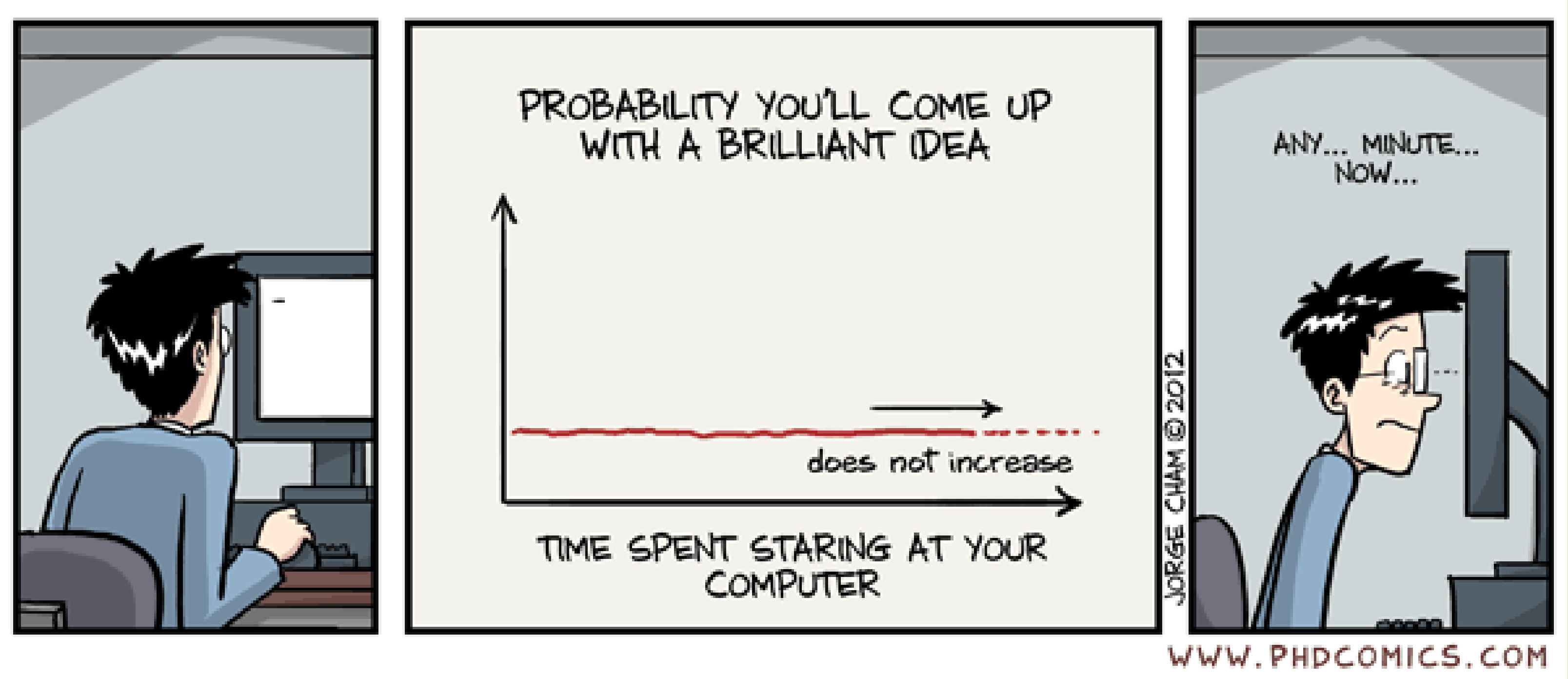

PROBABILITY YOU'LL COME UP WITH A BRILLIANT IDEA
does not increase
TIME SPENT STARING AT YOUR COMPUTER
JORGE CHAM © 2012
ANY... MINUTE... NOW...
WWW.PHDCOMICS.COM

Quantitative Thinking in the Life Sciences

October 24th – Linking probability, mathematical functions and data
Part 3

Today

- Concept maps – Data distributions
- Simple mathematical relationships and probability
- Assignment B
- More R fun!
    - R code questions?
    - Looking at snail vectors!

Housekeeping

- November 14^{th} absence
- After today only four class sessions left
- Homework A is due today
- Homework B is due on Nov 1^{st}
  - First attempt at simulating your data distributions
  - No new R chapter – catch up on existing R code!

My homework: Probability vs likelihood

- Data from a known distribution (normal) and parameters characterizing distribution(e.g., mean and sd)
- The **probability** of observing any data point would be based on the known parameters
- In our work, we will have data but will not know the exact distribution or the distribution parameters
- Given an assumed model distribution, the **likelihood** is defined as the probability of observed data as a function of the distribution parameters (e.g., mean and sd)
- In this case, the data are known, but distribution parameters are unknown
- The motivation for defining the likelihood is to determine the parameters of the distribution
- The likelihood function is not bound between 0 and 1 (unlike probabilities)
- The likelihood function is proportional to the probability of the observed data

stats.stackexchange.com/questions/2641/what-is-the-difference-between-likelihood-and-probability

Probability vs likelihood

- The likelihood of this model, given the data
- The probability of observing similar data given the model

Brief recap:
Probability to statistical modeling

Rolling two dice

- Two six-sided dice with sides numbered 1-6
- Likelihood of the dice landing on any of 6 numbers is equal
- All die rolls are independent

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Sum on dice

2: One possibility (1,1) probability = 1/36 options

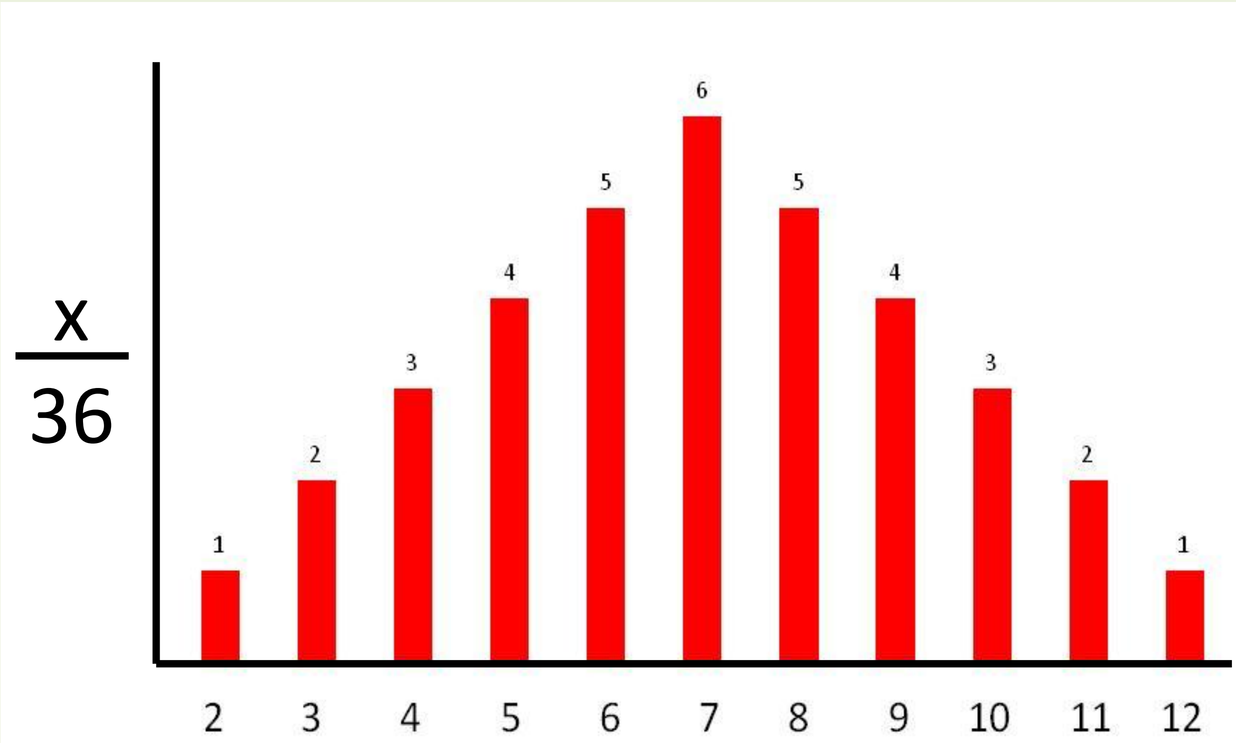
3: Two possibilities (1,2) & (2,1) probability = 2/36 options

4: Three possibilities (1,3), (2,2) & (3,1) probability = 3/36 options

⋮

7: Six possibilities (1,6), (2,5), (3,4), (4,3), (5,2) & (6,1) probability = 6/36 options

Probability space



(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

$1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 + 5/36 + 4/36 + 3/36 + 2/36 + 1/36 = \mathbf{1}$

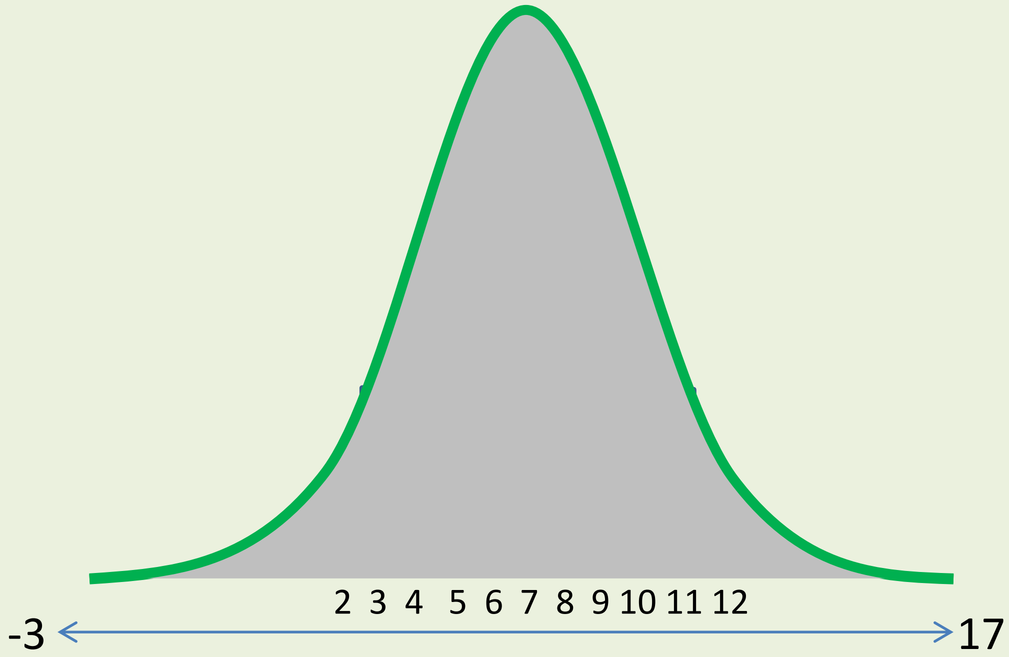
Probability space for rolling x dice

Dice	Combinations	Probability of any one combination	Range of values: Sum of dice
1 Die	6	0.167	Sum of dice: 1-6
2 Dice	36 	0.0278	Sum of dice: 2-12
3 Dice	216 	0.00463	Sum of dice: 3-18
4 Dice	1296	0.000772	Sum of dice: 4-24
5 Dice	7776	0.000129	.
6 Dice	46656	0.0000214	.
7 Dice	279936	0.00000357	.
8 Dice	1679616	0.000000595	
9 Dice	10077696	0.0000000992	
10 Dice	60466176	0.0000000165	
11 Dice	362797056	0.00000000276	
12 Dice	2176782336	0.000000000459	
13 Dice	13060694016	0.0000000000766	
14 Dice	78364164096	0.0000000000128	Sum of dice: 14-82

Combinations * Probability of occurrence of each = 1

78364164096 * 0.0000000000128 = 1

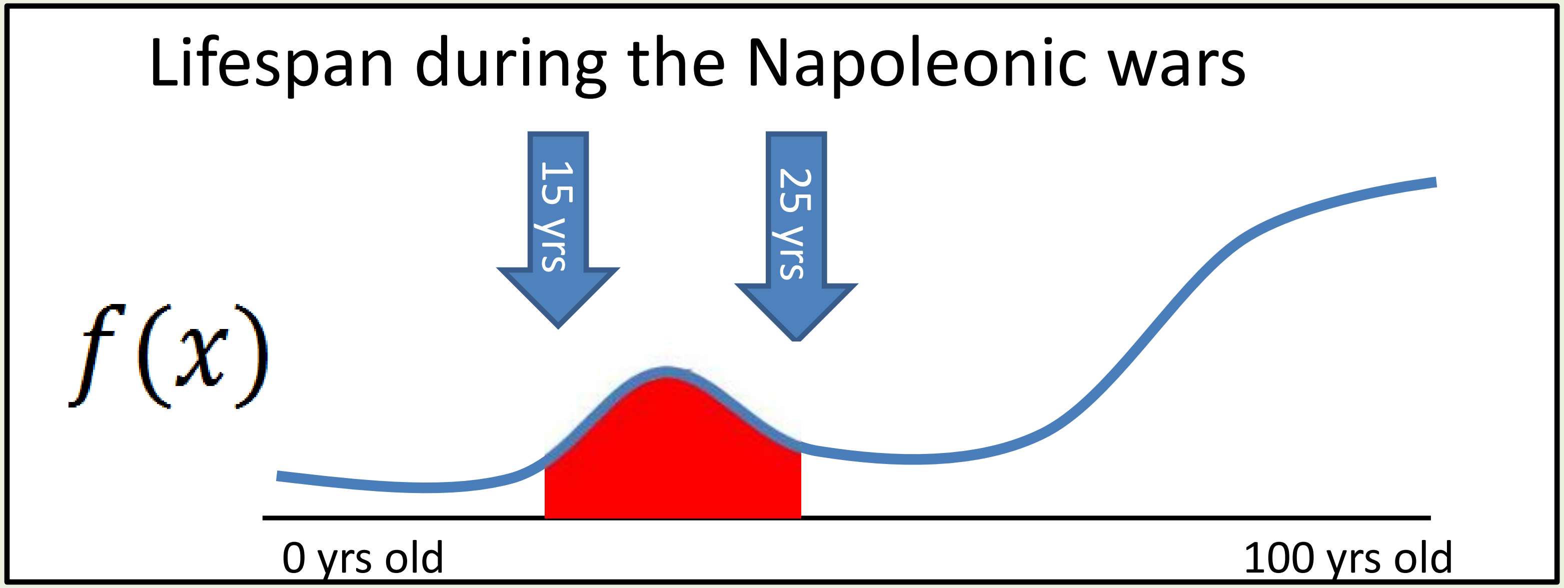
Discrete to continuous probability



Area under the curve is the continuous probability space

- Total area is equal to 1
- All the possible values are under the curve

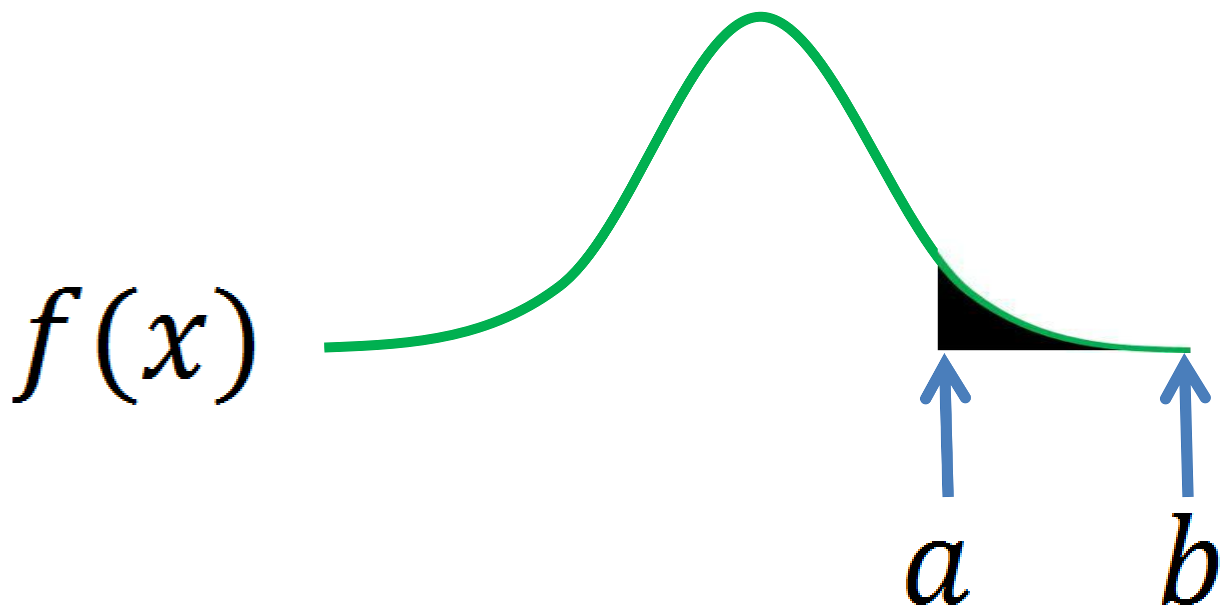
Probability example



$$\int_{15}^{25} f(x)\,dx$$ = probability of dying between 15 and 25 years old

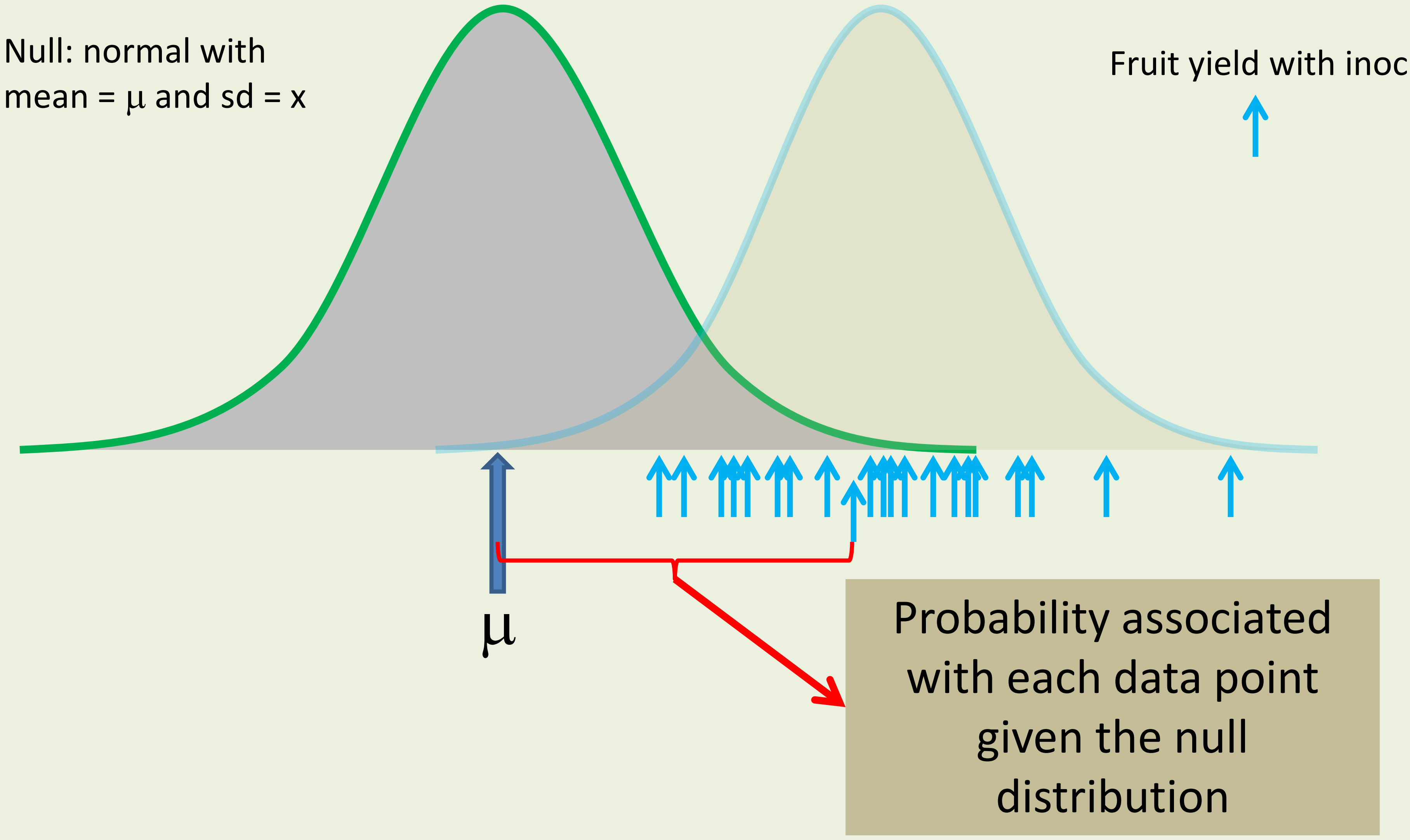
Hypothesis testing – frequentist approach

The **p-value** is the probability of obtaining a test statistic *at least* as extreme as the one that was actually observed, assuming that the null hypothesis is true.

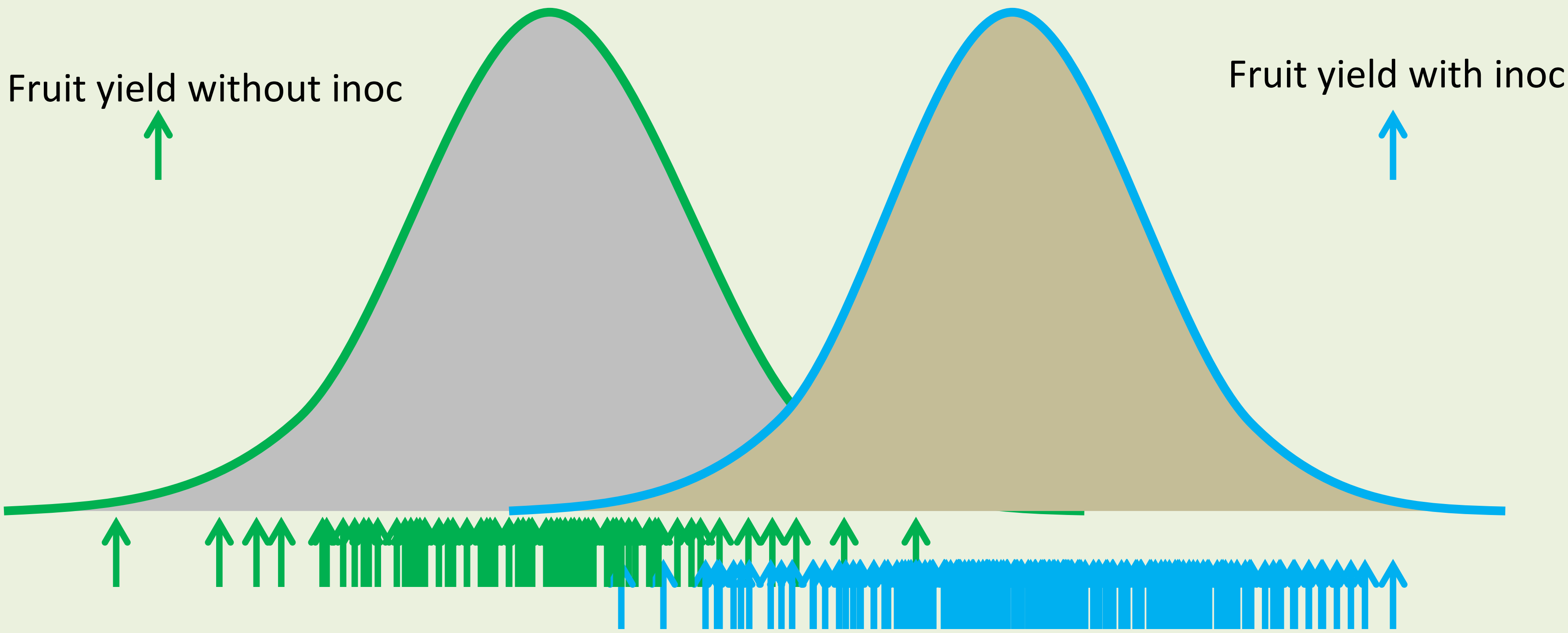


$$\int_a^b f(x)dx = \text{prob(test statistic)}$$

Linking data to the p-value

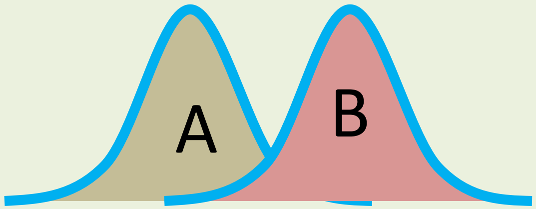


Using those data, and probabilities of observing those data, we can test if distribution A differs from distribution B

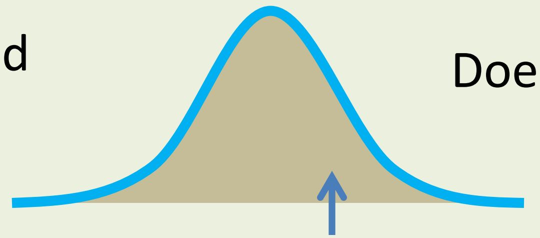


t-test will allow us to test

Test a null hypothesis that two normally distributed populations are equal



Test a null hypothesis that a normally distributed population has a specified mean value



Does $\mu = 5$?

Test a null that there is no difference between two paired or repeated measurements

Miticide Trial Data

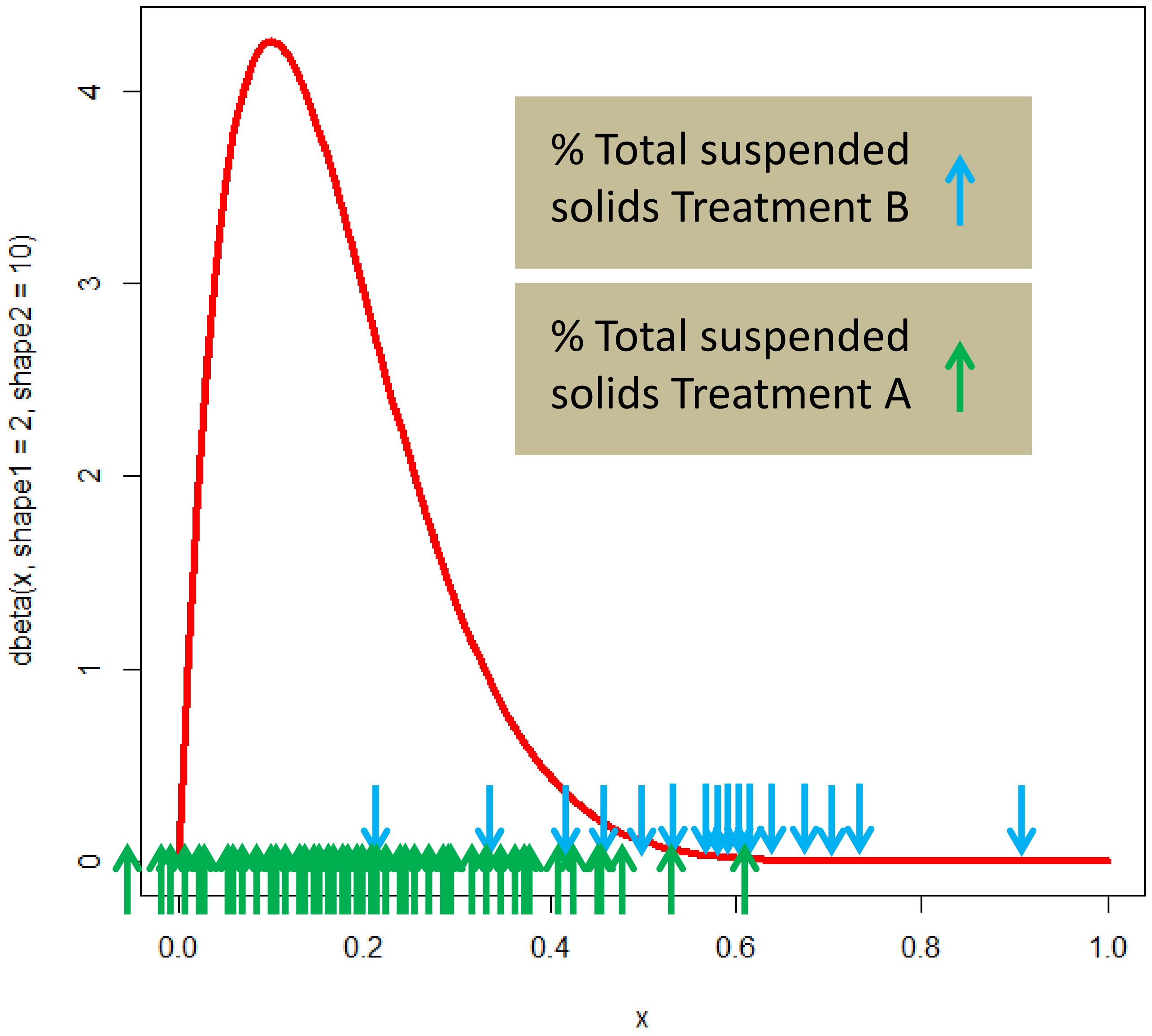
Mites/Plant	Before	After
Corn plot 1	0.500	22.967
Corn plot 2	10.657	29.364
Corn plot 3	43.469	15.972
Corn plot 4	7.045	7.683
Corn plot 5	9.626	10.089
Corn plot 6	18.534	14.059
Corn plot 7	34.237	23.093
Corn plot 8	38.291	28.351
Corn plot 9	11.959	4.898
Corn plot 10	1.582	13.964

Distributions matter!

Beta Distribution

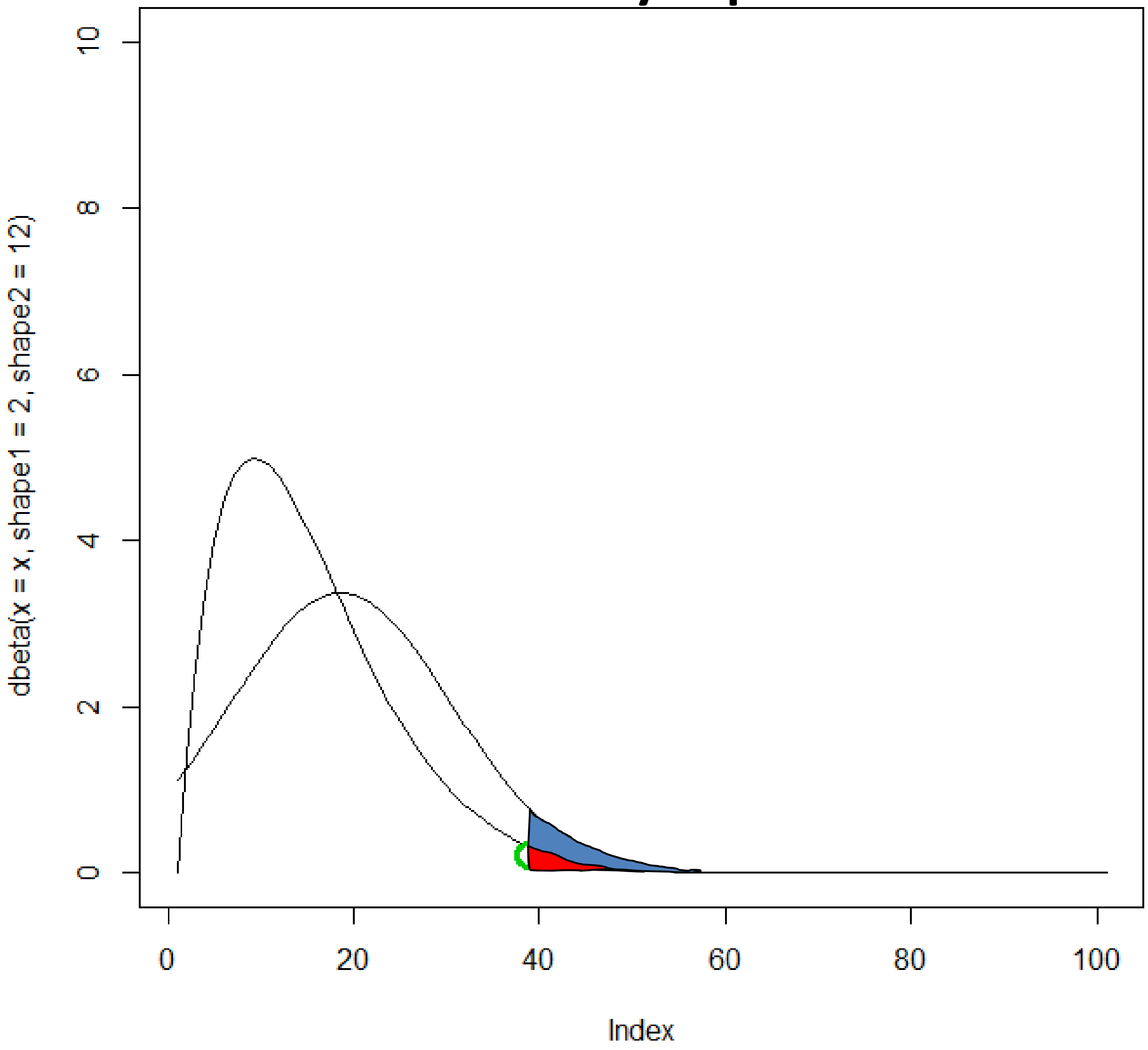
Shape 1 = 2

Shape 2 = 10



Probability space

$f(x)$

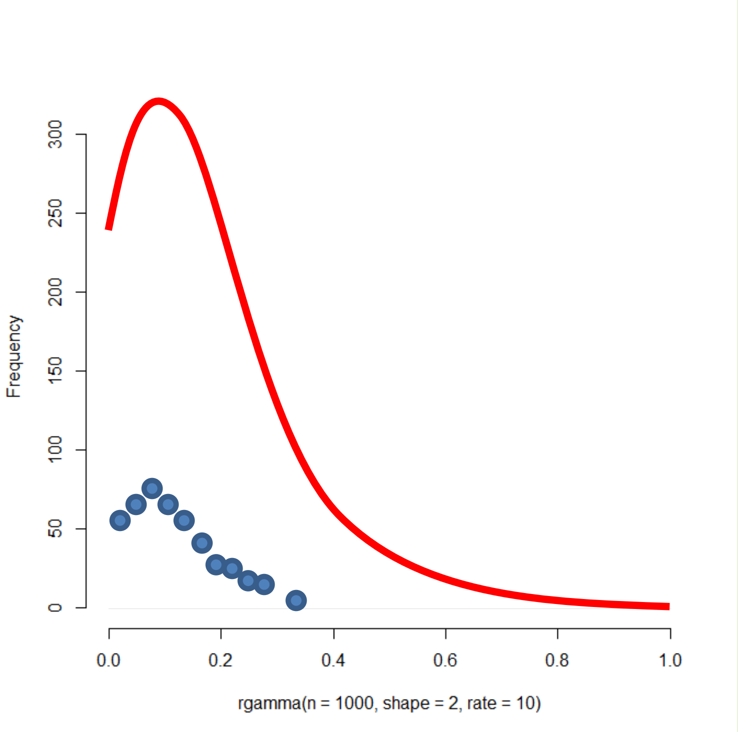


R Example: Oct 24_class notes Steel slag.R

Why are data distributions important?

Relative Elevation

- Most of the field is relatively flat but with a couple of terraces and a small hill
- That is, most plots will be at relatively low elevations with some exceptions

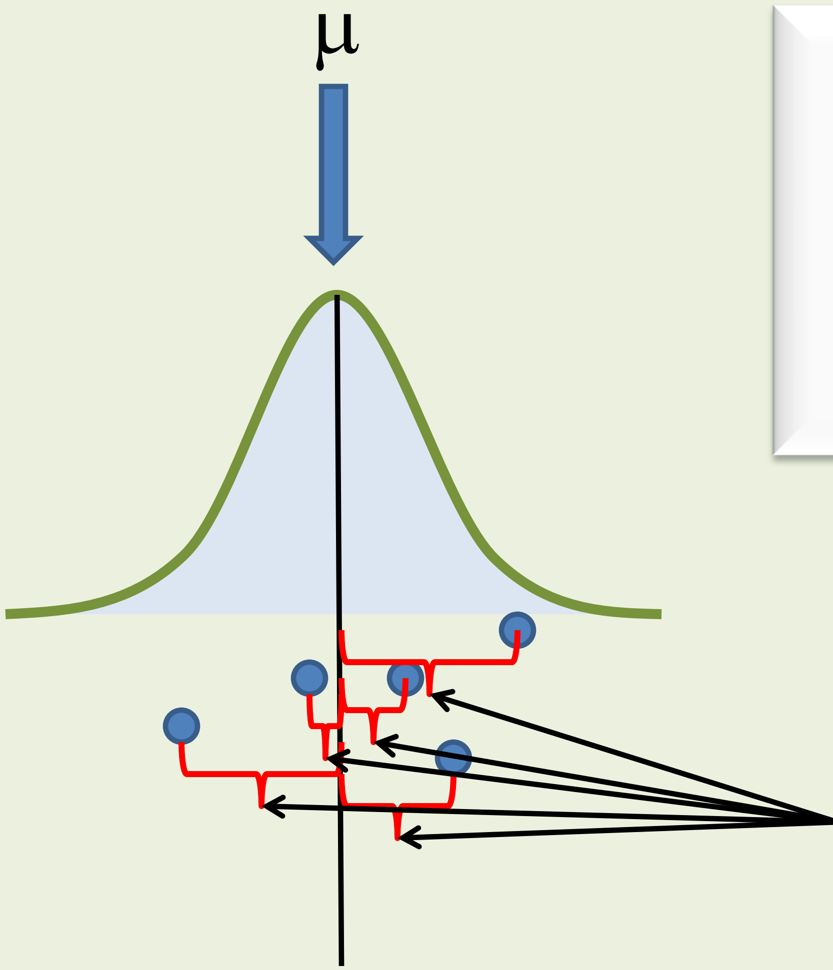


If we sampled randomly without stratifying, we could end up only sampling a very narrow range of relative elevation values.

It is hard to tell there if there is an effect of high relative elevation if no high relative elevation data were obtained.

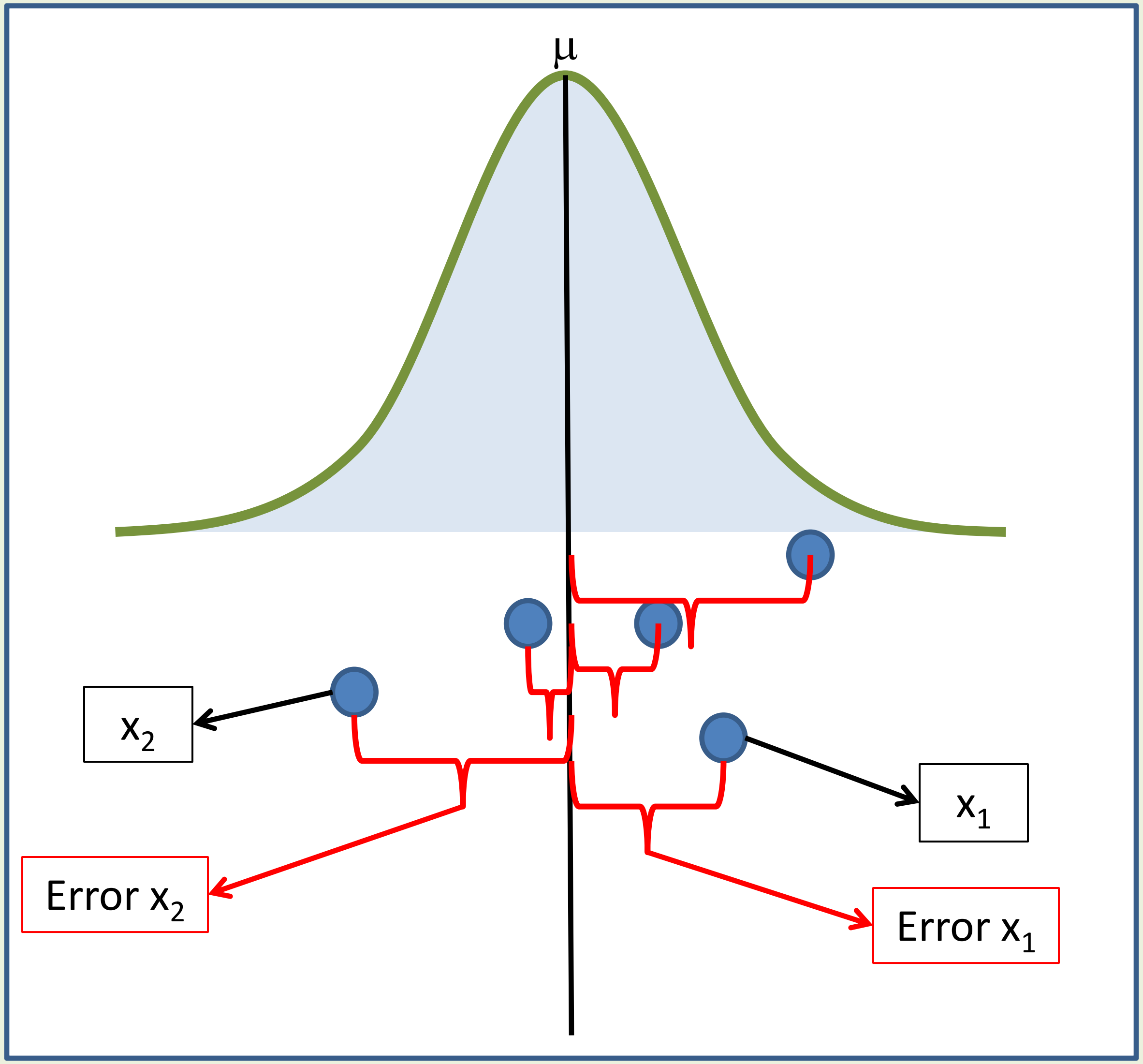
Time check!

Developing a test statistic with a normal distribution



Calculate probability for each data point, each error

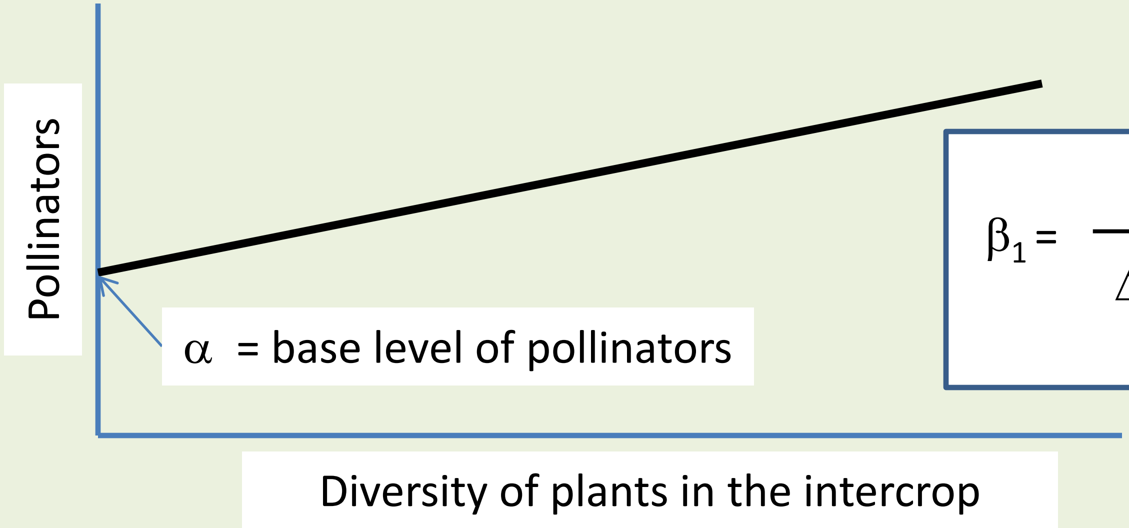
How far away are the data from their expected value(s)



$x_i - \mu$ = Distance or Error

Allows us to quantify the probability of x's occurrence

Linear model: $y = \alpha + \beta_1 * x$



$$\beta_1 = \frac{\Delta \text{Pollinators}}{\Delta \text{Intercrop.plant.diversity}}$$

Pollinators = $\alpha + \beta_1 *$ Intercrop.plant.diversity

Does $\beta_1 = 0$?

Example in R!

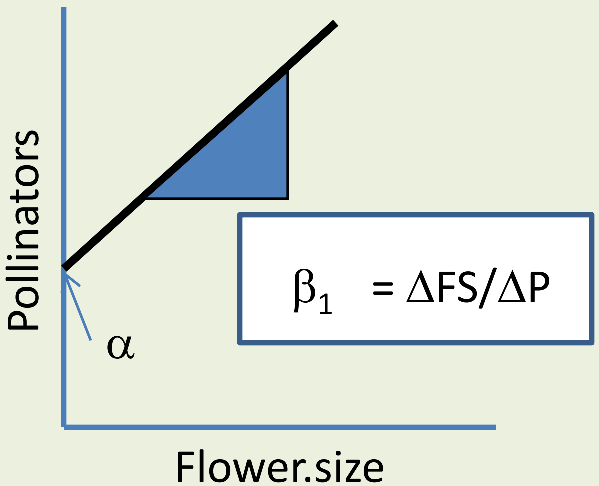
Linear regression: Assumptions about the data

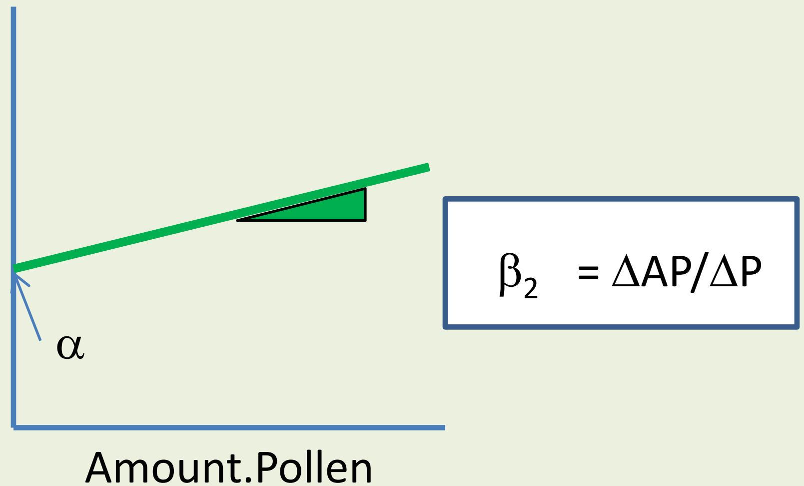
- There is no measurement error in your predictor variables (Ouch! – reinforces need for good design)
- Linearity (just witnessed in R)
- Constant variance in your errors (R example)
- Independence of errors in your response variable (y, e.g., # of pollinators)

Linear model multiple effects
multiple linear regression

$y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 \ldots$

Pollinators = $\alpha + \beta_1$* Flower.size + β_2* Amount.Pollen



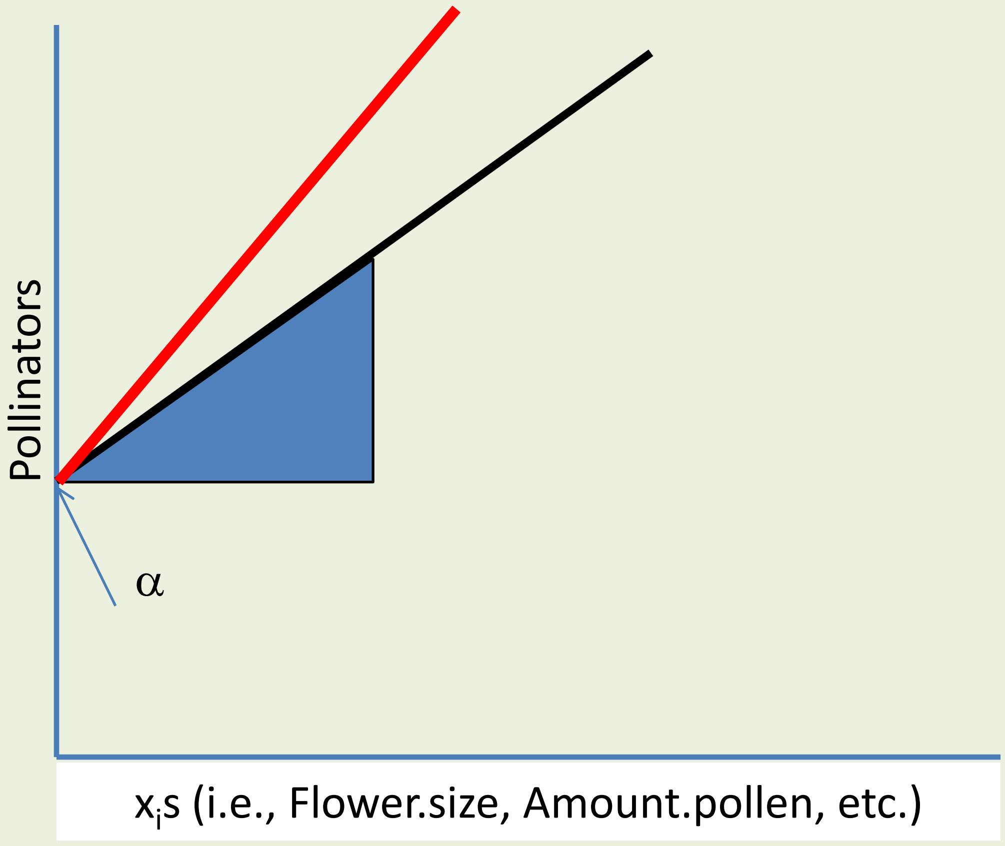


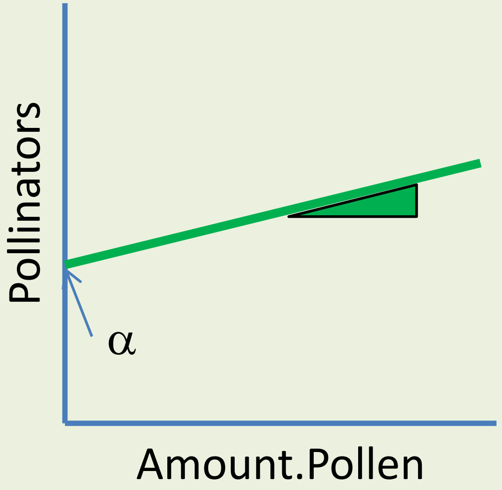
Does $\beta_1 = 0$?

Does $\beta_2 = 0$?

…and…

Pollinators = α + β_1* Flower.size **+** β_2* Amount.Pollen





What is the prob that the overall model slope = 0?
Could the slope of the red line be equal to zero?

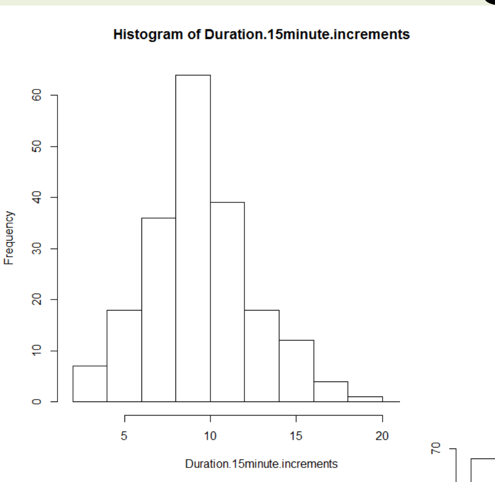
Assignment B

- Assignment B is due on November 1st
- Worth 50 points
- Simulation
  - Using the provided functions for distributions (from R Chapter 7) , take a first pass at simulating data for each of your components where you will be taking data. Assume that data will be measured perfectly (no measurement error).
  - Write up in manuscript form for a few of the components. That is, introduce the system (you can self-plagiarize but make it clean), describe how you will sample (or already sampled) components (Methods section), describe your simulation inputs, include output plots. Discuss in brief.

Steps

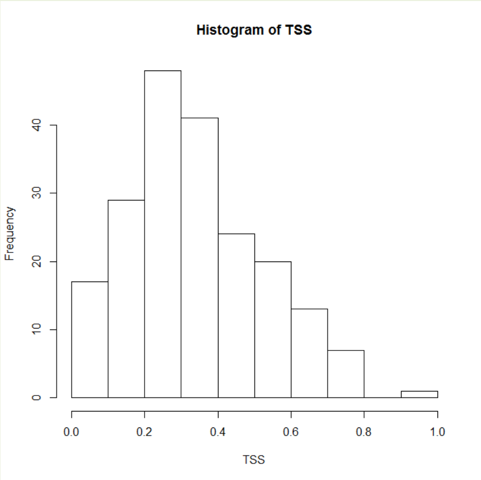
- Look at the data distributions that you have created for your concept map
- Look over the R Chapter 7 distributions
- Figure out one that looks like it fits
- Adjust the values so that distribution parameters fit your data

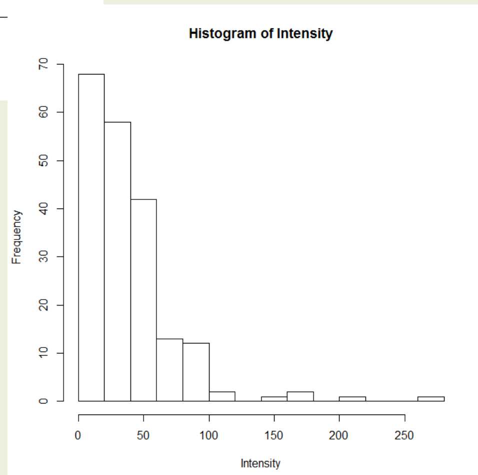
Testing a Bioretention systems: Total Suspended Solids



Poisson, most events around 2.5 hours

Duration

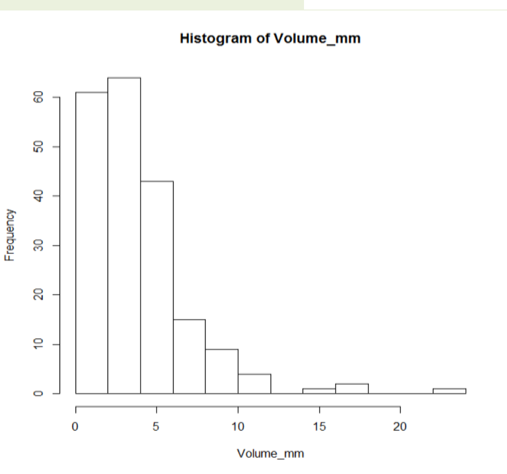




Intensity (precipitation events)

Percent of Total Suspended Solids (TSS) remaining in outflow

Intensity = Duration * Volume



Volume

log normal, 3mm mean, 2 mm standard deviation

Liner Strip

Cell vegetation

Assignment B

- Reintroducing the system
- Describing your actual sampling methodology (in brief)
- Describe with figures what you expect your data distributions to look like using histograms of your data
- Discuss in brief (or not)