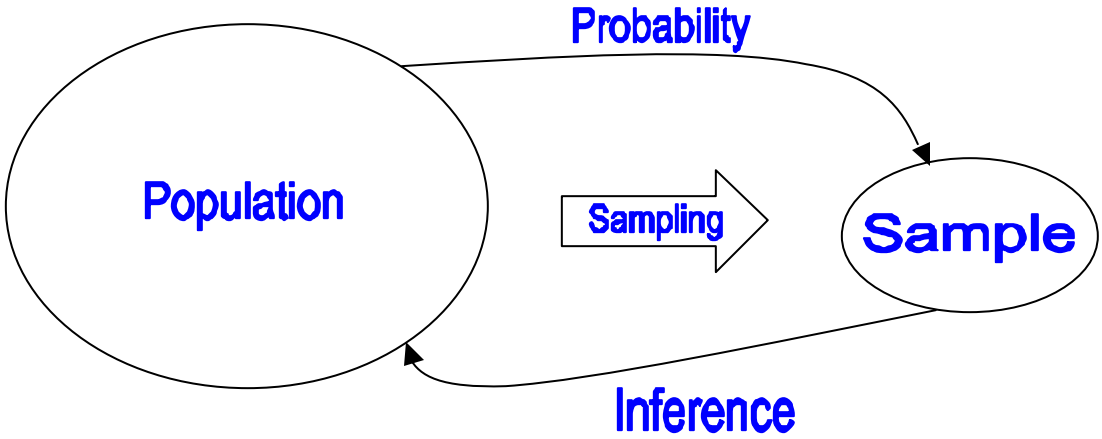
6. Probability

We have looked at methods for summarizing our data. We would like to use our data to say something about the population from which it was obtained. That is, we would like to infer characteristics of the population by using our sample data. First, we need some tools from the field of probability.



Roughly, probability asks questions like: Given a population with known characteristics, what can I say about subjects selected from the population? Statistics asks: given that I have selected a subset of subjects from the population (the sample), what can I say about the population?

We'll discuss probability in a somewhat "loose" fashion. It is beyond the scope of this class to pursue the technical details of probability. Our goal is to establish some basic notions and fundamental operations of probability that will help us later.

6.1. Operations on Events and Probability

First, some terminology:

- We will consider an experiment as the process by which we observe a data point from objects (subject, experimental unit). The result of one application of the experiment will be called an outcome.

E.g. In a study to test the effectiveness of a drug for treating headaches, you administer the drug or placebo (e.g. Sugar pill) to a patient (subject, experimental unit, unit) and, after some specified period of time you ask the patient if they experienced headache relief. Say you simply record the patient's response (the outcome) as "Yes" or "No".

- The sample space (denoted S) is the set of all possible outcomes of an experiment/measurement/observation.

E.g. Ask a patient if they smoke: S = {all possible outcomes of this experiment} = {smoke, not smoke}.

E.g. Measure the amount of zinc (µg/dl) in a patient's bloodstream S=[0,∞)

E.g. Toss a coin and observe the outcome: S={H, T}. Toss a coin twice: S={HH, TT, HT, TH}

- Event-An event is any possible collection of outcomes of an experiment; that is, any subset of the sample space S. We may use a verbal description for an event and will use

capital letters to denote events. We say an event "has occurred" if the outcome of an experiment is contained in the event.

E.g. A={observe at least 1 tail when tossing a coin twice}={HT, TT, TH}

E.g. Test for breast cancer in women patients from ages 35 to 40 years. Event A={a woman tests positive for breast cancer}. Event B={a woman does not test positive for breast cancer}. Note that B is simply the complement of A, or B is "not A".

Three basic operations on events (set operations):

- Complement of an event A is denoted A^C (read "A complement" or "not A"); or sometimes $\overline{A}$ (read "A-bar").
- The union of event A with event B is denoted $A \cup B$. Includes outcomes in A or in B or in both. Note $A \cup A^C = S$ (always for any event A).
- The intersection of two events A and B is denoted $A \cap B$. Includes only outcomes in both A and B

E.g. Tossing a die: A={observe 3 dots or greater}={3,4,5,6} and B={observe an even number of dots} = {2,4,6}. $A \cup B$ = $\{3,4,5,6\} \cup \{2,4,6\} = \{2,3,4,5,6\}$.

E.g. Die toss continued: $A \cap B = \{3,4,5,6\} \cap \{2,4,6\} = \{4,6\}$

- The null event or impossible event is denoted $\varnothing$. Doesn't include any outcomes--it's empty. Note that $A \cap A^C = \varnothing$ for any event A.

- Subset. An event A is a subset of event B, denoted $A \subseteq B$, if all outcomes that are in A are also in B. Note that $\varnothing \subseteq A$ by definition for any event A. You can also read $\subseteq$ as "is contained in.")

E.g. Breast cancer continued. Also, ask each patient if they smoke. Let A={woman tests positive for breast cancer} and let B={woman smokes}. Then $A \cap B$ is {woman smokes and tests positive for breast cancer} and is depicted using Venn diagrams as

- Frequentist definition of probability: If an experiment is repeated n times under identical conditions (i.e., on the same population of subjects or units), and if the event A occurs m times, then as n grows large, the ratio m/n approaches a fixed limit that is the probability of A.
  - P(A) = m/n
  - Interpretation: the relative frequency (proportion) of occurrence of an event in a large number of experiments or observations.
  - Intuitively, we see that probabilities should be between 0 and 1 (inclusive).
  - We can think of probability as the chance of seeing an event occur ("subjective" probability interpretation).

E.g. Tossing a fair coin.
http://www2.SPSU.edu/math/deng/m2260/stat/coin/TossCoin.html

Tossing a die.
http://www2.SPSU.edu/math/deng/m2260/stat/roll/RollDice.html

- Some basic rules of probability:
  - $0 \le P(A) \le 1$ for any event A
  - P(all events) = P(union of all possible events) = P(S) = 1
  - P(no events) = $P(\varnothing)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - If $(A \cap B) = \varnothing$, then $P(A \cap B) = P(\varnothing) = 0$, and we say A and B are disjoint or mutually exclusive events.
  - In the case of mutually exclusive events (or, more specifically, if $P(A \cap B) = 0$), we have the additive rule of probability:
    - $P(A \cup B) = P(A) + P(B)$
  - $P(A \cup A^C) = P(A) + P(A^C)$

Think about the above rules using Venn diagrams:
(See applet:
http://www2.SPSU.edu/math/deng/m2260/stat/venn2/prob.html)

6.2. Conditional Probability

E.g. Breast cancer continued: Say we know the event B={woman smokes} has occurred. Given that a woman smokes (i.e. given event B), what's the probability of the event A={woman tests positive for breast cancer}?
Use a Venn diagram to think about this.

- The conditional probability of an event A given an event B has occurred is denoted/defined as $P(A|B)=P(A\cap B)/P(B)$ (assuming P(B) is not zero!)
- The multiplicative rule of probability says $P(A\cap B)=P(A)P(B|A)=P(B)P(A|B)$.
- Note that we can get the conditional probability using the multiplicative rule (as long as the denominator is not zero!).

E.g. Assuming the life table (Table 5.1) on page 98 of POB describes the entire US population in 1992, we see that 92562 out of 100000 people live to be at least 50 years of age. Let A={a person lives to at least 50 years}. Let B={a person lives to at least 70 years}. Then, according to the frequentist definition of probability we have P(A) = 92562/100000 = 0.92562 and P(B) = 72063/100000 = 0.72063.

What's the probability of $A \cap B$?

Given that a person lived to at least 50 years, what's the probability that the person will live to at least 70?

Check out this simple applet demonstrating conditional probability:
http://www2.SPSU.edu/math/deng/m2260/stat/cond/cond.html

- When one event has no effect on the probability of another event (and vice versa), we say the events are independent. We define independence as $P(A \cap B) = P(A)P(B)$. This means $P(A|B) = P(A)$ (and $P(B|A) = P(B)$). That is, the probability of A is the same whether we know B or not, and vice versa. In other words, the fact that an event A does or does not occur has no bearing on the chance of seeing event B occur.

E.g. Toss coin.

Some questions:

If events are mutually exclusive are they necessarily independent?

If events are independent, are they necessarily mutually exclusive?

E.g. Natality: Mothers Age distribution for Births in 1992

Age	Probability
<15	0.003
15-19	0.124
20-24	0.263
25-29	0.290
30-34	0.220
35-39	0.085
40-44	0.014
45-50	0.001

What's the probability that a randomly selected woman who gave birth in 1992 was age 19 or younger?

What's the probability that she was 45 or older? 50 or older?

Given that a mother was 34 years or younger, what's the probability of her having been a teenager (or younger!)?

If you randomly select two mothers, what's the probability that they were both 20-24 years of age?

6.3. Bayes' Theorem—discussed along with 6.4
6.4. Diagnostic testing

If a person has a disease, we would like to detect the disease early in order to improve the chance of successful treatment. Often, a relatively inexpensive or simple procedure is used so as to be available to screen a large group of people. Such procedures are called diagnostic tests. Those who test positive for a disease would continue to undergo further, more intensive/expensive testing and subsequent treatment if necessary.

6.4.1. Sensitivity and Specificity

Two measures of the performance of a diagnostic test are sensitivity and specificity. These are illustrated in the following table.

		Test Result	
		T^+	T^-
Disease Status	D^+ $P(D^+)$= prevalence	$T^+\|D^+$ no error $P(T^+\|D^+)$= sensitivity	$T^-\|D^+$ false neg $P(T^-\|D^+)$= 1- $P(T^+\|D^+)$
	D^- $P(D^-)$= $1-P(D^+)$	$T^+\|D^-$ false pos $P(T^+\|D^-)$= $1-P(T^-\|D^-)$	$T^-\|D^-$ no error $P(T^-\|D^-)$= specificity

Note that in order to determine sensitivity and specificity, we would administer the test on several subjects and then use a more accurate (expensive) procedure to determine if each subject actually has the disease.

E.g. Mammogram test for beast cancer.

- suppose health records indicate 250 out of 100000 women have breast cancer
- assume 100000 is large enough to apply frequentist definition of probability:
  - prevalence=$P(D^+)$=250/100000= 0.0025
- A number of women are administered a mammogram and test results recorded (T^+ or T^-). These same women are examined further (with more intensive/expensive

procedures) to determine if they really do have breast cancer or not (D^+ or D^-) and the results indicate:
  - sensitivity=$P(T^+|D^+)=0.85$
  - specificity=$P(T^-|D^-)=0.80$
  - P(false negative)= $P(T^-|D^+)=1-P(T^+|D^+)=0.15$
  - P(false positive)= $P(T^+|D^-)=1-P(T^-|D^-)=0.20$
- In general, we want a test that has small probabilities of false negatives and false positives, or, equivalently, high sensitivity and high specificity, but there is typically a trade-off between these.

Of course, in practice, we normally do not have the results of the more intensive/expensive procedure to tell us whether a woman has breast cancer or not. That's the purpose of the diagnostic test, which is less intensive/expensive and can be applied to a larger number of women as a screening procedure.

In practice, we are not given D^+ or D^-. But, after administering the diagnostic test we do have T^+ or T^-. Thus, we "turn things around" and ask, "Given the test result, what's the probability that the patient has the disease or not?"

- Predictive value of a positive test = $P(D^+|T^+)$
- Predictive value of a negative test = $P(D^-|T^-)$

E.g. Breast cancer continued. We want to calculate the predictive value of a mammogram test using what we know about the test (sensitivity ($P(T^+|D^+)$) and specificity ($P(T^-|D^-)$)) and using what we know about the disease (prevalence $P(D^+)$):

- $PV^+ = P(D^+|T^+) = P(D^+ \cap T^+)/P(T^+)$
- $P(D^+ \cap T^+) = P(D+)P(T+|D+)$
- $P(T^+) = P(T^+ \cap (D^+ \cup D^-)) = P(T^+ \cap D^+) + P(T^+ \cap D^-) = P(D^+)P(T^+|D^+) + P(D^-)P(T^+|D^-)$
- $PV^+ = P(D+)P(T+|D+) / (P(D^+)P(T^+|D^+) + P(D^-)P(T^+|D^-)) = 0.0025*0.85/(0.0025*0.85 + 0.9975*0.20)=0.0105$

Note that we have "updated" the prior probability P(D+)=0.0025 using the test result, arriving at the posterior probability P(D+|T+)=0.0105. While 9895 out of 10000 people who test positive do not have the disease, the test result has increased the chance of properly diagnosing breast cancer (0.0105/0.0025=4.2).

E.g. Pulmonary Disease

- 60 yr old patient never smokes, complains of chronic cough and breathlessness
- Physician orders lung biopsy; results are consistent with either lung cancer or sarcoidosis (a fairly common, nonfatal lung disease)
- Let C={symptoms: chronic cough, positive lung biopsy result}
- Disease Status (mutually exclusive, exhaustive):
  - D1 = {normal lungs}
  - D2 = {lung cancer}
  - D3 = {sarcoid}
- Clinical experience, :
  - P(C|D1) = 0.001
  - P(C|D2) = 0.9
  - P(C|D3) = 0.9

- From age-smoking-disease prevalence rates:
  - P(D1) = 0.99
  - P(D2) = 0.001
  - P(D3) = 0.009
- We want to update these last 3 prior probabilities using Bayes rule. For example, determine P(D3|C):
  - P(D3|C)=P(D3)P(C|D3)/(P(D1)P(C|D1) + P(D2)P(C|D2)) + P(D3)P(C|D3))=0.009*0.9/(0.99*0.001 + 0.001*0.9 + 0.009*0.9)=0.0081/0.00999=0.811
  - P(D1|C)=0.099
  - P(D2|C)=0.090
- Conclusions:
  - The unconditional (prior) probability of sarcoid, P(D3)=0.009 is low (for 60 yr old nonsmoking man), but the conditional (posterior) probability is high, P(D3|C)=0.811
  - Sarcoid is much more likely than lung cancer for someone with these symptoms (in this age/sex/smoke group) even though the probability of these symptoms is the same given each disease (P(C|D2)=P(C|D3)=0.9)

6.4.3 ROC (Receiver Operator Characteristic) Curves

- In some cases, a test does not indicate one of 2 possible values (e.g. positive, negative), but will indicate one of several possible results. Sometimes the result is on a continuous scale (e.g. blood chemical measurement).
- We need to decide which category or level of the test result will be our cut-off between declaring a positive result as opposed to a negative result

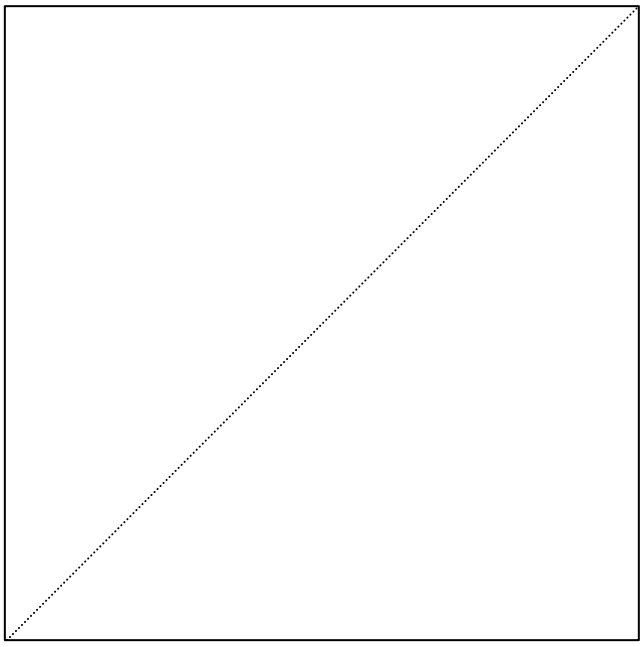
- ROC curves help us to evaluate the usefulness of a test and to select an appropriate cut-off value
- Plot sensitivity verses (1-Specificity)

E.g. Computed tomographic images (CT scans). A radiologist rated 109 CT scans for subject with known disease status to get the following table

	CT Rating					
True Disease	Definitely Normal 1	Probably Normal 2	Not sure 3	Probably Abnormal 4	Definitely Abnormal 5	Total
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

Cut-Off	Sensitivity	Specificity	False Positive
>=1 (always declare abnormal)	51/51= 1	1-58/58=0	58/58=1
>=2	46/51=0.94	1-25/58=0.57	25/58=0.43
>=3	0.90	0.67	0.33
>=4	0.86	0.78	0.22
>=5	0.65	0.97	0.03
never declare abnormal	0.00	1	0.00

ROC Curve Plot:



- Points near upper left corner are better
- Obviously for a test to be useful we should have $P(T^+|D^+) > P(T^+|D^-)$. That is, the test should indicate positive results with higher probability for those with the disease than for those without the disease! Hence, we want the ROC curve to be above the diagonal line.
- In above e.g., we are assuming that the number of CT scans studied is large enough to satisfy the frequentist definition of probability

6.5 Relative Risk and the Odds Ratio

We use relative risk (RR) to compare probabilities of disease given 2 or more different situations.

- RR = $P(D^+|\text{exposed})/P(D^+|\text{unexposed})$

- More generally, RR = $P(D^+|\text{situation A})/P(D^+|\text{situation B})$
- In words, RR is the chance (i.e., probability, risk) of having a disease for a member of one group relative to chance of having a disease for a member of another group
- "Risk" = probability
- Often choose the "baseline" group (in denominator) so that RR > 1

E.g. Socio-economic status (SES) and persistent respiratory symptoms (PRS) in children

SES	# Children	# Children with Symptoms	P(Symp|SES)
Low	79	31	0.39
Middle	122	29	0.24
High	192	27	0.14

- RR of PRS for children of low SES compared to those of high SES, RR=0.39/0.14 = 2.79.
- Thus, children of low SES are 2.79 times more likely to develop PRS than those of high SES
- Increased risk = (Change in Risk)/(Baseline Risk)*100% = (RR –1)*100% = (0.39 – 0.14)/0.14*100% = 179%.
- Thus, there is an 179% increase in the chance of having PRS for children of low SES compared to high.
- Note that we are assuming the number of children observed is large enough to satisfy the frequentist definition of probability

Remarks about Risk/RR/Increased Risk:

- No baseline risk reported (i.e., relative to what?)
- No time period identified
- Reported value is not your risk

The odds ratio (or relative odds), like RR, is used to measure the relative probability of disease.

- Odds in favor of an event having probability of occurrence, p, is, odds = p/(1-p)
- Thus, odds of having a disease = $P(D^+)/(1-P(D^+))$
- Odds ratio (OR) of disease: compute the odds of disease for 2 different groups and compute the ratio :-)
- For SES e.g.
  - odds of PRS for high SES =
  - odds of PRS for low SES =
  - odds ratio of low to high SES =
  - Interpretation

Summary of RR and OR

- If RR=OR=1,then there's no relationship between disease status and the "explanatory variable" (i.e., the conditioning variable or risk factor (e.g. SES))
- If probability of disease is small for both categories of conditioning variable, then RR ≈ OR (see 15.3)
- In a case-control study, individuals with an event or condition of interest (the cases with, e.g., lung cancer), are identified and then compared, with regard to one or more exposures or risk factors, to individuals without the event

or condition of interest (the controls, e.g., without lung cancer).

- A cohort study is one in which subjects (the cohort), initially disease free, are chosen based on exposure to some risk factor (e.g. smoking) thought to influence the occurrence of disease (e.g. lung cancer) and are followed up over a period of time. Some will develop the disease and some will not, and we may be interested in relating the risk factor to disease occurrence.
- Note that in the case of the cohort study, the exposure is initially known (i.e., "given"—we condition on this). But, with the case-control study, the proportion of patients with the disease and without the disease is chosen, so that it is the disease status that is fixed (i.e., "given"—we have to condition on this). Thus, we can calculate RR and OR for the cohort study (conditioning on exposure), but only OR for the case-control study using the mathematically equivalent definition:
  - $OR=P(\text{exposed}|D^+)/(1-P(\text{exposed}|D^+)) \div P(\text{exposed}|D^-)/(1-P(\text{exposed}|D^-))$