Describing Categorical Variables

David Gerard

2017-09-18

Learning Objectives

- Two-way tables.
- Conditional distributions.
- Bar Charts (and pie-charts)
- Section 1.7 of DBC

Recall: email dataset

These data represent incoming emails for the first three months of 2012 for an email account.

Some variables:

- `spam` Indicator for whether the email was spam.
- `to_multiple` Indicator for whether the email was addressed to more than one recipient.
- `viagra` The number of times "viagra" appeared in the email.
- `num_car` The number of characters in the email, in thousands.
- `number` Factor variable saying whether there was no number, a small number (under 1 million), or a big number.

Recall: Email dataset

```
library(tidyverse)
data("email", package = "openintro")
head(select(email, spam, to_multiple,
                viagra, num_char, number))
```

```
  spam to_multiple viagra num_char number
1    0           0      0   11.370    big
2    0           0      0   10.504  small
3    0           0      0    7.773  small
4    0           0      0   13.256  small
5    0           0      0    1.231   none
6    0           0      0    1.091   none
```

Distribution of categorical variable

- Recall: The distribution of a variable tells us what values it takes and how often it takes these values
- In terms of categorical variables, the distribution is just the counts of cases/proportions/percents in each category.
- A table of counts for a single variable is a frequency table.

```
table(email$number)
```

```
 none small   big
  549  2827   545
```

The relative frequency table

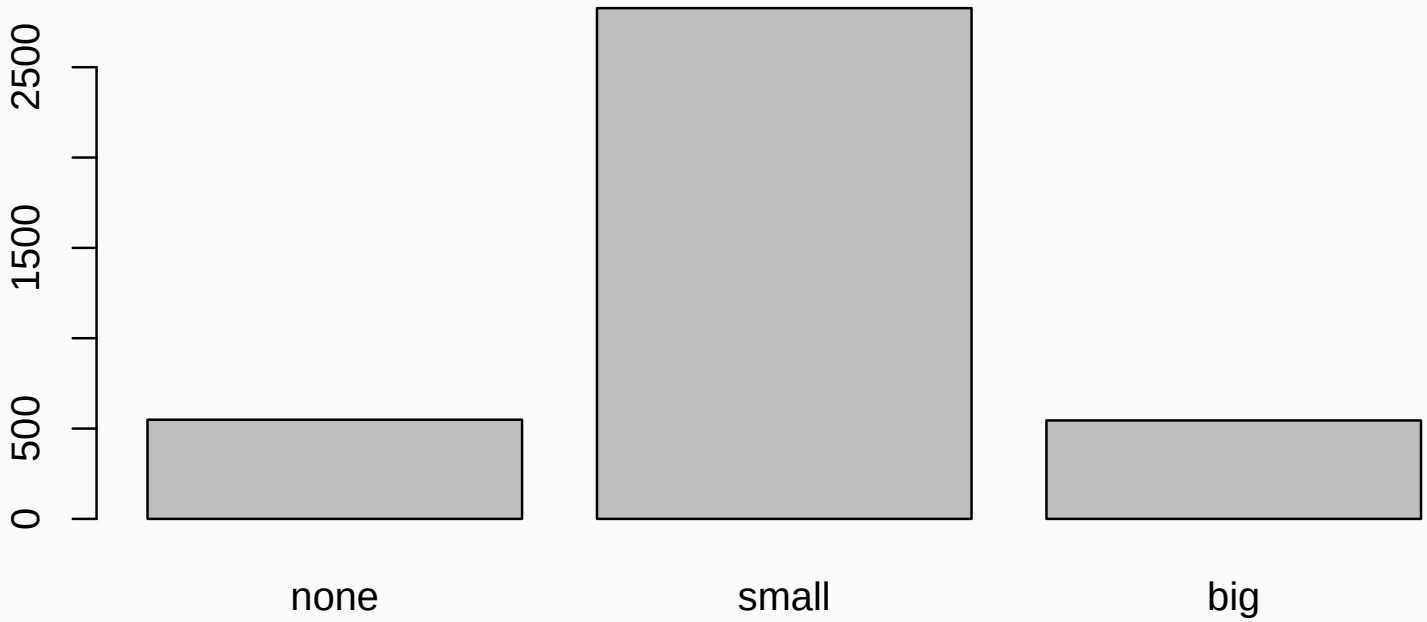
- A table of proportions/percentages for a single variable is a relative frequency table.

```
prop.table(table(email$number))
```

```
 none small   big
0.140 0.721 0.139
```

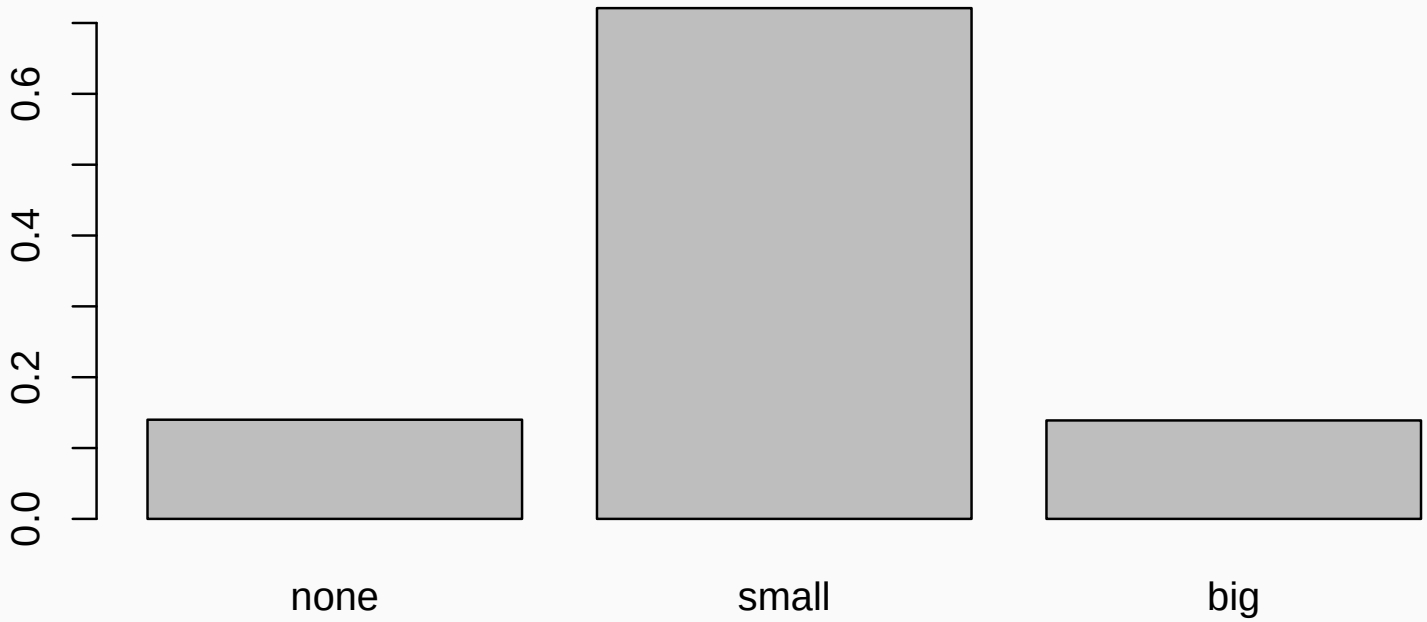
Barchart

```
barplot(table(email$number)) ## need table
```



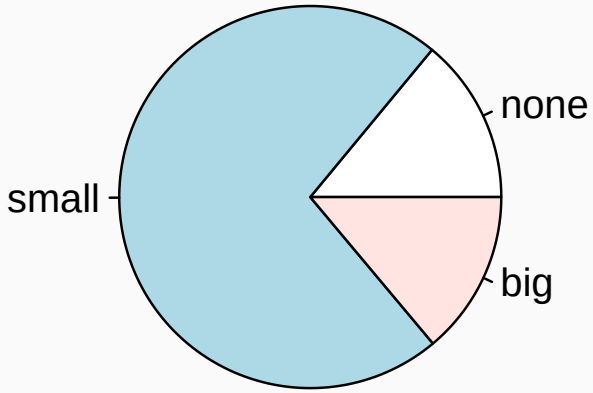
Barchart of proportions

```
barplot(prop.table(table(email$number))) ## need table
```



Piecharts

```
pie(table(email$number))
```

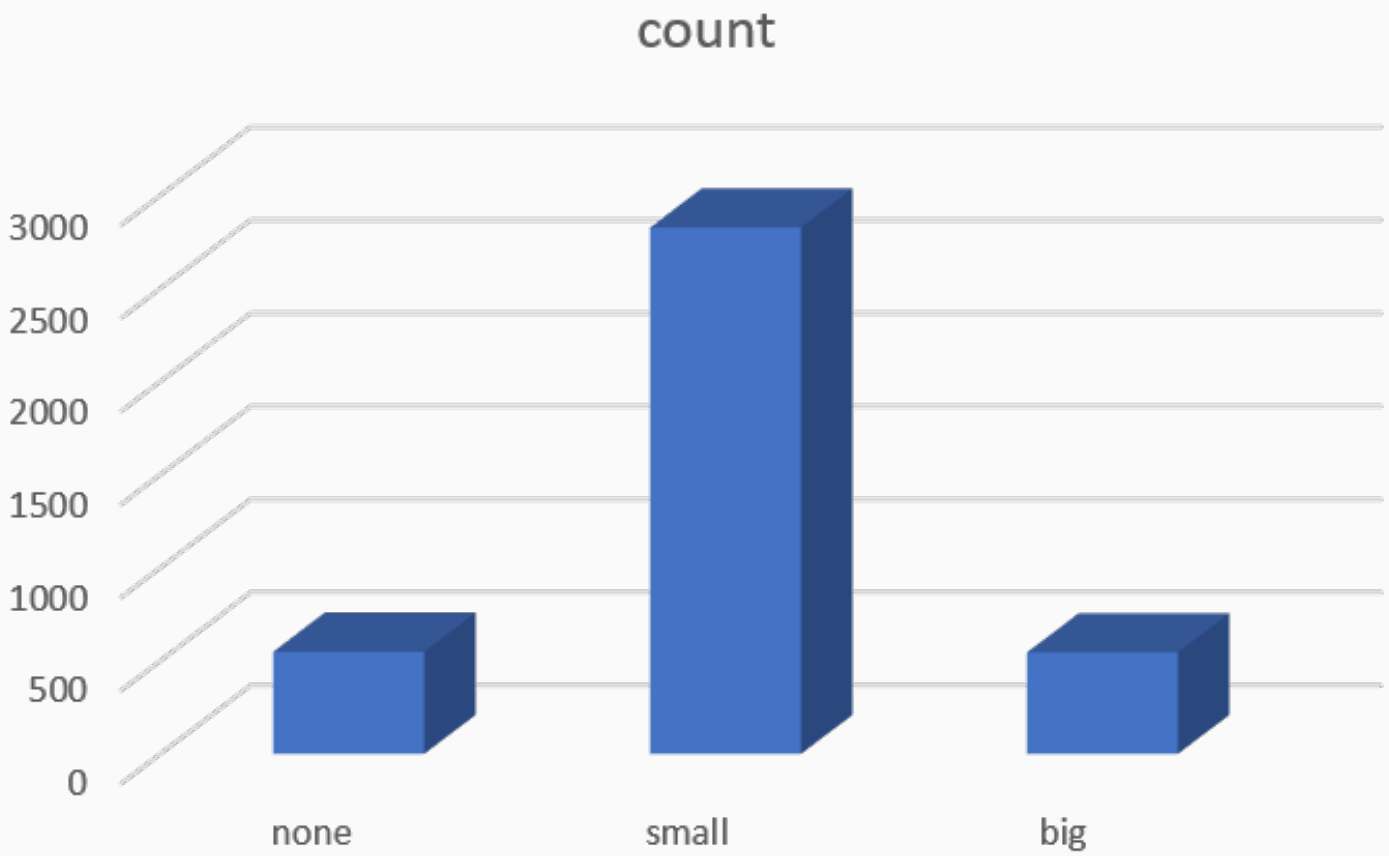


Never use picharts

- Humans find it easier to distinguish height rather than distinguish area.
- Which category has more emails: “big” or “none”.
- In which plot is it easier to see which category has more emails?

Never use 3D graphics to plot 2D data

They tend to distort/obscure the view of the data and are distracting.



Joint distribution

- What about the *joint* distribution of two categorical variables?
- The distribution of a variable tells us what values it takes and how often it takes these values.
- The joint distribution is just the counts of cases/proportions/percents in each possible combination of categories.
- A table of these counts is a contingency table, also called a two-way table.

First Contingency Table

```
tabdat <- table(email$spam, email$number)
rownames(tabdat) <- c("Not Spam", "Spam")
tabdat
```

```
          none small  big
  Not Spam  400  2659  495
  Spam      149   168   50
```

Often shown the row/column totals (or "margins")

	none	small	big	total
Not Spam	400	2659	495	3554
Spam	149	168	50	367
total	549	2827	545	3921

- What does 2659 represent?
- What does 495 represent?
- What does 3554 represent?
- What does 2827 represent?
- What does 3921 represent?

Joint Distribution

More informative: joint distribution in proportions:

```
prop.table(tabdat)
```

```
             none   small     big
Not Spam 0.10201 0.67814 0.12624
Spam     0.03800 0.04285 0.01275
```

- What does 0.6781 represent?
- What does 0.1262 represent?

Row Proportions

row proportions

The row proportions are computed as the counts divided by the row totals.

```
prop.table(tabdat, margin = 1)
```

```
              none  small    big
  Not Spam 0.1125 0.7482 0.1393
  Spam     0.4060 0.4578 0.1362
```

- What does 0.7482 represent?
- What does 0.1393 represent?

Column Proportions

column proportions

The column proportions are computed as the counts divided by the column totals.

```
prop.table(tabdat, margin = 2)
```

```
              none   small     big
  Not Spam 0.72860 0.94057 0.90826
  Spam     0.27140 0.05943 0.09174
```

- What does 0.9406 represent?
- What does 0.9083 represent?

Why do we care?

- Row/column proportions help us determine if two categorical variables are associated.
- E.g. Is the distribution of spam conditioned on seeing no numbers different from the distribution of spam conditioned on seeing small numbers? If so, then `number` and `spam` are associated.
- Would these be row or column proportions?
- Can also look for associations by checking the distribution of number conditioned on an email being spam and the distribution of number contioned on an email not being spam.
- Would these be row or column proportions?

Notice the word "conditioned"

```
prop.table(tabdat, margin = 2)
```

```
              none   small     big
Not Spam 0.72860 0.94057 0.90826
Spam     0.27140 0.05943 0.09174
```

- The row/column proportions represent conditional distributions.
- Each column is the distribution of spam conditioned on either no big number (column 1), a small number (column 2), or a big number (column 3).

Notice the word "conditioned"

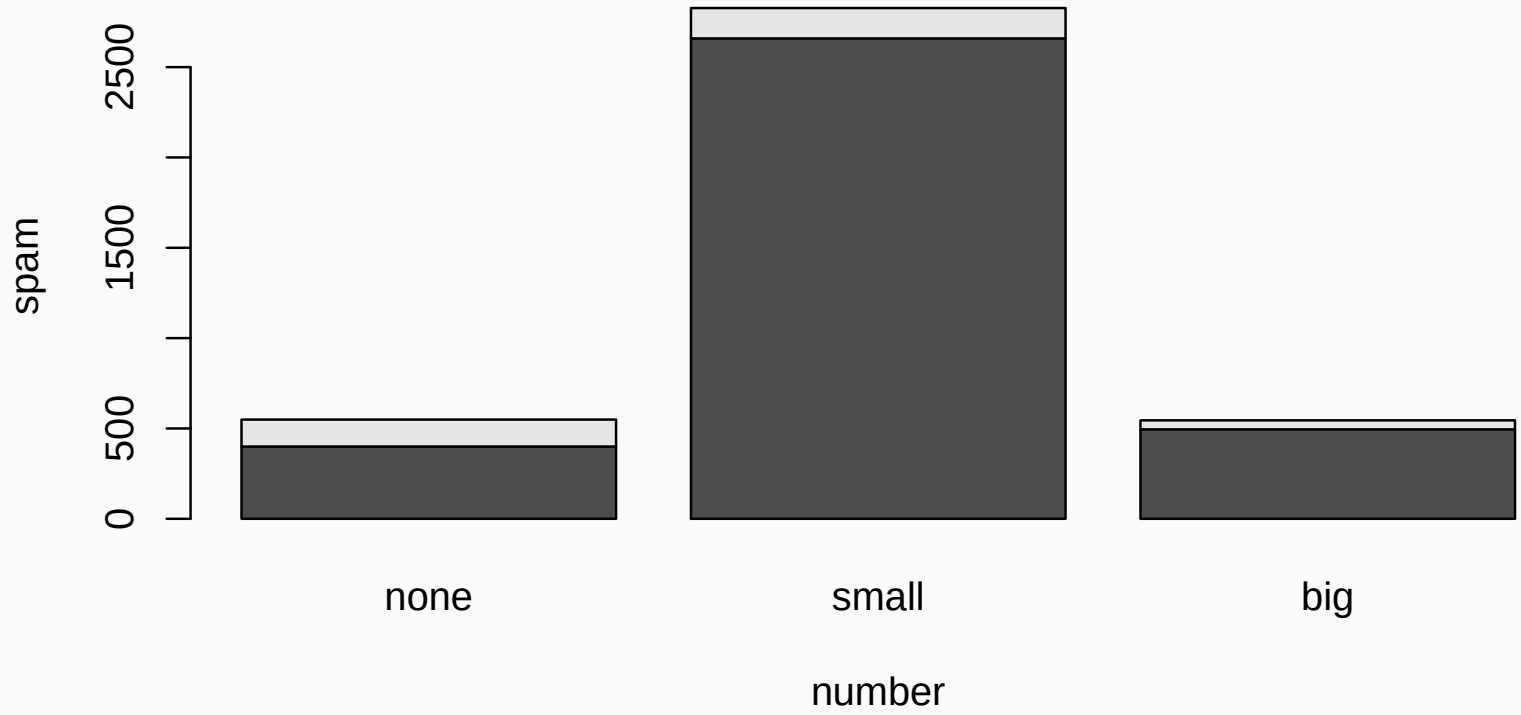
```
prop.table(tabdat, margin = 1)
```

```
              none  small    big
Not Spam 0.1125 0.7482 0.1393
Spam     0.4060 0.4578 0.1362
```

- The row/column proportions represent conditional distributions.
- Each row is the distribution of number conditioned on either an email being not spam (first row) or spam (second row).

Visualizing row proportions: segmented barplot

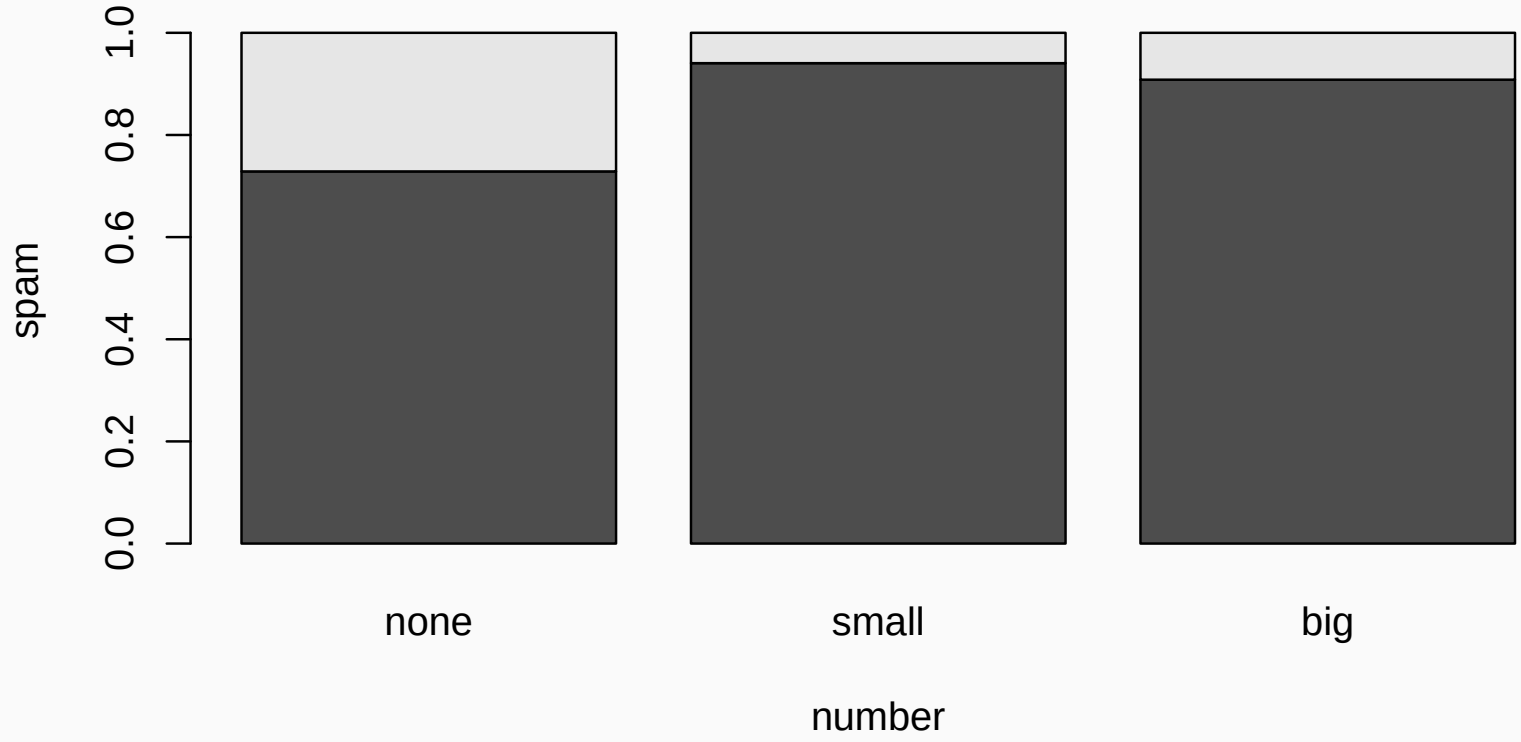
```
barplot(table(email$spam, email$number),
        xlab = "number", ylab = "spam")
```



What does the bottom left box represent?

Visualizing row proportions: standardized segmented barplot

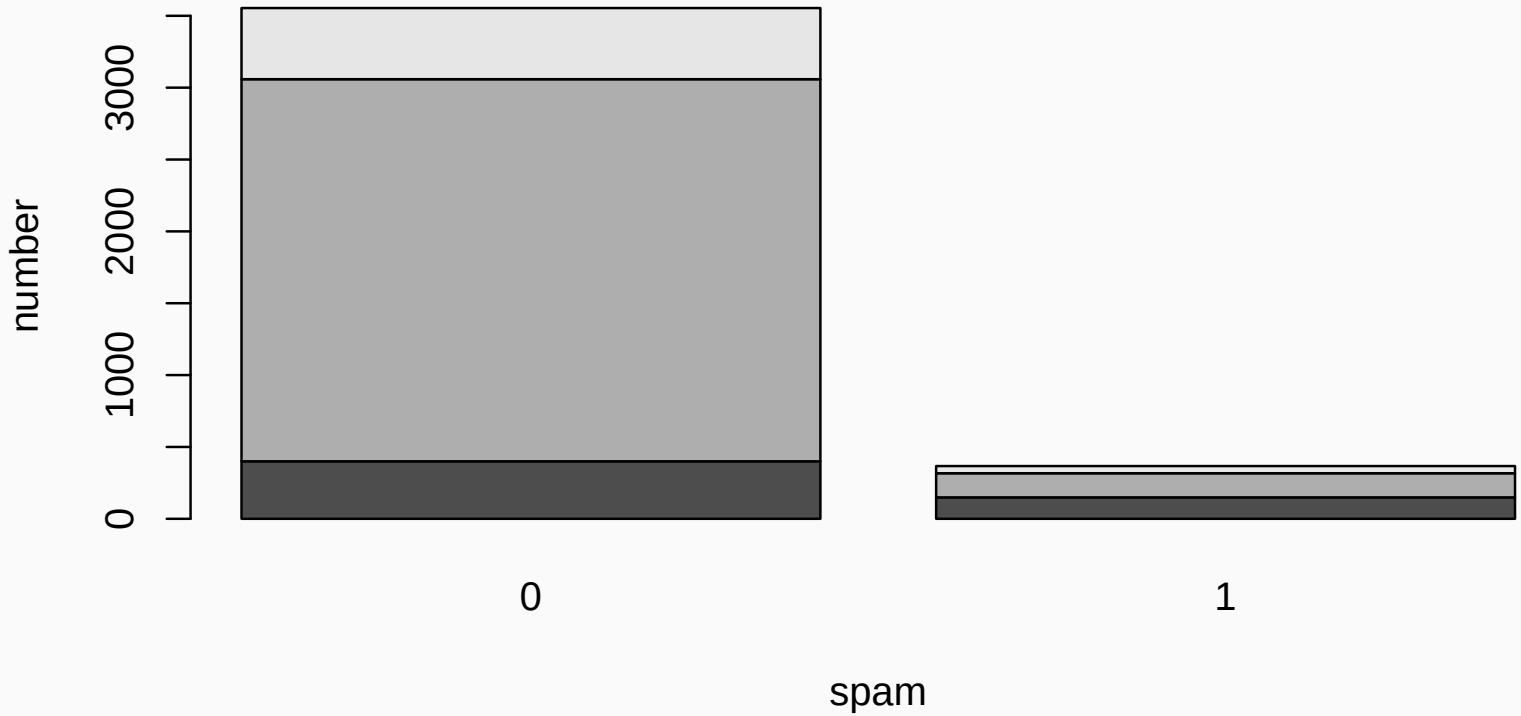
```
barplot(prop.table(table(email$spam, email$number),
                   margin = 2),
        xlab = "number", ylab = "spam")
```



What does the bottom left box represent?

Visualizing row proportions: segmented barplot

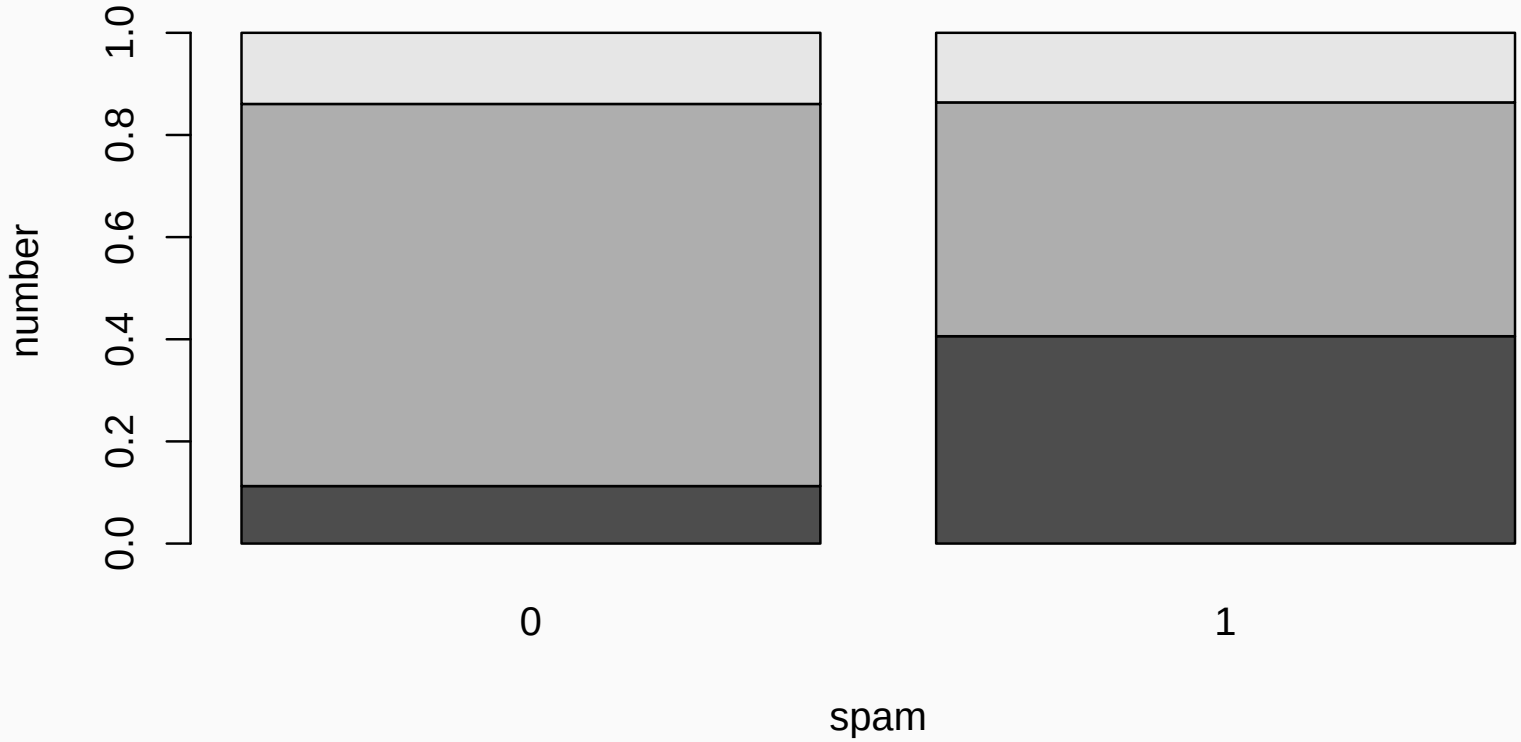
```
barplot(table(email$number, email$spam),
        xlab = "spam", ylab = "number")
```



What does the bottom left box represent?

Visualizing row proportions: standardized segmented barplot

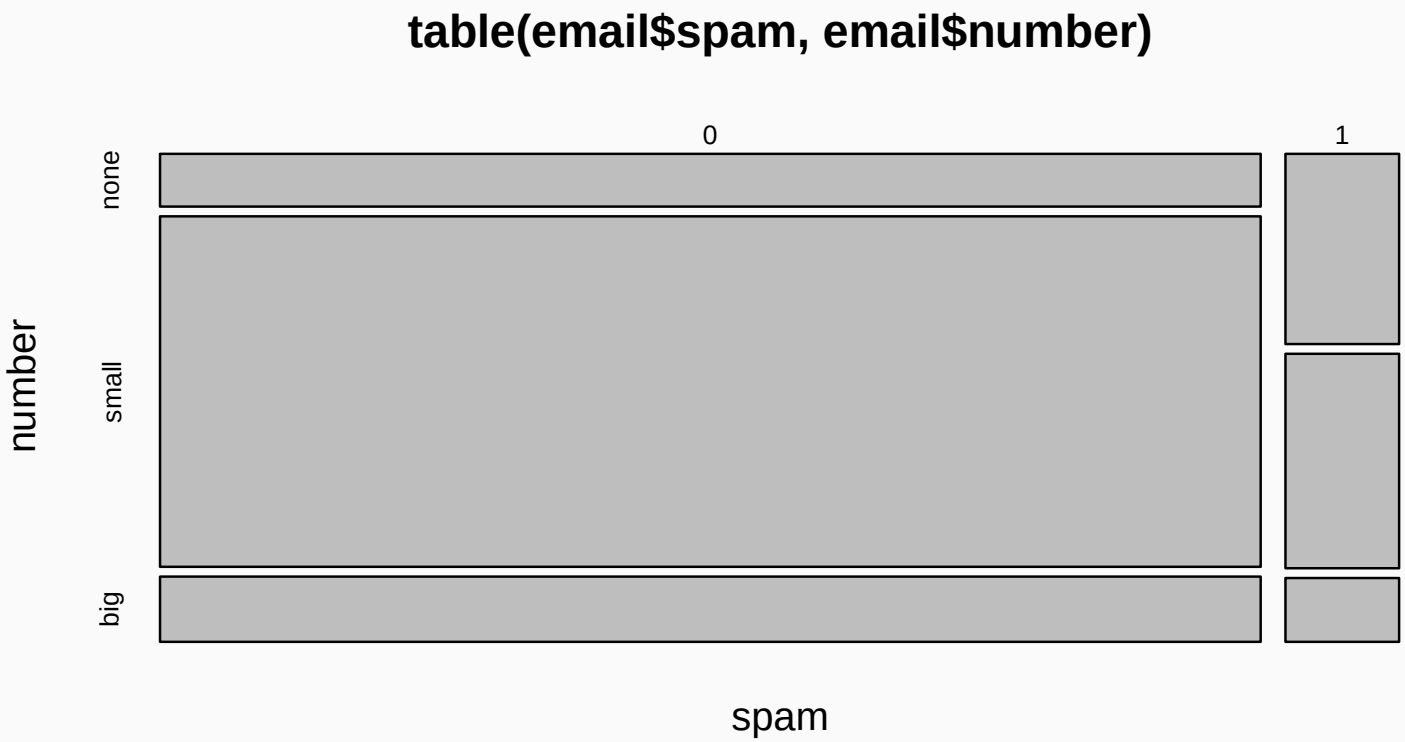
```
barplot(prop.table(table(email$number, email$spam),
                   margin = 2),
        xlab = "spam", ylab = "number")
```



What does the bottom left box represent?

Visualizing row proportions: mosaic plot

```
plot(table(email$spam, email$number),
     xlab = "spam", ylab = "number")
```

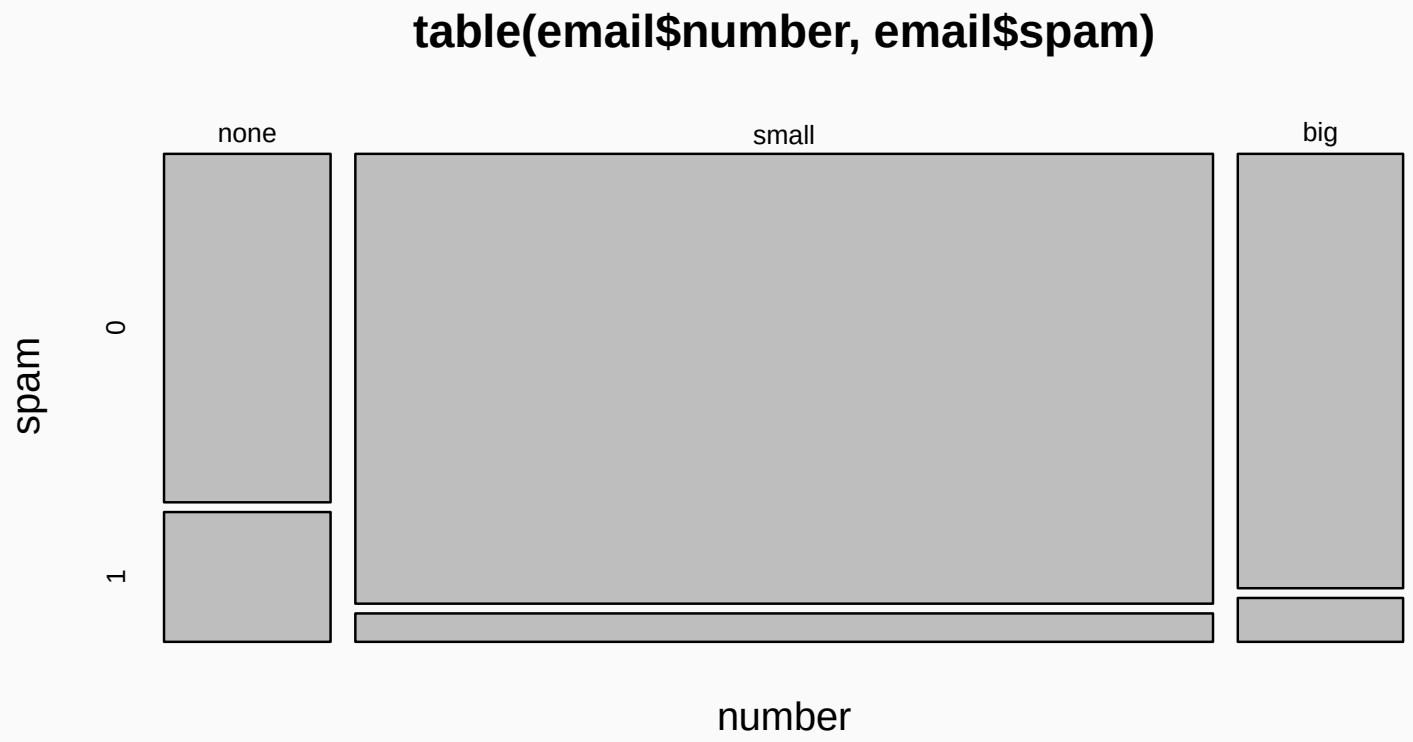


Width proportional to the counts in each `spam` category.

What does the bottom left box represent?

Visualizing row proportions: mosaic plot

```
plot(table(email$number, email$spam),
     xlab = "number", ylab = "spam")
```



Width proportional to the counts in each `number` category.

What does the bottom left box represent?

What's important in a mosaic plot?

- What in a mosaic plot are we looking for to see if two variables are associated?