

# **PubH 7405: REGRESSION ANALYSIS**



**Instructor: Chap T. Le, Ph.D.**

**Distinguished Professor of Biostatistics**

**Statistical Process & Probability**

# **A FAMILIAR SCENARIO:**

- **A Crime**
- **A Suspect**
- **Police's Investigation: Evidence against the suspect**
- **Prosecutor's Presentation: Exhibitions (summarized evidence)**

# **THE PROCESS:**

- **Jurors learn the rules & debate:**
  - Convict “beyond reasonable doubt”**
  - Unanimous decision**
- **The verdict: Guilty or Not Guilty**

**The Process is called TRIAL BY JURY**

# TRIAL BY JURY: THE NEED

**Why such an expensive process is needed ? (the cost is enormous!)**

**(1) The Truth is unknown, at least uncertain.**

**(2) May be only the suspect knows, but he/she often does not talk**

# TRIAL BY JURY: THE REASONS

The Truth is unknown/uncertain because:

(1) Variability: every case is different

(2) Incomplete Information: key evidence

(gun, knife, motive, etc...) may be missing!

# TRIAL BY JURY: RATIONALE

**“Trial by Jury” is the way our society deals with uncertainties. Its goal is to minimize errors/mistakes (not to eliminate them; mistakes are still being made every day!)**

# HOW DOES SOCIETY DEAL WITH UNCERTAINTIES ?

- We form **Assumption/Hypothesis**: “Every person is innocent until proven guilty” (written in our Constitution),
- We gather **data**: Evidence against Hypothesis- not against the suspect, then
- We **decide** whether Hypothesis should be rejected (If it is, the verdict is “Guilty”)

# ELEMENTS OF A SUCCESSFUL TRIAL

- A probable **CAUSE** (a crime and a suspect)
- A thorough **INVESTIGATION** (by police)
- An efficient **PRESENTATION** (by D.A.'s office/attorneys- including the organization and summarization of evidence)
- A fair & impartial **ASSESSMENT** by Jury where a decision is made.

# WHY TALKING ABOUT TRIALS BY JURY ?

Consider a few other examples:

- The crime is lung cancer & the suspect is (cigarette) smoking,
- The crime is leukemia & the suspect is the use of pesticides
- The crime is breast cancer & the suspect is certain defective gene or genes

# NEW TERMINOLOGIES

- In the trials of “Smoking”, of “Pesticides”, or of a “Defective Gene”, the Process is not called Trial by Jury; We call it **RESEARCH**.
- The “tool”, the **process**, to carry out the needed research is **BIOSTATISTICS**.
- **Biostatistics** is an essential component; there are no research without this needed process.

# DEFINITION

**BIOSTATISTICS is the Biomedical Version of the TRIAL BY JURY. It is “the science of dealing with uncertainties using incomplete information.” Obviously, it is an essential component of Biomedical Research; we have to face uncertainties and, most of the times, we have to rely on incomplete information.**

# THINKING OF MARRIAGE?

Statistics may help you to decide!

# FACTORS TO CONSIDER

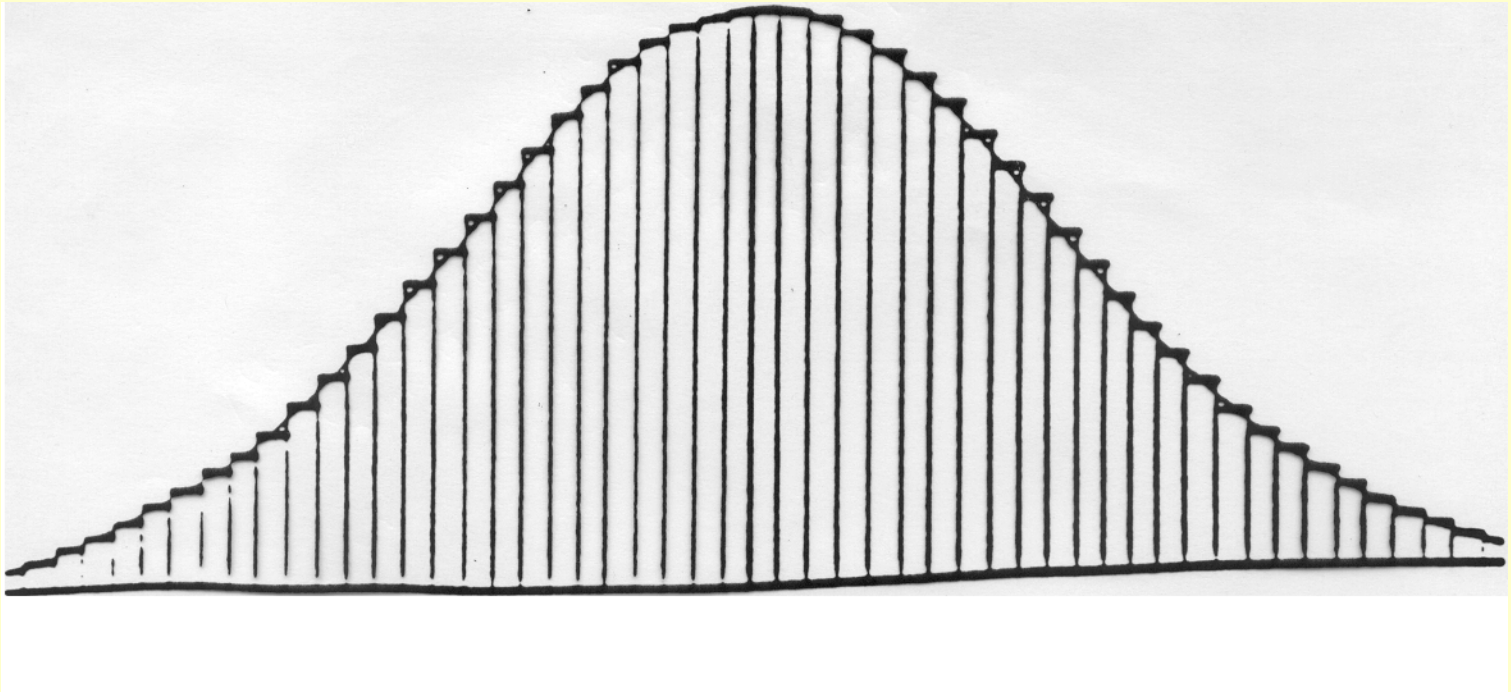
- Median age of first marriage
- Percent of all marriages (all first marriages) that end in divorce.
- Average length of a marriage (a first marriage)
- Percent of divorces that involve children
- Percent of married adults (single adults, divorced adults) who are satisfied with their life (and how are these numbers compared; are you still thinking of marriage after reading them?)

# MADD?: SOME FACTORS RELATED TO MOTOR VEHICLE FATALITIES

- **Driving age (% of drivers under 21)**
- **Alcohol-purchase age**
- **Average beer consumption**
- **Number of outlets selling alcohol**
- **Percentages of teen and male drivers**
- **Average mileage drive per year, etc...**

**Should you join the “MENSA”?**

**Intelligence test score, referred to as Intelligence Quotient - or IQ scores - are based on characteristics such as verbal skill, abstract reasoning power, numerical ability, and spatial visualization.**



If plotted on a graph with IQ scores on the horizontal axis, **the distribution of IQ scores forms a bell-shaped curve**, sort of like a handlebar moustache, referred to as “**IQ Curve**”.

# QUESTIONS TO PONDER

- **What would be average IQ? Are you “normal”?**
- **What is considered superior?**
- **How do schools defined/selected gifted children? Who join MENSA?**
- **Who would need or be qualified for special education?**

What do these examples  
have in common?

You need information to decide, but...

# THE CERTAINTY OF UNCERTAINTIES

Everything is uncertain including Science:

- **Different conclusions at different times**  
(effects of certain food ingredients- for example, eating eggs & exposure to low-level radioactivity)
- Classic Example: **Radical mastectomy vs. less drastic treatments**
- **Many studies are inconclusive!**

# **SAME REASONS FOR UNCERTAINTIES**

- **Variability**: Nature is complex, methods are imperfect, observers/investigators may be biased, subjects vary, measurements fluctuate; some explainable, some not.
- **Incomplete information**: cost, time, and future/moving target. We often rely on information gained from samples

# HOW DOES SCIENCE DEAL WITH UNCERTAINTIES ?

- **We form Assumption/Hypothesis: From experience & observations (The process leads to the so-called research questions)**
- **We gather data: Experiments & Trials, Surveys, Medical Records Abstractions.**
- **We make decision by performing DATA ANALYSIS, the “core” area of Biostatistics.**

# ELEMENTS OF GOOD RESEARCH

- A good **RESEARCH QUESTION** with well-defined objectives & endpoints,
- A thorough **INVESTIGATION**, lots of data
- An efficient **PRESENTATION**: data organization & summarization, and
- A proper **STATISTICAL INFERENCE** (the process & methods of drawing conclusions)

# AREAS OF BIOSTATISTICS

**Research is a three-step process:**

- (1) **Sampling/design**: Find a way or ways to collect data (going from population to sample).
- (2) **Descriptive statistics**: Learn to organize, summarize and present data which can shed light on the research question (investigating sample).
- (3) **Inferential statistics**: Generalize what we learn from the sample or samples to the target population and answer the research question (going from sample to population).

**Validity** is an important concept/component in research. It involves the assessment against accepted absolute standards which are often not available; or in a milder form, to see if the evaluation appears to cover its intended target or targets. Statistical contributions involve both Internal Validity and External Validity of any research project.

# STATISTICAL ISSUES

- **Statistics is a way of thinking**, thinking about ways to gather and analyze data.
- The gathering part (i.e. **data collection**) comes before the analyzing part; the first thing a statistician or a learner of statistics does when faced with a biomedical project is data collection (followed by **data management** and **data analysis**).
- Studies may be inconclusive because they were poorly planned, **not enough data** were collected to accomplish the goals and support the hypotheses.

# THE IMPORTANT PHASE

Just as in the case of “Trial by Jury”, the most **important stage** of the “Research Process” is the DESIGN: How & How Much data are collected! Also, It dictates how data should be analyzed. **May be it’s not the question of “how” to collect your data but the decision on “when to do what”!**

# EXPERIMENTAL DESIGNS

There are three different Designs (methods for data collection) depending on the timing (present, past, and future) and the focus (disease or exposure):

- **Cross-sectional, e.g. surveys**
- **Case-Control (retrospective)**
- **Cohort (prospective); clinical trials are of an important special form.**

# Cross-Sectional Design

	Factor Present	Factor Absent
Disease		
No Disease		

Take One Sample

The **cross-sectional** designs are very popular in social/behavioral studies, e.g. teen surveys. As for health research data, since **diseases are rare**, fundamental designs are case-control and cohort.

# Case-Control Design

	Factor Present	Factor Absent	
Disease			Sample 1: Cases
No Disease			Sample 2: Controls

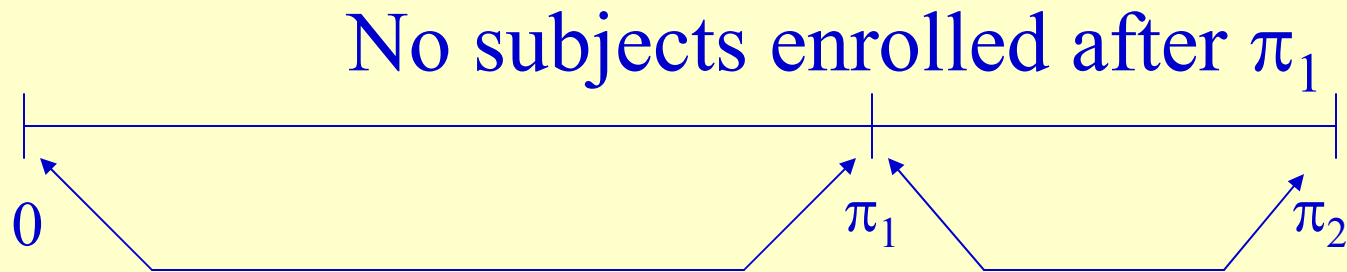
**Retrospective Studies** gather past data from selected cases (with disease) and controls (without disease) to **determine differences**, if any, in exposure to a suspected risk factor.

**Advantages:** Economical & Quick. **Major Limitations:** Accuracy of exposure histories & Appropriateness of controls

# A CLINICAL TRIAL

Study Initiation

Study Termination



Enrollment Period,  
e.g. three (3) years

Follow-up Period,  
e.g. two (2) years

**OPERATION:** Patients come sequentially; each is enrolled and randomized to receive one of two or several treatments, and followed for varying amount of time- between  $\pi_1$  &  $\pi_2$

# SOME TERMINOLOGIES

- **Research Designs**: Methods for data collection
- **Clinical Studies**: Class of all scientific approaches to evaluate Disease Prevention, Diagnostics, and Treatments.
- **Clinical Trials**: Subset of clinical studies that evaluates Investigational Drugs; they are in **prospective/longitudinal** form (the basic nature of trials is prospective).

# CANCER TRIALS

- **Phase I:** First human trial to focus on safety
- **Phase II:** Small trial to evaluate efficacy
- **Phase III:** Large controlled trial to demonstrate efficacy prior to FDA approval
- **Phase IV:** Optional, post-regulatory approval, to provide the medicine's more comprehensive safety and efficacy profile

# DESCRIPTIVE STATISTICS

**Tasks:** To organize, to summarize, and to present collected data. There are three different categories:

- **Tabular Methods:** Tables
- **Graphical Methods:** Graphs, Charts
- **Numerical Methods:** Few Statistics.

**Aims:** To communicate **more** effectively

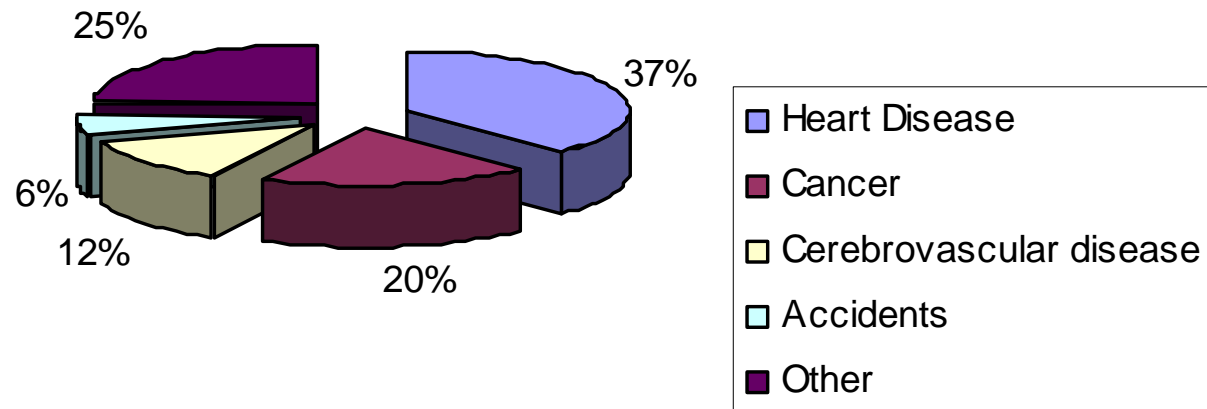
# EXAMPLE

- A study was conducted to determine whether the use of **Electronic Fetal Monitoring (EFM)** during labor affects the frequency of Caesarian Section delivery:

Caesarian	EFM	(%)	No EFM	(%)	Total
Yes	358	12.6	229	7.7	587
No	2,492		2,745		5,237
Total	2,850		2,974		5,824

- **Contingency tables**, or two-way tables, are popular methods for presenting data which are intended to show a possible relationship between two factors: exposure & outcome.

## Causes of Death for Minnesota Residents



**Pie Charts**: a popular device to present data which are intended to show the decomposition of a total into several components.

# STATISTICAL INFERENCE

## Two major Areas:

- **Tests of Significance**: For treatment evaluation, an analog of the trial by jury (New Therapy Vs. Standard Therapy)
- **Confidence Intervals**: For Parameter Estimation (e.g. Response Rate to a new drug, Toxicity Rate, Average SBP, Average Cholesterol Level)
- This is the topic of the next review hour

# Inferential Decision & Sampling

In order to make an inferential decision (step #3), that is to generalize what we learned about the sample to draw conclusion about the population (sample  $\longrightarrow$  population), we need to know what happened in step #1 (which goes population  $\longrightarrow$  sample).

**What could happen to the sample? For example, can sample mean  $\bar{x}$  (x-bar) be very away from population mean  $\mu$ ? How likely? How far?**

# Inferential Decision & Sampling

- Here is a simple analog: “I want to find your house. If I know my house is 2 blocks from your house, then I know your house is 2 blocks from my house”
- The population mean  $\mu$  is a parameter. It's fixed but unknown and we want estimate it. The sample mean  $\bar{x}$  is a statistic; it is known. If we know, from step #1, how far is  $\bar{x}$  from  $\mu$ ; then after step #3, we know how far is  $\mu$  from  $\bar{x}$ : problem in step #3 is solved!
- But what would happen in step #1 involves uncertainty.  
**How do we “measure” uncertainty?**

Basically, Science deals with “uncertainties” by inventing the concept of “Probability”. We can estimate probabilities so that we can **minimize the impact of uncertainties.**

# Proportion

- Proportion is defined as the “relative size” of the portion of the population with certain characteristic.

## For example:

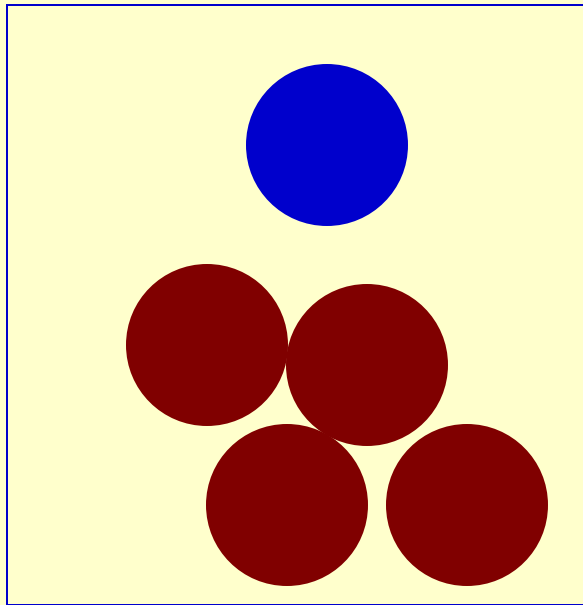
- (i) Disease prevalence is the proportion with disease
- (ii) Proportion of positive reactors to certain “test”
- (iii) Proportion of males in college (e.g. Public Health)
- **Used as a descriptive measure** with respect to a binary factor under investigation.
- A number between 0 and 1; the larger the number the larger the sub-population with the characteristic e.g. 70% male means more males (than 50% male).

# Random Drawing & Chance

- Consider a population with a certain characteristic. A **random draw/selection** is one in which each subject has an equal chance of being selected.
- What is “**the chance**” that an individual with the characteristic be selected? It depends on the “size” of the sub-population the he belongs to, i.e. the “proportion.” **The larger the proportion, the higher the chance.**
- The “chance” is measured by the proportion, a number between 0 and 1, called the “probability.”

# Proportion & Probability

- Proportion measures “size,” Probability measures “chance.” They are the same number.
- When we are concerned about the outcome (still uncertain): **Proportion** (static, no action) becomes **Probability** (action about to be taken).
- If we keep taking random selections, the long-term accumulated relative frequency the characteristic is observed is equal to the proportion associated with the characteristic. Because of that proportion and probability are sometimes used interchangeably.



FACT:

There are four red circles and one blue circle in this box; and suppose that you see them but I do not.

Q. Are there any red circles in the box?

A. Yes, 80% of the circles are red

The figure “80%” is a proportion; you use it to describe the content of the box.

**(Nothing is unknown, nor uncertain)**

Q. If I pick a circle at random, do you think that I would get a red circle?

A. Yes you would, 80% likely.

The figure “80%” is a probability; you use it to tell me my chance based on what you see in the the box.

**(Unknown and uncertain)**

(I chose a circle at random and hide it from you)

Q. What is the color of the circle I got?

A. Red, I'm 80% sure of that

The figure “80%” is your degree of confidence based on your knowledge of the content of the box.

**(Unknown -to you- but not uncertain)**

It's the same figure, 80%, in all three cases but it's:

(1) A **proportion**, used to describe

(2) A **probability**, used to measure a chance for an outcome (when action is about to be taken, **outcome is unknown and uncertain**)

(3) A **degree of confidence** about a decision or prediction after action has been taken. **The out come is unknown but not uncertain.**

WOULD PETE ROSE BET?

HIMSELF

Versus

JOE DIMAGGIO

Joe DiMaggio became a baseball legend with his famous 56-game hitting streak in 1941.

It has been considered in baseball circles as an **unbreakable major league standard** (including those who do not know how he was related to Marilyn Monroe).

Then came Pete Rose of the Cincinnati Reds with his own streak breaking many records in the summer of 1978.

When Pete set his 44-game record, Las Vegas odds makers started to focus on DiMaggio's 56-game record.

Would Pete be able to break Joe's unbreakable major league standard?

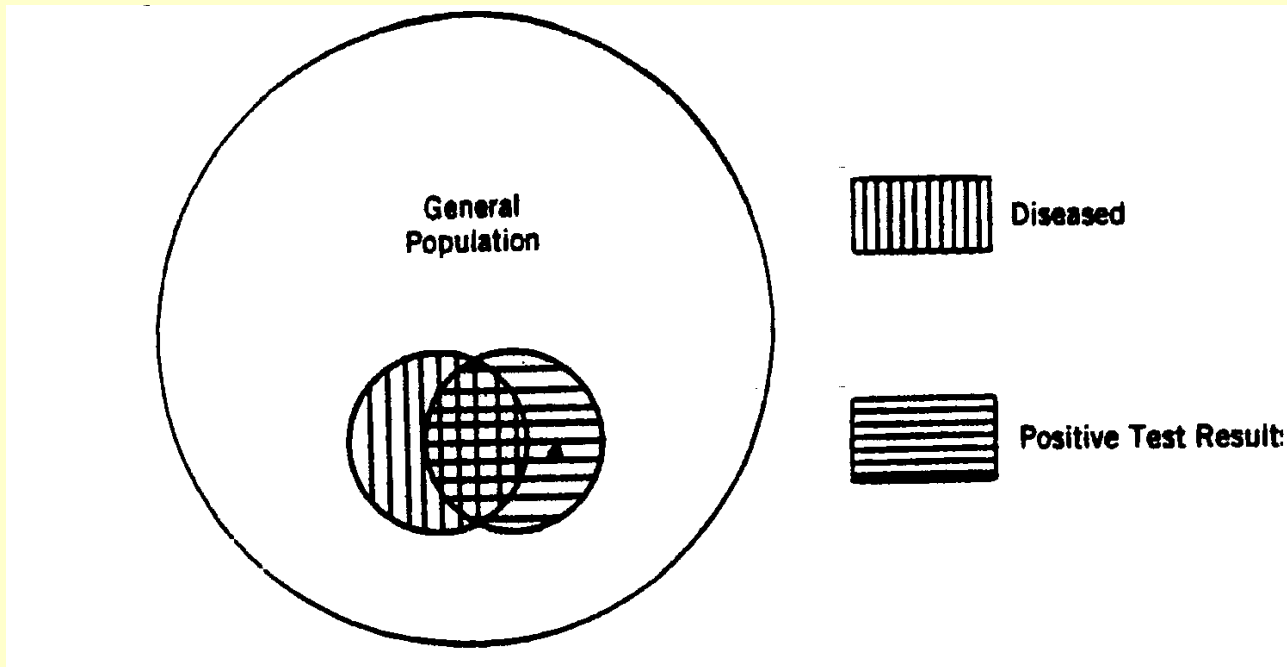
# WHAT'S THE ODDS?

- During the streak, Pete hit .376; assuming 4 at bats in a game, the **probability** of Pete having at least 1 hit is:  $1-(1-.376)^4 = .8484$
- The **probability** of hitting in 12 successive games is  $.8484^{12} = .1391$  or 13.91%
- So the **Odds** that Pete Rose could match Joe DiMaggio's performance **is less than 1-to-6**
- Pete's streak was ended when he went 0-for-4 in the next game with the Atlanta Braves.

# Developmental Stage of Screening Tests

- It starts with an idea and could be accidental or the result of a long search.
- **Stage I: Developmental**
- The question here is: Does the idea work?
- **Approach**: Test results versus truth; key parameters (conditional probabilities):
- **Sensitivity =  $\Pr(T=+|D+)$**
- **Specificity =  $\Pr(T=-|D=-)$**

# REASON: MISCLASSIFICATION



	Test=Positive	Test=Negative
Diseased	True Positive	False Negative
Healthy	False Positive	True Negative

# Applicational Stage of Screening Tests

- **Stage II: Applicational**
- The question here is: Does it work for “me”? (the testee; or when does it work?)
- **Key parameters** (conditional probabilities):
- **Positive Predictive Value =  $\Pr(D=+|T=+)$**
- **Negative Predictive Value =  $\Pr(D=-|T=-)$**   
(Users are more often concerned about PPV; Note: “D” is unknown in this stage)

# SHOULD WE CONDUCT “RANDOM TESTING” FOR AIDS?

## Assumptions:

- ❖ Complete privacy
- ❖ Complete confidentiality
- ❖ There a good/reliable screening procedure  
Let consider a low-risk sub-population  
and a high-risk sub-population.

# BAYES' RULE

$$\Pr(B | A) = \Pr(B \text{ and } A) / \Pr(A)$$

$$\Pr(B | A) = \frac{\Pr(A | B)\Pr(B)}{\Pr(A \text{ and } B) + \Pr(A \text{ and Not } B)}$$

$$\Pr(B | A) = \frac{\Pr(A | B)\Pr(B)}{\Pr(A | B)\Pr(B) + \Pr(A | \text{Not } B)\Pr(\text{Not } B)}$$

# APPLICATION

$$\Pr(B | A) = \frac{\Pr(A | B)\Pr(B)}{\Pr(A | B)\Pr(B) + \Pr(A | \text{Not } B)\Pr(\text{Not } B)}$$

Let  $A = (T=+)$  and  $B = (D=+)$ , we have:

$$\Pr(D = + | T = +) = \frac{\Pr(T = + | D = +)\Pr(D = +)}{\Pr(T = + | D = +)\Pr(D = +) + \Pr(T = + | D = -)\Pr(D = -)}$$

**Note:** “not B” = (D=-)

# RESULTS

**Both predictive values are functions of disease prevalence,  $\pi = \Pr(\mathbf{D} = +)$ :**

$$P^+ = \frac{S^+ \pi}{S^+ \pi + (1 - S^-)(1 - \pi)}$$

$$P^- = \frac{S^- (1 - \pi)}{S^- (1 - \pi) + (1 - S^+) \pi}$$

# EXAMPLES

- **Example A**:  $S^+ = .977$ ,  $S^- = .926$ , and  $\pi = .003$ :

$$P^+ = \frac{(.977)(.003)}{(.977)(.003) + (.074)(.997)} = .038 \text{ or } 3.8\%$$

- **Example B**:  $S^+ = .977$ ,  $S^- = .926$ , and  $\pi = .20$ :

$$P^+ = \frac{(.977)(.20)}{(.977)(.20) + (.074)(.80)} = .767 \text{ or } 76.7\%$$

- **Note**: Current Estimate for USA's AIDS: .3% as above and  $S^+$  and  $S^-$  are for ELISA in Weiss, 1985).

# Learned Lessons

- Predictive values of a screening test depend not only on sensitivity and specificity but on disease prevalence as well. The higher the prevalence, the higher predictive values.
- We should only “screen” high-risk sub-population; “random screening” does not do anyone any good! Not to testees, not to policy makers.

# VARIABLE & DISTRIBUTION

- Any rule that associates with each outcome of an experiment a corresponding number is called a **“variable”**
- The number that a variable associates with a particular outcome is called **the “value” of the variable** for that outcome.
- A list of possible values of the variable, together with their corresponding probabilities, is called **the “distribution” of the variable.**

# AN EXAMPLE

- **Experiment:** two students are selected at random from a college.
- **Variable:** the number  $X$  of females in the sample; possible values are: 0, 1, and 2
- **Distribution:** We need to know the proportion of female students of that college. Suppose 60% of students are women, then:
  - **$\Pr(X=0) = .16 = (.4)(.4)$**
  - **$\Pr(X=2) = .36 = (.6)(.6)$**
  - **$\Pr(X=1) = 1-.16-.36 = .48$**

# PROBABILITY DENSITY FUNCTION

For a “discrete” sample space and a random variable  $X$ , a Probability Density Function (pdf)  $f$  is defined so that:

$$f(k) = \Pr(X = k)$$

$$\sum_s f(k) = 1$$

# THE BINOMIAL DISTRIBUTION

- Let focus on a **“trial” with binary outcome**; for example the gender of a student selected at random, or the opinion (approved, not approved) of resident selected at random on a new piece of legislation.
- Suppose the experiment consists of repeating the above trial  **$n$  times independently**; e.g. asking the opinion of  $n$  person in a survey.
- Let  $X$  be the number of “approvals” from the survey;  $X$  is said to have a **“binomial distribution”**

# THE BINOMIAL DISTRIBUTION

- In general, let denote the 2 possible outcome of a binary trial by S (success) and F (failure). The number  $X$  of successes from  $n$  independent trials has a binomial distribution.
- Possible values of  $X$  are  $0, 1, 2, \dots, n$ ; the probability associated with each value depends on  $n$  and the probability of getting a success in a single trial. Let call this “ $\pi$ ”; we have  $B(n, \pi)$ .
- **Example:** we had simple one: number of females if 2 students are selected at random from college.

# BINOMIAL DISTRIBUTION

- Pdf for the binomial distribution  $B(n, \pi)$ :

$$\Pr(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

- **Example:** Let  $X$  be the number of females when 2 students are selected at random from a college with 60% women. We have:

$$\Pr(X = 1) = \binom{2}{1} (.6)^1 (1 - .6)^{2-1} = \frac{2!}{1!!} (.6)(.4) = .48$$

# CONTINUOUS VARIABLES

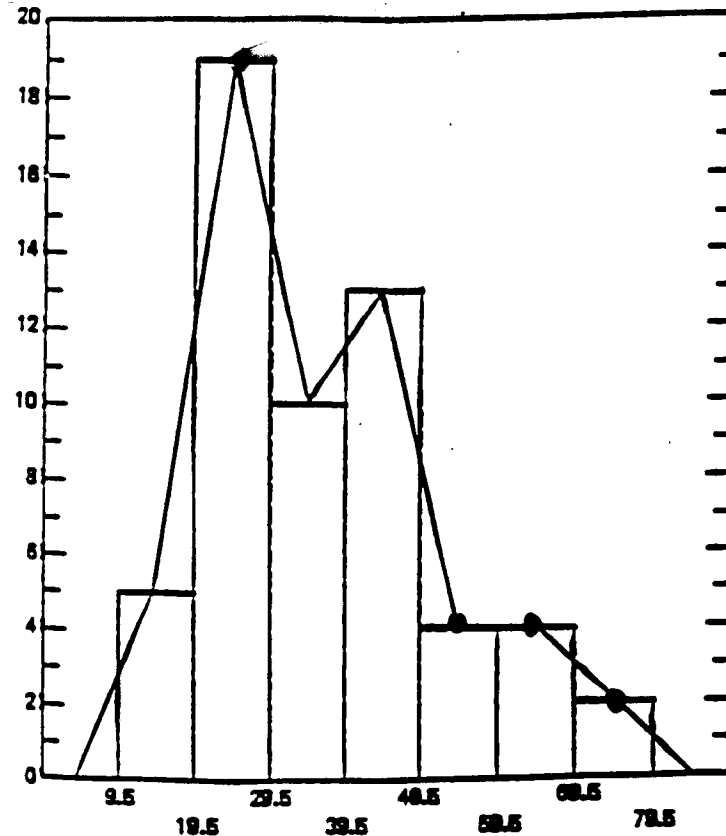
Recall:

## Histogram & Frequency Polygon

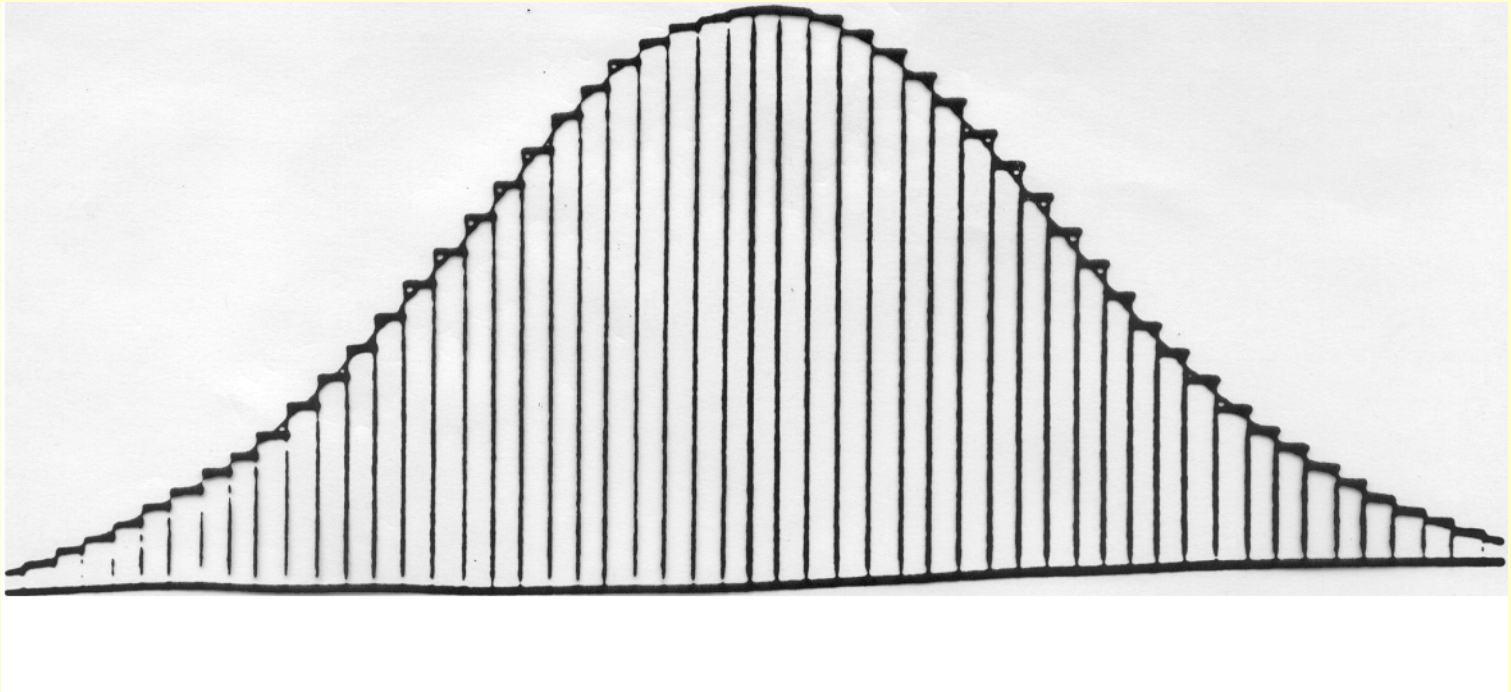
\* Horizontal axis:  
measurement, marked at  
boundaries.

\* Vertical Axis: density,  
defined as:

$$\frac{\text{Relative Frequency}}{\text{Interval Width}}$$



Distribution of Weights of 57 Children



Suppose we have data from all members of a population. There are a lot of numbers. Then we use intervals of very small width to construct a histogram and a frequency polygon. We would get an almost **smooth curve**, called **a density curve**.

# Density Curve

- A density curve is a smooth curve characterizing the distribution of a continuous variable for a population (whereas the frequency polygon plays the same role for a sample).
- Because we graph density versus measurement, we have for each (very thin) rectangle:

$$\text{Area} = (\text{Density})(\text{Width}) = \left(\frac{\text{Percentage}}{\text{Width}}\right)(\text{Width}) = \text{Percentage}$$

- Total area under a density curve is 1.0 or 100%

# PROBABILITY DENSITY FUNCTION

For a “continuous” sample space and a random variable  $X$ , a Probability Density Function (pdf)  $f$  is defined so that:

$$f(x)dx = \Pr(x \leq X \leq x + dx)$$

$$\int_S f(x)dx = 1, \text{ and}$$

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

# Normal Curves

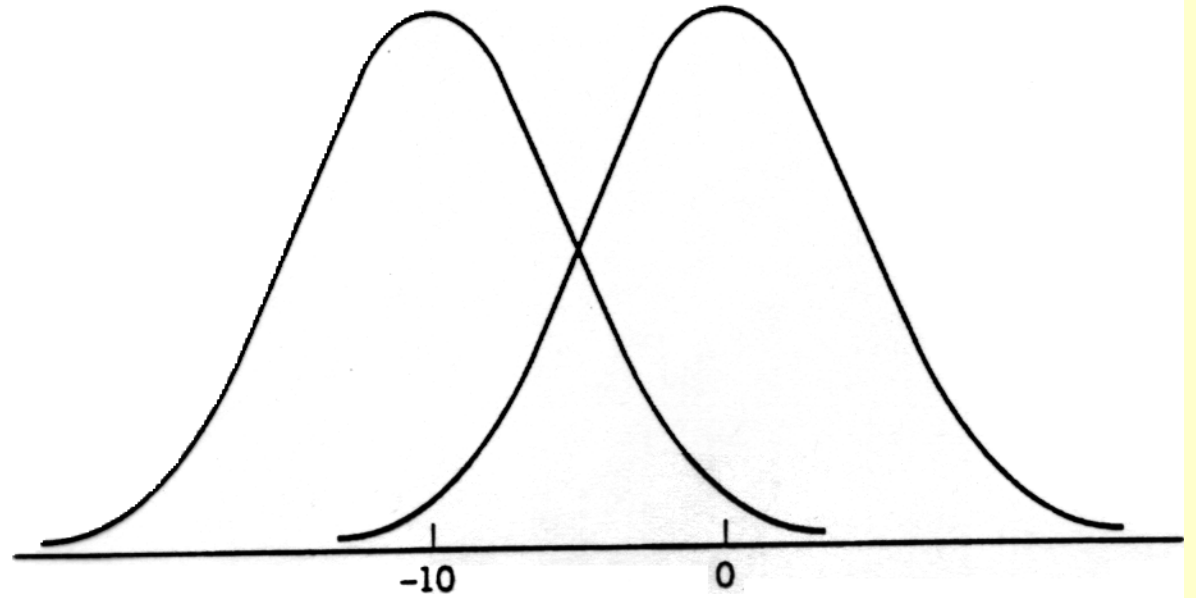
- Density curves: There are many possible shapes, including the normal curve (one of many possibilities).
- Each normal curve is a bell-shaped, symmetric curve. The total area under the curve is 100%.
- It is uni-modal. The location of the peak is the mean  $\mu$ . The height of the curve is inversely proportional to the standard deviation  $\sigma$ .
- Each curve is characterized by a mean  $\mu$  and a standard deviation  $\sigma$ . What we have is a family of curves.

## Normal Density $N(\mu, \sigma^2)$

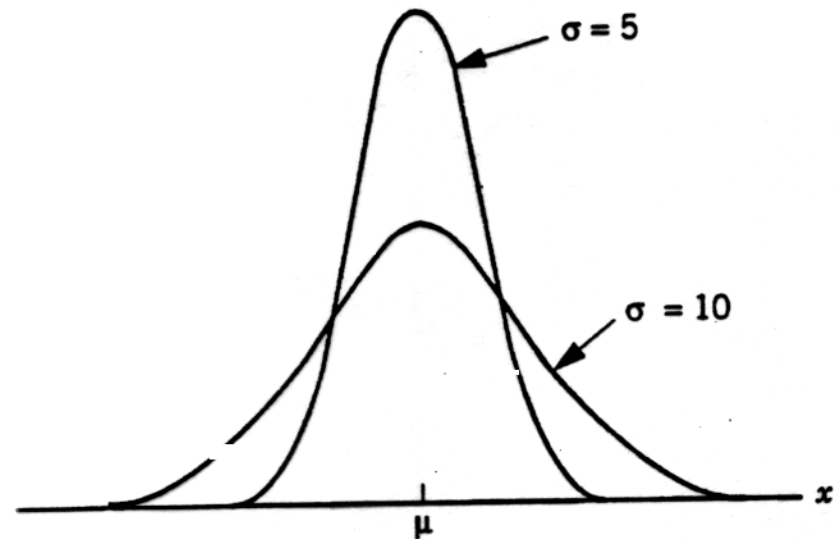
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}; \text{ for } -\infty < x < \infty$$

**An important special case is the  
“Standard Normal” Curve with Mean  
“0” and Standard Deviation “1.0”**

Two normal curves with same standard deviation but different means (different peak locations, but same height).



Two normal curves with same mean but different standard deviation (same peak location, but different heights).



# WHY NORMAL DISTRIBUTION?

- Many variables/characteristics are distributed almost as “normal”: symmetric bell-shaped curves (Height, Weight, Blood Pressure, Cholesterol, etc.)
- Sometimes a data transformation, such as taking log, would help to “normalize” a density curve, especially when it is positively skewed. Examples include income and antibody level.
- **Major reason: The Central Limit Theorem;** we save for the next review hour.