PubH 7405:
REGRESSION ANALYSIS

Instructor: Chap T. Le, Ph.D.

Distinguished Professor of Biostatistics

Review #1:

Biostatistics & Statistical Inference

GOAL OF FIRST LECTURE

- **Overview of the course, its objectives and contents, & its organization .**
- **A brief review of "Biostatistics & Statistical Inference"**
- **There are two more review sessions, one on Simple Regression and Correlation, and one on Introductory Experiment Design.**
- **These reviews are rather non-mathematical; there are homework assignments but they may or may not be linked directly to the three review lectures; they are aimed to review basic methods & computer implementation (t-test, ANOVA, Simple Regression & Correlation, & SAS).**

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.

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).

THE IMPORTANT PHASE

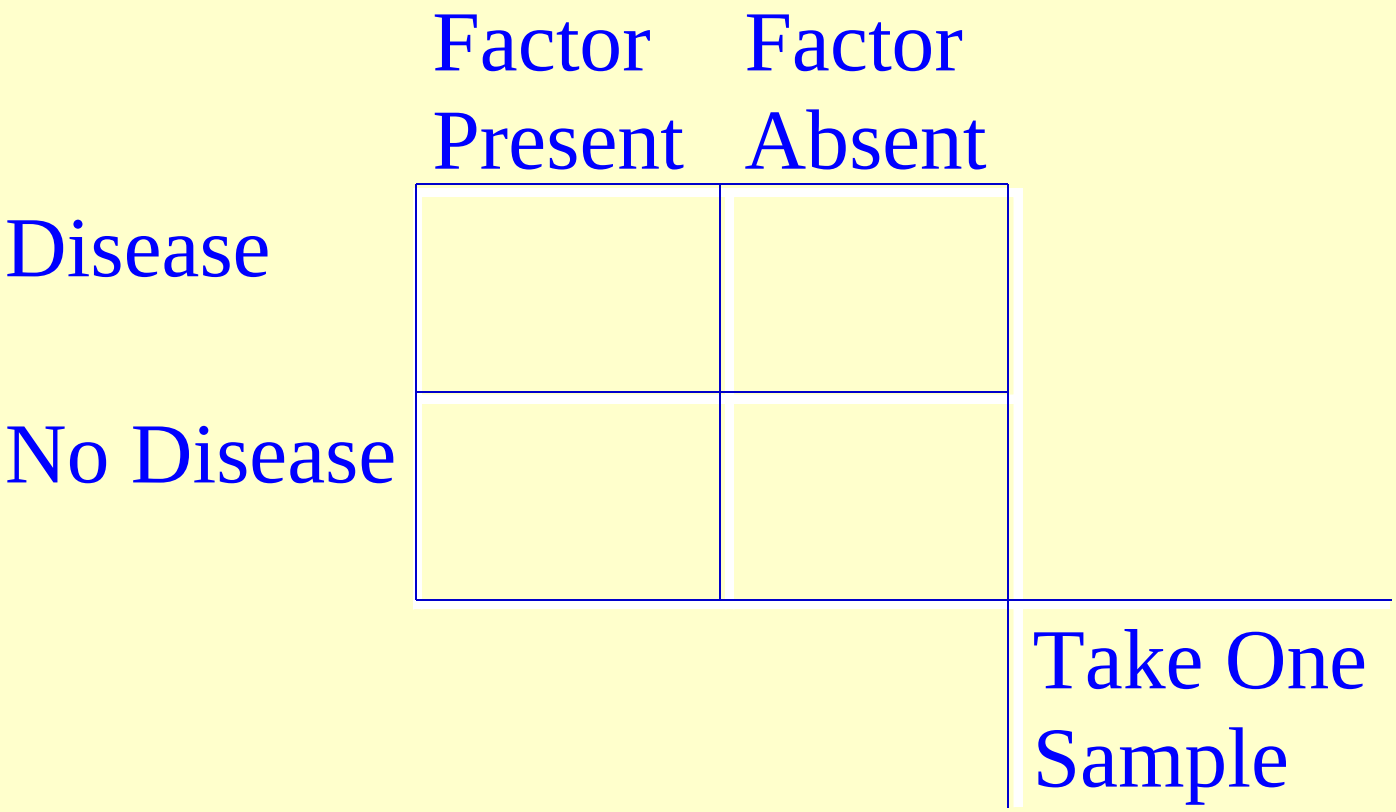
Just as in the case of "Trial by Jury", the most **important stage** of the "Research Process" is the DESIGN: How & How Much data are collected! Also, It dictates how data should be analyzed. **May be it's not the question of "how" to collect your data but the decision on "when to do what"!**

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



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

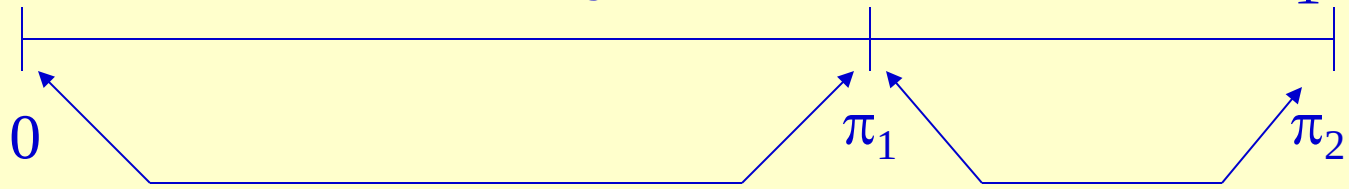
Retrospective Studies gather past data from selected cases (with disease) and controls (without disease) to determine differences, if any, in exposure to a suspected risk factor.
Advantages: Economical & Quick. **Major Limitations:** Accuracy of exposure histories & Appropriateness of controls

A CLINICAL TRIAL

Study Initiation

Study Termination

No subjects enrolled after π_1



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 π_1 & π_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.

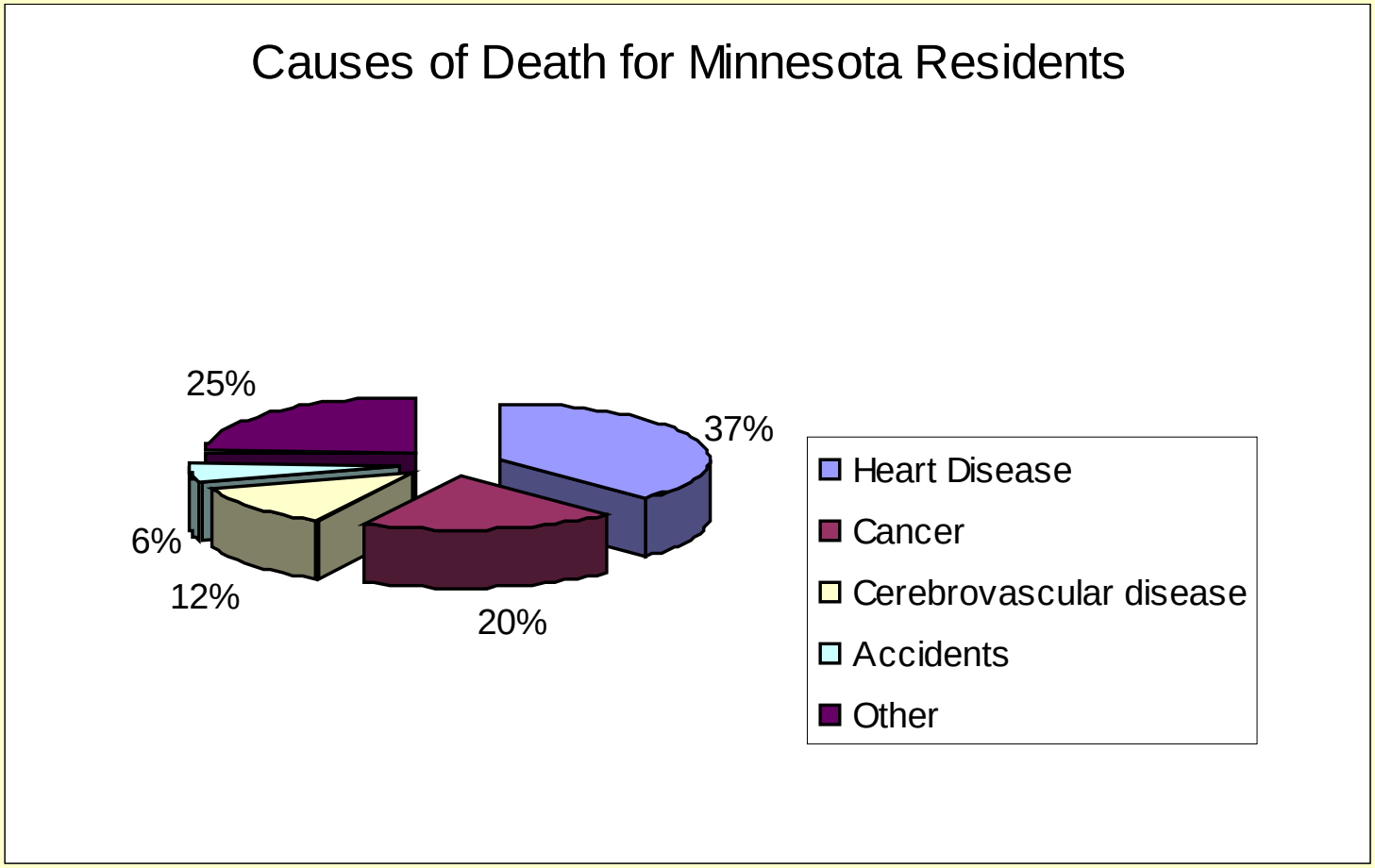
Aims: To communicate **more** effectively

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.



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

Statistical Inference

- **The last step of the data analysis process is 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 allows us to estimate a parameter (e.g. smoking rate, disease prevalence).

(2) Hypothesis Testing allows us to test hypotheses, i.e., to compare parameters (as in treatment evaluation – clinical trials - or evaluation of public health intervention programs).

SAMPLING DISTRIBUTION & STANDARD ERRORS

VARIABLE & DISTRIBUTION

- **A function or rule that maps or associates with each element in a domain (e.g. outcome of an experiment) a number is called a variable.**
- **A list of possible values of a variable, together with their corresponding probabilities, is called the distribution of that variable.**

VARIABLES IN ACTION

- In applications, a variable represents a characteristic or a class of measurement. It takes on different values on different subjects/persons. Examples include weight, height, race, sex, SBP, etc. The observed values, or observations, form items of a data set.
- On the micro scale, depending on the scale of measurement, we have different types of data (continuous, categorical, ordinal).
- On the macro scale, we have observed variables and calculated variables; a calculated variable is a statistic.

SAMPLING DISTRIBUTIONS & STANDARD ERRORS

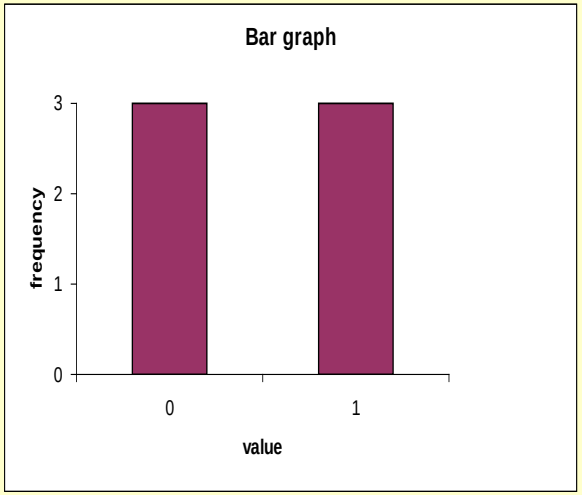
The distribution of a calculated variable or statistic, across all possible samples, is called a Sampling Distribution. We have, for example, sampling distribution of the mean and sampling distribution of proportion. The Standard Deviation of a sampling distribution is called the Standard Error of the corresponding statistic. The term "error" is used perhaps to emphasize the role of the statistic as an estimate/estimator.

Example 1: A Hypothetical Population

- 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

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

The mean of all possible sample means:

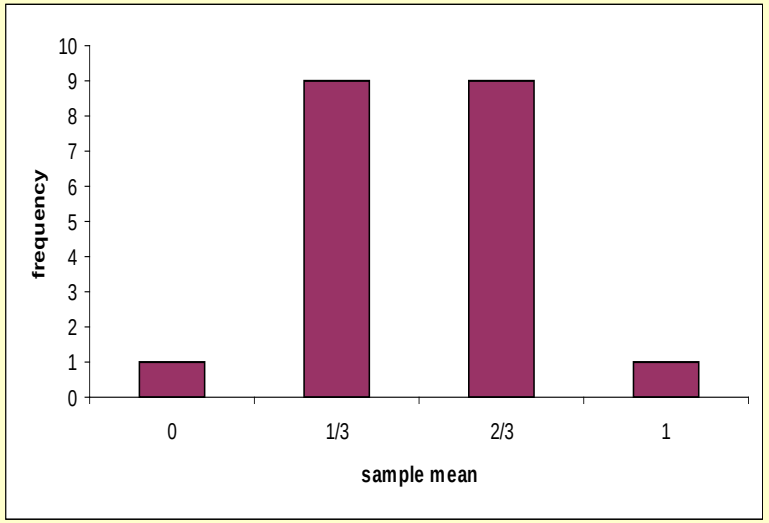
$$\mu_{\bar{x}} = \frac{1(0)+9(1/3)+9(2/3)+1(1)}{20} = 0.5\,(=\mu)$$

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!

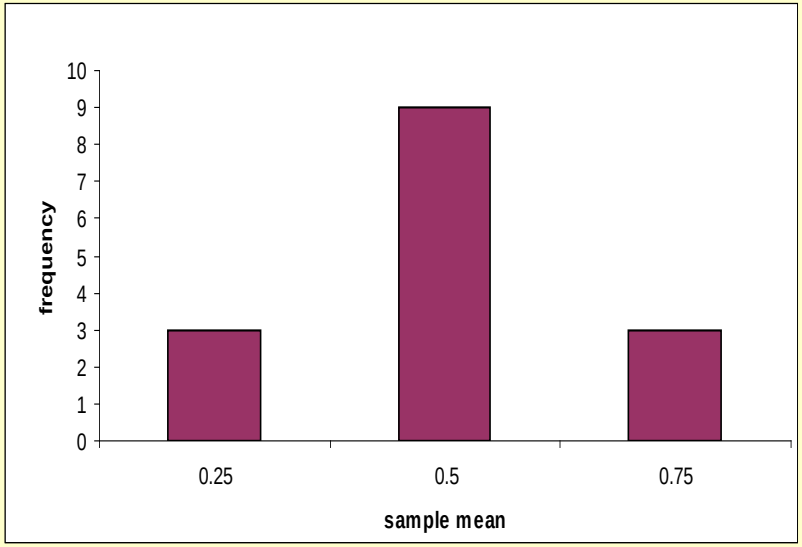


Increase the value of "n"

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)		
(A, B, C, D), (A, B, C, E), (A, B, C, F)	3	0.75
Total	15	

- If $n = 4$, the mean of all sample means is still 0.5.
- The shape is even more normal.

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



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

Example 2:
A Larger Population

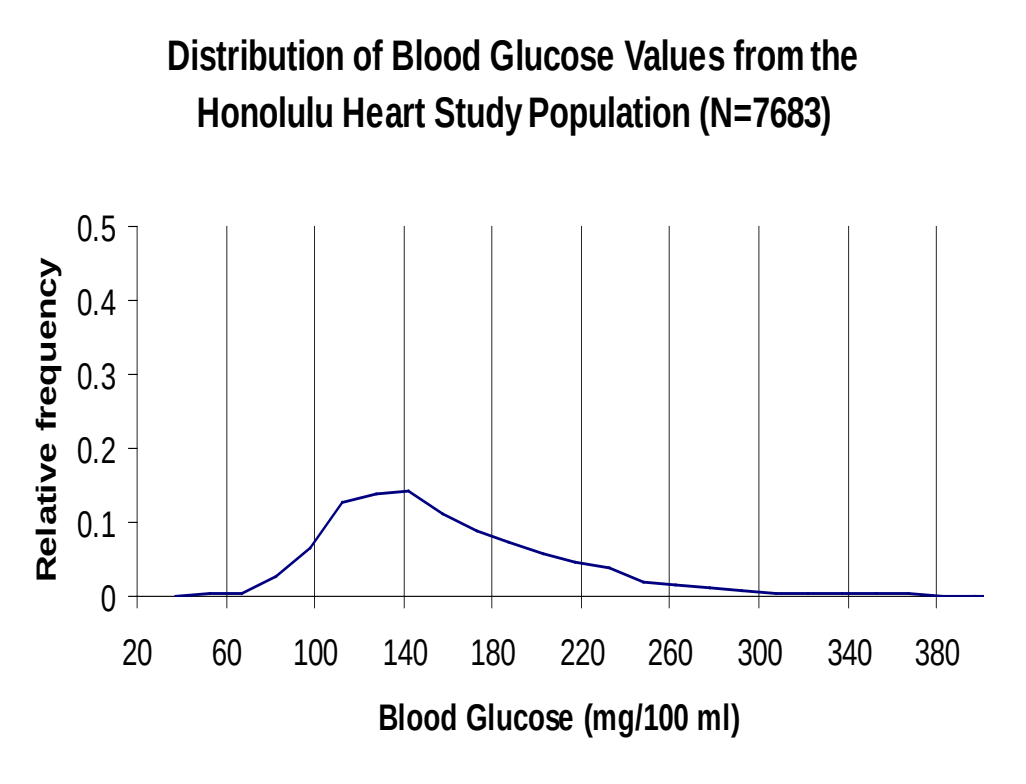
- **Blood glucose measurements from 7,683 men in Honolulu.**
- **Take 400 samples, 25 each. Sample means shown at left.**
- **Means of two distributions are approximately the same (There are many more than 400 possible samples).**
- **Variance of the distribution of sample means is smaller.**

Bllod 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 (Population) Blood Glucose Values

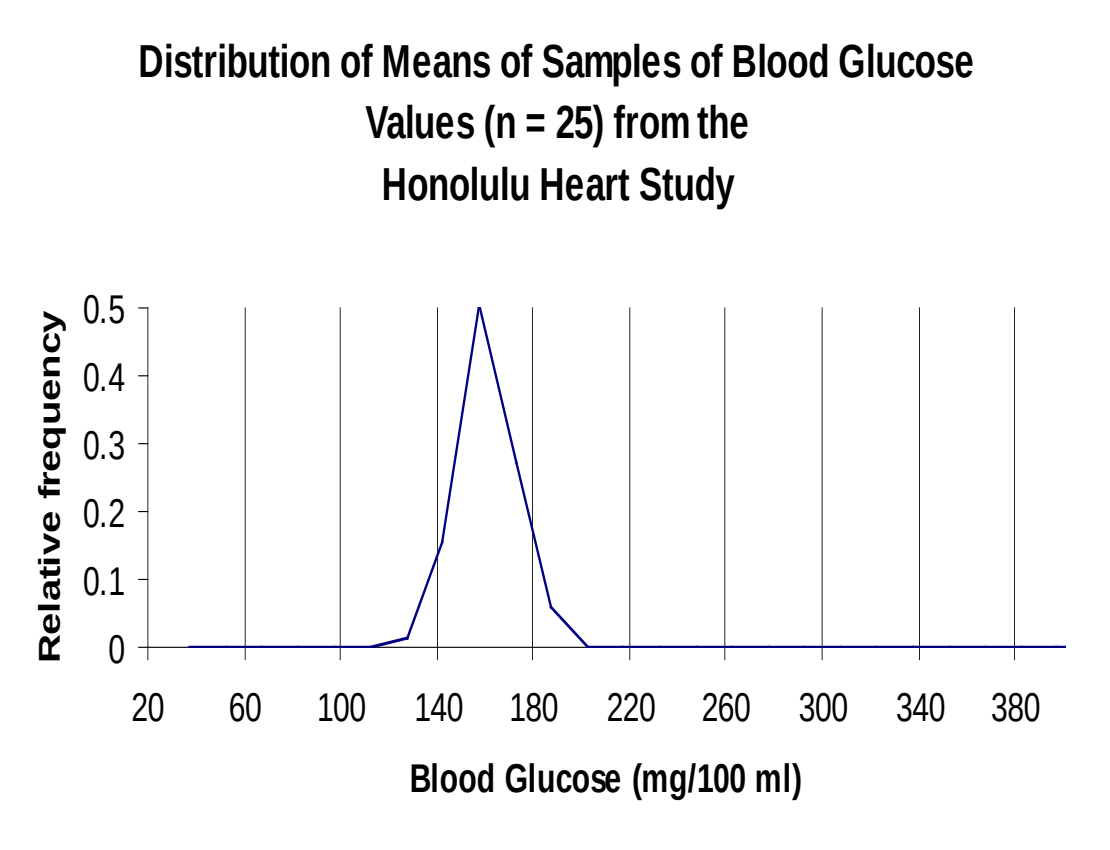
$\mu = 161.52$

$\sigma = 58.15$



Population distribution is not even symmetric!

Distribution of 400 Sample Means



μ = 160.66

σ = 12.24

The sampling distribution is a bit more symmetric & more normal!

CENTRAL LIMIT THEOREM

- **Given any population with Mean μ and Variance σ^2 (Standard Deviation σ): The Sample Mean is a "variable"; the (sampling) distribution of its possible values, with (large) sample size n being fixed, is normal with:**

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

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

- **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**
- **You have just seen two examples/illustrations**

Sample Proportion “p” is a special case of the Sample Mean (where measurements or sampled values are 0’s and 1’s if we use “1’ for success/presence and “0” for failure/absence). Therefore, Central Limit Theorem applies.

Sampling Distribution of Sample Proportion

- Let π be a "population proportion", the Sample Proportion is a "variable"; the (sampling) distribution of its possible values, with (large) sample size n being fixed, is normal with:

$$E(p) = \pi$$

$$Var(p) = \frac{\pi(1-\pi)}{n}$$

- The standard deviation of this distribution measures the variation among possible values of the sample proportion; it is called the "Standard Error" of the (sample) proportion.

Can we “find” Sampling Distributions in action?
How to “see” the impact of Sample Size n?

If one counts deaths from brain cancer, one should find more of them in California, Texas, New York, and Florida. Are these places unsafe? Not necessarily, these states have the most brain cancer because they have the most people. There are more people, there are more people with cancer – of any kind.

So, it's better to study rates: deaths as a proportion of total population.

Using proportion makes for a very different leaderboard. South Dakota takes the first place with 5.7 brain cancer deaths per 100,000 people per year (in 2008, compared to the national rate of 3.4). South Dakota is followed on the ranked list by Nebraska, Alaska, Delaware, and Maine. Are these states unsafe that you should avoid? Neighbors South Dakota and Nebraska suggest something?

Wait!
Scrolling down to the bottom of the list, you would find Wyoming, North Dakota, Hawaii, and the District of Columbia; Vermont is nearby in this end.
Why should South Dakota be prone to brain cancer and North Dakota nearly tumor free? Why would you be safe in Vermont and in trouble in its neighbor, Maine?

The five states at the top have something in common, and the five states at the bottom do, too. And it's the same thing in both ends: small states, small population sizes.

Why size matter?

Remember the sampling distribution of proportion; its variance is $\pi(1-\pi)/n$. The smaller n, the larger the variance, the more the proportion value swings (to both small and large ends).

Here is another example. If you rank all NBA players by shooting efficiency, you would find “bench warmers” at both ends. They took only a few shots a year; some make all or nearly all shots (100% or near 100%) and some missed all or nearly all shots (0% or near 0%). The NBA restrict the rankings to players who’ve reached certain threshold of playing time (This helps to improve but not eliminating possible problem)

And not everyone, every system are quantitative savvy. Many states institute incentive programs for schools that do well on standardized tests. For example, schools are ranked on the improvement of student test scores. Who win this kind of contest? Mostly smaller schools. One can argue that at smaller schools, teachers know the students and their families, and have time to craft and deliver individualized instruction. The fact is there are smaller schools at the other end of the ranking as well. The smaller n, the larger the variance, the more the proportion value swings to both small and large ends.

TEST OF SIGNIFICANCE
OR
HYPOTHESIS TESTING

Many scientific questions can be boiled down to a yes or no answer: Is something going on, or not? Does a psychological intervention make you happier or does it do anything at all? The "does nothing" scenario is called the "*Null Hypothesis*". That is, the Null Hypothesis is the hypothesis that the new Drug you're studying has no effect.

If you're the researcher who developed the new drug, the Null Hypothesis is the thing that keeps you up at night – *unless you can rule it out.*

The standard framework, called the Test of Significance, was developed by R. A. Fisher, the founder of the modern practice of statistics in the early twentieth century. It is an analog of the "Trial by Jury"

It goes like this. First you run an experiment. You might start with one hundred subjects, randomly select half to receive the New Drug while the other half gets a Placebo. Your hope is that the patients on the study drug will be less likely to die than the one getting the sugar pill.

From here, the protocol might seem simple: If you observe fewer deaths among the drug patients than the placebo patients, you would declare victory and file an application with FDA, right?

That's wrong! It's not enough to say that the data be consistent with your theory (called the Alternative Hypothesis); they have to be inconsistent with the negation of your theory, the Null Hypothesis. The Null Hypothesis is the hypothesis being tested, data are reality. If they are inconsistent, you must trust reality and reject the Null Hypothesis.

Let make this numerical.

Suppose we're in "the Null Hypothesis land", where the chance of death is exactly the same (say, 10%) for both groups. Let consider all the possibilities and compute the (binomial) probabilities and added up the three scenarios; we have:

13.3%: equally many drug and placebo patients die

43.3%: fewer placebo patients than drug patients die

43.3%: fewer drug patients than placebo patients die

Does your observation (result) of fewer deaths among drug patients insistent with the Null Hypothesis? Not quite, it could happen with 43.3% chance!

Seeing better results among the drug patients than the placebo patients says very little since this isn't at all unlikely, even under the Null Hypothesis that the drug doesn't work. You must show that the drug patients do a lot better to rule out the chance occurrence, to rule out (i.e. to reject) the Null Hypothesis.

So, here's the procedure for ruling out the Null Hypothesis:

(1) Run an experiment,
(2) Suppose the Null Hypothesis is true, and let "p" be the probability (under that Null hypothesis) of getting results as extreme as those observed,
(3) The number "p" is called the "p-value". It is a measure of compatibility between the Null Hypothesis (a theory that you assume) and data (the reality). If it is small, you're happy and say your results are "statistically significant".

How small is very small?
There is no magic threshold; a conventional choice for the threshold is 0.05 or 0.01.

Test of significance is popular because it captures our intuitive way of reasoning about uncertainty.
But we start with " Suppose the Null Hypothesis is true" while what we're trying to prove is that the Null Hypothesis isn't true. This is a very common line of logic: defeating a hypothesis by means of its own force; that a hypothesis implies a falsehood, then the hypothesis itself must be false.
In addition, tests of significance are popular because, using theses tests, you get the job done: You make a decision!

Now, let put the process in the context of a basic, popular procedure, the two-sample t-test.

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)

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**

COMPARISON OF TWO POPULATION MEANS

- In this type of problems, we have two independent samples (n_1, x_1, s_1^2) and (n_2, x_2, s_2^2); the n's being the sample sizes- may be different sizes, the 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", $x_2 - x_1$.
- 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 that 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 statistic 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 $(n_1, \bar{x}_1, s_1^2)$ and $(n_2, \bar{x}_2, s_2^2)$; 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) degrees of freedom if the n's are not large.**

DECISION

- **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) 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 are "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 = [(self-reported - measured)/measured] 100%

	Men			Women		
Education:	**n**	**mean**	**SD**	**n**	**mean**	**SD**
H. school:	**476**	**1.38**	**1.53**	**323**	**.66**	**1.50**
College:	**192**	**1.04**	**1.31**	**62**	**.41**	**1.46**

EXAMPLE

- **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**

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**

CONFIDENCE INTERVAL

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!

Anything's wrong with "Test of Significance"? To start with, that's the word itself: "significance".
In common language, it means something like "important" or "meaningful". But the test of significance that scientists use doesn't measure importance. When we're testing the effect of a new drug, the Null Hypothesis stipulates that there is no effect at all; so to reject the Null Hypothesis is merely to make the conclusion that the effect of the drug is not zero (that what we see is "real", not by chance). But the effect could still be very small – so small that the drug isn't effective in any sense that an ordinary person would call "significant" or "importance": a possible "insignificance of being significance"

Secondly, the Null Hypothesis – any null hypothesis, if we take it literally, is probably just about always false. When you drop a powerful drug into a patient's bloodstream, it's hard to believe that it has exactly zero effect.

Let make this numerical and suppose we are investigating the relationship between cigarette smoking (say, binary) and marriage; seemingly two unrelated "variables". Saying that "marital status" and "smoking status" are independent/uncorrelated (a Null Hypothesis) is simply to say that, in the population, the smoking rate of married people is the same as the smoking rate of unmarried people.

Now we see the problem: the chance is very small, for any population, that the smoking rate for married people and the smoking rate for unmarried people are exactly the same. Any null hypothesis is probably always false; everything are correlated with everything else.

A significance test is a scientific instrument; and like any other instrument, it has a certain degree of precision; by increasing the size of the studied sample, for example, you enable yourself to see ever-smaller effects. That's the power of the method, statistical power, but also its danger: the true is the hypothesis is probably always wrong. Therefore, it is much more important to know how wrong is a Null Hypothesis, how large is a drug effect, how strong is a correlation.

One simple needed strategy is to report confidence intervals in addition to p-values.

PARAMETER ESTIMATION

- **Process: Research Question leads to Endpoint, then Parameter of Interest; For example, Unemployment Rate which is an unknown number between 0 and 1 (or 100%).**
- **Question: How to estimate it ?**

(If Testing Hypothesis is the analog of Trial by Jury, Parameter Estimation is the Sentencing Phase; but we'll learn this phase first)

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 (from where the boy is.**

STORY OF UNEMPLOYMENT RATE

- **Lost "mother": True Unemployment Rate (say, in the state of Minnesota in September of 2016)**
- **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)

IMPLICATION OF "CLT"

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

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

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

- **which implies:**

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

CONFIDENCE INTERVAL FOR THE MEAN

- **We have previously:**

- **Also:**

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

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

- **Therefore:**

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

CONFIDENCE INTERVAL FOR THE MEAN

- **We have:** $\Pr(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = .95$

- **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 μ ; "95%" is the Degree of Confidence, how sure we are that μ 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 μ is (a,b) where a & b are obtainable from data:**

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

- **(a,b) is an "interval estimate"; we are "95% sure" that μ 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 (from your data).**
- **If you take many samples (of the same size), you have many 95% confidence intervals (one from each sample). Ninety five percent (95%) of these similarly constructed intervals do include μ (and 5% of them do not).**
- **In real-life, you have only one interval; yours may or may not include μ. Since 95% of similar intervals include μ, you "believe" that your interval does; you are 95% sure of that.**

MORE ABOUT ESTIMATION of μ

- **The Population Mean μ is unknown**
- **You estimate μ by 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.**

EXAMPLE #1

- **To assess physical condition of "joggers", a sample of n=25 joggers was selected and maximum volume of oxygen uptake was measured ~~f~~rom each. The results were: 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:**

$$47.5 \pm (1.96)(.96) = (45.62, 49.38) ml/kg$$

EXAMPLE #2

- **In the same study, a sample of n=26 "non-joggers" was selected and maximum volume of oxygen (VO2) uptake was measured from each. The results were: 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:**

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

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 <u>applicable only to larger samples</u>. The concept starts from:**

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

- **Therefore, it is valid if the <u>Population Variance σ^2 is known</u> or we can replace by a "good estimate" s which requires <u>large Sample Size n</u>**

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 the we still have the same likelihood of including μ 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.**

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 counter part 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

Confidence Interval versus Test of Significance:

The confidence interval tells you a lot more; it is even informative where you don't get a statistically result.

Let suppose you want to estimate some drug effect:

(1) If the confidence interval is [-0.5%,0.7%], then the reason you didn't get statistical significance is because you have good evidence the intervention doesn't do much anything;

(2) If the confidence interval is [-20%,22%], the reason you didn't get statistical significance is because you have no idea whether intervention has an effect, or which direction it goes.

Those two outcomes look the same from the viewpoint of significance test but have quite different reasons looking at confidence intervals: the first one showed a very small effect and the second one caused by a small sample size.

DUE AS HOMEWORK

None for today.

From the next lecture, there are two homework problems each day; Assignments for Monday and Wednesday are both due at the following week's recitation session & returned a week later – only in the labs; no late homework is accepted.