Hypothesis Testing

And the

Central Limit Theorem



## Central Limit Theorem

A very important result in statistics that permits use of the normal distribution for making inferences (hypothesis testing and estimation) concerning the population mean



## Central Limit Theorem

- If a variable X (with any distribution) has a population mean 💻 and standard deviation ☞, then: the distribution of sample means (from samples of size n taken from the population), has the following distribution as n tends to infinity:

$$\overline{X}_n \sim n(\mu, \frac{\sigma}{\sqrt{n}})$$



## Central Limit Theorem

- Insert CLT.pdf

## Central Limit Theorem

- Variable X, population mean=100, SD=15
- Samples of size 25 (for example)
  – Sample 1, mean=90
  – Sample 2, mean=115
  – Sample 3, mean=101
  – Sample 4, mean=94
  – .
  – .
  – .

5

## Central Limit Theorem

- Plot sample means (histogram), then:
- The sample means have mean 100
- The sample means have a SD of 15/5=3
- The distribution of sample means would tend to be normal as n gets large.

6

## Central Limit Theorem

- Thus we can combine this normality results from the CLT with standardization to make probabilistic statements about the population mean.

7

## Scientific Method

- Hypothesis testing is rooted in the scientific method.
- To prove something: assume that the opposite (complement) is true, and look for contradictory evidence.
- We then look for evidence to disprove the complement.

8

## Hypothesis Testing

- Hypothesis
  - A prediction
    - about a **population** parameter or
    - about the relationship between parameters from 2 or more populations
  - Formulate hypotheses BEFORE beginning an experiment
- Hypothesis Testing
  - Used to evaluate hypotheses

9

## Hypotheses

- Null hypotheses
  - $H_0$: No effect or no difference
  - Assume true but look for evidence to disprove
- **Alternative hypotheses**
  - **$H_A$:** Presence of an effect or difference
  - Try to prove
- Well-formulated hypotheses are quantifiable and testable
- Need to think about directionality (e.g., what are you trying to prove?)

10

## Hypothesis Testing

- Make a decision to reject (or fail to reject) the $H_0$ by comparing what is observed to what is expected if the $H_0$ is true (I.e., p-values)
- Note that the hypotheses concern (unknown) population parameters. We make decisions about these parameters based on the (observable) sample statistics.

11

## Hypothesis Testing vs. Court Trials

- Analogous to a court trial:
  - People are assumed innocent until proven guilty
    - $H_0$: Innocent
    - $H_A$: Guilty
  - If VERDICT = Guilty (i.e., reject $H_0$), then:
    - We can say that enough evidence was found to prove guilt (to some standard of proof)

12

## Hypothesis Testing vs. Court Trials

- If VERDICT = Not Guilty (i.e., do not reject $H_0$), then:
  - We cannot say that we have proven innocence
  - We can say that we failed to find enough evidence to prove guilt. (There is a subtle but important difference between the two).
- "Absence of evidence is not evidence of absence."

13

## Hypothesis Testing

- Evidence is used to disprove hypotheses
- We can prove the alternative hypothesis to some standard of proof.
- We cannot prove the null hypothesis (we can only fail to reject it) – we cannot prove what we have already assumed to be true.
- Thus we do not "accept" $H_0$. We simply fail to reject it.

14

## Hypothesis Testing

- The gap between the population and the sample always requires a leap of faith.
- The only statements that can be made with certainty are of the form:
  - "our conclusions fit the data"
  - "our data are consistent with the theory that …"

15

## P-values

- Used to make decisions concerning hypotheses
- Probability (Conditional)
  - Ranges from 0-1
- A function of sample sizes
  - Careful when N is very large (everything is significant)
  - Careful when N is very small (everything is non-significant)
- Based upon the notion of repeated sampling

16

## P-values

- Probability of observing a result (data) as extreme or more extreme than the one observed (due to chance alone) given that $H_0$ is true.
- $P(\text{data}|H_0)$
- It is NOT $P(H_0|\text{data})$

17

## P-values

- Thus if a p-value is small, then the probability of seeing data like that observed is very small if $H_0$ is true.
- In other words, we are either:
  - Seeing a very rare event (due to chance), or
  - $H_0$ is not true

18

## P-values

- P-value is small → reject $H_0$
- P-value is not small → fail to reject $H_0$
- What is "small"
  - Arbitrary (but strategic) choice by investigator
  - Often 0.05 (or 0.10 or 0.01)

19

## P-values

- Although commonly reported, p-values are controversial in some circles
  - Epidemiology
  - Bayesian statistics
  - vs. estimation
  - Draws an arbitrary line in the sand
    - P=0.051 vs. p=0.049

20

## P-values

- Random variable that varies from sample to sample.
- Not appropriate to compare p-values from difference experiments
- Do not reflect the strength of a relationship
- Do not provide information concerning the magnitude of the effect.

21

| | | Truth | |
|---|---|---|---|
| | | $H_o$ True | $H_o$ False |
| Test Result | Reject $H_o$ | Type I error (✂) | correct |
| | Do not reject $H_o$ | correct | Type II error (👓) |

22

## ✂ Error

- ✂
  - =P(making a Type I error|null hypothesis is true)
  - Probability of rejecting when we should not reject
  - = significance level
  - = Type I error
  - This is under our control
    - Conventionally, we often choose ✂=0.05

23

## 👓 Error

- Power = 1-👓
  - =P(not making a Type II error|$H_A$ is true)
  - Probability of rejecting when we should reject
  - Control with sample size
    - Larger sample size → higher power

24

## Determining ✂ and 👓

- ✂ is chosen by the investigator
- 👓 is chosen by the investigator in prospectively designed studies (e.g., clinical trials) where sample size is determined by the investigator.
- If the investigator does not control the sample size, then 👓 is not “controlled-for” and is simply a result of the sample size

25

## Determining ✂ and 👓

- Weigh the costs of making ✂ versus 👓 error
  - ✂ and 👓 are inversely related
  - Examples
    - Car alarms
    - Biosurveillance
    - Sensitivity/Specificity trade-off

26

## Statistical vs. Practical Significance

- Statistical significance does not imply practical relevance.
  - Results should be both: (1) statistically and (2) practically significant in order to influence policy.
    - Example: A drug may induce a statistically significant reduction in blood pressure. However, if this reduction is 1 mmHg in your systolic BP, then it is not a useful (practical and clinically relevant) drug.

27

## Hypothesis Testing

- Test only relevant hypotheses.
- Multiple testing → multiple places for error
  - Example: Food supplement for sexual dysfunction
- Specify hypotheses *a priori*
  - Post-hoc: some studies use data to identify hypotheses (hypothesis generating)
    - This is done as an exploratory analyses (but not as a confirmatory analyses)

28

## Example - Interpretation

- NCEP recommends LDL < 100 mg/dL (for that for patients with two or more CHD risk factors
- A new drug is being developed to (hopefully) lower LDL levels in such patients
- A group of 25 patients with high LDL levels and CHD risk factors are recruited for a study of the new drug

29

## Example - Interpretation

- The LDL level of each patient is measured before the drug is given and again after 12 weeks of treatment with the drug
- The difference (change=post-pre, ✆) is calculated for each patient

30

## Example - Interpretation

- The study investigators made the following statement: “data from this study are consistent with the hypothesis that this drug lowers LDL in patients with elevated LDL and >2 CHD risk factors (mean(✆)=-30, p=0.008)”.
- What does this mean?

31

## Example - Interpretation

- Wherever there is a p-value, there is a hypotheses also. Determine $H_0$ and $H_A$.
- In this case:
  - $H_0$: mean(✆)=0; I.e., the drug has no effect
  - $H_A$: mean(✆)⑤0
  - Although these hypotheses are not stated explicitly, they are implied by the language (this can be tricky when reading journal articles)

32

## Example - Interpretation

- Assuming that the drug has no effect (e.g., $H_0$ is true), then there are two possible reasons for observing Mean(✆)=-30
  - Chance
  - The drug truly decreases LDL
- P-values separate the two; indicating the probability that chance is producing the observed data

33

## Interpreting the p-value

- Assume that $H_0$ is true (i.e., the drug has no effect)
  - Hypothetically repeat the study a very large number of times and we observe the changes

34

## Interpreting the p-value

- Trial 1, Mean(✆)=-25
- Trial 2, Mean(✆)=10
- Trial 3, Mean(✆)=0
- Trial 4, Mean(✆)=-5
- Trial 5, Mean(✆)=18
- .
- .
- .

35

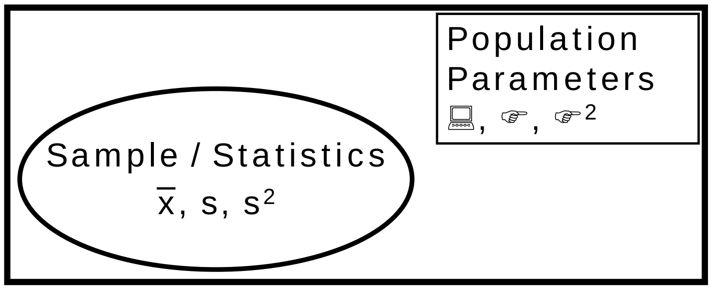
## Interpreting the p-value

- Note some differences are small, some are large, some are near 0. This is called ***sampling variability***.
  - Differences between the trials are simply due the the random samples chosen for each study (e.g., observed differences between studies are simply a function of who enrolled into the study. Remember, we assume $H_0$ is true … the drug has no effect.)

36

## Remember the Big Picture

Populations and Samples




37

## Interpreting the p-value

- P=0.008 says that if the drug has no effect (i.e., $H_0$ is true), the the probability of observing a result as extreme or more extreme that the one we observed in this study is 0.008
  - 0.8% of the hypothetical trials on our list would have differences that large

38

## Interpreting the p-value

- We often reject at 0.05 (our standard of proof)
  - If $p<0.05$ then the probability of observing a result this extreme if the drug had no effect, is too small (a very rare event)
- Thus we reject the null hypothesis that the drug has no effect and conclude that evidence suggests a drug effect
- Note that when the drug truly has no effect, we will incorrectly conclude that it does 5% of the time.

39

## Example

- The Neuropsychological and Neurologic Impact of HCV Co-Infection in HIV-Infected Subjects
- Insert HIV_HCV.pdf

40

## Example

- ACTG A5087: A prospective, multicenter, randomized trial comparing the efficacy and safety of fenofibrate versus pravastatin in HIV-infected subjects with lipid abnormalities
- Insert A5087.pdf



## Hypothesis Testing – Step by Step

1. State $H_0$ (i.e., what you are trying to disprove)
2. State $H_A$
3. Determine ✂ (at your discretion)
4. Determine the test statistic and associated p-value.
5. Determine whether to reject $H_0$ or fail to reject $H_0$