ANOVA:
One-way ANOVAs (theory)

Research Methods for Human Inquiry

Andrew Perfors

So far we've seen….

Chi-squared

- outcome is nominal
- e.g., frequency tables

Data:

A	B	C
35	76	24

Goodness of fit test
compares to a theoretical standard: is it "the same"?

X	Y	Z
33%	33%	33%

Test of independence
compares to another frequency table: are they "the same"?

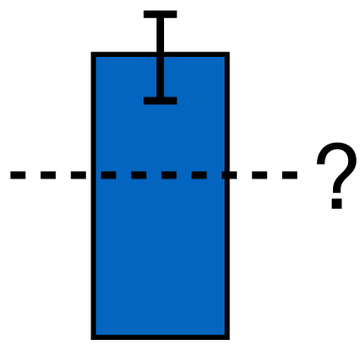
X	Y	Z
42	98	51

So far we've seen….

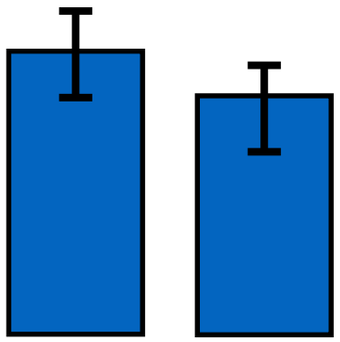
T-test

- outcome is numeric
- comparing two means: are they "the same"?

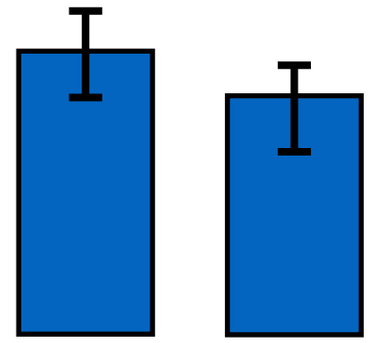
One-sample t-test compares a single mean to a theoretical standard: is it "the same"?



Independent-samples t-test compares two means to each other: are they "the same"?

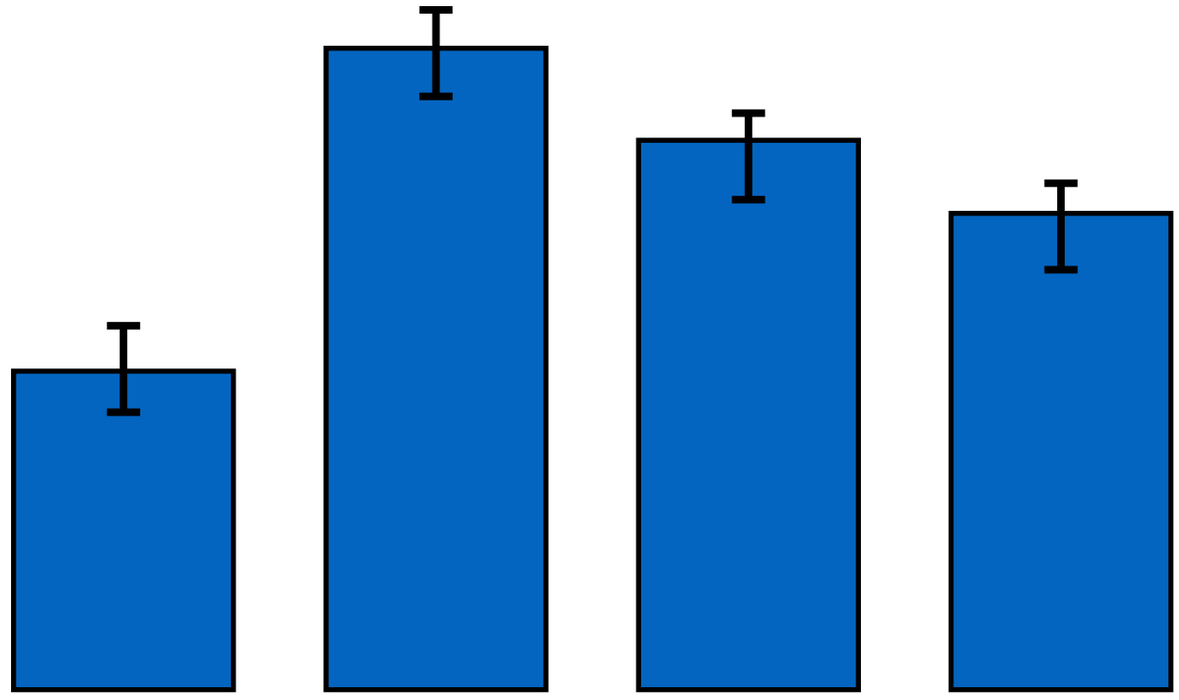


Paired-samples t-test compares two means representing two time points to each other: are they "the same"?



But what if we have more than two means?

Answer: Analysis of Variance (ANOVA)



What does ANOVA do?

- Simplest kind of ANOVA: One-way ANOVA
  - Groups are defined by a single variable: e.g.: `treatment` could be "placebo", "drug1" or “drug2"
  - It’s also possible to have groups defined by multiple variables (e.g., `treatment+gender` could have women-placebo, men-placebo, and so forth). We’ll talk about this a bit later…

- Examples...
  - e.g., `lifeSatisfaction ~ profession`
  - e.g., `extraversion ~ universityDegree`
  - e.g., `workingMemoryLoad ~ experimentalTask`
  - e.g., `killingSpeed ~ undeadSpecies`

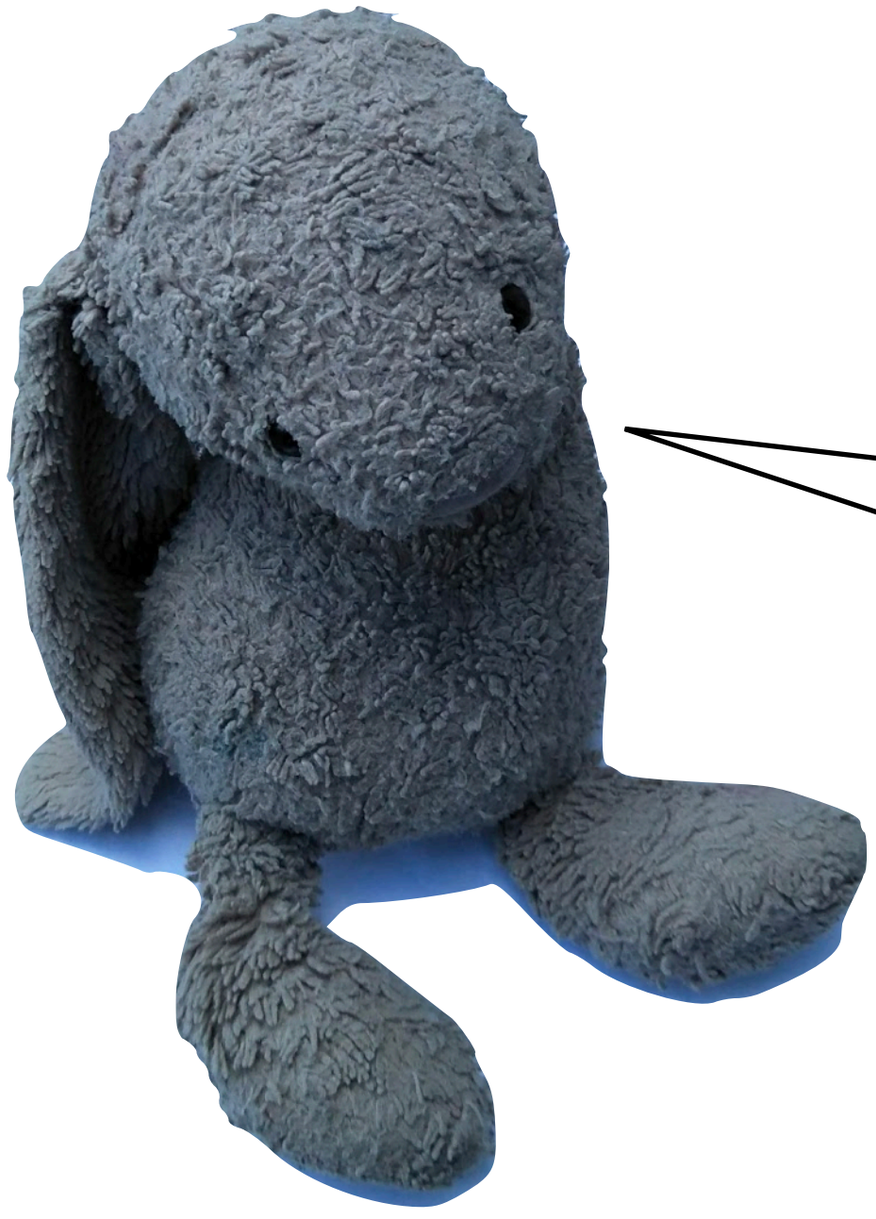
Let’s go back to our story first…

Gladly, Doggie and friends have gone to Otherland on a mission of rescue and attack!

Let’s go back to our story first…

Bunny, Flopsy, and Shadow are distressed to find them missing

Let’s go back to our story first…



We have to stop them! They’re might attack the Others even though the Others aren’t responsible

Let's go back to our story first…

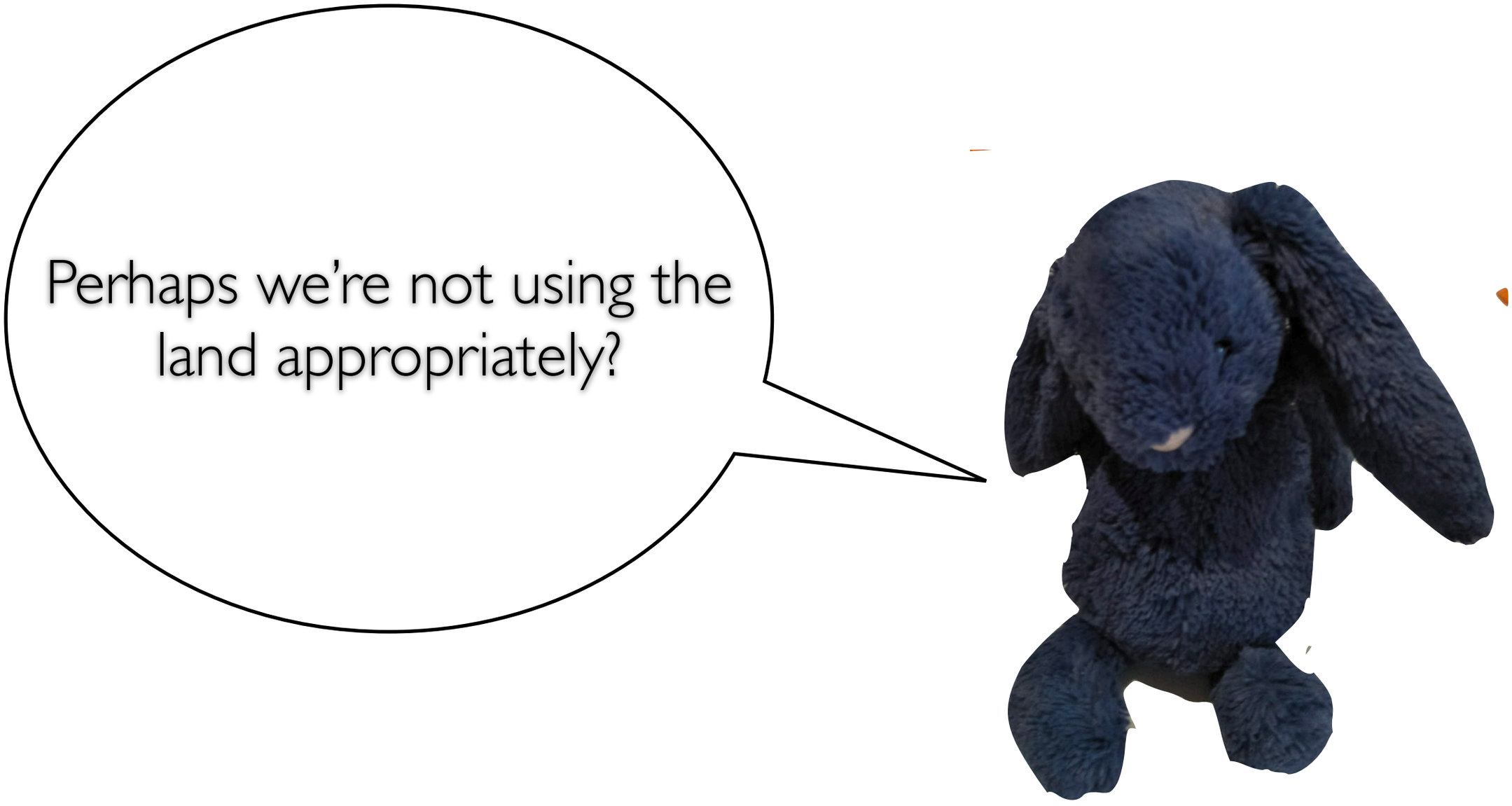
But we have to find a solution. We know that crop yields are doing down but we still don't know how to *fix* it. Without a solution first, how will we stop them?

Let’s go back to our story first…

Let’s go back to our story first…

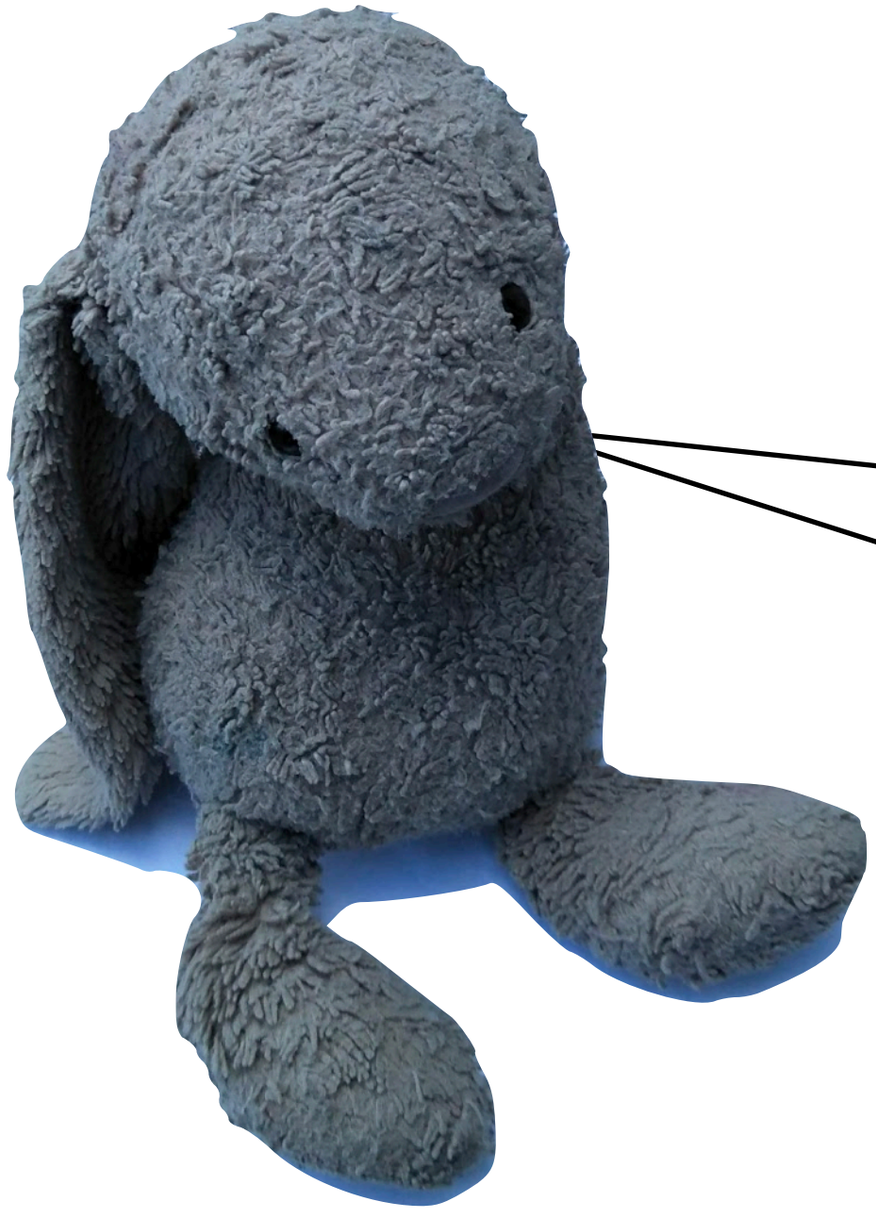
So they do; armed only with datasets, the group sets out to stop their friends before it’s too late!

They analyse frantically as they do so



They analyse frantically as they do so

They analyse frantically as they do so



We could see if we're using the land in the way we're supposed to? Ranching on pasture, growing grains on rich land, and fruit on hilly?

Good idea! Three kinds of land:

pasture

cows, sheep

rich

corn, wheat

hilly

berries, carrots

If they’re farming it appropriately, different foods should be concentrated on different kinds of land

The data

- One tibble, `d`
- Multiple variables:
  - `plot`... code uniquely identifying each plot of land
  - `type`... type of land: pasture, rich, or hilly
  - `cows, berries, corn`... units of each food
  - `time`... whether the data is from 15yrs ago (old) or now (new)

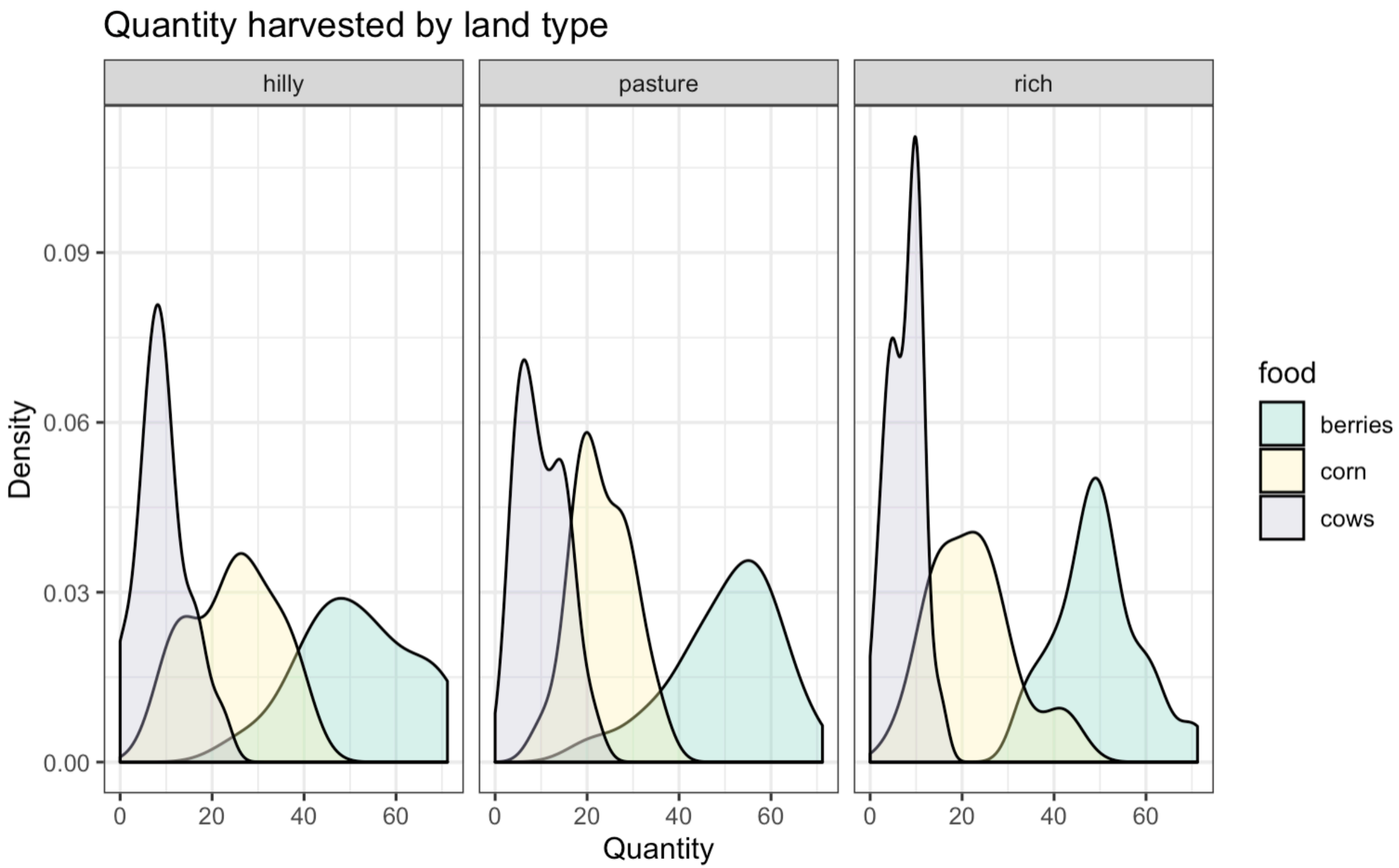
```
> head(d)
  plot     type    time   cows berries  corn
1 9Jx21zaa pasture old       6    58    9.96
2 Qp72PepB pasture old      12    33.0 26.2
3 5YIxvYbz pasture old      15    43.3 27.7
4 OnZuUW5M pasture old      14    55.0 23.2
5 rLMa3j90 pasture old       7    53.8 21.1
6 k3Hb2fUa pasture old      22    31.6 25.9
```

For our sanity (for now) let's break this up into two different tibbles

- Called `d_new` and `d_old`
- Each has five variables:
  - `plot`... code uniquely identifying each plot of land
  - `type`... type of land: pasture, rich, or hilly
  - `cows, berries, corn`... units of each food

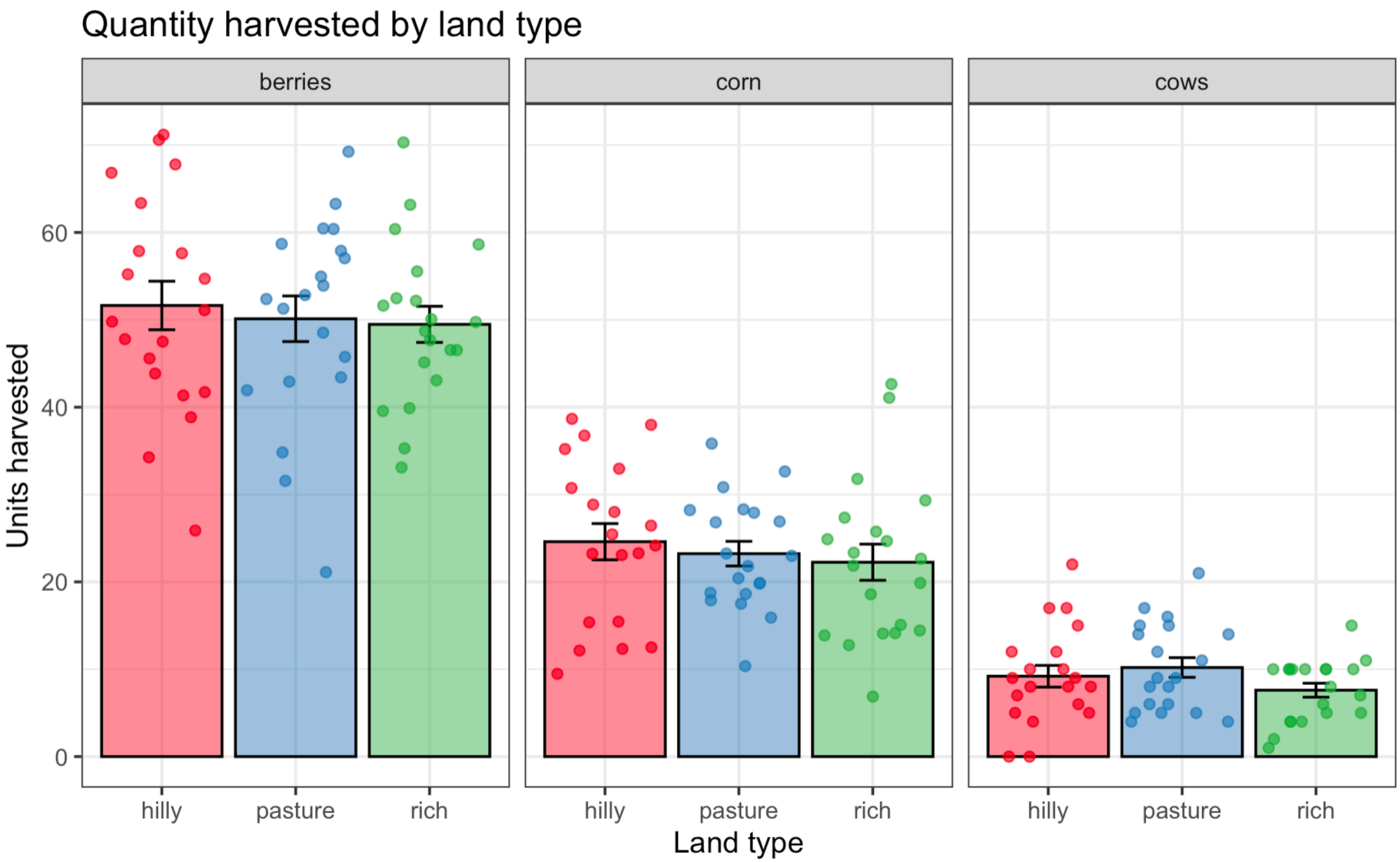
```
  plot     type     cows berries  corn
1 9Jx21zaa pasture    4    58.7  28.2
2 Qp72PepB pasture   14    57.1  28.3
3 5YIxvYbz pasture   11    31.6  27.9
4 OnZuUW5M pasture    9    34.8  19.9
5 rLMa3j90 pasture    4    52.8  23.0
6 k3Hb2fUa pasture   16    57.9  35.8
```

Plotting the data: most recent (new)



d_new

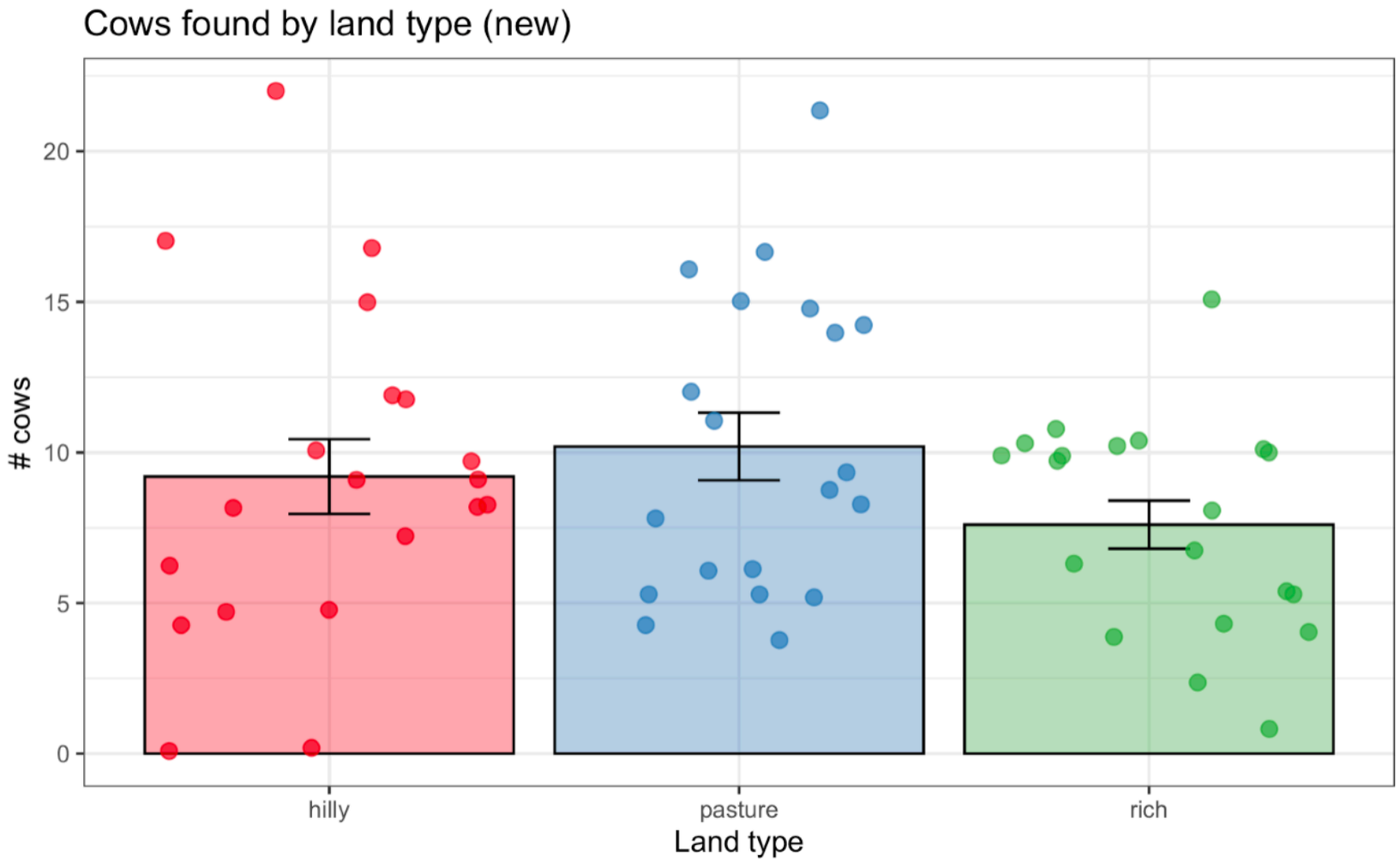
Plotting the data: most recent (new)



d_new

Zoom in on just one of these…

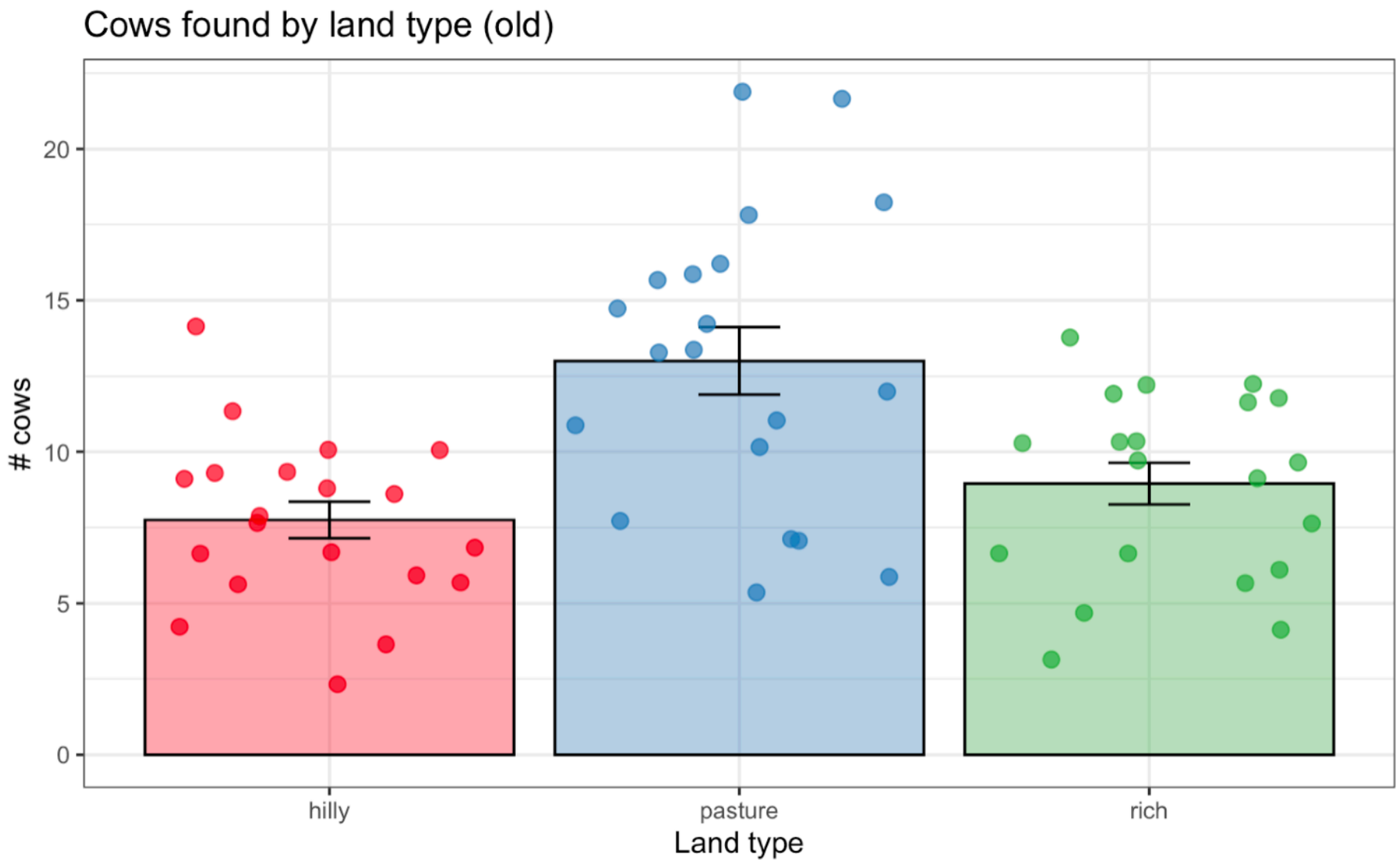
- Are there significantly more cows on pasture right now, as you'd expect if people were allocating land sensibly?



d_new

We can ask the same question for `d_old`

- Were there significantly more cows on pasture 15 years go, as you'd expect if people were allocating land sensibly?



`d_old`

…but for now let’s consider just the new data

- Are there significantly more cows on pasture right now, as you’d expect if people were allocating land sensibly?

- Okay... so, let's build a statistical test:
  - **Null**: all land types have the same # of cows
  - **Alternative**: some land types have more cows than others

Let’s construct a test

1) A diagnostic test statistic, *T*
2) Sampling distribution of *T* if the null is true
3) The observed *T* in your data
4) A rule that maps every value of *T* onto a decision (accept or reject H0)

Let’s construct a test

1) A diagnostic test statistic, *T*
2) Sampling distribution of *T* if the null is true
3) The observed *T* in your data
4) A rule that maps every value of *T* onto a decision (accept or reject H0)

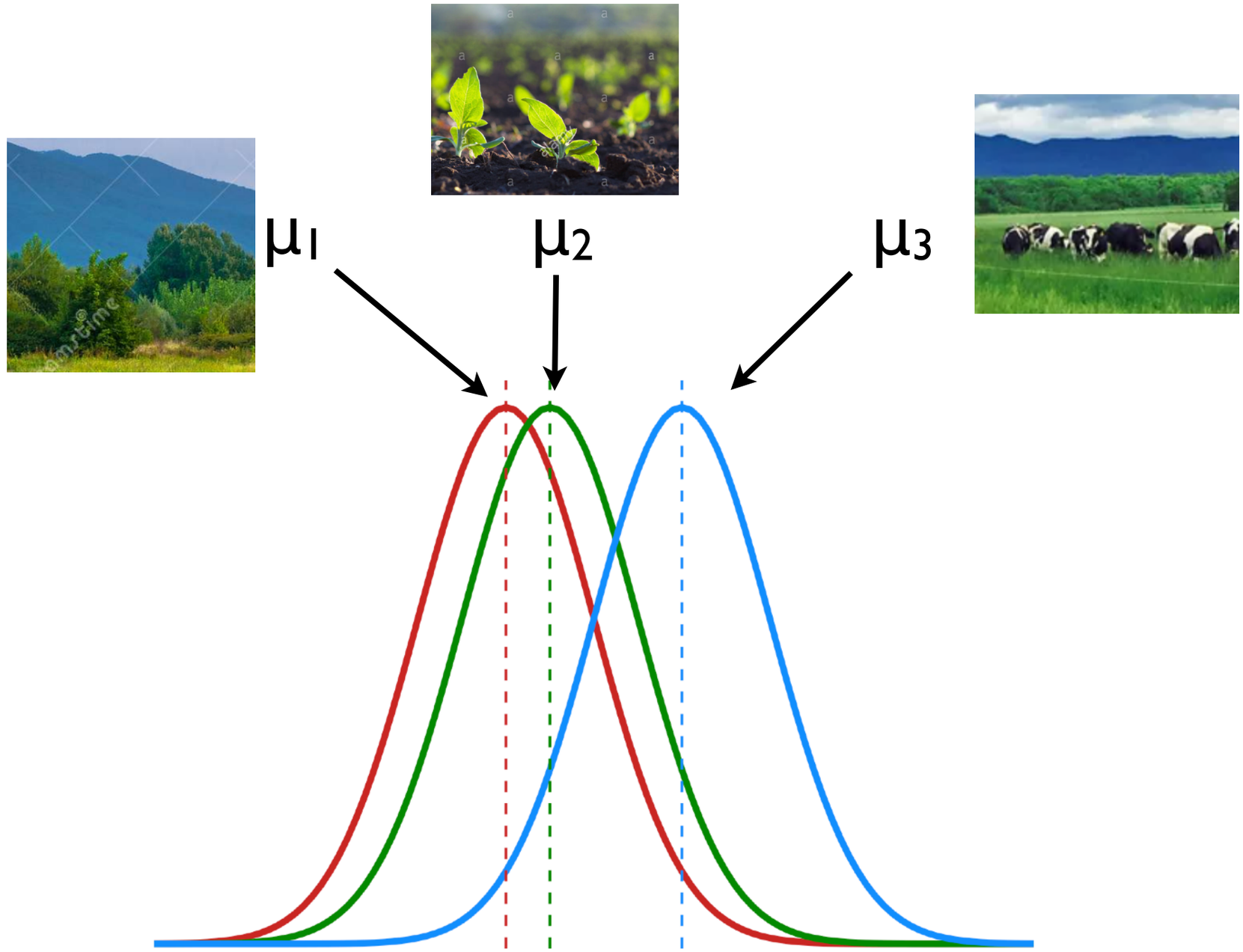
In the beginning was notation...

- Suppose we have G groups
- μ: population grand mean
  - in our case: mean number of cows across all land
- μ_g: population mean for group g
  - for instance, mean number of cows on pasture



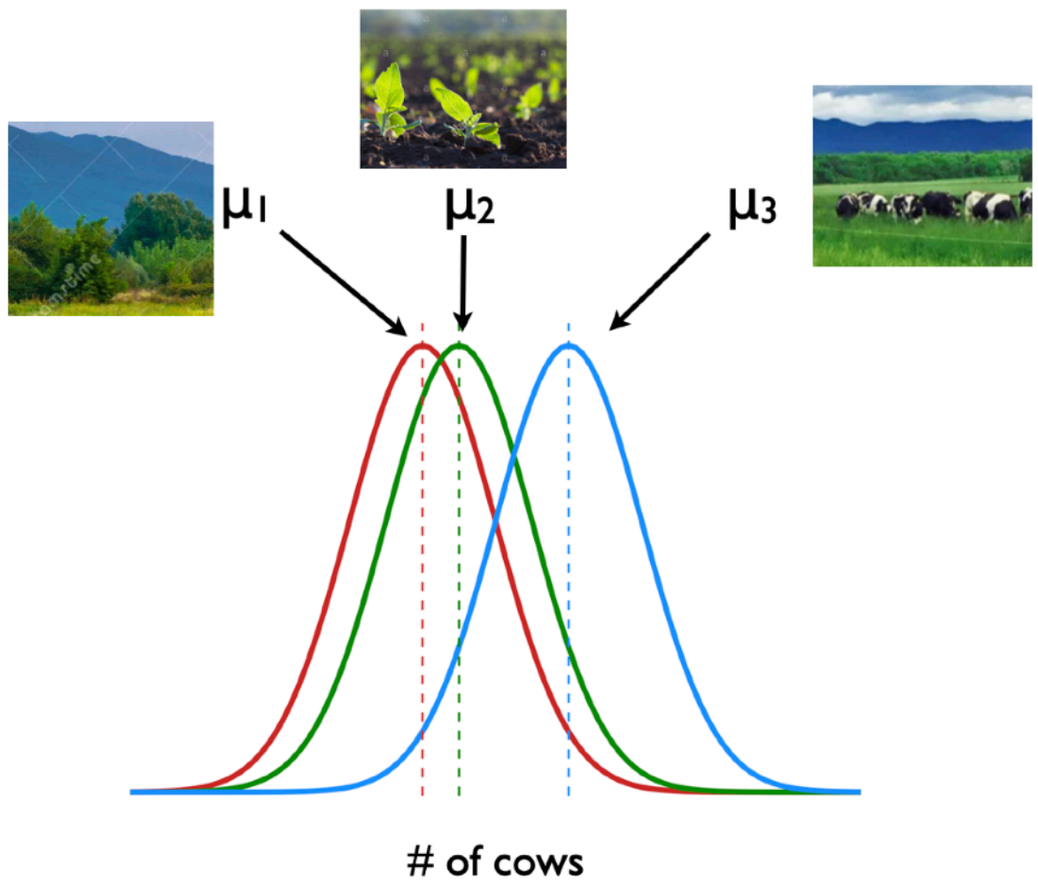
(G = 3 groups)

A picture to illustrate this:



Null and alternative hypotheses

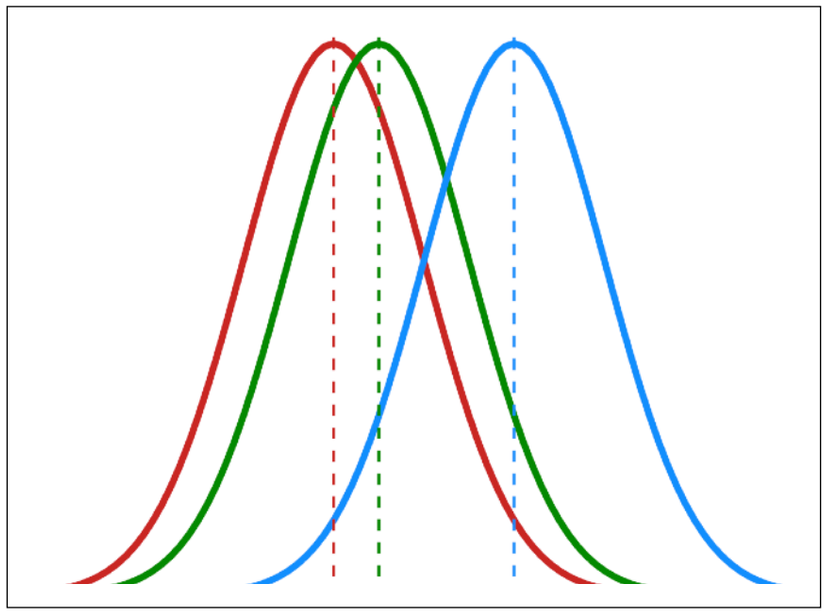
- The null:
  - Population means for all G groups are the same
  - $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$



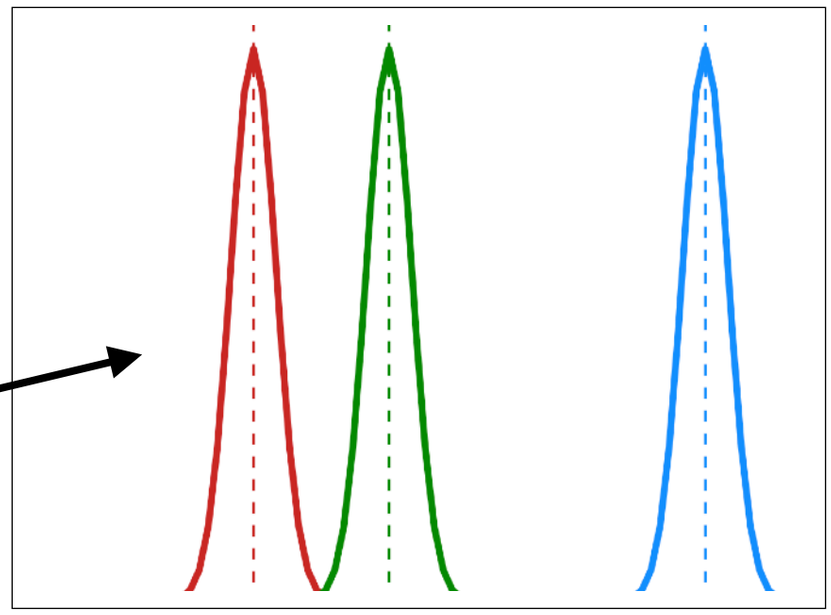
- The alternative:
  - Population means are not all identical
  - Doesn't have a "pretty" equation

Coming up with a test statistic

Intuitively, whether these means are *different* depends on both how far away the means are, as well as on how *variable* the distributions are.



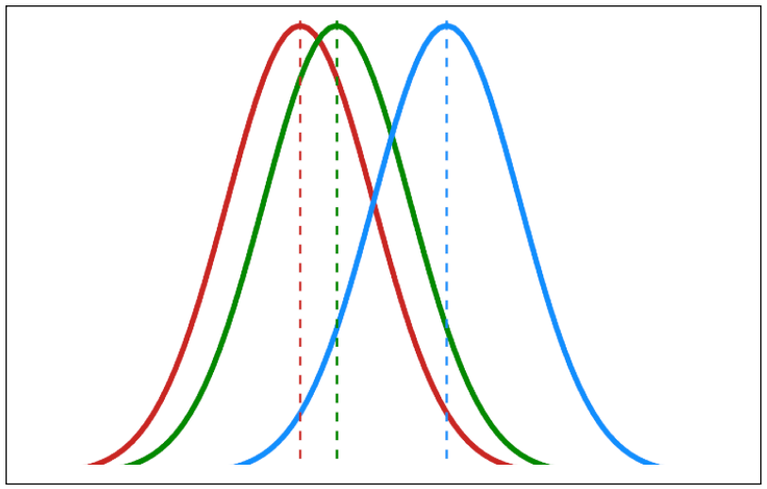
These seem more different



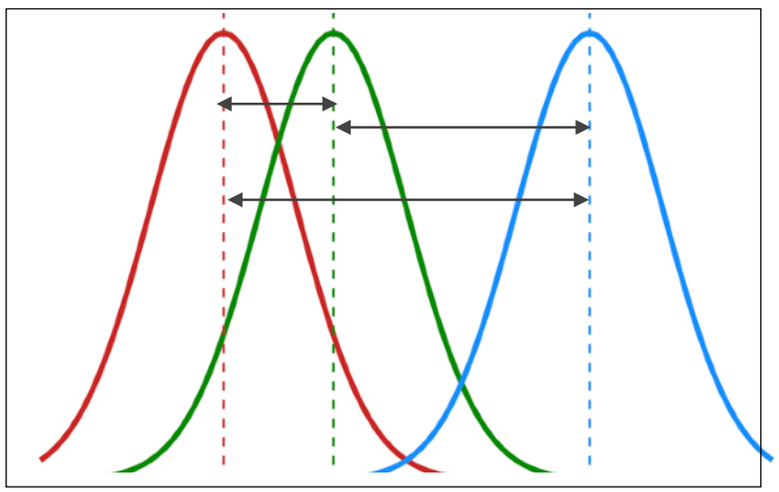
Coming up with a test statistic

Two kinds of variability, both captured with sum of squares (SS)

Between groups (SS_b): how different are the group means from one another?

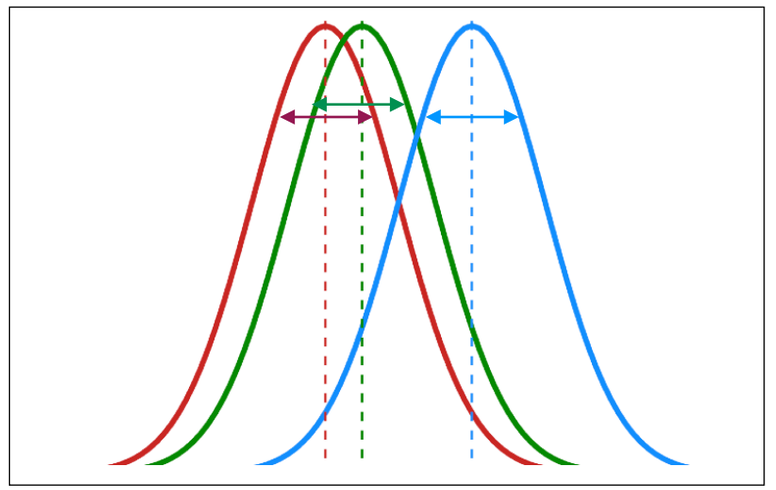


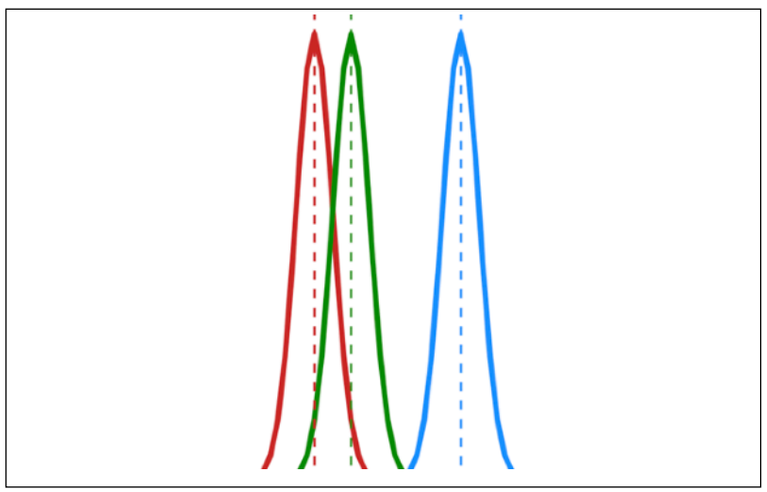
this has larger between-groups variability



Within groups (SS_w): how much do individuals within a group differ from the group mean?

this has larger within-groups variability





Between groups variability

$\bar{X}_1$

mean # of cows on plots of hilly land in the sample

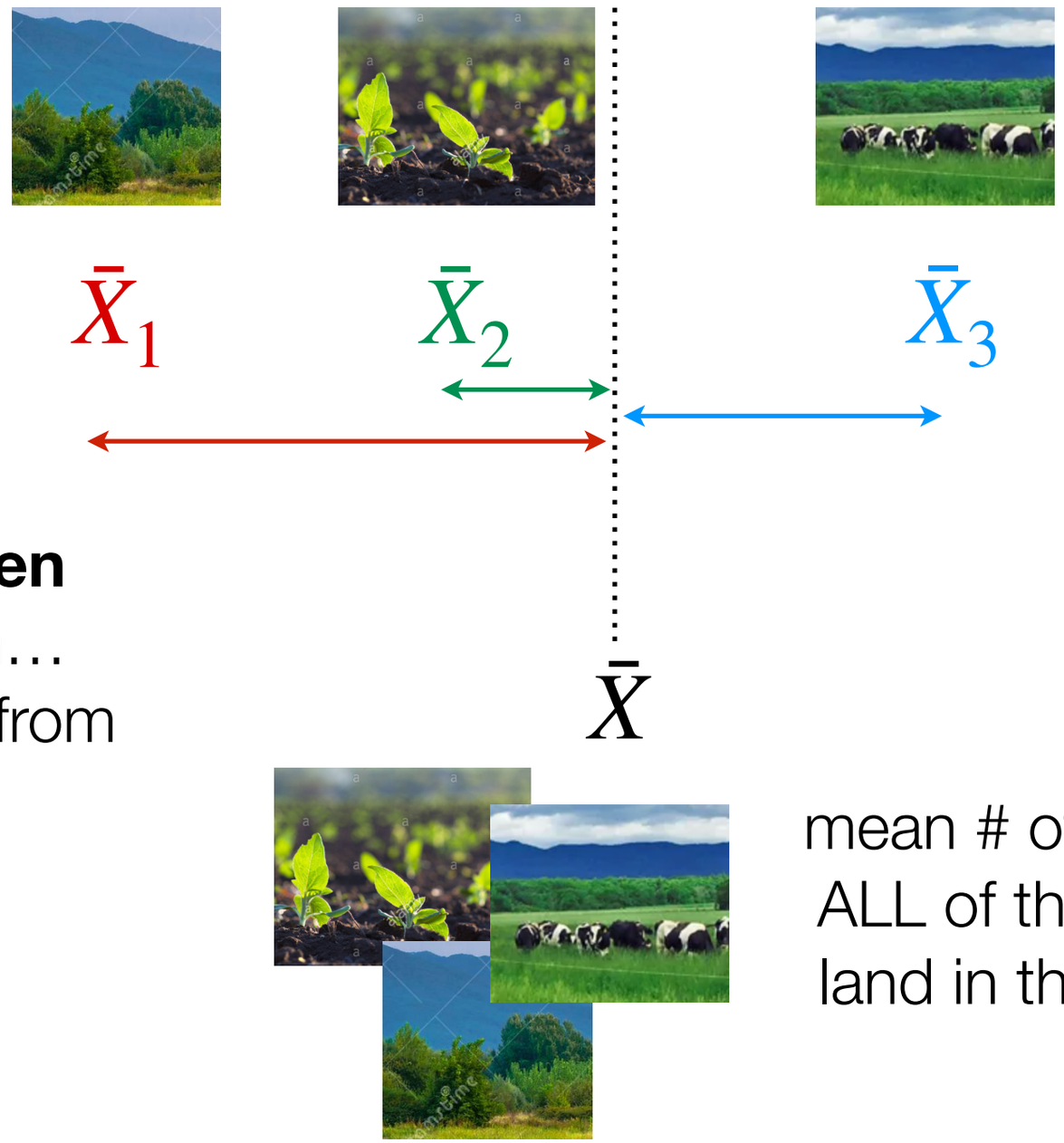
$\bar{X}_2$

mean # of cows on plots of rich land in the sample

$\bar{X}_3$

mean # of cows on plots of pasture in the sample

Between groups variability



This is the **between groups variation**… group means differ from grand means

mean # of cows for ALL of the plots of land in the sample

Between groups variability

Difference between each land type's mean and the global mean, squared

$$\mathrm{SS}_b = \sum_{k=1}^{G} N_k \left(\bar{X}_k - \bar{X}\right)^2$$

Sum this over all of the k land types

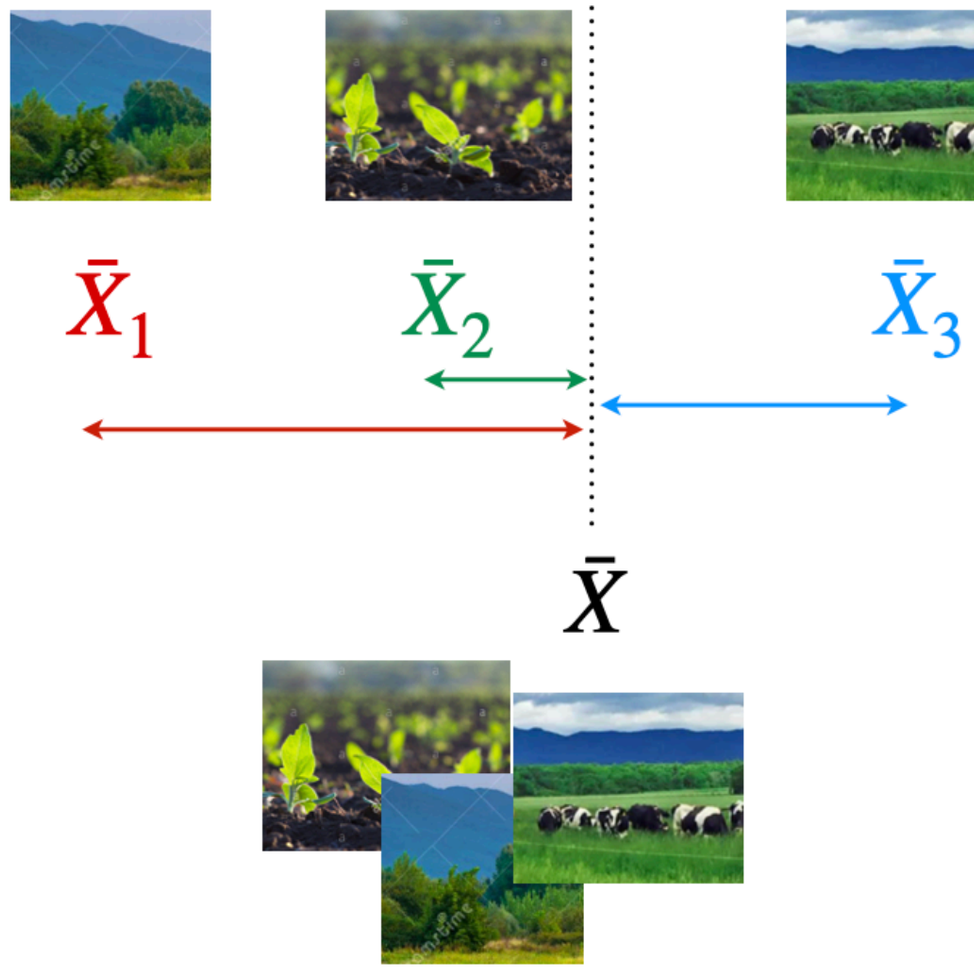
Weight by the sample size of land type k

Mean # of cows for k^{th} land type

Mean # of cows for all land types

Between groups variability

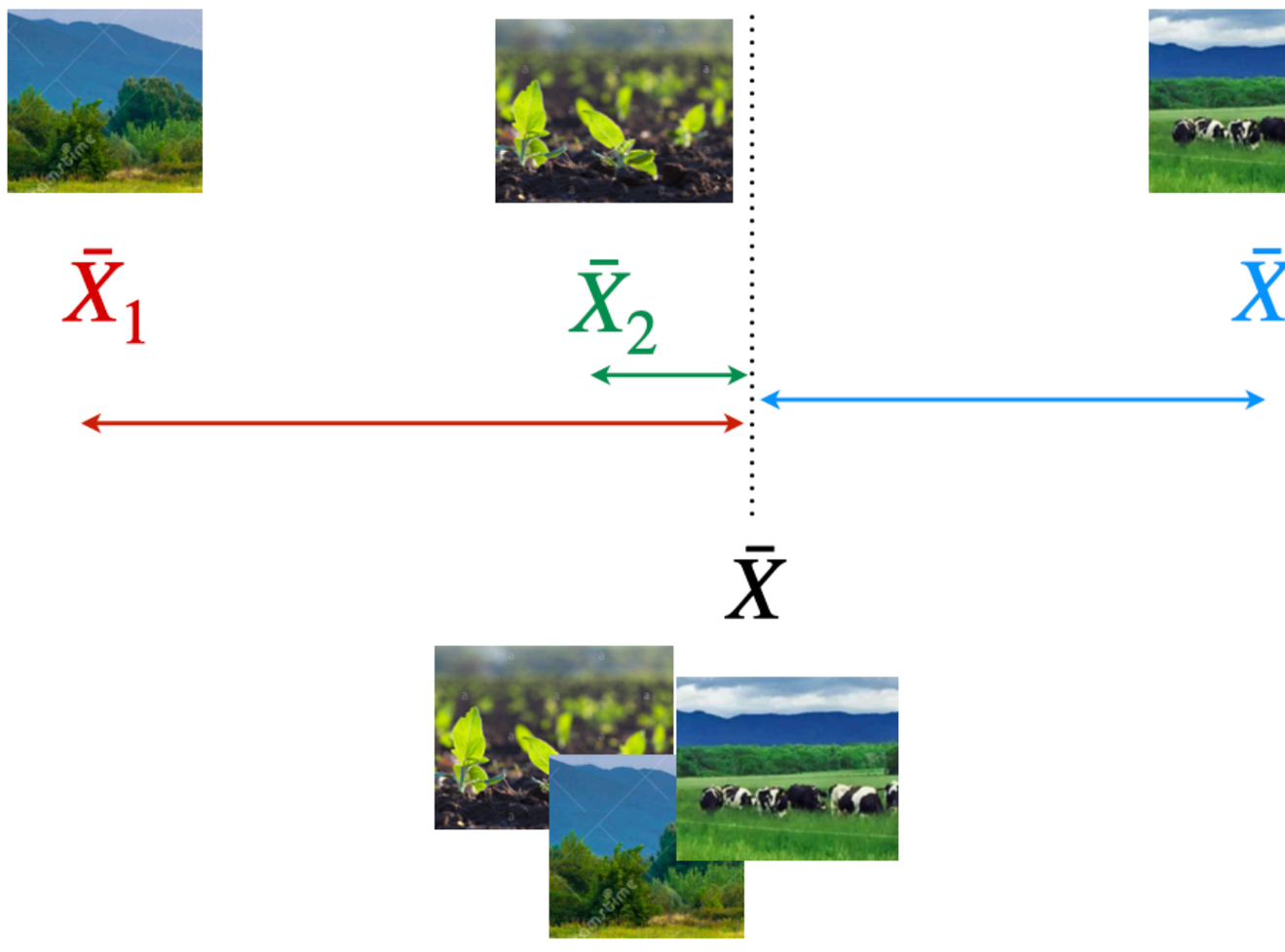
$$\mathrm{SS}_b = \sum_{k=1}^{G} N_k \left(\bar{X}_k - \bar{X}\right)^2$$



Between groups variability

$$\mathrm{SS}_b = \sum_{k=1}^{G} N_k \left(\bar{X}_k - \bar{X}\right)^2$$

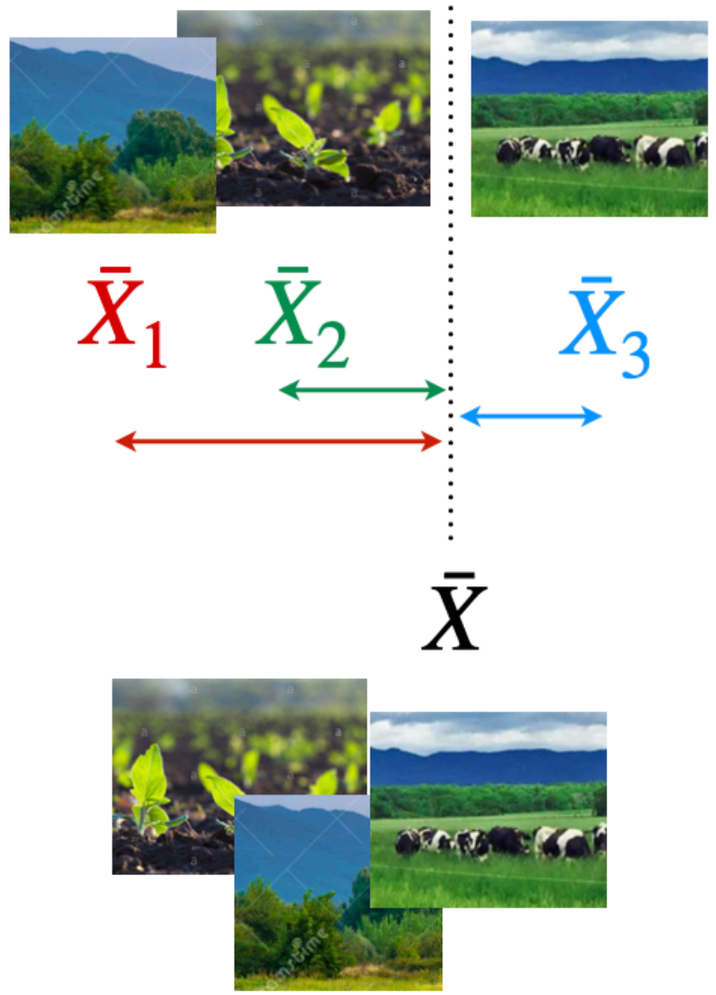
So SS_b is high if the difference between group means and the global mean is high



Between groups variability

$$\mathrm{SS}_b = \sum_{k=1}^{G} N_k \left(\bar{X}_k - \bar{X}\right)^2$$

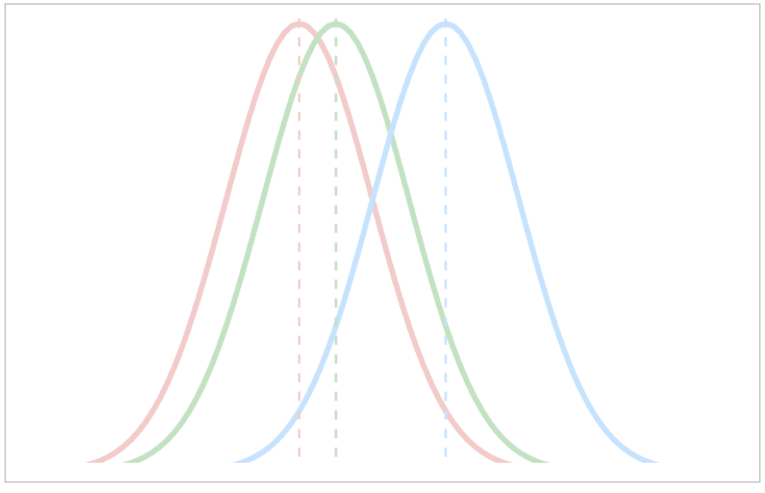
And SS_b is low if the difference between group means and the global mean is low



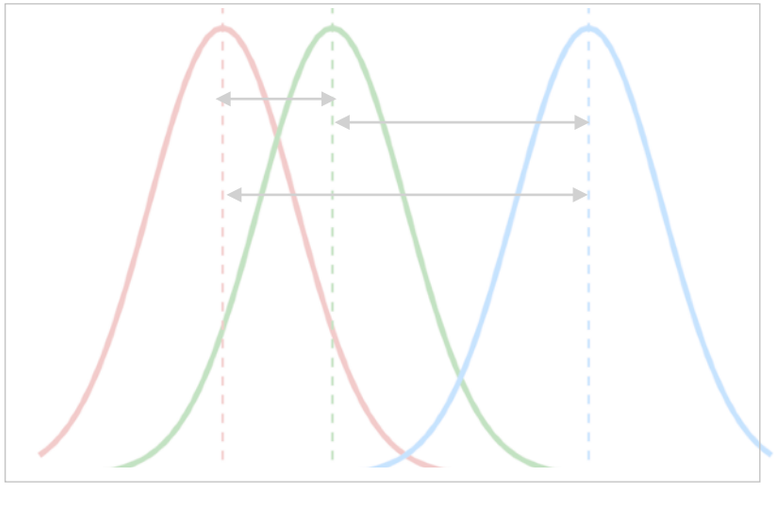
Coming up with a test statistic

Two kinds of variability, both captured with sum of squares (SS)

Between groups (SS_b): how different are the group means from one another?

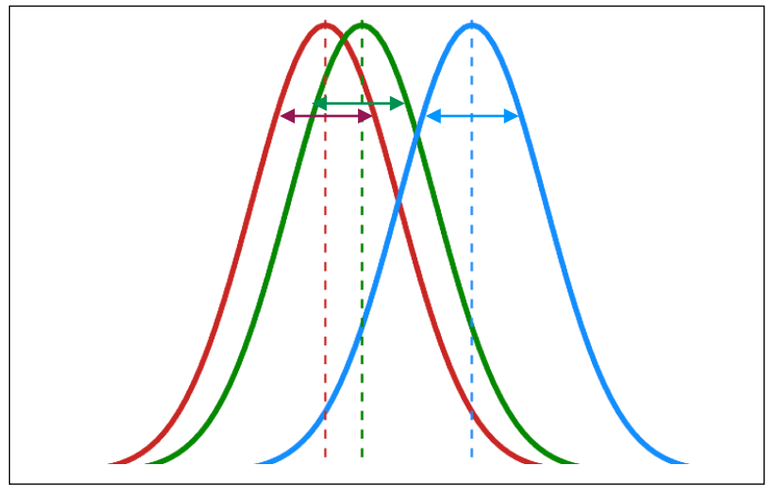


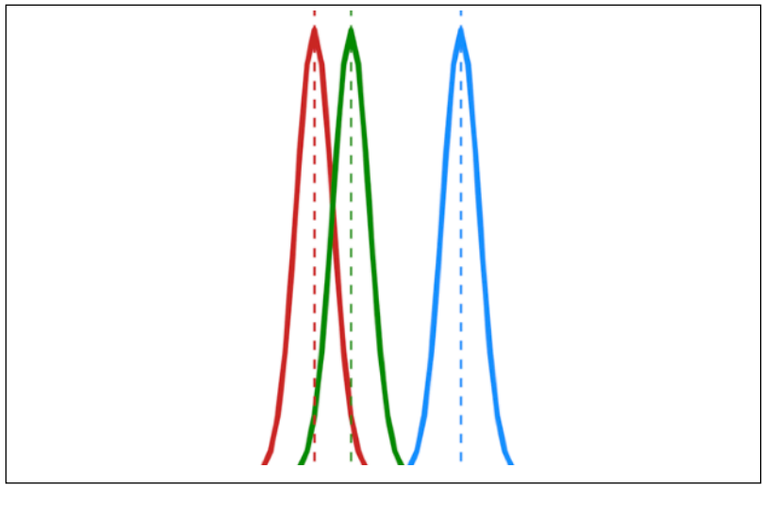
this has larger between-groups variability



Within groups (SS_w): how much do individuals within a group differ from the group mean?

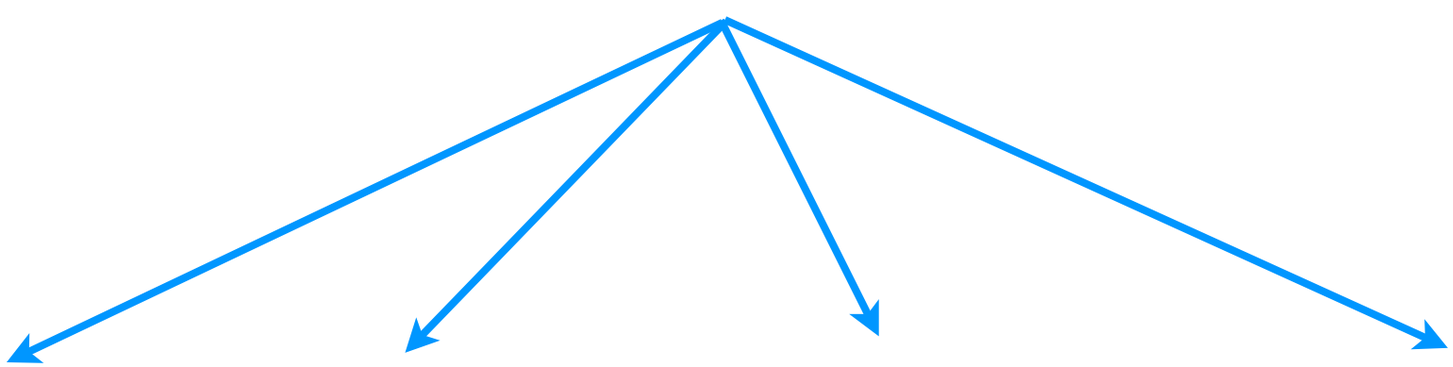
this has larger within-groups variability



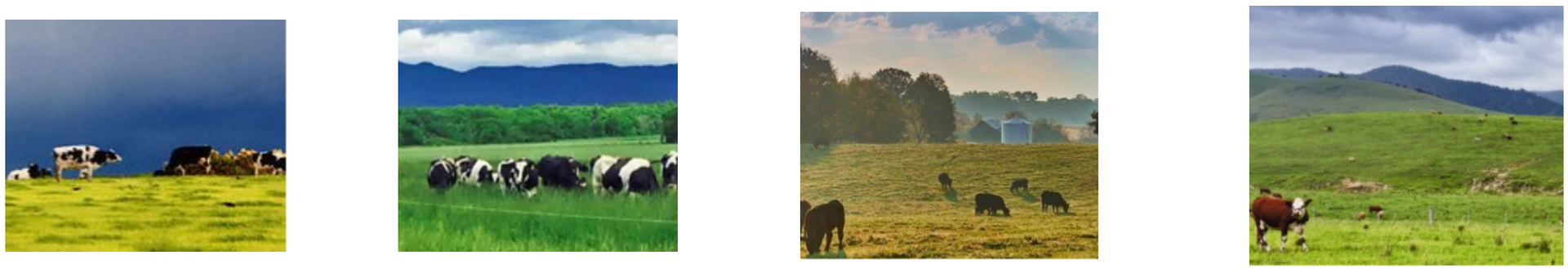


Within groups variability (residuals)

The # of cows on individual plots of pasture (group 3) aren't all the same



$\bar{X}_{31}$ $\bar{X}_{32}$ $\bar{X}_{33}$ $\bar{X}_{34}$



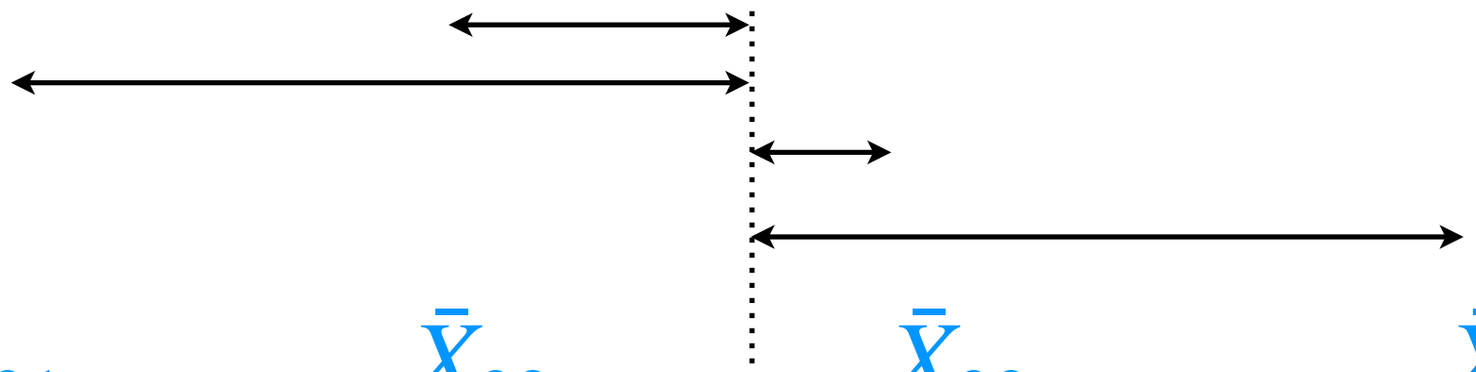
pasture 1 pasture 2 pasture 3 pasture 4

Within groups variability (residuals)

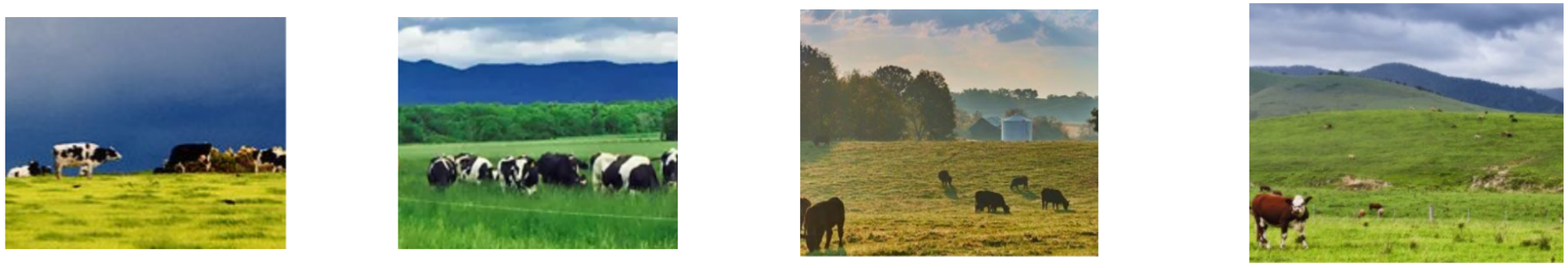
This is the **within groups variation**… individuals differ from the group mean

Here's the group mean:

$\bar{X}_3$



$\bar{X}_{31}$ $\bar{X}_{32}$ $\bar{X}_{33}$ $\bar{X}_{34}$



pasture 1 pasture 2 pasture 3 pasture 4

Within groups variability (residuals)

Difference between each individual in a species and the species mean, squared

$$\mathrm{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(X_i - \bar{X}_k\right)^2$$

Sum this over all of the k land types

Sum this over all individuals plots of land in land type k

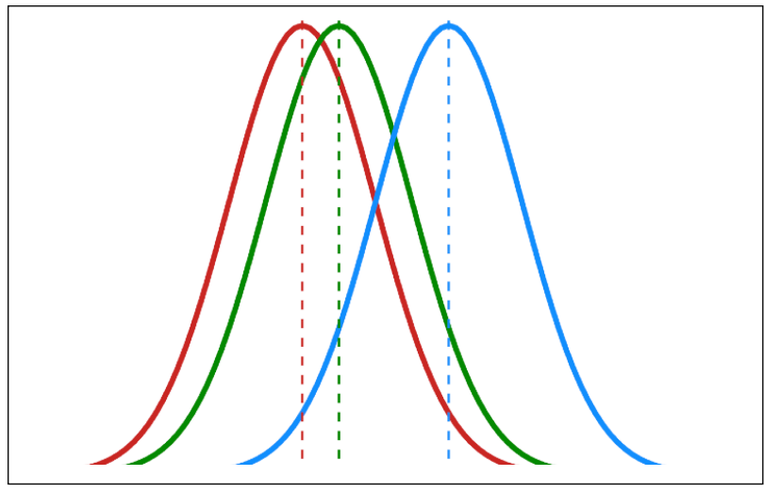
# of cows for the i^{th} individual plot of land in the k^{th} group

Mean # of cows for land type k

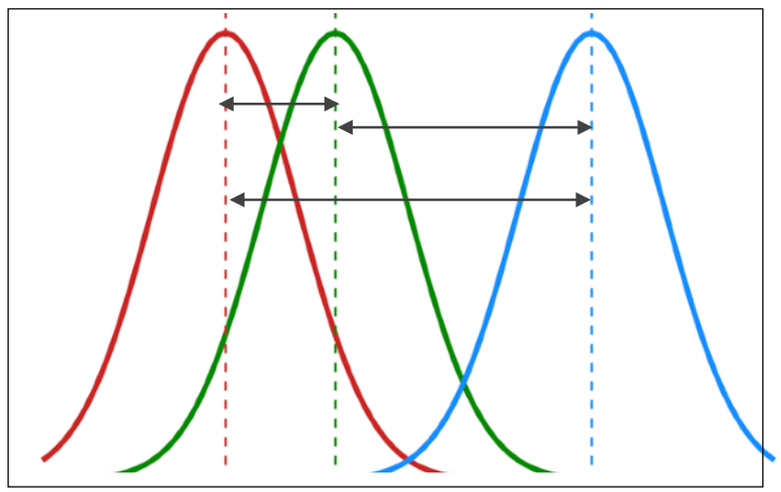
Coming up with a test statistic

Total variability: $SS_{tot} = SS_b + SS_w$

Between groups (SS_b): how different are the group means from one another?

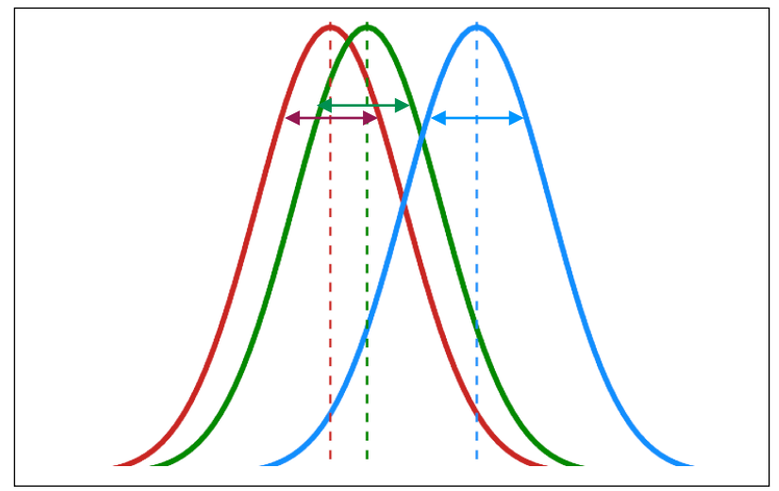


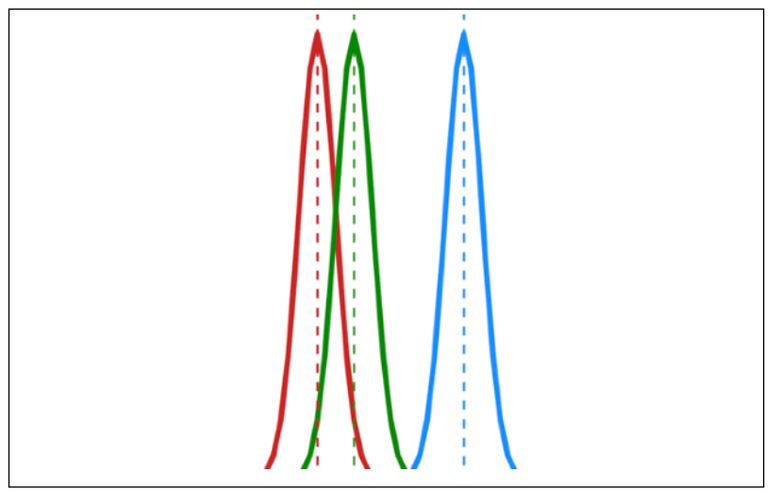
this has larger between-groups variability



Within groups (SS_w): how much do individuals within a group differ from the group mean?

this has larger within-groups variability





Coming up with a test statistic

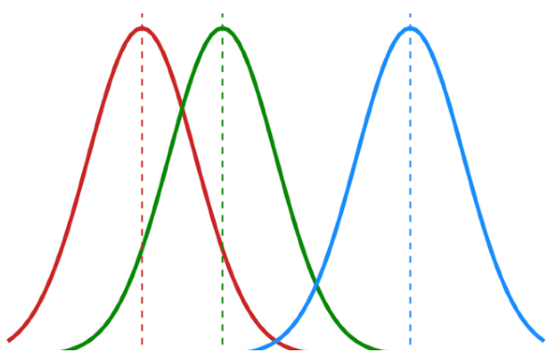
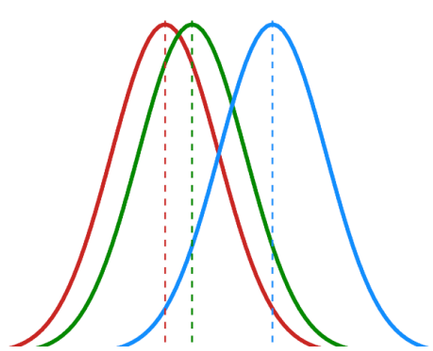
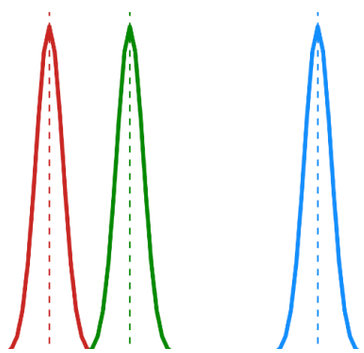
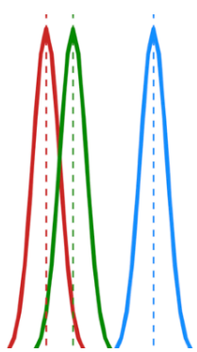
Total variability: $SS_{tot} = SS_b + SS_w$

Is this our test statistic?

NO. Intuitively, it is the *relationship* between them that matters

Between-groups variability

Within-groups variability

	high	low
high		
low	 *These means are different!*	

Maybe a ratio?

$$\frac{SS_b}{SS_w}$$

Close… but we have to correct for *degrees of freedom*

Coming up with a test statistic

Correcting for degrees of freedom

- **Between groups** (SS_b): ***G - 1***
  - G separate group means
  - 1 grand mean from which they deviate

Divide SS_b by degrees of freedom:

$$MS_b = \frac{SS_b}{G - 1}$$

- **Within groups** (SS_w): ***N - G***
  - N total observations
  - G group means from which they deviate

Divide SS_w by degrees of freedom:

$$MS_w = \frac{SS_w}{N - G}$$

Our test statistic is the ratio of these!!!

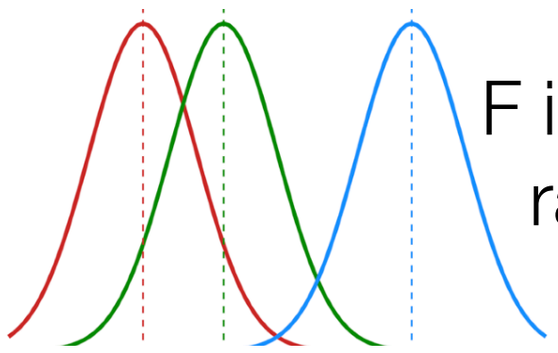
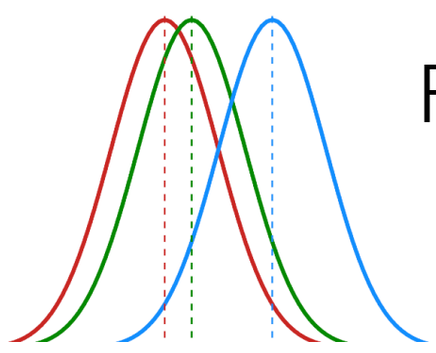
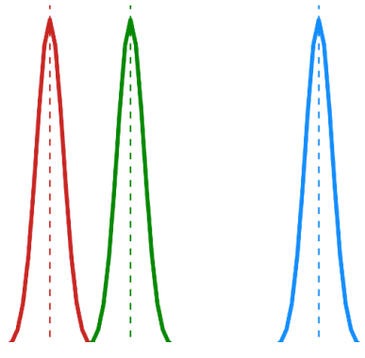
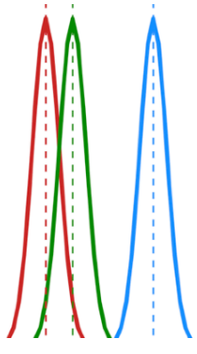
$$F = \frac{MS_b}{MS_w}$$

Coming up with a test statistic

Let's get some intuitions about *F*

$$F = \frac{MS_b}{MS_w}$$

Between-groups variability

Within-groups variability	high	low
high	 F is mid-range	 F is very small
low	 F is large *These means are different!*	 F is mid-range

So the means are more different when F is larger

Coming up with a test statistic

✔ 1) A diagnostic test statistic, *T*

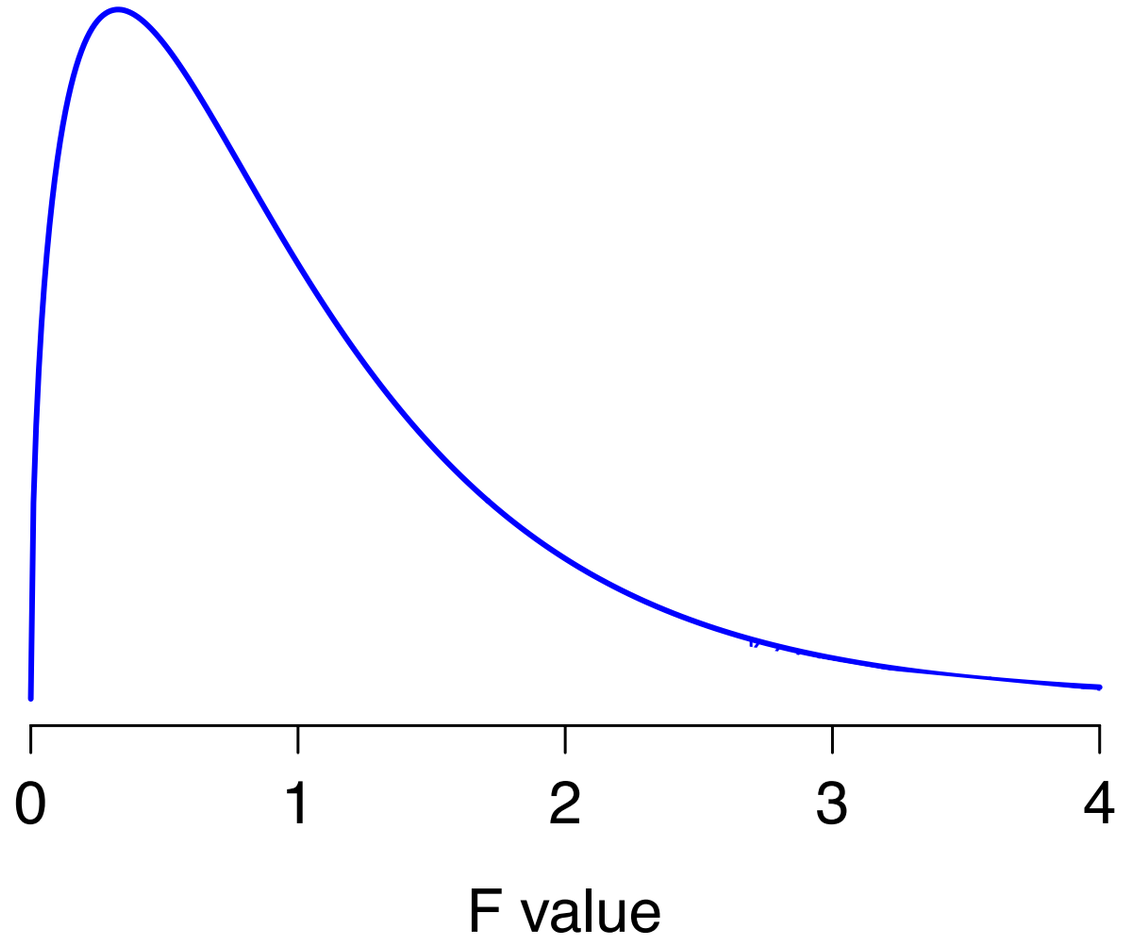
$$F = \frac{MS_b}{MS_w}$$

small when the null is true

2) Sampling distribution of *T* if the null is true
3) The observed *T* in your data
4) A rule that maps every value of *T* onto a decision (accept or reject H0)

Sampling distribution if the null is true

- If the null is true, the sampling distribution of the F-statistic is an F-distribution
  - With N and G as the associated degrees of freedom



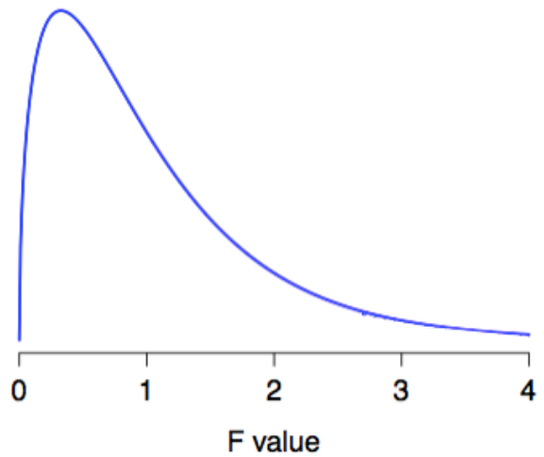
Coming up with a test statistic

✔ 1) A diagnostic test statistic, *T*

$$F = \frac{MS_b}{MS_w}$$

small when the null is true

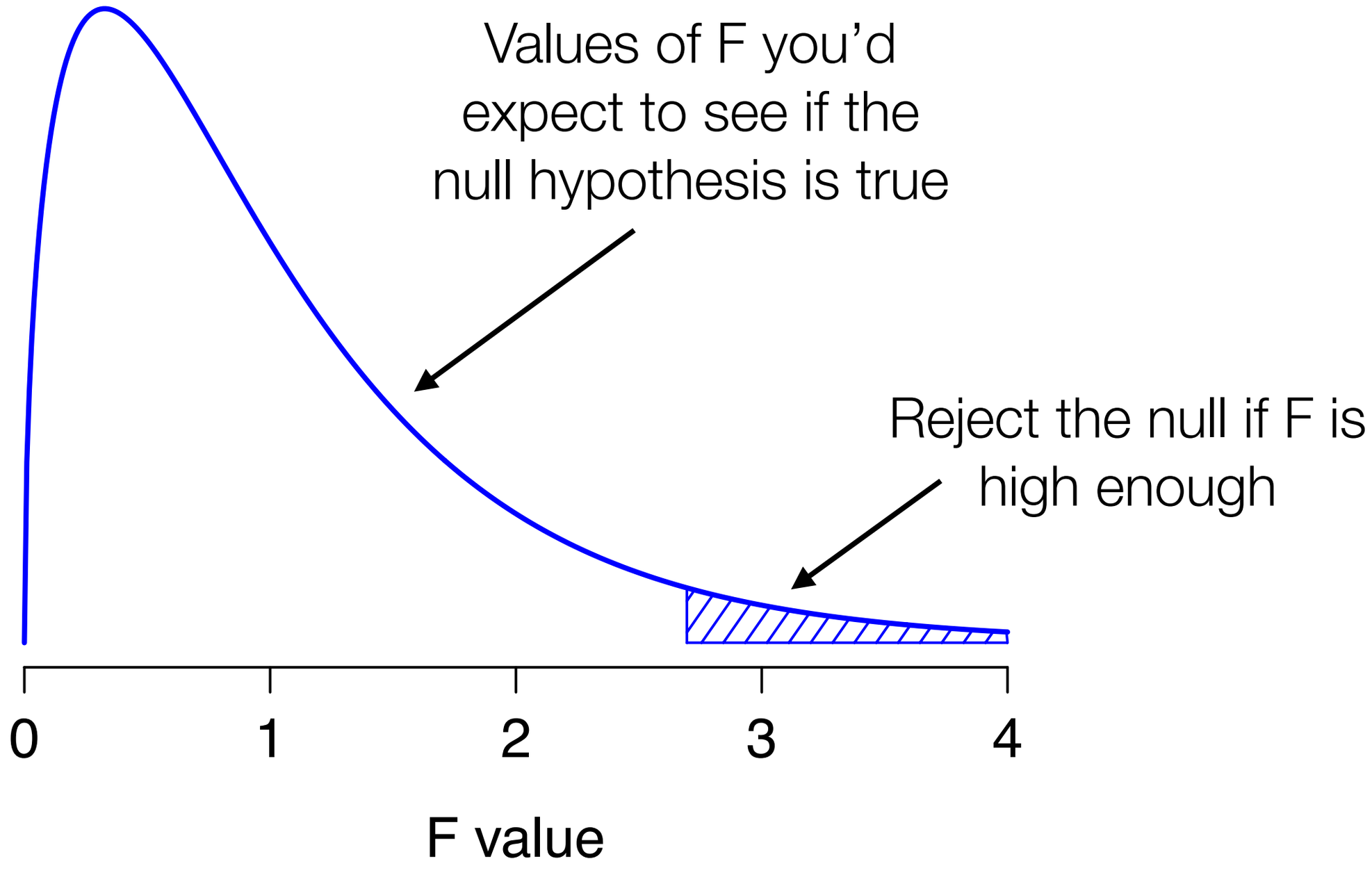
✔ 2) Sampling distribution of *T* if the null is true



F-distribution with N and G degrees of freedom

3) The observed *T* in your data
4) A rule that maps every value of *T* onto a decision (accept or reject H0)

Decision rule



How to write up the results...

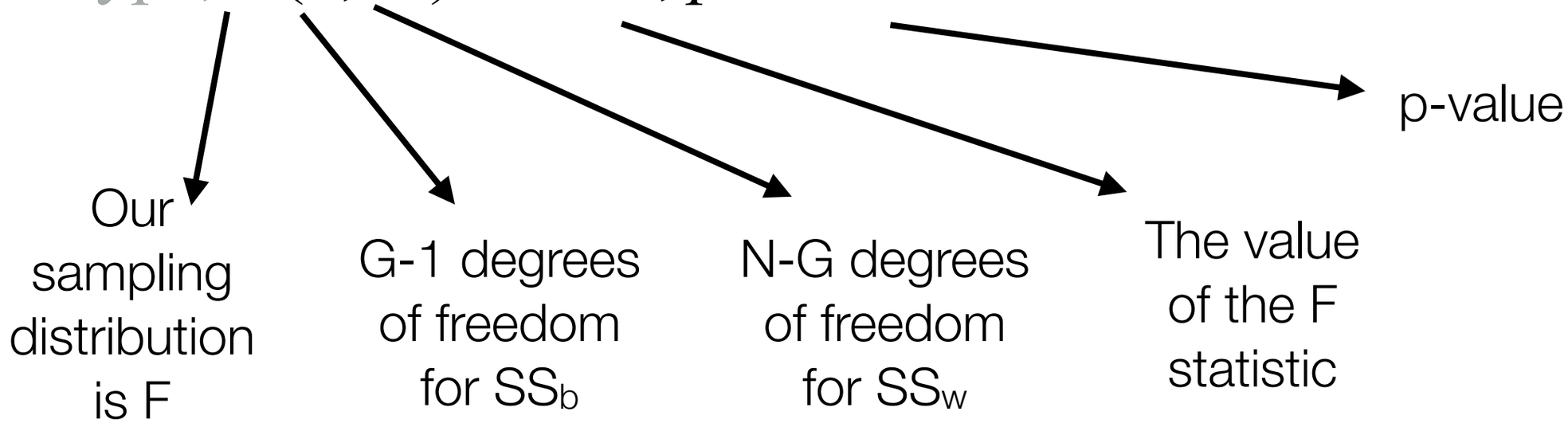
In our experiment, the mean number of cows on each plot of land was 9.2 on hilly land (SD=5.56), 7.6 on rich land (SD=3.57), and 10.2 on pasture (SD=5.03).

A one-way ANOVA indicated that there were not significant differences in the number of cows by land type, $F(2,57) = 1.50$, $p = .233$.

How to write up the results...

In our experiment, the mean number of cows on each plot of land was 9.2 on hilly land (SD=5.56), 7.6 on rich land (SD=3.57), and 10.2 on pasture (SD=5.03).

A one-way ANOVA indicated that there were not significant differences in the number of cows by land type, $F(2,57) = 1.50$, $p = .233$.



Exercises are in w8day1exercises.Rmd