Unit 4
Probabilities in Epidemiology

"All knowledge degenerates into probability"
- David Hume

So! Maybe there is no such thing as chance. Perhaps nothing is random, that there is no such thing as random, and there is no such thing as the probability of anything. Possibly, everything we encounter is actually the entirely predictable, net result, of a confluence of forces. Even the coin toss! That "heads" you see? Possibly, that "heads" landing is the net result of wind, torque, height, etc.

So why do we talk about chance? And how is the tool of chance useful to us? The answers have to do with the reality that, for the most part, we have no idea what the influences are, never mind how they interact, that result in whatever it is we encounter. Scientific inquiry seeks to discover and understand these influences.

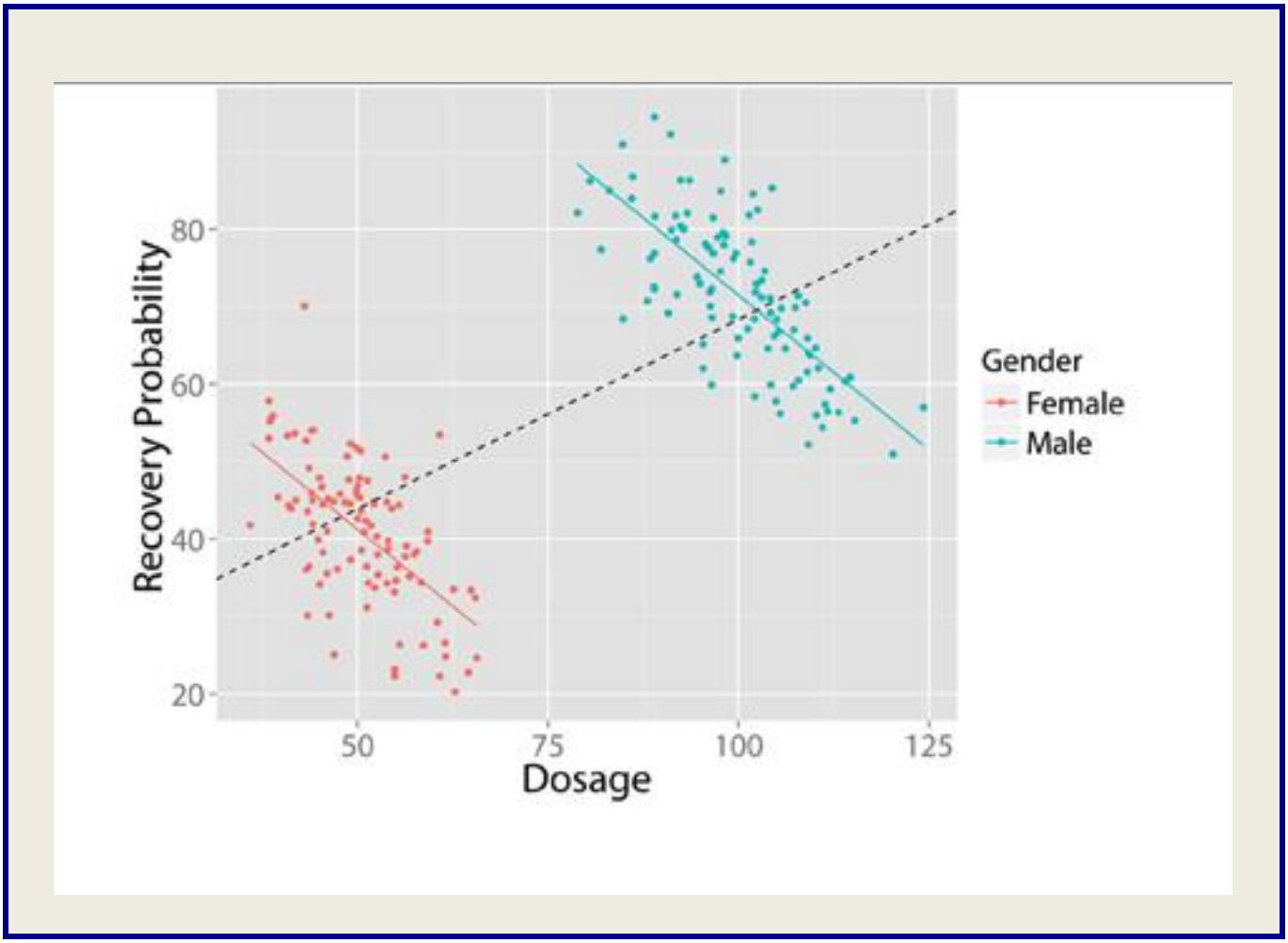
Along the way, researchers imagine (and provisionally assume is correct – at least for the time being), that some mechanism of chance (***a probability distribution which we call our model***) is what explains the variability (the "unexplainable") in what we observe. How does this imagined model comport with the reality of our observations is the question.! If the two do not comport, the imagined model is abandoned. If we get new information, the imagined model might be refined. For example, if we put our coin on the train tracks to produce a bend, we might provisionally assume that that coin is still fair. However, if after several tosses for which the percent "heads" is nowhere close to 50%, we might refine our model of "heads" for that coin. And we would then conclude "running over a coin with a train is associated with a greater (or lesser) likelihood of "heads".

And so it goes. The meaning we make of the world advances by progressing through a series of models of what we observe in nature. As new insights are acquired, we refine our models so as to accommodate their roles in producing what we have observed. Bit by bit, what's left over as residual chance (what we call the "unexplained") gets smaller and smaller.

Cheers!

Simpson's Paradox illustrated …

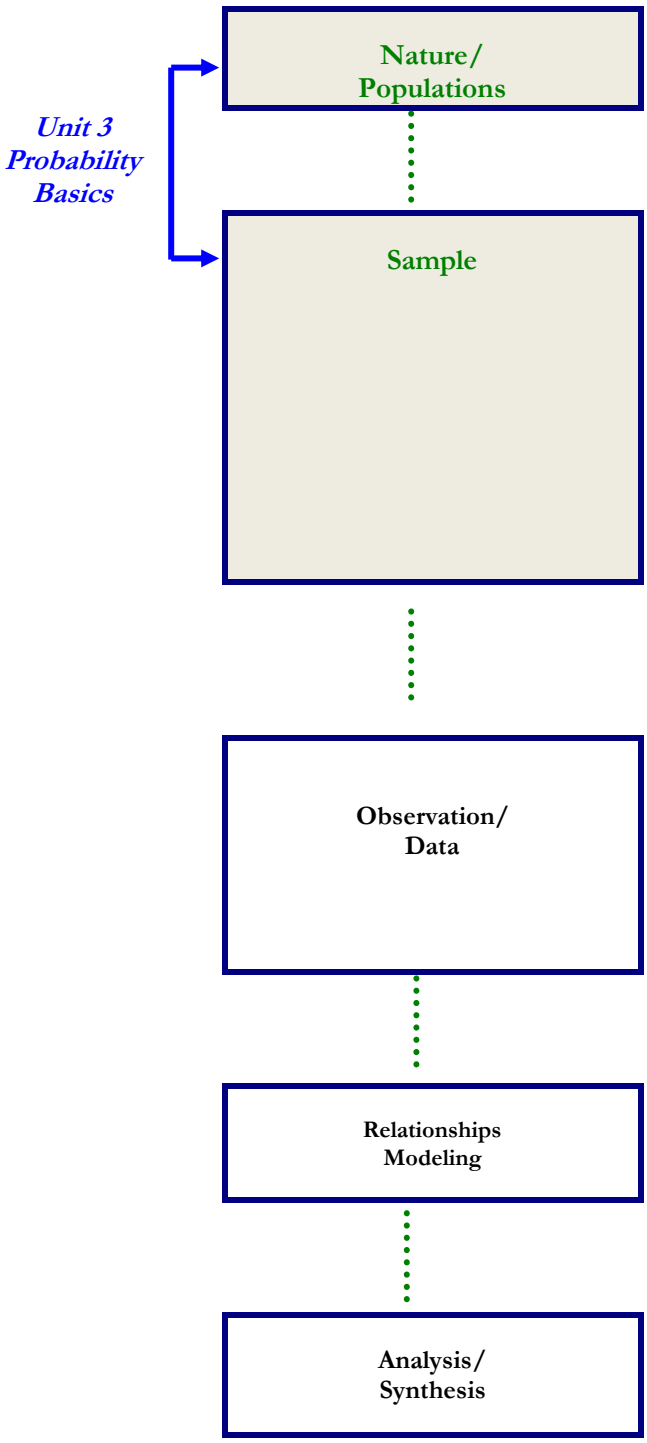
"Despite the fact that there exists a negative relationship between dosage and recovery in both males and females, when grouped together, there exists a positive relationship."



(Source: Kievit RA et al. 2013. Simpson's Paradox in Psychological Science A Practical Guide. Frontiers in Psychology. Volume 4. Article 513.

Table of Contents

1. Unit Roadmap



In Unit 1, we learned that a ***variable*** is something whose value can vary (e.g., age). A ***random variable*** is something that can take on different values ***depending on chance*** (e.g., roll of a die). In this context, we define a ***sample space*** as the collection of ***all possible outcomes*** for a random variable. **Note – It's tempting to call this collection a 'population.' We don't because we reserve that for describing nature. So we use the term "sample space" instead.**

In Unit 3, we learned that an ***event*** is one outcome or a set of outcomes. In one definition, Gerstman BB defines ***probability*** as the proportion of times an event is *expected* to occur in the population. It is a number between 0 and 1, with 0 meaning "never" and 1 meaning "always."

In Unit 4, we will learn some applications of probability basics that are useful in epidemiology.

2. Learning Objectives

When you have finished this unit, you should be able to:

- Define prevalence, incidence, sensitivity, specificity, predictive value positive, and predictive value negative.
- Calculate predictive value positive using Bayes Rule.
- Defined and explain the distinction between risk and odds. Define and explain the distinction between relative risk and odds ratio.
- Calculate and interpret an estimate of relative risk from observed data in a 2x2 table.
- Calculate and interpret an estimate of odds ratio from observed data in a 2x2 table.

3. Probabilities in Practice – Screening Tests and Bayes Rule

In your epidemiology courses, you are learning about:

- **diagnostic testing** (sensitivity, specificity, predictive value positive, predictive value negative);
- **disease occurrence** (prevalence, incidence); and
- **measures of association for describing exposure-disease relationships** (risk, odds, relative risk, odds ratio).

Most of these have their origins in notions of **conditional probability**. See again BIOSTATS 540 notes, *3. Probability Basics.*

3a. Prevalence ("existing")

https://s4be.cochrane.org/blog/2020/11/06/prevalence-vs-incidence-what-is-the-difference/

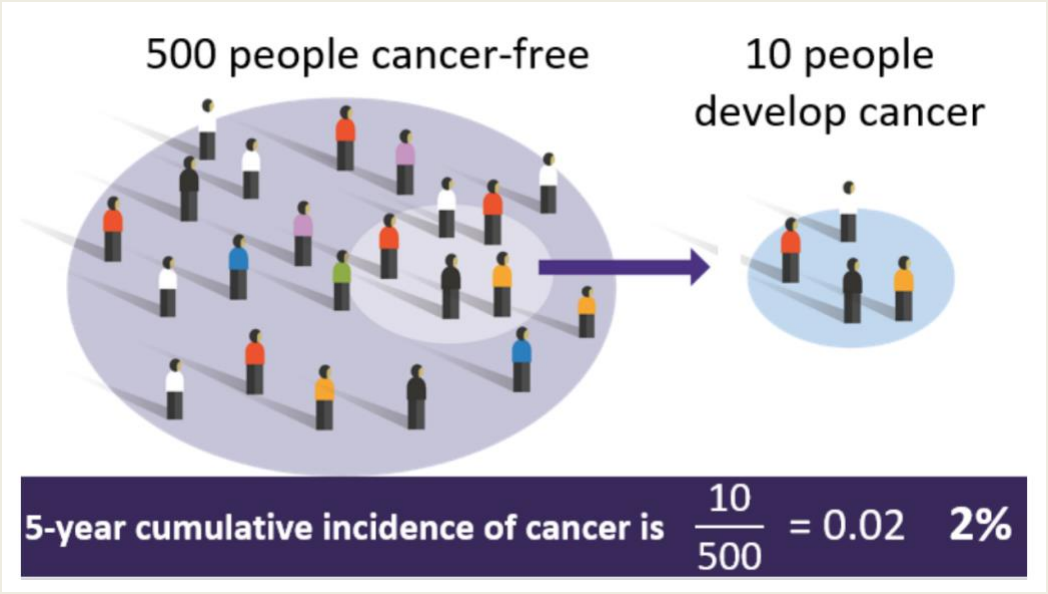
The **point prevalence of disease** is the proportion of individuals in a population that has disease at a single point in time (point), regardless of the duration of time that the individual might have had the disease. In actuality,

Prevalence is NOT a probability; it is a proportion.

Example -

A study of sex and drug behaviors among gay men is being conducted in Boston, Massachusetts. At the time of enrollment, 30% of the study cohort were sero-positive for the HIV antibody. Rephrased, the point prevalence of HIV sero-positivity was 0.30 at baseline.

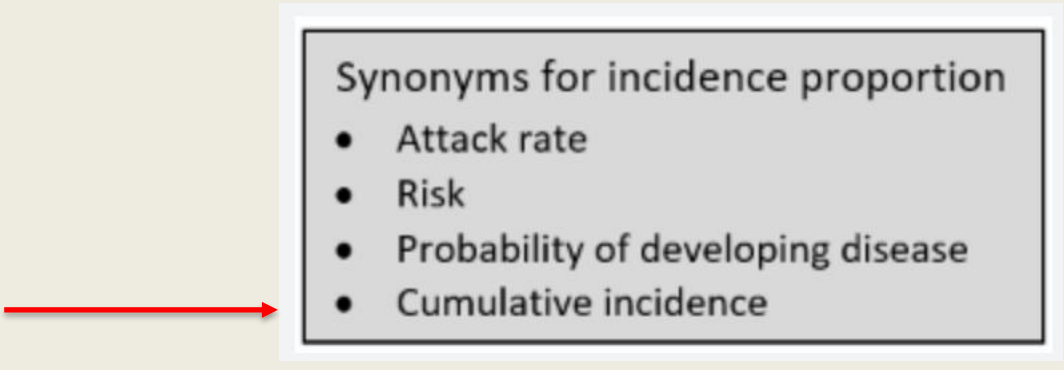
3b. Cumulative Incidence ("new")



Lecture, "Analytic Skills for Public Health 1 – Incidence and Cumlulative Incidence", UW Canvas – University of Washington

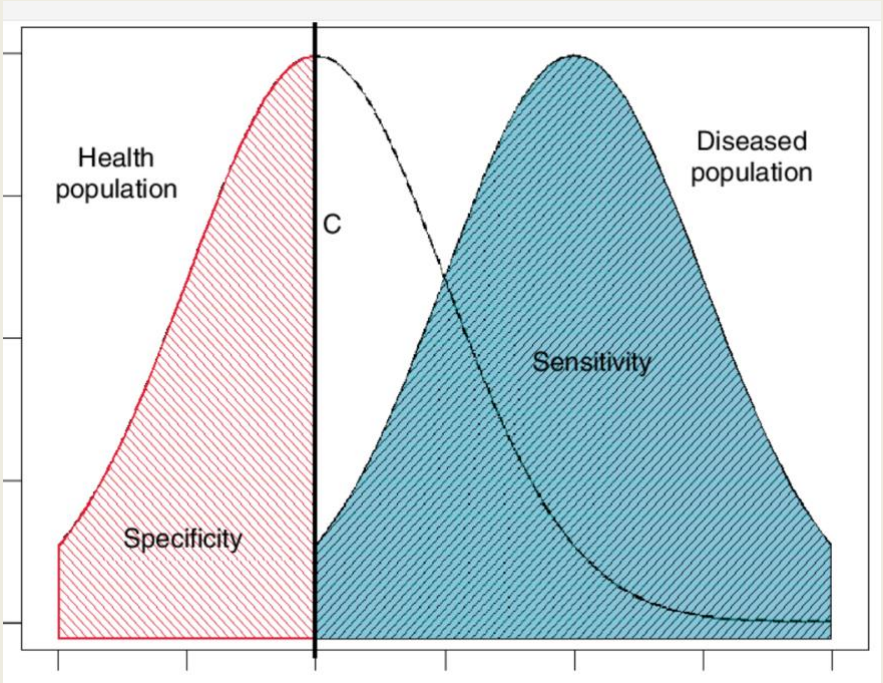
The **cumulative incidence** is a **probability**. The cumulative incidence of disease is the probability an individual who did not previously have disease will develop the disease over a specified time period.

Example 2 -
Consider again Example 1, the study of gay men and HIV sero-positivity. Suppose that, in the two years subsequent to enrollment, 8 of the 240 study subjects sero-converted. This represents a two-year cumulative incidence of 8/240 or 3.33%.



https://learning.eupati.eu/mod/book/tool/print/index.php?id=653

3c. Sensitivity, Specificity



https://www.researchgate.net/figure/Diagnostic-situation-illustrated-with-two-normal-distributions-with-variance-of-1-one_fig3_5478076

Sensitivity is about **persons with disease only**: What proportion of truly disease **correctly test positive for disease**?

Specificity is about the **disease free only**: What proportion of the truly disease free **correctly test negative for disease**?

Introduction to the 2x2 table notation for screening tests.

		Disease Status: Breast Cancer	Disease Status: Healthy	
Test Result	Positive	**800**	**9, 504**	**10,304**
	Negative	**200**	**89,496**	**89,696**
		1,000	**99,000**	**100,000**

		Disease Status: Present	Disease Status: Absent	
Test Result	Positive	a	b	a+b
	Negative	c	d	c+d
		a+c	b+d	a+b+c+d

In this table, a total of (a+b+c+d) individuals are cross-classified according to their values on two variables: disease (present or absent) and test result (positive or negative). It is assumed that a positive test result is suggestive of the presence of disease. The counts have the following meanings:

a = number of individuals who test positive AND have disease **(=800)**
b = number of individuals who test positive AND do NOT have disease **(=9,504)**
c = number of individuals who test negative AND have disease **(=200)**
d = number of individuals who test negative AND do NOT have disease **(=89,496)**

(a+b+c+d) = total number of individuals, regardless of test results or disease status **(= 100,000)**
(b + d) = total number of individuals who do NOT have disease, regardless of their test outcomes **(= 99,000)**
(a + c) = total number of individuals who DO have disease, regardless of their test outcomes **(= 1,000)**
(a + b) = total number of individuals who have a POSITIVE test result, regardless of their disease status. **(= 10,304)**
(c + d) = total number of individuals who have a NEGATIVE test result, regardless of their disease status. **(= 89,696)**

Sensitivity

Sensitivity is a conditional probability. Among those persons who are known to have disease, what are the chances that the diagnostic test will **correctly** yield a positive result?

Counting Solution: From the 2x2 table, we would estimate sensitivity using the counts "a" and "c":

representing an 80% chance

Conditional Probability Solution:

$$\text{sensitivity} = \Pr[+\text{test}|\text{disease}]$$

$$= \frac{\Pr[\text{"+test" AND "disease"}]}{\Pr[\text{disease}]}$$

$$= \frac{[a/(a+b+c+d)]}{[(a+c)/(a+b+c+d)]}$$

$$= \frac{[800/(100{,}000)]}{[(1{,}000)/(100{,}000)]}$$

$$= \frac{800}{1{,}000}$$

$$= .80$$

which matches the "counting" solution.

Specificity

Specificity is also conditional probability. Among those persons who are known to be disease free ("healthy"), what are the chances that the diagnostic test will **correctly** yield a negative result?

Counting Solution: From the 2x2 table, we would estimate sensitivity using the counts "b" and "d":

$$\text{specificity} = \frac{d}{b+d} = \frac{89,496}{99,000} = .904$$

Conditional Probability Solution:

$$\text{specificity} = \Pr[\text{-test} | \text{ NO disease}]$$

$$= \frac{\Pr[\text{"-test" AND "NO disease"}]}{\Pr[\text{NO disease}]}$$

$$= \frac{\left[d/(a+b+c+d) \right]}{\left[(b+d)/(a+b+c+d) \right]}$$

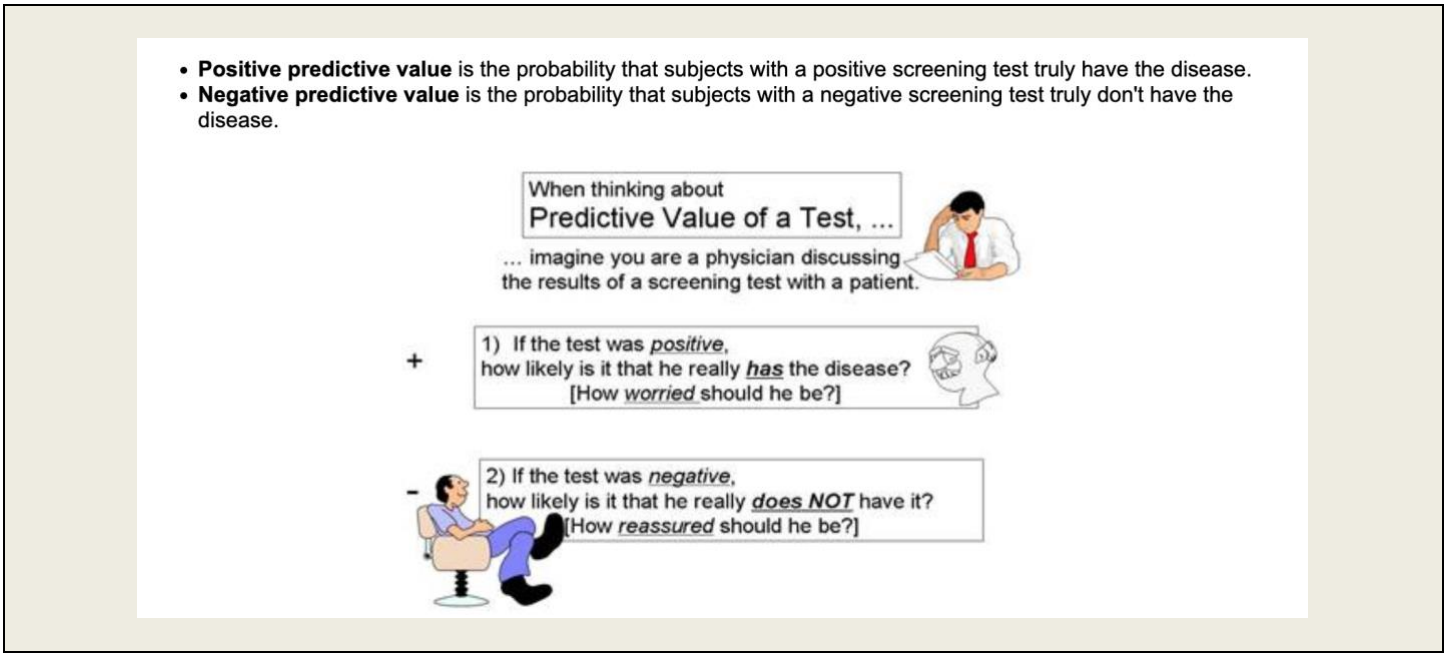
$$= \frac{\left[89,496/(100,000) \right]}{\left[(99,000)/(100,000) \right]}$$

$$= \frac{89,496}{99,000}$$

$$= .904$$

which matches the "counting" solution.

3d. Predictive Value Positive, Negative



https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_screening/ep713_screening5.html

Good to Know. ***The perspective is different.***

Perspective of sensitivity and specificity - We know disease status.
How often will the test give the right answer?
Screening test developers focus on this.

Perspective of predictive value of test – We know the test result but we do NOT know the disease status.
How often does the test result align with the actual disease status?
Clinicians focus on this

Of interest to the clinician: *"For the person who is found to test positive, what are the chances that they truly have disease?"*

- This is "predictive value positive test"

Of interest to the clinician: *"For the person who is found to test negative, what are the chances that they are truly disease free?*

"For the person who is known to test negative, what are the chances that he or she is truly disease free?".

- This is "predictive value negative test"

Predictive Value Positive Test

Predictive value positive test is a conditional probability. Among those persons who test positive for disease, what is the relative frequency of disease?

$$\text{Predictive value positive} = \frac{a}{a+b} = \frac{800}{10,304} = .078$$

Other Names for "Predictive Value Positive Test":

* posttest probability of disease given a positive test
* posterior probability of disease given a positive test

Also of interest to the clinician: Will unnecessary care be given to a person who does not have the disease?

Predictive Value Negative Test

Predictive value positive test is a conditional probability. Among those persons who test negative for disease, what is the relative frequency of disease-free?

$$\text{Predictive value negative} = \frac{d}{c+d} = \frac{89,496}{89,696} = .99$$

Other Names for "Predictive Value Negative Test":

* post-test probability of NO disease given a negative test
* posterior probability of NO disease given a negative test

HOMEWORK DUE Friday October 14, 2022
Question #1 of 4

Dear Class – This exercise asks you to draw upon what you learned in Unit 3 (Probability Basics), specifically: tree diagrams (see again p 26) and Bayes Rule (see again pp 35-37) - cb.

Enzyme immunoassay tests are used to screen blood specimens for the presence of antibodies to HIV, the virus that causes AIDS. The presence of antibodies indicates the presence of the HIV virus. The test is quite accurate but is not always correct. The following table gives the probabilities of positive and negative test results when the blood tested does and does not actually contain antibodies to HIV.

	Test Result	
	Positive (+)	Negative (-)
Antibodies present	0.9985	0.0015
Antibodies absent	0.0060	0.9940

Suppose that 1% of a large population carries antibodies to HIV in their blood.

(a) Draw a tree diagram for selecting a person from this population (outcomes: antibodies present or absent) and for testing their blood (outcomes: test positive or negative).

(b) Using your tree diagram, what is the probability that the test is positive for a randomly chosen person in this population?

(c) Again, using your tree diagram, what is the probability that a person in this population has the HIV virus, given that they test negative?

4. Probabilities in Practice – Risk, Odds, Relative Risk, Odds Ratio

Previously, we considered the 2x2 table notation for screening tests (see again, page 8).

Here we consider the 2x2 table for exploring the relationship between a yes/no (dichotomous) exposure variable and a yes/no (dichotomous) disease outcome variable.

	Outcome Present	Outcome Absent	Total
Risk Factor Present (Exposed)	a	b	a+b
Risk Factor Absent (Unexposed)	c	d	c+d
	a+c	b+d	n

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_multivariable/bs704_multivariable3.html

NOTE! Note that the rows now represent EXPOSURE status *(instead of "test result" as on p 8)*. The following is an example.

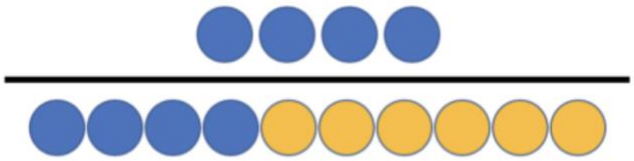
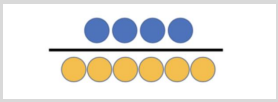
		Disease Status: Breast Cancer	Disease Status: Healthy	
Exposure Status	Exposed	**800**	**9, 504**	**10,304**
	Not exposed	**200**	**89,496**	**89,696**
		1,000	**99,000**	**100,000**

		Disease Status: Present	Disease Status: Absent	
Test Result	Exposed	a	b	a+b
	Not exposed	c	d	c+d
		a+c	b+d	a+b+c+d

a = number of exposed individuals who have disease **(=800)**
b = number of exposed individuals who do NOT have disease **(=9,504)**
c = number of NON exposed individuals who have disease **(=200)**
d = number of NON exposed individuals who do NOT have disease **(=89,496)**

(a+b+c+d) = total number of individuals, regardless of exposure or disease status **(= 100,000)**
(b + d) = total number of individuals who do NOT have disease, regardless of exposure **(= 99,000)**
(a + c) = total number of individuals who DO have disease, regardless of exposure **(= 1,000)**
(a + b) = total number of individuals who are EXPOSED regardless of their disease status **(= 10,304)**
(c + d) = total number of individuals who are NOT exposed regardless of their disease status. **(= 89,696)**

4a. Risk ("simple probability")

Risk p	
Odds $\frac{p}{1-p}$	

https://anthonybmasters.medium.com/odds-ratios-and-interpretations-f8fa4d1df577

Risk of disease, without referring to any additional information, is simply the **probability of disease**. An estimate of the probability or risk of disease is provided by the relative frequency:

$$\text{Overall risk} = \frac{(a+c)}{(a+b+c+d)} = \frac{1,000}{100,000} = .01$$

Typically, however, conditional risks are reported. For example, if it were of interest to estimate the risk of disease for persons with a positive exposure status, then attention would be restricted to the (a+b) persons positive on exposure. For these persons only, it seems reasonable to estimate the risk of disease by the relative frequency:

The straightforward calculation of the risk of disease for the persons known to have a positive exposure status is:

$$P(\text{Disease} \mid \text{Exposed}) = \frac{a}{(a+b)} = \frac{800}{10,304} = .078$$

Repeating the calculation using the definition of conditional probability yields the same answer.

$$\Pr[\text{ disease} \mid + \text{exposure}]$$

$$= \frac{\Pr[\text{"disease" AND "+ exposure"}]}{\Pr[+ \text{exposure}]}$$

$$= \frac{\left[a/\left(a+b+c+d\right)\right]}{\left[\left(a+b\right)/\left(a+b+c+d\right)\right]}$$

$$= \frac{\left[800/\left(100,000\right)\right]}{\left[\left(10,304\right)/\left(100,000\right)\right]}$$

$$= \frac{800}{10,304}$$

$$= .078$$

, ***which matches.***

HOMEWORK DUE Friday October 14, 2022
Question #2 of 4

In introductory epidemiology, one of the study designs that is introduced is the ***(prospective) cohort study***. In this type of study involving two groups, the investigator enrolls a pre-set number (set by design) of participants into each of the two groups that are generically described as "exposed" and "not exposed" and follows them forward to a designated end of the observation period, at which point one or more outcomes are measured.

The following table is from a cohort study of Danish men and women that investigated two outcomes, alcohol intake and mortality, in relationship to a number of possible influences: sex, age, body mass index, and smoking. Shown in this table is a cross-tabulation of alcohol intake and death, by sex and level of alcohol intake.

Table 8.2 The distribution of alcohol intake and deaths by sex and level of alcohol intake. Reproduced from *BMJ*, **308**, 302–6, courtesy of BMJ Publishing Group

	Men		Women	
Alcohol intake (beverages a week)*	No of subjects	No (%) of deaths	No of subjects	No (%) of deaths
<1	625	195 (31.2)	2472	394 (15.9)
1–6	1183	252 (21.3)	3079	283 (9.2)
7–13	1825	383 (21.0)	1019	96 (9.4)
14–27	1234	285 (23.1)	543	46 (8.5)
28–41	585	118 (20.2)	72	6 (8.3)
42–69	388	99 (25.5)	29	5 (17.2)
> 69	211	66 (31.3)	20	1 (5.0)
Total	6051	1398 (23.1)	7234	831 (11.5)

*One beverage contains 9–13 g alcohol.

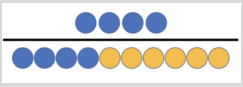
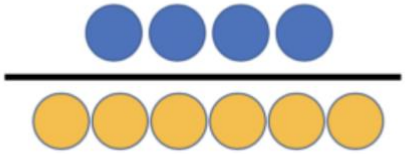
(a) From the information in the table, construct a table with 2 rows and 2 columns. Define your rows by sex and your columns by mortality. What you will have constructed is called a contingency table, and specifically, a 2x2 table.

(b) Next, construct the following contingency table, again with 2 rows and 2 columns. Define your first row to be persons who consume less than one beverage per week. Define your second row to be persons who consume more than 69 beverages per week. Define your columns by mortality.

(c) Using the information in your 2x2 table that you constructed in Question #2b, calculate the risk of death among persons who consume less than one beverage per week. Then calculate the risk of death among persons who consume more than 69 beverages per week.

(d) In 1-2 sentences, compare the two risk estimates you obtained in Question #2c.

4b. Odds ("comparison of two complementary (opposite) outcomes"):

Risk p	
Odds $\frac{p}{1-p}$	

https://anthonybmasters.medium.com/odds-ratios-and-interpretations-f8fa4d1df577

The **odds of an event "E"** compares the chances of the event occurring *(numerator)* to the chances of the same event NOT occurring, also known as the complement *(denominator)*.

$$\text{Odds} = \frac{\Pr(\text{Event occurs})}{\Pr(\text{Event does NOT occur})} = \frac{P(E)}{1 - P(E)} = \frac{P(E)}{(E^c)}$$

Example -

Perhaps the most familiar example of odds is reflected in the expression "the odds of a fair coin landing heads is 50-50". This is nothing more than:

$$\text{Odds(heads)} = \frac{P(\text{heads})}{P(\text{heads}^c)} = \frac{P(\text{heads})}{P(\text{tails})} = \frac{.50}{.50}$$

For the exposure-disease data in the 2x2 table,

$$\text{Odds(disease)} = \frac{P(\text{disease})}{P(\text{disease}^c)} = \frac{P(\text{disease})}{P(\text{NO disease})} = \frac{(a+c)/(a+b+c+d)}{(b+d)/(a+b+c+d)} = \frac{(a+c)}{(b+d)}$$

$$= \frac{1,000}{99,000}$$

$$= .0101$$

$$\text{Odds(exposed)} = \frac{\text{P(exposed)}}{\text{P(exposed}^c\text{)}} = \frac{\text{P(exposed)}}{\text{P(NOT exposed)}} = \frac{(a+b)/(a+b+c+d)}{(c+d)/(a+b+c+d)} = \frac{(a+b)}{(c+d)}$$

$$= \frac{10,304}{89,696}$$

$$= .1149$$

Conditional Odds.

What if it is suspected that exposure has something to do with disease, that there is a relationship between exposure and disease? To investigate this possibility, we might want to compare the odds of disease separately for persons who are exposed and with the odds of disease for persons who are not exposed. Now we're in the realm of conditional odds.

When we consider only those who are exposed, we are **conditioning on exposure** *(here on exposure = yes)*:

$$\text{Odds(disease | exposed)} = \frac{\Pr(\text{disease|exposed})}{\Pr(\text{NO disease|exposed})} = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$$

$$= \frac{800}{9,504}$$

$$= .084$$

When we consider only those who are NOTexposed, we are also **conditioning on exposure** *(but now on exposure = no)*:

$$\text{Odds(disease | NOT exposed)} = \frac{\Pr(\text{disease|not exposed})}{\Pr(\text{NO disease|not exposed})} = \frac{c/(c+d)}{d/(c+d)} = \frac{c}{d}$$

$$= \frac{200}{89,496}$$

$$= .002$$

Another way to investigate a possible relationship between exposure and disease wold be to compare the odds of exposure for diseased persons with the odds of exposure for NON-diseased persons:

Conditioning on disease status *(here on disease = yes)*:

$$\text{Odds(exposed | disease)} = \frac{\Pr(\text{exposed|disease})}{\Pr(\text{NOT exposed|disease})} = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$$

$$= \frac{800}{200}$$

$$= 4$$

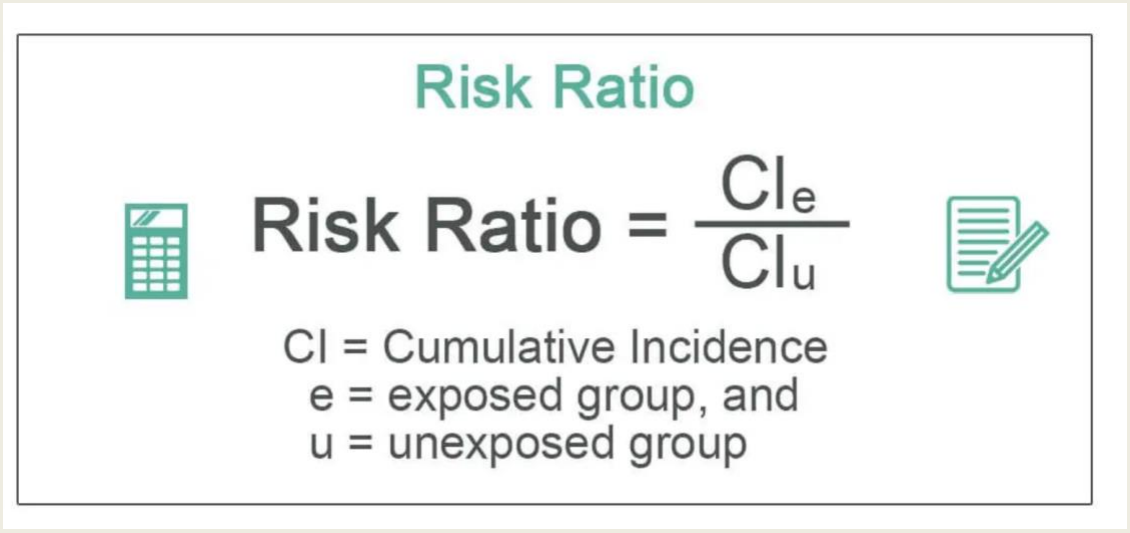
Conditioning on disease status *(here on disease = no)*:

$$\text{Odds(exposed | NO disease)} = \frac{\Pr(\text{exposed|NO disease})}{\Pr(\text{NOT exposed|NO disease})} = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

$$= \frac{9,504}{89,496}$$

$$= .1062$$

4c. Relative Risk ("comparison of two conditional probabilities")



https://www.wallstreetmojo.com/risk-ratio/

Various epidemiological studies (prevalence, cohort, case-control designs) give rise to data in the form of counts in a 2x2 table.

Recall the conventional 2x2 table structure for investigating an exposure-disease relationship.

	Disease	Healthy	
Exposed	a	b	a+b
Not exposed	c	d	c+d
	a+c	b+d	a+b+c+d

Example –
As an example, suppose we have (a+b+c+d) = 310 persons cross-classified by exposure and disease:

	Disease	Healthy	
Exposed	2	8	10
Not exposed	10	290	300
	12	298	310

Relative Risk

The relative risk is the ratio of the conditional probability of disease among the exposed to the conditional probability of disease among the non-exposed.

Relative Risk: The ratio of two conditional probabilities

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Example: In our 2x2 table, we have

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{2/10}{10/300} = 6$$

4d. Odds Ratio

$$Odds\ Ratio = \frac{Odds\ of\ an\ Event\ (Condition\ A)}{Odds\ of\ an\ Event\ (Condition\ B)} = \frac{\#\ Events\ (A) \Big/ \#\ Non\ Events\ (A)}{\#\ Events\ (B) \Big/ \#\ Non\ Events\ (B)}$$

https://statisticsbyjim.com/probability/odds-ratio/

Recall the meaning of an "odds"
Recall that if p = Probability[event] then Odds[event] = p/(1-p)

Let's look at the odds that are possible in our 2x2 table:

	Disease	Healthy	
Exposed	a	b	a+b
Not exposed	c	d	c+d
	a+c	b+d	a+b+c+d

Odds of disease among exposed $= \left[\frac{a/(a+b)}{b/(a+b)}\right] = \frac{a}{b} = \frac{2}{8} = .25$ ("cohort" study)

Odds of disease among non exposed $= \left[\frac{c/(c+d)}{d/(c+d)}\right] = \frac{c}{d} = \frac{10}{290} = .0345$ ("cohort")

Odds of exposure among diseased $= \left[\frac{a/(a+c)}{c/(a+c)}\right] = \frac{a}{c} = \frac{2}{10} = .20$ ("case-control")

Odds of exposure among healthy $= \left[\frac{b/(b+d)}{d/(b+d)}\right] = \frac{b}{d} = \frac{8}{290} = .0276$ ("case-control")

Students of epidemiology learn the following great result!

Odds ratio, OR

In a cohort study:

$$OR = \frac{\text{Odds disease among exposed}}{\text{Odds disease among non-exposed}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

In a case-control study:

$$OR = \frac{\text{Odds exposure among diseased}}{\text{Odds exposure among healthy}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Terrific!

The OR is the **same**, regardless of the study design, cohort (prospective) or case-control (retrospective)

Note: Come back to this later if this is too "epidemiological"!

Example: In our 2x2 table, we have

$$OR = \frac{a * d}{b * c} = \frac{(2)(290)}{(8)(10)} = 7.25$$

Thus, there are advantages of the Odds Ratio, OR.

1. Many exposure disease relationships are described better using ratio measures of association rather than difference measures of association.

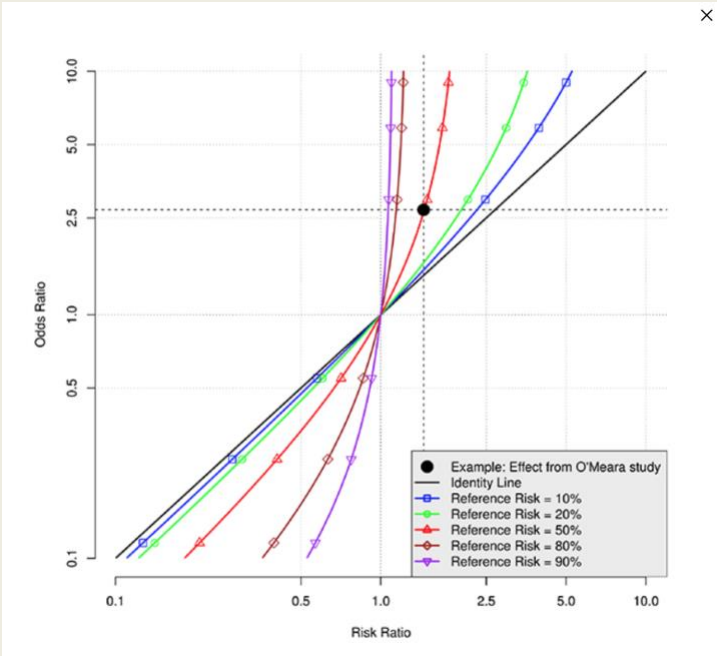
2. $OR_{\text{cohort study}} = OR_{\text{case-control study}}$

3. The OR is the appropriate measure of association in a case-control study.

   - Note that it is <u>not</u> possible to estimate an incidence of disease in a retrospective study. This is because we select our study persons based on their disease status.

4. When the disease is rare, $OR_{\text{case-control}} \approx RR$

Good to Know. The Odds Ratio Always "Overstates"



Exposure is "bad" (greater risk of disease)
IF the relative risk is **greater than 1,** THEN the odds ratio is even ***more*** **greater than 1**

Exposure is "good" (less risk of disease)
IF the relative risk is **less than 1,** THEN the odds ratio is even ***more*** **less than 1**

https://annalsofglobalhealth.org/articles/10.5334/aogh.2581/

HOMEWORK DUE Friday October 14, 2022
Question #3 of 4

Another study design that is introduced in introductory epidemiology is the (retrospective) case-control study. This study design also calls for the comparison of two groups. Here, however, the investigator enrolls set (again, set by design) numbers of participants, defined by their disease status at the start of the study. "Cases" are the enrollees with disease. "Controls" are the enrollees who do not have the disease under investigation. The investigation involves looking back in time ("retrospective review") at the histories of all study participants. The goal of this "back in time" look is to see if the cases are different from the controls with respect to their history of some exposure of interest.

The table below is from a case-control study that investigated the relationship of occurrences of Down Syndrome (cases) to history of exposure to maternal smoking during pregnancy. Shown in the table are some characteristics of the mothers, together with their status with respect to their history of smoking during pregnancy.

Table 8.3 Basic characteristics of mothers in a case-control study of maternal smoking and Down syndrome. Reproduced from *Amer. J. Epid.*, **149**, 442–6, courtesy of Oxford University Press

Selected characteristics of Down syndrome cases and birth-matched controls. Washington State, 1984–1994

	Cases (n = 775)		Controls (n = 7750)	
	No.	%	No.	%
Smoking during pregnancy				
Age < 35 years				
Yes	112	20.0	1411	20.2
No	421	75.0	5214	74.6
Unknown	28	5.0	363	5.2
Aged ≥ 35 years				
Yes	15	7.0	108	14.2
No	186	86.9	611	80.2
Unknown	13	6.1	43	5.6

(a) Using the information in the table, construct separate 2x2 contingency tables, one for mothers aged < 35 years and the other for mothers aged ≥ 35 years. Define rows by exposure (smoked during pregnancy versus not). Define columns by case status (cases versus controls).

(b) For each of the 2x2 tables you constructed in Question #3a, calculate two odds: (i) Odds of smoking during pregnancy among cases; and (ii) Odds of smoking during pregnancy among controls

(c) Using the calculations of odds that you obtained in Question #3b, calculate two odds ratios:
(i) Odds Ratio for history of maternal smoking among mothers age < 35; and (ii) Odds Ratio for history of maternal smoking among mothers age ≥ 35

(d) In 1-2 sentences, interpret your results in Question #3c.

HOMEWORK DUE Friday October 14, 2022
Question #4 of 4

This question is intended to re-enforce your appreciation of the distinction between the two study designs: prospective cohort versus case-control.

In 1-2 sentences, why can't you calculate risk in a case-control study?

Appendix
Computing Tools for Probabilities in Epidemiology

Bayes Rule

- **StatTrek**
  https://stattrek.com/online-calculator/bayes-rule-calculator

- **Social Science Statistics**
  https://www.socscistatistics.com/bayes/default.aspx

Diagnostic/Screening

- **VassarStats**
  1) At left, click **Clinical Research Calculators**; then
  2) At top, click **Calculator 1**
  http://vassarstats.net

- **MedCalc**
  https://www.medcalc.org/calc/diagnostic_test.php

Relative Risk and Odds Ratio

- **VassarStats**
  1) At left, click **Clinical Research Calculators**; then
  2) At top, click **Calculator 3**
  http://vassarstats.net

- **Social Science Statistics**
  https://www.socscistatistics.com/biostatistics/default2.aspx

2 x 2 Epidemiologic Table

- **VassarStats**
  1) At left, click **Clinical Research Calculators**; then
  2) At top, click **Calculator 3**
  http://vassarstats.net

- **EpiTools**
  https://epitools.ausvet.com.au/twobytwotable