Introductory Statistics
4: Data Editing and Summary

Richard Veale

Graduate School of Medicine
Kyoto University

https://youtu.be/QAG-glTTXds

Lecture Video at above link

Summary

Types of data (categorical, continuous, etc.)

Measures of location (mean, median, mode)

Histograms and Distributions

Box Plots

Outliers, Data Trimming

Measure of spread (variation)

Scatter Plots

Types of Data

What type of data is this?

Person	Like Ramen?	You a man?
1	Yes	Yes
2	Yes	Yes
3	Yes	No
4	No	Yes
5	No	No
6	No	Yes
7	Yes	Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes

Are you Man?

Like Ramen?	Yes	No
Yes	5	1
No	3	1

Types of Data

What type of data is this?

Person	Like Ramen?	You a man?
1	Yes	Yes
2	Yes	Yes
3	Yes	No
4	No	Yes
5	No	No
6	No	Yes
7	Yes	Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes

Like Ramen?

	Are you Man?	
	Yes	No
Yes	5	1
No	3	1

Types of Data

Categorical Data

Person	Like Ramen?	You a man?
1	Yes	Yes
2	Yes	Yes
3	Yes	No
4	No	Yes
5	No	No
6	No	Yes
7	Yes	Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes

Are you Man?

Like Ramen?	Yes	No
Yes	5	1
No	3	1

Types of Data

We choose to represent membership in the categories using symbols (“Yes” or “No”)

Person	Like Ramen?	You a man?
1	Yes	Yes
2	Yes	Yes
3	Yes	No
4	No	Yes
5	No	No
6	No	Yes
7	Yes	Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes

Are you Man?

Like Ramen?	Yes	No
Yes	5	1
No	3	1

Types of Data

We could use any symbol….
(1 for Yes, 0 for No). This works just the same!

Person	Like Ramen?	You a man?
1	1	1
2	1	1
3	1	0
4	0	1
5	0	0
6	0	1
7	1	1
8	1	1
9	0	1
10	1	1

Are you Man?

Like Ramen?	1	0
1	5	1
0	3	1

Types of Data

We could use “a” for Yes, “b” for no…
We could use 0 for Yes, 1 for No!!!
We could name our columns anything we want.

Person	Booboo?	BaaBaa?
1	0	0
2	0	0
3	0	1
4	1	0
5	1	1
6	1	0
7	0	0
8	0	0
9	1	0
10	0	0

	BaaBaa?	
Booboo?	0	1
0	5	1
1	3	1

Types of Data

We could use “a” for Yes, “b” for no…
We could use 0 for Yes, 1 for No!!!
We could name our columns anything we want.

Person	Booboo?	BaaBaa?
1	0	0
2	0	0
3	0	1
4	1	0
5	1	1
6	1	0
7	0	0
8	0	0
9	1	0
10	0	0

It is confusing…which is why we use “meaningful” names.

	BaaBaa?	
Booboo?	0	1
0	5	1
1	3	1

Types of Data

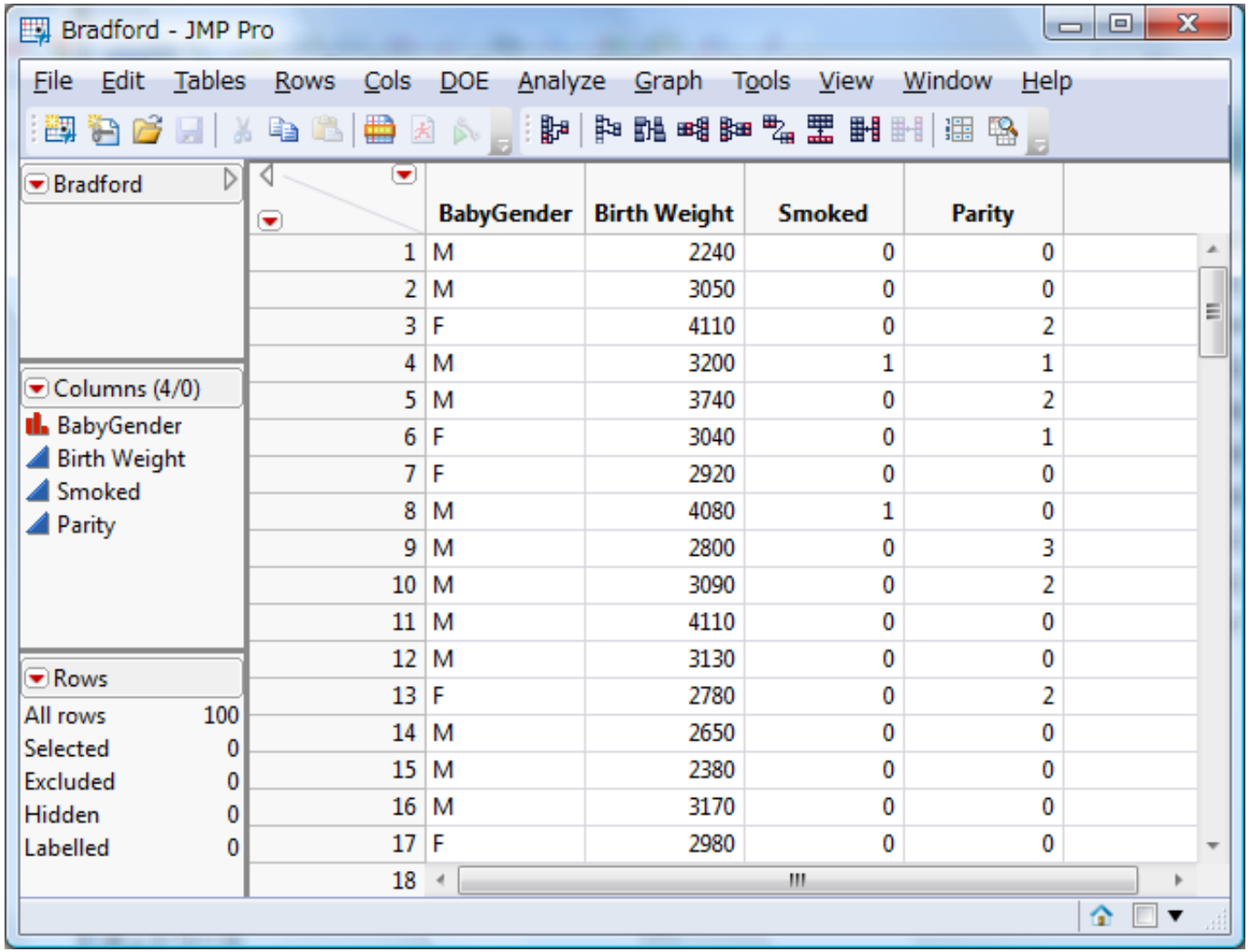
We could use “a” for Yes, “b” for no…
We could use 0 for Yes, 1 for No!!!
We could name our columns anything we want.

Person	Booboo?	BaaBaa?
1	0	0
2	0	0
3	0	1
4	1	0
5	1	1
6	1	0
7	0	0
8	0	0
9	1	0
10	0	0

People often think of numbers as representing “amounts”… don’t get confused! These are just symbols!

		BaaBaa?	
		0	1
Booboo?	0	5	1
	1	3	1

Types of Data

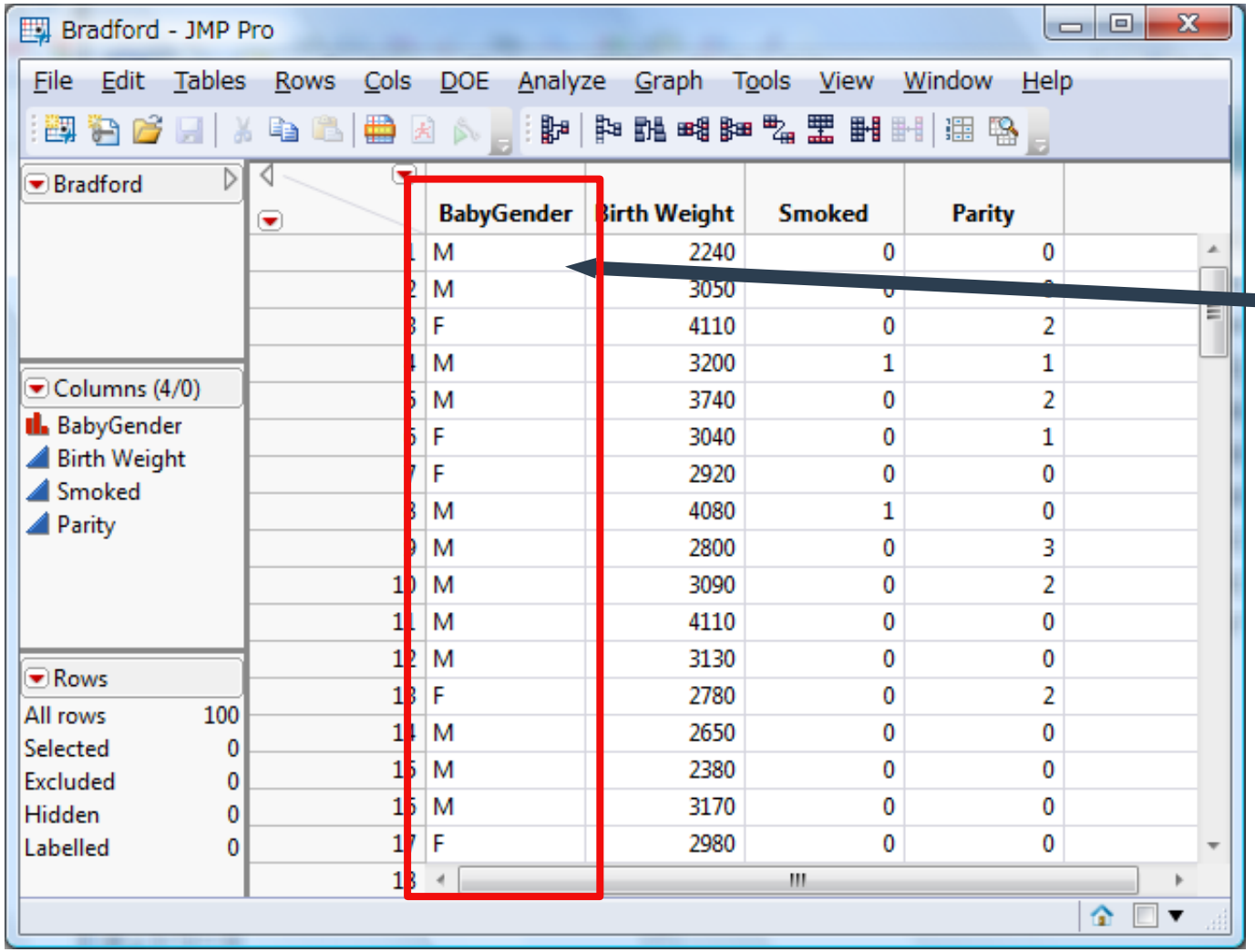


	BabyGender	Birth Weight	Smoked	Parity
1	M	2240	0	0
2	M	3050	0	0
3	F	4110	0	2
4	M	3200	1	1
5	M	3740	0	2
6	F	3040	0	1
7	F	2920	0	0
8	M	4080	1	0
9	M	2800	0	3
10	M	3090	0	2
11	M	4110	0	0
12	M	3130	0	0
13	F	2780	0	2
14	M	2650	0	0
15	M	2380	0	0
16	M	3170	0	0
17	F	2980	0	0
18				

What about this data?

What “types” of data do you see…?

Types of Data



	BabyGender	Birth Weight	Smoked	Parity
1	M	2240	0	0
2	M	3050	0	0
3	F	4110	0	2
4	M	3200	1	1
5	M	3740	0	2
6	F	3040	0	1
7	F	2920	0	0
8	M	4080	1	0
9	M	2800	0	3
10	M	3090	0	2
11	M	4110	0	0
12	M	3130	0	0
13	F	2780	0	2
14	M	2650	0	0
15	M	2380	0	0
16	M	3170	0	0
17	F	2980	0	0
18				

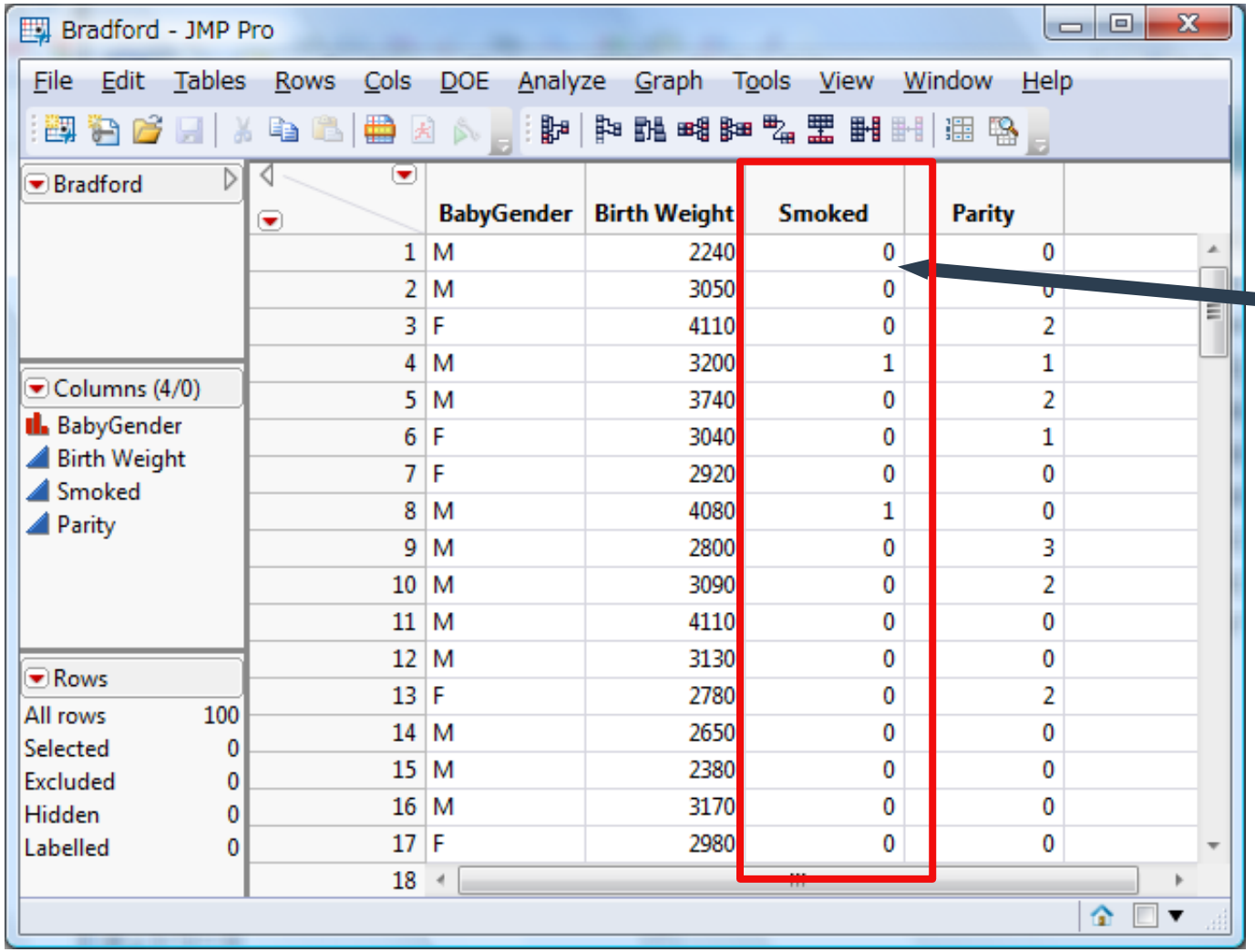
What about this data?

BabyGender
This looks like our "man or not a man".

Probably a category.

Uses "M" and "F" for the symbols for the 2 groups...

Types of Data



	BabyGender	Birth Weight	Smoked	Parity
1	M	2240	0	0
2	M	3050	0	0
3	F	4110	0	2
4	M	3200	1	1
5	M	3740	0	2
6	F	3040	0	1
7	F	2920	0	0
8	M	4080	1	0
9	M	2800	0	3
10	M	3090	0	2
11	M	4110	0	0
12	M	3130	0	0
13	F	2780	0	2
14	M	2650	0	0
15	M	2380	0	0
16	M	3170	0	0
17	F	2980	0	0
18				

What about this data?

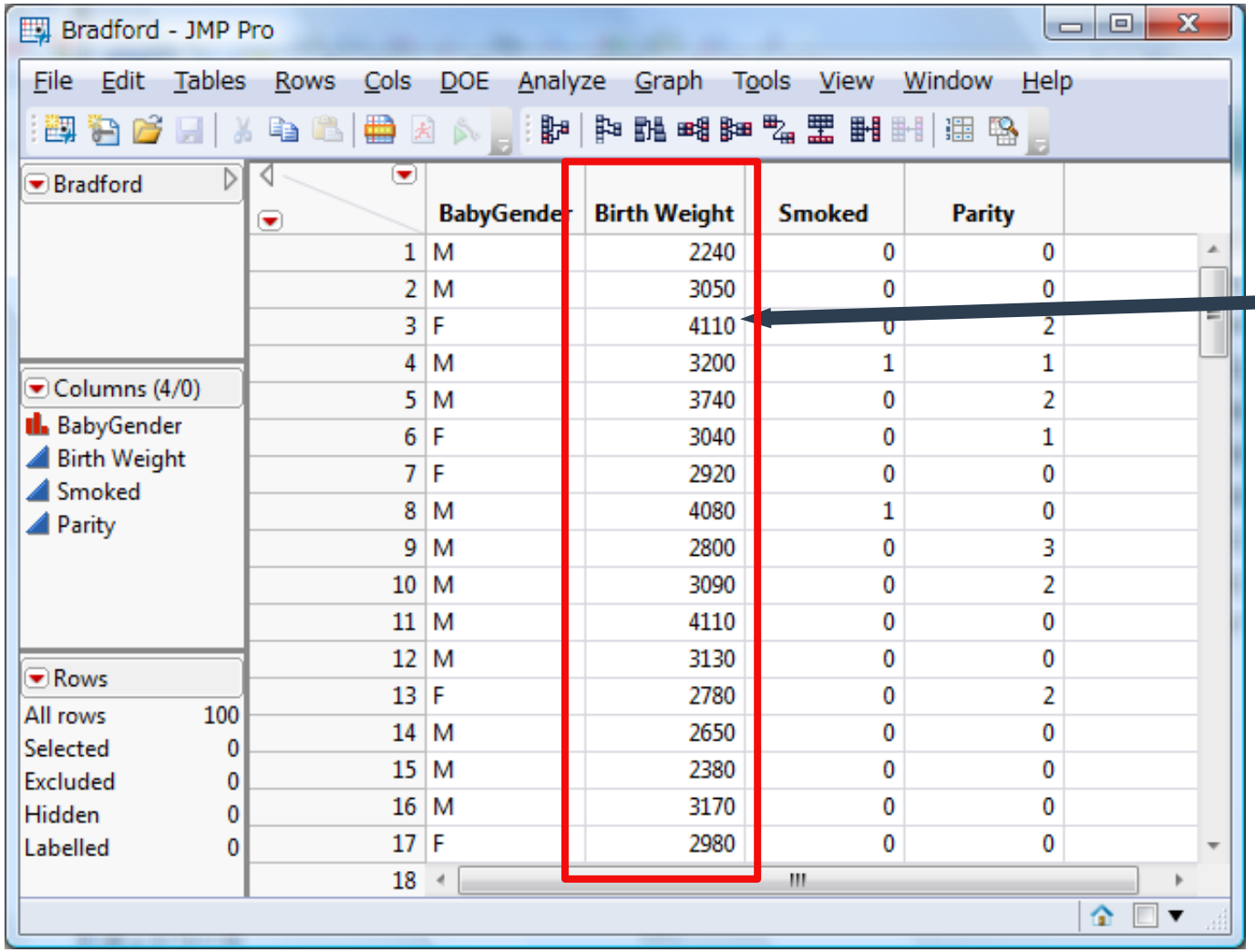
Smoked
This looks similar too.

Did X smoke.

We used “Yes” and “No”. It looks like they are using “1” and “0”

(this is common).

Types of Data



	BabyGender	Birth Weight	Smoked	Parity
1	M	2240	0	0
2	M	3050	0	0
3	F	4110	0	2
4	M	3200	1	1
5	M	3740	0	2
6	F	3040	0	1
7	F	2920	0	0
8	M	4080	1	0
9	M	2800	0	3
10	M	3090	0	2
11	M	4110	0	0
12	M	3130	0	0
13	F	2780	0	2
14	M	2650	0	0
15	M	2380	0	0
16	M	3170	0	0
17	F	2980	0	0
18				

What about this data?

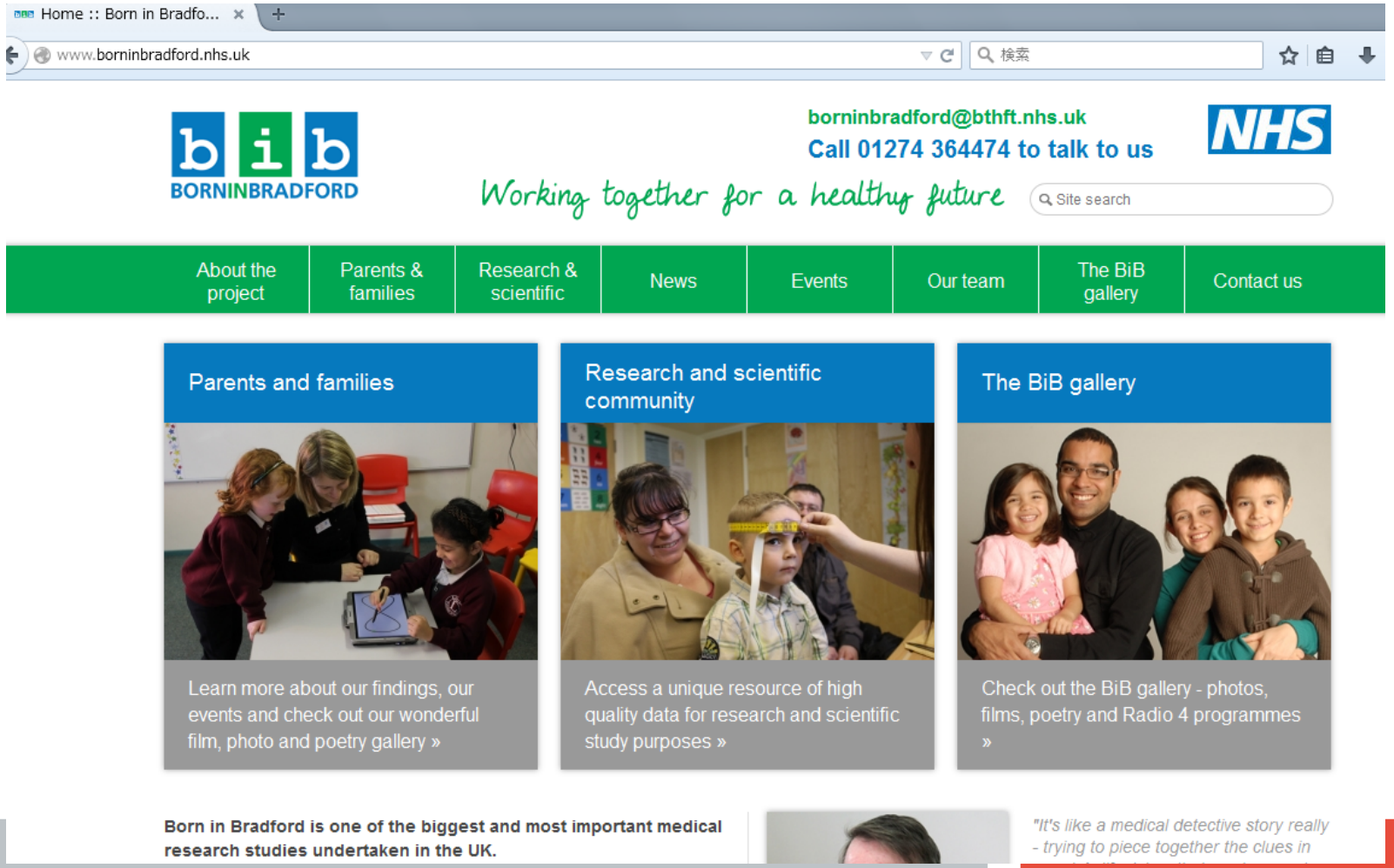
Birth Weight
This looks different!

In this case, the numbers *actually represent numbers.*

(how much the baby weighed at birth in grams?)

This is data from “Born in Bedford”

This study follows the health of around 13,500 babies born in Bradford between 2007-2010. Did mother smoke or not, etc…



Types of Data

Nominal data

Categorical data, cannot be put in any specific order.
E.g., gender, hair color. **Do you smoke? Yes or No.**

Ordinal data

Categorical data, can be ordered, but the numerical scale is arbitrary (differences between categories unknown).
E.g., pain on a 10-point scale, school grades. **Do you smoke (a) none (b) some (c) a lot?**

Continuous data

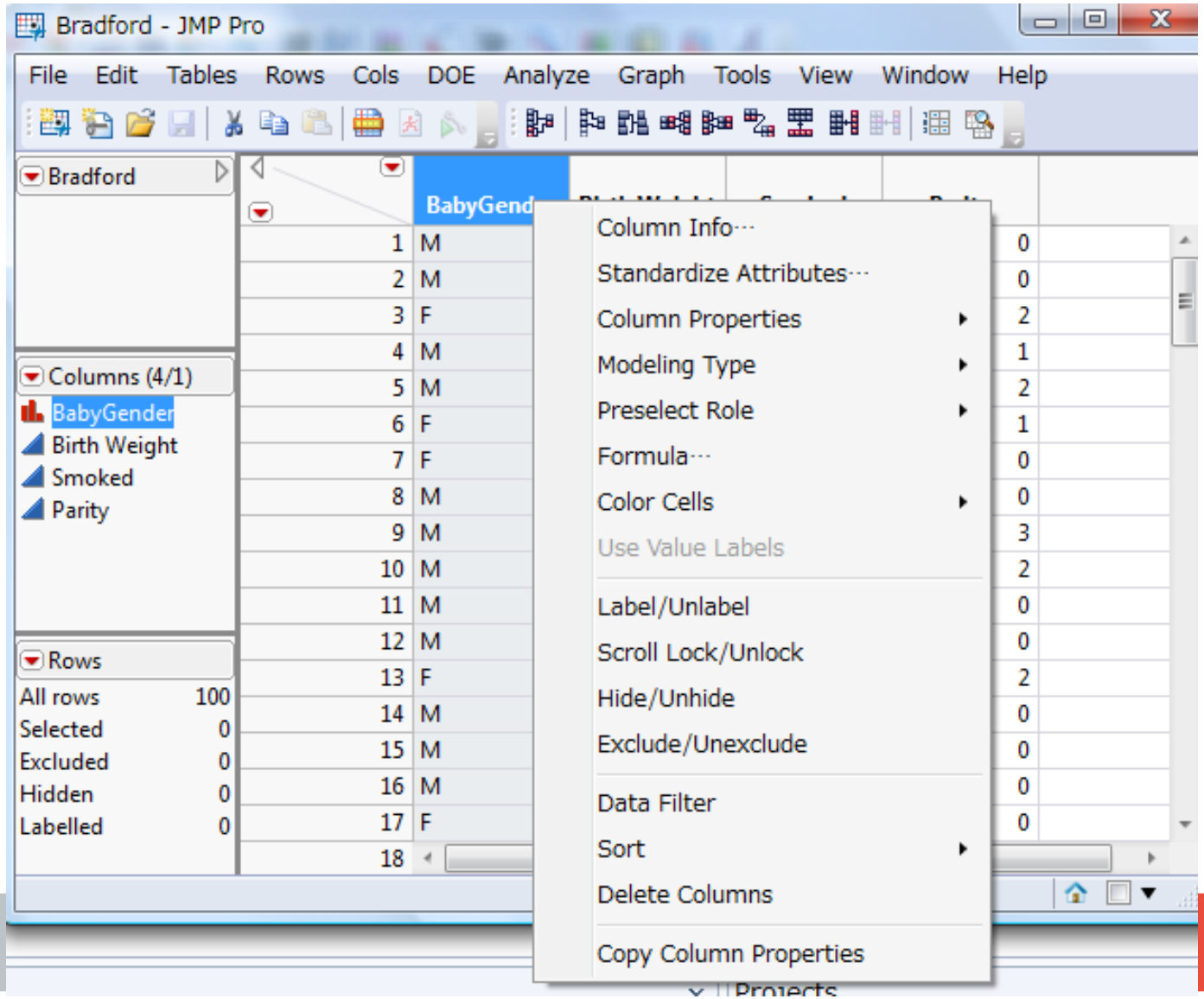
Metric data.
E.g., body temperature, height, air plane speed. **How many grams do you smoke?**

Discrete data

Also metric, but in integers.
E.g., age in years, number of children. **How many cigarettes per day?**

In JMP….

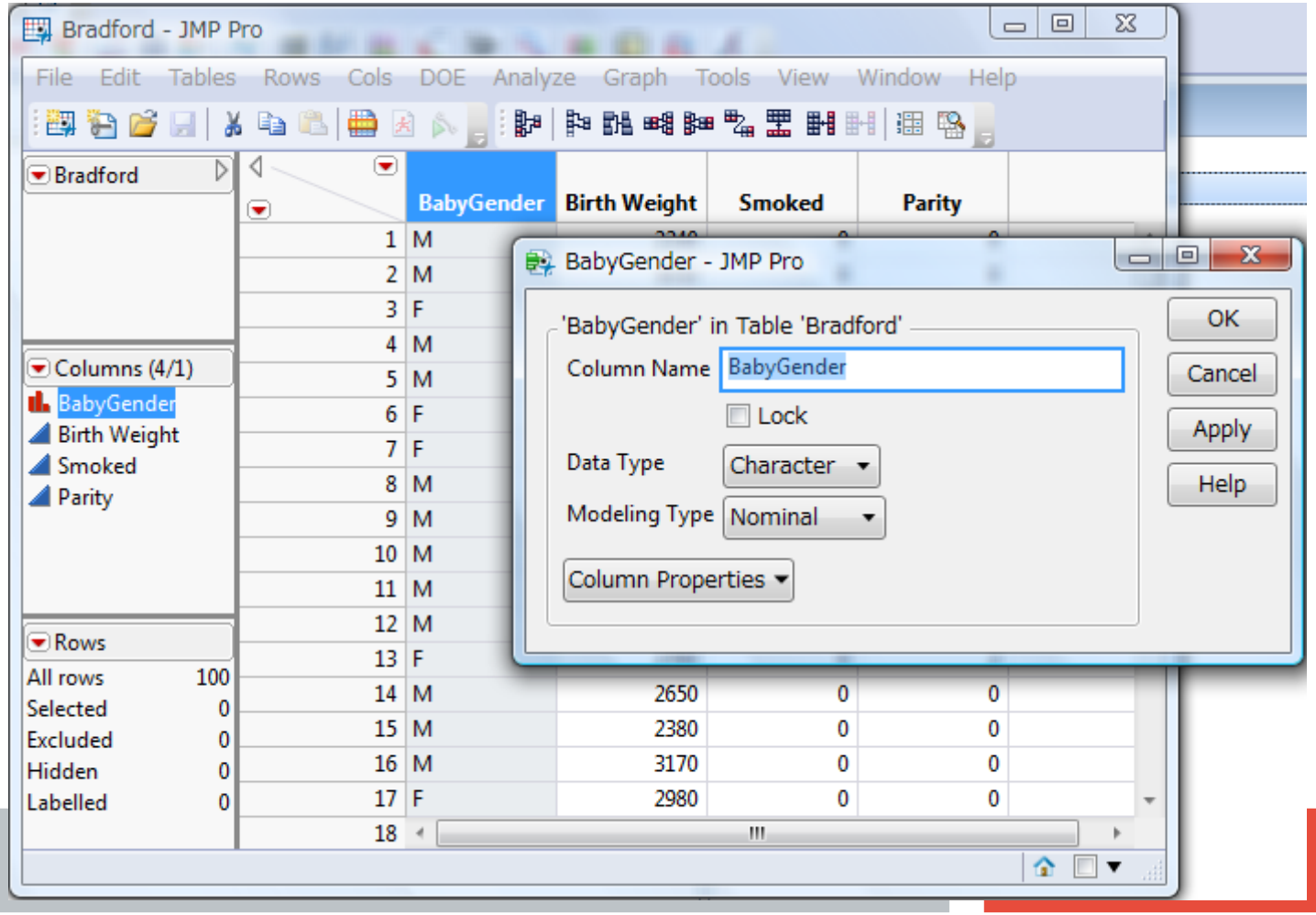
Right-click on column, set “data type”



In JMP….

→ “Data type” refers to the (computer) representation of the data.

→ “Modeling type” refers to should it be a category (nominal), or a continuous number, or a ranked ordering?



Measures of Location

$$\hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

“**Average**”:

Arithmetic Mean (of your sample…)

Measures of Location

$$\hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Doesn’t make sense for categorical (or ordinal) data.

→ What is the average of “red hair”, “blond hair” and “brown hair”?
→ What is the average of “none” and “sometimes”?

Median

Median

1) Sort the n observations from smallest to largest

2a) The median is the single middle value if n is odd.
2b) The median is the average of the two middle values if n is even.

E.g.:

Mode

Mode

Most frequent value in observations.

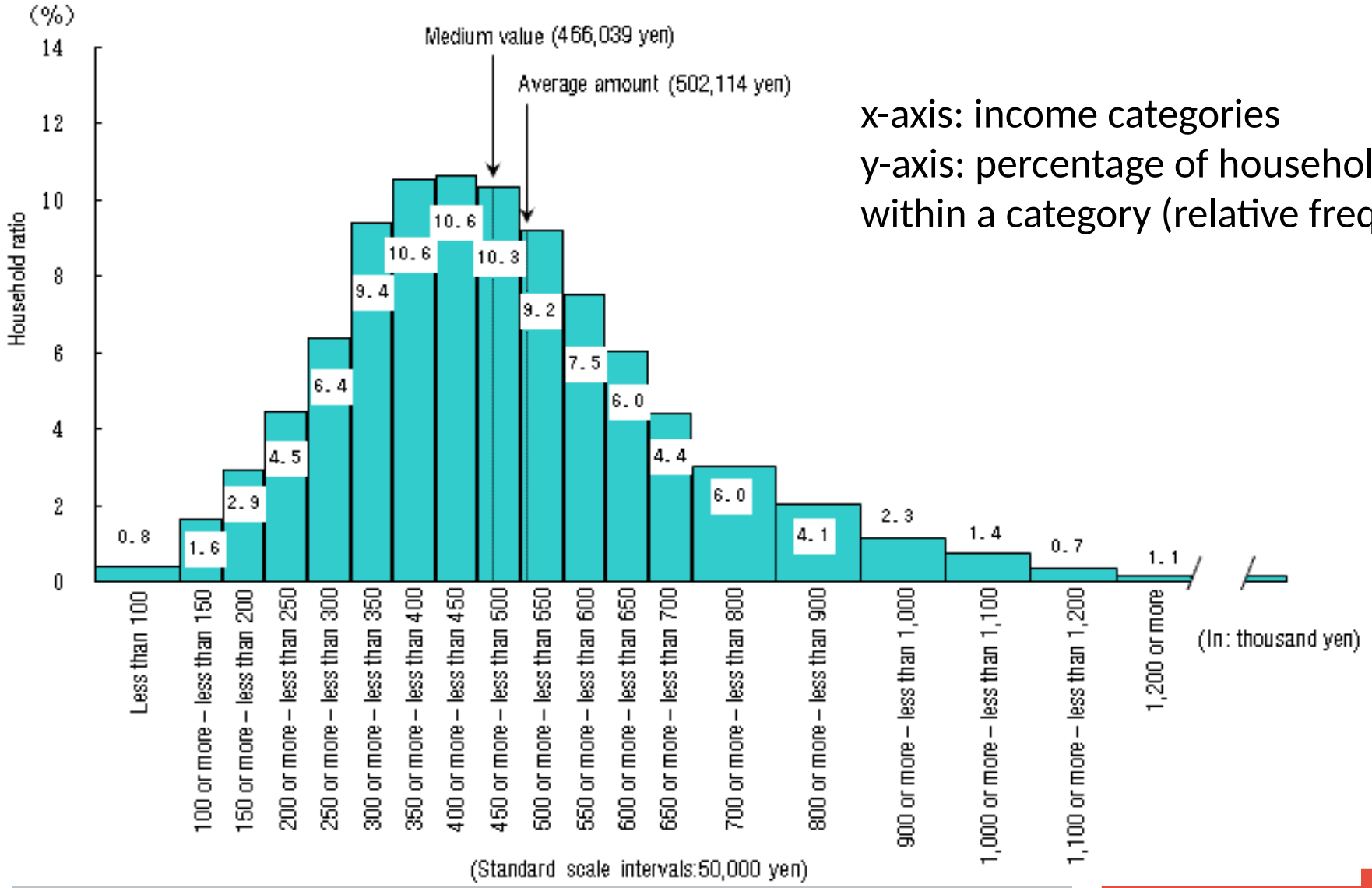
E.g.:

Data: 1 1 2 3 5 5 5 6 6 8

Mode = 5

Can be used for categorical, ordinal, and metric (continuous, discrete) data.

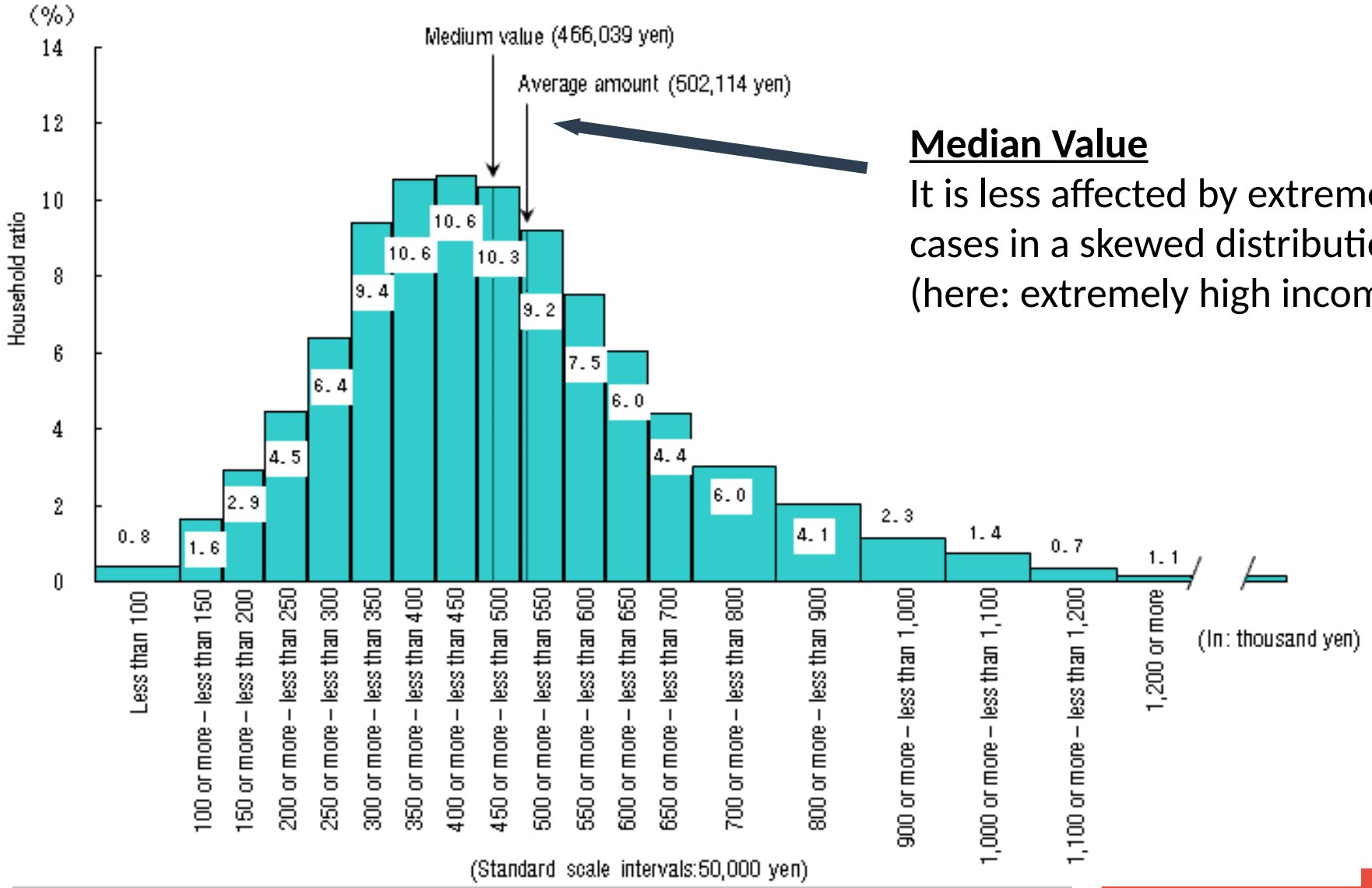
Visualizing Data: Histogram



x-axis: income categories
y-axis: percentage of households within a category (relative frequency)

Household income in Japan (www.stat.go.jp, 2004)

Visualizing Data: Histogram

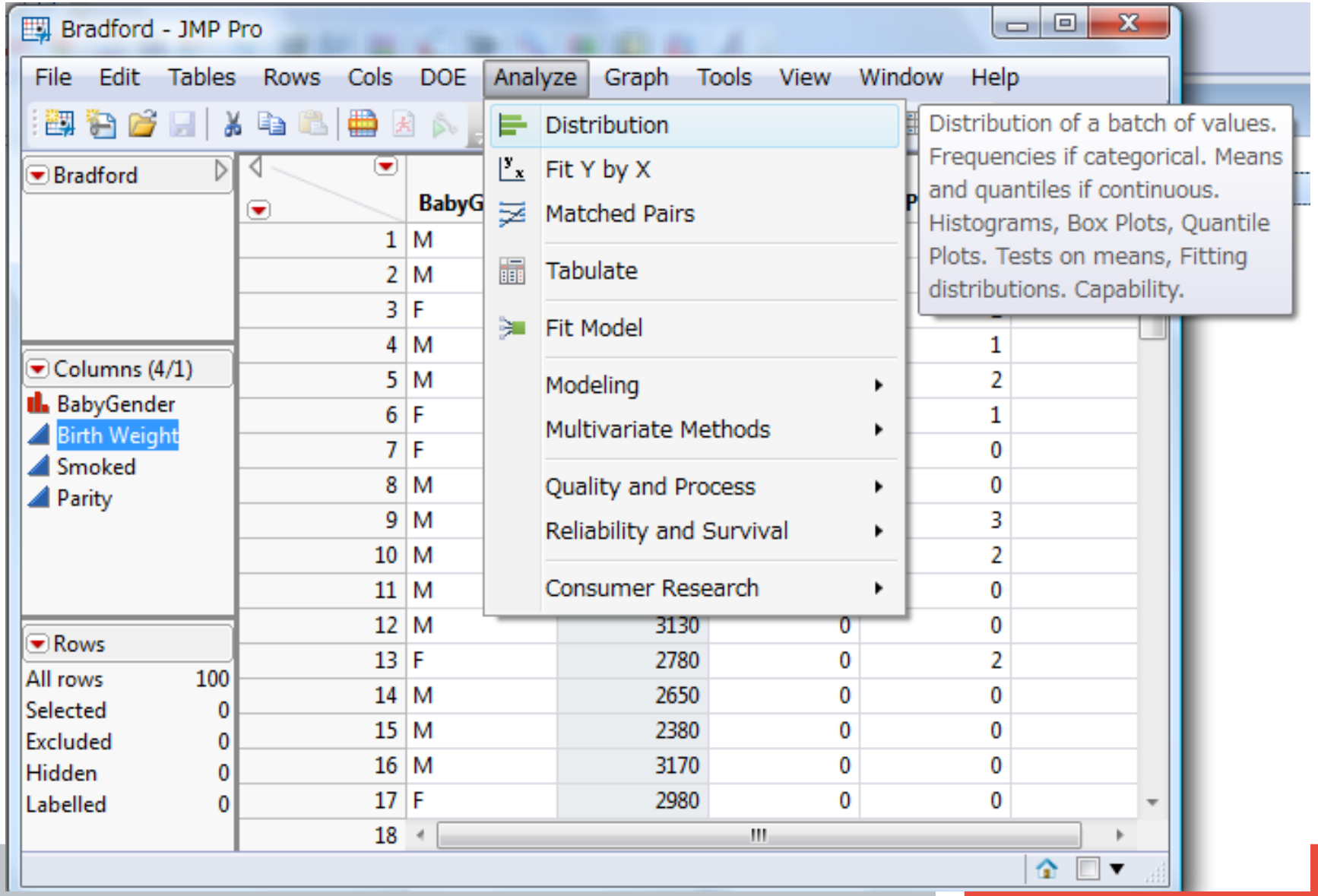


Median Value

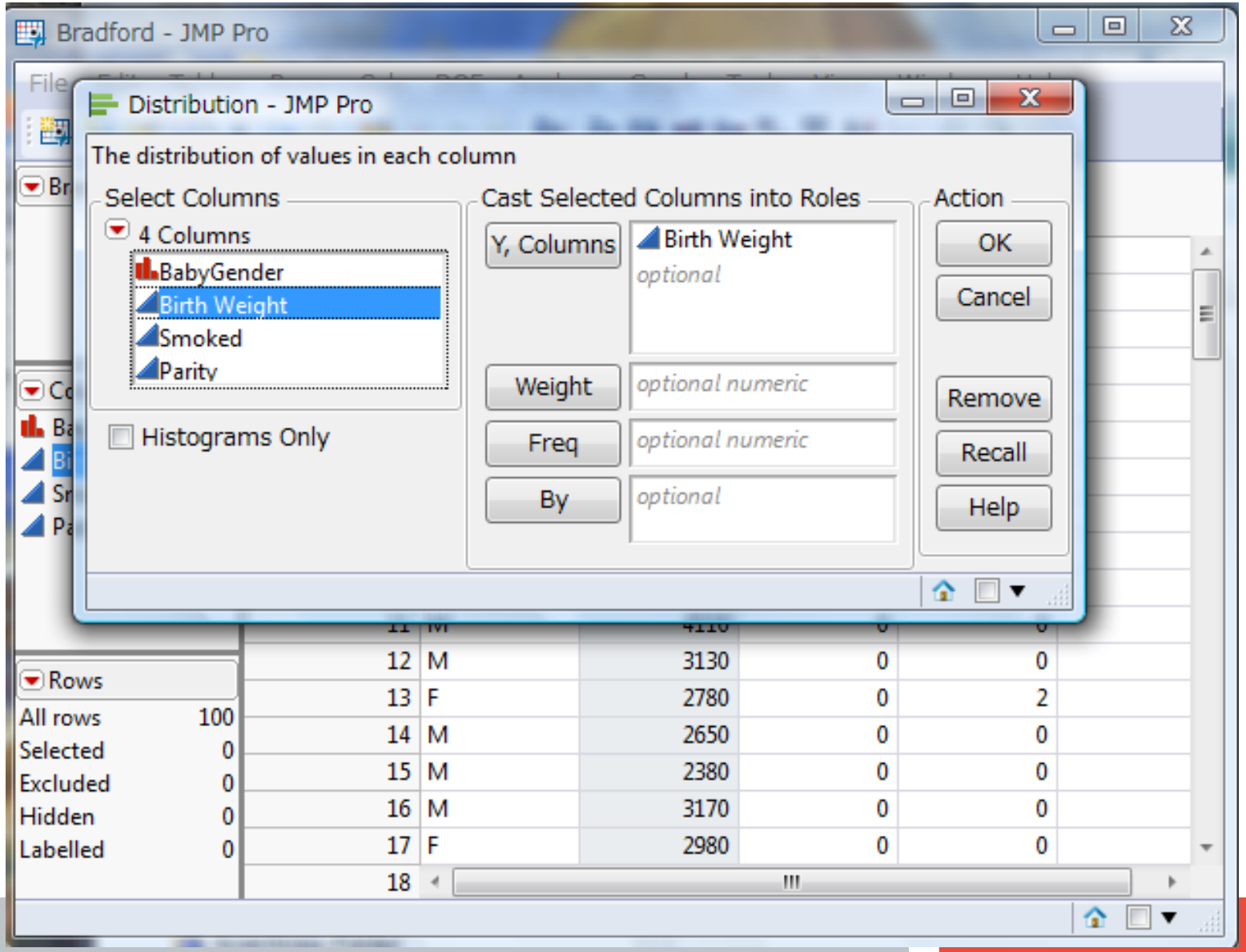
It is less affected by extreme cases in a skewed distributions (here: extremely high incomes)

Household income in Japan (www.stat.go.jp, 2004)

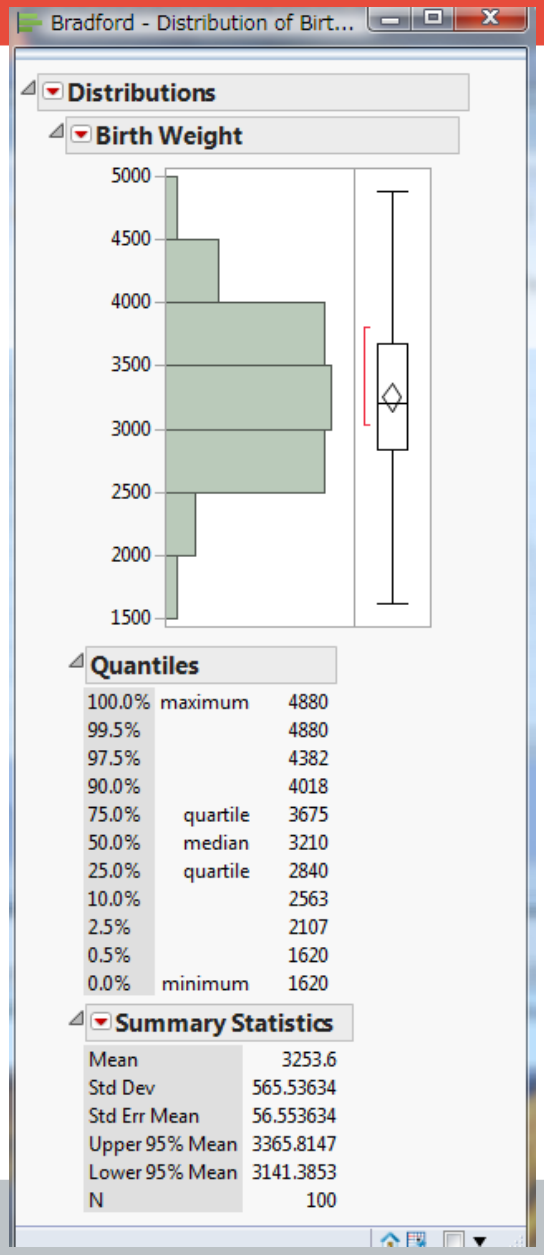
In JMP


Bradford - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Distribution
Fit Y by X
Matched Pairs
Tabulate
Fit Model
Modeling
Multivariate Methods
Quality and Process
Reliability and Survival
Consumer Research
Distribution of a batch of values. Frequencies if categorical. Means and quantiles if continuous. Histograms, Box Plots, Quantile Plots. Tests on means, Fitting distributions. Capability.
Bradford
Columns (4/1)
BabyGender
Birth Weight
Smoked
Parity
Rows
All rows 100
Selected 0
Excluded 0
Hidden 0
Labelled 0

In JMP


Bradford - JMP Pro
Distribution - JMP Pro
The distribution of values in each column
Select Columns
4 Columns
BabyGender
Birth Weight
Smoked
Parity
Histograms Only
Cast Selected Columns into Roles
Y, Columns
Birth Weight
optional
Weight
optional numeric
Freq
optional numeric
By
optional
Action
OK
Cancel
Remove
Recall
Help
Rows
All rows 100
Selected 0
Excluded 0
Hidden 0
Labelled 0
12 M 3130 0 0
13 F 2780 0 2
14 M 2650 0 0
15 M 2380 0 0
16 M 3170 0 0
17 F 2980 0 0
18

“Distribution” in JMP



Shows histogram + Boxplot

Shows quantiles

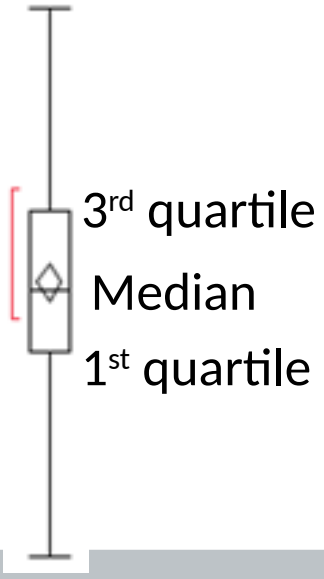
Presents summary statistics

What is a “quartile”?

1st quartile: 25% of data below, 75% above

1) Sort the n observations from smallest to largest value.

2) The value that separates the data into 25% below and 75% above is the 1st quartile.



Quartiles useful for: *Outliers*

1st quartile: 25% of data below, 75% above
3rd quartile: 75% of data below, 25% above

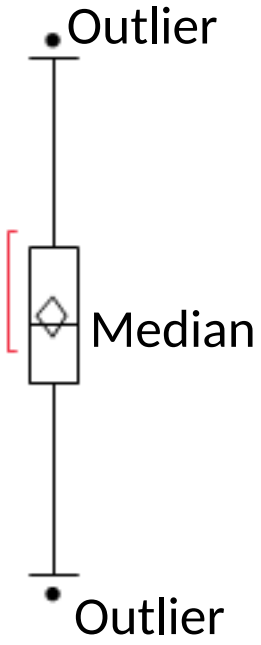
Inter-quartile range (IQR) = 3rd quartile - 1st quartile

Outliers can be defined as those data points:

$x < (1^{st}\text{ quartile} - 1.5 \times IQR)$

OR

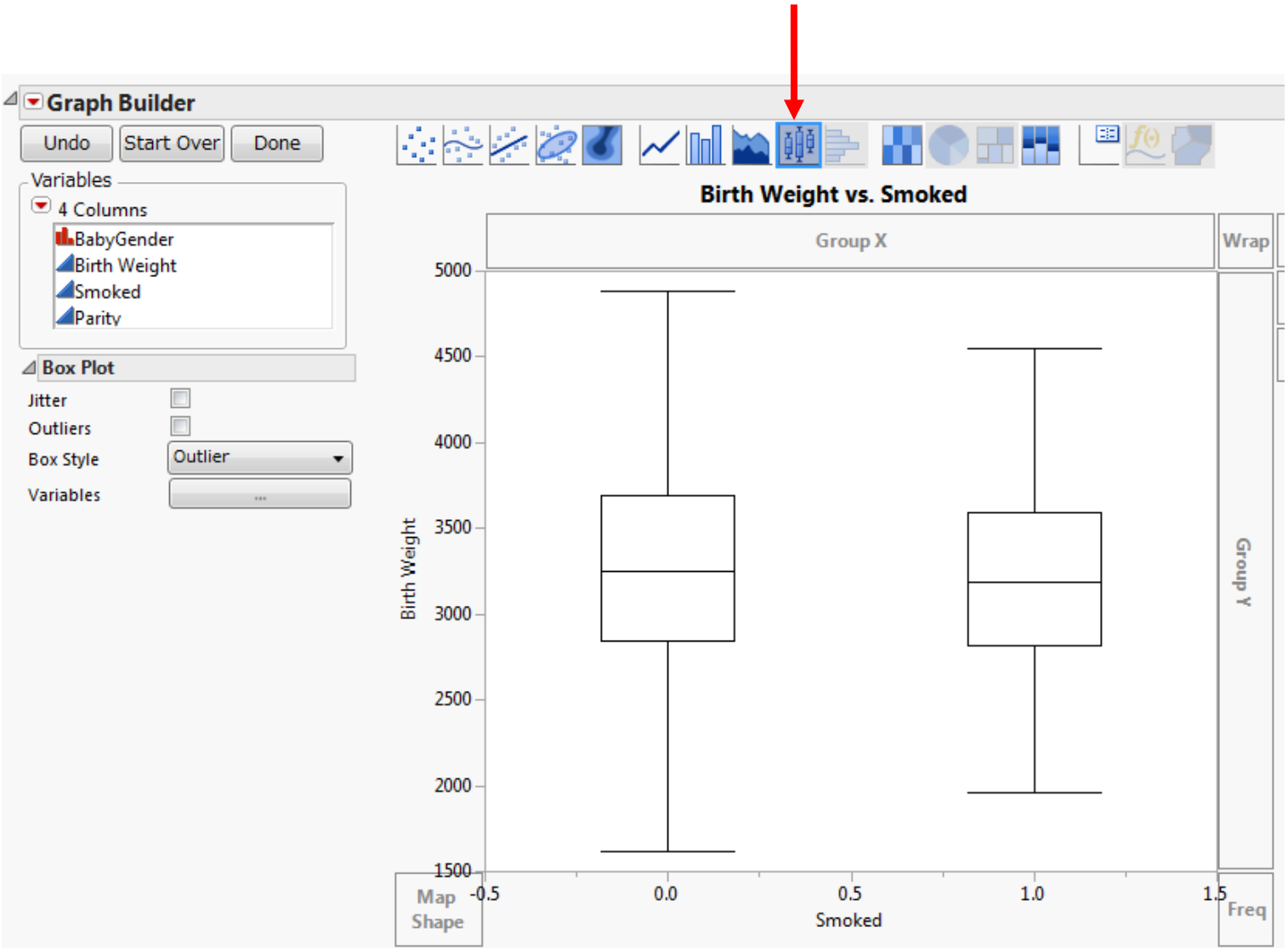
$x > (3^{rd}\text{ quartile} + 1.5 \times IQR)$



Box plot by variable in JMP

Menu-> Graph -> Graph builder

X: Smoked
Y: Baby weight



Spread of data (variance)

Population Variance
("Sigma Squared")

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

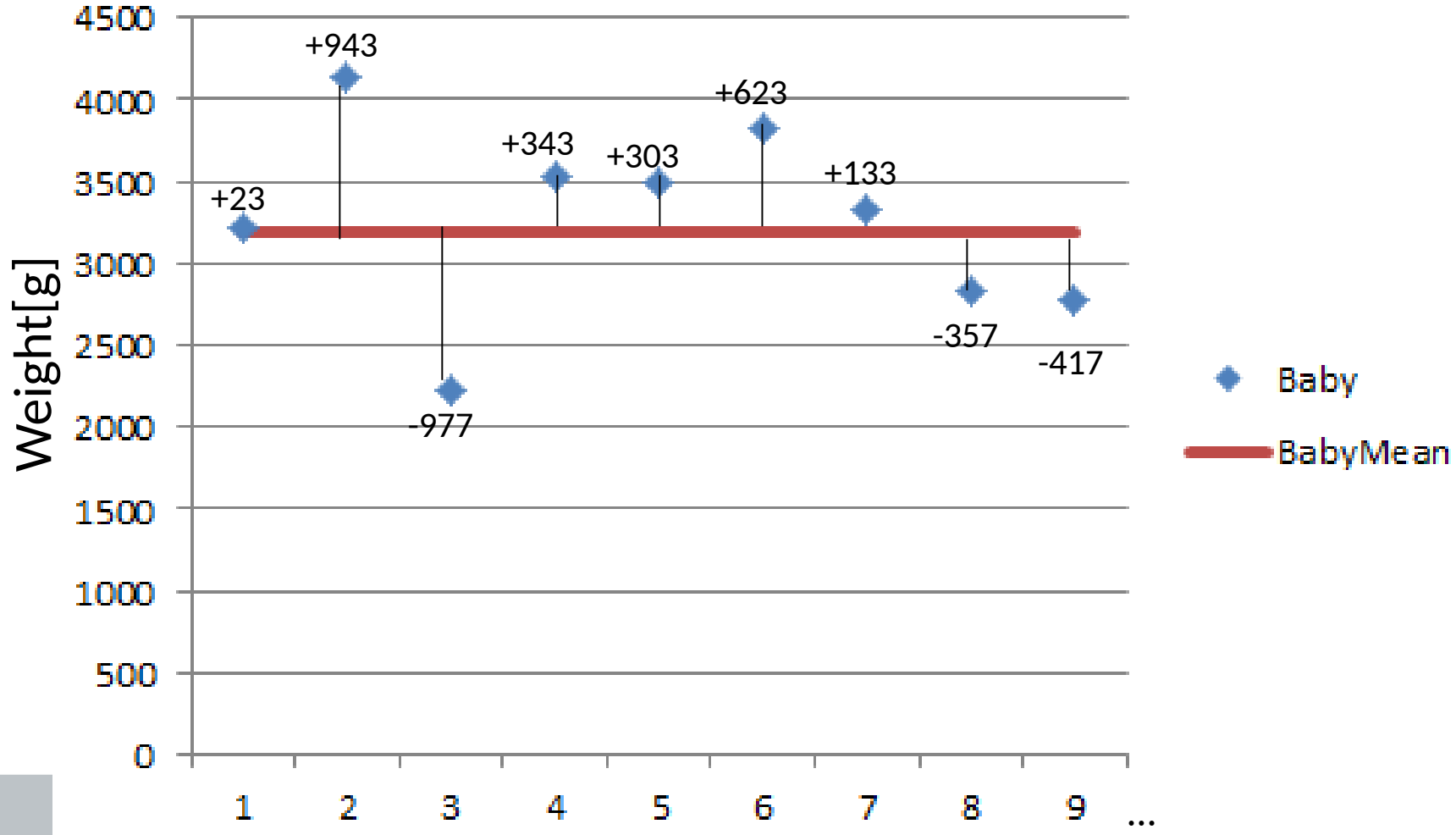
Sample Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2$$

Spread (variance)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2$$

How far is each individual data point from the average? Squared to make positive and over-weigh data further from average.



Variance versus Standard Deviation

Sample Variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \hat{x} \right)^2$$

Sample Standard Deviation “s”

$$s = \sqrt{s^2}$$

Statistic versus Parameter

I keep saying

→ “Sample standard deviation”

→ “Sample average”

Why do I say “sample”? What is so important?

Statistic versus Parameter

We differentiate between:

Statistic : We actually measure this. Using a sample of the population.

Parameter : Imaginary, (unknowable) theoretical property of the true population.

Statistic versus Parameter

We differentiate between:

Statistic : We actually measure this. Using a sample of the population.

Parameter : Imaginary, (unknowable) theoretical property of the true population.

The key theory is that we can *estimate* parameters using *samples*. And use it!

Sample Average is *Unbiased Estimator* of the Population Mean

No matter our sample size, the sample average will *underestimate* and *overestimate* the true population mean an equal amount!

→ In other words, the average of our sample averages will equal the true population mean…

Sample Average is *Unbiased Estimator* of the Population Mean

No matter our sample size, the sample average will *underestimate* and *overestimate* the true population mean an equal amount!

→ In other words, the average of our sample averages will equal the true population mean…

…meta.

Biased Estimator of Variance...

We take 1/(n-1) for sample variance.

But for population variance (parameter) it is 1/n.

Why?

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2$$

Biased Estimator of Variance…

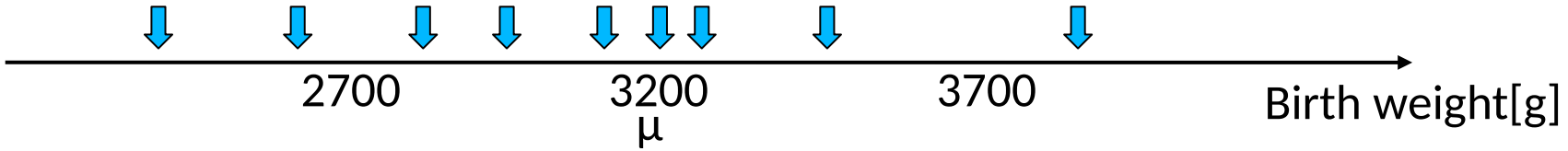
This is because unlike the mean, variance is a ***biased estimator.***

Using *n-1* rather than *n* is called Bessel's Correction – it is one method of correcting sample variance.

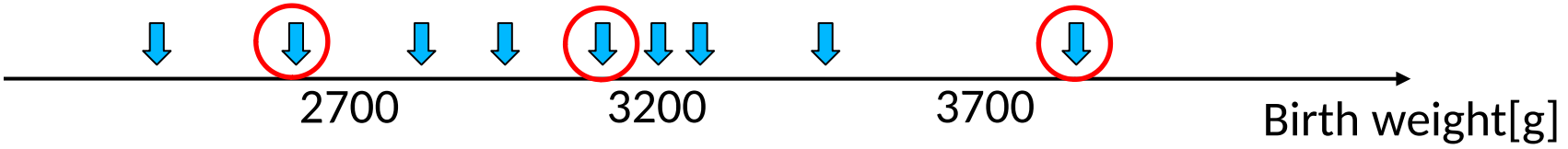
Unfortunately, square rooting the variance to get standard deviation "uncorrects" it some…oh well.

Why is variance underestimated...?

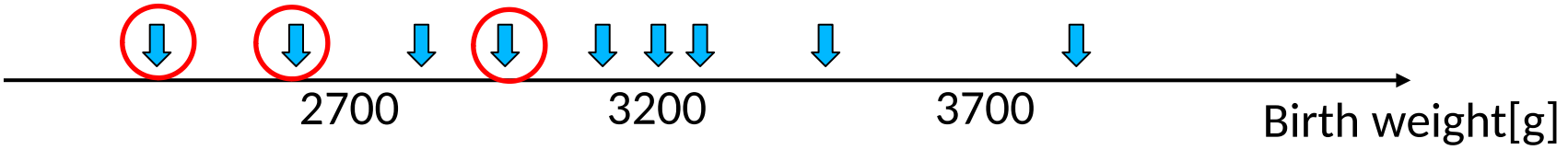
Consider these numbers as the values of the whole population:



If we take a small sample (n=3), we have a good chance that we estimate μ and σ well with sample mean (x^) and sample variance (s):

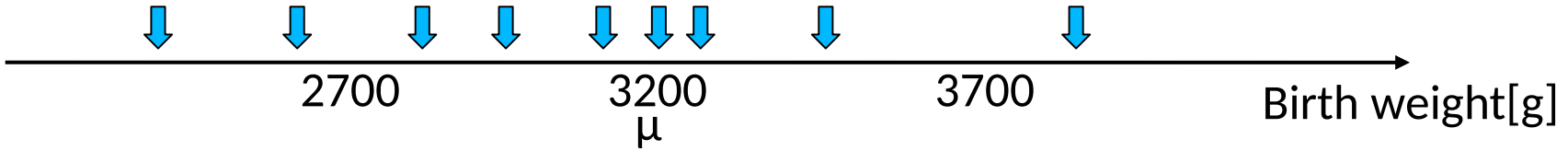


But sometimes, we will randomly pick sample elements that do not cover μ and in this case it is obvious that s will be much smaller than σ:

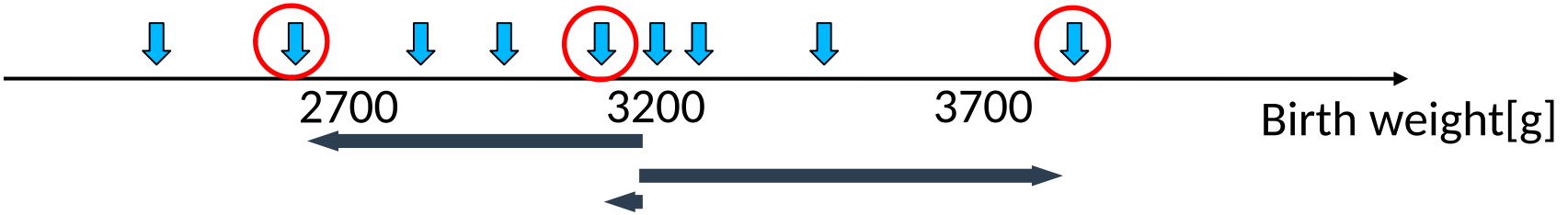


Why is variance underestimated...?

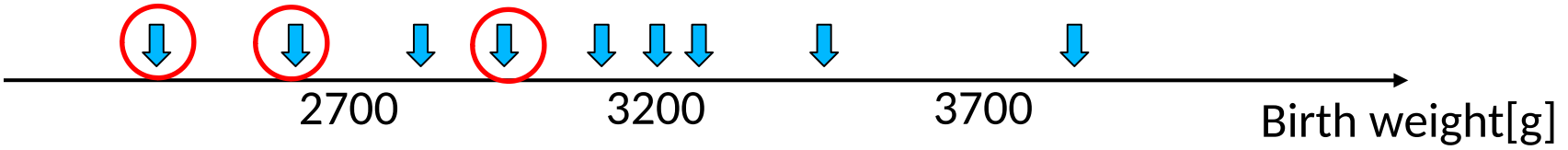
Consider these numbers as the values of the whole population:



If we take a small sample (n=3), we have a good chance that we estimate μ and σ well with sample mean (x^) and sample variance (s):

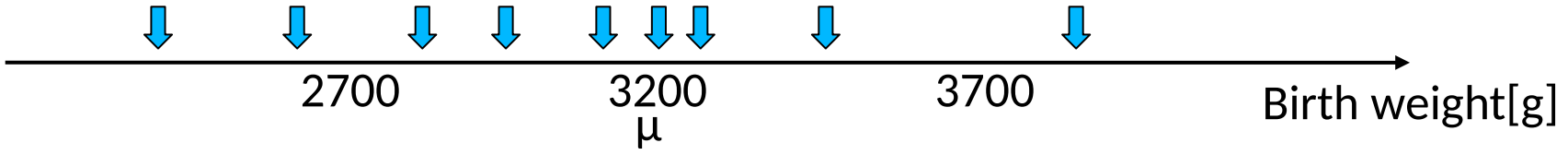


But sometimes, we will randomly pick sample elements that do not cover μ and in this case it is obvious that s will be much smaller than σ:



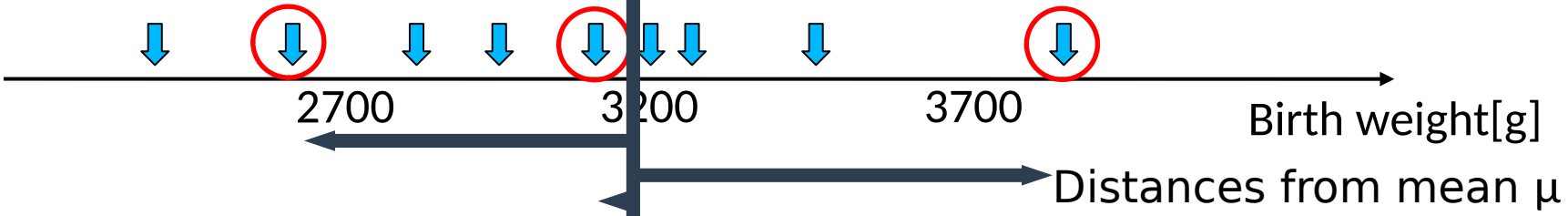
Why is variance underestimated...?

Consider these numbers as the values of the whole population:

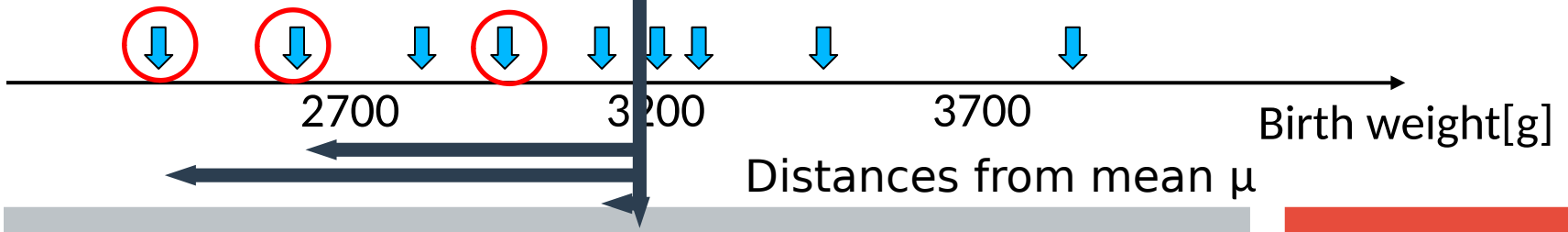


μ

If we take a small sample (n=3), we have a good chance that we estimate μ and σ well with sample mean (x^) and sample variance (s):

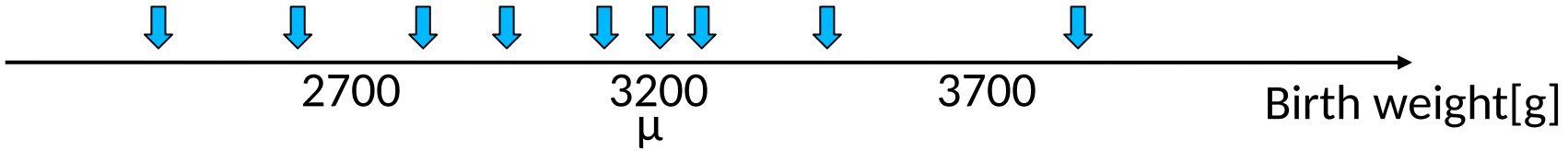


But sometimes, we will randomly pick sample elements that do not cover μ and in this case it is obvious that s will be much smaller than σ:



Why is variance underestimated...?

Consider these numbers as the values of the whole population:



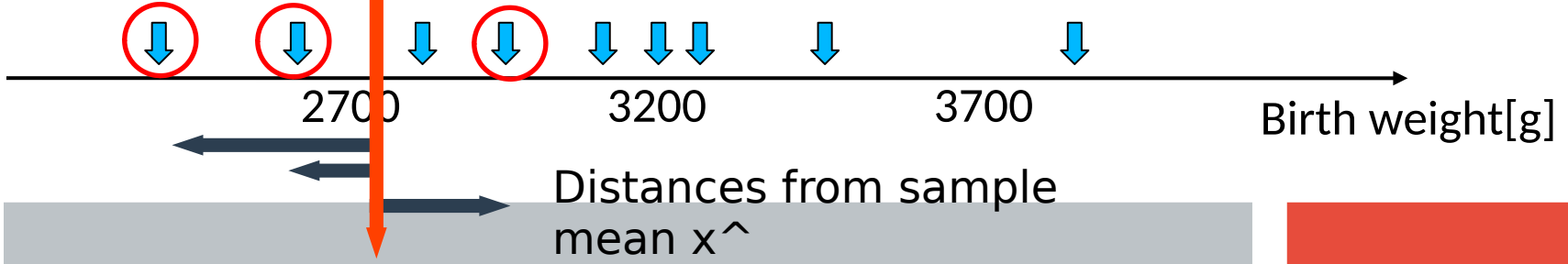
If we take a small sample (n=3), we have a good chance that we estimate μ and σ well with sample me

But we don't know the true mean!

Only our sample mean... which is calculated from our samples... which are all biased in one direction this time!

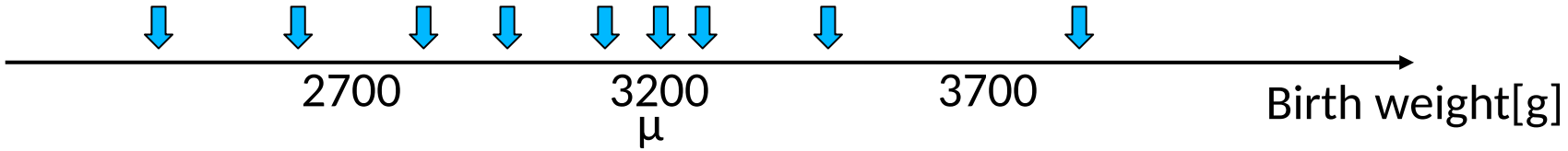
mean

But sometimes, we will μ and in this case it is obvious that s will be much smaller than σ:



Why is variance underestimated…?

Consider these numbers as the values of the whole population:



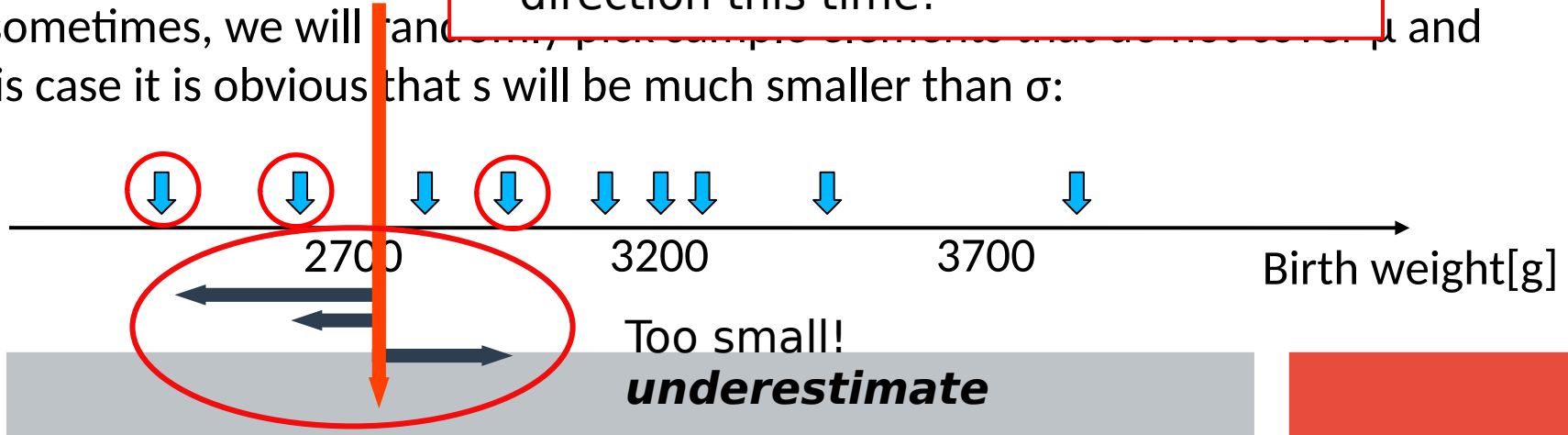
If we take a small sample (n=3), we have a good chance that we estimate μ and σ well with sample me

But we don't know the true mean!

Only our sample mean… which is calculated from our samples… which are all biased in one direction this time!

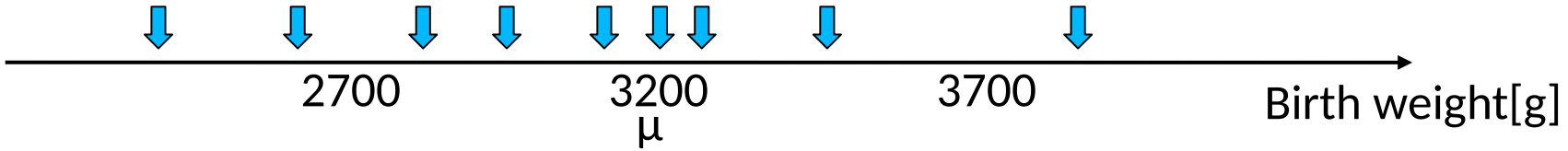
mean

But sometimes, we will [illegible] μ and in this case it is obvious that s will be much smaller than σ:

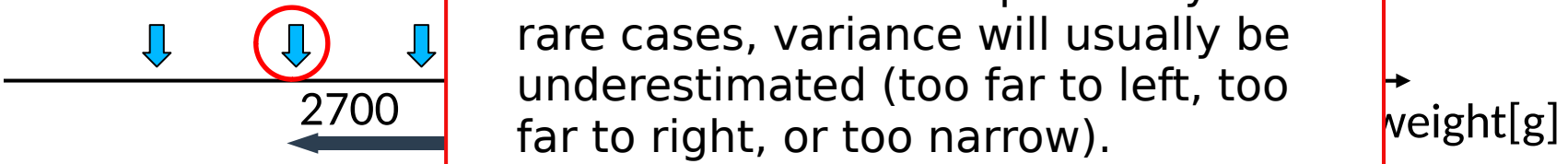


Why is variance underestimated...?

Consider these numbers as the values of the whole population:



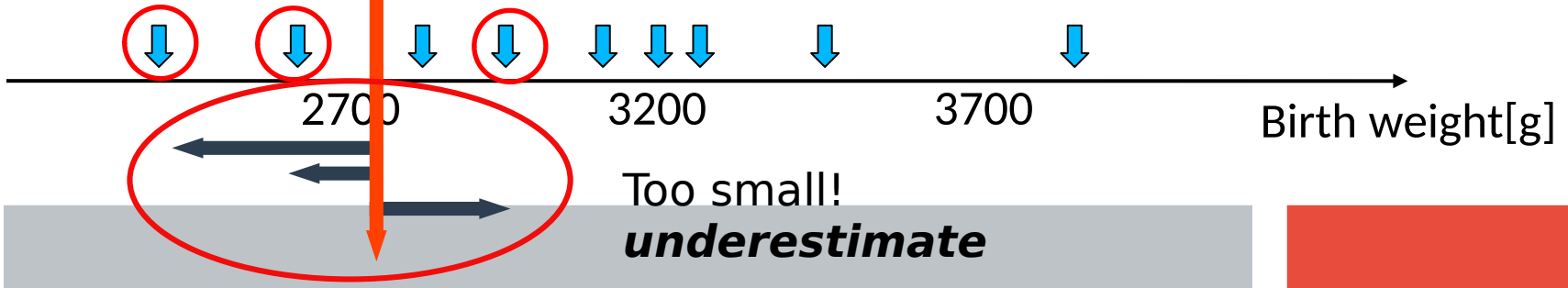
If we take a small sample (n=3), we have a good chance that we estimate μ and σ well with sample me



n mean

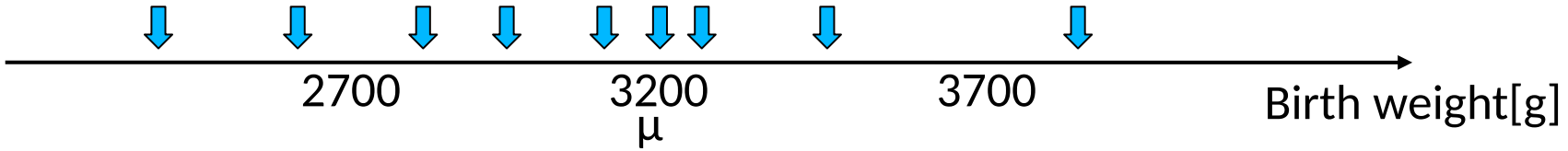
You can see that except in very rare cases, variance will usually be underestimated (too far to left, too far to right, or too narrow).

But sometimes, we will [illegible] μ and in this case it is obvious that s will be much smaller than σ:

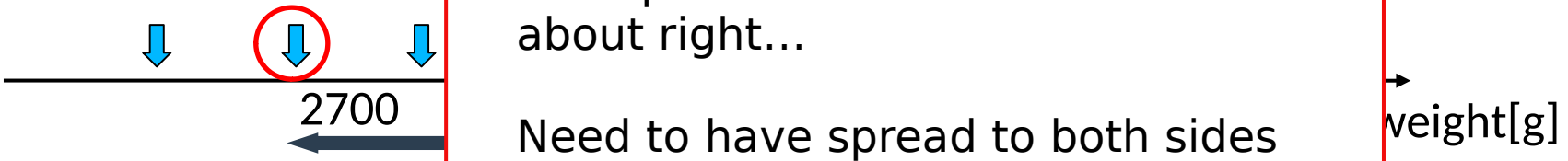


Why is variance underestimated…?

Consider these numbers as the values of the whole population:



If we take a small sample (n=3), we have a good chance that we estimate μ and σ well with sample me

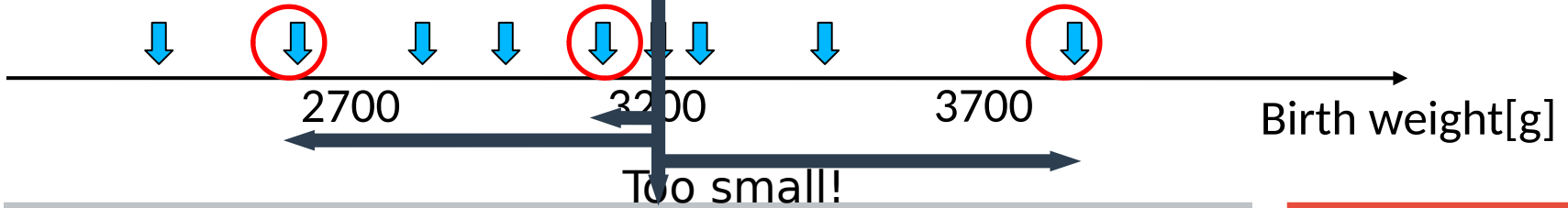


Example rare case where it is about right…

Need to have spread to both sides so that average balances, and also right spread to each side (duh).

n mean

But sometimes, we will ran μ and in this case it is obvious that s will be much smaller than σ:



underestimate

Biased Estimator of Variance...

Doing n-1 is a “hack” (rough correction) to get the values to underestimate the variance less in the limit.

It works because formally sample variance has one less degree of freedom (if we know the mean, and we know all but one sample, we can cheat and calculate the value of the last sample! Giving us information out of nowhere?

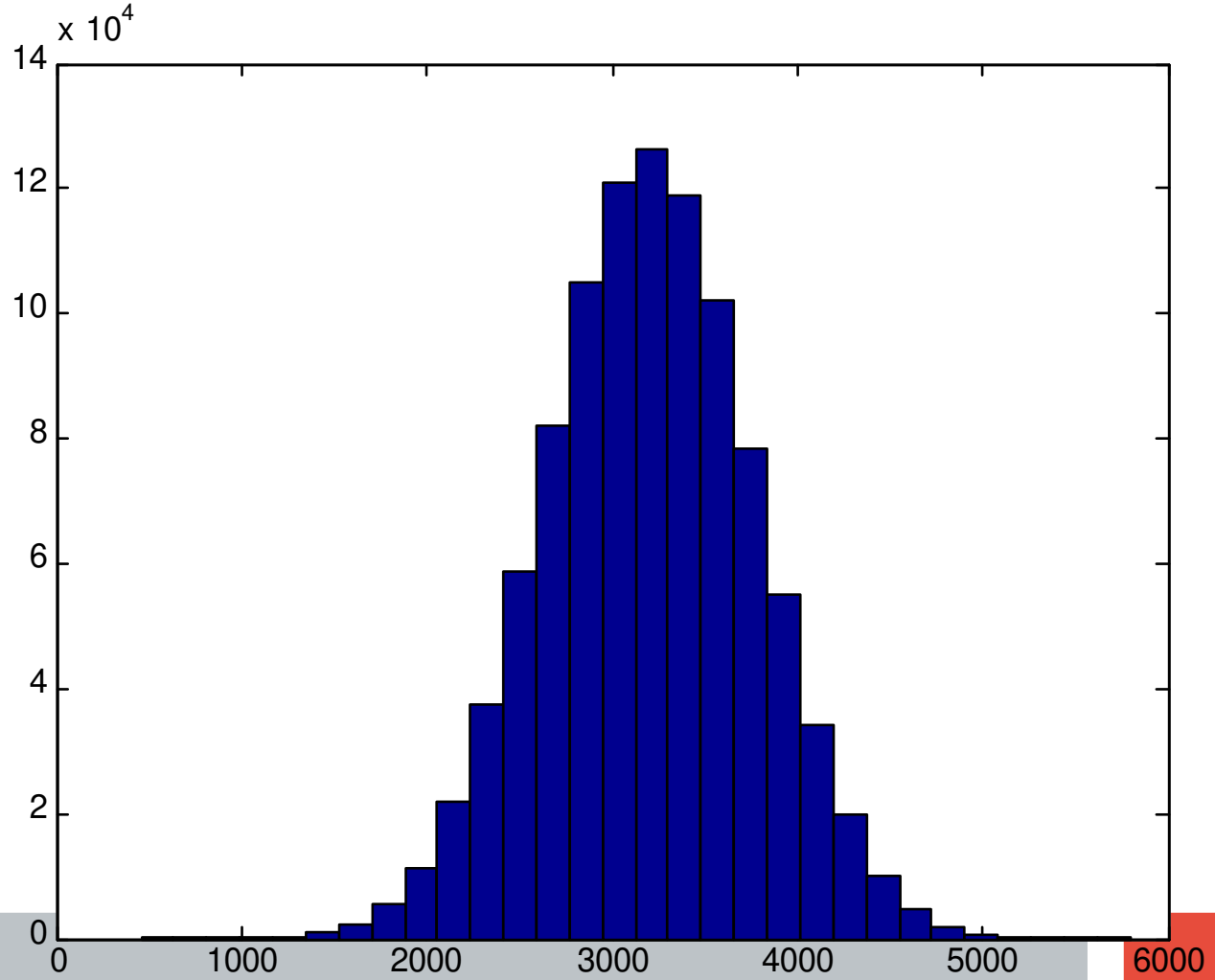
So, what that means is basically that we have less information than we thought.

- If we knew the true mean μ and used it → unbiased.

See it work

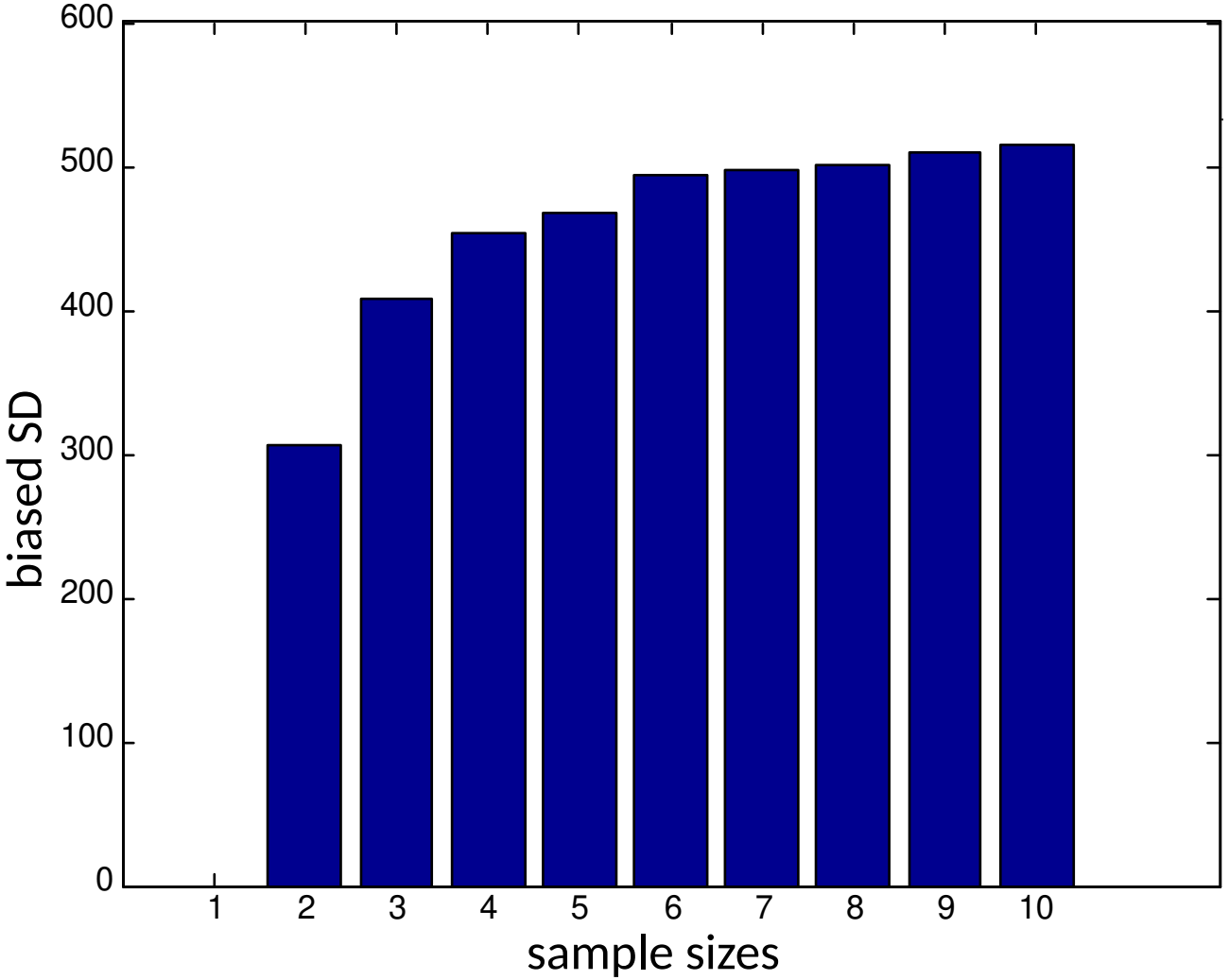
Let's say, we have 1,000,000 birth weight values [g]

μ=3200 g, σ=560 g (simulated with normal distribution)

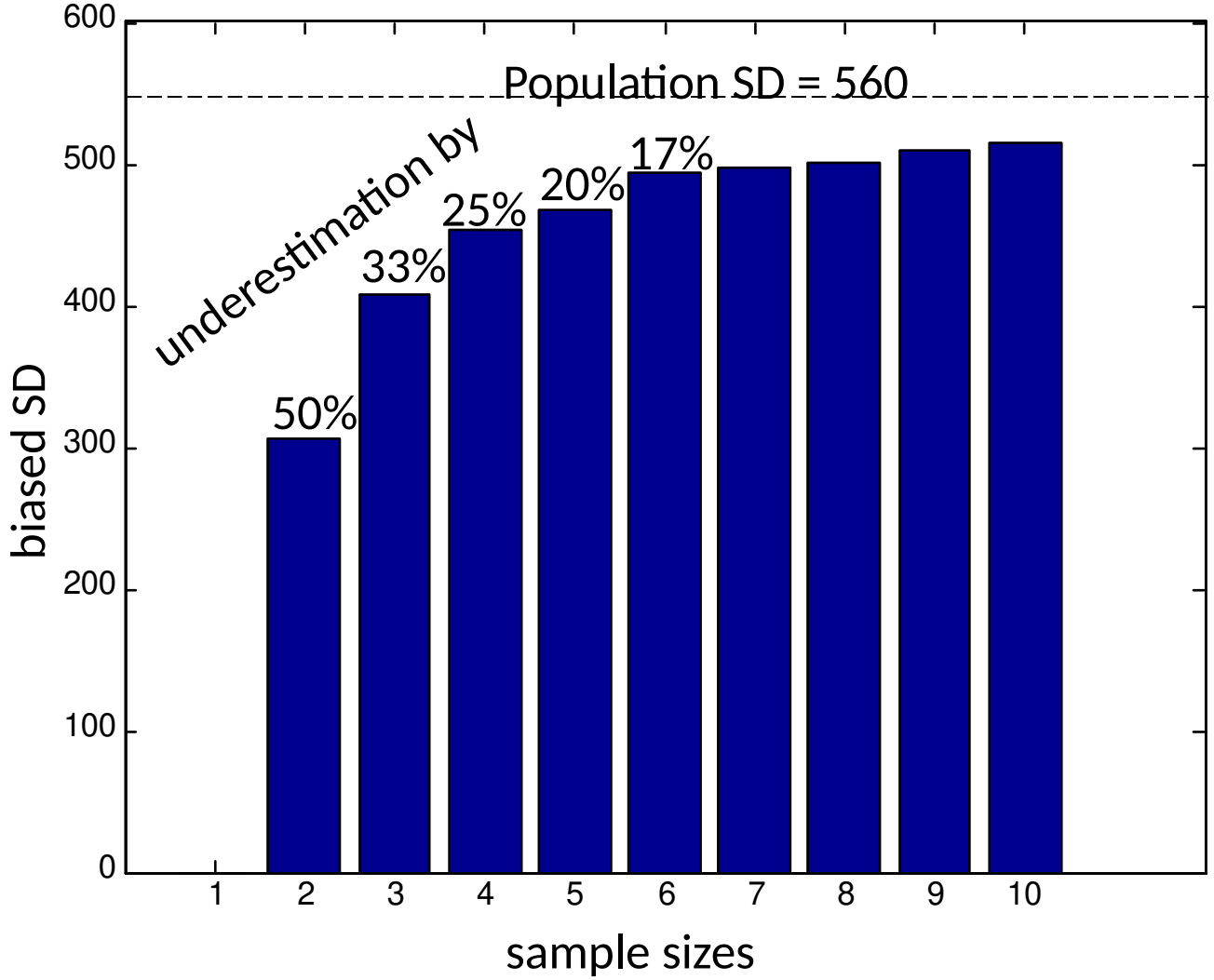


See it work

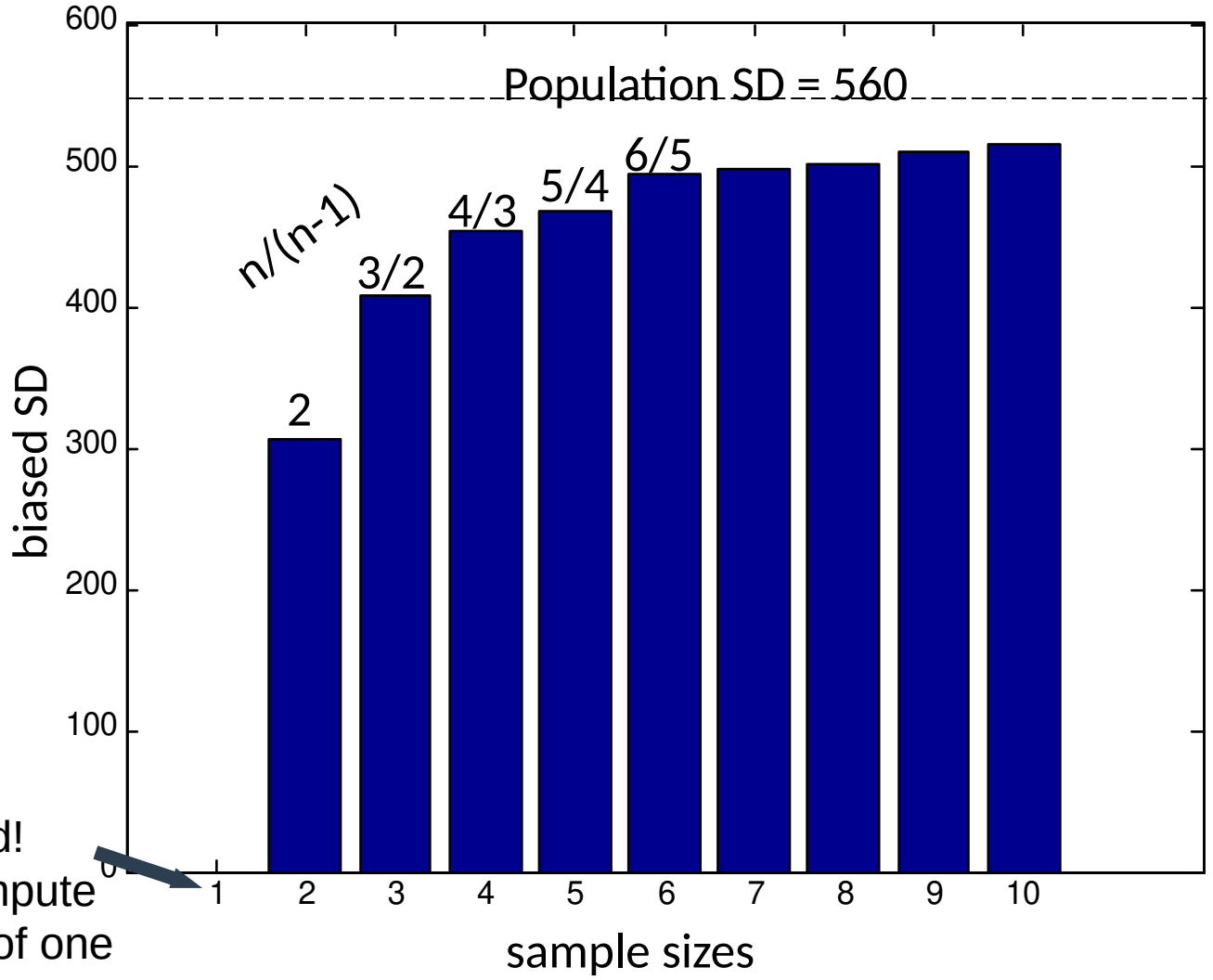
Mean of SD estimated for 100 samples with sample sizes 2 to 10.



See it work


600
500
400
300
200
100
0
Population SD = 560
underestimation by
50%
33%
25%
20%
17%
biased SD
1
2
3
4
5
6
7
8
9
10
sample sizes

Multiply by n/(n-1)

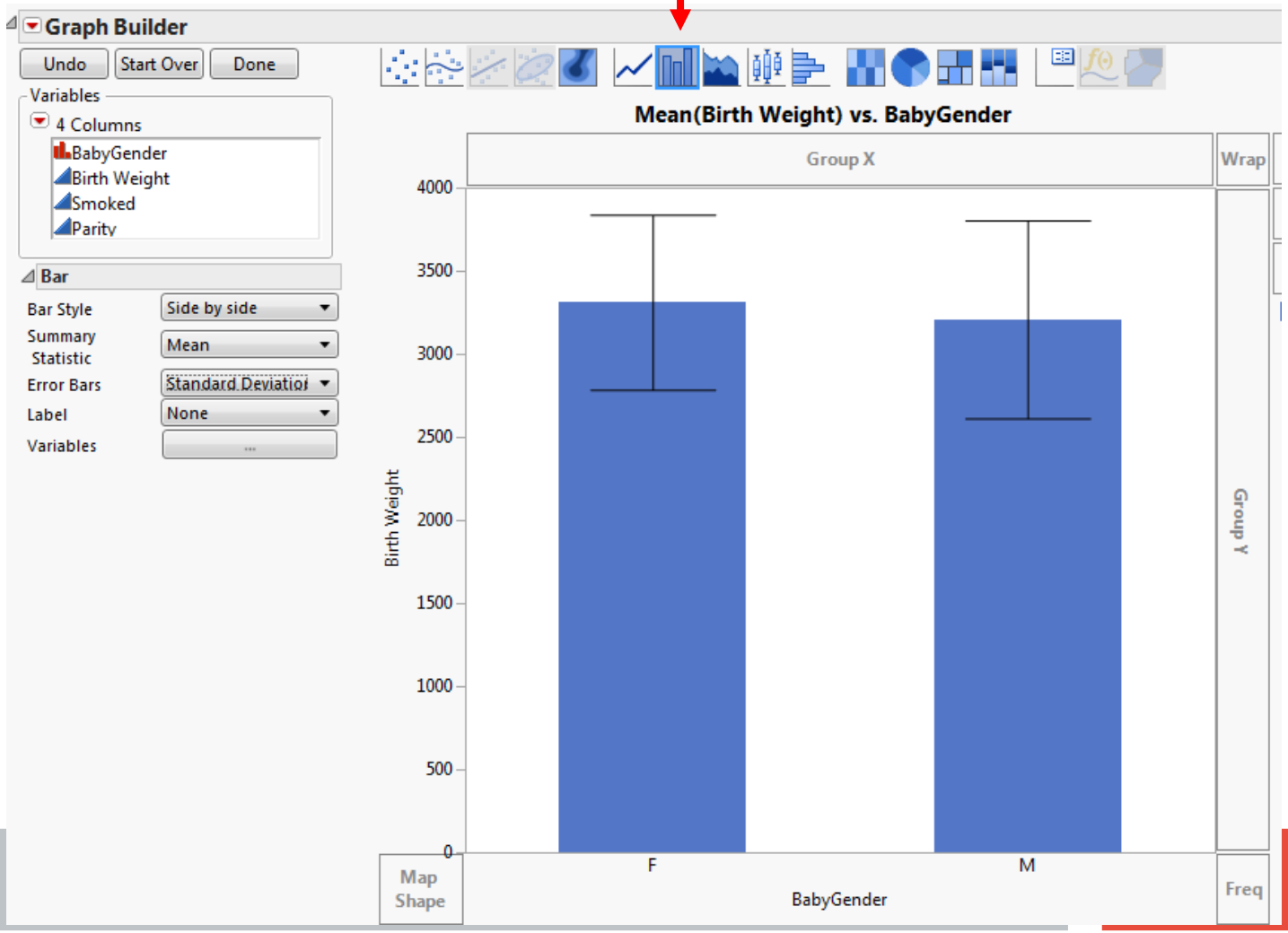


Undefined!
Can't compute variance of one sample anyways...

Bar Graph in JMP

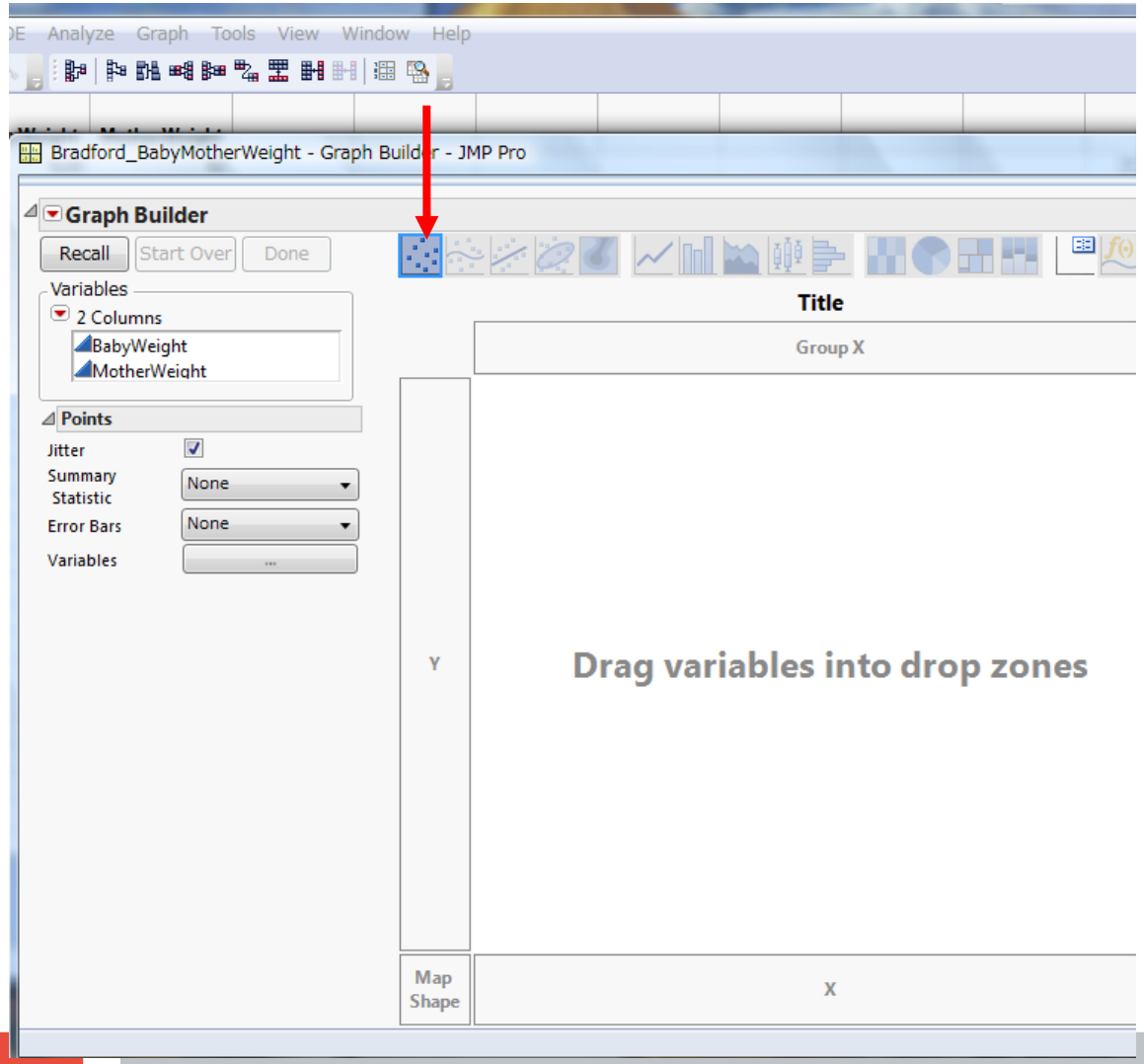
Menu → Graph → Graph builder

X: Baby gender; Y: Baby weight



Scatter Plot in JMP

Menu-> Graph -> Graph builder



Scatter plots can visualize correlations between continuous variables

Scatter Plot in JMP

X: Mother weight; Y: Baby weight