

# PubH 7405: REGRESSION ANALYSIS



**Basics of Statistical Inference**

# The Research Process

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

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

# WHERE ARE WE?

- We reviewed/described the “**Statistical Process**”
- We talked briefly about “**Study Designs**” (cross-sectional, case-control, and cohort).
- We mentioned briefly “**Descriptive Statistics**”
- We covered “**Probability & Probability Models**” as means to deal with uncertainties in research.
- Now, it’s last step: **Statistical Inference**, apply what we learned/knew to draw conclusions.

# Statistical Inference

- The last step of the data analysis process is called inferential statistics; statistical methods helping us to reach conclusions, **using what we learn from sample(s) to apply to the target population.**
- There are two sub-categories:
  - (1) **Interval Estimation**, which allows us to estimate a parameter (e.g. smoking rate, disease prevalence).
  - (2) **Hypothesis Testing**, which allows us to test hypotheses, i.e., **to compare parameters** (as in treatment evaluation or evaluation of public health intervention programs).

# THE STORY OF A LOST BOY

- **Scene**: A little boy crying “mommy!”
- **Task**: Help him to find his mother
- **Strategy**: Look around - **little kids can't go far**; But **HOW FAR** should we look?  
Factors to consider: His age, Traffic condition, **how SURE** do you want to be.  
The result? may be his mother is **a couple of blocks either way** .

# STORY OF UNEMPLOYMENT RATE

- Lost “mother”: True Unemployment Rate (say, in the state of Minnesota in September of 2008)
- The “Boy”? Unemployment Rate from a sample
- Strategy: Look around, but how far should we look. Factors to consider: sample size, chance variation, how SURE we want to be.

**RESULT: A Confidence Interval:**

**Estimate +/- Margin of Error**

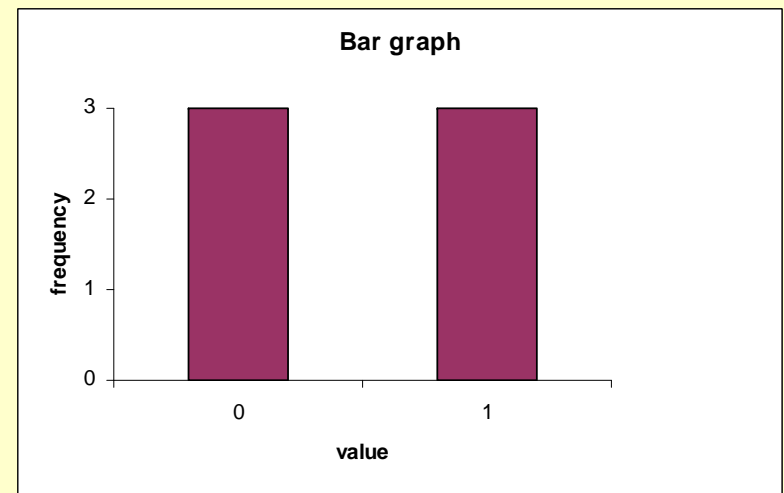
**(just like a couple of blocks either way)**

# Sampling Distribution

- For simplicity, consider a small population of size  $n = 6$ .
- Values are listed in the second column.
- The mean is 0.5.
- There is nothing special (i.e., **not normal**) about the shape of the histogram.

Subject	Value
A	1
B	1
C	1
D	0
E	0
F	0

$$\mu = \frac{3(1) + 3(0)}{6} = 0.5$$



Taking all possible samples of size  $n = 3$ :

The mean of all sample means is equal to the population mean (0.5)

Samples	Number of samples	Value of sample mean
(D,E,F)	1	0
(A, D, E), (A, D, F), (A, E, F)	9	1/3
(B, D, E), (B, D, F), (B, E, F)		
(C, D, E), (C, D, F), (C, E, F)		
(A, B, D), (A, B, E), (A, B, F)	9	2/3
(A, C, D), (A, C, E), (A, C, F)		
(B, C, D), (B, C, E), (B, C, F)		
(A, B, C)	1	1

**The mean of all possible sample means:**

$$\mu_{\bar{x}} = \frac{1(0) + 9(1/3) + 9(2/3) + 1(1)}{20} = 0.5 (= \mu)$$

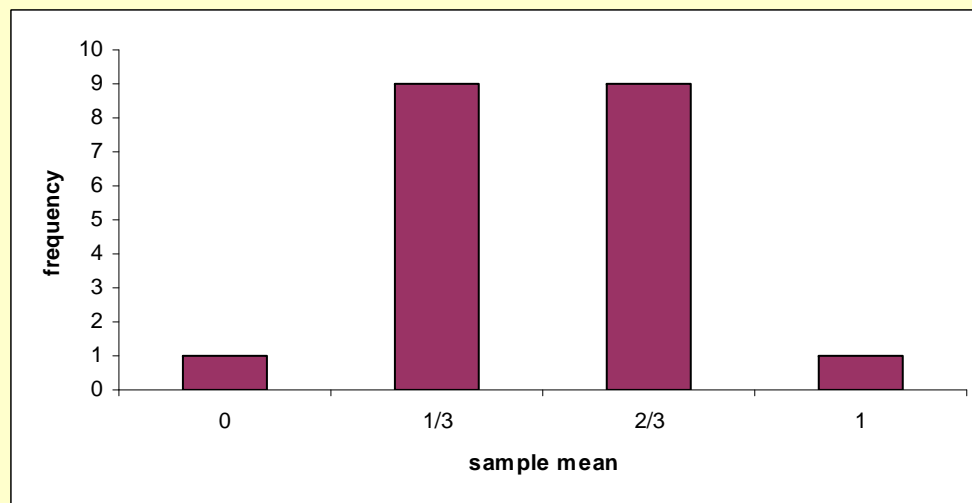
Subject	Value
A	1
B	1
C	1
D	0
E	0
F	0

The mean of all possible sample means:

$$\mu_{\bar{x}} = \frac{1(0) + 9(1/3) + 9(2/3) + 1(1)}{20} = 0.5 (= \mu)$$

We form a bar graph for this sampling distribution,

The “shape” of the histogram representing the distribution of all possible sample means looks **more “normal”** than the one for the population!



# Increase the $n$ value

• If  $n = 4$ , the mean of all sample means is still 0.5.

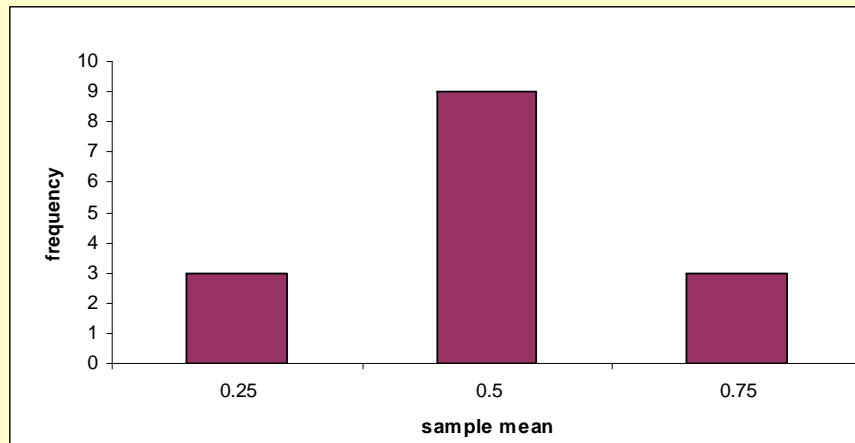
• The shape is even more normal.

Samples	samples	sample mean
(A, D, E, F), (B, D, E, F), (C, D, E, F)	3	0.25
(A, B, D, E), (A, B, D, F), (A, B, E, F)	9	0.5
(A, C, D, E), (A, C, D, F), (A, C, E, F)		
(B, C, D, E), (B, C, D, F), (B, C, E, F)	3	0.75
(A, B, C, D), (A, B, C, E), (A, B, C, F)		
<b>Total</b>	<b>15</b>	

The mean of all possible sample means:

$$\mu_{\bar{x}} = \frac{3(.25) + 9(.50) + 3(.75)}{15} = 0.5 (= \mu)$$

and the bar graph:



Subject	Value
A	1
B	1
C	1
D	0
E	0
F	0

$\bar{X}$  is an Unbiased Estimator for  $\mu$ ; If we estimate  $\mu$  by  $\bar{X}$ , we are correct on the average

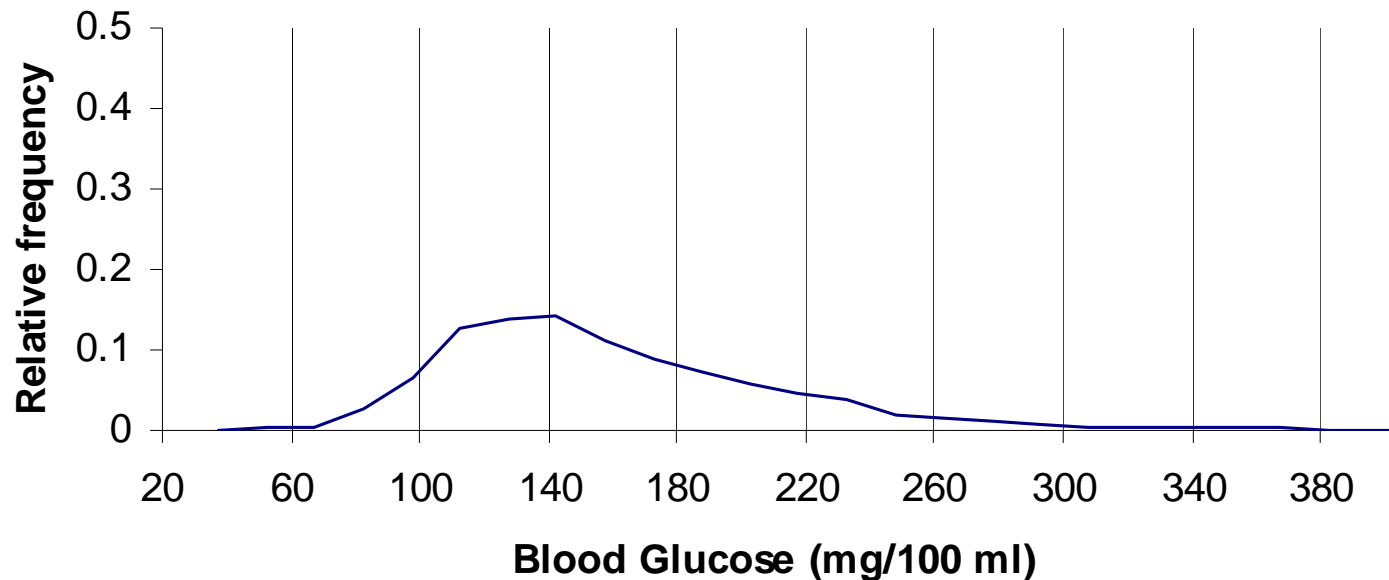
# Using a Larger Population

- Blood glucose measurements from 7,683 men in Honolulu.
- Take **400 samples**, 25 each. Sample means shown at left.
- Means of two distributions are approximately the same (There are many more than 400 possible samples).
- Variance of the distribution of sample means is smaller.

Blood glucose (mg/100ml)	Number of observations (frequency)	Sample means (n=25) (frequency)
30.1--45.0	2	
45.1--60.0	15	
60.1--75.0	40	
75.1--90.0	210	
90.1--105.0	497	
105.1--120.0	977	
120.1--135.0	1073	5
135.1--150.0	1083	62
150.1--165.0	849	201
165.1--180.0	691	109
180.1--195.0	569	23
195.1--210.0	440	
210.1--225.0	343	
225.1--240.0	291	
240.1--255.0	153	
255.1--270.0	115	
270.1--285.0	82	
285.1--300.0	60	
300.1--315.0	38	
315.1--330.0	18	
330.1--345.0	26	
345.1--360.0	19	
360.1--375.0	20	
375.1--390.0	9	
390.1--405.0	13	
405.1--420.0	11	
420.1--435.0	6	
435.1--450.0	5	
450.1--465.0	4	
465.1--480.0	24	
Total	7683	400

# Distribution of Blood Glucose Values

**Distribution of Blood Glucose Values from the Honolulu Heart Study Population (N=7683)**



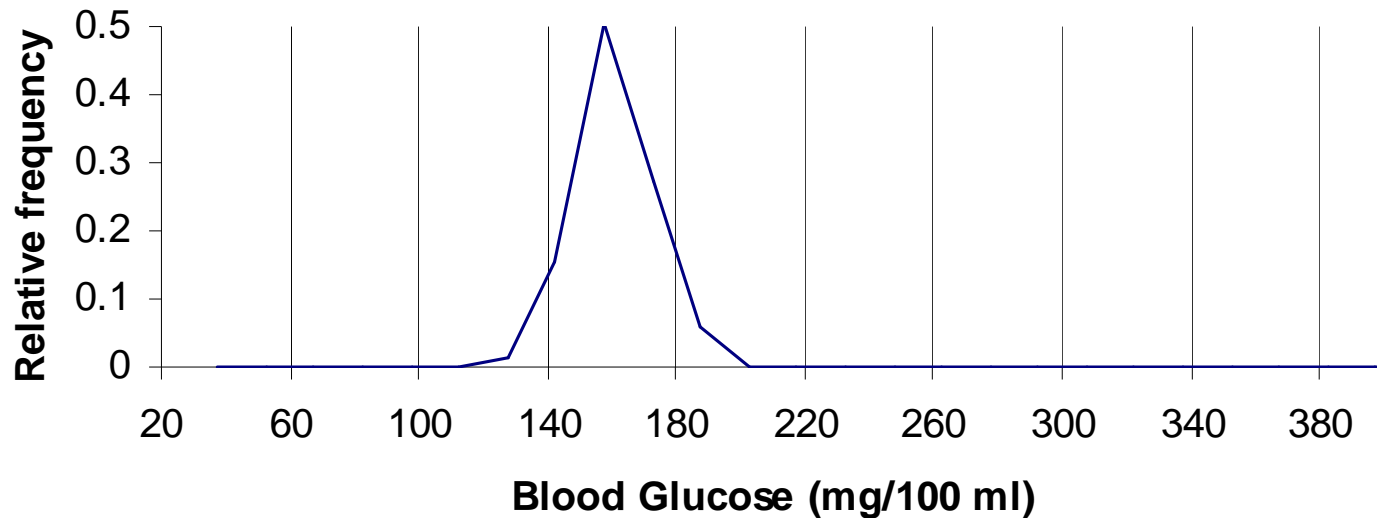
$$\mu = 161.52$$

$$\sigma = 58.15$$

Population distribution is not even symmetric!

# Distribution of 400 Sample Means

**Distribution of Means of Samples of Blood Glucose Values (n = 25) from the Honolulu Heart Study**



$$\mu = 160.66$$

$$\sigma = 12.24$$

The sampling distribution is a bit more normal!

# CENTRAL LIMIT THEOREM

- Given any population with Mean  $\mu$  and Variance  $\sigma^2$  (Standard Deviation  $\sigma$ ): The Sample Mean is a “variable”; the (sampling) distribution of its possible values, with (large) sample size  $n$  being fixed, is normal with:

$$E(\bar{X}) = \mu$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- The **standard deviation** of this distribution measures the **variation among possible values of the sample mean**; it is called the “**Standard Error**” of the (sample) mean

# IMPLICATION OF “CLT”

- Central Limit Theorem:  $\bar{X}$  is distributed as Normal with Mean and Variance given by:

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

- which implies:

$$\Pr(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = .95$$

# CONFIDENCE INTERVAL FOR THE MEAN

- We have previously:

$$\Pr(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = .95$$

- Also:

$$-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Leftrightarrow \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \Leftrightarrow \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu$$

- Therefore:

$$\Pr\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = .95$$

# CONFIDENCE INTERVAL FOR THE MEAN

- We have:

$$\Pr(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = .95$$

- After a sample has been taken:

$$a = \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \Rightarrow \bar{x} - 1.96 \frac{s}{\sqrt{n}} \quad \text{and} \quad b = \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \Rightarrow \bar{x} + 1.96 \frac{s}{\sqrt{n}}$$

- (a,b) is called a 95% confidence interval for the Population Mean  $\mu$  ; “95%” is the Degree of Confidence, **how sure we are** that  $\mu$  is in (a,b).

# 95% “C.I.”: INTERPRETATION

- After a sample has been taken, and data summarized, a 95% confidence interval for the (unknown) population mean  $\mu$  is (a,b) where:

$$a = \bar{x} - 1.96 \frac{s}{\sqrt{n}} \quad \boxed{\text{and}} \quad b = \bar{x} + 1.96 \frac{s}{\sqrt{n}} \quad \text{are obtainable from data}$$

- (a,b) is an “interval estimate”; we are “95% sure” **that  $\mu$  is between a and b.**
- $\bar{x}$  is “point estimate”, “margin of error”:  $1.96 \frac{s}{\sqrt{n}}$

# 95% “C.I.”: INTERPRETATION

- If you take one sample, you have one 95% confidence interval (obtainable from your data).
- If you take many many samples (of the same size), you have many many 95% confidence intervals (one from each sample). Ninety five percent (95%) of these similarly constructed intervals do include  $\mu$  (and 5% of them do not).
- In real-life, **you have only one interval; yours may or may not include  $\mu$ .** Since 95% of similar intervals include  $\mu$ , you “believe” that your interval does; you are 95% sure of that.

# MORE ABOUT ESTIMATION of $\mu$

- The Population Mean  $\mu$  is unknown
- You estimate  $\mu$  by  $\bar{x}$ , a Statistic
- You maybe wrong; the margin of error is

$$1.96 \frac{s}{\sqrt{n}}$$

- That “margin of error” involved 2 components:  
(1) number 1.96 (implied by your degree of confidence 95%) and (2)  $\frac{s}{\sqrt{n}}$  called “Standard Error” of the mean.

# Interpretation

- The “standard error” (of the sample mean),

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

measures how good  $\bar{x}$  is as an estimate of  $\mu$ , the smaller the better.

Note: If you increase the sample size four times, the standard error (of the sample mean) and (margin of) error are reduced only in half.

# EXAMPLE #1

- To assess physical condition of “joggers”, a sample of  $n=25$  joggers was selected and maximum volume of oxygen ( $VO_2$ ) uptake was measured from each. The results were:  
 $\bar{x} = 47.5$  ml/kg and  $s = 4.8$  ml/k  
$$SE(\bar{x}) = \frac{4.8}{\sqrt{25}} = .96$$
- A 95% confidence interval of the mean (of the “population of joggers”) is:  
$$47.5 \pm (1.96)(.96) = (45.62, 49.38) \text{ ml / kg}$$

# EXAMPLE #2

- In the same study, a sample of  $n=26$  “non-joggers” was selected and maximum volume of oxygen ( $VO_2$ ) uptake was measured from each. The results were:  $\bar{x} = 37.5$  ml/kg and  $s = 5.1$  ml/kg:

$$SE(\bar{x}) = \frac{5.1}{\sqrt{26}} = 1.0$$

- A 95% confidence interval of the mean (of the “population of non-joggers”) is:

$$37.5 \pm (1.96)(1.0) = (35.54, 39.46) \text{ ml / kg}$$

# WANT TO COMPARE?

- A 95% confidence interval of the mean (of the “population of joggers”) is:  
**(45.46 to 49.38) ml/kg**
- A 95% confidence interval of the mean (of the “population of non-joggers”) is:  
**(35.54 to 39.46) ml/kg**
- **Are they different?**  
**Note: intervals not over-lapped!**

# Forming Confidence Intervals

- In forming confidence intervals, the degree of confidence is determined by the investigator of a research project.
- Different investigators may prefer different confidence intervals.
- The coefficient to be multiplied with the standard error of the mean should be determined accordingly.
- A few **typical choices** are 90%, 95%, or 99%; **95% is the most conventional.**

# USE OF SMALL SAMPLES

- The Procedure we just learned for forming Confidence Intervals is applicable only to larger samples. The concept starts from:

$$\Pr(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = .95$$

- Therefore, it is valid if the Population Variance  $\sigma^2$  is known or we can replace by a “good estimate” s which requires large Sample Size n.

# USE OF SMALL SAMPLES

- The Population Variance is usually unknown, we need to estimate it by  $s$ . That estimation of  $s$  may not be good when  $n$  is small; we make up for that by changing the Coefficient to be multiplied by the Standard Error (so that we still have the same likelihood of including  $\mu$  in our Interval). For example, when we form 95% Confidence Interval, we need a Coefficient larger than 1.96; the smaller the Sample Size, the larger than 1.96 the Coefficient.
- These coefficients are from the “t-distributions” indexed by the “Degree of freedom”:  $df = n - 1$ .

When “Degrees of Confidence” are specified, here is how to pick up the Coefficient to be multiplied with the Standard Error from “Tables” in texts. Degree of Freedom is  $(n-1)$ .

	Area in upper tail				
df	.10	.05	.025	.01	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
<b>Confidence</b>	<b>80%</b>	<b>90%</b>	<b>95%</b>	<b>98%</b>	<b>99%</b>

# EXAMPLE #1 REVISITED!

- A sample of  $n=25$  “joggers” was selected and maximum volume of oxygen ( $VO_2$ ) uptake was measured from each. The results were:  $\bar{x} = 47.5$  ml/kg and  $s = 4.8$  ml/k

$$SE(\bar{x}) = \frac{4.8}{\sqrt{25}} = .96$$

- A 95% confidence interval of the mean (of the “population of joggers”) is:

$$47.5 \pm (2.064)(.96) = (45.52, 49.48) \text{ ml/kg}$$

(because  $n=25$  may not be large enough to use 1.96)

# Summary on “CI”

- The population mean  $\mu$  is unknown. You estimate  $\mu$  by  $\bar{x}$ , a statistic; you may be wrong; “the margin of error” is

$$(\text{coefficient}) \frac{s}{\sqrt{n}}$$

- (1) coefficient = 1.96 (if  $n$  is large & degree of confidence is 95%); for smaller sample sizes, use “t” curve with  $df = n-1$ ;

- (2)  $\frac{s}{\sqrt{n}}$  called “Standard Error” of the mean.

- (3) Finally,  $(\bar{x} \pm \text{margin of error})$  forms a (95%) confidence interval for the Population Mean  $\mu$

# ABOUT ESTIMATION

- A Parameter is a “Numerical Characteristic” of a population (a number: Mean, Proportion, Odds Ratio). It is fixed but unknown.
- A Parameter is estimated by a Statistic; its counterpart from sample(s). A statistic is known (from data) but varies from sample to sample. It serves as “Point Estimate”; We may be wrong with a Point Estimate, but we can determine its Margin of Error.
- Putting together Point Estimate & Margin of Error we form “Interval Estimate” called a Confidence Interval; one for each Degree of Confidence

The systolic blood pressures (in mmHg) of 12 women between the ages of 20 and 35 were measured before and after administration of a newly developed oral contraceptive

A SIMPLE  
APPLICATION

Subject	Before	After	After-Before
			Difference, di
1	122	127	5
2	126	128	2
3	132	140	8
4	120	119	-1
5	142	145	3
6	130	130	0
7	142	148	6
8	137	135	-2
9	128	129	1
10	132	137	5
11	128	128	0
12	129	133	4

Calculate the 95% confidence interval for the mean systolic blood pressure *change*. Does the oral contraceptive seem to change the mean systolic blood pressure?

The next step is focused on  
“Statistical Tests of Significance”:  
what are they and why are they  
needed? This is the second area of  
“inferential statistics”, can be used to  
compare parameters (the “conviction  
phase” of the “Trial by Jury”)

What makes the Trial by jury & Statistical Tests of Significance “attractive” is that we can usually reach “a conclusion”, a verdict; a simple and clear-cut conclusion - and get the job done!

# THE TASKS IN A TRIAL

To proceed through a “Trial” - a successful one, We need the following items:

- (1) (The Constitution) & The Charge
- (2) The Investigation & Evidence
- (3) Key Evidence (called “Exhibitions”)
- (4) (Legal Guidelines) & The Verdict  
(Then, of course, the Implications)

Note; newly mentioned terms:  
“exhibition” and “the charge”; “the  
charge” is particularly important!

# THE CERTAINTY OF UNCERTAINTIES

Even Science is uncertain:

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

# SAME REASONS FOR UNCERTAINTIES

- **Variability**: Nature is complex (methods are imperfect, subjects vary, measurements fluctuate; some explainable, some not). We “measure” variability by Variance/Standard Deviation.
- **Incomplete information**: cost, time, and future targets. We rely on information gained from samples, then apply what we know to Population.

# HOW DOES SCIENCE DEAL WITH UNCERTAINTIES ?

- We form Assumption/Hypothesis: From experience & observations (This leads to the so-called research questions)
- We gather data: Experiments & Trials, Surveys, Medical Records.
- We make decision by performing Data Analysis (and that's Why you are here!)

# ELEMENTS OF GOOD RESEARCH

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

# THE TASKS IN THE “TESTING” PROCESS

To proceed through the Testing Process - a successful one, We need the following items:

- (1) A Null and an Alternative Hypotheses
- (2) The Research Design & Data
- (3) Key Statistic (called “Test Statistic”)
- (4) (Statistical Guidelines) & The Conclusion  
(Then, of course, the Implications)

## Scientific Inquiries:

Want to explain some thing such as effects of smoking or of a therapy etc.

Start with a **research question** which turns into a **hypothesis** . A **hypothesis** is a statement – any statement - about a **parameter** (or parameters). eg:

- Men are taller than women (on the average)
- Use of OC elevates blood pressure (on the average)
- Smoking leads to higher Lung cancer rate (population-wise)

RESEARCH  
QUESTION



HYPOTHESIS

# HYPOTHESES

- Again, a “Hypothesis” is a Statement about a Distribution (“SBP is Normally distributed”), or its underlying parameter(s) (“mean  $\mu = 120$ ”), or about the relationship between Distributions (“Fat intake and Cholesterol level are related”).
- A **Hypothesis is just a statement**, usually by an investigator- but could be by anyone, true or false, with or without supports.

# NULL HYPOTHESIS

- Among the numerous possible hypotheses involved in a problem, there is a very special one- called the “Null hypothesis” and is denoted by  $H_0$ .
- The Null Hypothesis  $H_0$  is the counterpart of the Constitution statement stipulating “Innocence”. For example, when a researcher is concerned about the relationship between Oral Contraceptive (the “Pill”) and SBP; it is about the Means of two Populations: the populations of OC users and of OC non-users. The underlying Null Hypothesis is “ $H_0: \mu_1 = \mu_2$ ”.

# HYPOTHESIS TESTS

- A “Hypothesis Test” is a Decision-making Process that examines a set or sets of data and, on the basis of expectation under  $H_0$ , leads to a decision to “reject” or not to reject  $H_0$ .  $H_0$  is “rejected” (Guilty!) if the data show overwhelmingly that it is almost impossible to have the data that we already collected if  $H_0$  is true.
- Hypothesis Tests are also called “Tests of Significance”; the term “significant” only means “real”; the conclusion that, say, an observed difference is real- not happened “by chance”.

# REJECTION

- The Null Hypothesis  $H_0$  is a “Theory”. The Data are “Reality”. When they do not agree, then **we have to trust the reality**; That’s when  $H_0$  is rejected.
- How do we tell if Theory and Reality do not agree? When the data show overwhelmingly that it is almost impossible to have the data that we already collected if  $H_0$  is true (that it is possible but with a very small probability).

# ABOUT REJECTIONS

**Hypothesis  
Testing is similar  
to Trial by Jury**

A very important concept: when a null hypothesis is not rejected it does not necessarily lead to its acceptance, because a “not guilty” verdict is just an indication of “lack of evidence” and “innocence” is just one of its possibilities. That is, when a difference is not statistically significant, **there are still two possibilities:**

- (i) The null hypothesis is true,
- (ii) There is not enough evidence from sample data to support its rejection (i.e. sample size may be too small).

# ALTERNATIVE HYPOTHESIS

- The “Alternative Hypothesis”  $H_A$  is the counterpart of the “Charge” in a Trial by Jury (e.g. first-degree murder). It is important affecting the decision; the jury may see that the suspect is “kind of guilty” but the charge is more wrong, too severe that they may vote to acquit him/her.
- In the context of a research project, say, about the relationship between Oral Contraceptive and SBP with  $H_0: \mu_1 = \mu_2$ ; a possibility is  $H_A: \mu_1 > \mu_2$ .
- To a researcher,  $H_A$  is his/her (primary) Hypothesis.

# THE CHOICE

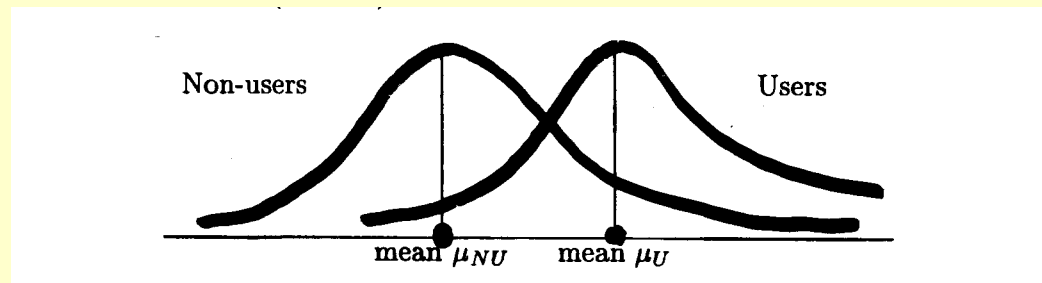
- In deciding whether to reject or not to reject a Null Hypothesis, the choice is not between  $H_0$  and “The Truth”; **because the Truth may not be relevant.**
- The choice is between  $H_0$  and  $H_A$ ; if there are enough data to support  $H_A$  then  $H_0$  is rejected.
- There are two forms/types of Alternatives: (1)  $H_A: \mu_1 > \mu_2$  is “**one-sided**” Alternative, (2)  $H_A: \mu_1 \neq \mu_2$  is a “**two-sided**” Alternative; For example, from the data  $\bar{x}_1 < \bar{x}_2$  showing that  $H_0$  may be wrong but the Alternative  $H_A: \mu_1 > \mu_2$  is even more wrong;  $H_0$  is not rejected (data support “the other side”!).

# TEST STATISTIC

- Test Statistic is the one key piece of evidence (summarized data) that measures the difference between the data and what is expected under the Null Hypothesis; so that if it is large, we would lean to rejecting the Null Hypothesis.
- For example, consider  $H_0: \mu = 100$ , an obvious choice for the test statistic is  $(\bar{x} - 100)$ .
- However, this evidence is “statistical evidence”, it varies from sample to sample. That’s why other statistic, such as variance, is still needed- even though the Null Hypothesis is about Mean  $\mu$ .

# Variability & Errors

In some medical cases such as infections, the presence or absence of bacteria and viruses are easier to confirm correctly. In other cases, it's not clear-cut. One possible model for these situations would be to think of the blood pressure  $X$  is distributed with differences means for users (U) and nonusers (NU). It can be seen from the figure below that errors are unavoidable when the two means  $\mu_{NU}$  and  $\mu_U$  may be close



# Errors

In making a decision concerning the Null Hypothesis to compare  $\mu_U$  versus  $\mu_{NU}$ , **errors are unavoidable**. Since a null hypothesis  $H_0$  may be true or false and our possible decisions are whether to reject or not to reject it, there are four possible outcomes combinations. Two of the four outcomes are correct decisions:

(i) not rejecting a true  $H_0$

(ii) rejecting a false  $H_0$

but there are also two possible ways to commit an error:

Type I: a true  $H_0$  is rejected

Type II: a false  $H_0$  is not rejected

# Types Of Errors

Truth	$H_0$ not rejected	$H_0$ is rejected
$H_0$ is true	Correct Decision	Type I Error
$H_0$ is false	Type II Error	Correct Decision

$$\alpha = \text{Pr}(\text{Type I Error})$$

$$\beta = \text{Pr}(\text{Type II Error})$$

Aim is to keep  $\alpha$  and  $\beta$ , the probabilities in the context of repeated sampling – types I and II errors respectively, as small as possible. However, if resources are limited, this goal requires a compromise because these actions are contradictory; **We fix  $\alpha$  at some specific conventional level- say .05 or .01 and  $\beta$  is controlled through the use of sample size(s).**

# ANALOGIES

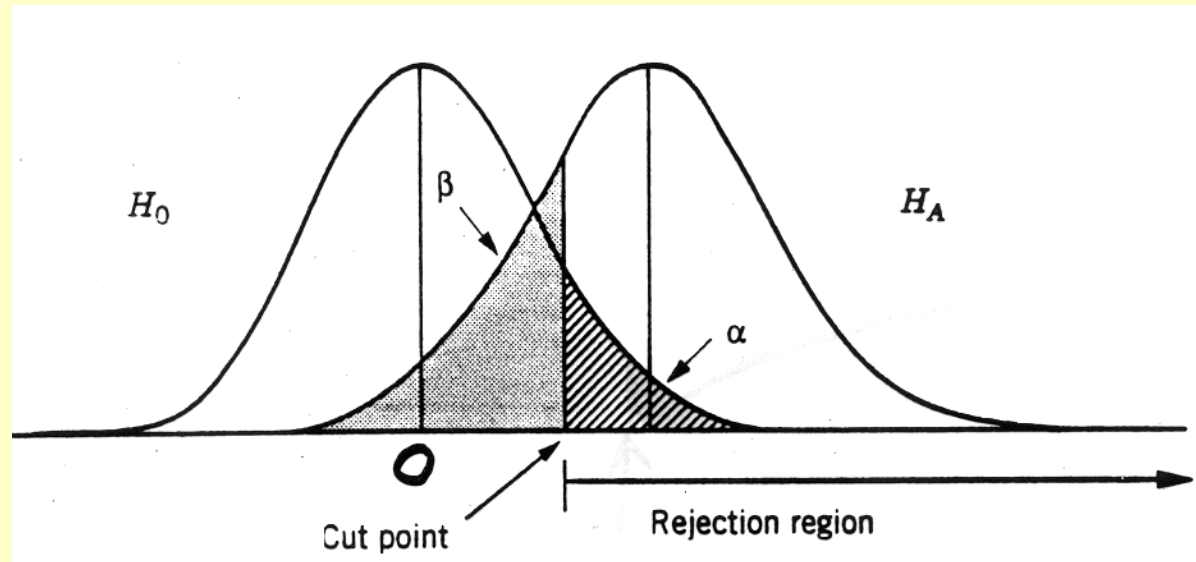
- **Type I error: Convicting an innocent man** (top priority: to keep the probability of committing this error low)
- **Type II error: Acquitting a guilty suspect** (also important but controlled earlier in the process, i.e. making sure to have enough evidence for a conviction by a thorough investigation. In research, that lies in the “design” stage, making sure to have a large study to collect enough data.

# DECISION

- Most common approach is the formation of Rejection/no-Rejection Regions. For example, the hypothesis is: Men are taller (on the average)
- Example:  $H_0: \mu_M = \mu_F$  &  $H_A: \mu_M > \mu_F$
- Summarized Evidence:  $\bar{x}_M - \bar{x}_F$ ; a large value of this supports  $H_A$ , leading to a rejection of  $H_0$ .
- The region of large values is called “Rejection Region”, the remaining part (that of small values) is the Non-Rejection Region.

(1) One-sided Alternative: Rejection Region is on one side, could be left or right side (see figure below).

(2) Two-sided Alternative: Rejection Region is on both sides; Test Statistic value is away from its hypothesized value, e.g. zero.



# p-VALUE

- An Alternative to the use of ‘Rejection Rule’ is to summarize data into a number called “p-value”.
- A p-value is the probability, that if  $H_0$  is true, to have an observed Test Statistic as far away from its hypothesized value, say zero, as the one we have.
- **Implication:** very small p-values imply that  $H_0$  is likely not true- so,  $H_0$  should be rejected.
- P-value can be viewed as a “measure of compatibility” between a Null Hypothesis (theory) and Data (reality); small values lead to rejection.

# COMMON INTERPRETATION/EXPRESSION (About p-Values)

- $p > .10$ : Result is not significant
- $.05 < p < .10$ : Result is marginally significant
- $.01 < p < .05$ : Result is significant
- $p < .01$  : Result is highly significant

# Calculating a p-value

- Most of the times, it is very hard- or impossible- to calculate a p-value by hand, even with the help of “statistical tables” in books and/or a powerful calculator.
- So, practice again again to gain enough knowledge, experience, and confidence in the use of **Excel** spreadsheet computer program – or any other more advanced products like SAS

# COMPARISON OF MEANS

- FOCUS: (Pop Mean of) Continuous Endpoint
- Often involved one or two groups of subjects
- PROBLEMS belong to one of three types:
  - (1) One-sample (versus Standard/Referenced)
  - (2) One-to-one matched sample
  - (3) Two independent samples
- Final Products: “t-tests”; one-sample and two-sample t-tests

# ONE-SAMPLE OF CONTINUOUS DATA

The Null Hypothesis considered is  $H_0: \mu = \mu_0$  and the data  $(n, \bar{x}, s^2)$ ;  $n$  being the sample size,  $\bar{x}$  the sample mean, and  $s^2$  the sample variance ; the test needed is the one-sample t-test.

## PAIR-MATCHED DATA

Let  $\{d_i\}_{i=1, \dots, n}$  be the sample of differences from the  $n$  pairs; this sample is summarized into the mean  $d$  and standard deviation  $s_d$ , respectively.

This is a special case of the “one-sample problem (i.e. with  $\mu_0 = 0$ ):

# COMPARISON OF TWO POPULATION MEANS

- In this type of problems, we have two independent samples  $(n_1, \bar{x}_1, s_1^2)$  and  $(n_2, \bar{x}_2, s_2^2)$ ; the  $n$ 's being the sample sizes- may be different sizes, the  $\bar{x}$ 's the sample means, and the  $s^2$ 's the sample variances (the  $s$ 's are standard deviations).
- The Null Hypothesis considered is  $H_0: \mu_1 = \mu_2$  or equivalently,  $H_0: \mu_2 - \mu_1 = 0$ .

# GENERAL APPROACH

- In general, the Null Hypothesis of a “Statistical Test” is concerned with a Parameter or Parameters (Population Proportion, Population Mean, or Coefficient of Correlation). In the current problem, the Difference of 2 Population Proportions:  $\mu_2 - \mu_1$ .
- Sample data are summarized into a Statistic which is used to estimate the Parameter under investigation. Therefore, in the current problem, we focus on the difference of two sample means  $\bar{x}_2 - \bar{x}_1$ .

# GENERAL APPROACH

- We have a Parameter,  $\mu_2 - \mu_1$ , involved in the Null Hypothesis and its “Estimator”,  $\bar{x}_2 - \bar{x}_1$ .
- The next step is to measure the distance from the “observed value” of the estimator (representing “reality”) to its hypothesized value under  $H_0$  (representing the “theory”). In the current problem, it is the difference between  $\bar{x}_2 - \bar{x}_1$  and  $\mu_2 - \mu_1 = 0$ ; if the “discrepancy” is larger than what can be explained (by chance), then we have to “trust” the reality and reject the theory. That is to reject  $H_0$ .

# GENERAL APPROACH

- We are measuring the “distance” from the estimate of a Parameter and its hypothesized value under  $H_0$ .
- An estimate is a Statistic which is itself a Variable (in the context of repeated sampling, its value varies from sample to sample). In that sampling distribution the variation (representing its “reproducibility”) of the Statistic is measured by its “Standard Error”.
- The “distance” between Statistic & its hypothesized value under  $H_0$  is “converted” to a standard unit: “Number of standard errors” that the Statistic is away from its hypothesized value under  $H_0$ .

# GENERAL APPROACH

- We are measuring the “distance” from the estimate of a Parameter (a Statistic) and its hypothesized value under  $H_0$  and expressed it as the “Number of standard errors” of that Statistic.
- If the Statistic involved has “Normal” as its sampling distribution (in this case, this is backed by the CTL if the  $n$ 's are large); the above “Number of standard errors” is on “the Standard Normal scale which we can determine how likely to occur under the assumption that is true. The larger the “Number of standard errors” the less likely that  $H_0$  is true.

# TWO-SAMPLE t-TEST

- **Null Hypothesis**  $H_0: \mu_1 = \mu_2$ , or  $H_0: \mu_2 - \mu_1 = 0$ .
- **Data & Test statistic:** 2 independent samples of data  $\bar{x}_1(n_1, s_1^2)$  and  $\bar{x}_2(n_2, s_2^2)$ ; Standard Error & “t” Test Statistic:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{SE(\bar{x}_2 - \bar{x}_1)}$$

- The statistic  $\bar{x}_2 - \bar{x}_1$  is “t standard errors away from its hypothesized value” of 0; This “t” is on the Standard normal scale if the n’s are large and “t-scale” with  $(n_1 + n_2 - 2)$  dfs if the n’s are not large.

# TWO-SAMPLE WITH CONTINUOUS DATA

- The Null Hypothesis considered is  $H_0: \mu_1 = \mu_2$ , or  $H_0: \mu_2 - \mu_1 = 0$ . The statistic  $\bar{x}_2 - \bar{x}_1$  is “t standard errors away from its hypothesized value” of 0 :

$$t = \frac{\bar{x}_2 - \bar{x}_1}{SE(\bar{x}_2 - \bar{x}_1)}$$

- There are two ways to form a decision:
  - (1) Choose a level of Type I error, form a Rejection Region, then decide whether or not  $H_0$  is rejected,
  - (2) Summarize the finding into a “p-value”

# P-VALUE

- Instead of saying that an observed value of the test statistic is significant (i.e., falling into the rejection region for a given choice of  $\alpha$ ) or is not significant, many writers in the research literature prefer to report findings in terms of a p-value.
- The p-value is the probability of getting values of the test statistic as extreme as, or more extreme than, that observed if the null hypothesis is true. For the current problem, it is the Area to the “left” of  $t$  for  $H_A: \mu_2 < \mu_1$  & the area to the right of  $t$  for  $H_A: \mu_2 > \mu_1$  and it is the area “beyond  $\pm t$  for two-sided  $H_A: \mu_2 \neq \mu_1$  where the degree of freedom is  $(n_1 + n_2 - 2)$ .

# EXAMPLE

- Data in epidemiologic studies are sometimes self-reported. The following table gives the percent discrepancy between self-reported and measured height:

$$x = [(\text{self-reported} - \text{measured})/\text{measured}] 100\%$$

	Men			Women		
Education:	n	& mean	& SD	n	& mean	& SD
H.school:	476	& 1.38	& 1.53	& 323	& .66	& 1.5
College:	192	& 1.04	& 1.31	& 62	& .41	& 1.46

# EXAMPLE #1

- Focusing on men with high-school education, we have: (n=476, mean=1.38, st. deviation=1.53)
- The difference (between means of self-reported and measured height) is significant at the .05 level; two-sided p-value < .001.
- It's the one-sample t-test

# EXAMPLE #2

- Comparing Men versus Women, both groups with High-school education, we have:
- The difference is significant at the .05 level; two-sided p-value  $< .001$
- It's the two-sample t-test