

# Variables and data

## data (variables)

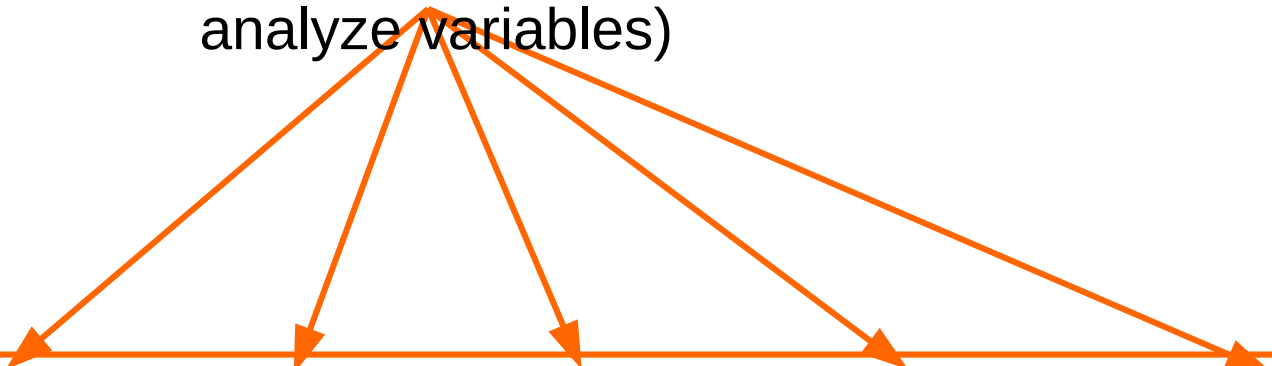
**Variables** are attributes that can vary (we measure values for, and analyze variables)

**Data** is a set of measurements, counts, observations, etc. for variables of interest (we collect data)

Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

data  
(variables)


**Variables** are attributes that can vary (we measure values for, and analyze variables)



Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

## data (variables)

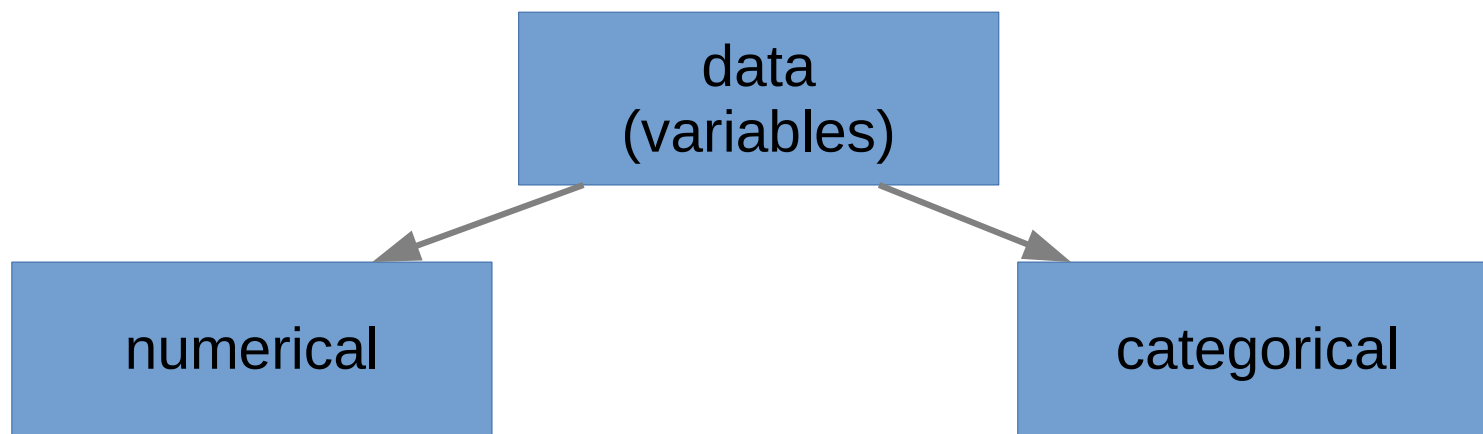
**Data** is a set of measurements, counts, observations, etc.  
for variables of interest (we collect data)

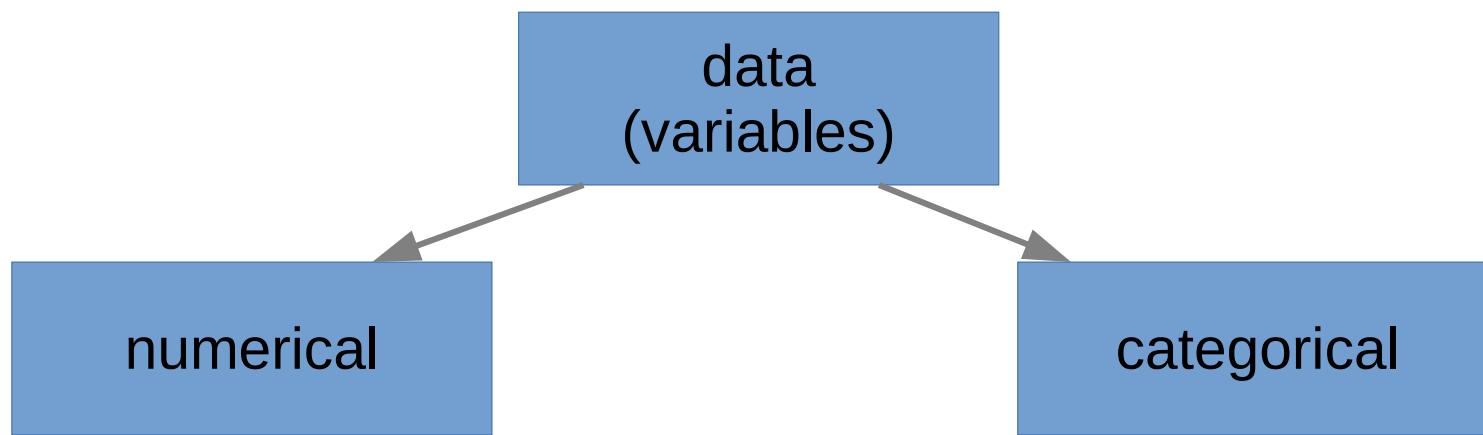


Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

# Types of variables

data  
(variables)





Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).



data  
(variables)

```
graph TD; A["data (variables)"] --> B["numerical"]; A --> C["categorical"]; B --> D["continuous"]; B --> E["discrete"];
```

numerical

Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

continuous

discrete

categorical

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).

data  
(variables)

```
graph TD; A["data (variables)"] --> B["numerical"]; A --> C["categorical"]; B --> D["continuous"]; B --> E["discrete"];
```

numerical

Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

continuous

Can take on any value and are often measured (height, weight). Math results make sense.

discrete

categorical

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).

data  
(variables)

```
graph TD; A["data (variables)"] --> B["numerical"]; A --> C["categorical"]; B --> D["continuous"]; B --> E["discrete"];
```

numerical

Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

continuous

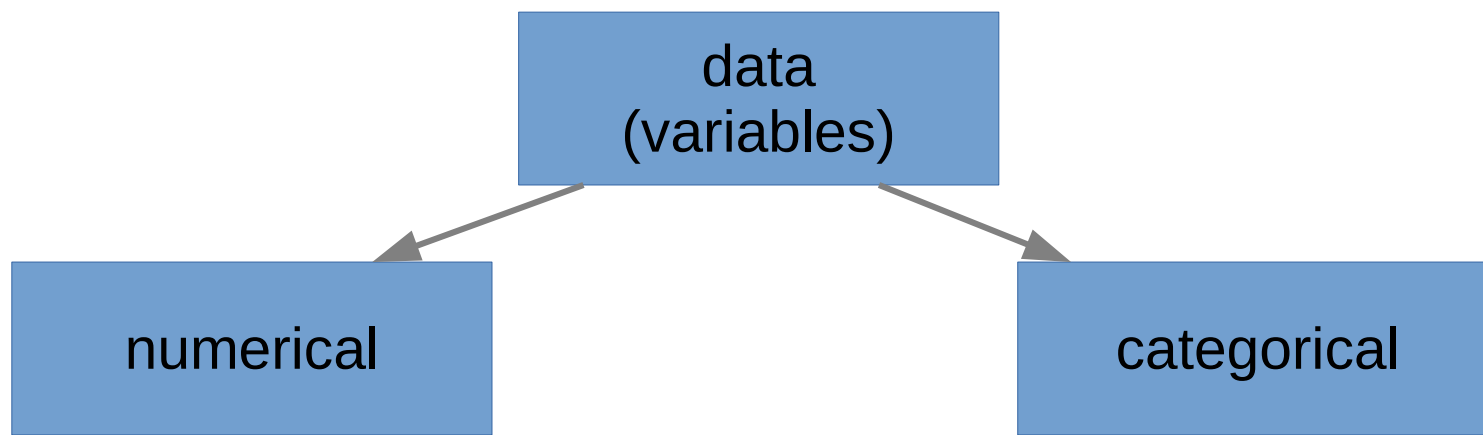
Can take on any value and are often measured (height, weight). Math results make sense.

discrete

Only has specific values, often counted (coin flips, number of people). Results of math are OK, but sometimes don't make 100% sense (average of 2.2 people per group is useful sometimes).

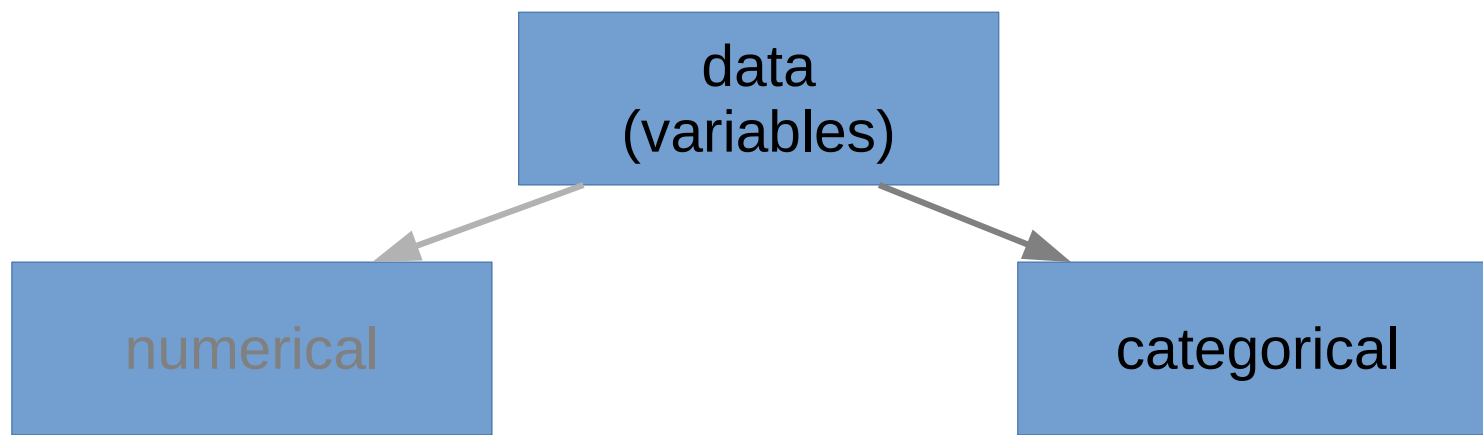
categorical

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).



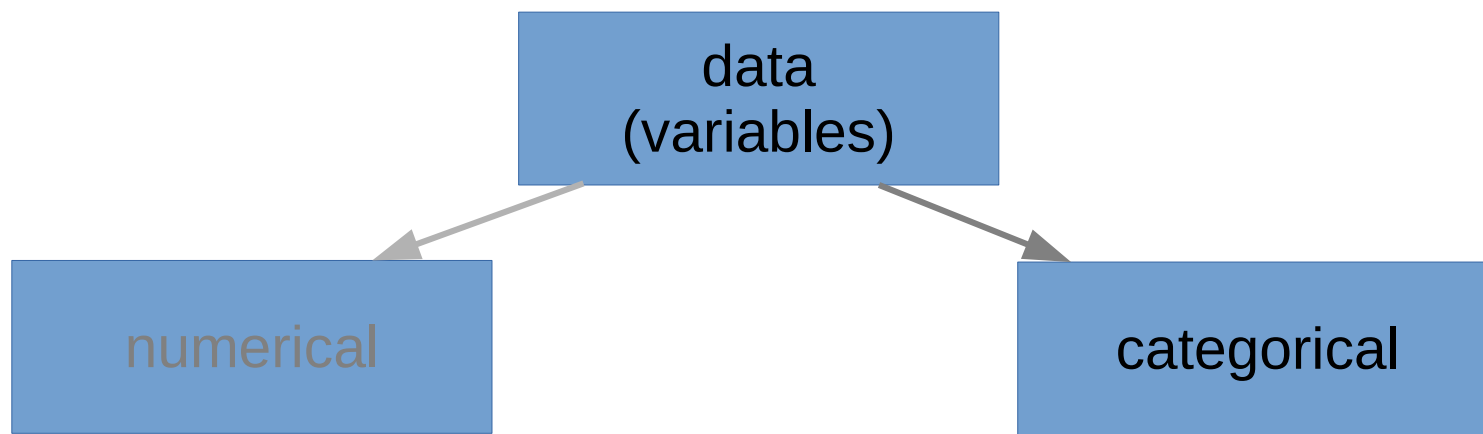
Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).



Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).

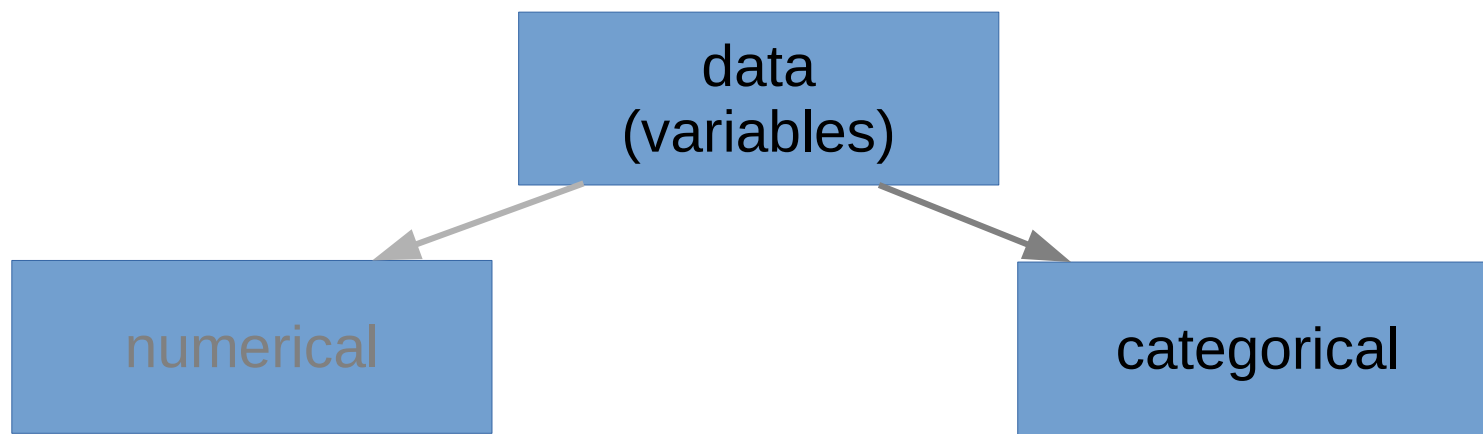


Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).

nominal

ordinal



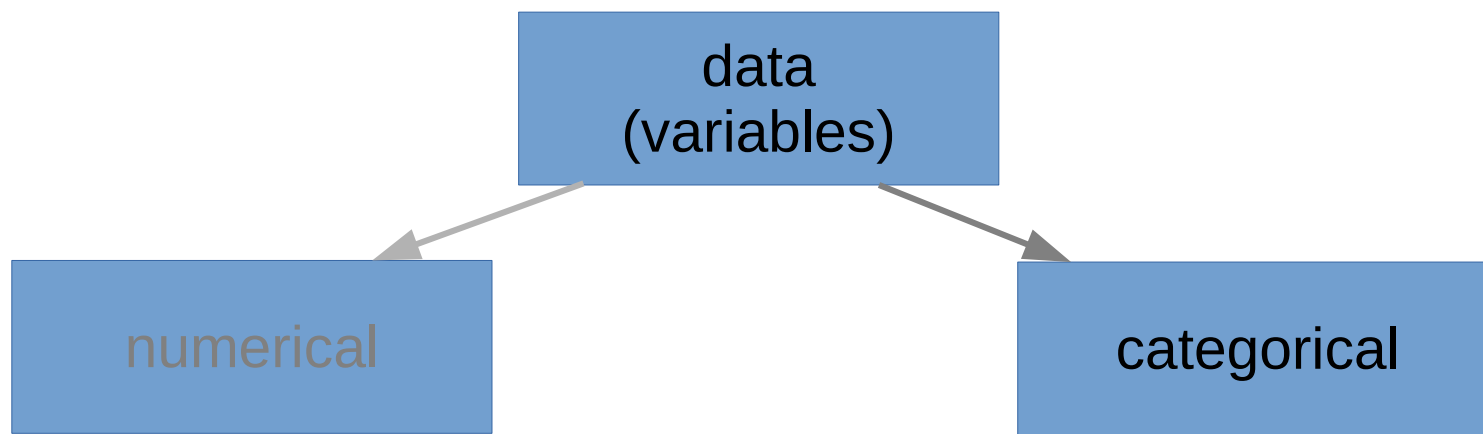
Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).

nominal

ordinal

Categories that are only names without any inherent/meaningful order (M/F, names, colors, dog breeds), often genotypes (when you don't know or understand their order).



Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).

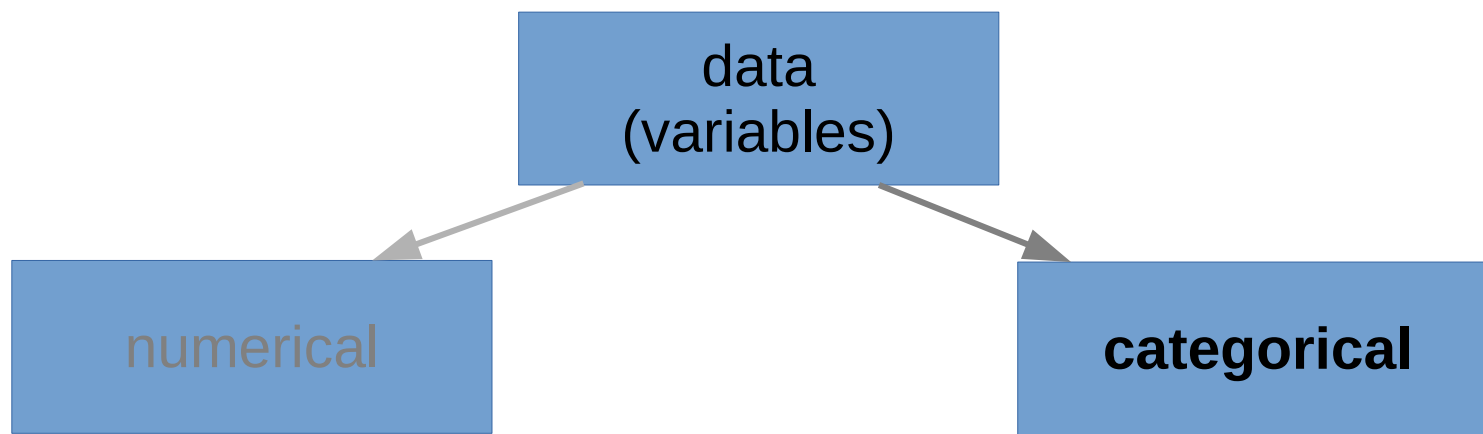
nominal

Categories that are only names without any inherent/meaningful order (M/F, names, colors, dog breeds), often genotypes (when you don't know or understand their order).

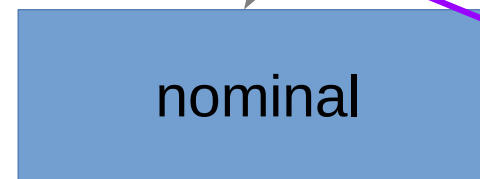
ordinal

Named categories that do have a meaningful order. Our most common example is Likert or preference scores (a lot, a little, not at all), genotypes are ordinal sometimes (you know that AA is tallest, then Aa, then aa).

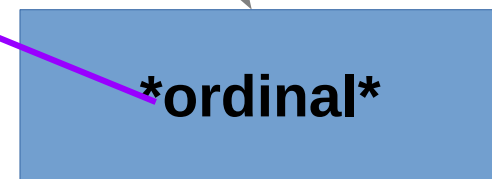




Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).



Categories that are only names without any inherent/meaningful order (M/F, names, colors, dog breeds), often genotypes (when you don't know or understand their order).



Named categories that do have a meaningful order. Our most common example is Likert or preference scores (a lot, a little, not at all), genotypes are ordinal sometimes (you know that AA is tallest, then Aa, then aa).

***\*note\* \*note\* \*note\* \*note\* \*note\* \*note\****

Sometimes researchers “code” variables using numbers and even analyze them as discrete numerical! For example, yes and no can be coded as 1 and 0, but it does not make sense to analyze them as numerical since they are nominal. Coding is used to make the analysis easier. Another common example is the Likert scale (ask the google about it) which is often coded on a 1-5 scale. The most **important** part of deciding whether your coded variable can be treated as a discrete numerical variable during analysis is whether the distances between the categories are roughly the same (is the difference between 1 and 2 the same as 4 and 5), if it is, then you can analyze the variable as discrete numerical (e.g. a 3.5 on a Likert scale tells you something useful).

data  
(variables)

```
graph TD; A["data (variables)"] --> B["numerical"]; A --> C["categorical"]; B --> D["continuous"]; B --> E["discrete"]; C --> F["nominal"]; C --> G["ordinal"];
```

numerical

Have number values that are reasonable to add, subtract, take an average, etc. They are often measurements (e.g. height) or counts.

continuous

Can take on any value and are often measured (height, weight). Math results make sense.

discrete

Only has specific values, often counted (coin flips, number of people). Results of math are OK, but sometimes don't make 100% sense (average of 2.2 people per group is useful sometimes).

categorical

Are categories, usually a small number of options. Math doesn't make sense. Genotypes are an example of categories (you're either AA, Aa, or aa).

nominal

Categories that are only names without any inherent/meaningful order (M/F, names, colors, dog breeds), often genotypes (when you don't know or understand their order).

ordinal

Named categories that do have a meaningful order. Our most common example is Likert or preference scores (a lot, a little, not at all), genotypes are ordinal sometimes (you know that AA is tallest, then Aa, then aa).

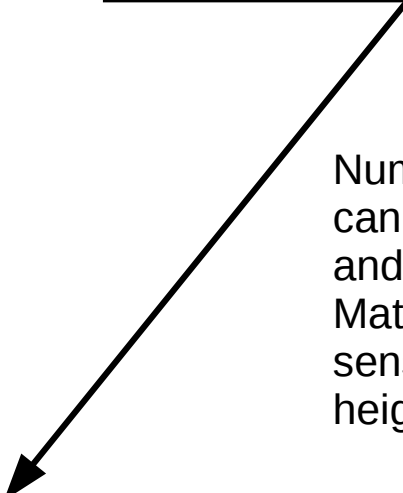
Types of variables: example

Which variables in the data are continuous (numerical)? And Why?

Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

Which variables in the data are continuous (numerical)? And Why?

Number values that  
can take on any value  
and are measured.  
Math results make  
sense (average  
height).



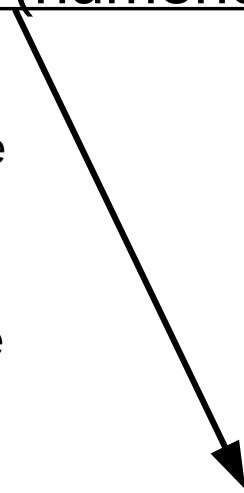
Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

Which variables in the data are discrete (numerical)? And Why?

Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

Which variables in the data are discrete (numerical)? And Why?

Number values that have a limited number of values and are counted. Math results make sense in this case to a point (2.5 average servings/day, but small decimals would be strange since people are estimating the #).



Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

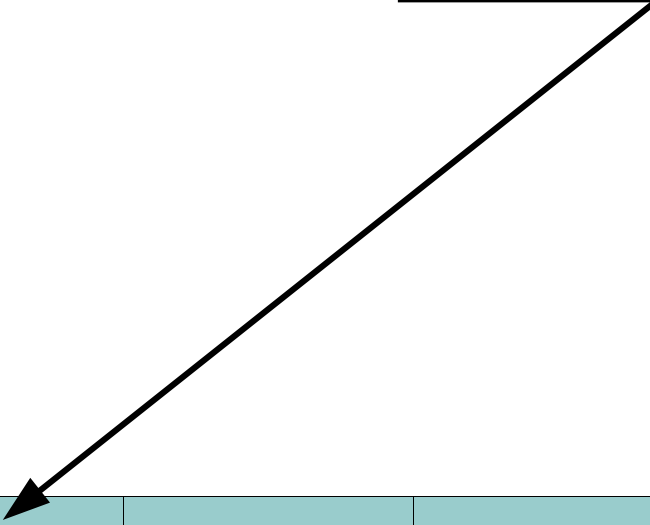
Which variables in the data are nominal (categorical)? And Why?

Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...



Which variables in the data are nominal (categorical)? And Why?

Genotype could be considered nominal (but not ordinal, it is certainly categorical) especially if we don't know much about what it does to the phenotype.



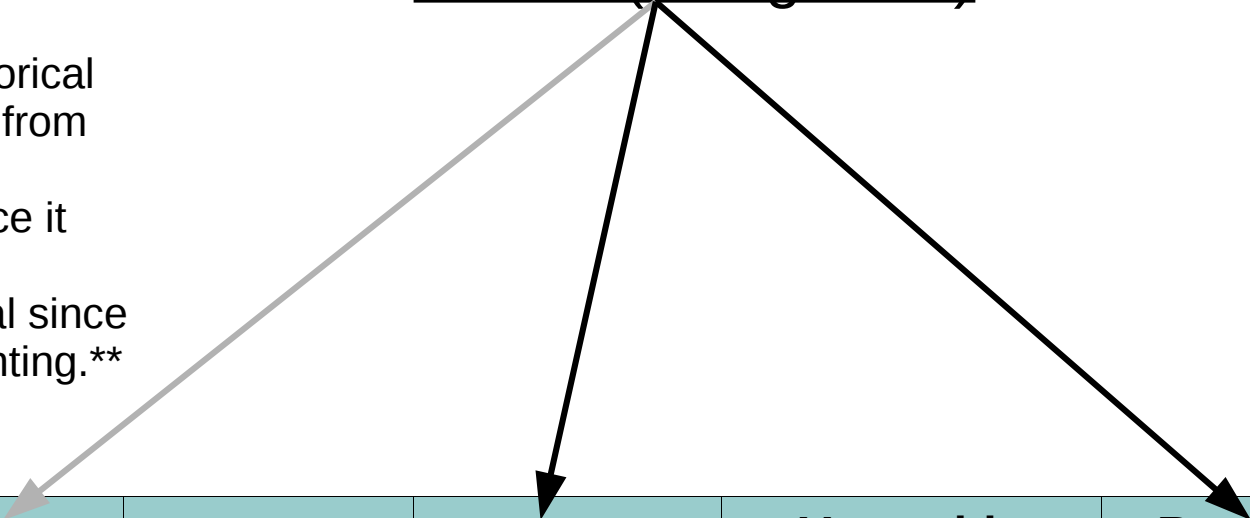
Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

Which variables in the data are ordinal (categorical)? And Why?

Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

# Which variables in the data are ordinal (categorical)? And Why?

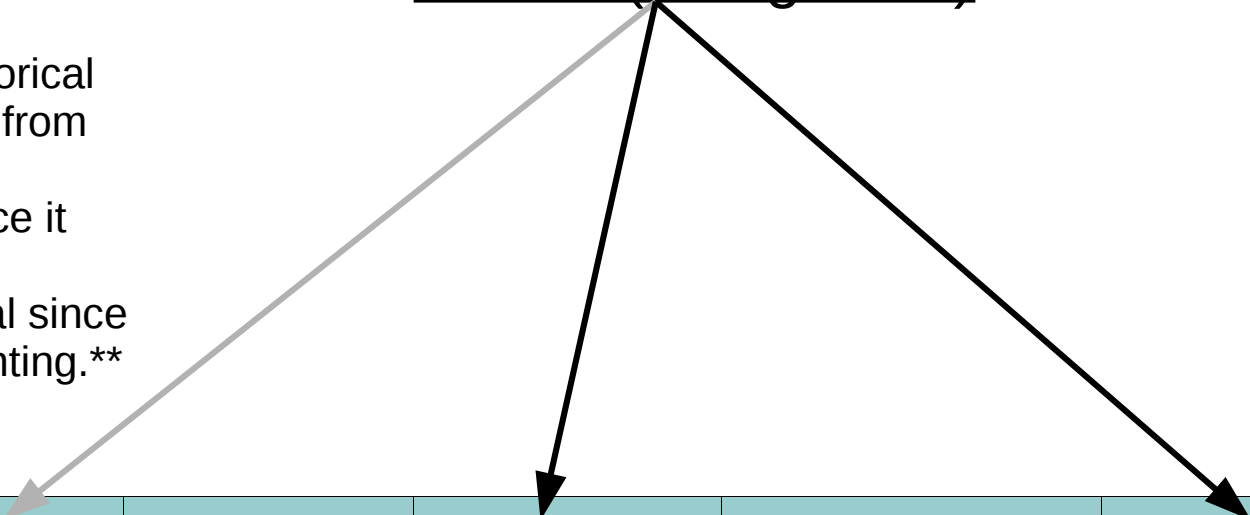
PTC taste is an ordinal categorical variable since it has an order from none to very bitter. Broccoli preference is also ordinal since it has an order from not to lots. Neither of these are numerical since we are not measuring or counting.\*\*



Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

# Which variables in the data are ordinal (categorical)? And Why?

PTC taste is an ordinal categorical variable since it has an order from none to very bitter. Broccoli preference is also ordinal since it has an order from not to lots. Neither of these are numerical since we are not measuring or counting.\*\*



Observation #	Genotype	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

\*\* A few notes about this category. Genotype could also be considered here if we knew that there was an order to the genotypes. If in PTC taste for example, TT usually perceived PTC to taste the most bitter Tt was in the middle, then tt was on the lowest end we would consider genotype as ordinal. Remember it's a categorical variable either way. Another note, researchers often “code” variables using numbers but they are **still categorical**! See the slide on ordinal categorical data and the next page for more info.

Observation #	Genotype*	Height (m)	PTC taste	Vegetable servings/day	Broccoli preference
1	TT	1.55	bitter	0	not
2	Tt	1.73	very bitter	1	not
3	Tt	1.86	slightly bitter	1	little
4	tt	1.70	none	3	some
5	tt	1.77	none	3	lots
...	...	...	...	...	...

### ***A few notes about this category...***

\* Genotype could also be considered here if we knew that there was an order to the genotypes. If in PTC taste for example, TT usually perceived PTC to taste the most bitter Tt was in the middle, then tt was on the lowest end we would consider genotype as ordinal. Remember it's a categorical variable either way.

\*\*Sometimes researchers “code” variables using numbers and even analyze them as discrete numerical! Here we could code *PTC taste* and *Broccoli preference* on a numerical scale of 1 to 5 (or however many levels there are). Remember that the most ***important*** part of deciding whether your coded variable can be treated as a discrete numerical variable during analysis is whether the distances between the categories are roughly the same (is the difference between 1 and 2 the same as 4 and 5), if it is, then you can analyze the variable as discrete numerical. So in the case of these 2 variables, let's assume that *PTC taste* has the following levels: 1 = none, 2 = slightly bitter, 3 = bitter, 4 = very bitter, and 5 = extremely bitter. we need to decide whether the difference between bitter and very bitter is the same as very bitter and extremely bitter (and all the other categories) if it is the same we can consider *PTC taste* as a discrete numerical variable for analysis.