Week 3: Descriptives and Probability

Outline

1. Review some descriptive statistics (Ch. 2 OpenIntro)
2. Describing variables: central tendency and dispersion
3. Contingency tables or crosstabs for analyzing categorical data (ordinal and nominal variables)
4. Probability
   - Conditional probability, joint probability, and marginal probability.
   - Using contingency tables/crosstabs
5. Random variables (continue into next week)

1. Descriptive Inference

- Different forms of “description.”

- Generalization (inference) v. understanding particular observations.
  - Producing general knowedge v. collecting facts

- Examples of descriptive inference:
  - Candidate evaluation
  - Issue polarization

2. Describing Variables

- Summarize and describe our variables.
- Measurement process:
  - *Concept* → *Measurement strategy* → ***VARIABLE***

- What types of "descriptive inferences" can we make about our variable?
  - *Measures of central tendency* (mean, median, mode)
  - *Dispersion* in our variable; how is it distributed across our units of analysis? How **different** are the observations?
  - *Frequency distributions* (this also gets leverage on how dispersed it is)
  - *Graphical displays* (visualizing central tendency and dispersion)

- **Importance**

Describing Variables – Central Tendency

- *Mode*, *median*, and *mean*.

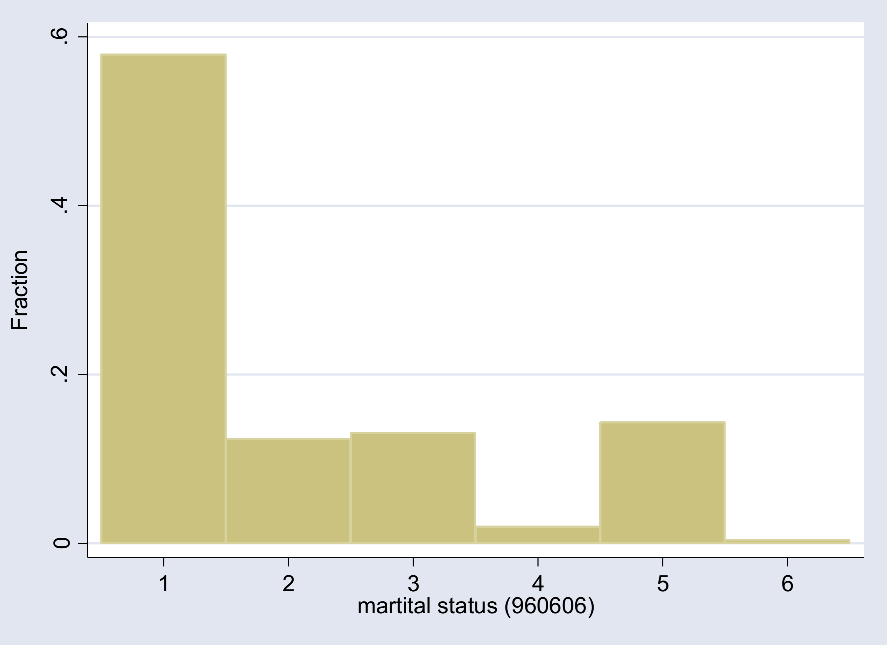
- The level of measurement influences which measure to use.

- Think about measures of central tendency as a “best guess” of a typical person’s level of trust, or any other variable.

- It’s a **generalization**.

Mode

- *Mode*: the value of the variable that contains the most units of analysis; i.e., the value that contains the highest frequency.
  - What's the mode of marital status?



Marital Status	Freq.	Percent
1. married	319	57.89
2. widowed	68	12.34
3. divorced	72	13.07
4. separated	11	2.00
5. never married	79	14.34
6. partners, not married	2	0.36
Total	551	100

Median

- The median is the middle value of the variable if you were to sort values from smallest to largest. Half of the cases fall below the median and half fall above it; 50th percentile.

- Let’s say we have 9 individuals, and I want to find the median level of trust in government. Here are the data:

ID	Trust
1	0.25
2	0.33
3	0.67
4	0.39
5	0.29
6	0.21
7	0.54
8	0.40
9	0.21

I need to sort the data from smallest to largest.

ID	Trust
6	0.21
9	0.21
1	0.25
5	0.29
2	0.33
4	0.39
8	0.40
7	0.54
3	0.67

Median

- What if our data looked like this instead?

ID	Trust
6	0.21
9	0.21
1	0.25
5	0.29
10	0.31
2	0.33
4	0.39
8	0.40
7	0.54
3	0.67

Median

- What if our data looked like this instead?

ID	Trust
6	0.21
9	0.21
1	0.25
5	0.29
10	0.31
2	0.33
4	0.39
8	0.40
7	0.54
3	0.67

•What's the median? Since there are **10** observations, there isn't one exact value where exactly half fall below the value, and half fall above it.

Median

- What if our data looked like this instead?

ID	Trust
6	0.21
9	0.21
1	0.25
5	0.29
10	0.31
2	0.33
4	0.39
8	0.40
7	0.54
3	0.67

•What's the median? Since there are **10** observations, there isn't one exact value where exactly half fall below the value, and half fall above it.

• Take the average of the two middle values

•Median = (.31 + .33)/2 = **.32**

Mean

- The *mean* of a variable is the “average” value of it. It’s the sum of the values of the variable across units of analysis divided by the total number of observations.

- “Continuous” variables

Mean

- Return to our trust example.

ID	Trust
6	0.21
9	0.21
1	0.25
5	0.29
2	0.33
4	0.39
8	0.40
7	0.54
3	0.67

- Mean level of trust is:

$$\frac{0.21+0.21+0.25+0.29+0.33+0.39+0.4+0.54+0.67}{9}=0.37$$

Dispersion

- Upon calculating a variable’s central tendency, we’ll also want to understand its dispersion, that is, how it is distributed.
  - How much variation is in our variable?
  - How *different* are the observations in the values of the variable?
  - What’s the spread of observations look like?

Frequency Distributions

- Best for ordinal and nominal variables.

Frequency Distributions; Dispersion

- Compare median, mean, and mode. Similar or different?

- Why does it matter?

  - "Single-peaked," unimodal distribution versus bimodal or multimodal variables.

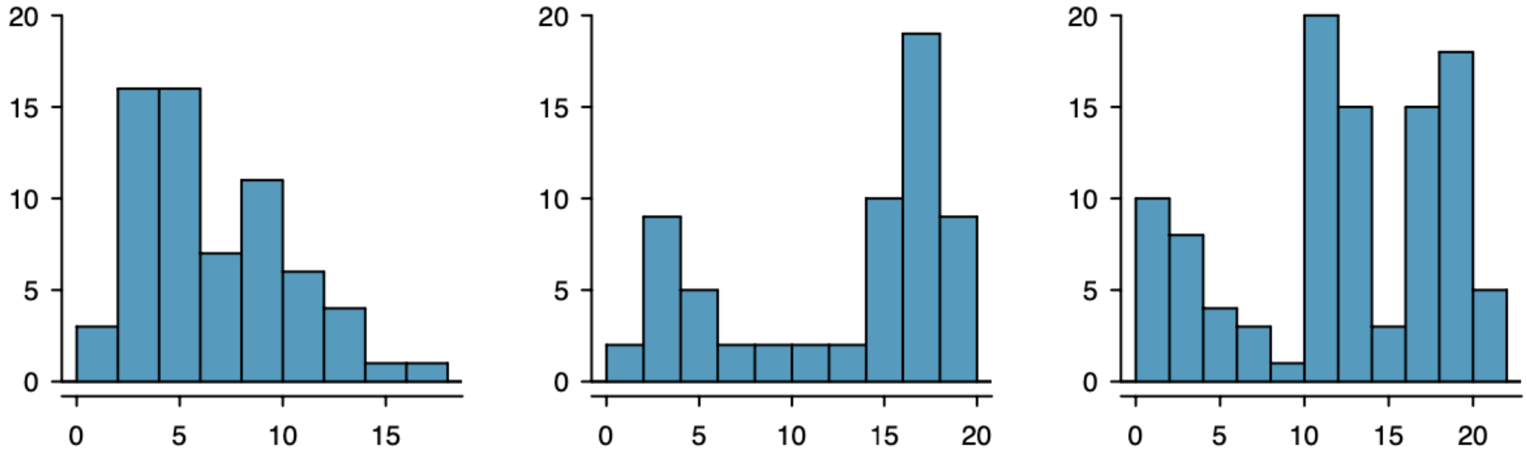


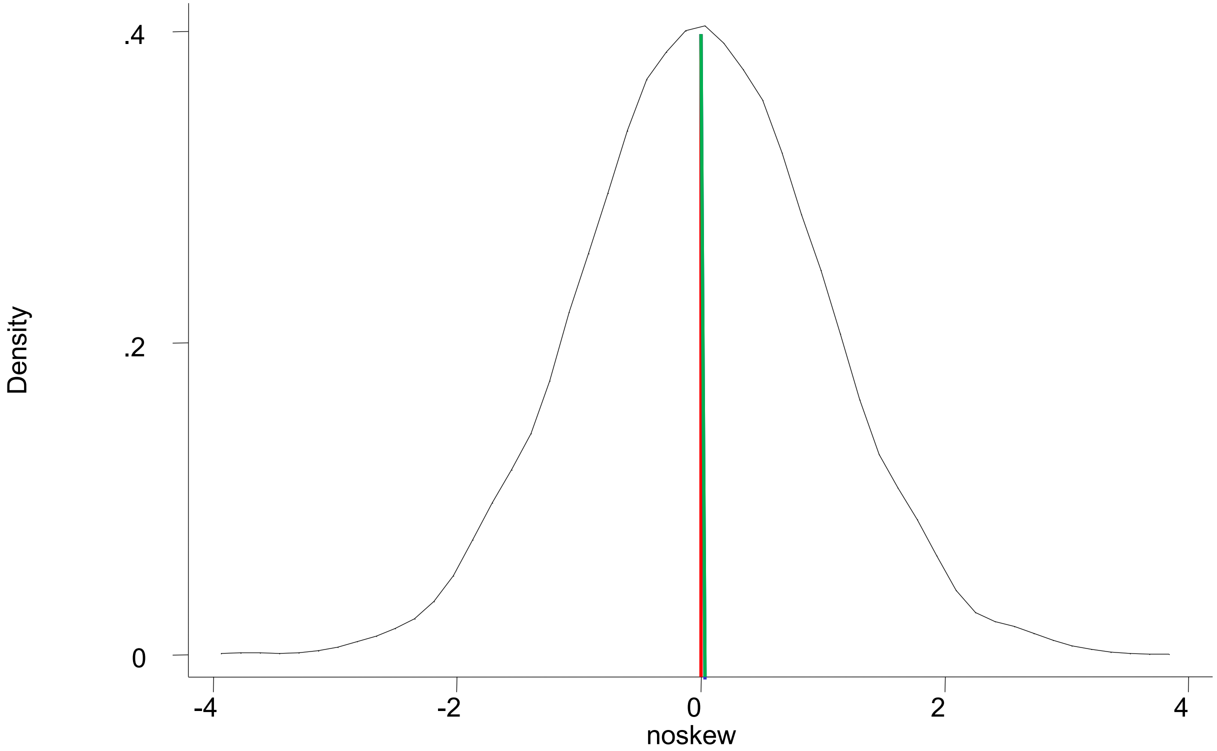
Figure 2.7: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal. Note that we've said the left plot is unimodal intentionally. This is because we are counting *prominent* peaks, not just any peak.

Skewness

- *Skewness*: The extent to which a variable's distribution deviates from a nice, bell-shaped form.

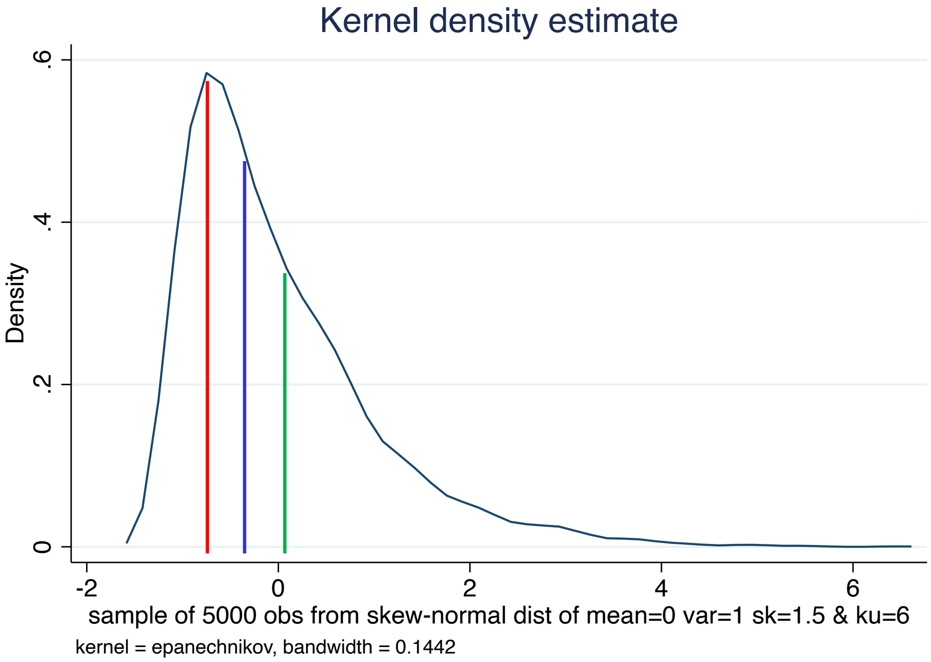
Absence of Skewness

- Mean = median = mode



Skewness

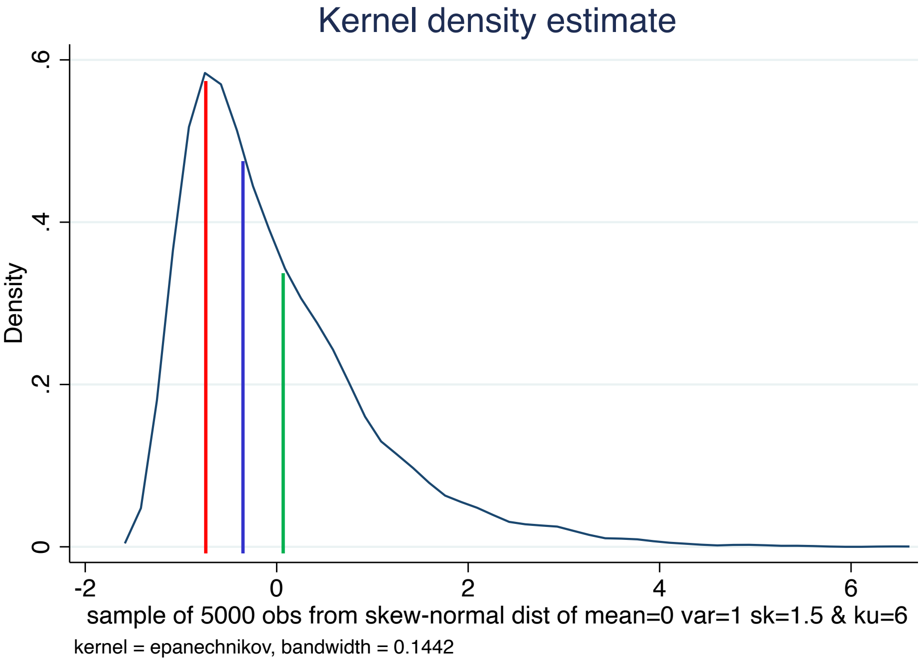
Positive skew / skewed to the right



Mean > median > mode

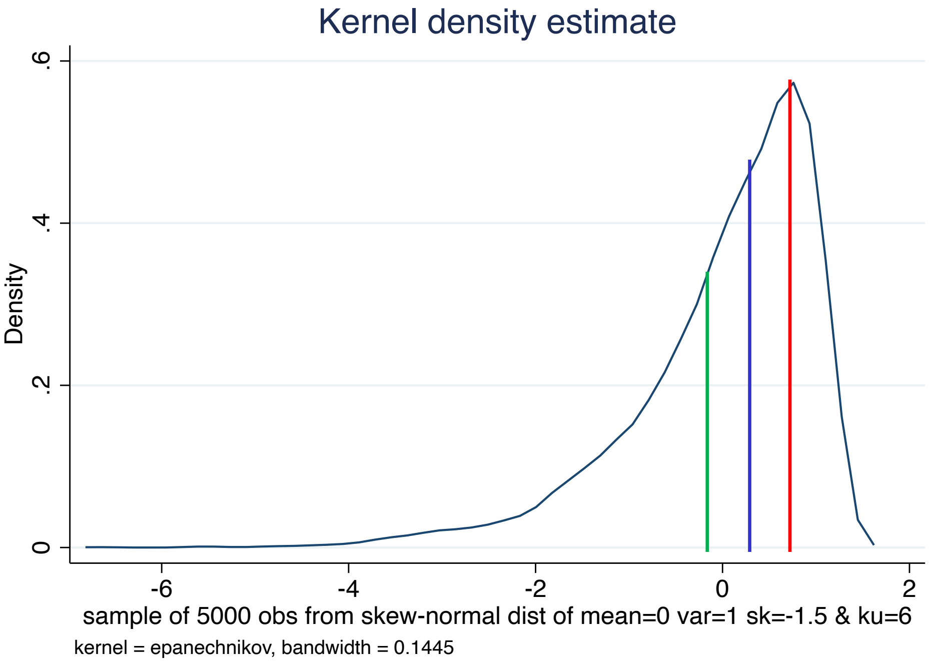
Skewness

Positive skew / skewed to the right



Mean > median > mode

Negative skew / skewed to the left



Mean < median < mode

Skewness: Comparing Mean and Median

- What kind of variables tend to be skewed?

- In skewed distributions, the mean is pulled in the direction of the skewness.

- Importantly, the median and mode are not as sensitive to skewness. The median in general is *unfazed by dispersion*.
  - Again, levels of measurement important…..

- Bottom line: *For skewed distributions, the median offers a more accurate depiction of a variable's central tendency than the mean.*

- The mean is "biased" by skewness, the median is not.

Bimodal Distributions

- Back to bimodal (and multimodal) distributions

- Is central tendency even meaningful?

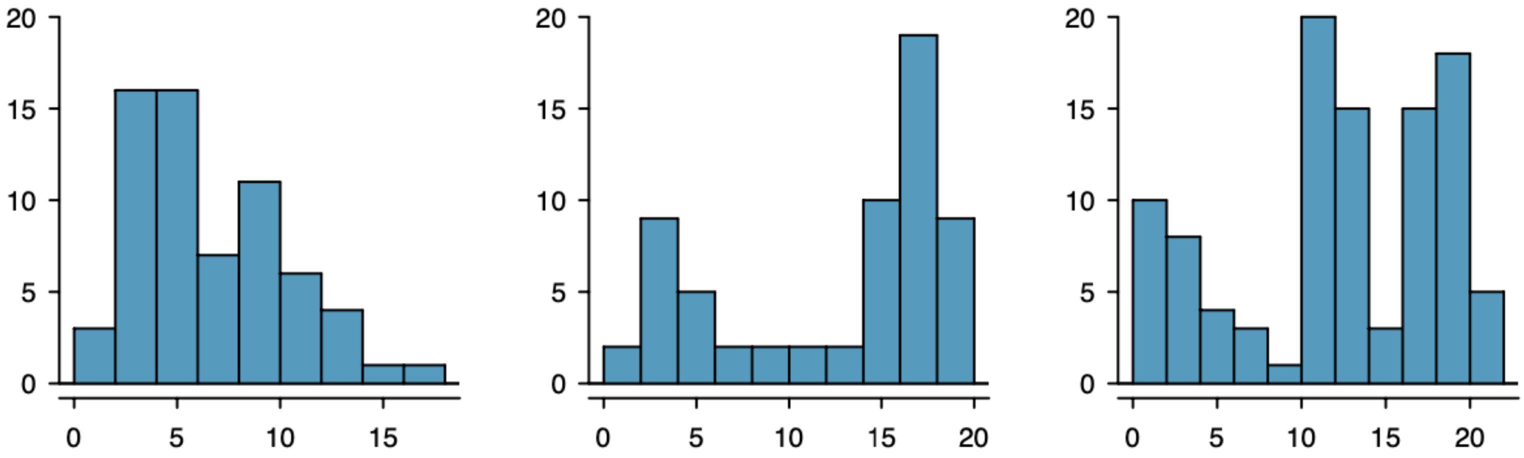


Figure 2.7: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal. Note that we've said the left plot is unimodal intentionally. This is because we are counting *prominent* peaks, not just any peak.

Central Tendency and Dispersion

- Think about the nature and degree of dispersion in the variable when making descriptive inferences.

- A variable’s dispersion has consequences for how we describe central tendency.

- Is dispersion the dominant feature?

- Another issue: Outliers?

Variance and Standard Deviation

- How much “spread” is there around the mean?
- The “meaning of the mean.”
- Var and s.d. best for variables with normal-like, single-peaked distribution.
- Calculating var and s.d.
- Interpreting s.d.

3. Contingency Tables/Crosstabs

- Best for categorical variables, particularly ordinal variables (with small number of categories) or also nominal variables.
  - Nominal: Race -> vote choice

- Useful for examining relationships between variable, and specifically the effect of X on Y.
  - How experiencing an event depends on another condition or event.
  - Can also be used for associations between variables where there isn't necessarily causal effect posited.

- Rules of crosstabs when examining the effect of X on Y:
  - Y is the row variable.
  - X is the column variable.
  - Report column percentages.
  - Compare column percentages across columns
  - Example: Education on voter turnout.

4. Probability

- Probability as a long-run relative frequency

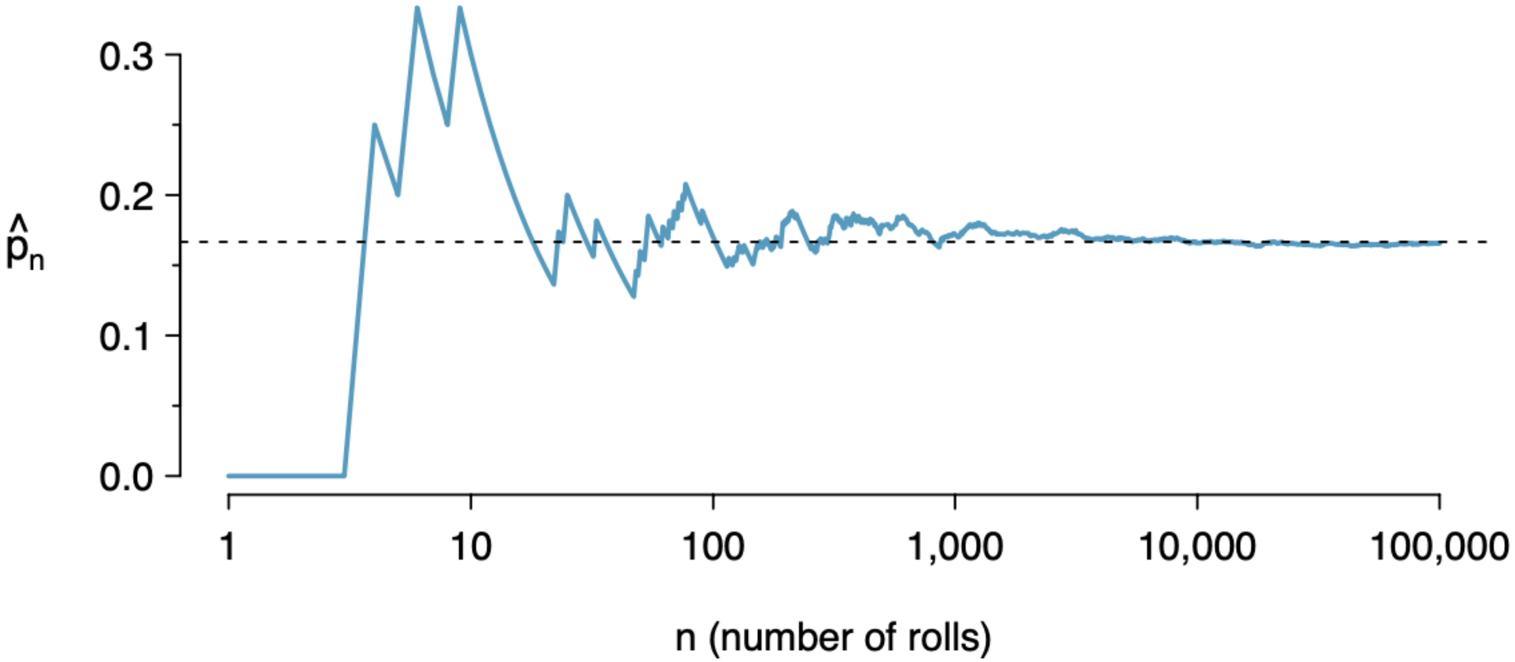
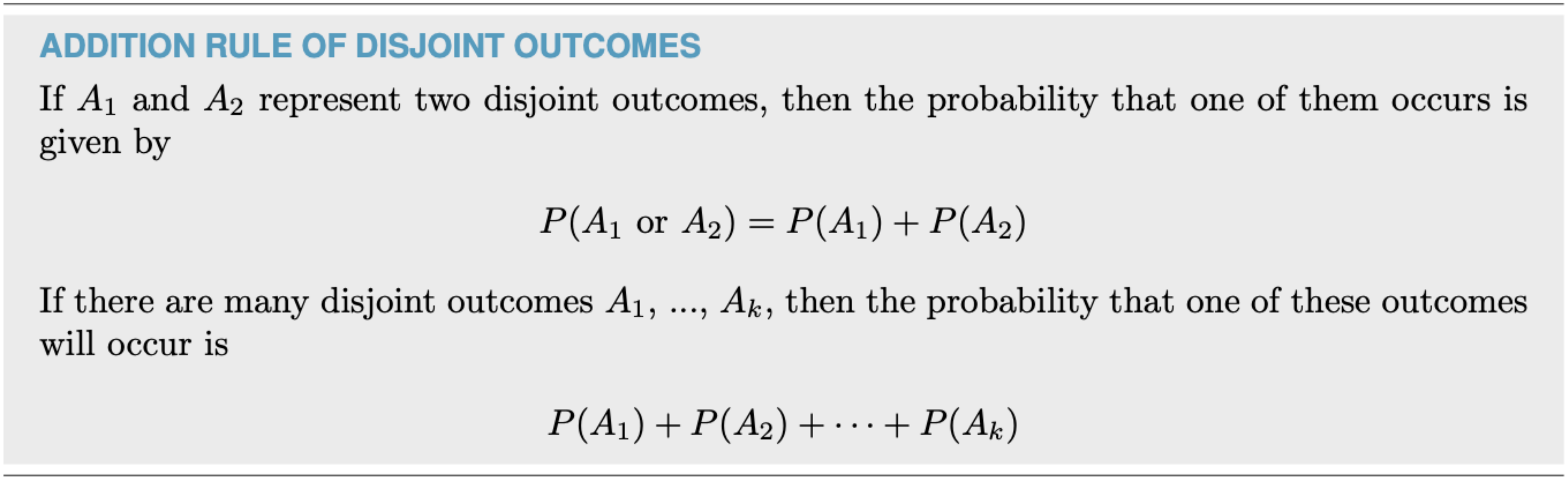


Figure 3.1: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

Disjoint / Mutually Exclusive Events

- Disjoint: When two events can't both occur. Dice rolling example.



ADDITION RULE OF DISJOINT OUTCOMES

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

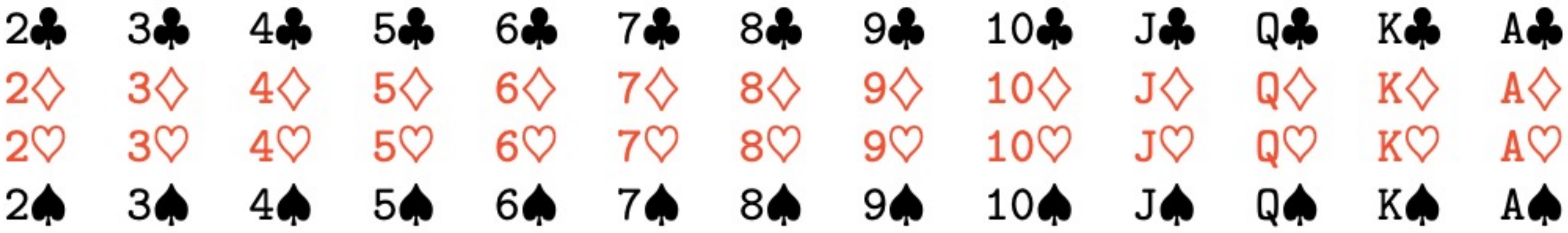
If there are many disjoint outcomes A_1, ..., A_k, then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k)$$

- A=rolling a 1; B= rolling a 2. Disjoint?
- A=rolling a 1; B=rolling an odd number. Disjoint?

When events are not disjoint; Venn diagram

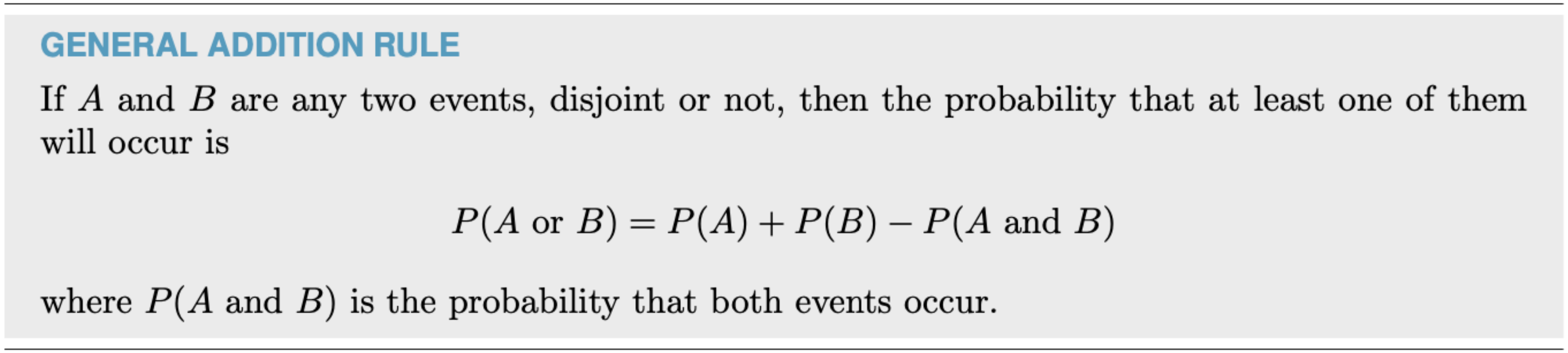
- Deck of cards



2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
2♢	3♢	4♢	5♢	6♢	7♢	8♢	9♢	10♢	J♢	Q♢	K♢	A♢
2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡	A♡
2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠

Exercise: A=draw a single-digit card (Ace through 9)
B=draw a spade

What's **P(A or B)?**



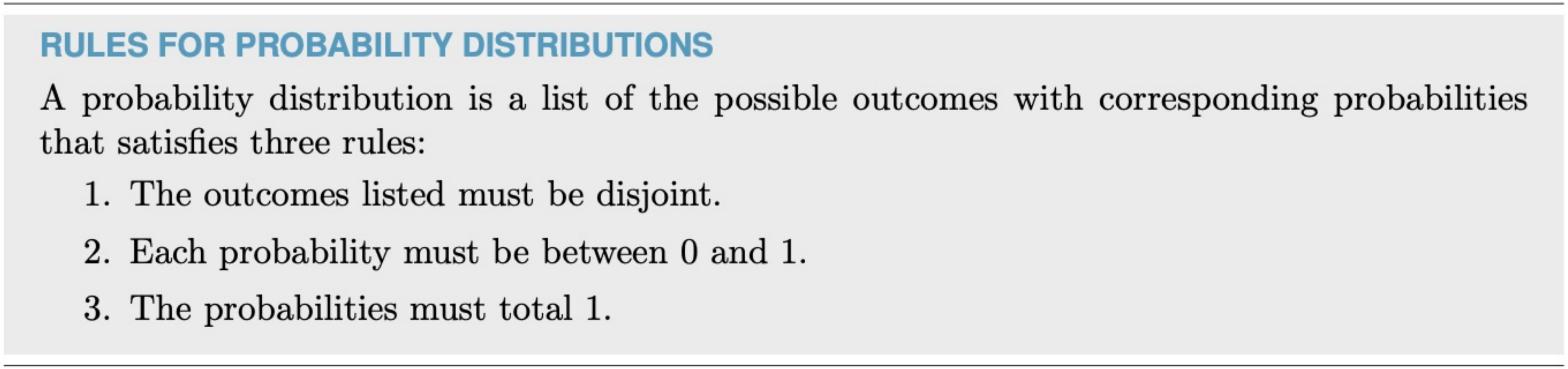
GENERAL ADDITION RULE

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

where $P(A \text{ and } B)$ is the probability that both events occur.

Probability Distributions



RULES FOR PROBABILITY DISTRIBUTIONS

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

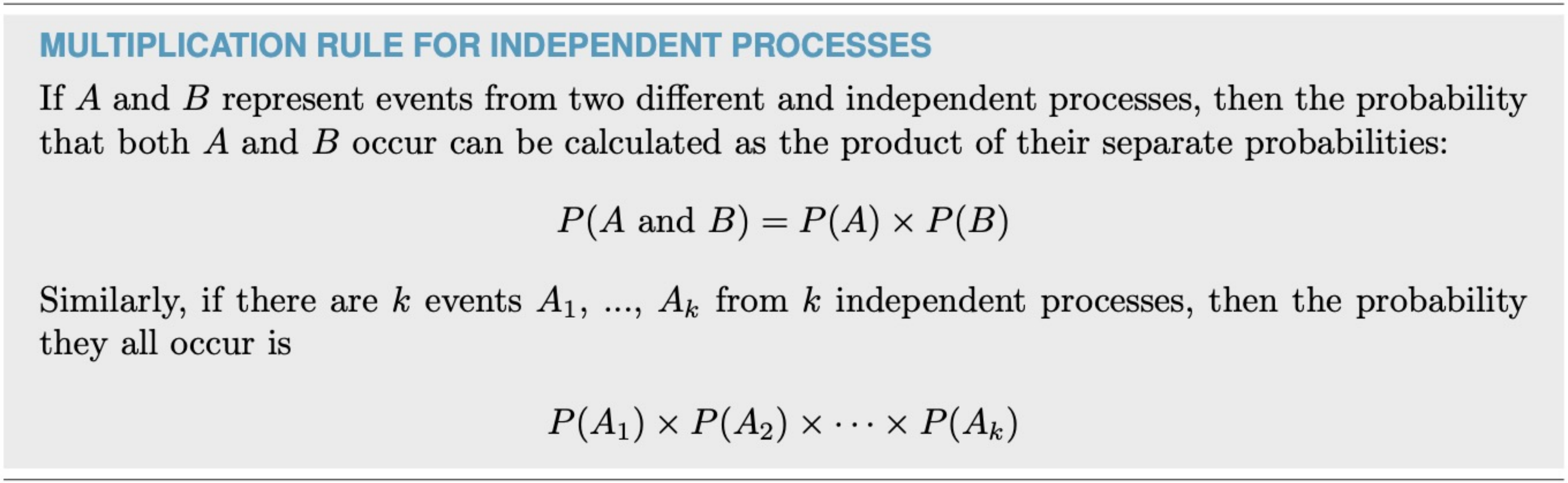
A simple frequency distribution is a probability distribution.

- Education example

Independence of Events

- “Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.”

- Coin flip, dice roll, sampling individuals in a survey poll.



MULTIPLICATION RULE FOR INDEPENDENT PROCESSES

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Similarly, if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

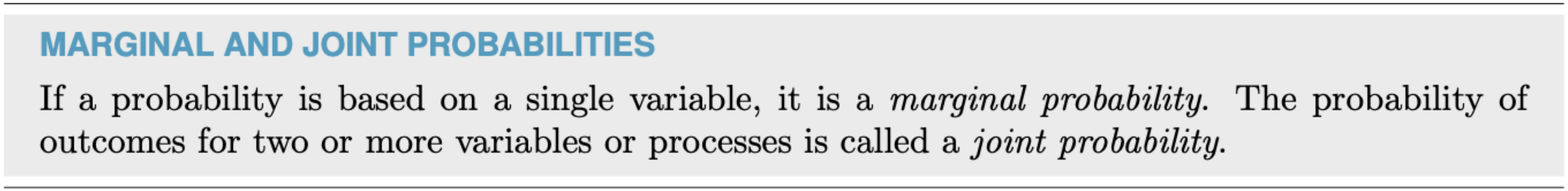
Conditional Probability

- P(A | B) = probability that some event, A, occurs *given* some other event, B, occurs.

- Let's test your intuition by looking at a contingency table (in R) between ideology and party identification in survey data (NES).

- Calculate:
  - P(Liberal | Democrat)
  - P(Liberal | Republican)
  - What feature of crosstabs can we call on to get those quantities?
- Calculate:
  - P(Democrat | Liberal)
  - P(Republican | Liberal)
  - What feature of crosstabs can we call on to get those quantities?

Marginal and Joint Probabilities



MARGINAL AND JOINT PROBABILITIES

If a probability is based on a single variable, it is a *marginal probability.* The probability of outcomes for two or more variables or processes is called a *joint probability.*

Marginal probability:	P(A)
Joint probability:	P(A and B) or P(A, B)
Conditional probability:	P(A | B)

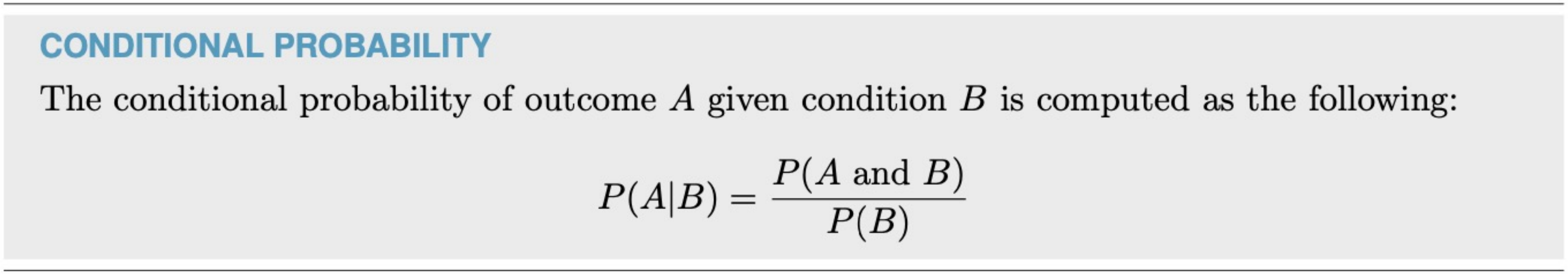
- Open “ideopid.xlsx.” Study this contingency table.

- Which cells are associated with **joint probabilities** and which are associated with **marginal probabilities**?

1. Calculate all joint probabilities. (How many are there?)
2. Calculate all marginal probabilities (How many are there?)

Conditional Probability Equation

- Formalize the logic we used earlier:

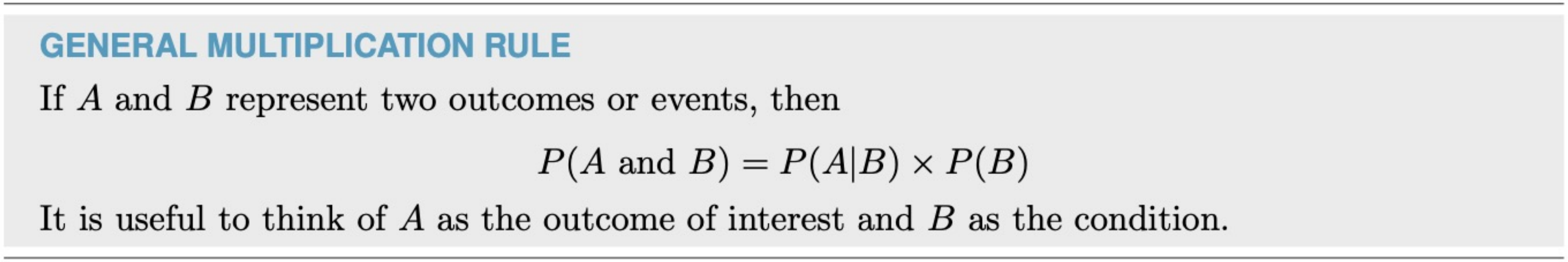


CONDITIONAL PROBABILITY

The conditional probability of outcome A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

- And it follows that joint probability can be formalized as:



GENERAL MULTIPLICATION RULE

If A and B represent two outcomes or events, then

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

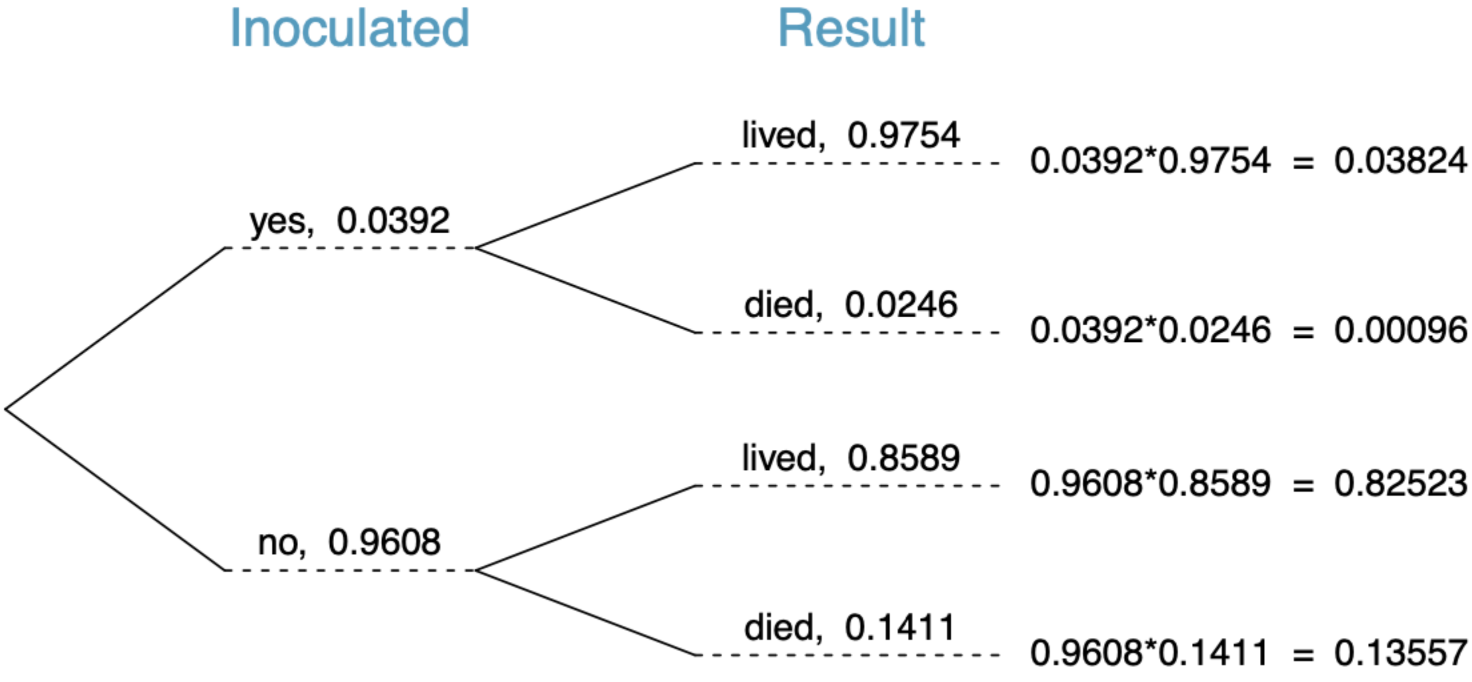
It is useful to think of A as the outcome of interest and B as the condition.

Tree Diagrams

- Similar logic as contingency tables – p. 99 & 102 in OpenIntro

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000



Inoculated

Result

yes, 0.0392

lived, 0.9754

0.0392*0.9754 = 0.03824

died, 0.0246

0.0392*0.0246 = 0.00096

no, 0.9608

lived, 0.8589

0.9608*0.8589 = 0.82523

died, 0.1411

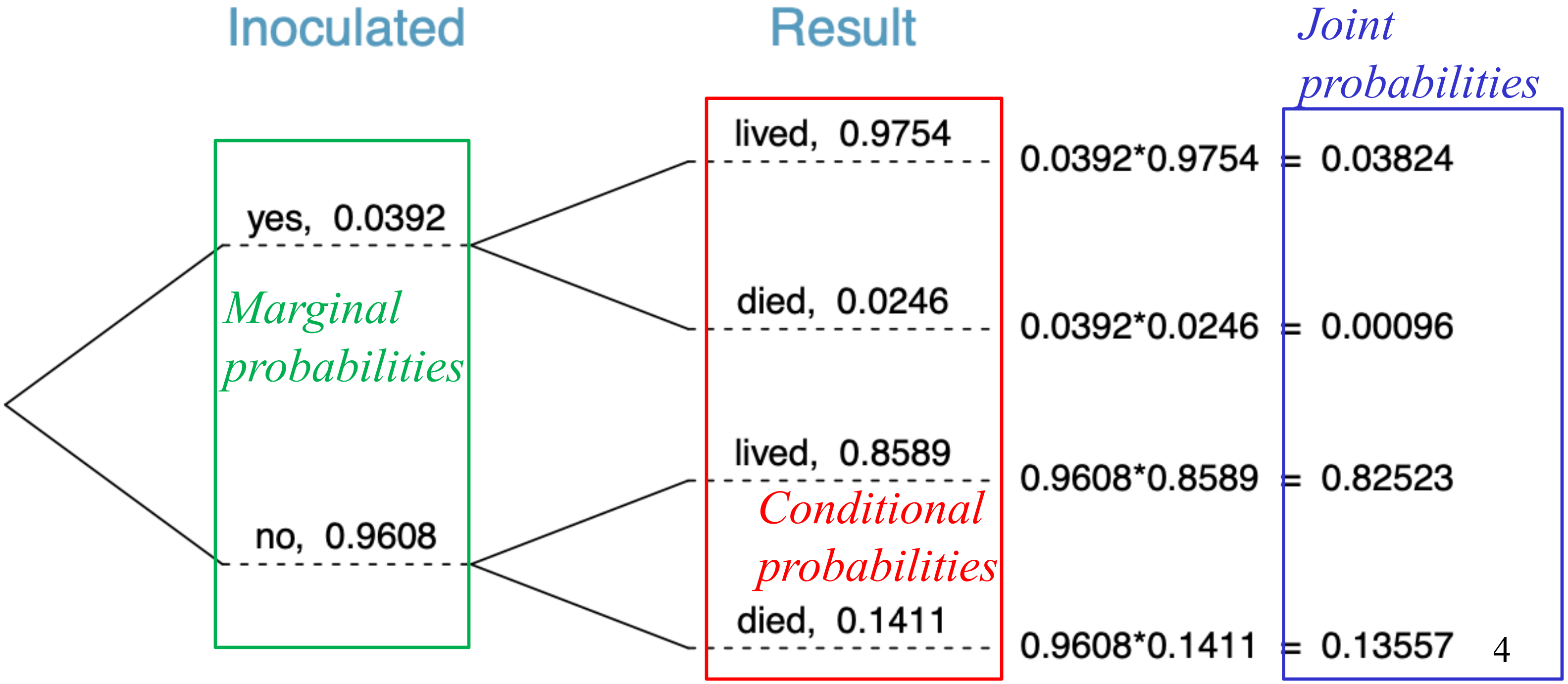
0.9608*0.1411 = 0.13557

Tree Diagrams

- Similar logic as contingency tables – p. 99 & 102 in OpenIntro

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000



Inoculated

Result

Joint probabilities

yes, 0.0392

no, 0.9608

Marginal probabilities

lived, 0.9754

died, 0.0246

lived, 0.8589

died, 0.1411

Conditional probabilities

0.0392*0.9754 = 0.03824

0.0392*0.0246 = 0.00096

0.9608*0.8589 = 0.82523

0.9608*0.1411 = 0.13557

Tree Diagrams

- 3.22 on p. 111. “Exit Poll.” Next assignment.

Bayes' Theorem

- Notion of inverse probability and updating.

- We can learn about P(A | B) by working with P(B | A)

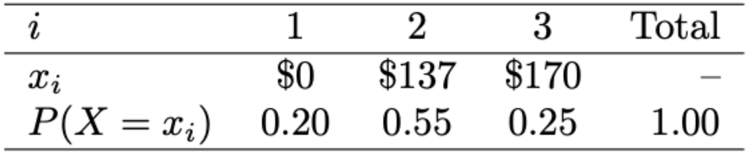
$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)}$$

Random Variables

- Relevant for sampling theory and distributions.
  - In regression modeling, we assume the dependent variable, Y, follows a conditional distribution. Y is treated as a random variable.

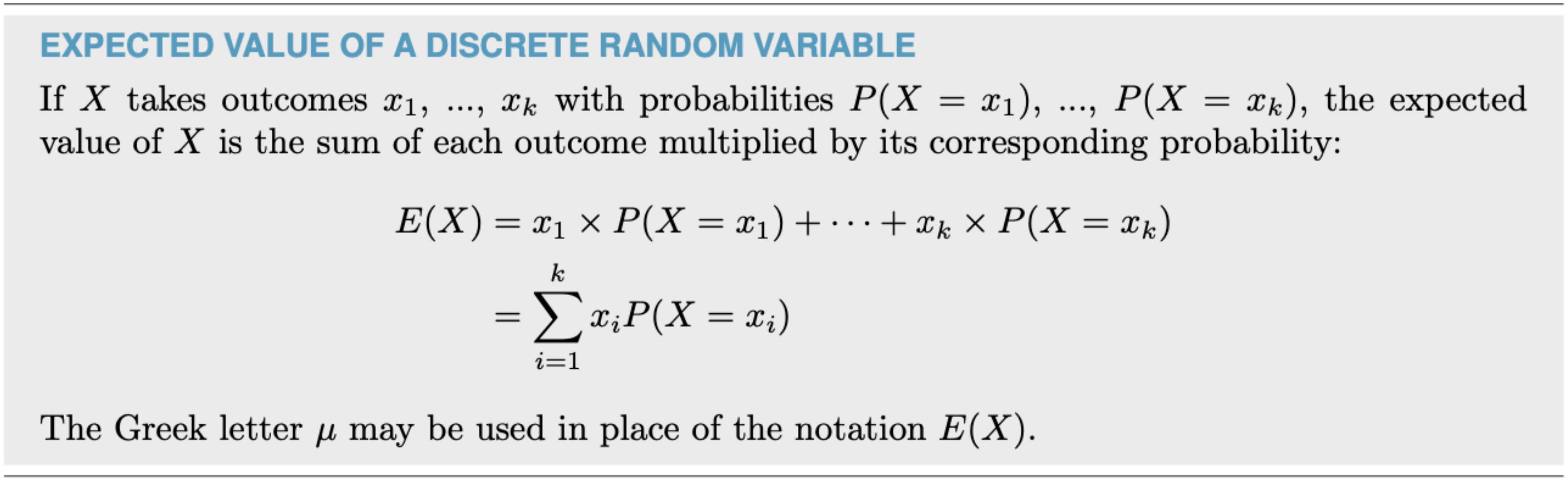
- Random variable: Random process whereby outcomes are randomly generated (with stochastic and/or systematic components); follows a **probability distribution**.

- Textbook example: P(a student, *i*, spends *x* amount of money)



i	1	2	3	Total
x_i	\$0	\$137	\$170	–
$P(X = x_i)$	0.20	0.55	0.25	1.00

Expected Value (Wtd. Avg. across outcomes)



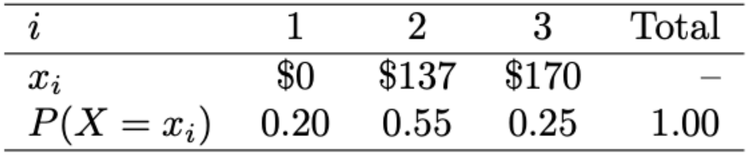
EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + \cdots + x_k \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i P(X = x_i) \end{aligned}$$

The Greek letter μ may be used in place of the notation $E(X)$.

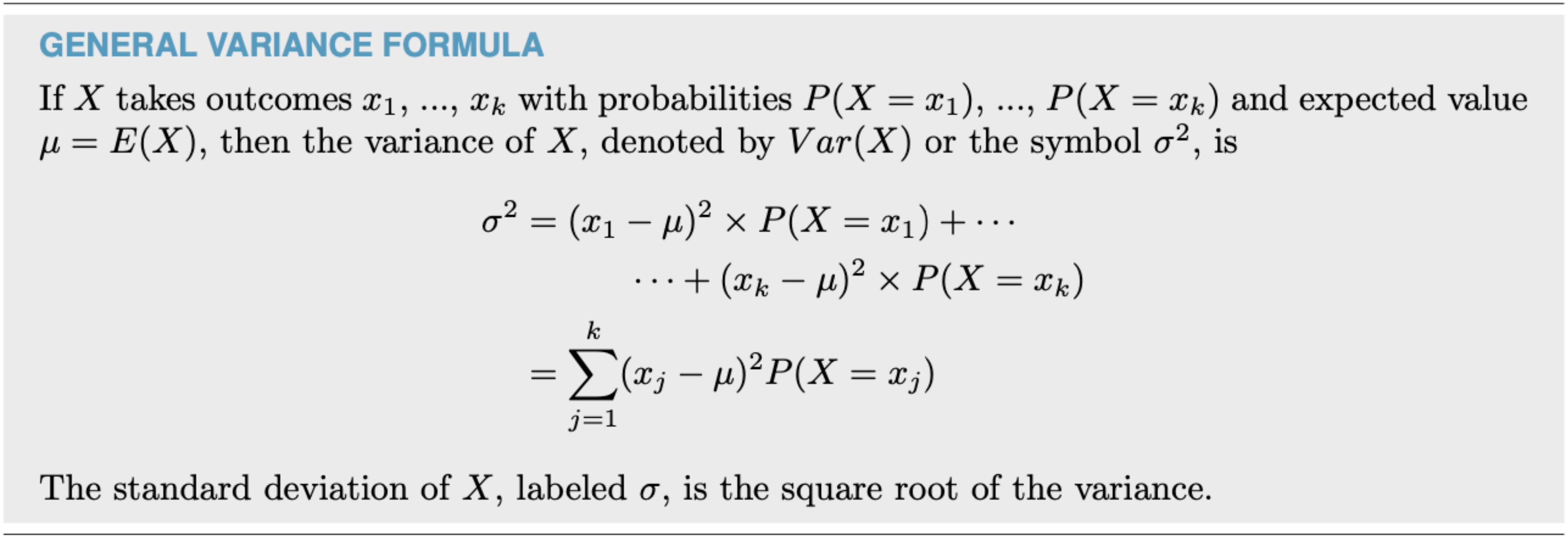
- Weighted average
- Expectation=average across possible outcomes weighted by likelihood.



i	1	2	3	Total
x_i	\$0	\$137	\$170	–
$P(X = x_i)$	0.20	0.55	0.25	1.00

Variance

- Once we get into probability distributions more broadly (the notion of repeated sampling), what does variance represent?



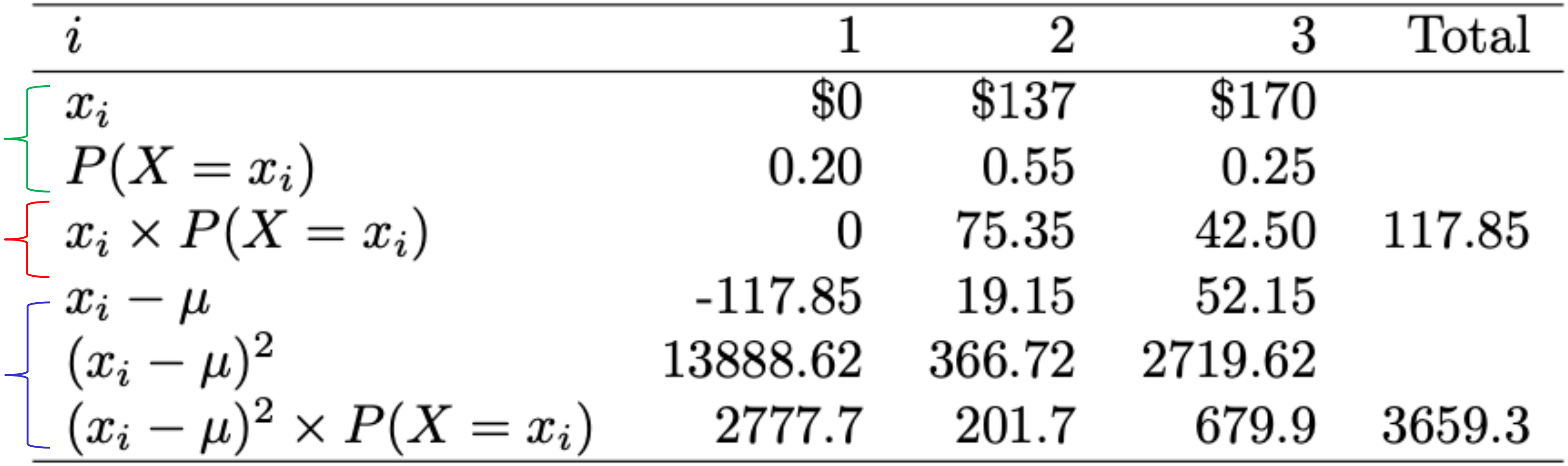
GENERAL VARIANCE FORMULA

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X, denoted by $Var(X)$ or the symbol σ^2, is

$$\begin{aligned}\sigma^2 &= (x_1 - \mu)^2 \times P(X = x_1) + \cdots \\ &\qquad \cdots + (x_k - \mu)^2 \times P(X = x_k) \\ &= \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j)\end{aligned}$$

The standard deviation of X, labeled σ, is the square root of the variance.

Variance (Build a table)



	i	1	2	3	Total
Data given	x_i	\$0	\$137	\$170	
	$P(X = x_i)$	0.20	0.55	0.25	
Exp. value	$x_i \times P(X = x_i)$	0	75.35	42.50	117.85
Variance	$x_i - \mu$	-117.85	19.15	52.15	
	$(x_i - \mu)^2$	13888.62	366.72	2719.62	
	$(x_i - \mu)^2 \times P(X = x_i)$	2777.7	201.7	679.9	3659.3

sd = ?