Welcome to STA 101!

9/10/2024 checkpoint

Statistics is a confrontation with uncertainty.

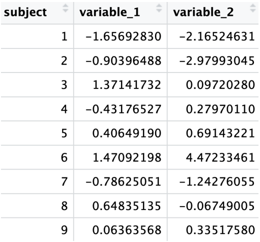
Statistics confronts uncertainty by quantifying it.

Data analysis

Transforming messy, incomplete, imperfect data into knowledge.

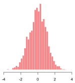
What form does that knowledge usually take?

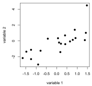
- pictures;
- a concise set of numerical summaries.

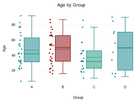


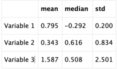
subject	variable_1	variable_2
1	−1.65692830	−2.16524631
2	−0.90396488	−2.97993045
3	1.37141732	0.09720280
4	−0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	−0.78625051	−1.24276055
8	0.64835135	−0.06749005
9	0.06363568	0.33517580

⟹







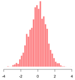


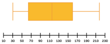
	mean	median	std
Variable 1	0.795	−0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

What kind of picture do I make?

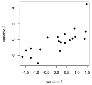
It depends on the data type and the question:

One numerical variable

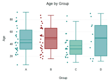




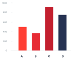
Two numerical variables



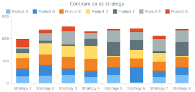
Many numerical variables



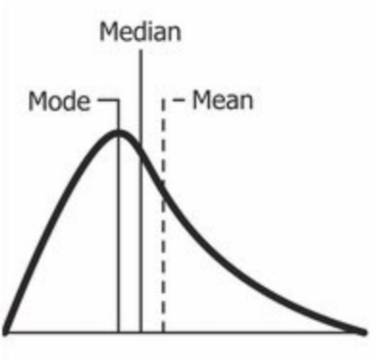
One categorical variable



Two categorical variables



What kind of summaries do I compute?



- **Center**: mean, median, mode
- **Spread**: standard deviation
- **Association strength**: correlation coefficient

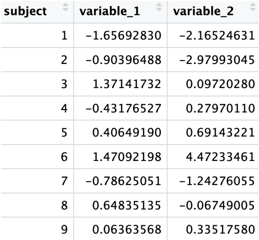
And on and on.

Data analysis

Transforming messy, incomplete, imperfect data into knowledge.

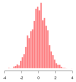
What form does that knowledge usually take?

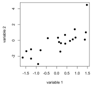
- pictures;
- a concise set of numerical summaries.

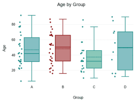


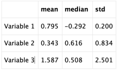
subject	variable_1	variable_2
1	−1.65692830	−2.16524631
2	−0.90396488	−2.97993045
3	1.37141732	0.09720280
4	−0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	−0.78625051	−1.24276055
8	0.64835135	−0.06749005
9	0.06363568	0.33517580

$\Longrightarrow$







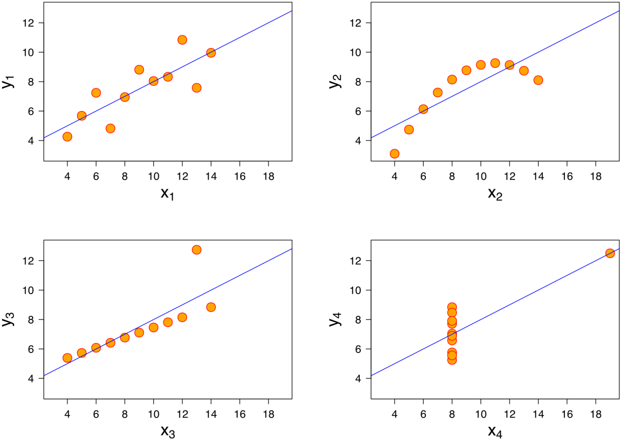


	mean	median	std
Variable 1	0.795	−0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

Theme: pictures and summaries need to work together!

Anscombe's quartet

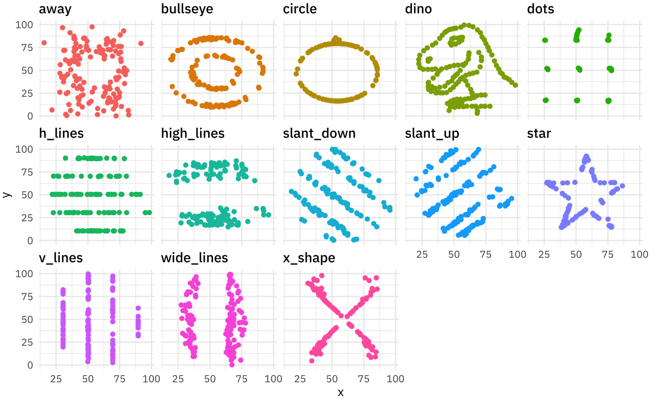
These all have the same basic summary statistics:



ABV: **A**lways **B**e **V**isualizing

DatasauRus dozen (Lab 1)

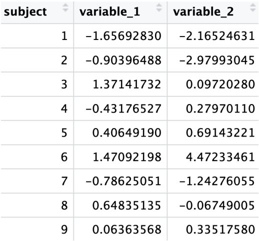
These all have similar summary statistics:



ABV: **A**lways **B**e **V**isualizing

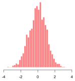
Data analysis

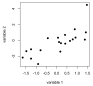
Transforming messy, incomplete, imperfect data into knowledge.

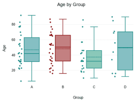


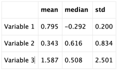
subject	variable_1	variable_2
1	−1.65692830	−2.16524631
2	−0.90396488	−2.97993045
3	1.37141732	0.09720280
4	−0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	−0.78625051	−1.24276055
8	0.64835135	−0.06749005
9	0.06363568	0.33517580

$\Longrightarrow$









	mean	median	std
Variable 1	0.795	−0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

Statistical inference

Quantifying uncertainty about that knowledge.

Statistical inference

You ask a quantitative question:

- What is the "typical" lead level in the Flint MI drinking water?
- What is the probability that Kamala Harris wins the 2024 presidential election?
- How many jobs does a $1.00 increase in the minimum wage create or destroy?

The answer would take the form of a single number.

Statistical inference

Question: What's the number?

Answer: use the data to come up with a best guess.

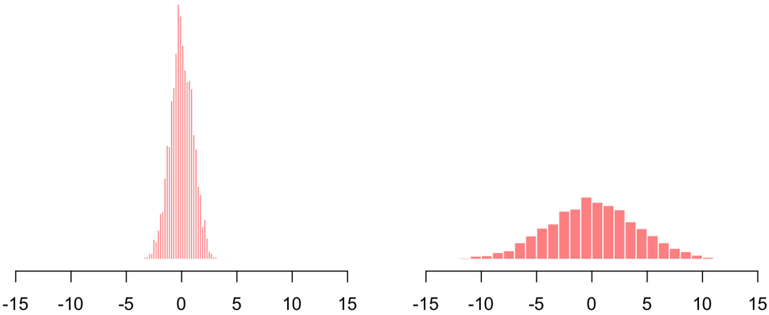
Statistics: Compute a *margin of error* for the guess:

$$\text{best-guess} \pm \text{margin-of-error.}$$

- Gives a *range* of likely values, not just a single guess;
- The size of the margin quantifies uncertainty.

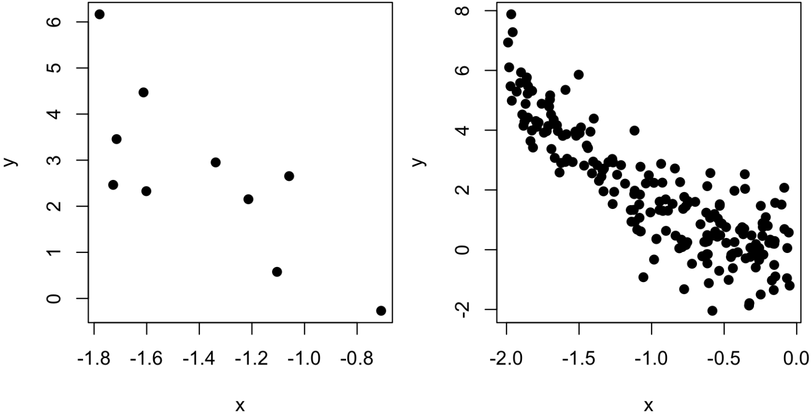
But where does the margin of error come from?

Which dataset provides stronger conclusions?



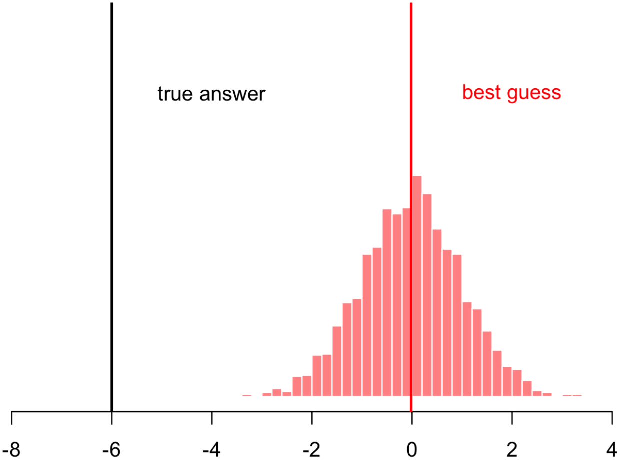
The one that is less (more) variable might give lower (higher) margin of error.

Which dataset provides stronger conclusions?



The one that is bigger (smaller) might give lower (higher) margin of error.

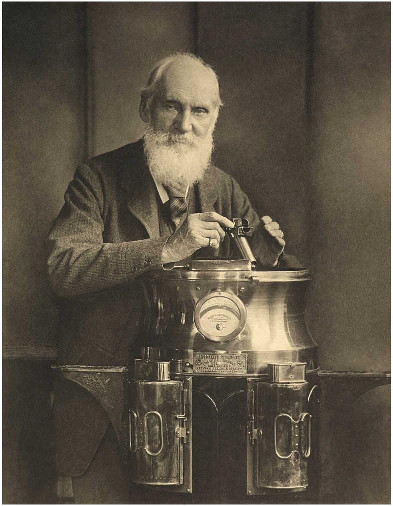
What if this happens to you?



Two themes:

- you *need* domain knowledge;
- beware a false sense of precision.

Beware false precision

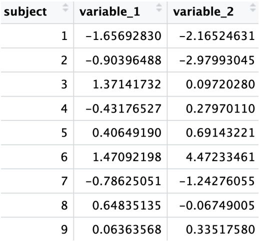


"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind." - Lord Kelvin

- **Maybe**: without quantification, you cannot fully understand;
- **But also**: just because you are quantifying does not mean you understand.

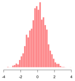
Data analysis

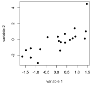
Transforming messy, incomplete, imperfect data into knowledge.

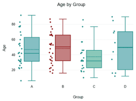


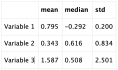
subject	variable_1	variable_2
1	−1.65692830	−2.16524631
2	−0.90396488	−2.97993045
3	1.37141732	0.09720280
4	−0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	−0.78625051	−1.24276055
8	0.64835135	−0.06749005
9	0.06363568	0.33517580

$\Longrightarrow$









	mean	median	std
Variable 1	0.795	−0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

Statistical inference

Quantifying uncertainty about that knowledge:

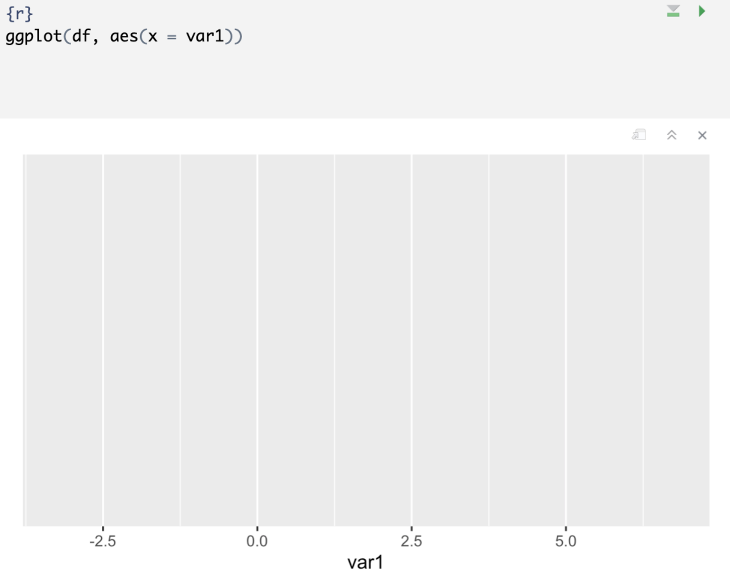
best-guess ± margin-of-error

Margin based on sample size, data variability, etc.

But how do you actually *do* these things?

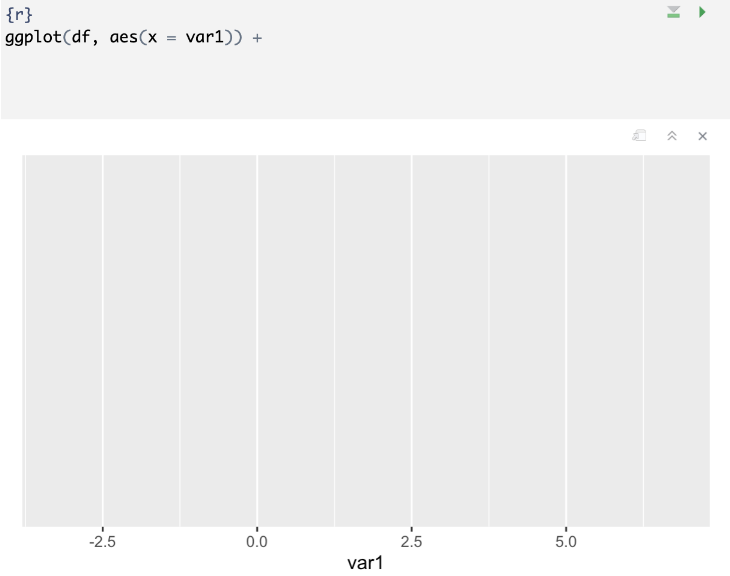
- Use software like `R`/`RStudio`
  - the learning curve is steep;
  - people actually use this in the “real world.”
- There are two main skills we need to master:
  - `ggplot` layering;
  - piping (|>).

`ggplot`: building plots in layers



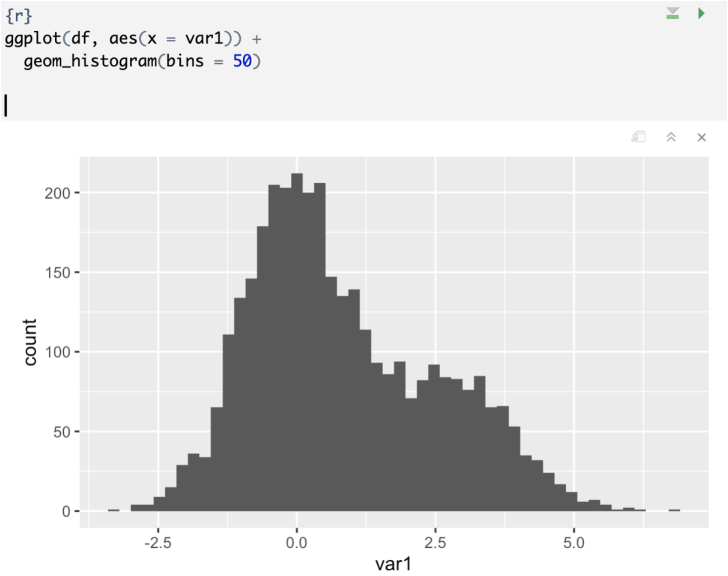
```
{r}
ggplot(df, aes(x = var1))
```

ggplot: building plots in layers



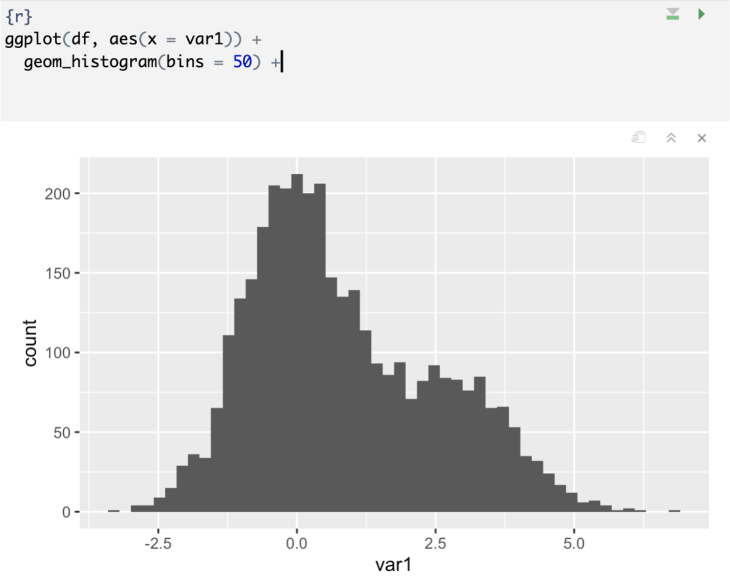
```
{r}
ggplot(df, aes(x = var1)) +
```

ggplot: building plots in layers



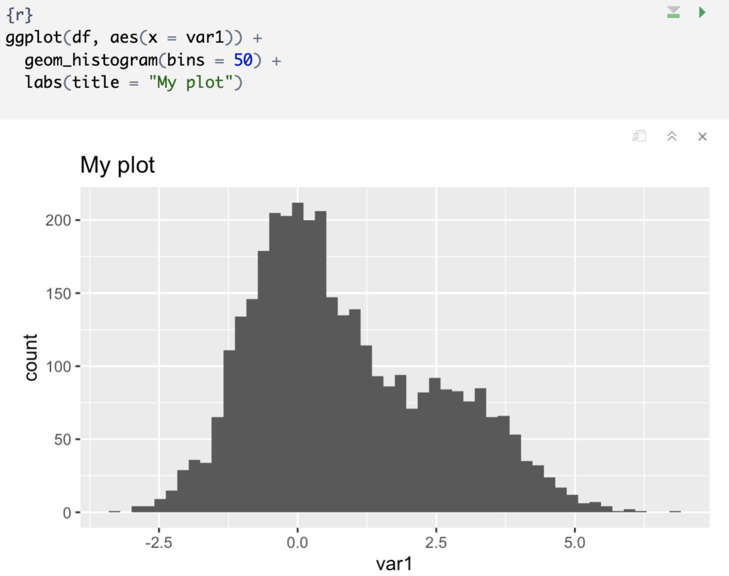
```
{r}
ggplot(df, aes(x = var1)) +
  geom_histogram(bins = 50)
```

ggplot: building plots in layers



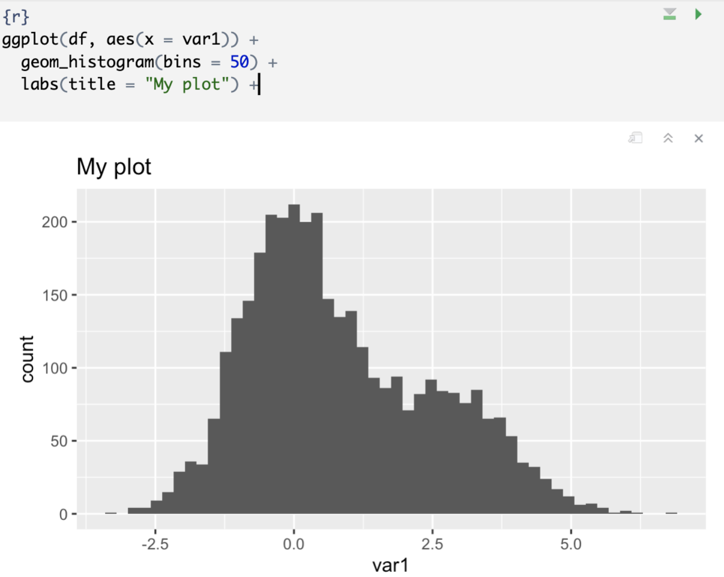
```
{r}
ggplot(df, aes(x = var1)) +
  geom_histogram(bins = 50) +
```

`ggplot`: building plots in layers



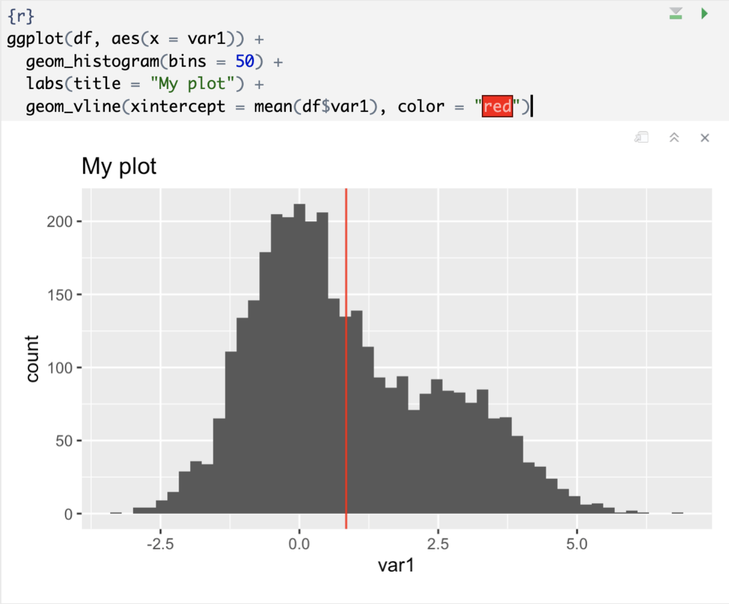
```
{r}
ggplot(df, aes(x = var1)) +
  geom_histogram(bins = 50) +
  labs(title = "My plot")
```

ggplot: building plots in layers



```
{r}
ggplot(df, aes(x = var1)) +
  geom_histogram(bins = 50) +
  labs(title = "My plot") +
```

ggplot: building plots in layers

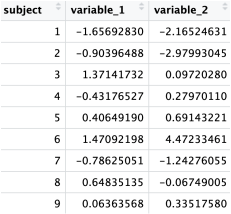


```
{r}
ggplot(df, aes(x = var1)) +
  geom_histogram(bins = 50) +
  labs(title = "My plot") +
  geom_vline(xintercept = mean(df$var1), color = "red")
```

ggplot

pipes |>

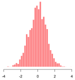
Data analysis

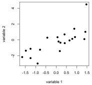


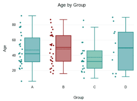
subject	variable_1	variable_2
1	−1.65692830	−2.16524631
2	−0.90396488	−2.97993045
3	1.37141732	0.09720280
4	−0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	−0.78625051	−1.24276055
8	0.64835135	−0.06749005
9	0.06363568	0.33517580

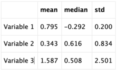
`ggplot` ⟹

`|>` ⟹









	mean	median	std
Variable 1	0.795	−0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

Statistical inference

Question: What's the number?

Answer: best-guess ± margin-of-error