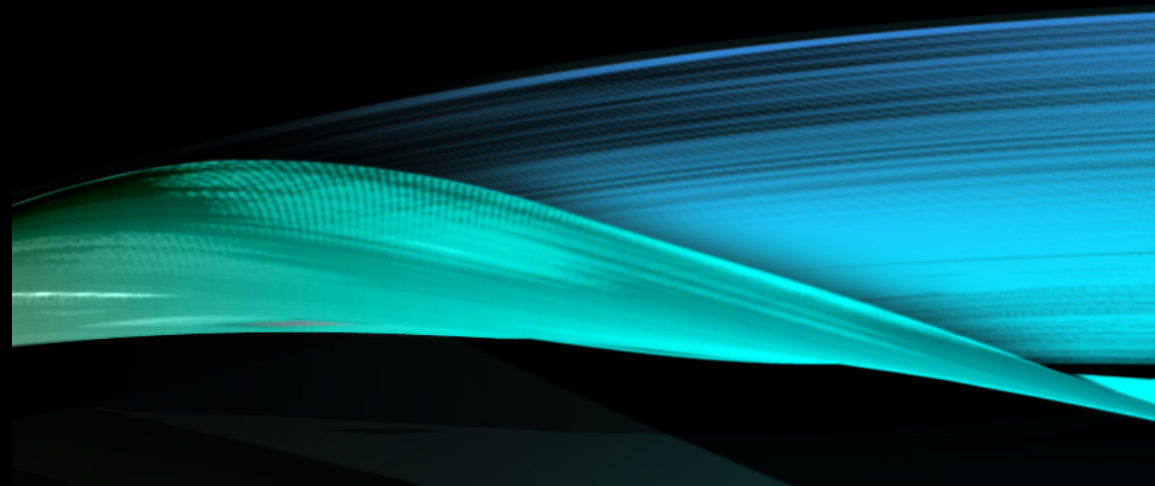
eZcZ Stat Network

Statistics Made EZ

2. CATEGORICAL DATA



Data Analysis, an Initial Approach

- Examine each variable in a data set individually and look for possible relationships among variables.
- Start with graphs and numerical summaries.

Example

- Data on medical malpractice (link in the comments).
- 118 observations in the original dataset.
- Variable of interest: Specialty (20 different observed)

	Amount	Severity	Age	Private Attorney	Marital Status	Specialty	Insurance	Gender
1	426500	6	37	Private	Married	Pediatrics	Private	Female
2	64000	4	66	Private	Married	Plastic Surgeon	Medicare/…	Female
3	17604	7	69	Private	Married	Internal Medicine	Private	Male
4	1986	3	56	Not Private	Married	Urological Surgery	Private	Female
5	41500	5	42	Private	Married	General Surgery	Unknown	Male
6	33708	7	69	Private	Married	General Surgery	Private	Male
7	177498	3	34	Private	Married	OBGYN	Private	Female
8	64000	4	45	Private	Single	Orthopedic Surgery	Private	Female
9	5250	5	42	Private	Married	Internal Medicine	Private	Female
10	28000	5	73	Private	Unknown	Ophthamology	Unknown	Female
11	9500	4	2	Private	Single	Emergency Medicine	Unknown	Female
12	4000	3	31	Private	Married	Orthopedic Surgery	Private	Male
13	38100	3	36	Private	Married	OBGYN	No Insurance	Female
14	34000	3	24	Private	Married	Anesthesiology	Unknown	Female
15	21500	5	42	Private	Divorced	Neurology/…	Private	Male
16	16500	3	29	Private	Married	Family Practice	No Insurance	Female
17	48500	3	61	Private	Married	Orthopedic Surgery	Workers …	Male
18	136500	9	49	Private	Married	General Surgery	Unknown	Female
19	376500	7	0	Private	Single	OBGYN	Private	Female
20	311500	3	28	Private	Married	General Surgery	Private	Female

Example

Level	Count	Prop
Anesthesiology	13	0.11017
Cardiology	4	0.03390
Dermatology	2	0.01695
Emergency Medicine	7	0.05932
Family Practice	17	0.14407
General Surgery	14	0.11864
Internal Medicine	8	0.06780
Neurology/Neurosurgery	7	0.05932
OBGYN	13	0.11017
Occupational Medicine	1	0.00847
Ophthamology	5	0.04237
Orthopedic Surgery	11	0.09322
Pathology	1	0.00847
Pediatrics	2	0.01695
Physical Medicine	1	0.00847
Plastic Surgeon	2	0.01695
Radiology	3	0.02542
Resident	3	0.02542
Thoracic Surgery	1	0.00847
Urological Surgery	3	0.02542
Total	118	1.00000

eZcZ Stat Network

Statistics Made EZ

Anesthesiology

$\frac{Anesthesiology}{Total}$ to obtain proportion (relative frequency)

The proportion column should add to 1

$$\frac{13}{118} = 0.11017$$

$\frac{Anesthesiology}{Total} * 100$ to obtain percentage

The percentage column should add to 100

0.11017*100%=11.017%

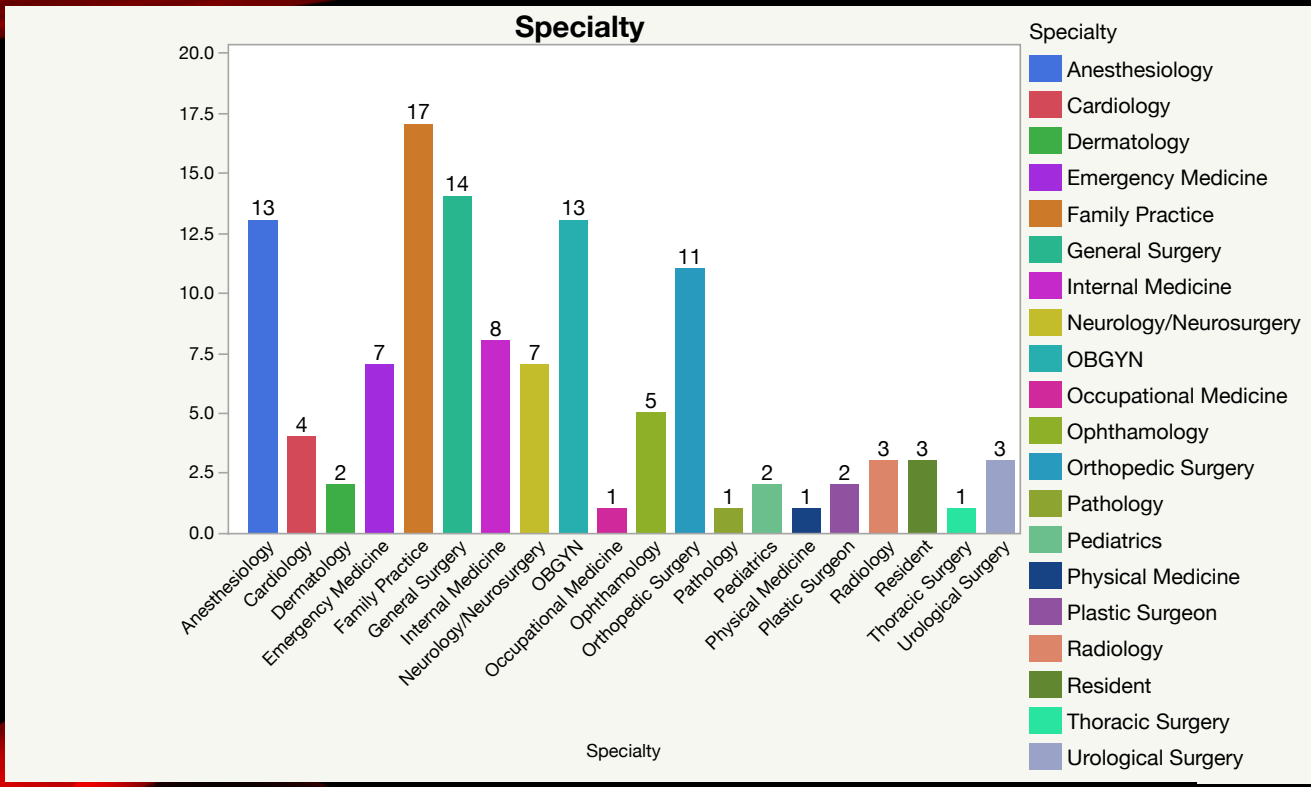
Bar charts

- A bar chart shows the amount of data that belong to each category as proportionally sized rectangular areas.
- Thus it is a display of the distribution of a categorical variable.
- Bar charts are more flexible than pie charts because we don't need to account for all possible categories of the variable.

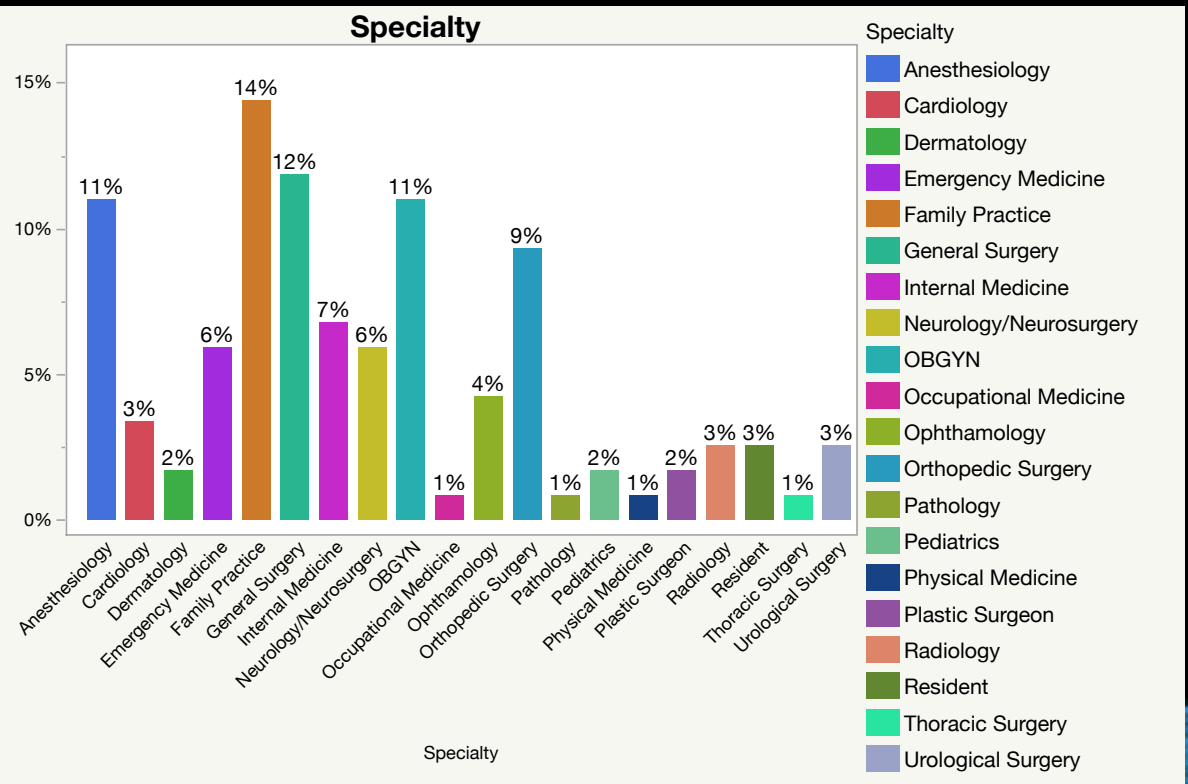
Bar charts

- Useful to display categorical data.
- Valuable presentation tools for reinforcing differences in magnitudes. (bars should be equally spaced).
- Bar charts can be horizontal or vertical.

Bar chart with Frequencies



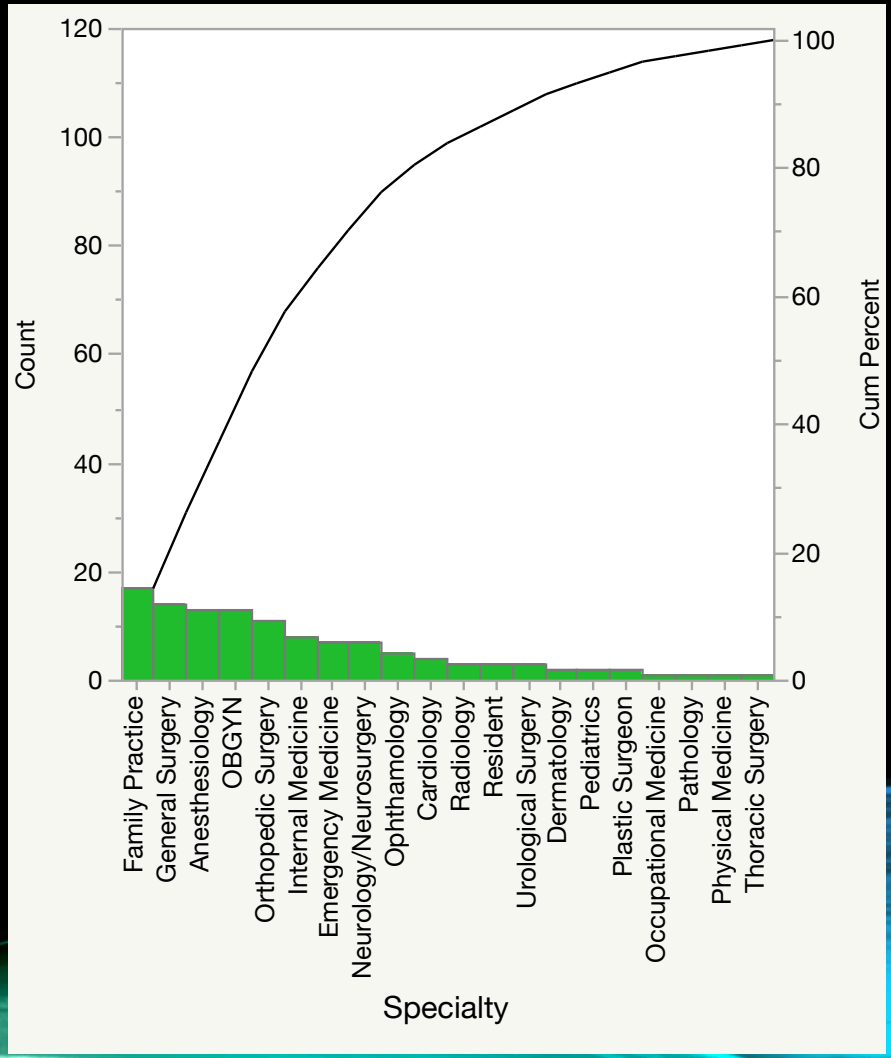
Bar chart with Percentages



eZcZ Stat Network
Statistics Made EZ

Pareto Charts

- Used for natural ordering (i.e., freshman, sophomore, …)
- Arrange bars with respect to order of magnitude.
- Quality control to identify problems in a business process.
- Cumulative line often omitted.

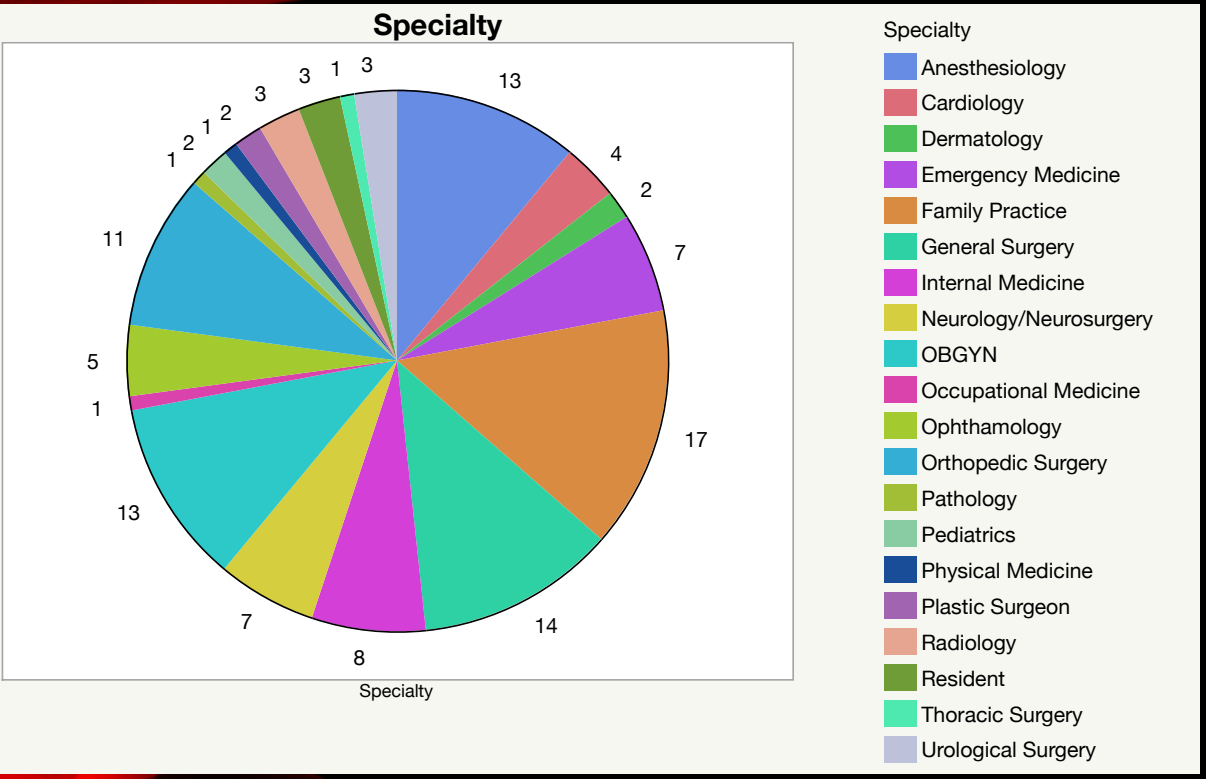


Pie Charts

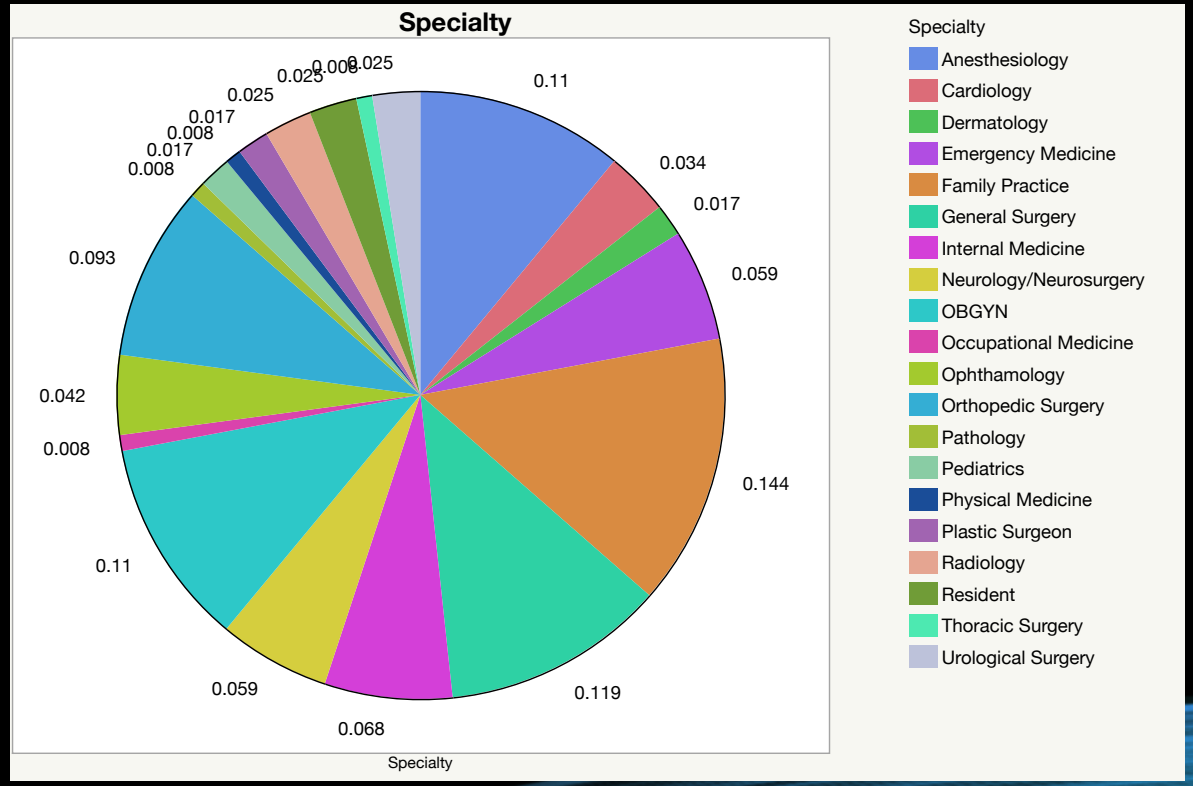
Shows the amount of data that belong to each category as a proportional slice of the circle.

- **Useful for one variable.**
- **Larger slices → Larger relative frequencies.**
- **Total percentage of all displayed categories equals 100.**

Pie Chart with Frequencies



Pie chart with Proportions



Mode

- The mode is the value of the variable that occurs the most.
- Most useful for categorical data with a relatively small number of possible values.
- When two or more categories tie for the highest frequency the distribution of data is multimodal.
- Example: Grades on a stat exam for some class.

A: 6 B:4 C:3 D:1 F:1

Median

- The median of an ordinal variable is the category of the middle observation of the sorted values.
- NOT appropriate for nominal data.

Example:

A: 6 B:4 C:3 D:1 F:1

A, A, A, A, A, A, B, B, B, B, C, C, C, D, F

Median=B

UP NEXT….

DESCRIBING NUMERICAL DATA

SUBSCRIBE