

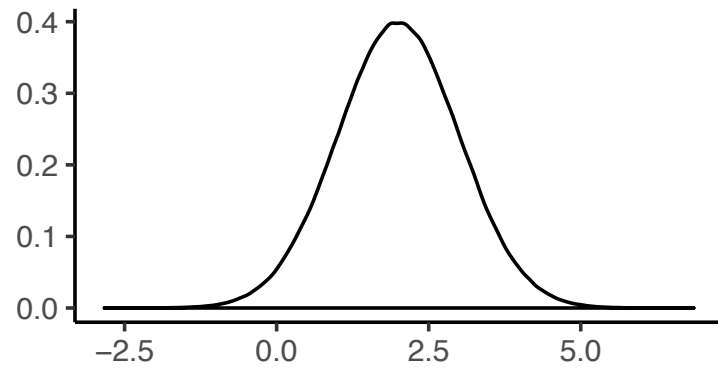
Probability

BIO5312 FALL2017

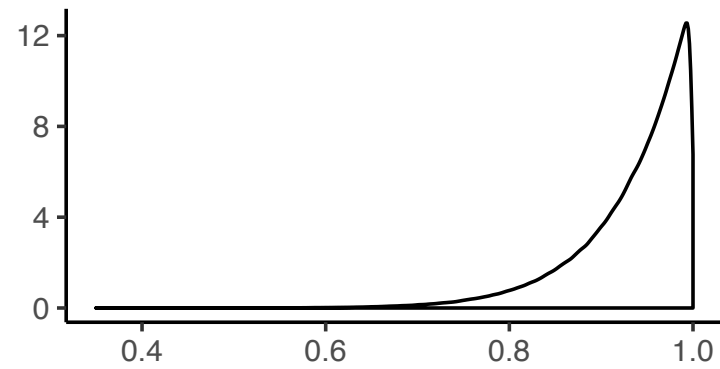
STEPHANIE J. SPIELMAN, PHD

Skew

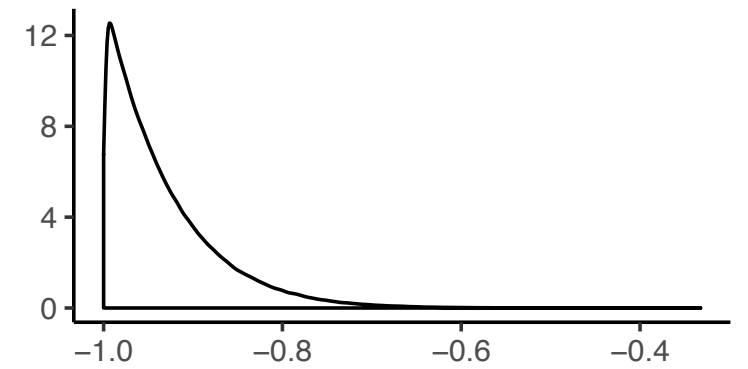
Symmetric



Left-skew

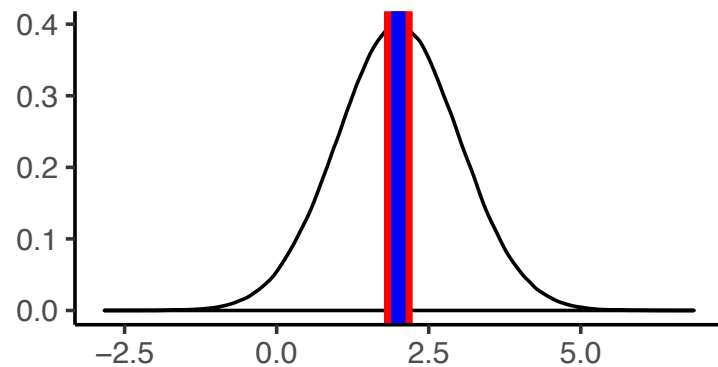
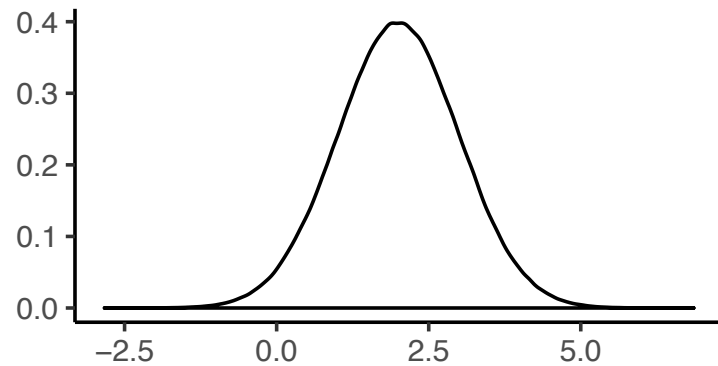


Right-skew

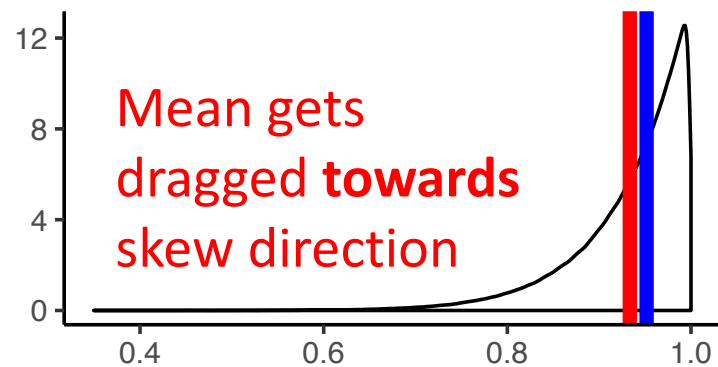
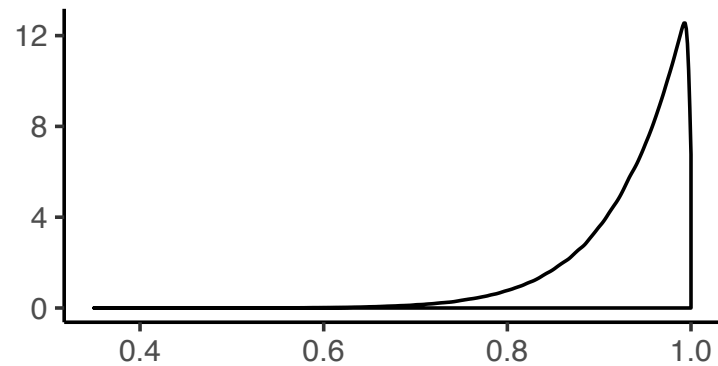


Mean vs median

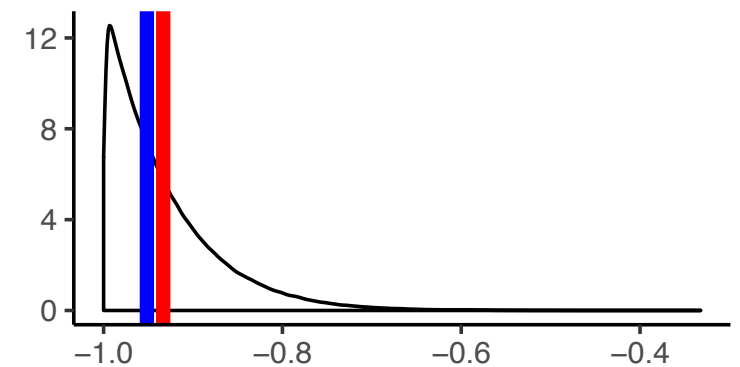
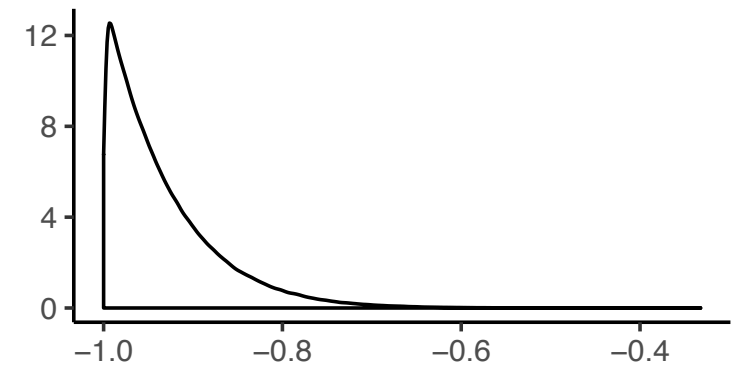
Symmetric



Left-skew



Right-skew

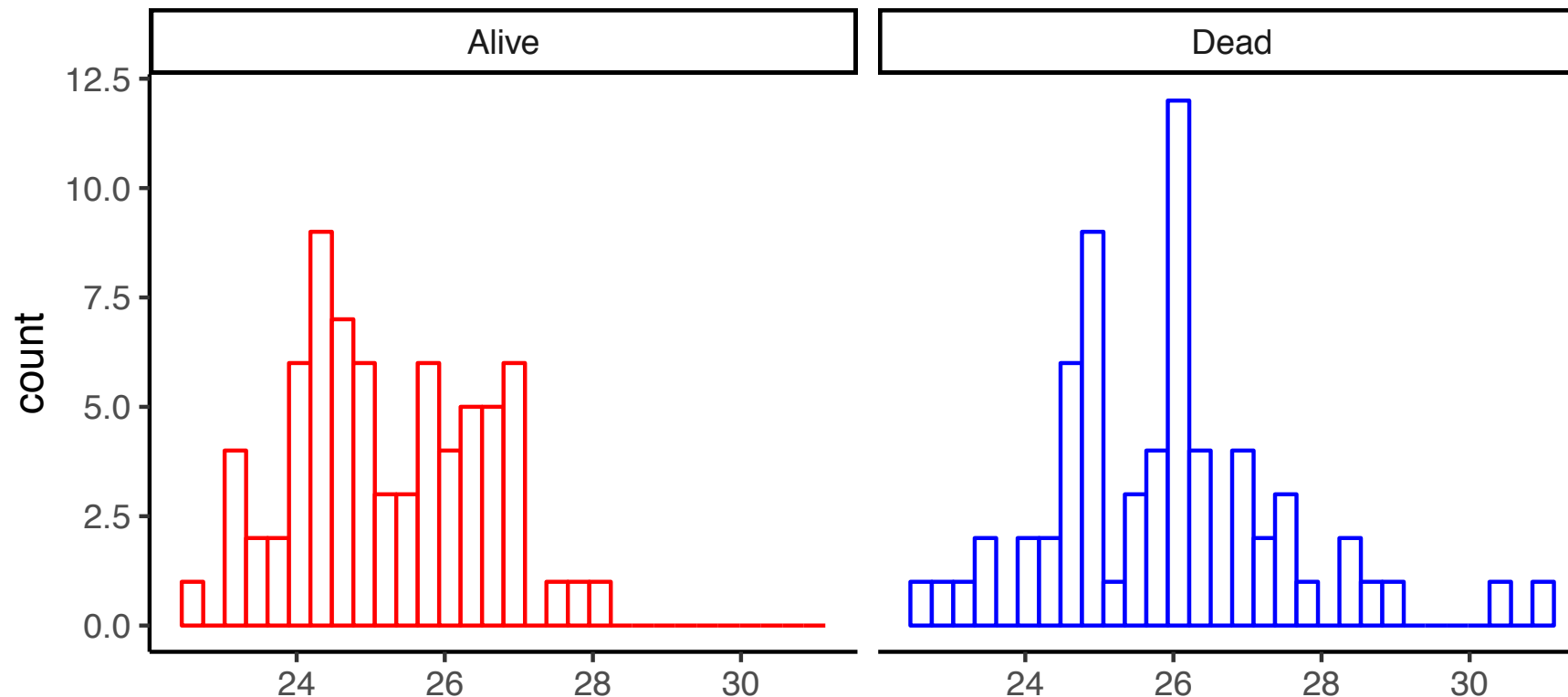


Mean vs median

When it is difficult to tell which might be "better", default to median.

This is particularly true for **small sample sizes** (more on why in coming weeks)

Does sparrow weight influence survival?



```
> summary(sp$Weight[sp$Survival == "Alive"])  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
22.60  24.20  24.90  25.21  26.30  28.00
```

```
> summary(sp$Weight[sp$Survival == "Dead"])  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
22.60  24.80  25.95  25.86  26.58  31.00
```

Probability vocabulary

Sample space

Event

Probability

Mutually exclusive

Probability distribution

Independent

Sample space and event

Sample space is the set of all possible outcomes of a random trial

Event is a subset of this set

Example: Roll a die

Sample space is $\{1, 2, 3, 4, 5, 6\}$

Events: roll a 4, roll something ≥ 5 , etc.

Probability

Probability of an event is the proportion of times the event would occur., i.e. event frequency, in an infinite number of trials

Empirical probabilities are based on a finite amount of data. If sample size expanded indefinitely, probabilities are measured with increasing precision and approach the true event probability. *This is pretty much what we can measure.*

Probability: roll a die

Theoretical probability

- $P[\text{roll a 5}] = 1/6$
- $P[\text{roll an even number}] = 1/2$

Empirical probability

- **After rolling 10x, we got:** 5 5 6 1 4 2 3 1 1 5 2 1
- $P[\text{roll a 5}] = 3/10$
- $P[\text{roll an even number}] = 4/10 = 2/5$

Basic properties of probabilities

Probabilities are always between 0 and 1

$$0 \leq P[\textit{event}] \leq 1$$

The sum of probabilities for all events equals 1

$$\sum_i P_i = 1$$

Mutually exclusive

Two events are **mutually exclusive** if they cannot both occur simultaneously

Mutually exclusive events: roll a 4 and a 1

Not mutually exclusive events: roll an even # and a 2

Probability distribution

The list of probabilities for all *mutually exclusive* outcomes of a random trial

A fair die has this distribution:

$$P[\text{roll } 1] = 1/6$$

$$P[\text{roll } 2] = 1/6$$

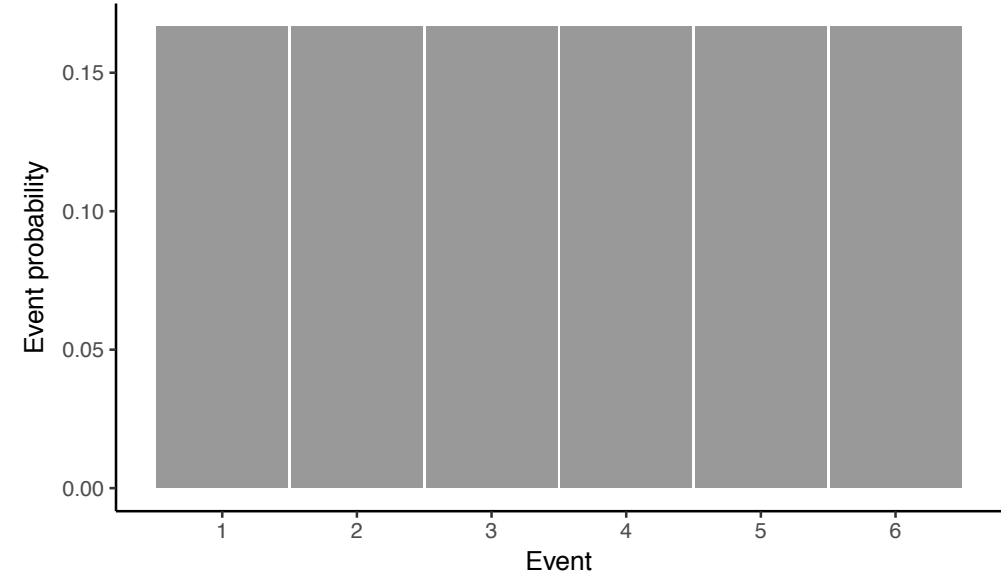
$$P[\text{roll } 3] = 1/6$$

$$P[\text{roll } 4] = 1/6$$

$$P[\text{roll } 5] = 1/6$$

$$P[\text{roll } 6] = 1/6$$

This is a **discrete probability distribution**



Independent

Two events are **independent** if the occurrence of one *does not change* the occurrence of another.

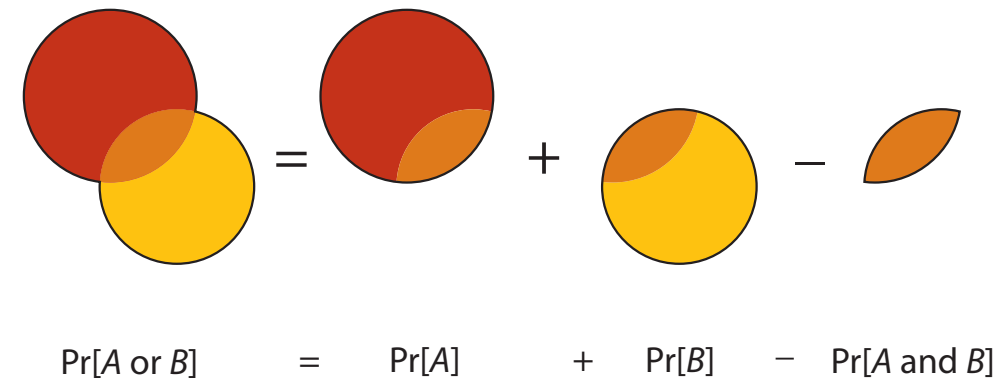
Probability rules

The probability of two *mutually exclusive* events A **or** B:

$$P[A \text{ or } B] = P[A] + P[B]$$

The probability of two *not mutually exclusive* events A **or** B:

$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$



What is the probability of rolling a 2 or a 5 on a fair die?

Are these events mutually exclusive? **Yes.**

$$P[2 \text{ or } 5] = P[\text{roll } 2] + P[\text{roll } 5] = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

What is the probability of rolling a 2 or an even number on a fair die?

Are these events mutually exclusive? **No.**

$$\begin{aligned} P[2 \text{ or even}] &= P[\text{roll } 2] + P[\text{roll even}] - P[2 \text{ and even}] \\ &= \frac{1}{6} + \frac{1}{2} - \frac{1}{6} = \frac{1}{2} \end{aligned}$$

Probability rules

The probability of two *mutually exclusive* events A **or** B:

$$P[A \text{ or } B] = P[A] + P[B]$$

The probability of two *not mutually exclusive* events A **or** B:

$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$

The probability of two *independent* events A **and** B:

$$P[A \text{ and } B] = P[A] \times P[B]$$



We add "or"



We multiply
"and"

Event independence

Mendel's experiment yielded **1600** pea pods:

- **900** were tall and green
- **300** were tall and yellow
- **300** were short and green
- **100** were short and yellow

Are tall and green pods independent?

Yes, **if** $P[A \text{ and } B] = P[A] \times P[B]$

Event independence

$$P[A \text{ and } B] = P[A] \times P[B]$$

Mendel's experiment yielded **1600** pea pods:

- **900** were tall and green
- **300** were tall and yellow
- **300** were short and green
- **100** were short and yellow

$$P[\text{green and tall}] = \frac{900}{1600} = \frac{9}{16}$$

$$P[\text{green}] \times P[\text{tall}] = \frac{(900 + 300)}{1600} \times \frac{(900 + 300)}{1600} = \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}$$

Yes, green and tall are independent events.

Question

Assume that a long (~infinite) stretch of DNA has A, C, G, T's in equal proportions, randomly occurring throughout.

What is the probability of seeing 10 A nucleotides in a row?

$$P[A] = 0.25$$

$$P[A \text{ and } A \text{ and } A \dots \text{and } A] = 0.25 \times 0.25 \dots = 0.25^{10} = 9.56 \times 10^{-7}$$

Question

Assume that a long (~infinite) stretch of DNA has A, C, G, T's in equal proportions, randomly occurring throughout.

What is the probability of **not** seeing 10 A nucleotides in a row?

$$1 - P[10 A's] = 1 - 9.56 \times 10^{-7} = 0.9999$$

We can calculate empirical probabilities directly from data

Example: A study assessed HIV risk associated with intravenous drug users and found these results:

	HIV+	HIV-	Total
Intravenous user	8	12	20
Not intravenous user	2	13	15
Total	10	25	35

Q1: What is the probability that a randomly chosen study participant is HIV+?

	HIV+	HIV-	Total
user	8	12	20
not user	2	13	15
Total	10	25	35

$$\begin{aligned} P(\text{HIV+}) &= (\text{number of HIV+}) / (\text{number participants}) \\ &= 10 / 35 = 2/7 \end{aligned}$$

Q2: What is the probability that a randomly chosen study participant who is HIV- is a user?

	HIV+	HIV-	Total
user	8	12	20
not user	2	13	15
Total	10	25	35

$$= 12 / 25$$

Q3: What is the probability that a randomly chosen study participant is either HIV+ or user but not both?

	HIV+	HIV-	Total
user	8	12	20
not user	2	13	15
Total	10	25	35

$$= (2+12)/35 = 14/35 = \mathbf{2/5}$$

Calculating probabilities directly from data frames

What is the probability of an iris being virginica, in the iris dataset?

```
# The denominator  
> nrow(iris)  
[1] 150
```

```
# The numerator  
> iris %>% filter(Species == "virginica") %>% tally()  
      n  
1  50
```

```
## The probability is 50/150 = 0.3333
```

Calculating probabilities directly from data frames

What is the probability of an iris being virginica and having petal lengths less than 5?

```
# The denominator  
> nrow(iris)  
[1] 150
```

```
# The numerator  
> iris %>% filter(Species == "virginica", Petal.Length  
< 5) %>% tally()  
      n  
1 6
```

```
## The probability is 6/150 = 0.04
```

Dependent events

Recall the probability of two independent events A and B:

$$P[A \text{ and } B] = P[A] \times P[B]$$

The probability of two dependent events A and B:

$$P[A \text{ and } B] = P[A|B] \times P[B]$$

Conditional Probability: Probability of A given B

Conditional probability, $P[A | B]$

Probability that a sick person is coughing

Probability that a person is coughing and sick

Probability that coughing person is sick

Conditional probability, $P[A | B]$

Probability that a sick person is coughing **$P[\text{coughing} | \text{sick}]$**

Probability that a person is coughing and sick **$P[\text{coughing and sick}]$**

Probability that coughing person is sick **$P[\text{sick} | \text{coughing}]$**

Conditional probabilities condition on *a priori* information

Example: Theoretical probabilities

A seed blows around a complex habitat. It can land on one of three (high-quality, medium-quality, poor-quality) soil types.

The probability of landing on each habitat is:

High-quality, 30%, Medium-quality, 20%, Low-quality, 50%

The probability of surviving each habitat is :

High-quality, 80%, Medium-quality, 30%, Low-quality, 10%

Question: What the probability a seed survives?

Example: Theoretical probabilities

Step 1: Convert text to probability statements

Step 2: Determine probability equation to solve the problem

Step 3: Plug in and solve

Convert text to prob. statements

The probability of landing on each habitat is:

High-quality, 30%, Medium-quality, 20%, Low-quality, 50%

The probability of surviving each habitat is :

High-quality, 80%, Medium-quality, 30%, Low-quality, 10%

$P[\text{land on high quality}] = 0.3$

$P[\text{land on med quality}] = 0.2$

$P[\text{land on low quality}] = 0.5$

$P[\text{survive on high quality}] = 0.8$

$P[\text{survive on med quality}] = 0.3$

$P[\text{survive on low quality}] = 0.1$

Determine probability equation

Seed can survive in three *mutually exclusive* ways:

- Land on high quality and survive
- Land on medium quality and survive
- Land on low quality and survive

$P[\text{seed survives}] =$

$P[\text{high qual. \& survives}] + P[\text{med qual. \& survives}] + P[\text{low qual. \& survives}] =$

$P[\text{high qual}] * P[\text{survives} | \text{high qual}] + \dots$



Survival is **dependent** on
land quality

Step 3: Plug in and solve

$P[\text{land on high quality}] = 0.3$
 $P[\text{land on med quality}] = 0.2$
 $P[\text{land on low quality}] = 0.5$

$P[\text{survive on high quality}] = 0.8$
 $P[\text{survive on med quality}] = 0.3$
 $P[\text{survive on low quality}] = 0.1$

$P[\text{seed survives}] =$

$P[\text{high qual. \& survives}] + P[\text{med qual. \& survives}] + P[\text{low qual. \& survives}] =$

$P[\text{high qual}] * P[\text{survives} | \text{high qual}] + \dots =$

$0.3 * 0.8 + 0.2 * 0.3 + 0.5 * 0.1 = \mathbf{0.35}$

Followup: What is the probability that a seed **does not** survive?

Part II

Now assume there is a 0.2 chance of not landing on any habitat, and therefore the seed will die. What is the new probability of survival assuming the seed lands?

Step 1: Text to probabilities

$P[\text{lands}] = 0.8$

$P[\text{does not land}] = 0.2$

Enter, Bayes Theorem

Recall:

- $P[A \text{ and } B] = P[B|A] \times P[A]$

Therefore, this is also true *and equal to the above*:

- $P[B \text{ and } A] = P[A|B] \times P[B]$

Put them together to derive **Bayes Theorem**:

$$P[A|B] = \frac{P[B|A] * P[A]}{P[B]}$$

Step 2: Probability equation

$$P[\text{seed survives} \mid \text{seed lands}] = \frac{(P[\text{lands} \mid \text{survives}] * P[\text{survives}])}{P(\text{lands})}$$

TAKE NOTICE:

THIS IS THE TYPE OF STATEMENT YOU WILL HAVE TO WRITE ON HW3

Step 3: Plug in and solve

P[land on high quality] = 0.3
P[land on med quality] = 0.2
P[land on low quality] = 0.5

P[survive on high quality] = 0.8
P[survive on med quality] = 0.3
P[survive on low quality] = 0.1

$$\begin{aligned} P[\text{seed survives} \mid \text{seed lands}] &= \\ & (P[\text{lands} \mid \text{survives}] * P[\text{survives}]) / P(\text{lands}) \\ &= (1 * 0.35) / 0.2 \\ &= \mathbf{0.175} \end{aligned}$$

Example: Theoretical probability and Bayes

Mammograms have a 7% false positive rate and a 25% false negative rate.

Assume women in the general population, have a 0.5% chance of having cancer at any time.

Probability statements:

$$P[\text{positive result} \mid \text{healthy}] = 0.07$$

$$P[\text{negative result} \mid \text{cancer}] = 0.25$$

$$P[\text{cancer}] = 0.005$$

Example 1

$$\begin{aligned}P[\text{positive result} \mid \text{healthy}] &= 0.07 \\P[\text{negative result} \mid \text{cancer}] &= 0.25 \\P[\text{cancer}] &= 0.005\end{aligned}$$

What is the probability that a healthy woman who gets a mammogram is given a negative result?

1. $P[\text{negative} \mid \text{healthy}]$
2. $P[\text{negative} \mid \text{healthy}] = 1 - P[\text{positive} \mid \text{healthy}]$
 - Remember, possible events sum to 1.
3. $P[\text{negative} \mid \text{healthy}] = 1 - 0.07 = \mathbf{0.93}$

Example 2

$$\begin{aligned}P[\text{positive result} \mid \text{healthy}] &= 0.07 \\P[\text{negative result} \mid \text{cancer}] &= 0.25 \\P[\text{cancer}] &= 0.005\end{aligned}$$

A woman gets a positive result from her mammogram. What is the probability she has cancer?

1. $P[\text{cancer} \mid \text{positive result}]$

2. $P[\text{cancer} \mid \text{positive result}] =$

$$(P[\text{positive result} \mid \text{cancer}] * P[\text{cancer}]) / P[\text{positive result}]$$

Example 2

$$\begin{aligned} P[\text{positive result} \mid \text{healthy}] &= 0.07 \\ P[\text{negative result} \mid \text{cancer}] &= 0.25 \\ P[\text{cancer}] &= 0.005 \end{aligned}$$

$$P[\text{cancer} \mid \text{positive result}] =$$

$$(P[\text{positive result} \mid \text{cancer}] * P[\text{cancer}]) / P[\text{positive result}]$$

When solving Bayes Theorem, the denominator generally requires a bit more work – Must consider **all** situations where it applies (remember seed survival?)

$$P[\text{positive result}] = P[\text{positive and cancer}] + P[\text{positive and healthy}]$$

Solving the denominator

$$\begin{aligned}P[\text{positive result} \mid \text{healthy}] &= 0.07 \\P[\text{negative result} \mid \text{cancer}] &= 0.25 \\P[\text{cancer}] &= 0.005\end{aligned}$$

$$P[\text{positive}] = P[\text{positive and cancer}] + P[\text{positive and healthy}]$$

$$\text{Recall: } P[A \text{ and } B] = P[B|A] \times P[A]$$

Therefore:

$$= P[\text{positive} \mid \text{cancer}] * P[\text{cancer}] + P[\text{positive} \mid \text{healthy}] * P[\text{healthy}]$$

$$= 0.75 * 0.005 + 0.07 * 0.995$$

$$= 0.0734$$

Put it all together

$$\begin{aligned} P[\text{positive result} \mid \text{healthy}] &= 0.07 \\ P[\text{negative result} \mid \text{cancer}] &= 0.25 \\ P[\text{cancer}] &= 0.005 \end{aligned}$$

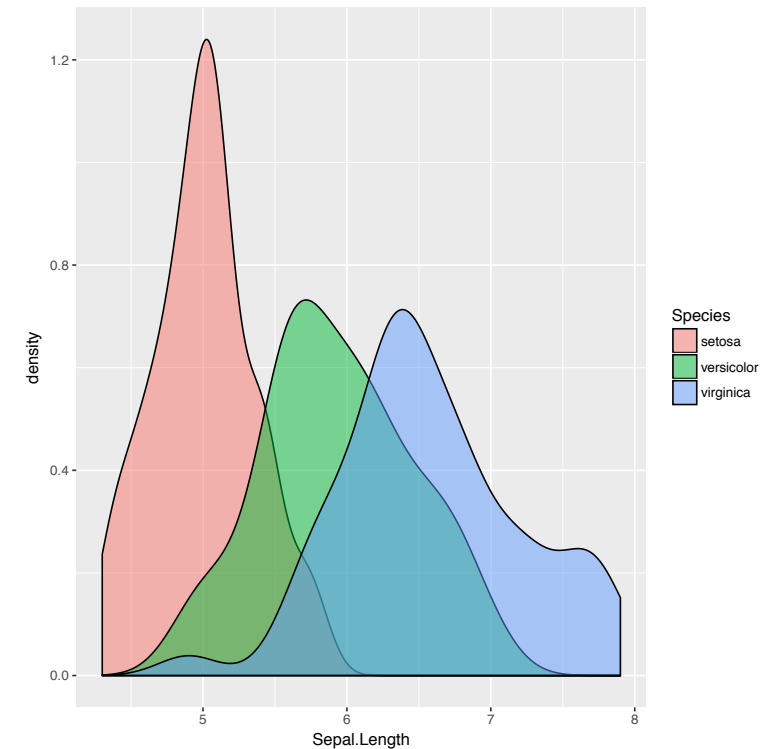
$$P[\text{cancer} \mid \text{positive result}] =$$

$$\begin{aligned} & (P[\text{positive result} \mid \text{cancer}] * P[\text{cancer}]) / P[\text{positive result}] = \\ & (0.75 * 0.005) / 0.0734 = \mathbf{0.514} \end{aligned}$$

BREAK

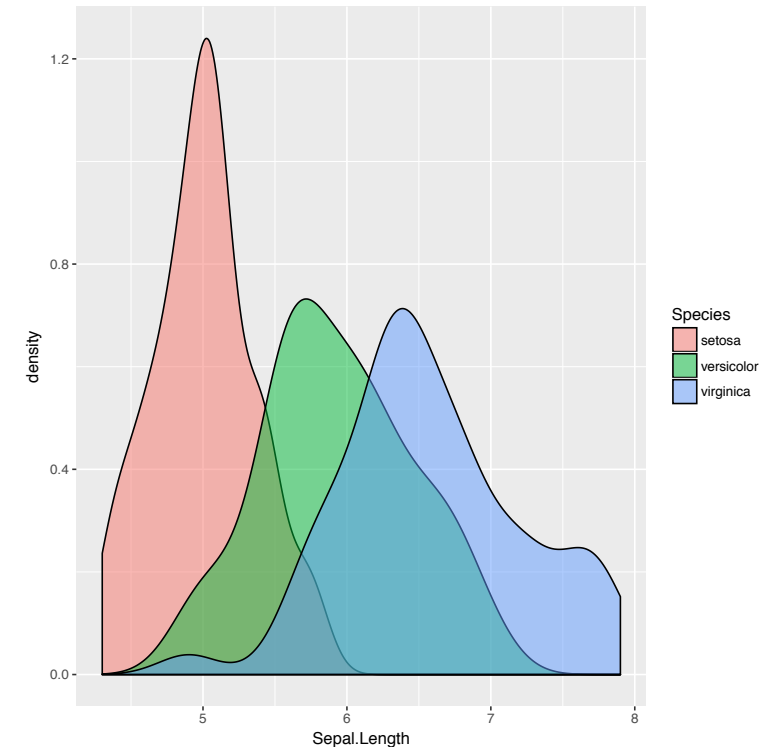
ggplot2: saving plots

```
> ggplot(iris, aes(x = Sepal.Length, fill = Species)) + geom_density( alpha = 0.5 )
```



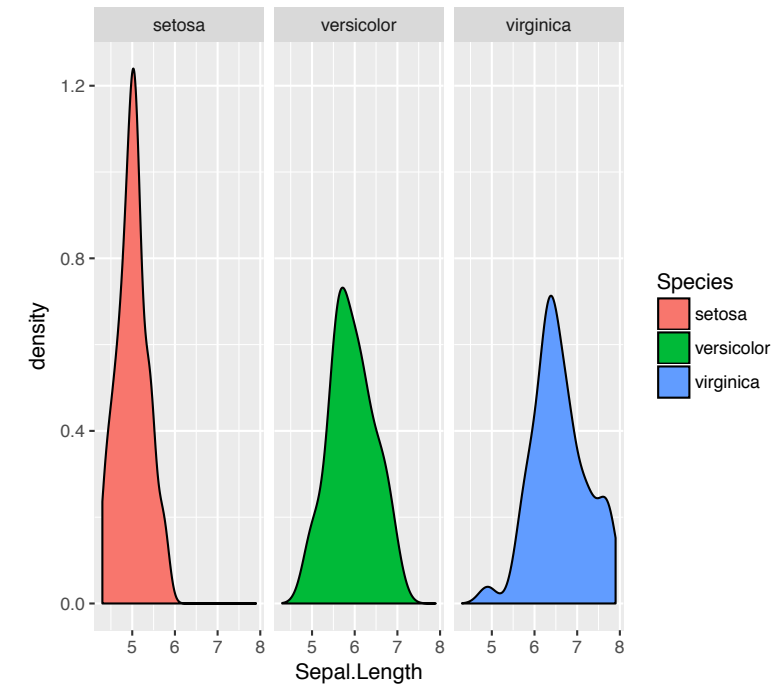
ggplot2: saving plots

```
> P <- ggplot(iris, aes(x = Sepal.Length, fill = Species)) + geom_density( alpha = 0.5 )  
  
> ### Save as PNG  
> ggsave("plot.png", P)  
  
> ### Save as PDF  
> ggsave("plot.pdf", P)
```



ggplot2: Faceting plots

```
> ggplot(iris, aes(x = Sepal.Length, fill = Species)) + geom_density() + facet_grid(~Species)
```



ggplot2: Faceting plots

```
> head(iris2)
```

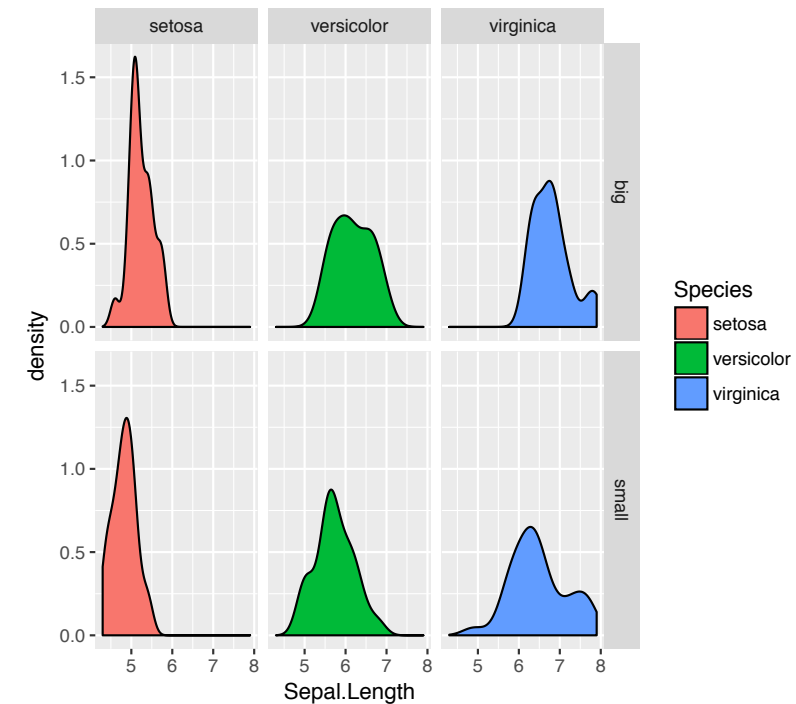
```
Source: local data frame [150 x 6]
```

```
Groups: Species [3]
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	size
	<dbl>	<dbl>	<dbl>	<dbl>	<fctr>	<chr>
1	5.1	3.5	1.4	0.2	setosa	big
2	4.9	3.0	1.4	0.2	setosa	small
3	4.7	3.2	1.3	0.2	setosa	small
4	4.6	3.1	1.5	0.2	setosa	small
5	5.0	3.6	1.4	0.2	setosa	big
6	5.4	3.9	1.7	0.4	setosa	big

ggplot2: Faceting plots

```
> ggplot(iris2, aes(x = Sepal.Length, fill = Species)) + geom_density() + facet_grid(size~Species)
```



dplyr: Joining related dataframes

```
> data1
```

	x	y	z
1	3.108060	61.48849	
2	8.976264	55.68174	
3	11.673850	56.32225	
4	8.551282	58.53424	
5	5.819844	61.71424	

```
> data2
```

	x	a	b
1	-2.76205636	112.9588	
2	-1.44485264	149.3682	
3	-1.14390532	132.8789	
4	-2.86488120	143.6860	
5	-2.91982194	121.3927	

```
> left_join(data1, data2)
```

Joining, by = "x"

	x	y	z	a	b
1	3.108060	61.48849		-2.76205636	112.9588
2	8.976264	55.68174		-1.44485264	149.3682
3	11.673850	56.32225		-1.14390532	132.8789
4	8.551282	58.53424		-2.86488120	143.6860
5	5.819844	61.71424		-2.91982194	121.3927

left_join() creates NA's when missing

```
> data1
```

	x	y	z
1	3.108060	61.48849	
2	8.976264	55.68174	
3	11.673850	56.32225	
4	8.551282	58.53424	
5	5.819844	61.71424	

```
> data3
```

	x	a	b
1	-2.76205636	112.9588	
2	-1.44485264	149.3682	
3	-1.14390532	132.8789	
5	-2.91982194	121.3927	

Missing
x=4

```
> left_join(data1, data3)
```

Joining, by = "x"

	x	y	z	a	b
1	3.108060	61.48849		-2.762056	112.9588
2	8.976264	55.68174		-1.444853	149.3682
3	11.673850	56.32225		-1.143905	132.8789
4	8.551282	58.53424		NA	NA
5	5.819844	61.71424		-2.919822	121.3927

left_join() only preserves what is in the left data frame

```
> data1
```

	x	y	z
1	3.108060	61.48849	
2	8.976264	55.68174	
3	11.673850	56.32225	
4	8.551282	58.53424	
5	5.819844	61.71424	

```
> data3
```

	x	a	b
1	-2.76205636	112.9588	
2	-1.44485264	149.3682	
3	-1.14390532	132.8789	
5	-2.91982194	121.3927	

```
> left_join(data3, data1)
```

```
Joining, by = "x"
```

	x	y	z	a	b
1	3.108060	61.48849		-2.762056	112.9588
2	8.976264	55.68174		-1.444853	149.3682
3	11.673850	56.32225		-1.143905	132.8789
5	5.819844	61.71424		-2.919822	121.3927

right_join() is the opposite

```
> data1
```

	x	y	z
1	3.108060	61.48849	
2	8.976264	55.68174	
3	11.673850	56.32225	
4	8.551282	58.53424	
5	5.819844	61.71424	

```
> data3
```

	x	a	b
1	-2.76205636	112.9588	
2	-1.44485264	149.3682	
3	-1.14390532	132.8789	
5	-2.91982194	121.3927	

```
> right_join(data1, data3) ## Equivalent to left_join(data3, data1)
```

```
Joining, by = "x"
```

	x	y	z	a	b
1	3.108060	61.48849		-2.762056	112.9588
2	8.976264	55.68174		-1.444853	149.3682
3	11.673850	56.32225		-1.143905	132.8789
5	5.819844	61.71424		-2.919822	121.3927

inner_join() only joins what the tables have in common

```
> data4
```

**Missing
x=2** →

	x	y	z
1	3.108060	61.48849	
3	11.673850	56.32225	
4	8.551282	58.53424	
5	5.819844	61.71424	

```
> data3
```

	x	a	b
1	-2.76205636	112.9588	
2	-1.44485264	149.3682	
3	-1.14390532	132.8789	
5	-2.91982194	121.3927	

```
> inner_join(data4, data3)
```

Joining, by = "x"

	x	y	z	a	b
1	3.108060	61.48849		-2.762056	112.9588
3	11.673850	56.32225		-1.143905	132.8789
5	5.819844	61.71424		-2.919822	121.3927

Joins galore

See this vignette if you're extra curious (not required):

<https://cran.r-project.org/web/packages/dplyr/vignettes/two-table.html>