Chi-squared tests: Goodness of fit 1

Research Methods for Human Inquiry

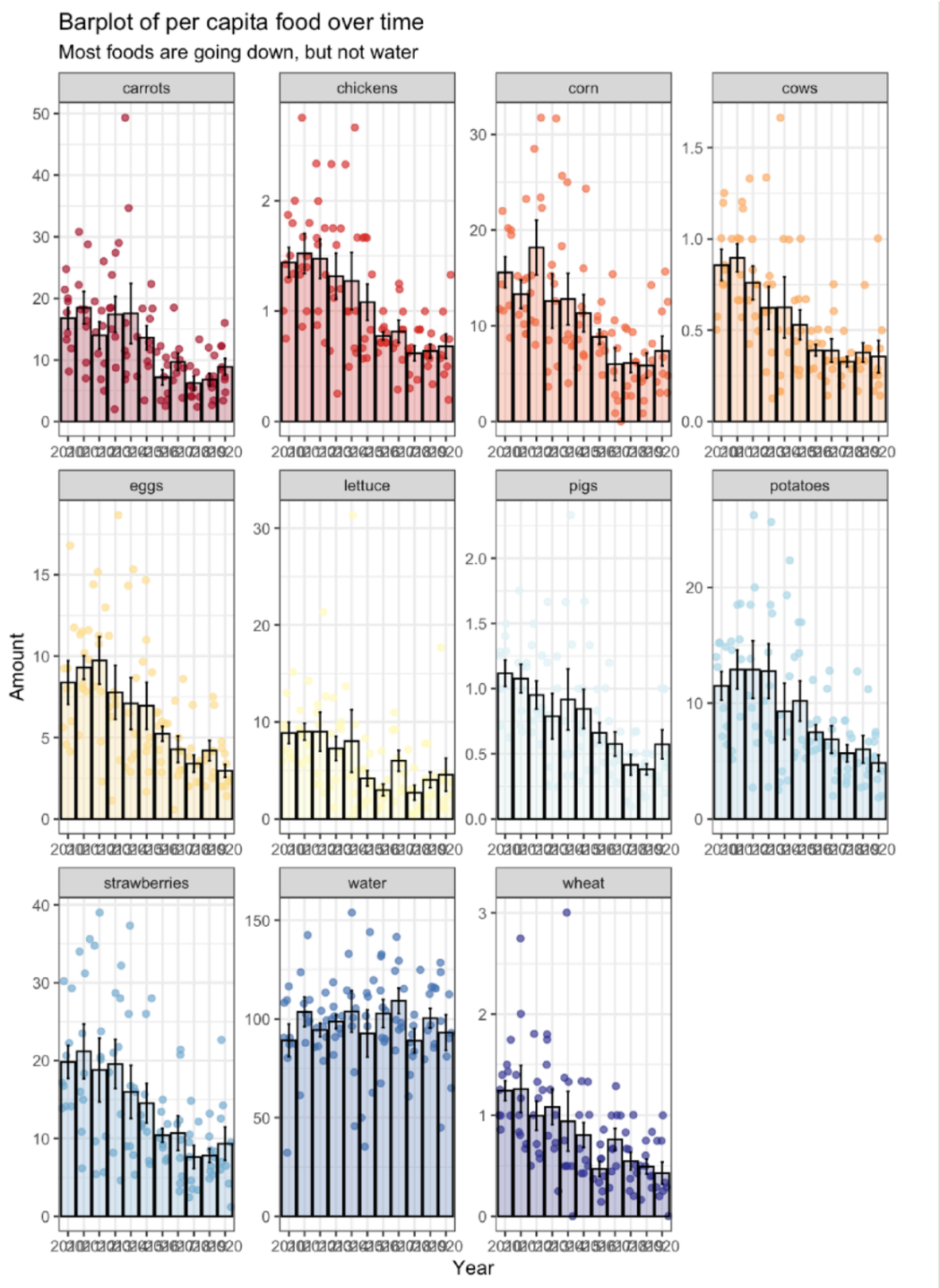
Andrew Perfors

Today's story…

Things look dire! People in Bunnyland are getting hungrier and hungrier

Today's story…

It seems like the food in Otherland is decreasing too!



Today's story…

And LFB and Foxy are gone, lost on the mission to Otherland :(

???

Today's story…

Everyone is fighting about what to do next

We need to go back to Otherland and rescue them, even if it takes several tries

Today's story…

Everyone is fighting about what to do next

We'll just lose other people and not be addressing our big problem: the food. We need to get more data to figure out what is going on

Today's story…

Everyone is fighting about what to do next

Data! You always just want data and never *do* anything. We don't need data, we need to attack the Others. They are behind this somehow and we can steal their food.

Today's story…

Everyone is fighting about what to do next

That's dangerous!
And I think they didn't cause this; it's probably the weather or something. I think we need to think about moving somewhere better.

Today's story…

Everyone is fighting about what to do next

QUACK! I don't know what to do, but I'm scared and I wish everyone would stop fighting

Let’s vote

100 people voted between each option

BUNNY	DOGGIE	GLADLY	SHADOW
			
leave	attack	rescue	analyse
2	55	36	7

Objection

How do we know that's what people chose for the actual option, rather than just reflecting how much they support the *person*?

Luckily…

These are the exact people who ran in a recent council election

BUNNY	DOGGIE	GLADLY	SHADOW
			
0.125	0.455	0.334	0.086

Our question:

Are these two things different?

BUNNY	DOGGIE	GLADLY	SHADOW
0.125	0.455	0.334	0.086

Leave (B)	Attack (D)	Rescue (G)	Analyse (S)
2	55	36	7

For this, we're going to use a
Chi-squared Goodness of Fit test

Chi-Squared tests

- Outcome variable *x* is **nominal**

Goodness of fit

- Comparing *x* against a theoretical prediction *p*

BUNNY	DOGGIE	GLADLY	SHADOW
0.125	0.455	0.334	0.086

Here's my "actual" data

w6daylanalysis.Rmd

```
> loc <- here("votingresults.csv")
> d <- read_csv(file=loc)
> d

# A tibble: 100 x 2
   person   vote
   <chr>    <chr>
 1 person1  gladly
 2 person2  gladly
 3 person3  doggie
 4 person4  doggie
 5 person5  doggie
 6 person6  gladly
 7 person7  gladly

BLAH BLAH BLAH

100 person100 gladly
```

And the R Markdown file already contains the election data to compare it to:

```
> ed
 bunny doggie gladly shadow
 0.125  0.455  0.334  0.086
```

```
> votingTable <- table(d$vote)
> votingTable

 bunny doggie gladly shadow
     2     55     36      7
```

What are our hypotheses?

- Research hypothesis: "people voted this time at least in part based on the option and not just the person"

- Statistical hypotheses:
  - Null, $\mathbf{H_0}$: The counts that we see in the `d` data (about what to do) reflect the same probabilities that the previous election data `ed` suggest
  - Alternative, $\mathbf{H_1}$: The `d` data are systematically different from the `ed` probabilities in some manner.

Let's be a little more precise

- Let θ refer to the true (but unknown) probability of endorsing each of the options.

- We have four population parameters:
  - θ_1 "probability of choosing to leave (Bunny)"
  - θ_2 "probability of choosing to attack (Doggie)"
  - θ_3 "probability of choosing to rescue (Gladly)"
  - θ_4 "probability of choosing to analyse (Shadow)"

Let's be a little more precise

- The null hypothesis claims that these population parameters are identical to the data on how people voted in the election last time

- So, $\mathbf{H_0}$ states that:
  - $\theta_1 = 0.125$ (Bunny)
  - $\theta_2 = 0.455$ (Doggie)
  - $\theta_3 = 0.334$ (Gladly)
  - $\theta_4 = 0.086$ (Shadow)

Let's be a little more precise

- We'll use $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ as notation to refer to the four different probabilities.

- Null, $\mathbf{H_0}$ ...
  - $\boldsymbol{\theta} = (\ 0.125,\ \ 0.455,\ \ 0.334, 0.086\ )$

- Alternative, $\mathbf{H_1}$ ...
  - $\boldsymbol{\theta} \neq (\ 0.125,\ \ 0.455,\ \ 0.334, 0.086\ )$

Notation to refer to the data

- Let $\mathbf{O}$ refer to the "observed" frequencies
- That is:
  - $O_1 = 2$ (i.e. 2 people voted to leave (Bunny))
  - $O_2 = 55$ (i.e., 55 people voted to attack (Doggie))
  - $O_3 = 36$ (i.e., 36 people voted to rescue (Gladly))
  - $O_4 = 7$ (i.e., 7 people voted to analyse (Shadow))

- Vector of observed frequencies:
  - $\mathbf{O} = (2, 55, 36, 7)$

Summary so far:

- Our statistical hypotheses:
  - $H_0 : \boldsymbol{\theta} = (0.125, 0.455, 0.334, 0.086)$
  - $H_1 : \boldsymbol{\theta} \neq (0.125, 0.455, 0.334, 0.086)$

- Our data:
  - $\mathbf{O} = (2, 55, 36, 7)$

Next: building a statistical test

Exercises are in w6day1exercises.Rmd