Discussion 4

DSC 80

2024-04-26

Section 1

FA23 Midterm Problem 3

Problem

We have a DataFrame `donkeys` with 544 rows of measurements. The `WeightAlt` column has values for 30 donkeys that were **reweighed** a day later.

	id	BCS	Age	Weight	WeightAlt
0	d01	3.0	<2	77	NaN
1	d02	2.5	<2	100	NaN
2	d03	1.5	<2	74	NaN

`id`	A unique identifier for each donkey (`d01`, `d02`, etc.).
`BCS`	Body condition score: from 1 (emaciated) to 3 (healthy) to 5 (obese) in increments of 0.5.
`Age`	Age in years: <2, 2–5, 5–10, 10–15, 15–20, and over 20 years.
`Weight`	Weight in kilograms.
`WeightAlt`	Second weight measurement taken for 30 donkeys. NaN if the donkey was not reweighed.

We have to look at the following scenarios and tell if the missingess for `WeightAlt` is NMAR, MAR, or MCAR.

Refresher on Missingness Types

- **Missing By Design (MD)**: Missingness is exactly determined by other columns, for intended reasons
- **Not missing at random (NMAR)**: Missingness depends on the values *of the column that's missing*
- **Missing at random (MAR)**: Missingness depends on the values of *other* columns that are present
- **Missing completely at random (MCAR)**: Missingness in a column does not depend on the values in that column, or in other columns

If you're curious, the terminology comes from `Rubin, Donald B. "Inference and missing data." Biometrika 63.3 (1976): 581-592.`

Part A

- The researchers chose the 30 donkeys with the largest `Weight` values to reweigh.

Part A

- The researchers chose the 30 donkeys with the largest `Weight` values to reweigh.
- This is **MAR**: The missingness depends on the values in the `Weight` column.

Part B

- The researchers drew 30 donkeys uniformly at random without replacement from the donkeys with BCS score of 4 or greater.

Part B

- The researchers drew 30 donkeys uniformly at random without replacement from the donkeys with BCS score of 4 or greater.
- This is **MAR**: The missingness depends on the values in the BCS column. The pattern of that dependence is different, but it's still MAR.

Part C

- The researchers set `i` as a number drawn uniformly at random between 0 and 514, then reweighed the donkeys in `donkeys.iloc[i:i+30]`.

Part C

- The researchers set `i` as a number drawn uniformly at random between 0 and 514, then reweighed the donkeys in `donkeys.iloc[i:i+30]`.
- This one... could go either way. The intended answer was **MCAR** because this process means we're selecting a random length-30 slice of the data to reweigh.

Part C

- The researchers set `i` as a number drawn uniformly at random between 0 and 514, then reweighed the donkeys in `donkeys.iloc[i:i+30]`.
- This one... could go either way. The intended answer was **MCAR** because this process means we're selecting a random length-30 slice of the data to reweigh.
- But you might notice that with this procedure, not every value has an equal likelihood of being in that slice, so you could also say it's **MAR** based on index/position.

Part D

- The researchers reweighed all the donkeys, but deleted all the values in `WeightAlt` except for the 30 lowest values.

Part D

- The researchers reweighed all the donkeys, but deleted all the values in `WeightAlt` except for the 30 lowest values.
- This is **NMAR**: The missingness depends on the values in the column that is missing!

Part E

- The researchers split up the donkeys into the 6 different age groups, then sampled 5 donkeys uniformly at random without replacement within each age group.

Part E

- The researchers split up the donkeys into the 6 different age groups, then sampled 5 donkeys uniformly at random without replacement within each age group.
- This one... also depends on some assumptions about the data.

Part E

- The researchers split up the donkeys into the 6 different age groups, then sampled 5 donkeys uniformly at random without replacement within each age group.
- This one... also depends on some assumptions about the data.
- If the age groups are not evenly distributed, then this is definitely **MAR**: the missingness will be more likely for some values than others, depending on age.

Part E

- The researchers split up the donkeys into the 6 different age groups, then sampled 5 donkeys uniformly at random without replacement within each age group.
- This one... also depends on some assumptions about the data.
- If the age groups are not evenly distributed, then this is definitely **MAR**: the missingness will be more likely for some values than others, depending on age.
- If the age groups are evenly distributed, then every value has the same probability of being chosen, so it's **MCAR**.

Section 2

Scatterplots (1)

Scatterplots (1)

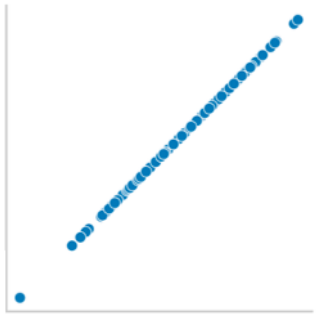
For this next question, assume that the researchers chose the 30 donkeys to reweigh by drawing a simple random sample of 30 underweight donkeys: donkeys with `BCS` values of 1, 1.5, or 2. The researchers weighed these 30 donkeys one day later and stored the results in `WeightAlt`.

First: **what kind of missingness is this?**

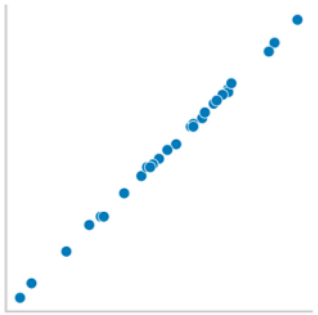
Problem

Which of the following shows the scatter plot of `WeightAlt - Weight` on the y-axis and `'Weight'` on the x-axis? Assume that missing values are not plotted.

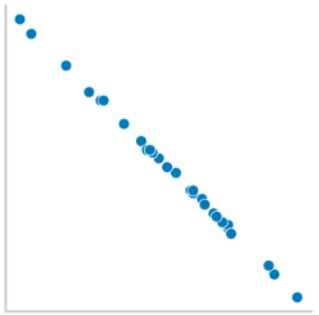
A)



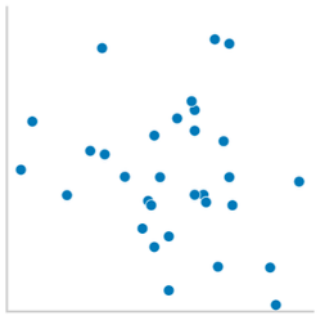
B)



C)



D)



Solution

- Note that this is the scatter plot of `WeightAlt - Weight`, not just `WeightAlt`!

Solution

- Note that this is the scatter plot of `WeightAlt - Weight`, not just `WeightAlt`!
- Can we say anything from the problem description about the relationship between `WeightAlt - Weight`?

Solution

- Note that this is the scatter plot of `WeightAlt - Weight`, not just `WeightAlt`!
- Can we say anything from the problem description about the relationship between `WeightAlt - Weight`?
- The answer is, pretty much, no – for available values, `WeightAlt` is just `Weight` but measured on the next day. So the difference between the two should be pretty much random, and the answer is D.

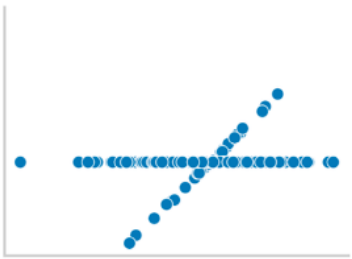
Section 3

Scatterplots (2)

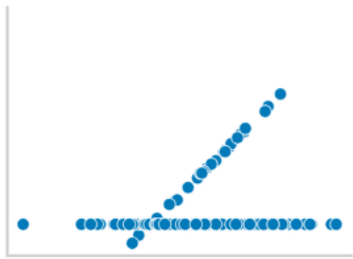
Scatterplots (2)

Suppose we use mean imputation to fill in the missing values in `WeightAlt`. Select the scatter plot `WeightAlt` on `Weight` after imputation.

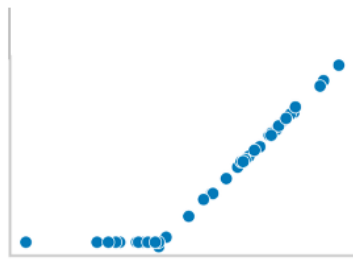
A)



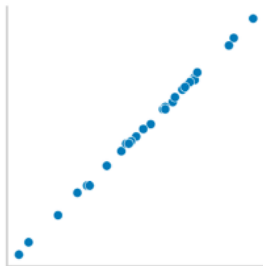
B)



C)



D)



Solution

- First: what is mean imputation?

Solution

- First: what is mean imputation?
- So, what should a scatterplot for a mean-imputed column of data look like?

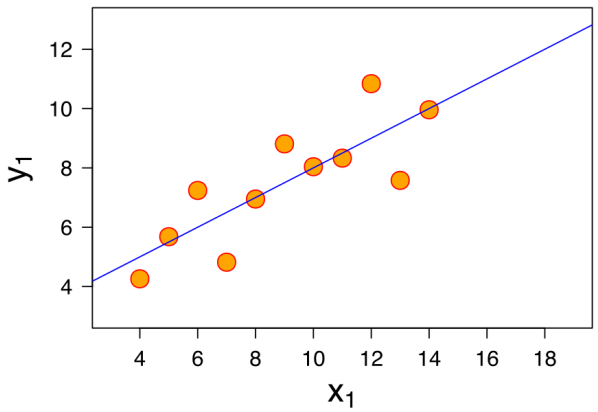
Solution

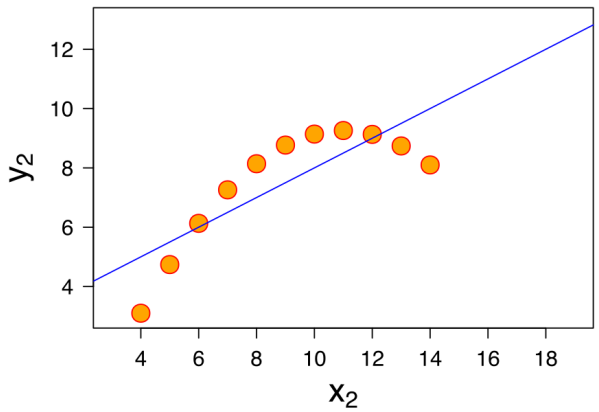
- First: what is mean imputation?
- So, what should a scatterplot for a mean-imputed column of data look like?
- The answer is A: A noisy line of values that were present in the dataset, and then a bunch of values that, no matter the `Weight` value, now all have the value of the mean.

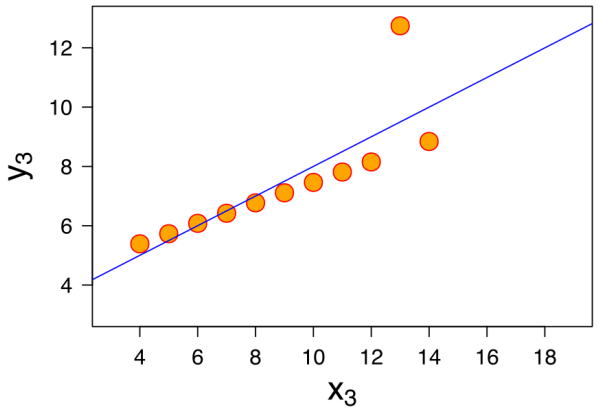
Section 4

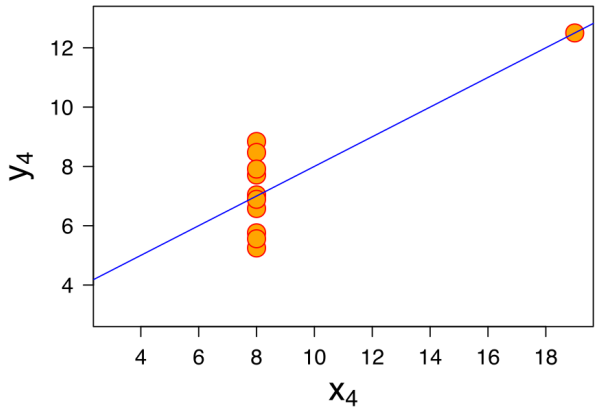
Choosing Test Statistics

Anscombe's Quartet





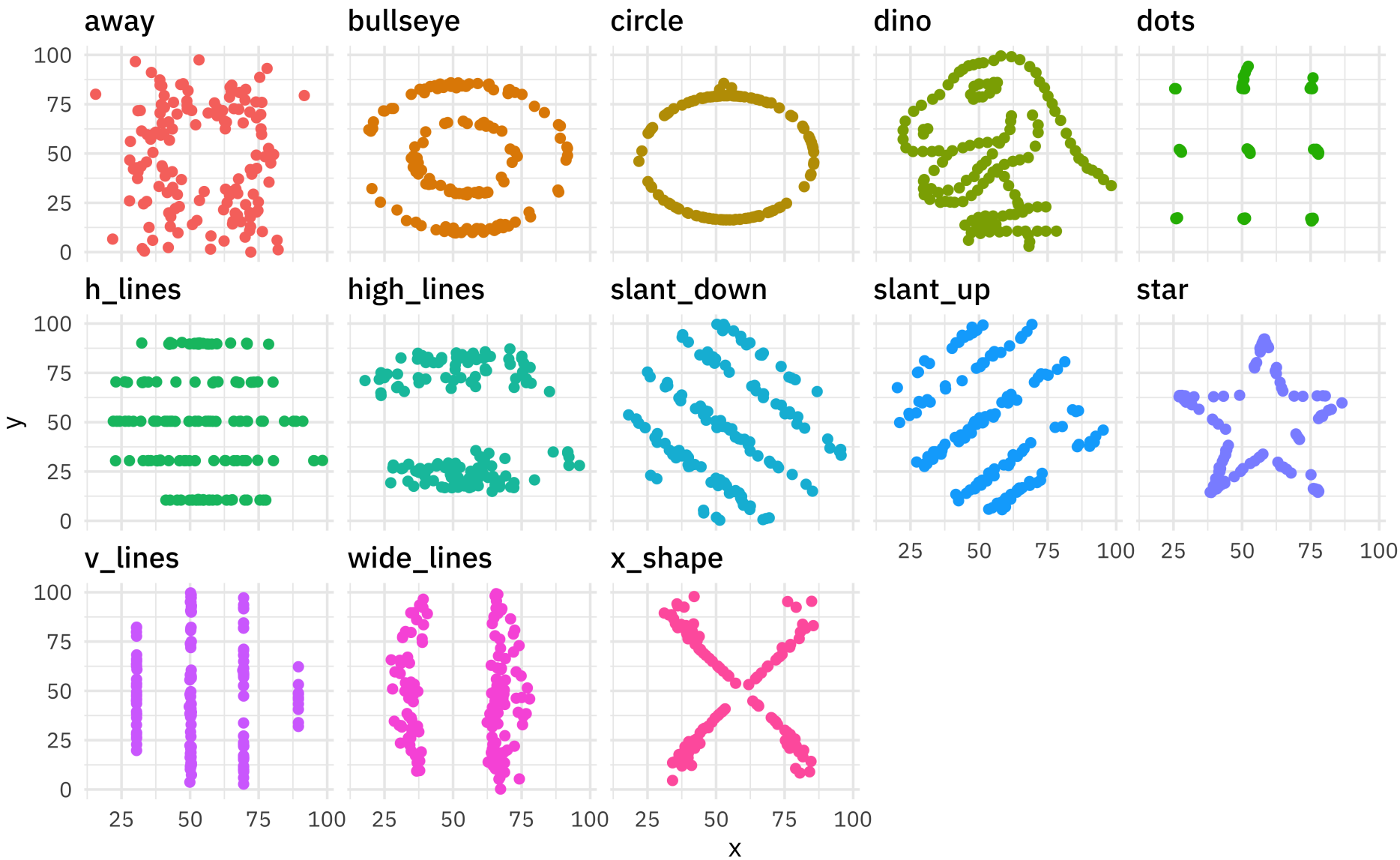




Anscombe's Quartet Stats

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

Datasaurus Dozen



Datasaurus Stats

```
datasaurus_dozen %>%
  group_by(dataset) %>%
  summarise(across(c(x, y), list(mean = mean, sd = sd)),
    x_y_cor = cor(x, y)
  )
```

```
## # A tibble: 13 x 6
##    dataset    x_mean  x_sd y_mean  y_sd x_y_cor
##    <chr>       <dbl> <dbl>  <dbl> <dbl>   <dbl>
##  1 away         54.3  16.8   47.8  26.9 -0.0641
##  2 bullseye     54.3  16.8   47.8  26.9 -0.0686
##  3 circle       54.3  16.8   47.8  26.9 -0.0683
##  4 dino         54.3  16.8   47.8  26.9 -0.0645
##  5 dots         54.3  16.8   47.8  26.9 -0.0603
##  6 h_lines      54.3  16.8   47.8  26.9 -0.0617
##  7 high_lines   54.3  16.8   47.8  26.9 -0.0685
##  8 slant_down   54.3  16.8   47.8  26.9 -0.0690
##  9 slant_up     54.3  16.8   47.8  26.9 -0.0686
## 10 star         54.3  16.8   47.8  26.9 -0.0630
## 11 v_lines      54.3  16.8   47.8  26.9 -0.0694
```

Section 5

FA23 Final Problem 3

Data

The `bus` table (left) records bus arrivals over 1 day for all the bus stops within a 2 mile radius of UCSD. The data dictionary (right) describes each column.

	time	line	stop	late
0	12pm	201	Gilman Dr & Mandeville Ln	-1.1
1	1:15pm	30	Gilman Dr & Mandeville Ln	2.8
2	11:02am	101	Gilman Dr & Myers Dr	-0.8
3	8:04am	202	Gilman Dr & Myers Dr	NaN
4	9am	30	Gilman Dr & Myers Dr	-3.0

`time`	Time of arrival (`str`). Note that the times are inconsistently entered (e.g. `12pm` vs. `1:15pm`).
`line`	Bus line (`int`). There are multiple buses per bus line each day.
`stop`	Bus stop (`str`).
`late`	The number of minutes the bus arrived after its scheduled time. Negative numbers mean that the bus arrived early (`float`). Some entries in this column are missing.

Problem 1

Question: Are buses equally likely to be early or late?

How would we generate one sample from the null hypothesis?

- `np.random.choice([-1, 1], bus.shape[0])`
- `np.random.choice(bus['late'], bus.shape[0], replace = True)`
- Randomly permute the `late` column

Solution

- Is this a (standard) hypothesis test or a permutation test?

Solution

- Is this a (standard) hypothesis test or a permutation test?
- What is the null hypothesis? Which option generates from the null hypothesis?

Solution

- Is this a (standard) hypothesis test or a permutation test?
- What is the null hypothesis? Which option generates from the null hypothesis?
- The answer is `np.random.choice([-1, 1], bus.shape[0])`

Problem 1 (cont.)

What test statistic should we use?

Note: while the problem says there is only one solution, post-exam two options for the test statistic were given credit. Pick one of the two.

- Number of values below 0
- `np.mean`
- `np.std`
- TVD
- K-S statistic

Solution

- What information do we want to capture about each sample?

Solution

- What information do we want to capture about each sample?
- Which test statistic captures it?

Solution

- What information do we want to capture about each sample?
- Which test statistic captures it?
- The answer is either # values below 0, or `np.mean`

Problem 2

Question: Is the `late` column MAR dependent on the `line` column?

How would we generate one sample from the null hypothesis?

- `np.random.choice([-1, 1], bus.shape[0])`
- `np.random.choice(bus['late'], bus.shape[0], replace = True)`
- Randomly permute the `late` column

Solution

- Is this a (standard) hypothesis test or a permutation test?

Solution

- Is this a (standard) hypothesis test or a permutation test?
- What is the null hypothesis? Which option generates from the null hypothesis?

Solution

- Is this a (standard) hypothesis test or a permutation test?
- What is the null hypothesis? Which option generates from the null hypothesis?
- The answer is to randomly permute the `late` column

Section 6

A note on coding questions on the exams

A note on coding questions on the exams

- What we've been teaching you so far is a toolkit of things you can do with DataFrames and data.

A note on coding questions on the exams

- What we've been teaching you so far is a toolkit of things you can do with DataFrames and data.
- When we ask you coding questions, we're trying to get you to match those tools with problems

A note on coding questions on the exams

- What we've been teaching you so far is a toolkit of things you can do with DataFrames and data.
- When we ask you coding questions, we're trying to get you to match those tools with problems
- Try to break down each problem into subproblems to go from point A to point B, and then translate subproblems into tools

A note on coding questions on the exams

- What we've been teaching you so far is a toolkit of things you can do with DataFrames and data.
- When we ask you coding questions, we're trying to get you to match those tools with problems
- Try to break down each problem into subproblems to go from point A to point B, and then translate subproblems into tools
- Good luck!

Section 7

Attendance

Attendance

Once I give you a number, fill out the following Google form: https://forms.gle/wz1r6G2oiwn8cigU8