Estimation of Statistical Parameters

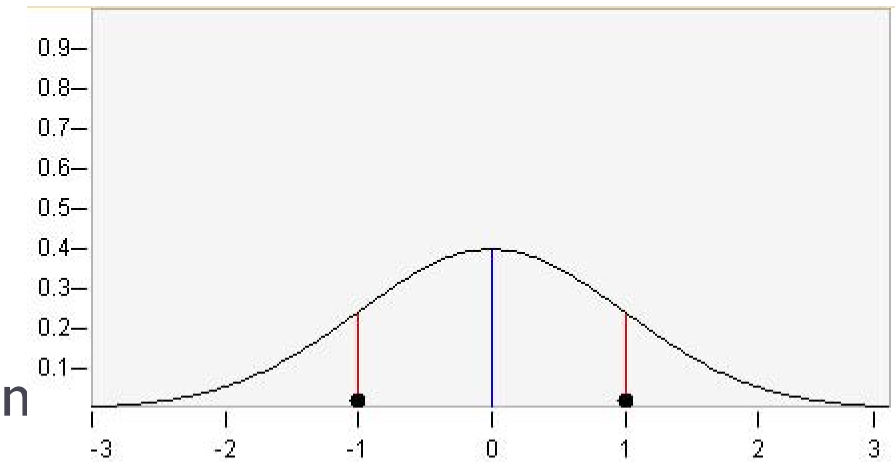
Sorana D. BOLBOACĂ

OUTLINE

- Estimation & Confidence Intervals:
  - Normality analysis based on statistical terms
  - Confidence intervals for means
  - Confidence intervals for frequencies
- Hypothesis testing:
  - Concept and Practice

Estimation & Confidence Intervals

- Normal distribution:
  - Gaussian distribution
  - Symmetric
  - Not skewed
  - Unimodal
  - Described by two paran
    - Probability density function:
  - μ & σ are parameters
  - μ = mean
  - σ = standard deviation
  - π, e = constants



$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Estimation of Confidence Intervals

- Normal distribution: Why do we use it!
  - Many biological variables follow a normal distribution
  - The normal distribution is well-understood, mathematically

- Punctual estimation
  - Is a value for estimated theoretical parameter
    - m (sample mean) is a punctual estimation of μ (population mean)
  - Is influenced by the fluctuations from sampling
  - Could be very far away from the real value of the estimated parameter

Why Confidence Intervals?

- It is recommended to estimate a theoretical parameter by using a range of value not a single value
  - It is called confidence intervals
  - The estimated parameter belong to the confidence intervals with a high probability.

Definitions

- A range around the sample estimate in which the population estimate is expected to fall with a specified degree of confidence, usually 95% of the time at a significance level of 5%.
  - **P[lower critical value < estimator < higher critical value] = 1-α**
  - α = significance level
- The range defined by the critical values will contains the population estimator with a probability of 1-α
- It is applied when variables are normal distributed!

CONFIDENCE INTERVAL: INTERPRETATION

- If 0 is contains by the confidence intervals for the distance between an observed and theoretical mean, the difference between the two investigated means is 0.
- If 0 is NOT contains by the confidence intervals for the distance between an observed and theoretical mean, the difference between the two investigated means is NOT 0.

Confidence Interval: Interpretation

- Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter 95% of the time.
- The confidence interval represents values for the population parameter for which the difference between the parameter and the observed estimate is not statistically significant at the 5% level.

Confidence Interval

- It is calculated taking into consideration:
  - The sample or population size
  - The type of investigated variable (qualitative OR quantitative)

- Formula of calculus comprised two parts
  - One estimator of the quality of sample based on which the population estimator was computed (standard error)
    - Standard error: is a measure of how good our best guess is.
    - Standard error: the bigger the sample, the smaller the standard error.
    - Standard error: i always smaller than the standard deviation
  - Degree of confidence (Z_α score)

- It is possible to be calculated for any estimator but is most frequent used for mean

Confidence Interval for Mean

- Standard error of mean is equal to standard deviation divided by square root of number of observations:
  - If standard deviation is high, the chance of error in estimator is high
  - If sample size is large, the chance of error in estimator is small

$$\left[\overline{X} - Z_{\alpha}\frac{s}{\sqrt{n}}, \overline{X} + Z_{\alpha}\frac{s}{\sqrt{n}}\right] \qquad \left[m - Z_{\alpha}\frac{s}{\sqrt{n}}, m + Z_{\alpha}\frac{s}{\sqrt{n}}\right]$$

Confidence Interval for Mean

- Z_α score is the score of normal distribution for a mean of 0 and a standard deviation of 1. Any distribution could be transform in a Z score using the following formula.

$$Z = \left(X - \overline{X}\right)/s \qquad Z = \left(X - m\right)/s$$

  - Lower confidence limit is smaller than the mean
  - Upper confidence limit is higher than the mean
  - **For the 95% confidence intervals: $Z_{5\%}$ = 1.96**
  - For the 99% confidence intervals : $Z_{1\%}$ = 2.58

$$\left[m - Z_\alpha \frac{s}{\sqrt{n}}, m + Z_\alpha \frac{s}{\sqrt{n}}\right] \qquad \left[\overline{X} - Z_\alpha \frac{s}{\sqrt{n}}, \overline{X} + Z_\alpha \frac{s}{\sqrt{n}}\right]$$

Confidence Interval for Mean

- The mean of blood sugar concentration of a sample of 121 patients is equal to 105 and the variance is equal to 36.
- Which is the confidence levels of blood sugar concentration of the population from which the sample was extracted?
- Use a significance level of 5% (Z = 1.96). It is considered that the blood sugar concentration is normal distributed.

- n = 121
- s^2 = 36
- s = 6
- m = 105

$$\overline{X} = 105$$

$$\left[105 - 1.96\frac{6}{\sqrt{121}}; 105 + 1.96\frac{6}{\sqrt{121}}\right]$$

- [105-1.07; 105+1.07]
- [103.93; 106.07]
- [104;106]

Comparing Means using Confidence Intervals

http://www.biomedcentral.com/content/pdf/1471-2458-12-1013.pdf

Table 1 Living conditions of the MS-MV and the immigrant population (CASEN survey 2006)

	IMMIGRANT POPULATION 1% total sample, n = 154 431 weighted population (1877 real observations)		**MS-MV GROUP 0.67% total sample, n = 108 599 weighted population (1477 real observations)**	
	% or mean	**95% CI**	**% or mean**	**95% CI**
DEMOGRAPHICS				
Mean age**	$X = 33.41$	31.81–35.00	X = 26.13	23.41–28.26
Age categories:				
<16 years old**	13.60	11.29–16.28	45.25	39.53–51.10
16-65 years old**	79.08	75.92–81.93	47.26	41.64–52.94
>65 years old	7.32	5.33–9.97	7.49	5.31–10.46
Sex (female = 1)	45.21	41.74–48.72	51.27	47.99–55.41
Marital status:				
Single**	45.81	42.06–49.62	64.30	59.36–68.95
Married**	45.49	41.66–49.36	29.39	25.09–34.10

Confidence Interval for Frequency

- Could be computed if:
  - n*f > 10, where n = sample size, f = frequency

$$\left[f - Z_{\alpha}\sqrt{\frac{f(1-f)}{n}} ; f + Z_{\alpha}\sqrt{\frac{f(1-f)}{n}} \right]$$

Confidence Interval for Frequency

- We are interested in estimating the frequency of breast cancer in women between 50 and 54 years with positive family history. In a randomized trial involving 10,000 women with positive history of breast cancer were found 400 women diagnosed with breast cancer.
- What is the 95% confidence interval associated frequently observed?

- f = 400/10000 = 0.04

$$\left[0.04-1.96\sqrt{\frac{0.04\cdot 0.96}{10000}};\,0.04+1.96\sqrt{\frac{0.04\cdot 0.96}{10000}}\right]$$

- [0.04-0.004; 0.04+0.004]
- [0.036; 0.044]

Comparing ORs using Confidence Intervals

http://www.biomedcentral.com/content/pdf/1471-2458-12-1013.pdf

Table 3 Odds Ratio (OR) of presenting **any disability and any chronic condition or cancer**, adjusted by different sets of factors separately (CASEN survey 2006)

	ANY DISABILITY				**ANY CHRONIC CONDITION OR CANCER**			
	International immigrants		**MS-MV**		**International immigrants**		**MS-MV**	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
DEMOGRAPHICS								
Age	**1.04***	**1.02-1.06**	**1.04***	**1.02-1.06**	**1.05***	**1.02-1.08**	**1.02***	**1.01-1.04**
Sex (female = 1)	0.56	0.25-1.25	**0.39***	**0.20-0.75**	**2.78****	**1.26-6.71**	1.05	0.46-2.36

REMEMBER!

- Correct estimation of a statistical parameter is done with confidence intervals.
- Confidence intervals depend by the sample, size and standard error.
- The confidence intervals is larger for:
  - High value of standard error
  - Small sample sizes

Hypothesis Testing

Objective:

- Understand the principles of hypothesis-testing
- To be able to interpret **P** values correctly
- To know the steps needed in application of a statistical test

DEFINITIONS

- **Statistical hypothesis test** = a method of making statistical decisions using experimental data.
- A result is called **statistically significant** if it is unlikely to have occurred by chance.
- Statistical hypothesis = an assumption about a population parameter. This assumption may or may not be true.
- Clinical hypothesis = a single explanatory idea that helps to structure data about a given client in a way that leads to better understanding, decision-making, and treatment choice.

  [Lazare A. The Psychiatric Examination in the Walk-In Clinic: Hypothesis Generation and Hypothesis Testing. Archives of General Psychiatry 1976;33:96-102.]

DEFINITIONS

- **Clinical hypothesis**:
  - A proposition, or set of propositions, set forth as an explanation for the occurrence of some specified group of phenomena, either asserted merely as a provisional conjecture to guide investigation (working hypothesis) or accepted as highly probable in the light of established facts
  - A tentative explanation for an observation, phenomenon, or scientific problem that can be tested by further investigation.
  - Something taken to be true for the purpose of argument or investigation; an assumption.

From Probability to Hypothesis Testing

Population:

The set of all individuals of interest (e.g. all women, all college students)

Sample:

A subset of individuals selected from the population from whom data is collected

probability

From Probability to Hypothesis Testing

What we Learned from Probability

1) The mean of a sample can be treated as a random variable.

2) By the central limit theorem, sample means will have a normal distribution (for n > 30) with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

3) Because of this, we can find the probability that a given population might randomly produce a particular range of sample means.

$$P(\overline{X} > something) = P(Z > something) = \text{Use standard table}$$

Inferential Statistics

Inferential Statistics

Inferential statistics

Population:

The set of all individuals of interest (e.g. all women, all college students)

Sample:

A subset of individuals selected from the population from whom data is collected

Inferential Statistics

- Once we have got our sample
- The key question in statistical inference:
  - Could random chance alone have produced a sample like ours?
- Distinguishing between 2 interpretations of patterns in the data:

Inferential statistics separates

Random Causes:

Fluctuations of chance

Systematic Causes Plus Random Causes:

True differences in the population

Bias in the design of the study

Reasoning of Hypothesis Testing

1. Make a statement (the null hypothesis) about some unknown population parameter.

2. Collect some data.

3. Assuming the null hypothesis is TRUE, what is the probability of obtaining data such as ours? (this is the "p-value").

4. If this probability is small, then reject the null hypothesis.

Hypothesis Testing: Step 1

- State the research question in terms of a statistical hypothesis
  - Null hypothesis (the hypothesis that is to be tested): abbreviated as H_0
    - Straw man: "Nothing interesting is happening"
  - Alternative hypothesis (the hypothesis that in some sense contradicts the null hypothesis): abbreviated as H_a or H_1
    - What a researcher thinks is happening
    - May be one- or two-sided

Hypothesis Testing: Step 1

- Hypotheses are in terms of population parameters

One-sided	Two-sided
H_0: $\mu=110$ $H_{1/a}$: $\mu < 110$ OR $H_{1/a}$: $\mu > 110$	H_0: $\mu = 110$ $H_{1/a}$: $\mu \neq 110$

Hypothesis Testing: Step 2

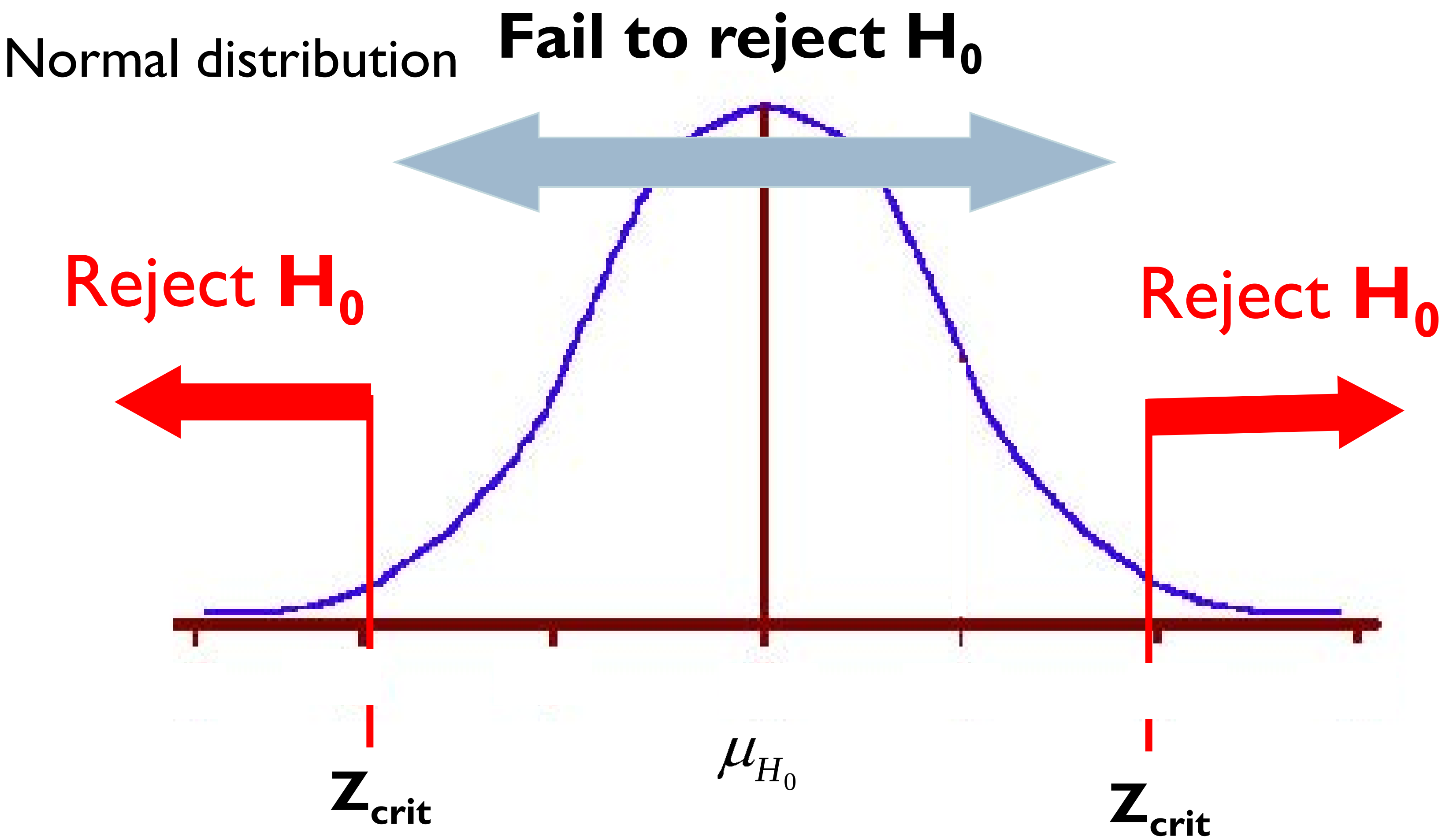
- Set decision criterion:
  - Decide what p-value would be "too unlikely"
  - This threshold is called the alpha level.
  - When a sample statistic surpasses this level, the result is said to be significant.
  - Typical **alpha levels** are **0.05** and **0.01**.
- Alpha levels (level of significance) = probability of a type I error (the probability of rejecting the null hypothesis even that H0 is true)
- The probability of a type II error is the probability of accepting the null hypothesis given that H_1 is true. The probability of a Type II error is usually denoted by β.

Hypothesis Testing: Step 3

- Setting the rejection region:
  - The range of sample mean values that are "likely" if H_0 is true.
  - If your sample mean is in this region, retain the null hypothesis.

  - The range of sample mean values that are "unlikely" if H_0 is true.
  - If your sample mean is in this region, reject the null hypothesis

Hypothesis Testing: Step 3

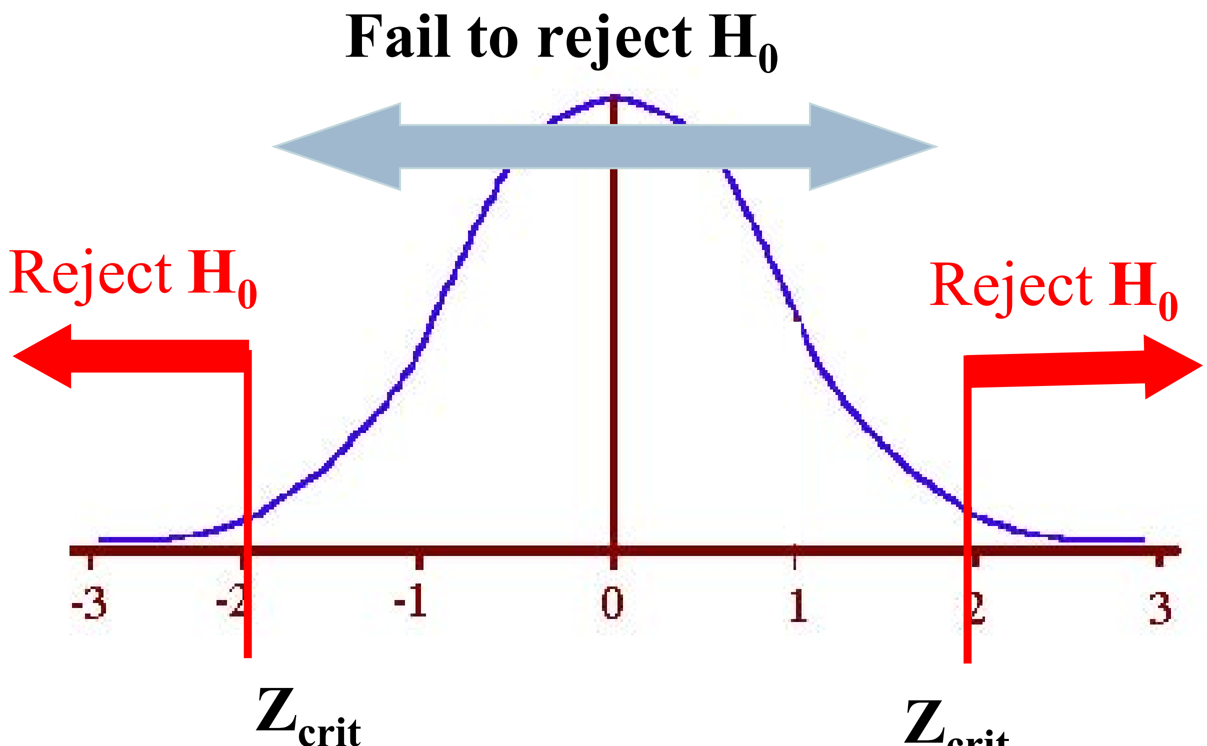


HYPOTHESIS TESTING: STEP 4

- Compute sample statistics
- A **test statistic** (e.g. Z_{test}, T_{test}, or F_{test}) is information we get from the sample that we use to make the decision to reject or keep the null hypothesis.

- A test statistic converts the original measurement (e.g. a sample mean) into units of the null distribution (e.g. a z-score), so that we can look up probabilities in a table.

Hypothesis Testing: Step 4



- If we want to know where our sample mean lies in the null distribution, we convert X-bar to our test statistic Z_{test}
- If an observed sample mean were lower than z=-1.65 then it would be in a critical region where it was more extreme than 95% of all sample means that might be drawn from that population

Hypothesis Testing: Step 5

- State the test conclusion:
  - If our sample mean turns out to be extremely unlikely under the null distribution, maybe we should revise our notion of μ_{H0}
  - We never really "accept" the null. We either reject it, or fail to reject it.

Steps in Hypothesis Testing

1. Step 1: State hypothesis (H_0 and H_1/H_a)
2. Step 2: Choose significance level
3. Step 3: Setting the regression region
4. Step 4: Compute test statistic (Z_{test}) and get a p-value
5. Step 5: Make a decision

One- vs. Two-tailed Tests

- In theory, should use one-tailed when
  1. Change in opposite direction would be meaningless
  2. Change in opposite direction would be uninteresting
  3. No rival theory predicts change in opposite direction

- By convention/default in the social sciences, two-tailed is standard

- Why? Because it is a more stringent criterion (as we will see). A more conservative test.

One- vs. Two-tailed Tests

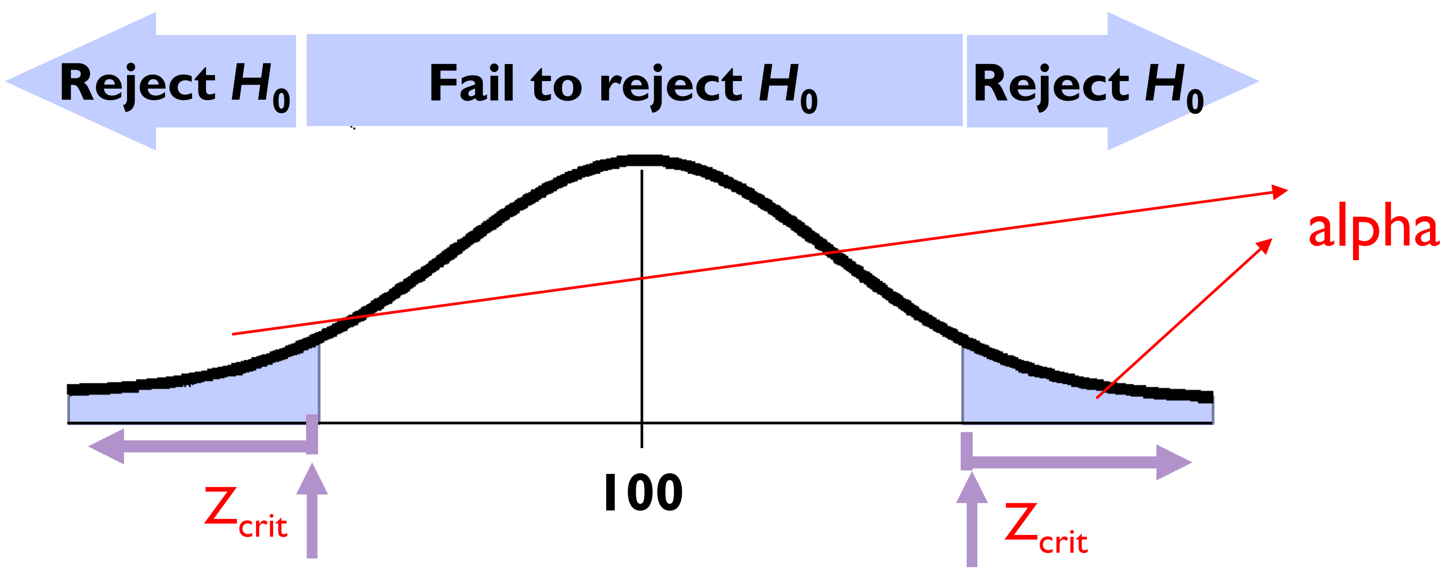
- H_a is that μ is *either* greater or less than μH0
  - H_a: μ ≠ μH0

- α is divided equally between the two tails of the critical region

Two-Tailed Hypothesis Testing

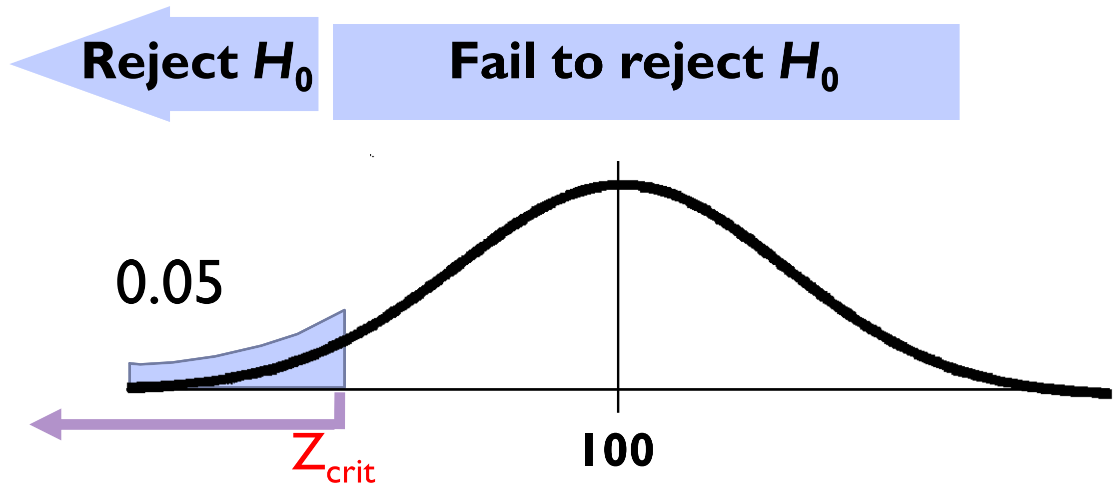
H_0: μ = 100

H_1: μ ≠ 100



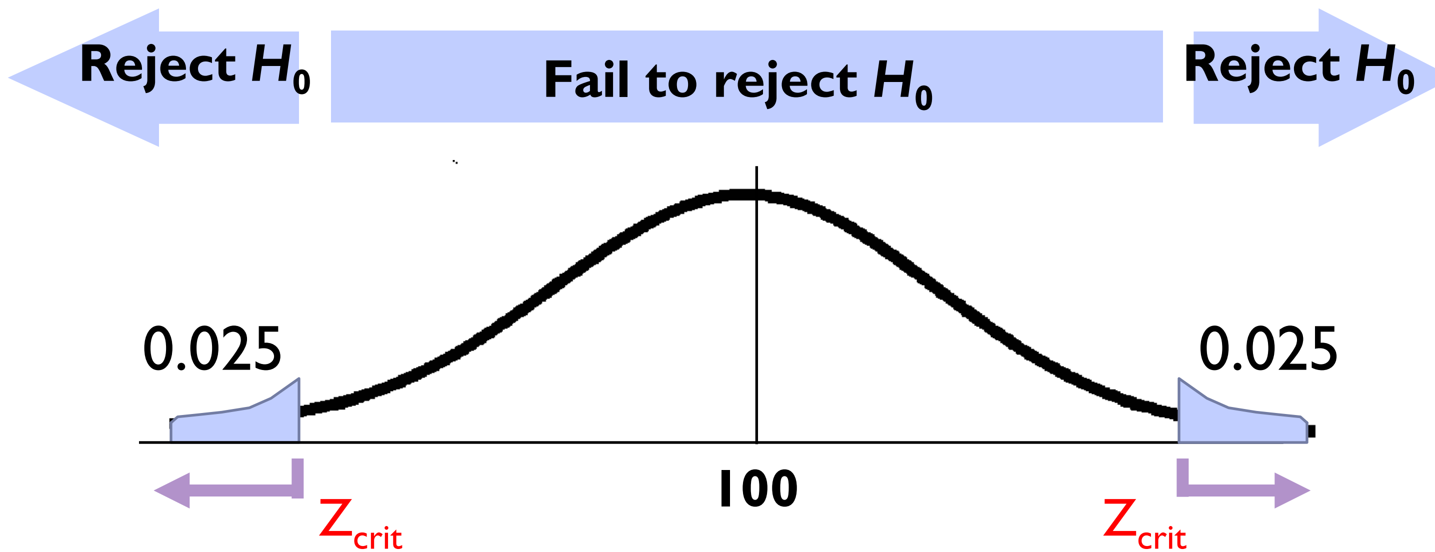
Values that differ significantly from 100

One tail



Values that differ "significantly" from 100

Two tails



Values that differ significantly from 100

Difference between P values and Confidence Intervals

- A **P** value measures the strength of evidence against the null hypothesis.
- A **P** value is the probability of getting a result as, or more, extreme if the null hypothesis were true.
- It is easy to compare results across studies using **P** values
- **P** values are measures of **statistical significance**
- Confidence intervals give a plausible range of values in clinically interpretable units
- Confidence intervals enable easy assessment of **clinical significance**

Relation of Confidence Intervals with Hypothesis Testing

- A general purpose approach to constructing confidence intervals is to define a $100(1-\alpha)\%$ confidence interval to consist of all those values θ_0 for which a test of the hypothesis $\theta=\theta_0$ is not rejected at a significance level of $100\alpha\%$.
  - Such an approach may not always be available since it presupposes the practical availability of an appropriate significance test.
  - Naturally, any assumptions required for the significance test would carry over to the confidence intervals.

Relation of Confidence Intervals with Hypothesis Testing

- It may be convenient to make the general correspondence that parameter values within a confidence interval are equivalent to those values that would not be rejected by an hypothesis test, but this would be dangerous.
- In many instances the confidence intervals that are quoted are only approximately valid, perhaps derived from "plus or minus twice the standard error", and the implications of this for the supposedly corresponding hypothesis tests are usually unknown.

SUMMARY

- Estimation of statistical parameters:
  - Confidence intervals for means
  - Confidence intervals for frequencies
- Hypothesis testing:
  - Concept and Practice