Lecture 04

Bill Perry

# **Lecture 3: Review**

* We covered:
  + Understand why statistics is vital in biology
  + Distinguish between different types of biological variables
  + Learn about accuracy, precision, and bias in measurements
  + Calculate and interpret measures of central tendency (mean, median, geometric mean)
  + Calculate and interpret measures of spread (standard deviation, variance, IQR)
  + Understand data transformations for skewed distributions
  + Visualize descriptive statistics for our data
  + Learn how to handle uncertainty in our data

add to this for eye candy

# **Lecture 4:** Probability distributions

## Introduction to probability distributions

* What is a frequency distribution?
* What is a probability distribution?
* Distributions for variables and for statistics

## Estimation

* Populations and samples
* Parameters and statistics

we are going to use some sculpin data that is real!



[Wikipedia](https://en.wikipedia.org/wiki/Slimy_sculpin)

# Lecture 4: Frequency distributions

Example data we will use will be a combination of data from Toolik Alaska LTER

[source](https://portal.edirepository.org/nis/mapbrowse?scope=knb-lter-arc&identifier=10577)

We will specifically look at fishes like

### Lake Trout



[source](https://news.orvis.com/fly-fishing/fish-facts-lake-trout-salvelinus-namaycush)

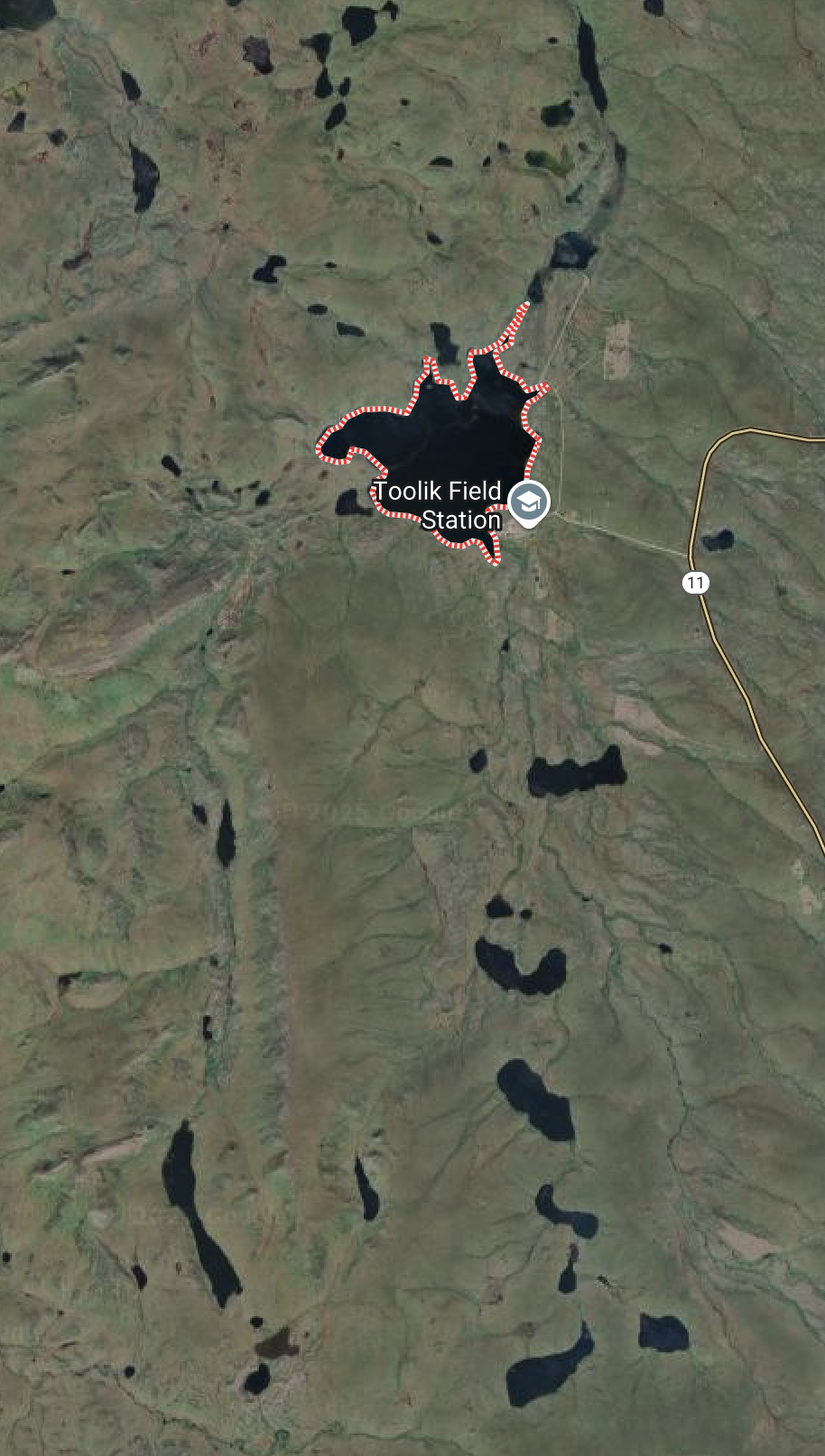
### Grayling



### Slimy Sculpin



[source](https://en.wikipedia.org/wiki/Slimy_sculpin)



# Lecture 4: Frequency distributions

* Data - has been cleaned in terms of lake names and species names
* Slimy Sculpin - Toolik Lake

sculpin\_df %>%   
 filter(lake == "Toolik") %>%   
 summarize(  
 mean = mean(total\_length\_mm, na.rm = TRUE),  
 sd = sd(total\_length\_mm, na.rm = TRUE),  
 se = sd(total\_length\_mm, na.rm = TRUE)/(sum(!is.na(total\_length\_mm))^0.5),  
 count = sum(!is.na(total\_length\_mm)),   
 .groups = "drop")

# A tibble: 1 × 4  
 mean sd se count  
 <dbl> <dbl> <dbl> <int>  
1 51.7 12.0 0.834 208

Note in the quarto code we use things to control what we see like  
#What does this mean

* # | echo: false
* # | message: false
* # | warning: false
* # | fig-height: 4 # | fig-width: 3
* # | paged-print: false)

|  |
| --- |
| Practice Exercise 1: Reading Slimy Sculpin - Toolik Lake |
| * Data - has been cleaned in terms of lake names and species names * Slimy Sculpin - Toolik Lake   # Write your code here to read in data # Remember to use tidy coding skills and comment the HOOI #  #  # library(tidyverse) # library(patchwork) # sculpin\_df <- read\_csv("data/sculpin.csv") # now look at what is there |

|  |
| --- |
| Practice Exercise 2: Now lets look at descriptive statistics |
| Let’s try looking at what the summary of the data tell us  # now do the summary statistics please |

# Lecture 4: Frequency Distributions

What is a frequency distribution?

* Display of number of observations in certain intervals
* e.g., the number of sculpin per interval in Toolik Lake
* as a table like below or histogram

sculpin\_df %>%   
 filter(lake == "Toolik") %>%  
 filter(!is.na(total\_length\_mm)) %>%   
 mutate(length\_bin = cut\_interval(total\_length\_mm, length = 2)) %>%  
 count(length\_bin)

# A tibble: 29 × 2  
 length\_bin n  
 <fct> <int>  
 1 [10,12] 1  
 2 (12,14] 3  
 3 (18,20] 1  
 4 (22,24] 1  
 5 (26,28] 1  
 6 (28,30] 1  
 7 (30,32] 2  
 8 (32,34] 3  
 9 (34,36] 4  
10 (36,38] 3  
# ℹ 19 more rows

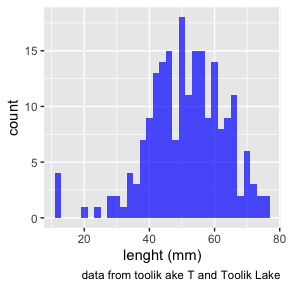
|  |
| --- |
| Practice Exercise 3: Now try to modify this so it is in 5 mm lenghts |
| Let’s try looking at what the summary of the data tell us  # now try different bins sculpin\_df %>%   filter(lake == "Toolik") %>%  filter(!is.na(total\_length\_mm)) %>%   mutate(length\_bin = cut\_interval(total\_length\_mm, length = 2)) %>%  count(length\_bin)  # A tibble: 29 × 2  length\_bin n  <fct> <int>  1 [10,12] 1  2 (12,14] 3  3 (18,20] 1  4 (22,24] 1  5 (26,28] 1  6 (28,30] 1  7 (30,32] 2  8 (32,34] 3  9 (34,36] 4 10 (36,38] 3 # ℹ 19 more rows |

# Lecture 4: Frequency Distributions

The alternative is to use a histogram

* the y axis is the count
* the x axis is the bin range
* each bin 0 - 5 and 5 - 10 and 10 - 15 or as you choose
* in ggplot the code looks like

dataframe %>% ggplot(aes(thing\_to\_count))+  
 geom\_histogram(   
 binwidth = increments\_to\_work\_with  
 )

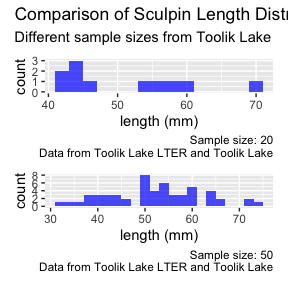


|  |
| --- |
| Practice Exercise 4: This is something you shoudl do |
| Let’s try stuffing frogs in our pockets  # Write your code here to create funny plot # Remember to use tidy coding skills and comment the HOOI |

# Lecture 4: Frequency Distributions

What happens as sample size changes…

* Sampls size
  + Low sample number - 15
  + High sample number - 70
* Frequency distribution takes on “bell-shape”…

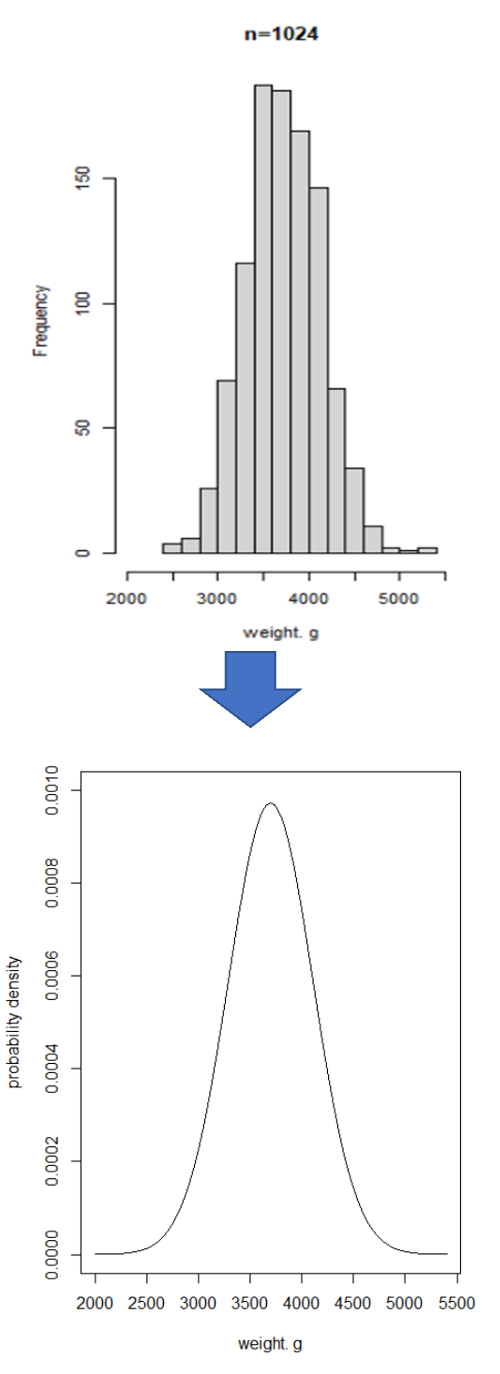


# Lecture 4: Probability distributions

Can we make assumption about distribution of random variable weight in population?

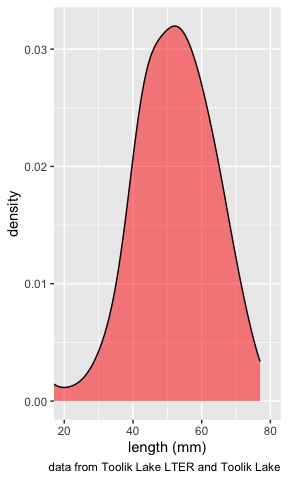
Probability distribution:

* theoretical frequency distribution in population



# Lecture 4: Probability distributions

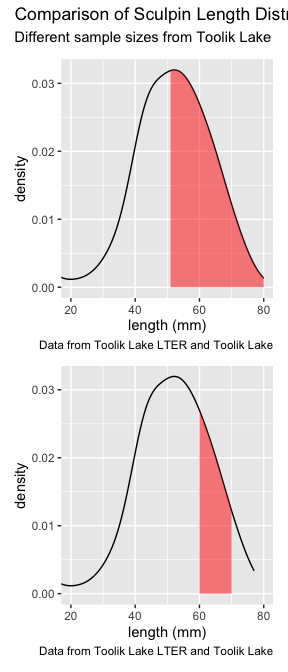
* For continuous random var: probability density function (PDF)
* PDF: mathematical expression of probabilities associated with getting certain values of random variable
* Area under curve = 1
* i.e., probability of length between 10 and 80 = 1



# Lecture 4: Probability distributions

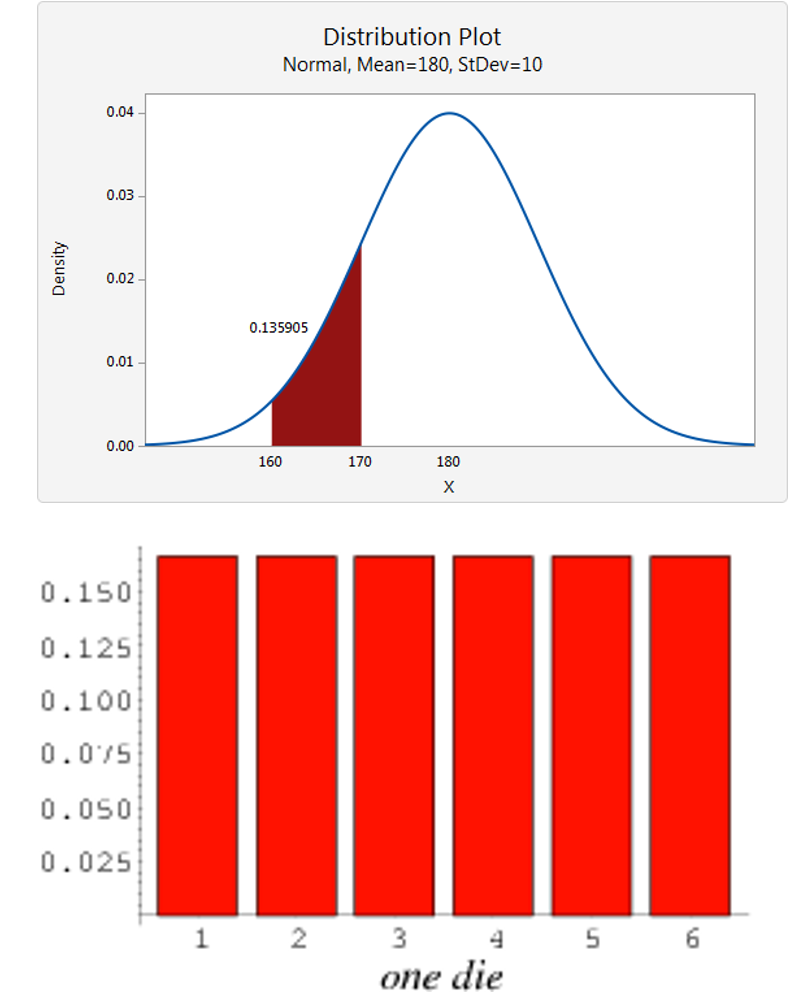
Now we could look at a lot of different ranges of lengths

* probability of the length larger than the mean
* probability of the length larger than 70 mm
* probability of the length between two numbers



# Lecture 4: Probability distributions

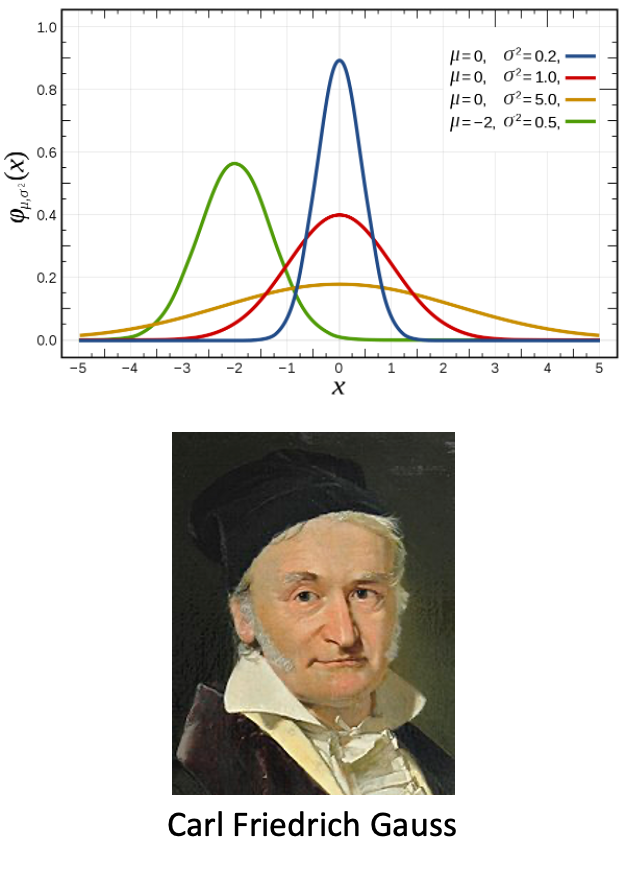
* Usually need to know probability distribution of random variables in statistical analyses
* Can define many distributions; some do reasonable job especially whit continuous variables
* Different distributions for continuous, discrete variables like a single die



# Lecture 4: Probability distributions

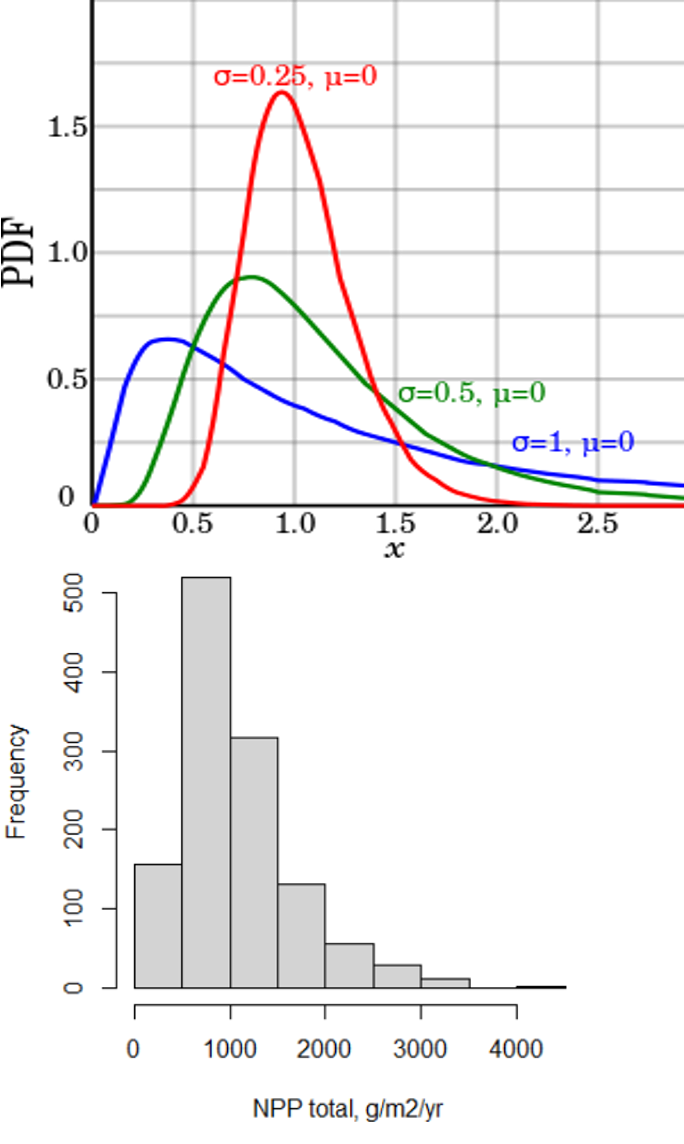
Normal (Gaussian): symmetrical, bell-shaped

* Defined in terms of mean and variance (μ, 𝜎2)
* SND (z-distribution) has mean μ=0 , 𝜎2 =1



# Lecture 4: Probability distributions

* Lognormal: right-skewed distribution
* Logarithm of random variable is normally distributed
* Common in biology.
* Why would this occur or be common in biology?

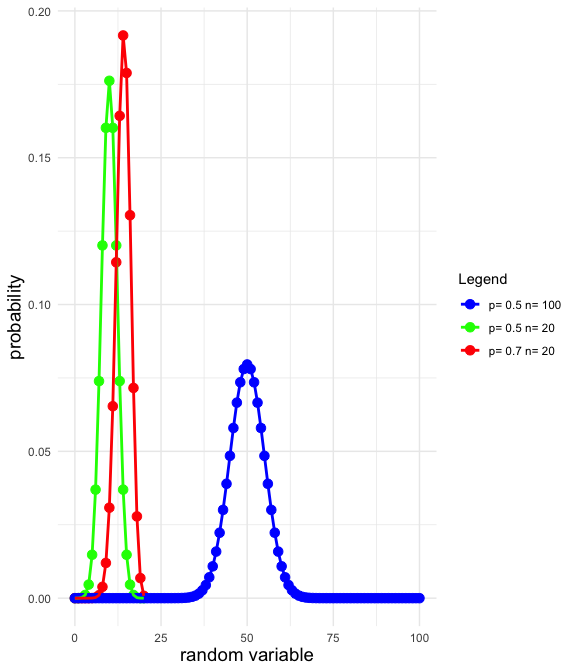


# Lecture 4: Probability distributions

## Binomial (multinomial):

* probability of event that have two outcomes (heads/ tails, dead/alive)
* Defined in terms of “successes” out of set number of trials

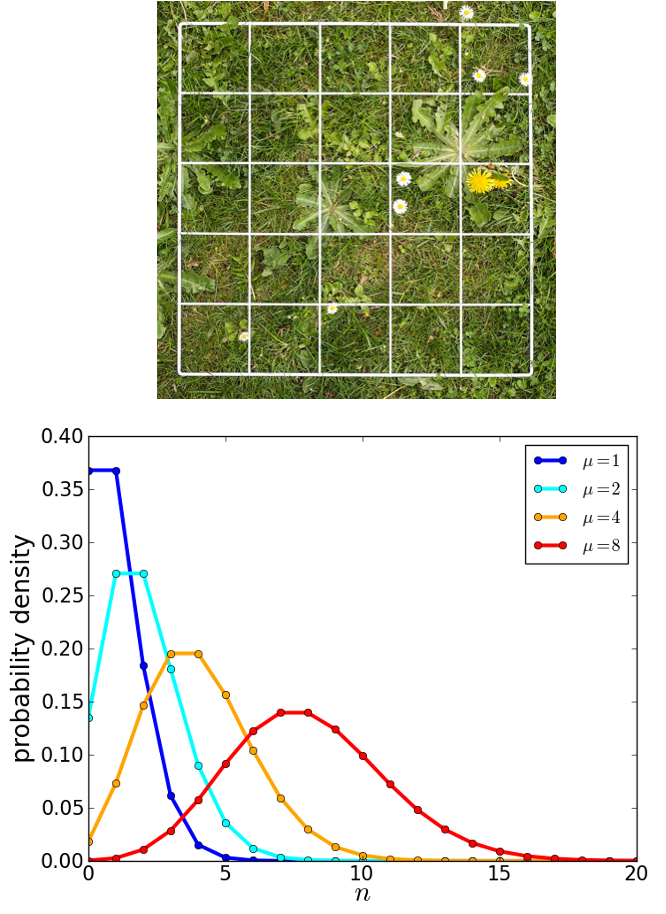
### In large number of trials: approximately normal distribution



# Lecture 4: Probability distributions

## Poisson: occurrences of (rare) event in time/space

* E.g., number of
  + ***Taraxacum officinale*** **- common dandelion** in quadrat
  + copepod eaten per minute
  + cells in field of view
* Measures Probability(y= certain integer value)
  + defined in terms of μ or mean
  + Right-skewed at small μ
  + more symmetrical at higher μ



# Lecture 4: Data gathering - managing

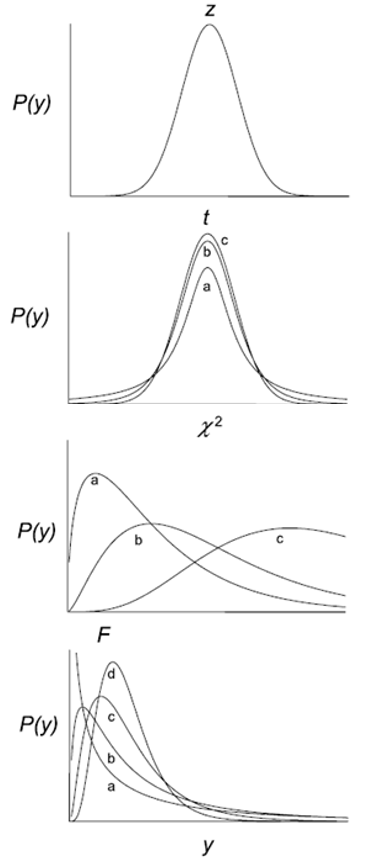
## Also have distributions of test statistics

### Test statistics:

* summary values calculated from data used to test hypotheses
* is your result due to chance?

### Different test statistics:

* different, well-defined distributions
* allows estimation of probabilities associated with results
* Examples:
  + z-distribution, student’s t-distribution, χ2-distribution, F-distribution



# Lecture 4: Samples and populations

### Inferential statistics:

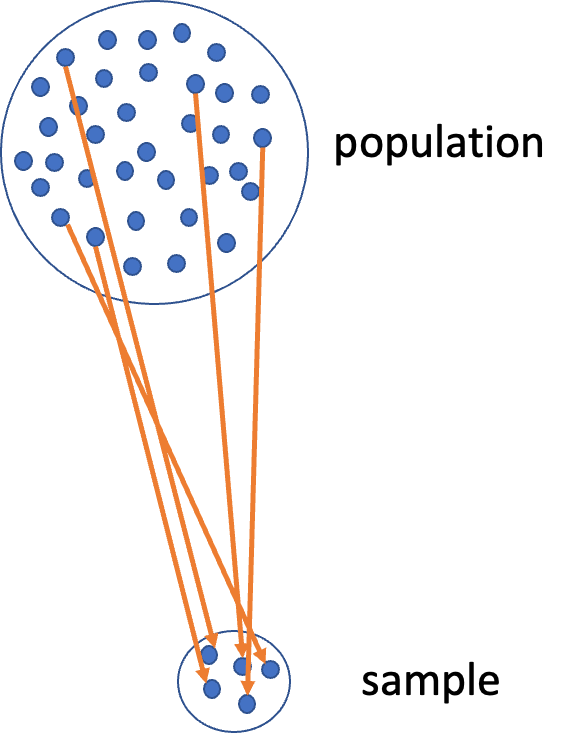
* inference from samples to populations

### Statistical population:

* All possible observations of interest
* Normally: populations too large to census

### Populations are defined in time + space

### Examples of statistical populations from you research area?



# Lecture 4: Samples and populations

### Key characteristic of sample is

* size (n observations; n = sample size)

### Characteristics of population - called parameters

* Parameters - Greek letters

### Characteristics of samples - statistical estimates of parameters

* statistics= Latin letters

**Random sampling crucial for**

**sample -> population**

**inference statistics -> parameters**



# Lecture 4: Parameters and statistics

Two main kinds of summary statistics: - center and spread

Center: - Mean (µ, ȳ): sum of sampled values divided by n - Mode: the most common number in dataset - Median: middle measurement of data; = mean for normal distributions

**Mean**

**Formula for n odd**

**Formula for n even**

# Lecture 4: Parameters and statistics

## Spread

* Range: from highest and lowest observation
* Variance (σ2, s2): sum of squared differences of observations from mean, divided by n-1

E.g., fish lengths = 20, 30, 35, 24, 36 g

# A tibble: 1 × 1  
 mean  
 <dbl>  
1 29

# Lecture 4: Parameters and statistics

Spread

(20 -29)^2+ (30 -29)^2 + (35 -29)^2 + (24 -29)^2 + (36 -29)^2 = 57,104

192 / (5-1) = 48 mm^2 *Problem: weird units!*

# A tibble: 1 × 2  
 mean variance  
 <dbl> <dbl>  
1 29 48

# Lecture 4: Parameters and statistics

## Spread

* Standard Deviation(σ, s): square root of variance.
  + In same units as observations
  + In example: √48 = 6.9 mm
* Coefficient of variation: SD as % of mean.
  + Useful for comparing spread in samples with different means
  + In example: (6.9/29)\*100= 23.8 %

# Lecture 4: Estimation

Problem: - don’t know the values of parameters

Goal: - estimate parameters from empirical data (samples)

3 general methods of parameter estimation: - Maximum Likelihood Estimation (MLE) - Ordinary Least Squares (OLS) - Resampling techniques

* MLE general method to estimate parameters in a way that maximizes the likelihood of the observed data given the parameter values.
* aims to find the parameter values that make the observed data most probable under the assumed statistical model.
* OLS specific method to estimate parameters of a linear regression model.
* minimizes the sum of the squared differences between observed and predicted values

# Lecture 4: Estimation

