This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.

Copyright 2008, The Johns Hopkins University and Marie Diener-West. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided “AS IS”; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Probability Concepts

Marie Diener-West, PhD
Johns Hopkins University

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section A

Definitions and Examples

What Is Probability?

- **Probability** provides a measure of the **uncertainty** (or certainty) associated with the occurrence of events or outcomes
- Probability is useful in exploring and quantifying relationships

What Do We Mean by "Probability of 1 in 14 Million"?

- Winning the Maryland Lottery
- Heads every time in 24 tosses of a coin
- 350 people in the world with a characteristic
- 17 people in the United States with a characteristic

Useful Set Notation and Definitions

- A **set** is a collection of objects
- The objects in a set are called the **elements** of the set
- The **joint occurrence** of two sets (the **intersection**) contains the common elements
- Two sets are **mutually exclusive** if the sets have **no common elements**

Intersection of Sets

- The **intersection** of two sets, A and B, is another set which consists of the common elements of both A and B (all elements belonging to **both** A and B)
- The **intersection** is the **joint** occurrence of A and B
- Notation: A **and** B, $A \cap B$

Example of an Intersection of Events

- 40 women (**characteristic A**)
- 30 individuals aged 30 or older (**characteristic B**)
- Suppose we are told that there are 20 women aged 30 years or older
- The **intersection** of sets A and B is the set of 20 women aged 30 or older (A and B)
  - The intersection is the joint occurrence of individuals who are women and of age 30 years or older

Example of Mutually Exclusive Events

- 60 women aged 60 or older (**characteristic A**)
- 20 women aged less than 60 (**characteristic B**)
- The intersection of the two sets, A and B, is empty because a woman cannot belong to both age groups
- Sets A and B are **mutually exclusive** (e.g., the two age groups are mutually exclusive) because individuals cannot belong to both age groups at the same time

Union of Sets

- The **union** of two sets, A and B, is another set which consists of **all** elements belonging to set A or to set B (or some may belong to both sets)
- Notation: A or B, $A \cup B$

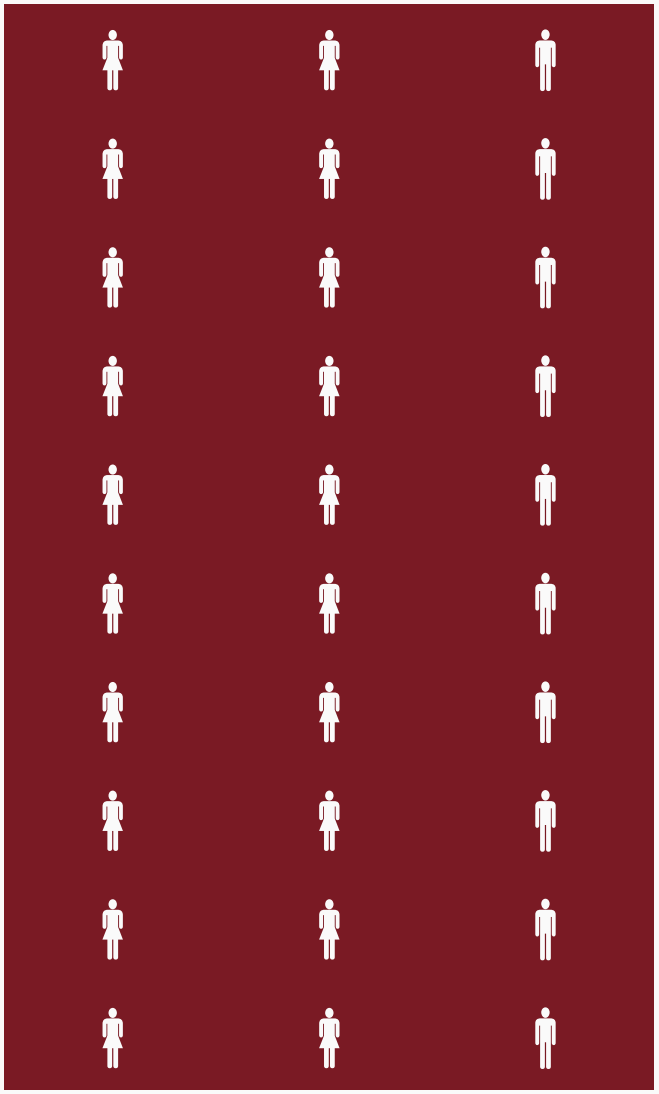
Example of a Union of Events

- 40 women (**characteristic A**)
- 30 individuals aged 30 or older (**characteristic B**)
- The **intersection** of sets A and B is the set of 20 women aged 30 or over (A and B)
- Sets A and B total to 70
  - But 20 individuals are common to both sets
- The **union** of sets A and B is the set of all individuals who are women **or** aged 30
- The **union** of the these two sets (A or B) consists of the 50 unique individuals who have one or both characteristics (A alone *or* B alone *or* both A and B)

Intersections and Unions

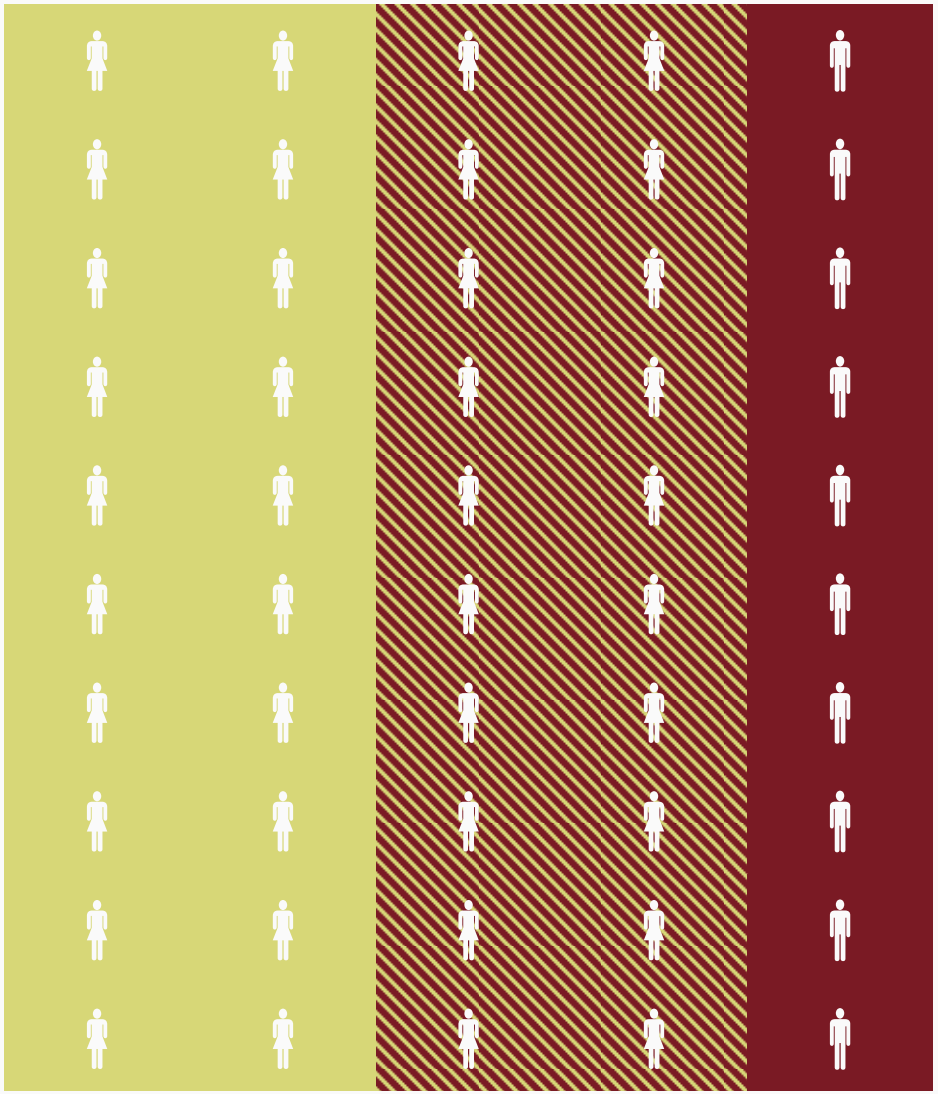
- If you have a group of 40 women (set A)

- And you have a group of 30 individuals aged 30 or over (set B)



Intersections and Unions

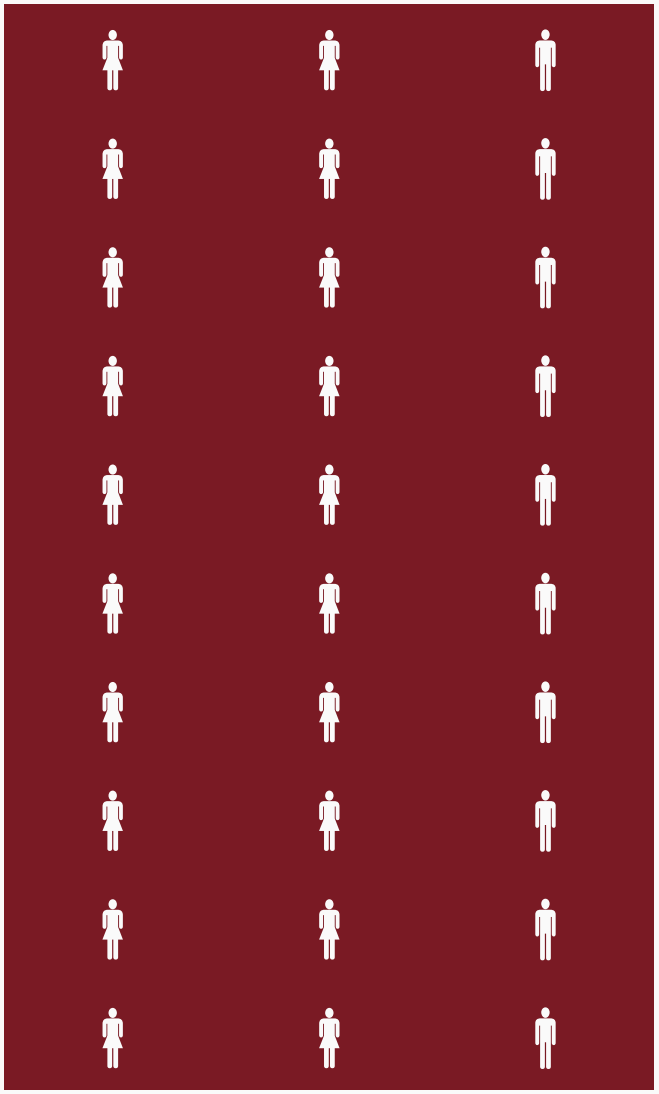
- The intersection of sets A and B are the 20 women who are aged 30 or over



Intersections and Unions

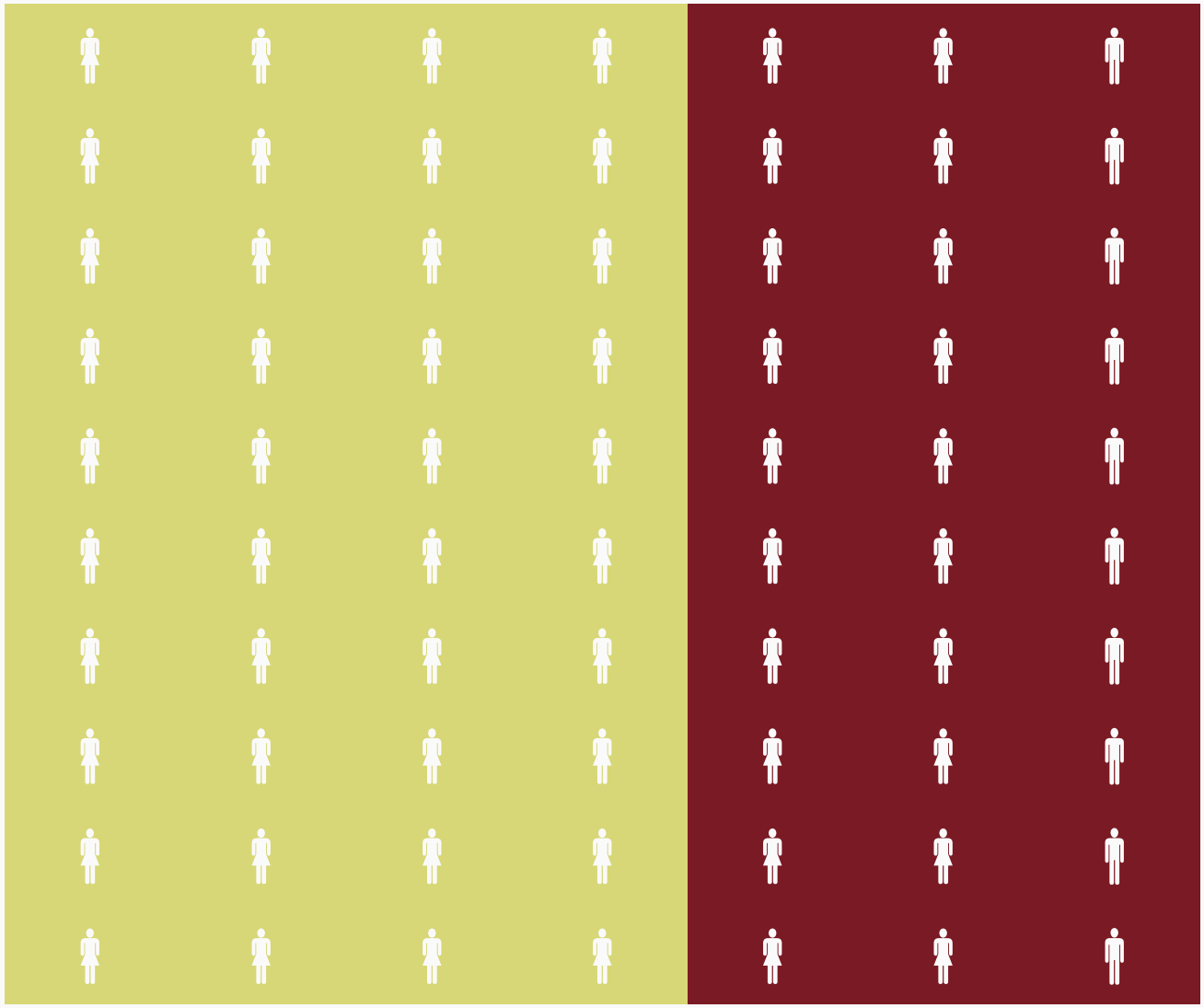
- If you have a group of 40 women (set A)

- And you have a group of 30 individuals aged 30 or over (set B)



Intersections and Unions

- The union of sets A and B is determined by adding up the total numbers of each set . . .



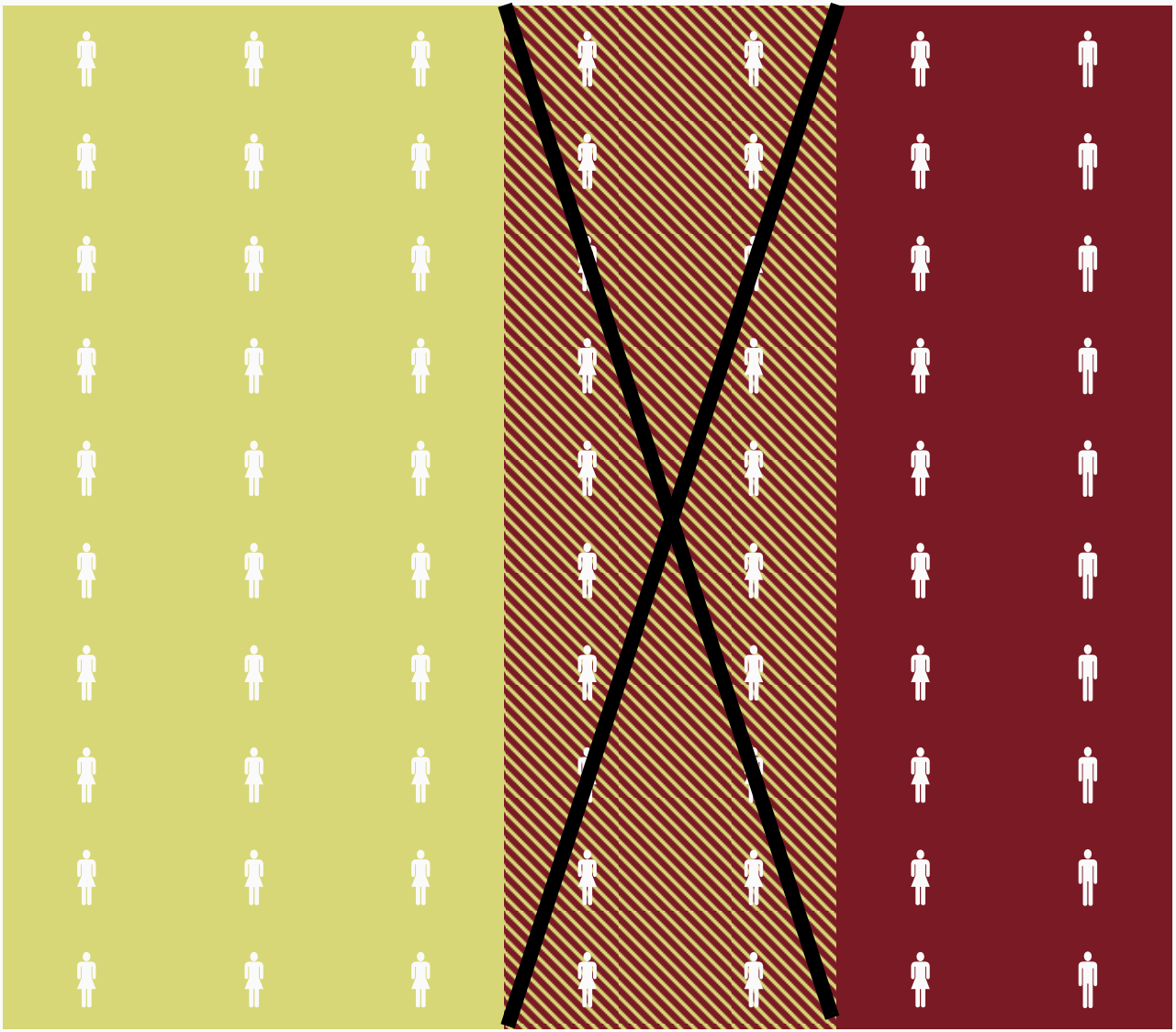
Set A = 40

+ Set B = 30

Total = 70

Intersections and Unions

- . . . And then subtracting out the individuals common to both sets



Set A = 40

+ Set B = 30

Total = 70

- Common = 20

Union: 50 unique individuals

Classical Interpretation of Probability

- An event occurs in N mutually exclusive and equally likely ways
- If m events possess a characteristic, E, then the probability of the occurrence of E is equal to m/N
- $P(E) = m/N$

Example of the Classical Interpretation of Probability

- Flipping two fair coins (equal chances of obtaining a head or a tail) results in N=4 equally likely and mutually exclusive outcomes
- $E_1 = HH, E_2 = HT, E_3 = TH, E_4 = TT$
- The probability of each outcome is 1/4 or 0.25
- $P(E_1) = P(E_2) = P(E_3) = P(E_4) = 0.25$

Relative Frequency Interpretation of Probability

- Some process or experiment is repeated for a large number of times, n
- If m events possess the characteristic, E, then the relative frequency of occurrence of E, m/n will be **approximately equal** to the probability of E

Example 1: Flip Two Coins 100 Times

Possible Outcomes	Frequency
$E_1 = HH$	22
$E_2 = HT$	28
$E_3 = TH$	23
$E_4 = TT$	27
Total	**100**

- $P(E_1)$ = P(2 heads) is 0.25 using the classical interpretation of probability
- $P(E_1)$ is approximately 22/100 = **0.22** using the relative frequency interpretation of probability

Example 2: Flip Two Coins 10,000 Times

Possible Outcomes	Frequency
E_1 = HH	2,410
E_2 = HT	2,582
E_3 = TH	2,404
E_4 = TT	2,604
Total	**10,000**

- $P(E_1)$ = P(2 heads) is 0.25 using the classical interpretation of probability
- $P(E_1)$ is approximately 2,410/10,000 = **0.24** using the relative frequency interpretation of probability

Rules of Probability with Mutually Exclusive Events

- When there are n mutually exclusive events (cannot occur together), the probability of any event is nonnegative
  - $P(\text{Event i}) \geq 0$
- The sum of the probabilities of all mutually exclusive events equals **1**
  - $P(\text{Event 1}) + P(\text{Event 2}) + \ldots + P(\text{Event n}) = 1$
- If Event i and Event j are mutually exclusive events, then the probability of either Event i or Event j is the sum of the two probabilities
  - $P(\text{Event i or Event j}) = P(\text{Event i}) + P(\text{Event j})$

Addition Rule of Probability

- **General rule**: if two events, A and B, **are not** mutually exclusive, then the probability that event A or event B occurs is:
  - P(A or B) = P(A) + P(B) – P(A and B)
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Special case**: when two events, A and B, **are** mutually exclusive, then the probability that event A or event B occurs is:
  - P(A or B) = P(A) + P(B)
  - $P(A \cup B) = P(A) + P(B)$

  since P(A and B) = 0 for mutually exclusive events

Joint Probability

- The **joint probability** of an event A and an event B is $P(A \cap B) = P(A \text{ and } B)$
- When events A and B are mutually exclusive, then $P(A \text{ and } B) = 0$

Conditional Probability

- The **conditional probability** of an event A given an event B is present is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ where } \quad P(B) \neq 0$$

Multiplication Rule of Probability

- **General rule**: the multiplication rule specifies the joint probability as:

$$P(A \cap B) = P(B)P(A \mid B)$$

- **Special case**: When events A and B are independent, then:

$$P(A \mid B) = P(A)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

Example: Relationship between Gender and Age

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

Marginal Probabilities Associated with the Example

- **Marginal probabilities** can be calculated:
  - P(Male) = 50/200 = 0.25
  - P(Female) = 150/200 = 0.75
  - P(Young) = 70/200 = 0.35
  - P(Older) = 130/200 = 0.65

Probability of Selecting an Older Female?

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- **Joint probability**
  - = P(Female and Older)= P(Female $\cap$ Older)
  - = 110/200
  - = 0.55

Probability of Selecting a Female or Older Individual?

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- By inspecting the table, we can see that the **probability of the union:**
  - P(Female or Older) = P(Female $\cup$ Older)
  - = (40+110+20)/200 = 0.85

Probability of Selecting a Female or Older Individual?

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- Using the addition rule of probability:
  - P(Female or Older)
  - = P(Female) + P(Older) – P(Female and Older)
  - = 150/200 + 130/200 – 110/200 = 170/200=0.85

Probability of Selecting a Younger Male?

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- **Joint Probability**
  - = P(Male and Younger) = 30/200 = 0.15

Probability of Selecting a Male or a Younger Individual?

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- By inspecting the table, we can see that the **probability of the union**:
  - P(Male or Younger) = (30+20+40)/200 = 0.45

Probability of Selecting a Male or a Younger Individual?

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- Using the addition rule of probability:
  - P(Male or Younger)
  - = P(Male) + P(Younger) – P(Male and Younger)
  - = 50/200 + 70/200 – 30/200 = 90/200=0.45

Are Two Characteristics, Sex and Age, Independent?

- Is sex **independent** of age in this group of patients?
- If sex and age are independent:
  - Then the probability of being in a particular age group should be the **same** for both sexes
    - In other words, the conditional probabilities should be equal
- Assess whether:
  - P(Older given Male) = P(Older given Female) = P(Older)
  - P(Older | Male) = P(Older |Female) = P(Older)

Conditional Probability for Males

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- P(Older given Male) = P(Older | Male) =20/50 = 0.40

Conditional Probability for Females

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- P(Older given Female) = P(Older | Female) =110/150= 0.73

Overall (Marginal Probability)

Patients with Disease X			
	Age		
Gender	Young	Older	Total
Male	30	20	50
Female	40	110	150
Total	70	130	200

- P(Older) = 130/200= 0.65

Comparing Conditional Probabilities

- For males:
  - $P(\text{Older} \mid \text{Male}) = 0.40$
- For females:
  - $P(\text{Older} \mid \text{Female}) = 0.73$
- For all patients:
  - $P(\text{Older}) = 0.65$
- In this group of patients, age and sex are not independent because the probability of being in the older age group depends on gender!
  - $P(\text{Older}|\text{Male}) \neq P(\text{Older} \mid \text{Female}) \neq P(\text{Older})$

Summary

- Probabilities can describe certainty associated with an event or characteristic
- Types of events:
  - Mutually exclusive events
  - Independent events
- Addition and multiple rules of probability
- Types of probabilities:
  - Marginal, joint, conditional
- Future applications of probability (e.g. screening tests)

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

Using Probability in an Epidemiology Word Problem

Probability Word Problem

- In a certain population of women:
  - 4% have had breast cancer
  - 20% are smokers
  - 3% are smokers who have had breast cancer
- From this problem, **three probability statements** may be stated
- A 2x2 table of cell counts and marginal totals can be constructed

Probability Statement 1

- 4% have had breast cancer
- *P* (breast cancer) = 0.04

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes			
No			
Total	4	96	100

Probability Statement 2

- 20% are smokers
- *P* (smoker) = 0.20

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes			20
No			80
Total	4	96	100

Probability Statement 3

- Three percent are smokers who have had breast cancer
- *P* (breast cancer and smoker) = 3/100 = 0.03

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes	3		20
No			80
Total	4	96	100

Completing the 2 x 2 Table for Word Problem

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes	3	17	20
No	1	79	80
Total	4	96	100

Marginal Probabilities

- P (breast cancer) = 0.04
- P (no breast cancer) = 0.96
- P (smoker) = 0.20
- P (non-smoker) = 0.80

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes	3	17	20
No	1	79	80
Total	4	96	100

Question: Breast Cancer or Smoking? Method 1

- What is the probability that a woman selected at random from this population has had breast cancer or smokes?
- **Method 1:** one can see from the table that:
  - P(Breast Cancer or Smoker)
  - = P(Breast Cancer $\cup$ Smoker)
  - = (3+1+17)/100 = 0.21

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes	3	17	20
No	1	79	80
Total	4	96	100

Question: Breast Cancer or Smoking? Method 2

- What is the probability that a woman selected at random from this population has had breast cancer **or** smokes?
- **Method 2:** using the addition rule:
  - P(Breast Cancer or Smoker)
  - = P(Breast Cancer) + P(Smoker) – P(Breast Cancer and Smoker)
  - = 0.04 + 0.20 – 0.03 = 0.21 = 21%

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes	3	17	20
No	1	79	80
Total	4	96	100

Answer: Breast Cancer or Smoking?

- The probability that a woman has had breast cancer or smokes is **0.21 or 21%**
- This probability may be derived from either:
  - Inspection of the 2 x 2 table
  - The probability statement

Question: Is Breast Cancer Independent of Smoking?

- Is breast cancer independent of smoking in this population?
  - In other words, is the probability of breast cancer the same for both smokers and non-smokers?
- Compare the separate (conditional) probabilities of breast cancer for the two smoking groups

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes	3	17	20
No	1	79	80
Total	4	96	100

Comparing Conditional Probabilities of Breast Cancer

- For smokers: P(Breast Cancer | Smoker) = 3/20=0.15
- For non-smokers: P(Breast Cancer | Non-Smoker) = 1/80=0.01
- For all women: P(Breast Cancer) = 4/100 = 0.04

	Women with Breast Cancer		
Smoker	Yes	No	Total
Yes	3	17	20
No	1	79	80
Total	4	96	100

Answer: Breast Cancer Is Not Independent of Smoking

- In this group of patients, breast cancer and smoking are **not independent** because the probability of having breast cancer differs by smoking group
- Semantics—it's the same to say:
  - "Breast cancer and smoking are **not independent"**
  - "Breast cancer and smoking are **dependent**" (for example, there appears to be an association
- The probability of breast cancer is increased in smokers
- P(Breast Cancer|Smoker) ≠P(Breast Cancer | Non-Smoker) ≠ P(Breast Cancer)

Other Questions?

- What is the probability of having breast cancer? (marginal probability)
- What is the probability of being a smoker for a woman without breast cancer? (conditional probability)
- What is the joint probability of not smoking and not having breast cancer? (intersection)
- What is the probability of being a smoker or not being a smoker? (union)
- What is the joint probability of being both a smoker and a non-smoker? (intersection)