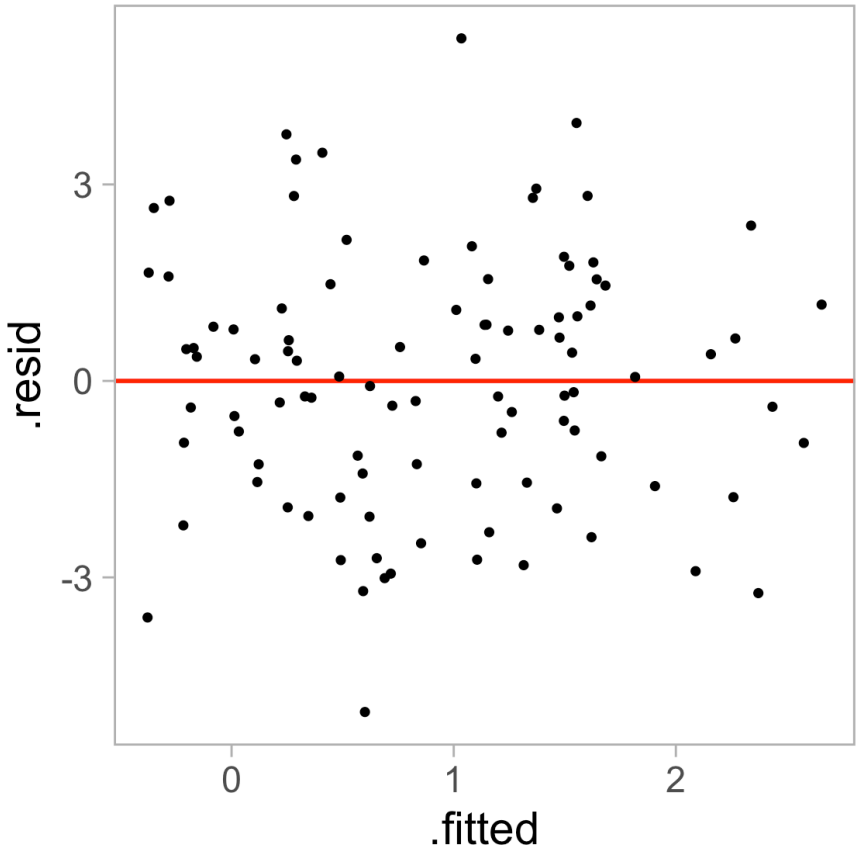
Residual plot (2/3)

Is there a problem with the model?



Residual plot (3/3)

Is there a problem with the model?

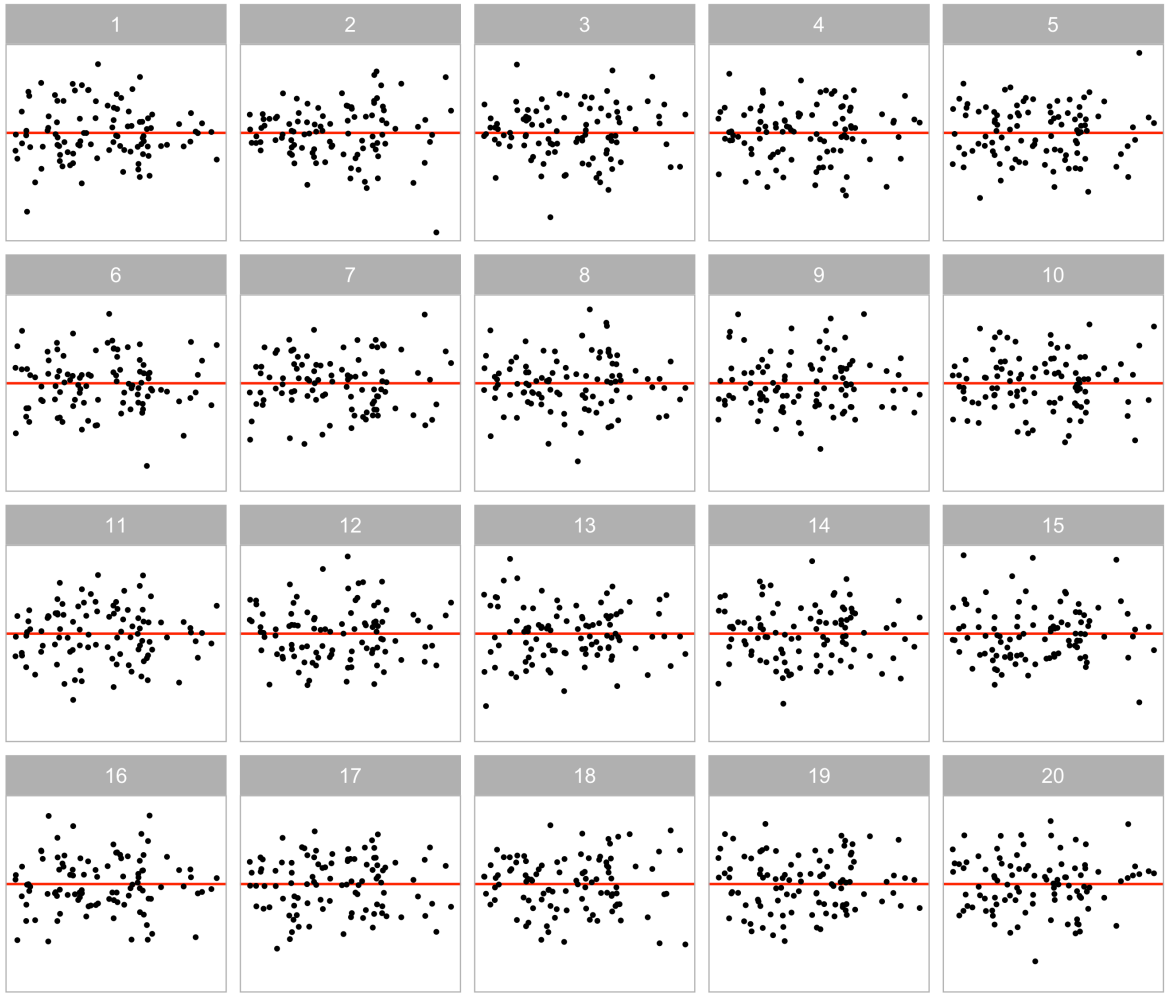


Residual plots need context

You are asked to decide IF THERE IS NO PATTERN. This is hard!

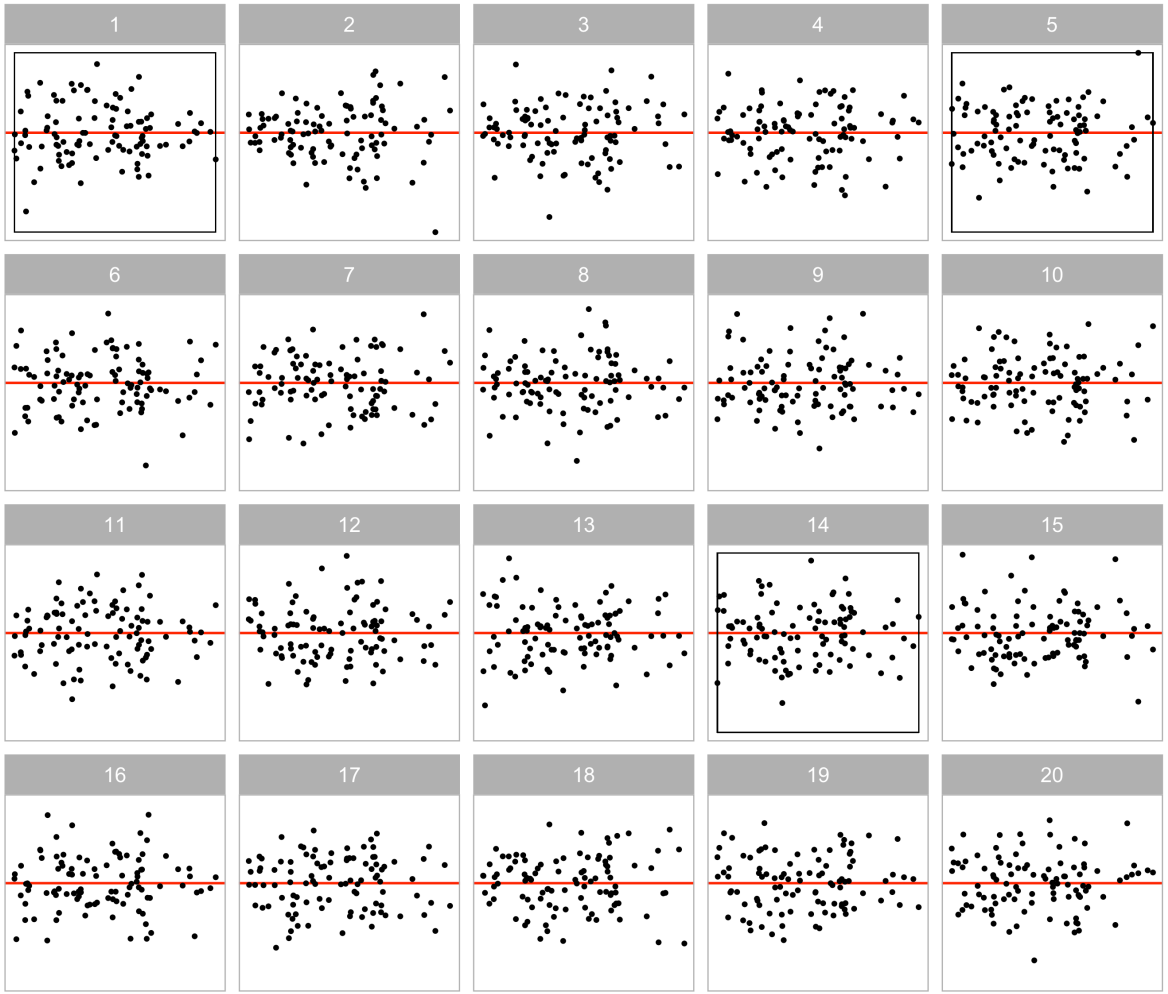
Residual plots are better when viewed in the context of good residual plots, where we know the assumptions of the model are satisfied.

Which is the worst residual plot?

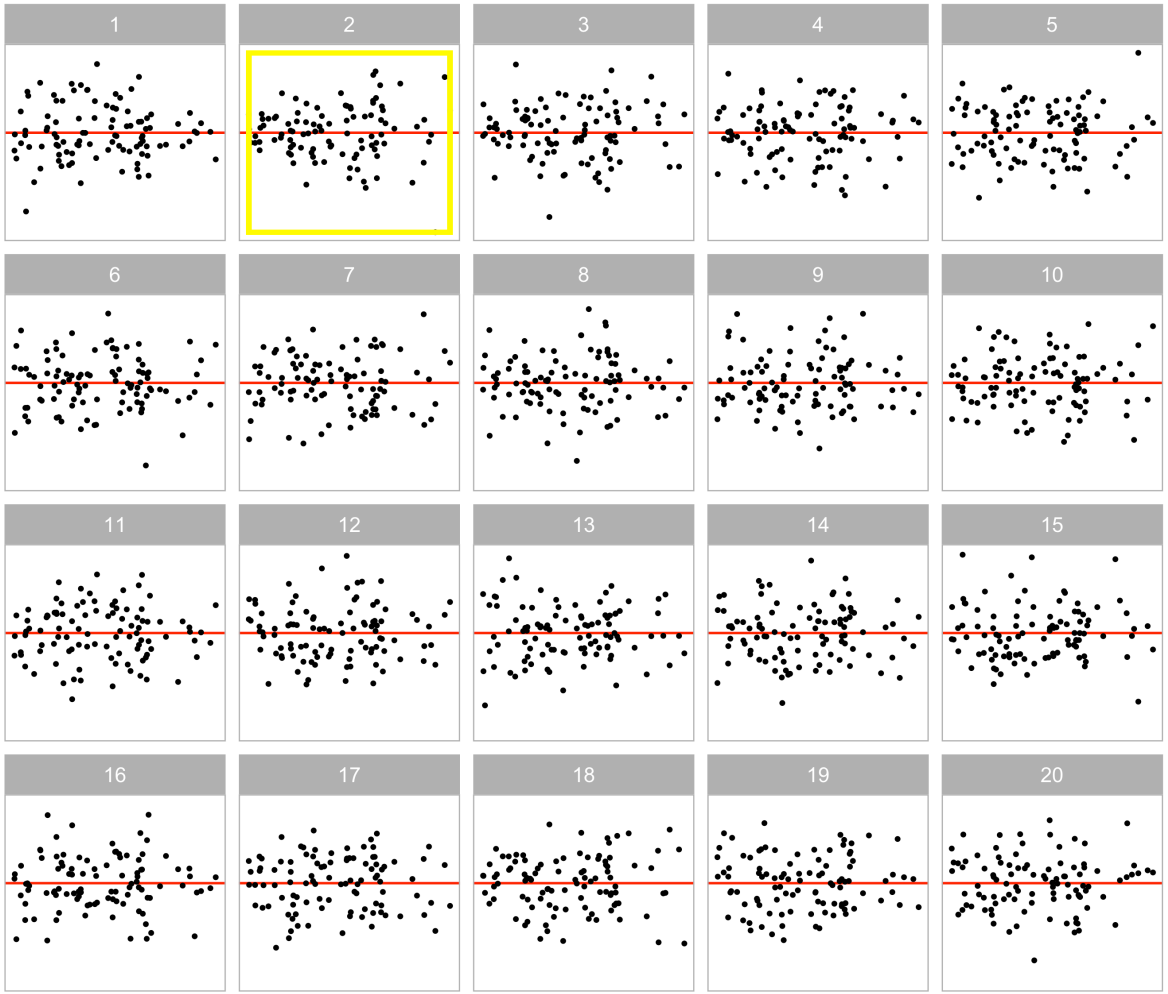


19 of these plots are good residual (null) plots.

All of the residual plots shown slides 22-24 were NULL plots.



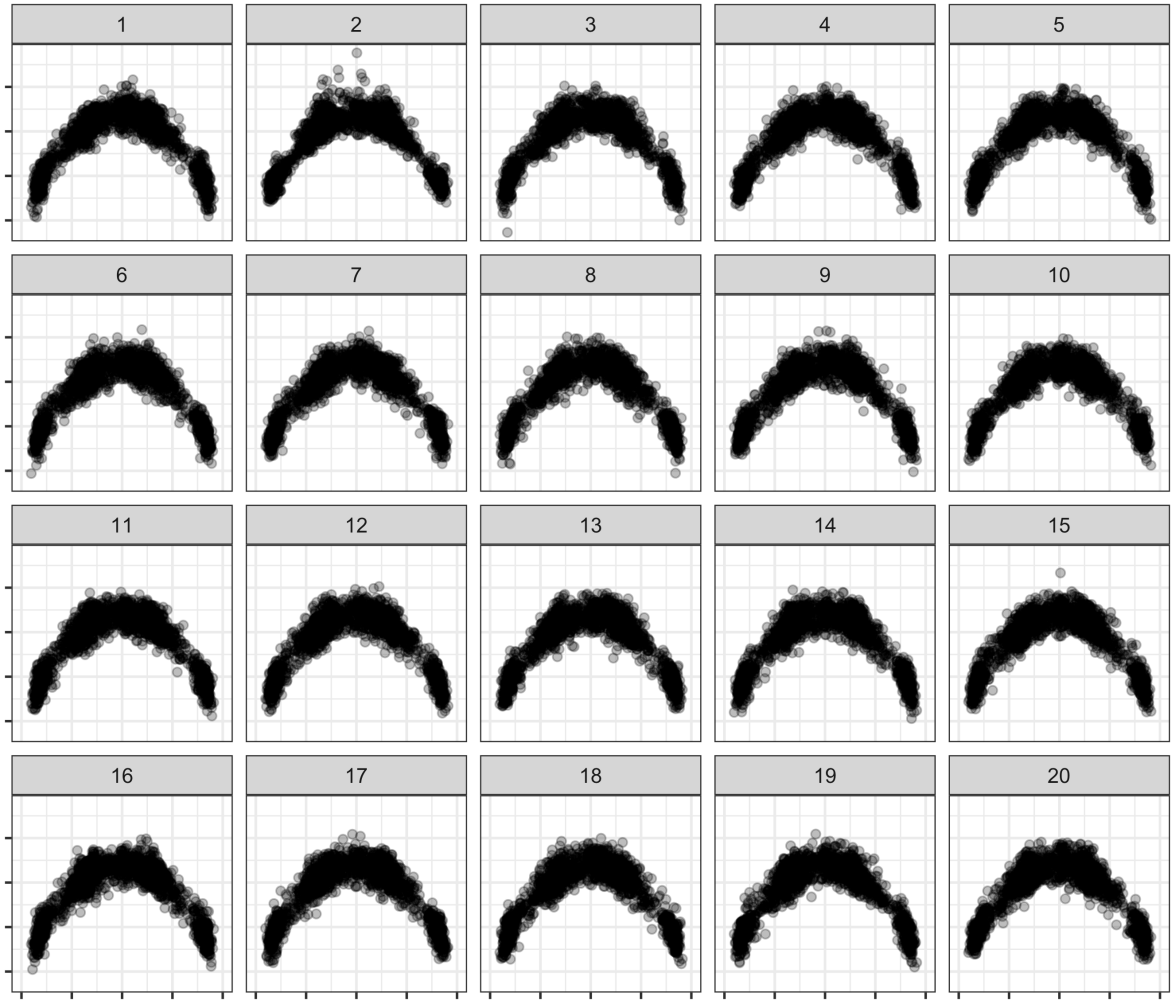
The actual residual plot is



It’s not only for residual plots

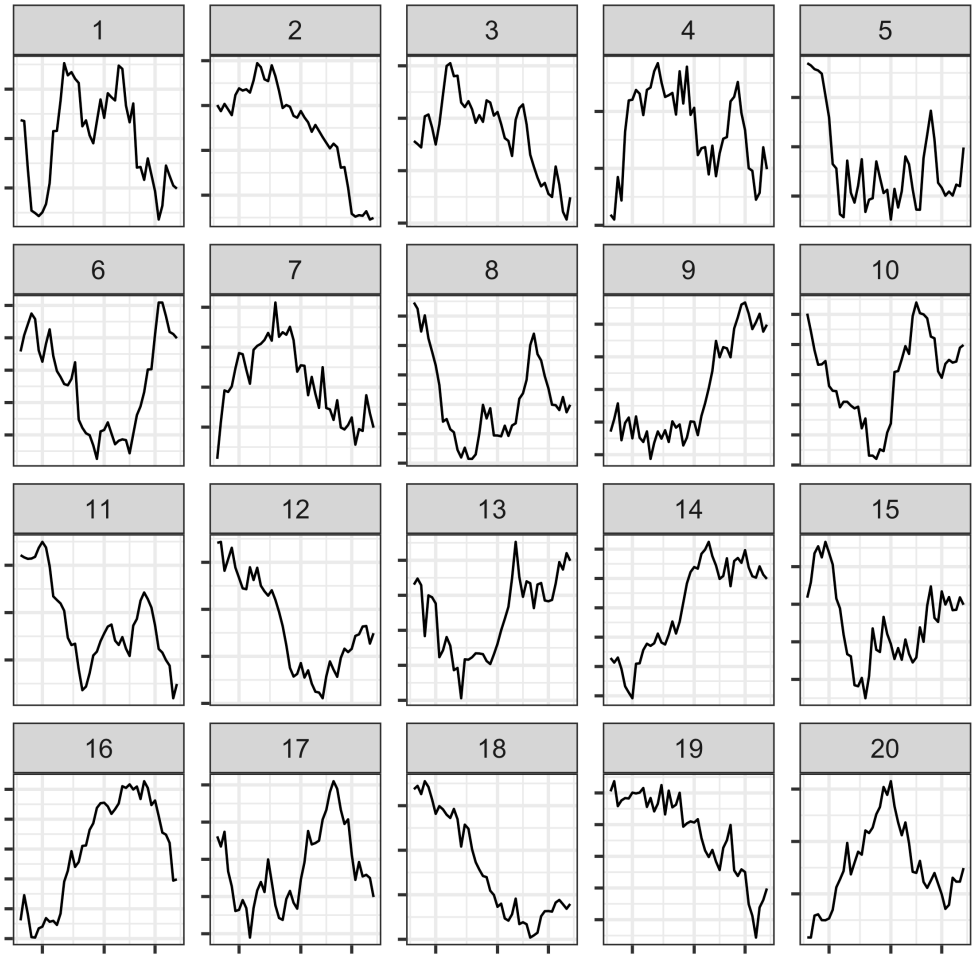
Sports analytics: basketball

Which plot is most different?



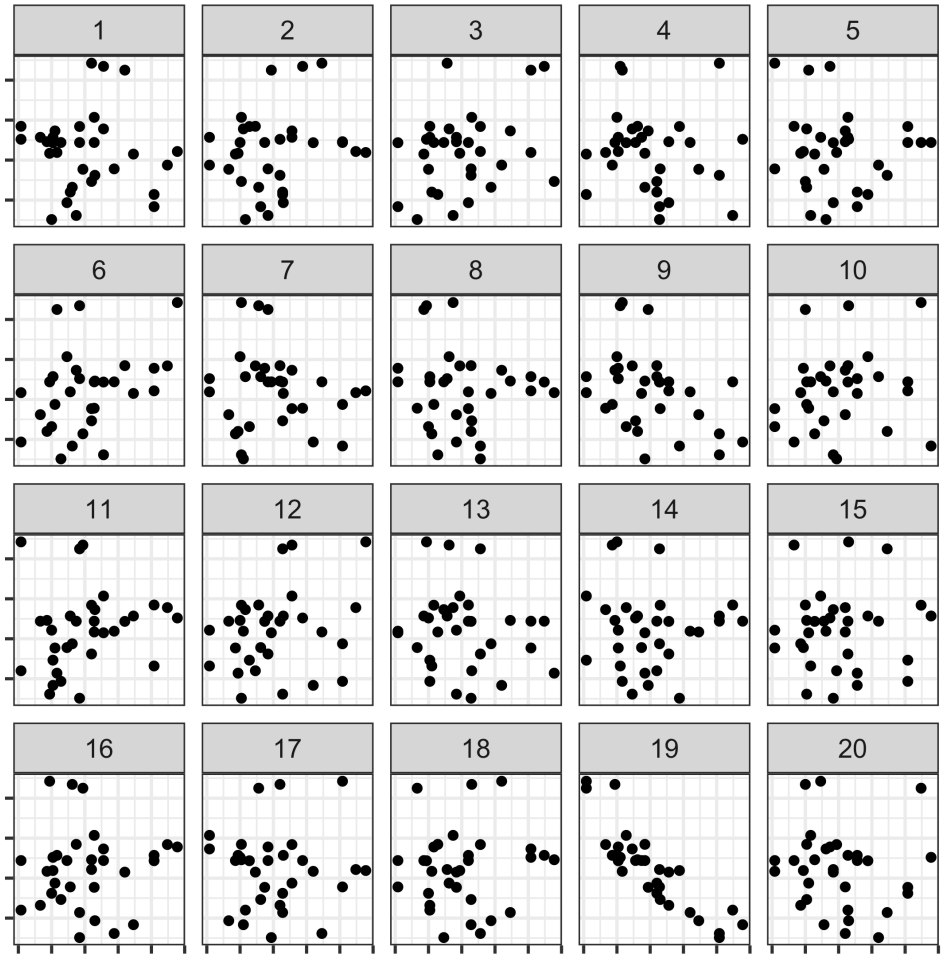
Time series: cross-currency rates

Which plot is most different?



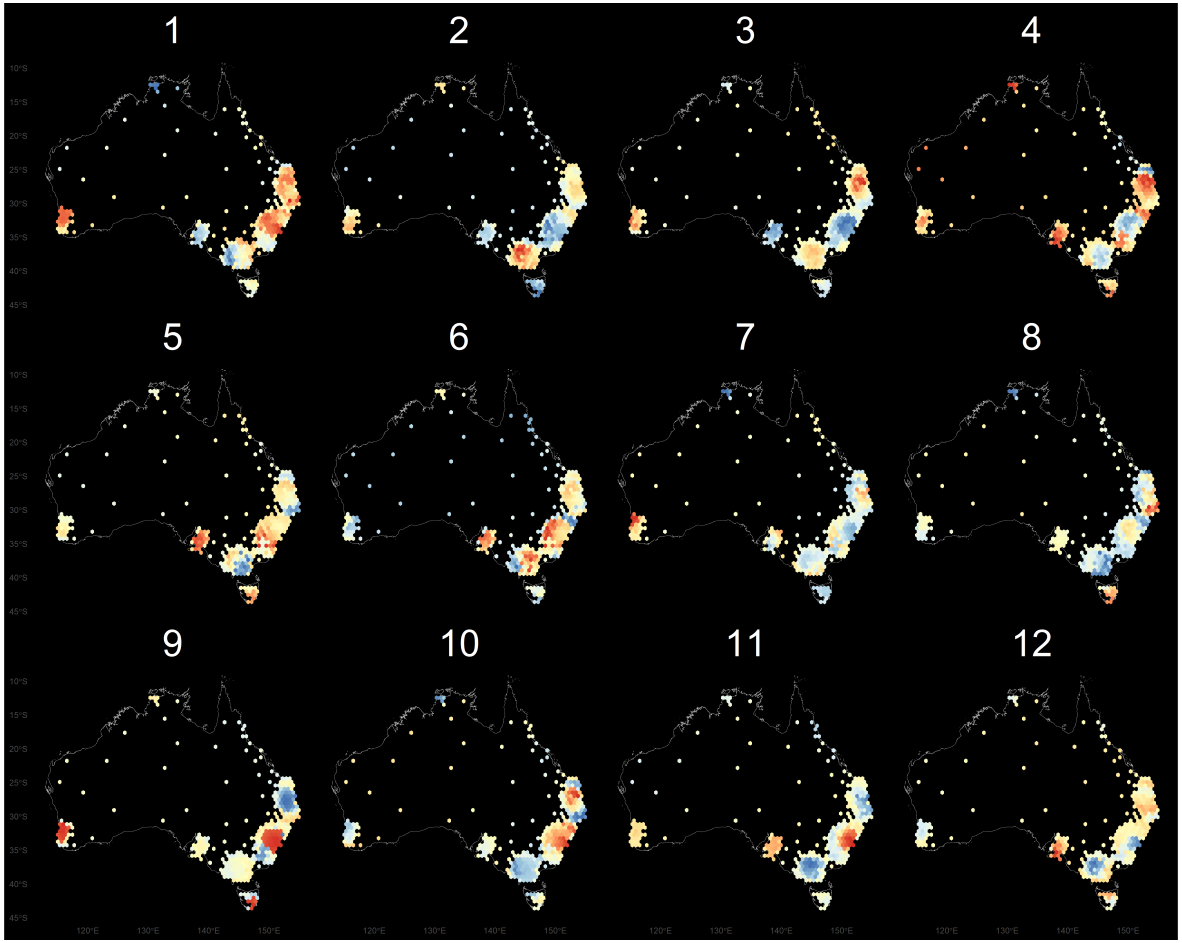
Association: cars

Which plot is most different?



Spatial analysis: cancer incidence

Which plot is most different?



From Steff Kobakian's Master's thesis

Reading any plot is easier in the context of null plots

Why is a data plot a statistic?

Why is a data plot a statistic? (1/2)

- The concept of tidy data matches elementary statistics
- Tabular form puts variables in columns and observations in rows

$$X = \begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$$
$$= \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$$

- Variables can have distributions, e.g. $X_1 \sim N(0,1),\ \ X_2 \sim \text{Exp}(1) \dots$

Why is a data plot a statistic? (2/2)

- A statistic is a function on the values of items in a sample, e.g. for n iid random variates $\bar{X}_1 = \sum_{i=1}^{n} X_{i1}$, $s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{i1} - \bar{X}_1)^2$
- We study the behaviour of the statistic over all possible samples of size n.
- The grammar of graphics is the mapping of (random) variables to graphical elements, making plots of data into statistics

Example 1:

```
ggplot(threept_sub,
       aes(x=angle, y=r)) +
  geom_point(alpha=0.3)
```

`angle` is mapped to the x axis

`r` is mapped to the y axis

Example 2:

```
ggplot(penguins,
      aes(x=bl,
          y=fl,
          colour=species)) +
  geom_point()
```

`bl` is mapped to the x axis

`fl` is mapped to the y axis

`species` is mapped to colour

Example 3:

```
ggplot(aud, aes(x=date, y=rate))
+
  geom_line()
```

`date` is mapped to the x axis

`rate` is mapped to the y axis

displayed as a line `geom`

Determining the null hypothesis

What is the null hypothesis? (1/2)

To determine the null hypothesis, you need to think about what pattern would NOT be interesting.

A

```
ggplot(data) +
  geom_point(aes(x=x1, y=x2))
```

B

```
ggplot(data) +
  geom_point(aes(x=x1,
    y=x2, colour=cl))
```

C

```
ggplot(data) +
  geom_histogram(aes(x=x1))
```

D

```
ggplot(data) +
  geom_boxplot(aes(x=cl, y=x1))
```

🤔 Which of these plot definitions would most match to a null hypothesis stating *there is no difference in the distribution between the groups?*

What is the null hypothesis? (2/2)

A

H_o : no association between `x1` and `x2`

B

H_o : no difference in association of between `x1` and `x2` between levels of `cl`

C

H_o : the distribution of `x1` is XXX

D

H_o : no difference in the distribution of `x1` between levels of `cl`

How do you generate null samples

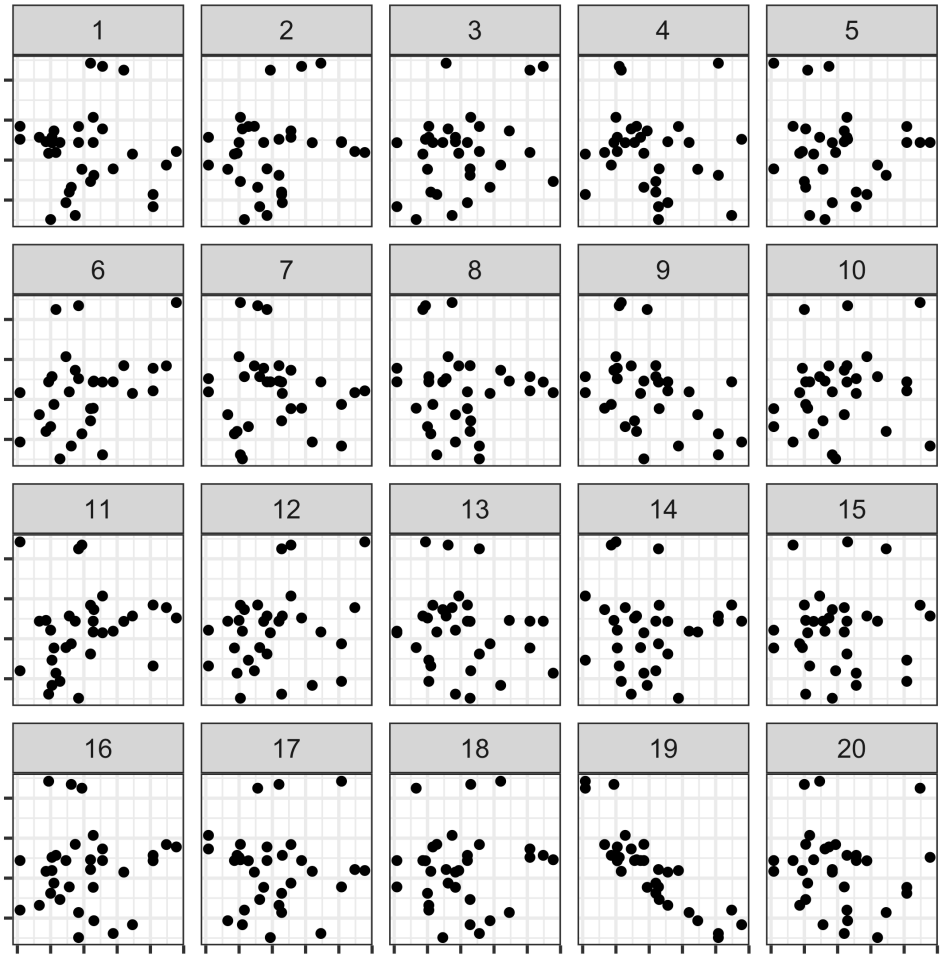
Primary null-generating mechanisms

Null samples can be generated using two basic approaches:

- Permutation: randomizing the order of one of the variables breaks association, but keeps marginal distributions the same.
- Simulation: from a given distribution, or model. Assumption is that the data comes from that model.

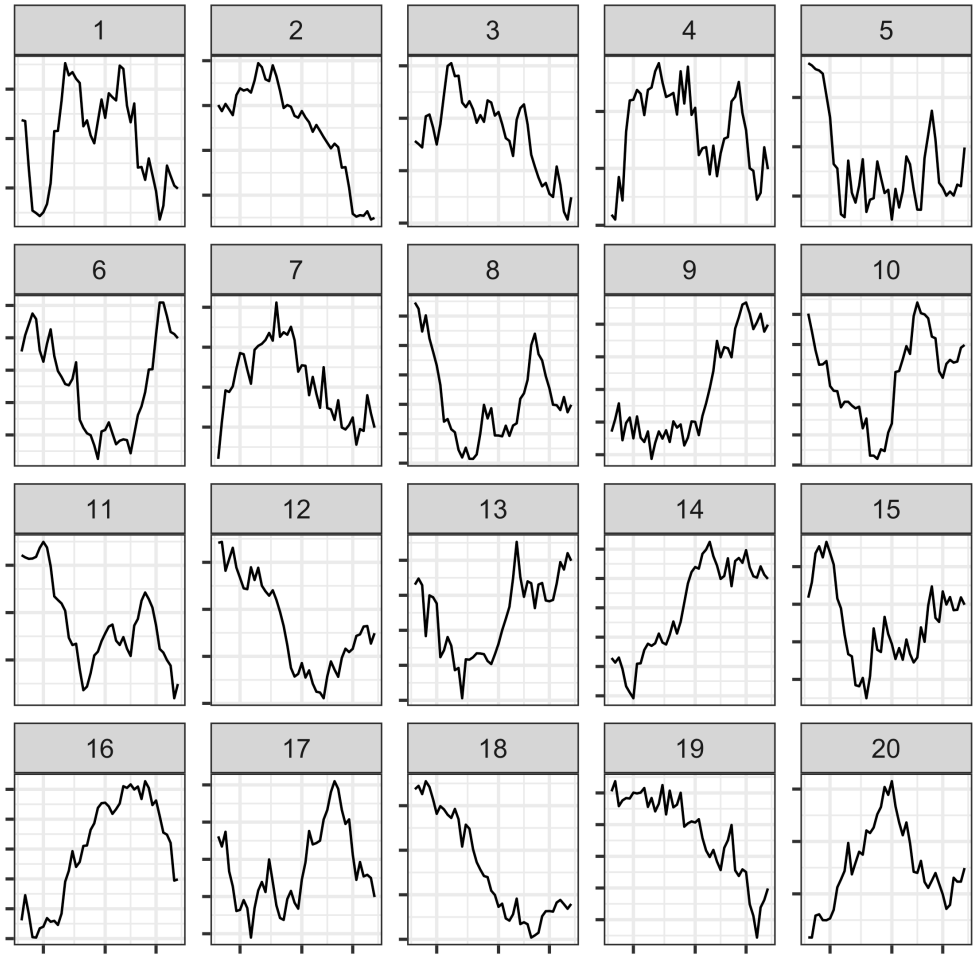
applied to subsets, or conditioning on other variables. Simulation may require computing summary statistics from the data to use as parameter estimates.

Association: cars



Null plots generated by permuting x variable.

Time series: cross-currency rates



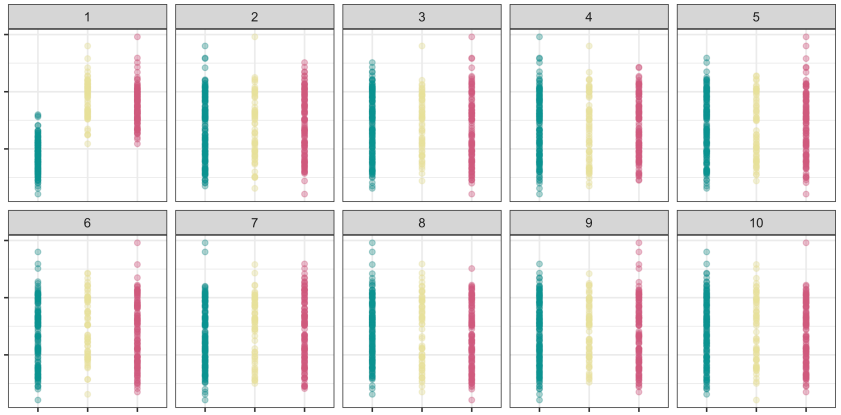
Nulls generated by simulating from an ARIMA model.

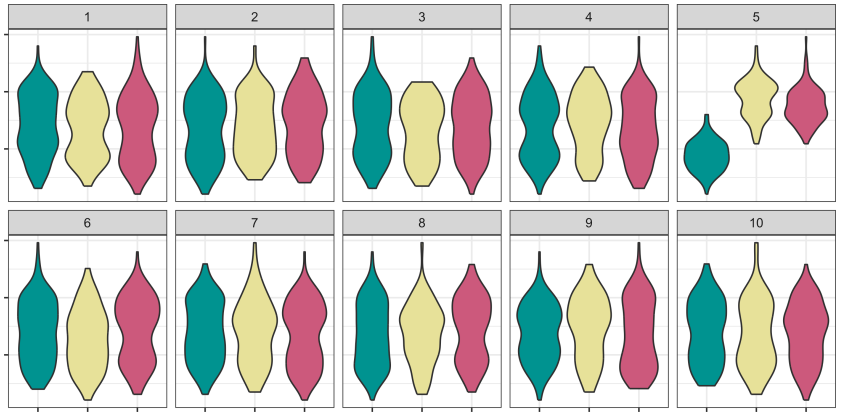
Beyond p-value to power

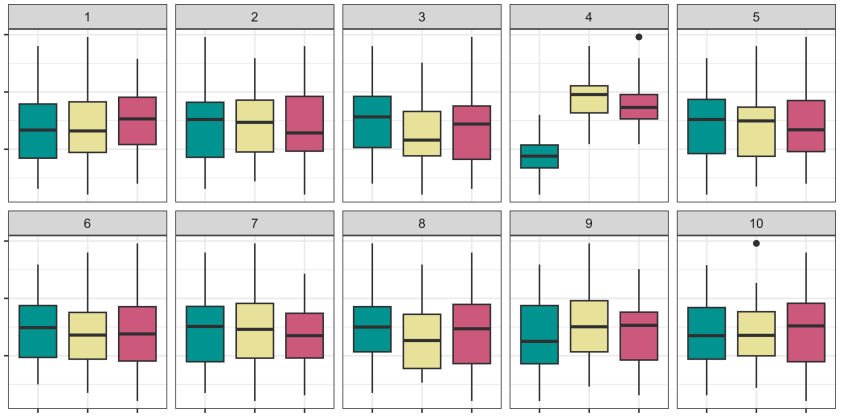
What is power?

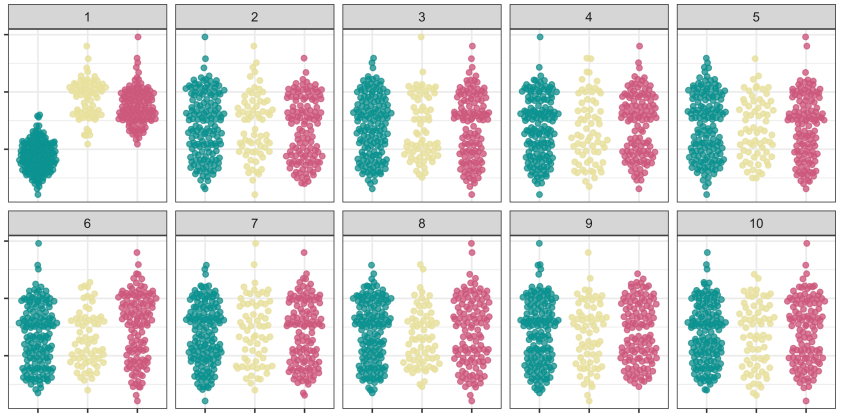
- A statistic is said to be more powerful than another statistic if it has a higher probability of correctly rejecting the null hypothesis when the alternative hypothesis is true.
- The effectiveness of two plots designs for the same data can be compared by computing power from a lineup.
- The power of a lineup is calculated as x/n where x is the number of people who detected the data plot out of n people.

Which of these plots is more effective for assessing difference between groups?


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10

Computing the power

Note: Different people evaluated each lineup.

Plot type	x	n	Power
`geom_point`	$x_1 = 4$	$n_1 = 23$	$x_1/n_1 = 0.174$
`geom_boxplot`	$x_2 = 5$	$n_2 = 25$	$x_2/n_2 = 0.185$
`geom_violin`	$x_3 = 6$	$n_3 = 29$	$x_3/n_3 = 0.206$
`ggbeeswarm::geom_quasirandom`	$x_4 = 8$	$n_4 = 24$	$x_4/n_4 = 0.333$

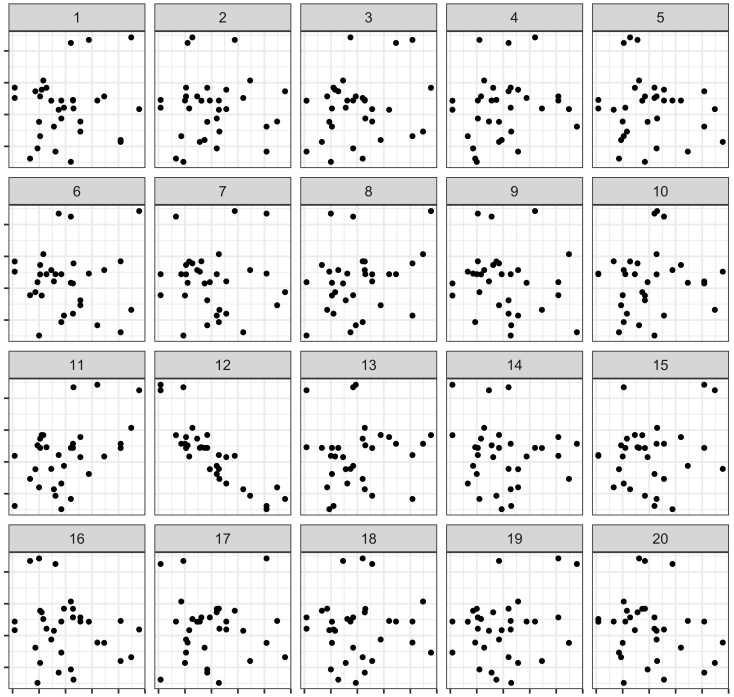
- The plot type with a higher power is preferable
- You can use this framework to find the optimal plot design

Using the nullabor 


```r
set.seed(20190709)
ggplot(lineup(null_permute('mpg'), mtcars),
  aes(x=mpg, y=wt)) +
  geom_point() +
  facet_wrap(~ .sample) +
  theme(axis.text=element_blank(),
        axis.title=element_blank())
```

When you run the example yourself, you get a `decrypt` code line, that you run after deciding on a plot to print the location of the data plot amongst the nulls.

- plot is a scatterplot, null hypothesis is *there is no association between the two variables mapped to the x, y axes*
- null generating mechanism: permutation



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 and other identifying information from plots
- These methods can be used whenever formal inference is not possible/available, for EDA or IDA or diagnosing models.
- The data collection is vital for good inference: bad data leads to bad inference.
- Determining how to generate null samples can be complicated. We'll see more examples throughout the next few weeks.

Resources

- 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.