STAT 110: Week 3

University of Otago

Outline

- Data summaries: sample mean and standard deviation
- Summaries are limited
  - To go further we needed statistical models
    - Use probability to describe the variation in the data
- Had an introduction to probability
- Today we will introduce idea of a random variable
  - Useful in helping us use probability to describe data

Example: bovine leptospirosis

- An inspector visits cattle & dairy farms for signs of bovine leptospirosis
- If they visit three farms, the sample space has eight possible outcomes
  - LLL, LLC, LCL, LCC, CLL, CLC, CCL, CCC
    - L: evidence of leptospira at farm
    - C: farm is clear
  - Each outcome has an associated probability
- If the inspector visits 30 farms, there are 1 073 741 824 possible outcomes
- The way the problem is expressed makes it difficult to answer questions:
  - How many farms would we expect to have evidence of leptospira?
  - How likely is it that 24 or more farms will have evidence of leptospira?
- We need a better way of writing/expressing things

Random variable

- A random variable assigns a numerical value to each outcome in sample space
- For our purposes, we can use a simpler definition:
  - A random variable is a (random) process with a numerical outcome
- Common to represent a random variable with capital letter
  - e.g. X or Y or Z
- The possible values are given with lowercase letters
  - e.g. x, y, z

Random variables: leptospirosis example

- Y represents the number of farms with evidence of leptospira
- Visit three farms
  - Four possible values: $y_1 = 0$, $y_2 = 1$, $y_3 = 2$, $y_4 = 3$
- Visit 30 farms
  - 31 possible values: $y_1 = 0$, $y_2 = 1$, ..., $y_{31} = 30$.
- We may use i (or j) as an index of possible values
  - e.g. $i = 2$ is the second possible value; $y_i = y_2 = 1$
- We use the k to represent the number of possible values
  - $k = 4$ if we visit three farms
  - $k = 31$ if we visit 30 farms

Probability distribution

- A random variable has an associated probability distribution
- For the leptospirosis example

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

- $\Pr(Y = y_i)$: the probability that (the random variable) Y takes the value y_i
  - e.g. for $i = 3$: $\Pr(Y = 2) = 0.4$, the probability that Y takes the value 2

Probability distribution: example

- Suppose we open an online store that sells two products
- A given online visitor may:
  - With probability 0.4 buy nothing: we receive \$0
  - With probability 0.3 buy item A: we receive \$20
  - With probability 0.2 buy item B: we receive \$35
  - With probability 0.1 buy item A and B: we receive \$50
- If Y represents the money we receive from an online visitor

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

Using probability distributions

- With these definitions we can start to ask useful questions
  - How likely is it that 2 or more farms will have evidence of leptospira?
  - How likely is it that we will receive $20 or below from an online visitor?

Using probability distributions

- With these definitions we can start to ask useful questions
  - How likely is it that 2 or more farms will have evidence of leptospira?
  - How likely is it that we will receive $20 or below from an online visitor?
- We use results from last week to answer those questions
- Using the online store as an example
  - Think of the y values as events: $y_1 = 0$, $y_2 = 20$, $y_3 = 35$, $y_4 = 50$
  - The events are mutually exclusive
  - $\Pr(Y \leq 20) = \Pr(Y = 0 \text{ or } Y = 20) = \Pr(Y = 0) + \Pr(Y = 20) = 0.4 + 0.3 = 0.7$

Expectation

- We can't yet answer the other question from earlier
  - How many farms would we expect to have evidence of leptospira? or
  - How much money do we expect to receive from an online visitor?
- We want to find $E[Y]$, the expected value of the random variable Y
  - The expected value is the same as the mean and is often represented by μ
- To find this, we weight each possible value by its corresponding probability

$$E[Y] = \sum_{i=1}^{k} y_i \Pr(Y = y_i)$$

- k is the number of possible values (in both our examples $k = 4$)
  - $E[Y] = y_1 \Pr(Y = y_1) + y_2 \Pr(Y = y_2) + y_3 \Pr(Y = y_3) + y_4 \Pr(Y = y_4)$

Expectation: leptospirosis example

- How many farms would we expect to have evidence of leptospira?

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

$$E[Y] = \underbrace{0 \times 0.25}_{0} + \underbrace{1 \times 0.15}_{0.15} + \underbrace{2 \times 0.4}_{0.8} + \underbrace{3 \times 0.2}_{0.6}$$

$$= 1.55$$

- We expect to find 1.55 farms with evidence of leptospira infection

Expectation: online store

- How much money do we expect to receive from an online visitor?

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

$$E[Y] = \underbrace{0 \times 0.4}_{0} + \underbrace{20 \times 0.3}_{6} + \underbrace{35 \times 0.2}_{7} + \underbrace{50 \times 0.1}_{5}$$

$$= 18$$

- We expect to receive \$18 from an online visitor

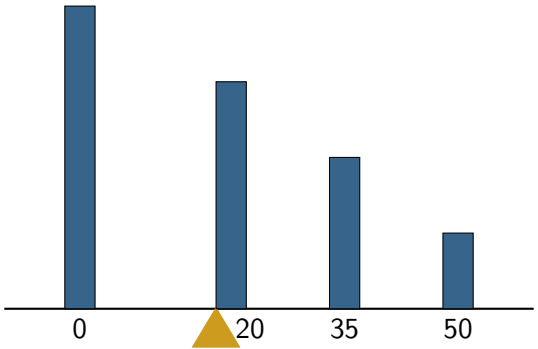
Expection: intuition

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

- If we saw 100 online visitors
  - We would expect 40 of them to spend nothing: receive \$0
  - We would expect 30 of them to spend \$20: receive \$600
  - We would expect 20 of them to spend \$35: receive \$700
  - We would expect 10 of them to spend \$50: receive \$500
- We would expect to receive \$1800 per 100 visitors = \$18 per visitor
- Multiplying y_i by $\Pr(Y = y_i)$ is taking a 'direct route' to this answer

Expectation: intuition

- Another way we can look at expectation is by thinking of the probability distribution as a old-fashioned scale
- The expected value balances the probability distribution (gold triangle)



Variance

- We could also ask questions that relate to variability
  - How much would we expect income from our store to vary from one day to the next?
- For small problems (like those we have been looking at)
  - Probably preferable to base this off the probability distribution
- For larger problems (which we are moving toward)
  - We need a measure of variability
  - Typically use variance / standard deviation

Variance

- The variance of the random variable Y is $\text{Var}(Y)$
  - Find the average of squared deviations from the mean
  - Weight the squared deviations by their probability

$$\text{Var}(Y) = \sum_{i=1}^{k} (y_i - E[Y])^2 \Pr(Y = y_i)$$

- For $k = 4$
  - $\text{Var}(Y) = (y_1 - E[Y])^2 \Pr(Y = y_1) + (y_2 - E[Y])^2 \Pr(Y = y_2) +$
    $(y_3 - E[Y])^2 \Pr(Y = y_3) + (y_4 - E[Y])^2 \Pr(Y = y_4)$

Variance: leptospirosis example

- What is the variance in the number of farms that have evidence of leptospira?
  - We know $E[Y] = 1.55$

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

$$\begin{aligned}\text{Var}(Y) &= \underbrace{(0-1.55)^2 \times 0.25}_{2.4025\times 0.25} + \underbrace{(1-1.55)^2 \times 0.15}_{0.3025\times 0.15} + \underbrace{(2-1.55)^2 \times 0.4}_{0.2025\times 0.4} + \underbrace{(3-1.55)^2 \times 0.2}_{2.1025\times 0.2} \\ &= 1.1475\end{aligned}$$

Standard deviation

- The standard deviation is the square root of variance
  - $\text{sd}(Y) = \sqrt{\text{Var}(Y)}$
- For the leptospirosis example
  - $\text{sd}(Y) = \sqrt{1.1475} = 1.07$
- The standard deviation is (approximately) the average deviation from the mean
- Often the variance will be represented by σ^2
  - The standard deviation as σ

Example: online visitors

- What is the variance in the amount we receive from an online visitor?
  - We know $E[Y] = 18$

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

$$\begin{aligned}
\mathsf{Var}(Y) &= \underbrace{(0-18)^2 \times 0.4}_{324\times 0.4} + \underbrace{(20-18)^2 \times 0.3}_{4\times 0.3} + \underbrace{(35-18)^2 \times 0.2}_{289\times 0.2} + \underbrace{(50-18)^2 \times 0.1}_{1024\times 0.1} \\
&= 291 \\
\mathsf{sd}(Y) &= 17.1
\end{aligned}$$

We've seen this before

- We saw expectation (mean), standard deviation, and variance in Week 1
  - Sample mean, sample variance, sample standard deviation
  - These are summaries of a particular data set (a sample)
- Today we've found these quantities for a distribution
  - Summaries of a random variable
  - Tells us something about what realizations from the distribution should look like

Summary

- Introduced random variables
- Probability distribution of random variable
- Saw several summaries of random variables
  - Mean
  - Variance
  - Standard deviation

Outline

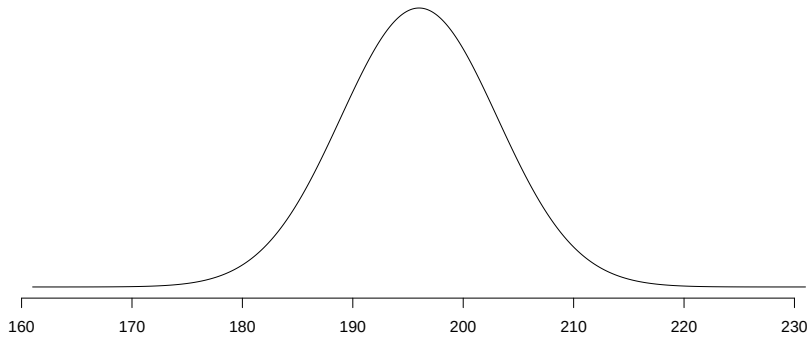
- We saw random variables in the last lecture
  - Probability distribution
  - Expectation
  - Variance
- Continue learning about random variables today
  - Can we have continuous random variables?
  - What happens when we combine random variables?

Discrete vs continuous

- The random variables we looked at in the last lecture were all discrete
  - Countable number of distinct values
- Discrete random variables are useful in a range of problems, e.g.
  - Number of eggs in a nest
  - Number of tasks completed in fixed time
  - Number of bugs in a piece of computer code
  - Number of voters who prefer candidate X
- There are other situations where things aren't discrete, e.g.
  - The flipper length of a gentoo penguin
  - The time taken in reflex test
  - The pH of seawater
- These can take continuous values

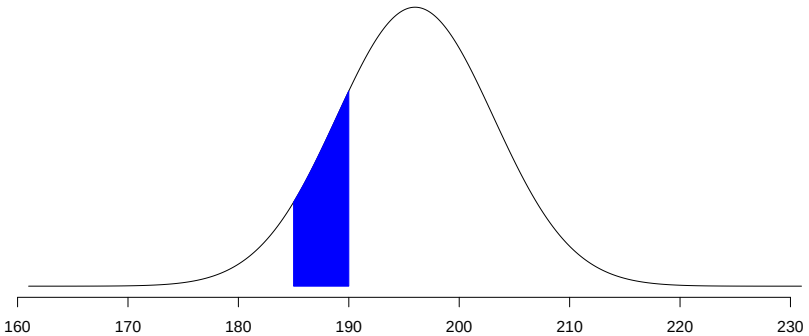
Continuous random variables

- An infinite (and uncountable) number of possible values
- Each value has a probability density
  - Best seen graphically (e.g. for flipper length)



Probability density

- This curve is called a probability density function (pdf)
- Probability is given by the area under the curve (pdf)
  - The total area under the curve (pdf) is 1
- The probability of flipper length between 185 and 190 mm is given by:



Continuous vs discrete

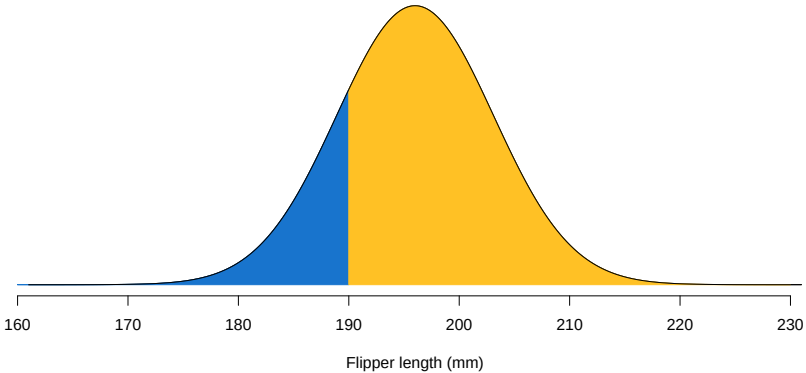
- Much of what we have already learned applies to continuous random variables
  - We can find expectation, variance, standard deviation
  - The calculations are more complex (sums are replaced by integrals)
  - Explore in more detail in more advanced courses (e.g. STAT 270)

Continuous

- Much of what we have already learned applies to continuous random variables
  - We can find expectation, variance, standard deviation
  - The calculations are more complex (sums are replaced by integrals)
  - Explore in more detail in more advanced courses (e.g. STAT 270)
- Look at examples on the next two slides

Complement

- Suppose we know the probability that flipper length is less than 190 mm (blue)
  - $\Pr(\text{flipper length} < 190) = 0.2$
- What is $\Pr(\text{flipper length} > 190)$? (gold)
  - It is a complement!



Combinations of random variables

- We may be interested in the combination of several random variables
  - Adélie penguins: feeding trip time
    - Random variables: time spent (i) feeding, (ii) resting, (iii) transit, in a trip
    - Combination: total trip time
  - Genetic linkage (crossover[1])
    - Random variables: number of crossovers in each chromosome
    - Combination: total number of crossovers
  - Cricket: runs scored
    - Random variables: number of singles, twos, threes, fours, sixes in an innings.
    - Combination: total score
  - Finance: portfolio value
    - Random variables: share prices for spark (SPK) and port of Tauranga (POT)
    - Combination: portfolio value (e.g. portfolio: 5 SPK, 10 POT)

[1] segments of DNA from one parent's chromosome swap with corresponding segments on the other parent's chromosome during meiosis

Combination of random variables

- Suppose we have random variables X and Y
  - To guide the development, we will think about
    - X: value of one SPK share in one months time
    - Y: value of one POT share in one months time
- We may be interested in a linear combination of X and Y
  - $aX + bY$
- What is the expected value of $aX + bY$?
- What is the variance of $aX + bY$

Expected value of combination

- If we owned shares: 5 SPK and 10 POT
  - Linear combination represents the value of our portfolio in one months time
  - $5X + 10Y$
    - Here, a is the number of SPK shares: 5
    - Here, b is the number of POT shares: 10
- How do we find the expected value of the linear combination?

$$E[aX + bY] = aE[X] + bE[Y]$$

- If $E[X] = 3$ and $E[Y] = 6.3$ then, the expected portfolio value is

$$\begin{aligned} E[5X + 10Y] &= 5E[X] + 10E[Y] \\ &= 5 \times 3 + 10 \times 6.3 \\ &= 78 \end{aligned}$$

Expected value of combination

- Ice cream is sold from 16 L containers in NZ
  - Expect that there is 16 L when opened
  - Can vary: let's say a standard deviation of 0.1 L (variance 0.01)
  - Let X be the amount of ice cream in a container: $E[X] = 16$, $Var(X) = 0.01$
- A new container of goldrush icecream is opened for the person ahead of us in line.
- They get a scoop of gold rush
  - Expect each scoop to get 0.1 L of ice cream
  - Standard deviation of 0.01 L (variance 0.0001).
  - Let Y be the amount in a scoop of ice cream: $E[Y] = 0.1$, $Var(Y) = 0.0001$
- The amount of goldrush icecream when we come to order is $X - Y$
  - What is $E[X - Y]$?

Variance of combination

- Can also be important to have a measure of variability for the combination of random variables
  - Trip time for Adélie penguins
  - Number of crossovers
  - Runs in cricket innings
  - Value of portfolio
- If X and Y are independent, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

- If X and Y are not independent
  - The variance is more complicated (additional term needed)
  - Considered in higher level courses

Variance of combination

- What is $Var(X - Y)$ for ice cream example?
  - $a = 1$
  - $b = -1$

$$\begin{aligned} Var(X - Y) &= 1^2 Var(X) + (-1)^2 Var(Y) \\ &= Var(X) + Var(Y) \\ &= 0.01 + 0.0001 \\ &= 0.0101 \end{aligned}$$

- Portfolio: what is $Var(5X + 10Y)$?
  - Assume that share prices are independent (unlikely to be the case in reality)

Variance of combination

- We saw that $Var(X - Y) = Var(X) + Var(Y)$
  - We are subtracting Y from X. Why do the variances add?
- A server with low variability
  - Each scoop has is consistent in terms of the amount of ice cream
- A server with high variability
  - Each scoop can vary greatly (small or large or anywhere in between)
- If server is highly variable, will amount left in container be highly variable?

Variance of combination

- We saw that $Var(X - Y) = Var(X) + Var(Y)$
  - We are subtracting Y from X. Why do the variances add?
- A server with low variability
  - Each scoop has is consistent in terms of the amount of ice cream
- A server with high variability
  - Each scoop can vary greatly (small or large or anywhere in between)
- If server is highly variable, will amount left in container be highly variable?
- The variability in the amount of ice cream is the same if:
  - Add a scoop of ice cream to the container, or
  - Took a scoop of ice cream away

Abstract example

- Look at another example: somewhat abstract
  - Provide some useful results that we will use in coming weeks
- Let Y_1 and Y_2 be independent observations from a distribution
  - Mean μ
  - Standard deviation σ
- What is the mean and variance of $\frac{Y_1+Y_2}{2}$?
  - Sample mean of two values from a distribution

Abstract example: expected value

- The expected value of the sample mean is

$$\begin{aligned} E\left[\frac{Y_1+Y_2}{2}\right] &= \frac{1}{2}E[Y_1] + \frac{1}{2}E[Y_2] \\ &= \frac{1}{2}\mu + \frac{1}{2}\mu \\ &= \mu \end{aligned}$$

- The variance of the sample mean is

$$\begin{aligned} Var\left(\frac{Y_1+Y_2}{2}\right) &= \frac{1}{4}Var(Y_1) + \frac{1}{4}Var(Y_2) \\ &= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \\ &= \frac{\sigma^2}{2} \end{aligned}$$

Abstract example: extension

- This can be extended to when we have n independent observations: $Y_1, Y_2, \ldots, Y_n$
- The expected value of the sample mean is

$$E\left[\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right] = \frac{1}{n}E[Y_1] + \frac{1}{n}E[Y_2] + \ldots + \frac{1}{n}E[Y_n]$$
$$= \mu$$

- The variance of the sample mean is

$$Var\left(\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right) = \frac{1}{n^2}Var(Y_1) + \frac{1}{n^2}Var(Y_2) + \ldots + \frac{1}{n^2}Var(Y_n)$$
$$= \frac{\sigma^2}{n}$$

Summary

- Looked at continuous random variables
  - There are differences, but much remains the same
- Looked at combination of random variables
  - Expectation
  - Variance
- Next lecture: start developing models for data

Outline

- Introduction to statistical modeling
  - Populations and parameters
  - Samples and statistics
  - Estimation of parameters
  - Introduce the normal distribution

Big picture

- We may be interested in flipper lengths of gentoo penguins in the Palmer archipelago
  - e.g. what is the mean flipper length?
- How could we find the mean flipper length of gentoos on Palmer?

Big picture

- We may be interested in flipper lengths of gentoo penguins in the Palmer archipelago
  - e.g. what is the mean flipper length?
- How could we find the mean flipper length of gentoos on Palmer?
- Problem: question refers to population (of gentoos in the archipelago)
  - We cannot answer it unless we measure every individual in the population
  - Likely impossible
- Formulate a statistical model
  - Use a sample to tell us about the population

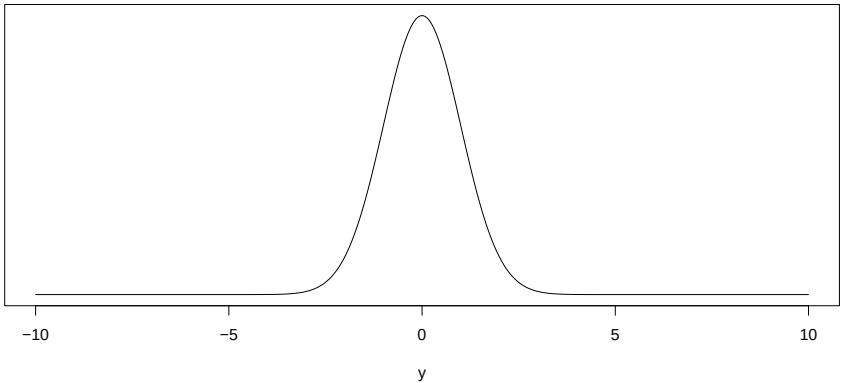
Statistical model

- Idea: assume the population values follow some distribution
  - The distribution tell us how the values vary in the population
- The distribution has unknown parameters
  - Parameter: any quantity that describes a population
- It is the parameter(s) that are of interest
  - Tell us about the population
- Abstract concepts
  - Introduce an example to make the idea more concrete
- We have seen probability distributions in simple 'generic' cases
  - Introduce specific case: normal distribution

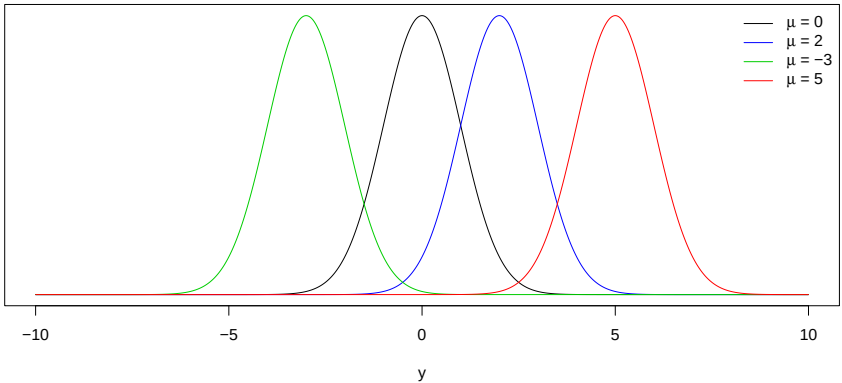
Normal model

- Assume that the data are normally distributed
  - We might also say we are using a normal model
- The normal distribution is sometimes called:
  - Bell-shaped curve
  - Gaussian model
- Described by two parameters:
  - Mean μ (Greek letter mu)
  - Standard deviation σ (Greek letter sigma)
    - Often refer to the variance σ^2 instead of the standard deviation
- We will spend some time familiarizing ourselves with the normal distribution

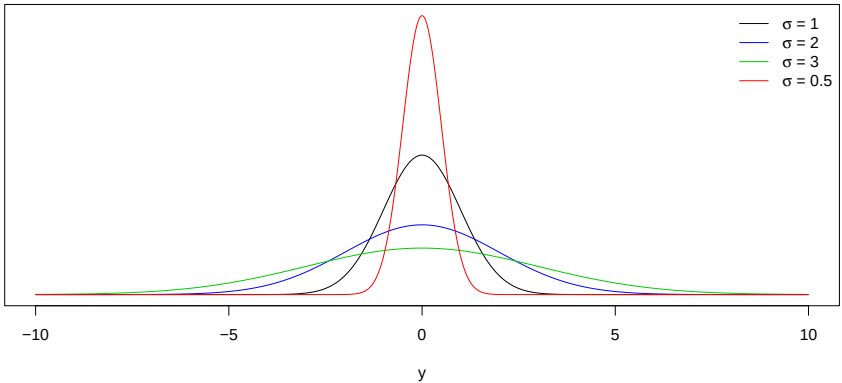
Probability density function (pdf) of normal distribution: $\mu = 0$, $\sigma = 1$



Pdf of normal distribution: different μ



Pdf of normal distribution: different σ



Model for gentoo

- We assume flipper lengths follow a normal distribution
  - This is an assumption about the population of gentoo penguins in Palmer archipelago
  - Parameters μ and σ are unknown
    - μ: mean flipper length (population level)
    - σ: standard deviation of flipper lengths (population level)
- Typically use greek letters for parameters
  - Here we are using μ and σ

Populations and samples

- Big idea: use a sample (and statistics) to estimate parameters
  - The estimate is an educated guess at the parameter value
- We have flipper length measurements from 68 gentoo penguins (cf. week 1)
- How could we use this sample to estimate μ?

Populations and samples

- Big idea: use a sample (and statistics) to estimate parameters
  - The estimate is an educated guess at the parameter value
- We have flipper length measurements from 68 gentoo penguins (cf. week 1)
- How could we use this sample to estimate μ?
- The sample mean $\bar{y}$ could be used to estimate the population mean μ
  - The sample mean $\bar{y}$ is an example of a statistic
  - Statistic: any quantity computed from values in a sample

That's easy ... are we done?

- Our example: finding a 'suitable' statistic is straightforward
  - Not always the case
- Let's imagine a more extensive penguin study:
  - Interested in understanding how feeding patterns, spatial structure (within a colony and between colonies), time of year, (and other factors) might influence penguin condition
    - What statistic(s) should we use for that?
- Later in semester we will (hopefully) think more about general strategies for finding suitable statistics (estimators)

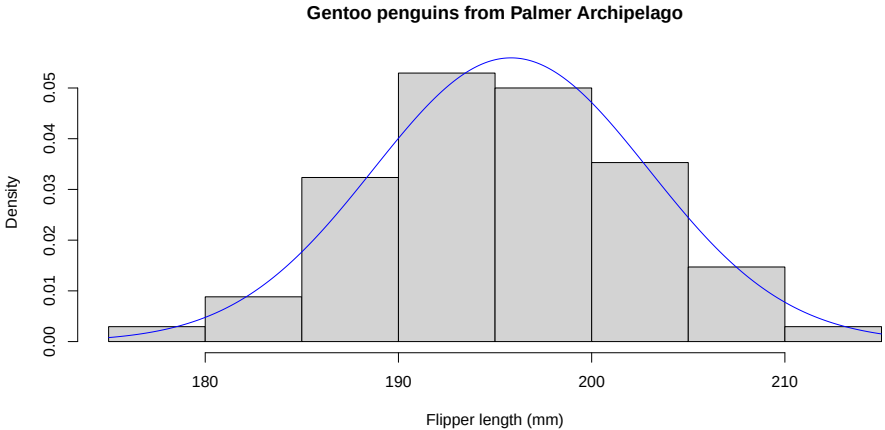
Model fitting

- For our model, we have
  - $\hat{\mu} = \bar{y}$
  - $\hat{\sigma} = s$
  - The population std deviation (σ) is being estimated by the sample std deviation (s)
- We have used the hat symbol $\hat{}$ to represent that we are estimating a parameter
  - $\hat{\mu}$ is said "mu-hat"
  - $\hat{\mu} = \bar{y}$: the parameter μ is being estimated by $\bar{y}$ (a statistic)

Fitted model

- Look at the fitted model (graphically)
  - (Normal) model at the estimated parameter values
- Compare the fitted model to the data
  - Load the gentoo penguin data into R (for the next few slides)
  - Revise material from week 1

Fitted model

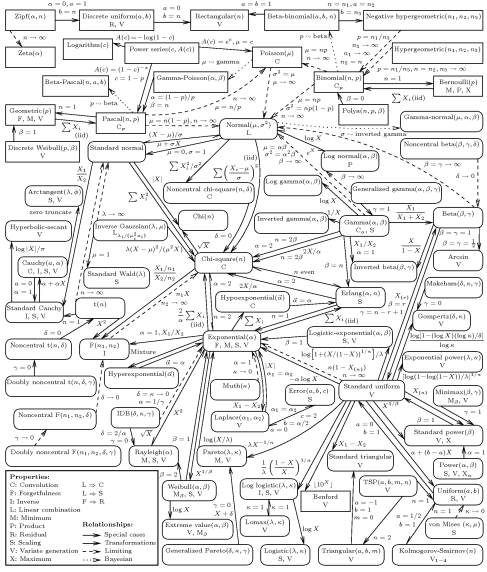


Statistical models

- Common mistakes:
  - Believing that μ is 195.8 mm (the sample mean)
    - This is the estimate of μ
    - It is impossible to know the value of μ
  - Believing that the model is true
    - Hope the model chosen is a reasonable approximation to reality
    - We should check this
- Checking the model fit
  - Looking to see if the model and the data are 'out of sync'
    - Plot: normal appears to describe the data reasonably well
  - Think a lot more about model fit in a couple of weeks (regression)

Statistical models

- In this example we have used a normal model for the data
- Should be reasonable for what we believe about flipper length:
  - Continuous values (flipper lengths can take any (positive) value)
  - Reasonably close to symmetric
    - An adult gentoo is unlikely to have a flipper four times the size of another adult gentoo
    - Cf. income
- Not all data looks like this!
  - Different types of data (yes/no, count, categories, time, space, ...)
  - Different characteristics (e.g. income)
  - Different complexity
- Many probability distributions with different characteristics
  - Next figure is for illustration, we don't need to learn it!



From: Leemis and McQuestion (2008). Univariate Distribution Relationships. *The American Statistician*, **62**, 45 – 53.

Looking forward I

- We will be working with a normal model for a few weeks
  - Look more at the normal distribution
    - Use it to describe (and model) data
    - Want to understand it as much as possible
- Explore a strategy for estimating parameters
  - Barely scratch the surface
  - Cover in more depth in higher level courses STAT 270, 370, 371
- Explore 'extensions' to normal model (e.g. regression)
- Explore models for other types of data: yes/no data

Looking forward II

- What does our estimate tell us about the parameter?
  - We have an estimate of μ from the sample of size 68
  - How 'close' to the true value of μ is it likely to be?
- Is the estimate likely to be better / worse if it were from:
  - A sample of size 6?
  - A sample of size 600?
- Explore how to determine how precise/uncertain the estimate is
- Also important is how were the data were collected?
  - e.g. does our sample consist of only adults?
  - We'll come back to this later in the semester
- Use the model for prediction (regression)

Summary

- Introduction to statistical modeling
  - Fit a normal model
  - Estimated the parameter μ with the statistic $\bar{y}$
- Next: get a better understanding of the normal distribution