Categorical data

Professor Jarad Niemi

STAT 226 - Iowa State University

August 22, 2018

Outline

- Distribution, frequency table
- Bar chart
- Pareto chart
- Pie chart
- Mode and median
- Contingency table

Distribution and frequency tables

Definition

The distribution of a variable is the collection of possible values the variable can take and how often each value occurs.

Definition

A frequency table is a summary that shows the distribution of a variable.

For categorical variables, we can use the following to understand the distribution of a variable: frequency table, bar chart, Pareto chart, pie chart, mode, median, and contingency table.

Majors

	Major	College	Clsfn.Year
1	P BUS	M	3
2	P BUS	M	3
3	AG B	A	2
4	MKT	M	2
5	ACCT	M	2
6	P BUS	M	2
7	ACCT	M	2
8	AG B	A	3
9	BUS U	M	2
10	P BUS	M	2

Majors summary

```
     Major     College   Clsfn.Year
P BUS  :65   A:25     Min.   :1.000
AG B   :24   H: 1     1st Qu.:2.000
ACCT   : 3   M:74     Median :2.000
MKT    : 3   S: 1     Mean   :2.485
BUS U  : 2           3rd Qu.:3.000
A M D  : 1           Max.   :4.000
(Other): 3
```

Recode Year

```
    Major     College   Clsfn.Year            Year
P BUS  :65   A:25     Min.   :1.000   Freshman : 2
AG B   :24   H: 1     1st Qu.:2.000   Sophomore:54
ACCT   : 3   M:74     Median :2.000   Junior   :39
MKT    : 3   S: 1     Mean   :2.485   Senior   : 6
BUS U  : 2            3rd Qu.:3.000
A M D  : 1            Max.   :4.000
(Other): 3
```

Tabular summary

Majors:

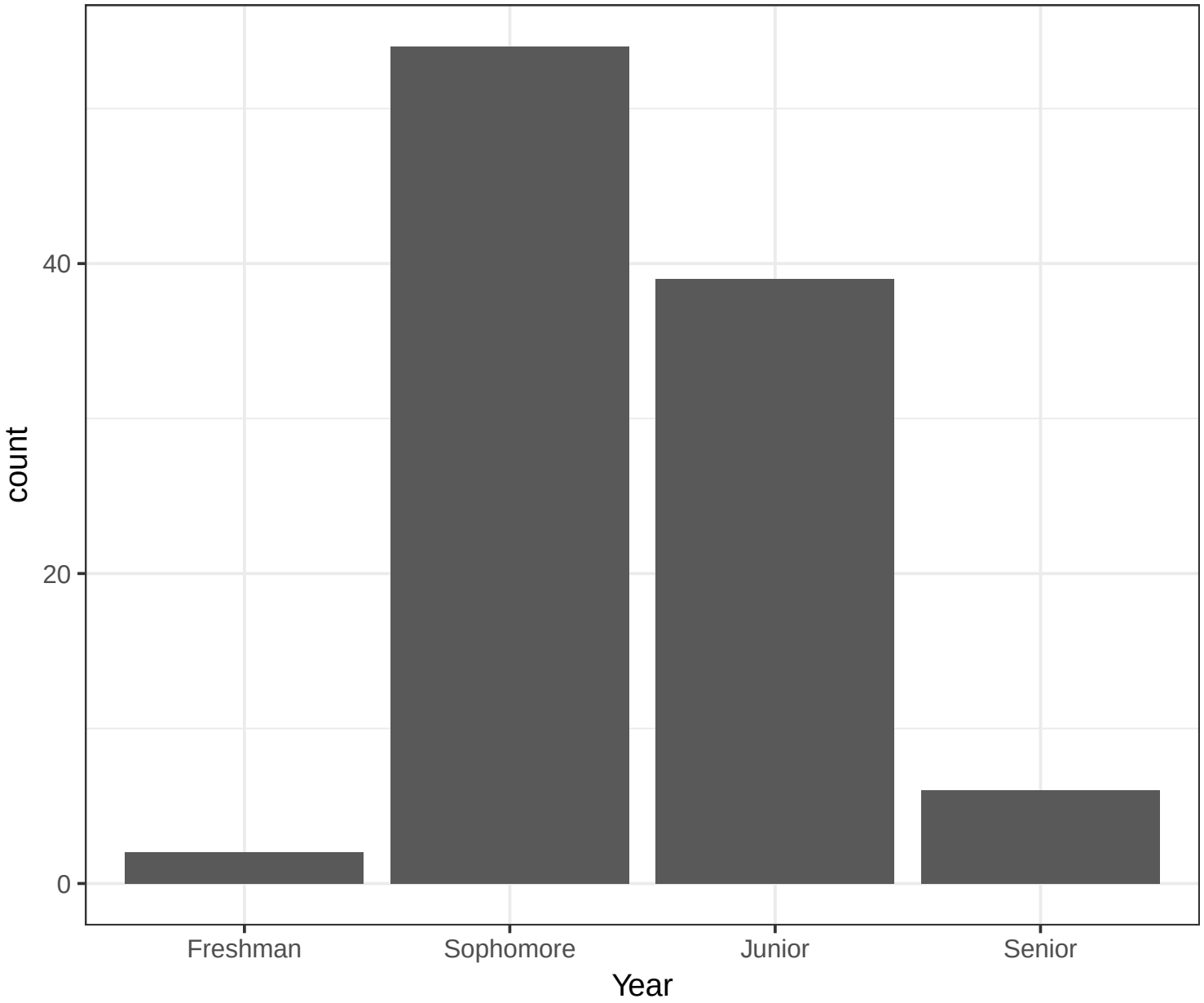
```
A M D ACCT  AG B  AN S  BUS U ECON  FIN   MKT   P BUS
    1    3   24    1    2    1    1    3   65
```

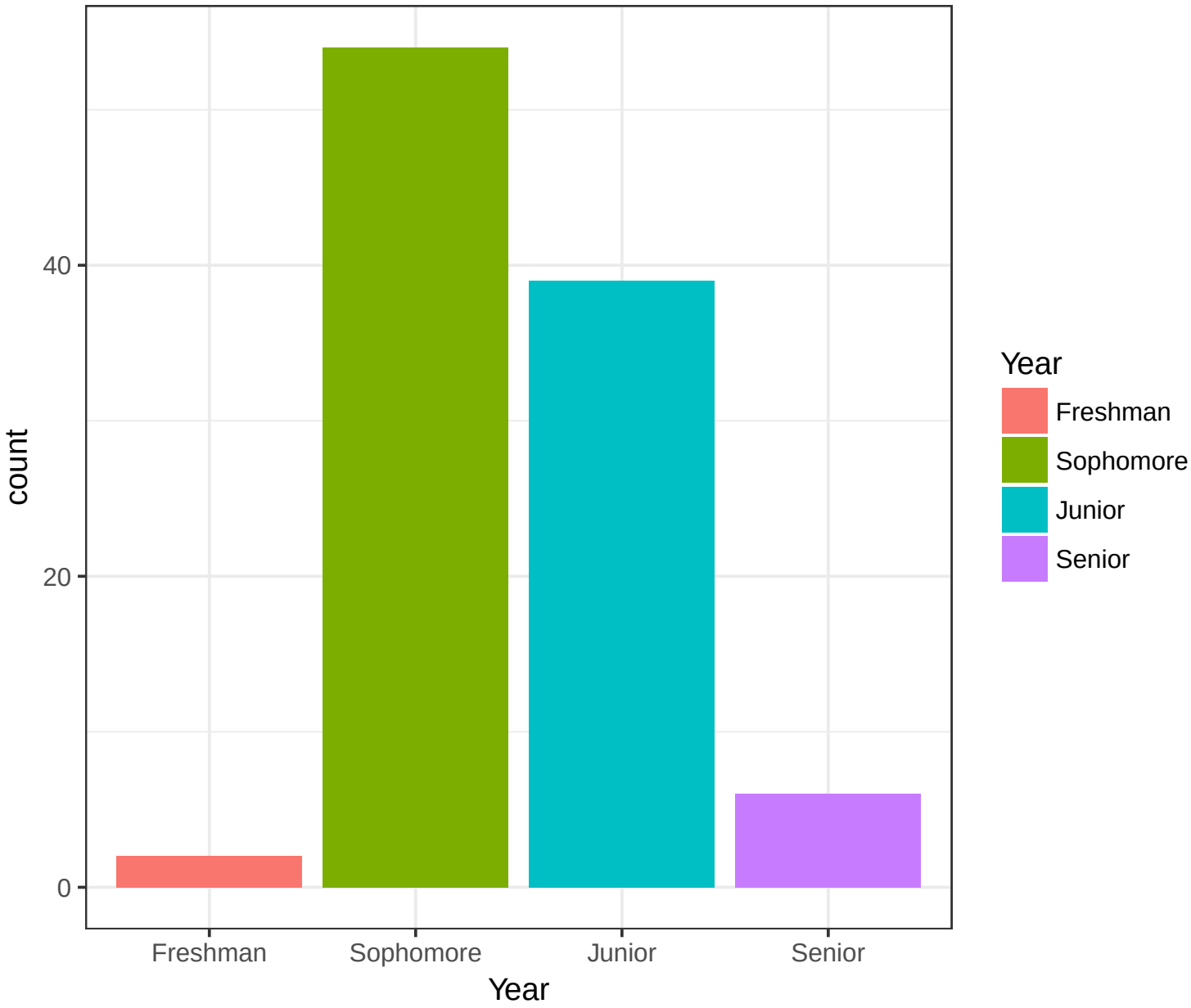
College:

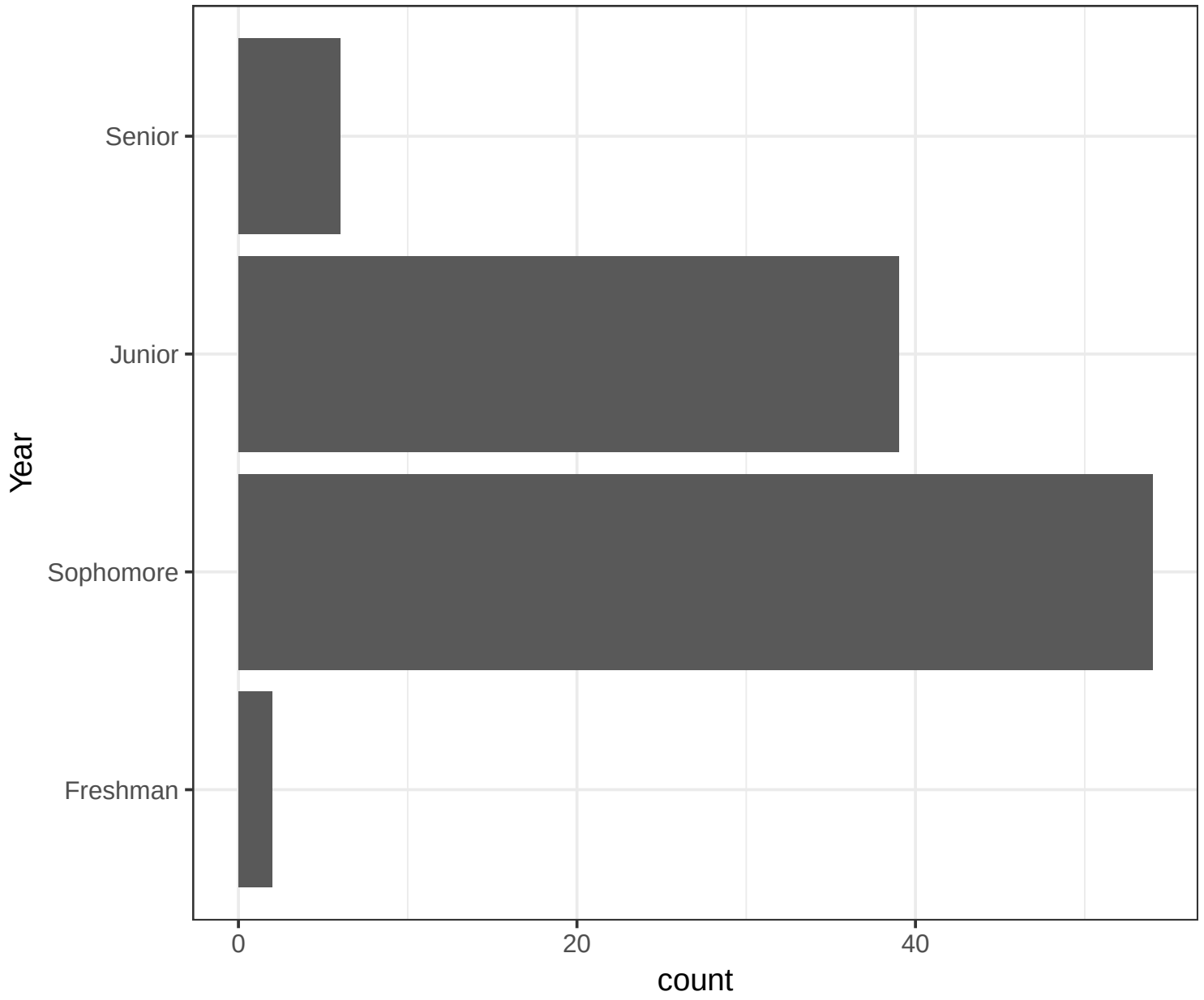
```
 A  H  M  S
25  1 74  1
```

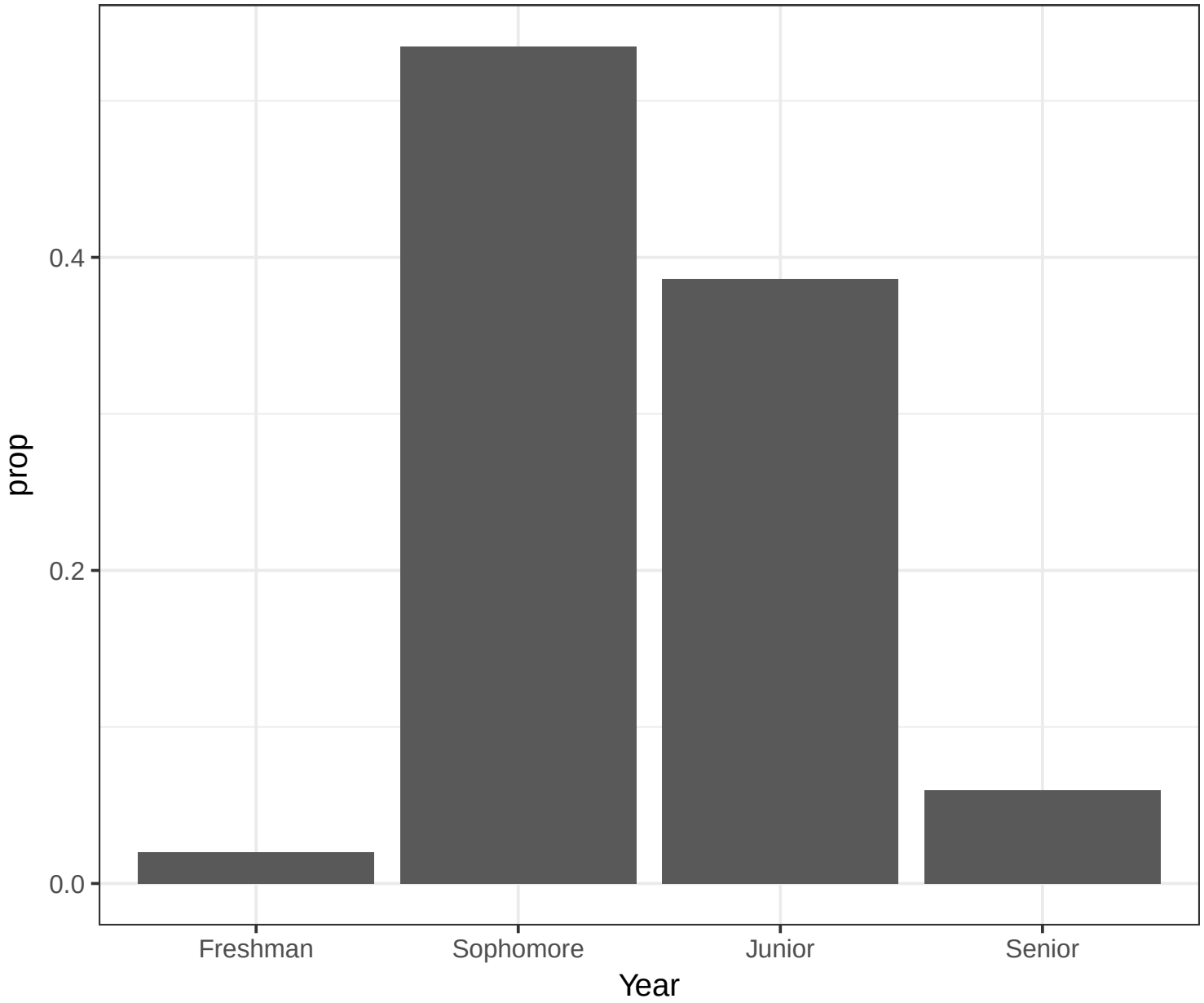
Year:

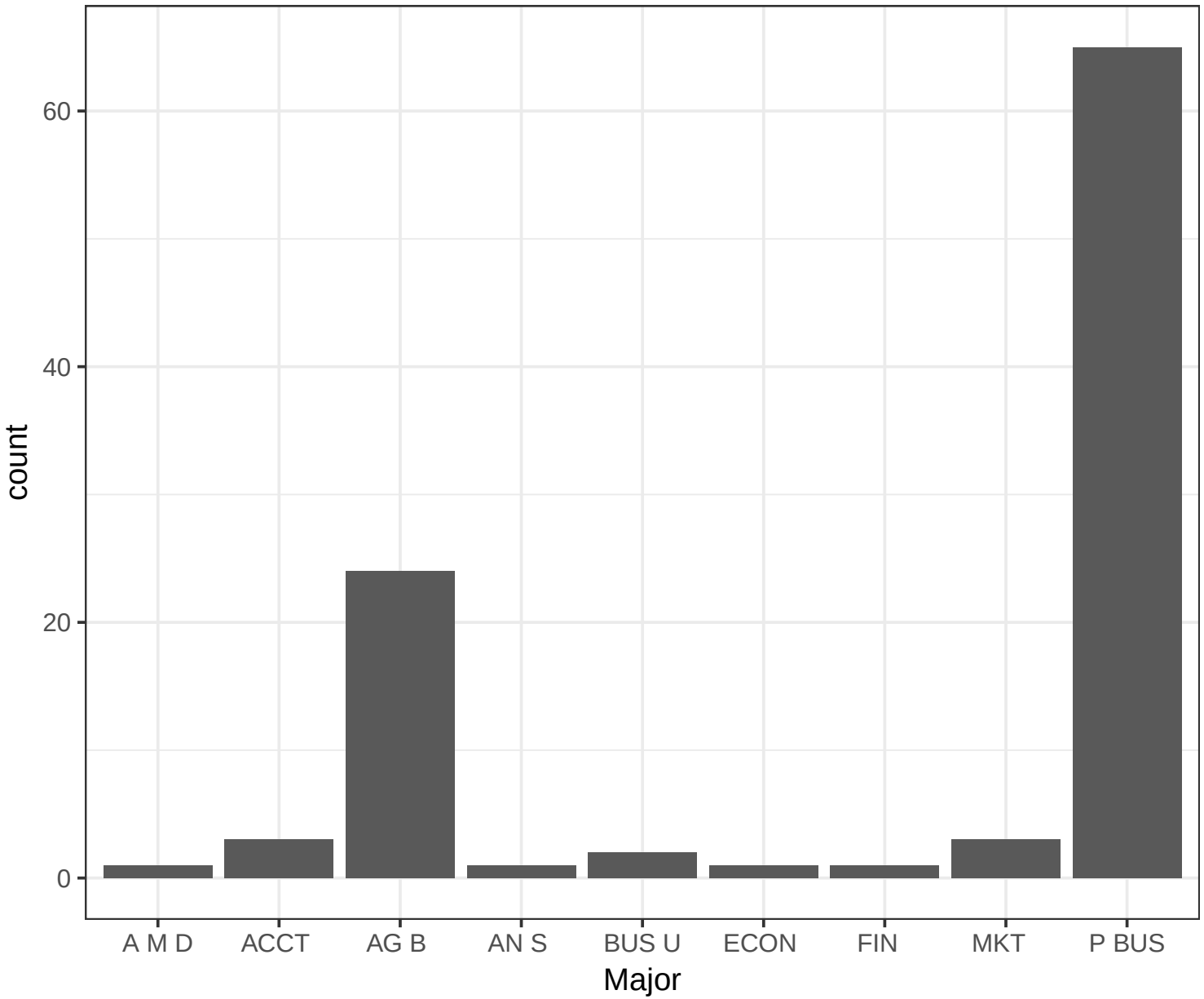
```
 Freshman Sophomore    Junior    Senior
        2        54        39         6
```

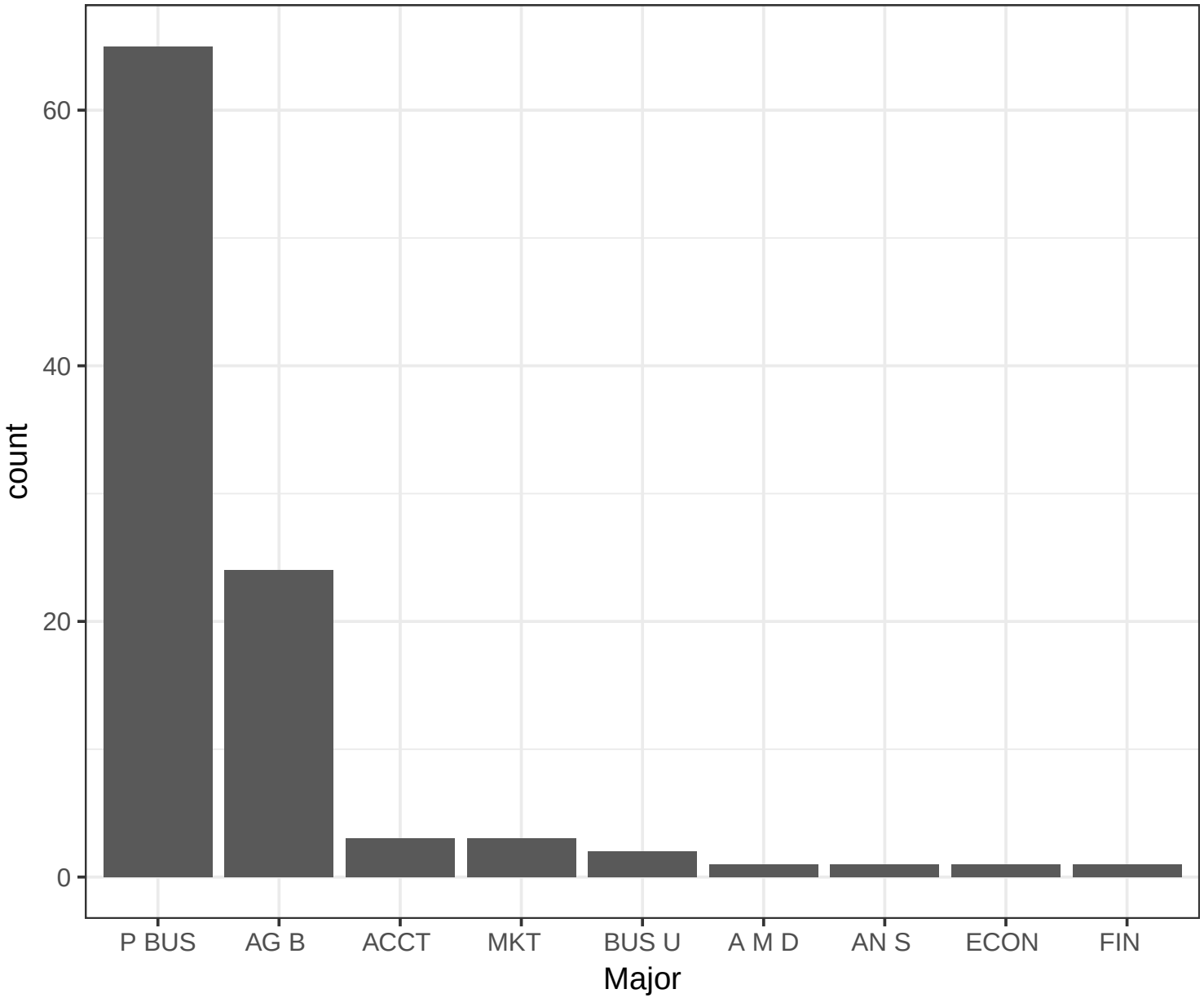

40
20
0
count
Freshman
Sophomore
Junior
Senior
Year

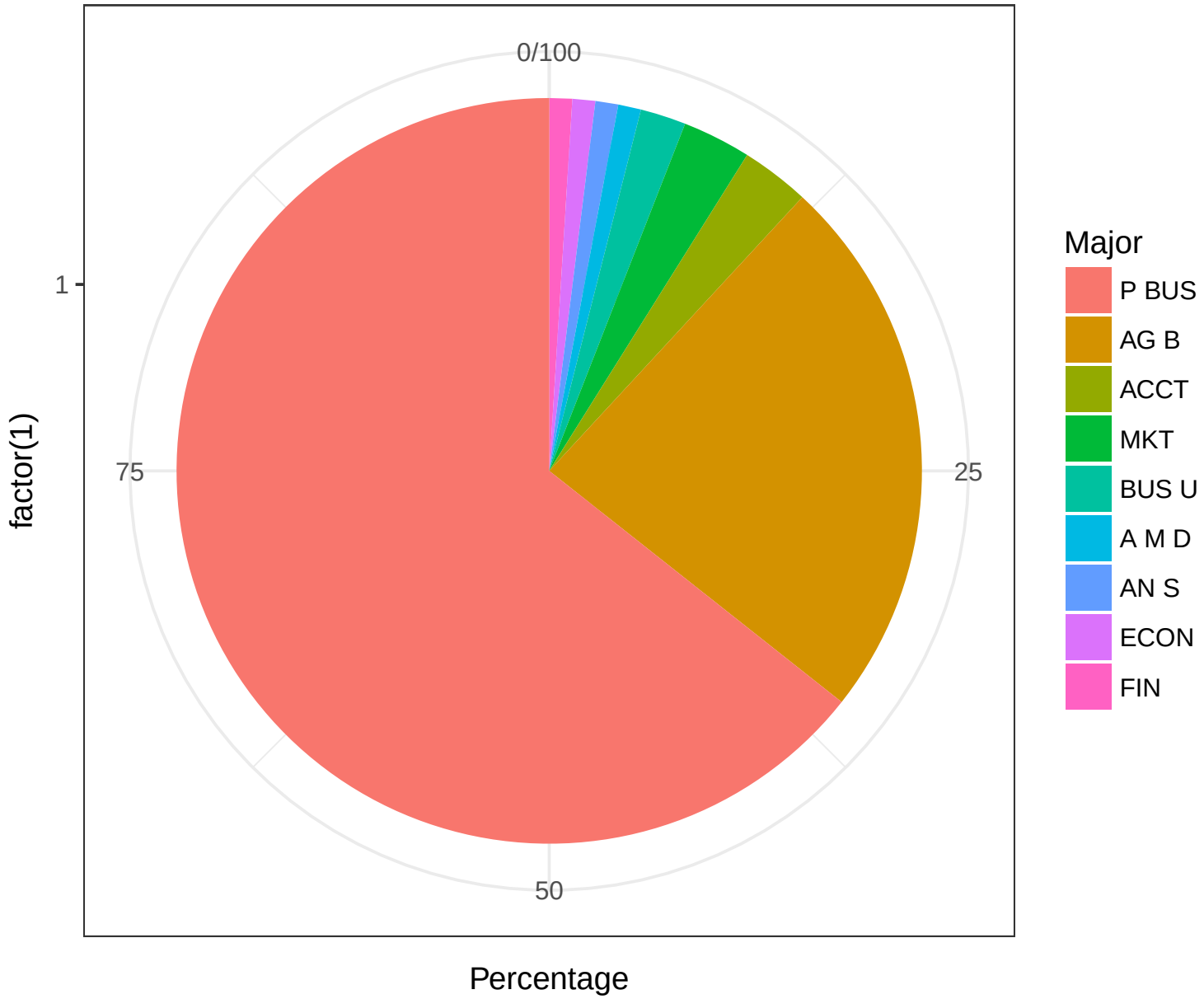

count
40
20
0
Freshman
Sophomore
Junior
Senior
Year
Year
Freshman
Sophomore
Junior
Senior


Senior
Junior
Sophomore
Freshman
Year
0
20
40
count


0.4
0.2
0.0
prop
Freshman
Sophomore
Junior
Senior
Year

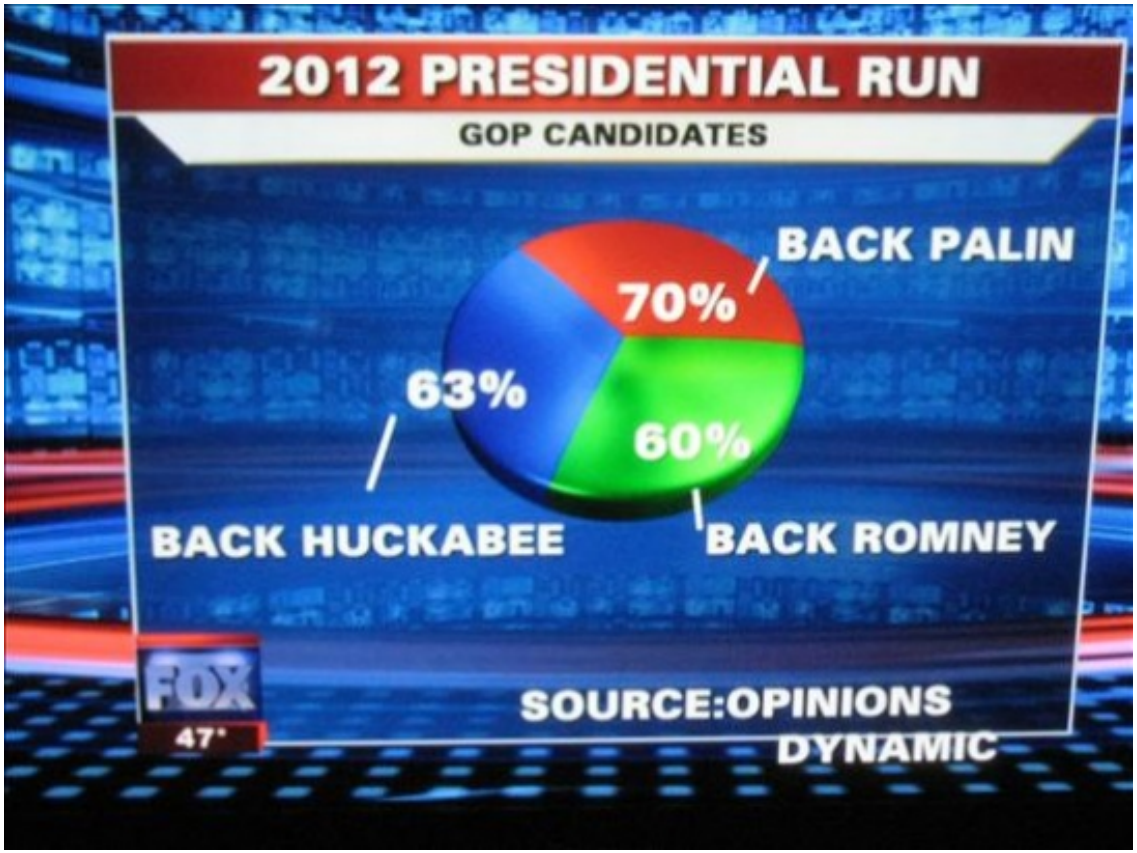

60
40
20
0
count
A M D
ACCT
AG B
AN S
BUS U
ECON
FIN
MKT
P BUS
Major


60
40
20
0
count
P BUS
AG B
ACCT
MKT
BUS U
A M D
AN S
ECON
FIN
Major


0/100
25
50
75
1
factor(1)
Percentage
Major
P BUS
AG B
ACCT
MKT
BUS U
A M D
AN S
ECON
FIN

Arguments against pie charts

https://www.darkhorseanalytics.com/blog/salvaging-the-pie


2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47'
SOURCE:OPINIONS
DYNAMIC

Mode

Definition

The mode is the most common value. The mode may not be unique.

```
P BUS AG B  ACCT  MKT   BUS U A M D AN S  ECON  FIN
   65    24     3     3     2     1     1     1     1

 A  H  M  S
25  1 74  1

 Freshman Sophomore    Junior    Senior
        2        54        39         6
```

Definition

The median of an ordinal variable is the middle value when the values are ordered.

```
  [1] Freshman  Freshman  Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sopho
 [12] Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sopho
 [23] Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sopho
 [34] Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sopho
 [45] Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sophomore Sopho
 [56] Sophomore Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junio
 [67] Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junio
 [78] Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junior    Junio
 [89] Junior    Junior    Junior    Junior    Junior    Junior    Junior    Senior    Senior    Senior    Senio
[100] Senior    Senior
Levels: Freshman Sophomore Junior Senior
```

Contingency table

Definition

A contingency table shows the distribution of one variable in the rows and another in the columns.

	Year			
Major	Freshman	Sophomore	Junior	Senior
P BUS	2	33	27	3
AG B	0	12	10	2
ACCT	0	2	1	0
MKT	0	3	0	0
BUS U	0	2	0	0
A M D	0	1	0	0
AN S	0	0	1	0
ECON	0	0	0	1
FIN	0	1	0	0